Experimental Research in Evolutionary Computation

Thomas Bartz-Beielstein
Mike Preuss

Algorithm Engineering
Universität Dortmund

July, 9th 2006

Overview

1. Introduction
   - Goals
   - History
2. Comparison
   - Observed Significance Plots
   - Beyond the NFL
3. Sequential Parameter Optimization (SPO)
   - Similar Approaches
   - Basics
   - Overview
   - Models
   - Heuristic
4. SPO Toolbox (SPOT)
   - Demo
   - Community
   - Discussion
5. Parametrized Performance
   - Parametrized Algorithms
   - Parameter Tuning
   - Performance Measuring
6. Reporting and Visualization
   - Reporting Experiments
   - Visualization

Scientific Goals?

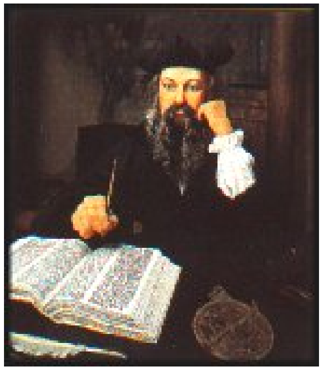


Figure: Nostradamus

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

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

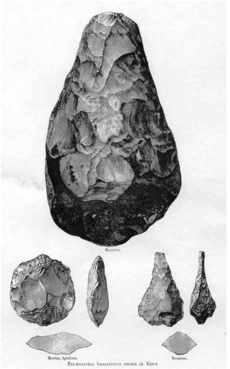
(RG-3) *Conjecture.* Good: demonstrate performance. Better: explain and understand performance

(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

Population	Dimension	Generation	Fitness
20	10	1000	96,1725
20	20	1500	214,6764

- 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, ...
- Comparisons mainly on standard benchmark sets
- Questionable assumptions (NFL)

Yesterday: Mean Values and Simple Statistics

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

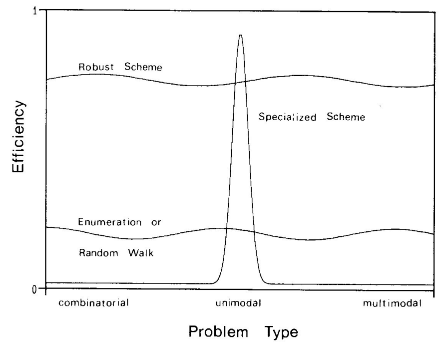


Figure: [Gol89]

Theorem (NFL)

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

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`)
  - 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

Test functions	SGA mean best (std. dev.)	FDGA: OGA mean best (std. dev.)	FDGA: MGA mean best (std. dev.)	FDGA: KGA mean best (std. dev.)	t-value between SGA to the best FDGA	Best algorithm
f_1	8.0607e+000 1.5537e+000	8.5689e+000 1.6671e+000	8.6545e+000 1.5069e+000	8.2722e+000 1.5728e+000	-6.76 *	SGA
f_2	7.8010e-001 4.5832e+000	4.2479e+000 1.3211e+000	3.5444e+000 2.0873e+000	3.5093e+000 1.4978e+000	-4.00 *	SGA
f_3	6.4957e+000 1.8003e+000	9.2728e+000 1.8037e+000	8.6608e+000 1.8614e+000	8.6378e+000 1.9866e+000	-5.65 *	SGA
f_4	1.3506e+002 3.3349e+002	9.2200e+002 2.8070e+002	8.2073e+002 2.5999e+002	8.2272e+002 2.4858e+002	-11.69 *	SGA
f_5	2.7476e-002 3.0828e-002	6.8234e-002 5.4773e-002	8.2052e-002 5.2042e-002	6.2478e-002 5.5991e-002	-3.87 *	SGA
f_6	2.0791e-003 9.1846e-004	2.7050e-005 3.5287e-006	2.5915e-005 3.3210e-006	2.5830e-005 2.7375e-006	15.81 *	FDGA
f_7	2.0791e-003 9.1846e-004	4.3338e-011 7.5496e-012	4.0195e-011 8.0494e-012	4.0062e-011 8.3297e-012	1.91 *	FDGA
f_8	7.1211e+001 7.1211e+001	5.0154e+001 4.1123e+001	5.1774e+001 3.7574e+001	4.0649e+001 4.1068e+001	3.13 *	FDGA
f_9	1.4856e-001 6.2373e-002	5.1283e-002 4.1936e-003	4.6518e-002 1.6727e-002	4.6306e-002 1.2852e-002	11.33 *	FDGA
f_{10}	9.2123e-002 6.1055e-002	7.2324e-002 2.1381e-002	6.4803e-002 2.1804e-002	6.4846e-002 2.4023e-002	2.94 *	FDGA

* The value of t with 49 degree 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

- 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

Definition (p value)

The p value is the probability that the null hypothesis is true. No!

Definition (p value)

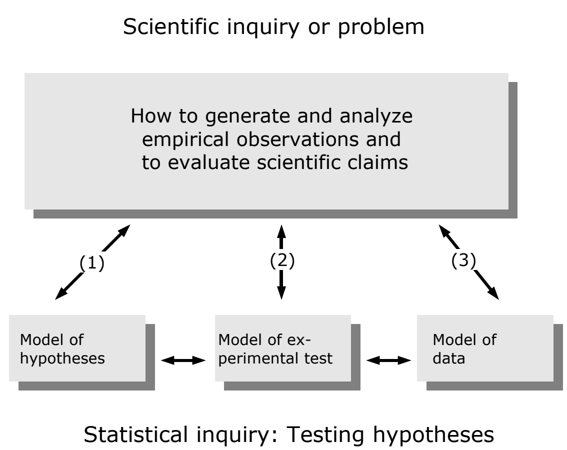
The p value is $p = P\{$ result from test statistic, or greater $|$ null model is true $\}$

- $\Rightarrow$ The p value is not related to any probability whether the null hypothesis is true or false

Tomorrow: Correct Statistics and Correct Conclusions

- 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



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 $\overline{x}$
- Alternative hypothesis $J : \theta > \theta_0$

Definition (Observed significance level)

The observed significance level is defined as

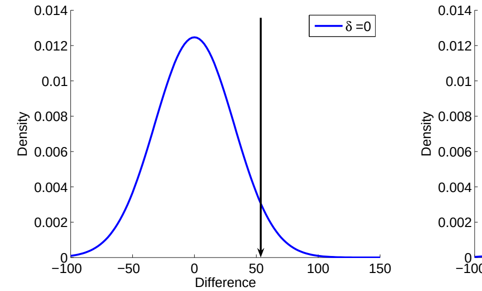
$$\alpha(\overline{x}, \theta) = \hat{\alpha}(\theta) = P(\overline{X} \geq \overline{x} | \theta) \tag{1}$$

Plots of the Observed Significance

- Observed significance level

$$\alpha(\overline{x}, \theta) = \hat{\alpha}(\theta) = P(\overline{X} \geq \overline{x} | \theta)$$

- Observed average $\overline{x} = 51.73$



- Rejection of the null hypothesis

$$H : \theta = \theta_0 = 0$$

by a test T^+ in favor of an alternative

$$J : \theta > \theta_0$$

Then $\hat{\alpha}(\theta) = 0.0530$

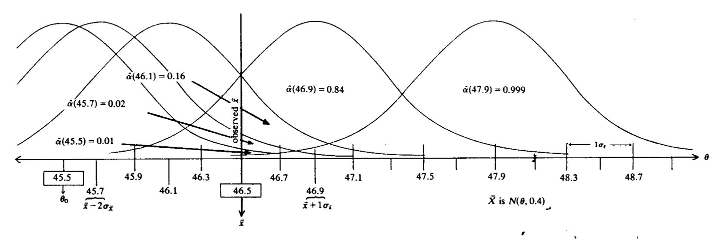
- Interpretation: Frequency of erroneously rejecting H ("there is a difference in means as large as θ_0 or larger") with such an $\overline{x}$

Small α Values

- Rejecting H with a T^+ test with a small size α indicates that $J : \theta > \theta_0$
- If any and all positive discrepancies from θ_0 are scientifically important $\Rightarrow$ small size α ensures that construing such a rejection as indicating a scientifically important θ would rarely be erroneous
- Problems if some θ values in excess of θ_0 are not considered scientifically important
- Small size α does not prevent a T^+ rejection of H from often being misconstrued when relating it to the scientific claim
- $\Rightarrow$ Small α values alone are not sufficient

Largest Scientifically Unimportant Values

- [May83] defines θ_{un} the largest scientifically unimportant θ value in excess of θ_0
- But what if we do not know θ_{un}?
- Discriminate between legitimate and illegitimate construals of statistical results by considering the values of $\hat{\alpha}(\theta')$ for several θ' values



OSL Plots

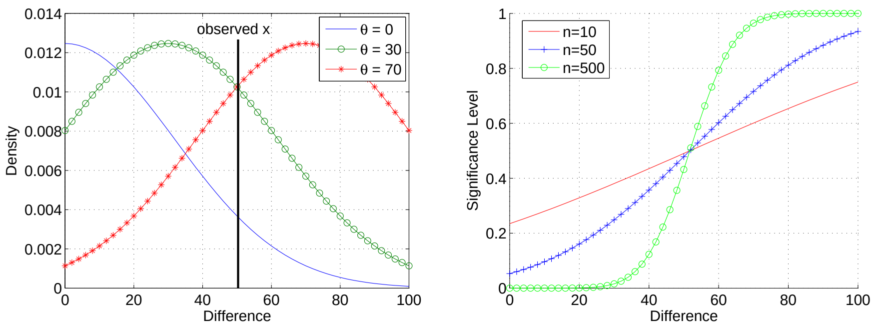


Figure: Plots of the observed difference. *Left*: This is similar to Fig. 4.3 in [May83]. Based on $n = 50$ experiments, a difference $\overline{x} = 51.3$ has been observed, $\hat{\alpha}(\theta)$ is the area to the right of the observed difference $\overline{x}$. *Right*: The $\hat{\alpha}(\theta)$ value is plotted for different n values.

OSL Plots

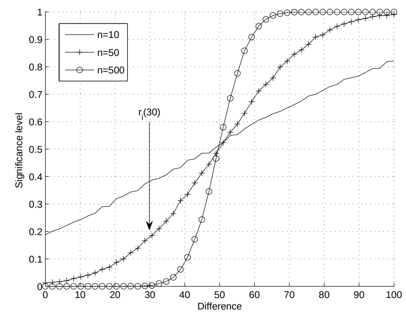


Figure: Same situation as above, bootstrap approach

- 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

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:

  Which algorithm is better?

  to

  What exactly is the algorithm good for?

[1] no free lunch theorem

The Art of Comparison

Efficiency vs. Adaptability

Most existing experimental studies focus on the efficiency of optimization algorithms, but:

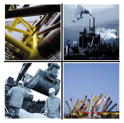
- Adaptability to a problem is not measured, although
- It is known as one of the important advantages of EAs

Interesting, previously neglected aspects:

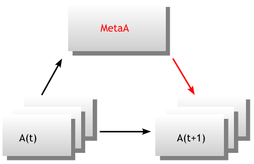
- Interplay between adaptability and efficiency?
- How much effort does adaptation to a problem take for different algorithms?
- What is the problem spectrum an algorithm performs well on?
- Systematic investigation may reveal inner logic of algorithm parts (operators, parameters, etc.)

Similarities and Differences to Existing Approaches

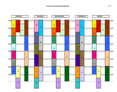
- Agriculture, industry: Design of Experiments (DoE)

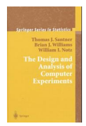
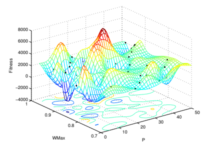
- Evolutionary algorithms: Meta-algorithms



- Algorithm engineering: Rosenberg Study (ANOVA)



- Statistics: Design and Analysis of Computer Experiments (DACE)

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

Algorithm Designs

Example (Algorithm design)

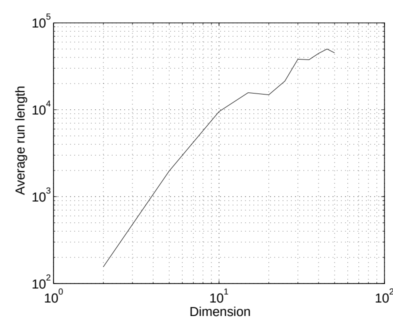
Particle swarm optimization. Set of exogenous strategy parameters

- Swarm size s
- Cognitive parameter c_1
- Social parameter c_2
- Starting value of the inertia weight w_{max}
- Final value of the inertia weight w_{scale}
- Percentage of iterations for which w_{max} is reduced
- Maximum value of the step size v_{max}

Problem Designs

Example (Problem design)

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



- Tuning (efficiency):
  - Given one problem instance $\Rightarrow$ determine improved algorithm parameters
- Robustness (effectivity):
  - Given one algorithm $\Rightarrow$ test several problem instances

SPO Overview

- Pre-experimental planning
- Scientific thesis
- Statistical hypothesis
- Experimental design: Problem, constraints, start-/termination criteria, performance measure, algorithm parameters
- Experiments
- Statistical model and prediction (DACE). Evaluation and visualization
- Solution good enough?
  Yes: Goto step 1
  No: Improve the design (optimization). Goto step 1
- Acceptance/rejection of the statistical hypothesis
- Objective interpretation of the results from the previous step

Statistical Model Building and Prediction

Design and Analysis of Computer Experiments (DACE)

- 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)

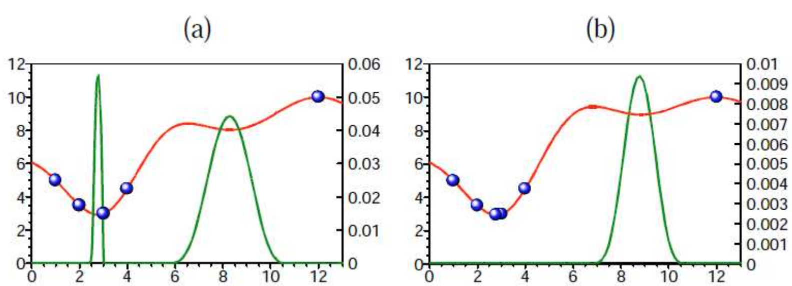


Figure: Axis labels left: function value, right: expected improvement. Source: [JSW98]

(a) Expected improvement: 5 sample points
(b) Another sample point $x = 2.8$ was added

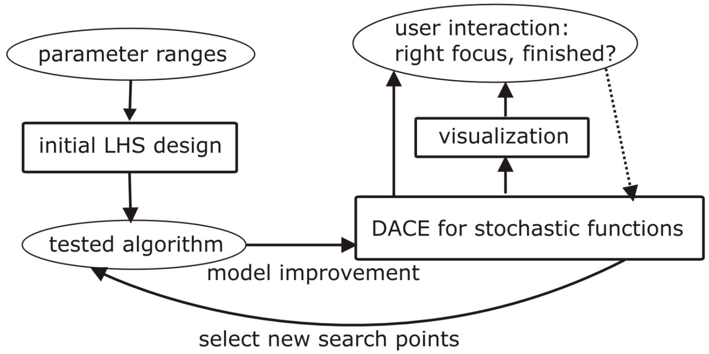
Heuristic for Stochastically Disturbed Function Values

- 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 MSE)
- Budget-concept: Best search point are re-evaluated
- Fairness: Evaluate new candidates as often as the best one

Table: SPO. Algorithm design of the best search points

Y	s	c_1	c_2	w_{max}	w_{scale}	w_{iter}	v_{max}	Conf.	n
0.055	32	1.8	2.1	0.8	0.4	0.5	9.6	41	2
0.063	24	1.4	2.5	0.9	0.4	0.7	481.9	67	4
0.066	32	1.8	2.1	0.8	0.4	0.5	9.6	41	4
0.058	32	1.8	2.1	0.8	0.4	0.5	9.6	41	8

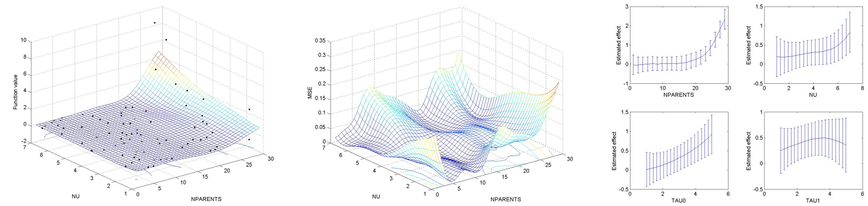
Data Flow and User Interaction



- 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

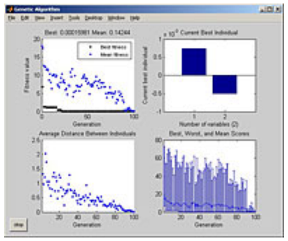
- Sequential Parameter Optimization Toolbox (SPOT)
- Introduced in [BB06]



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

SPOT Community

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



Program	Language	
Evolution Strategy	JAVA, MATLAB	`http://www.springer.com/3-540-32026-1`
Genetic Algorithm and Direct Search Toolbox	MATLAB	`http://www.mathworks.com/products/gads`
Particle Swarm Optimization Toolbox	MATLAB	`http://psotoolbox.sourceforge.net`

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

  $\Rightarrow$ 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, cannot learn too many parameter values at a time, and not necessarily reduce the number of parameters

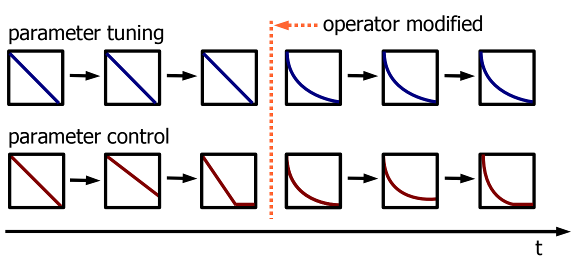
$\Rightarrow$ We can reduce number of parameters, but usually at the cost of either performance or robustness

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

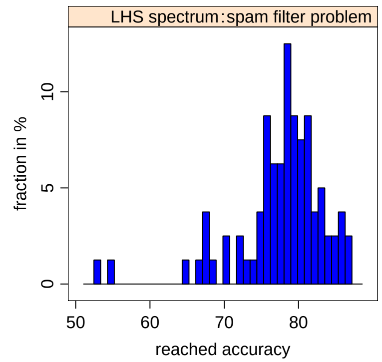


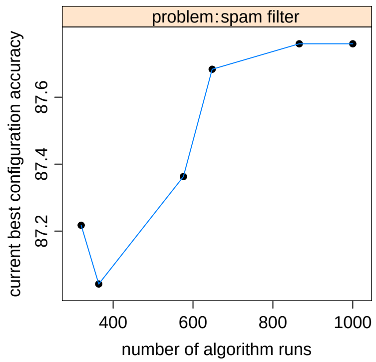
And: How to find meta-parameter values for parameter control?
$\Rightarrow$ Parameter control *and* parameter tuning

Tuning and Comparison

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

- A best configuration from $\{perf(alg(arg_t^{exo}))|1 \leq t \leq T\}$ 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:

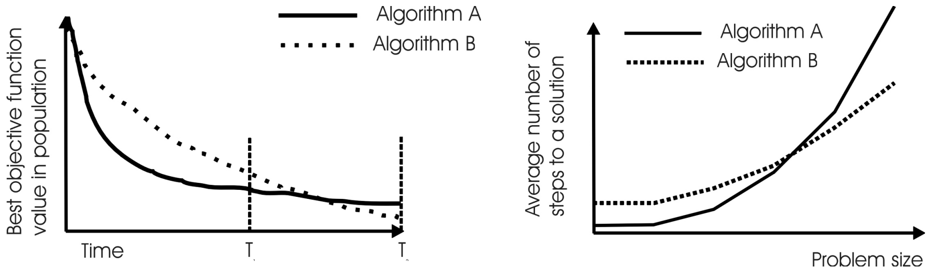
- How much amount to put into tuning (fixed budget, until stagnation)?
- Where shall we be on the spectrum when we compare?
- Can we compare spectra ($\Rightarrow$ adaptability)?

“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)

Aggregated Measures

Especially Useful for Restart Strategies

Success Performances:

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

$$SP1 = \frac{\mathbb{E}(T_A^s)}{p_s} \tag{2}$$

- SP2 [AH05] for different expected lengths, unsuccessful runs are stopped at FE_{max}:

$$SP2 = \frac{1 - p_s}{p_s} FE_{max} + \mathbb{E}(T_A^s) \tag{3}$$

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

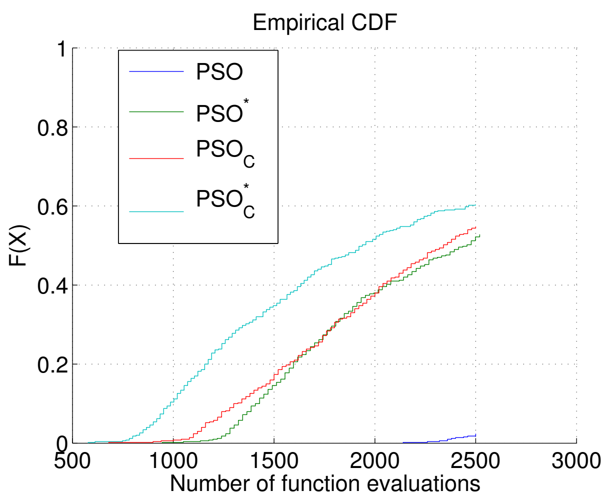
Suggested Report Structure

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

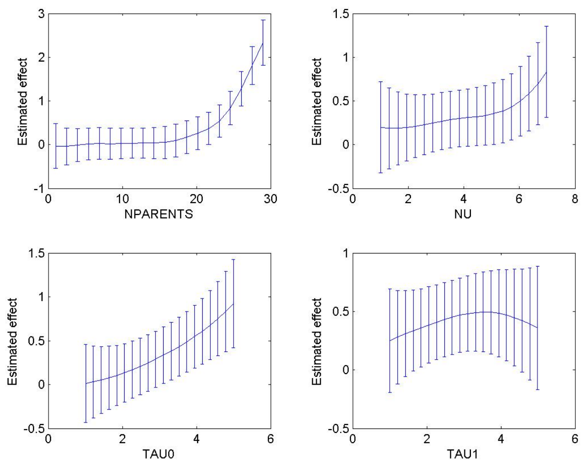
Objective Interpretation of the Results

Comparison. Run-length distribution



(Single) Effect Plots

Useful, but not Perfect

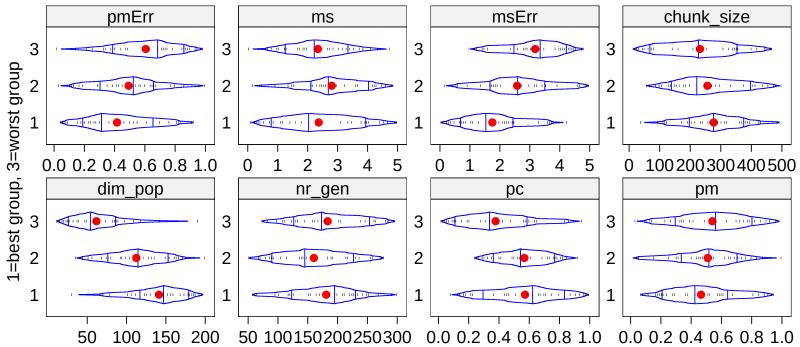


- Large variances originate from averaging
- The τ_0 and especially τ_1 plots show different behavior on extreme values (see error bars), probably distinct (averaged) effects/interactions

One-Parameter Effect Investigation

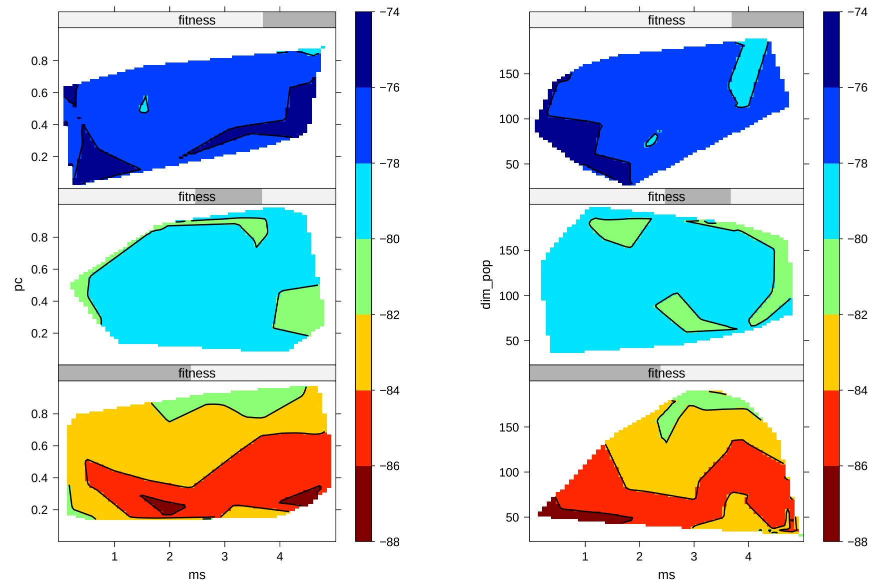
Effect Split Plots: Effect Strengths

- 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` for updates, software, etc.

Anne Auger and Nikolaus Hansen.
Performance Evaluation of an Advanced Local Search Evolutionary Algorithm.
In B. McKay et al., editors, *Proc. 2005 Congress on Evolutionary Computation (CEC'05)*, Piscataway NJ, 2005. IEEE Press.

Thomas Bartz-Beielstein.
Experimental Research in Evolutionary Computation—The New Experimentalism.
Springer, Berlin, Heidelberg, New York, 2006.

Kit Yan Chan, Emin Aydin, and Terry Fogarty.
An empirical study on the performance of factorial design based crossover on parametrical problems.
In *Proceedings of the 2004 IEEE Congress on Evolutionary Computation*, pages 620–627, Portland, Oregon, 20-23 June 2004. IEEE Press.

David E. Goldberg.
Genetic Algorithms in Search, Optimization, and Machine Learning.
Addison-Wesley, Reading MA, 1989.

Nikolaus Hansen and Stefan Kern.

Evaluating the cma evolution strategy on multimodal test functions.
In X. Yao, H.-P. Schwefel, et al., editors, *Parallel Problem Solving from Nature – PPSN VIII, Proc. Eighth Int'l Conf., Birmingham*, pages 282–291, Berlin, 2004. Springer.

D.R. Jones, M. Schonlau, and W.J. Welch.
Efficient global optimization of expensive black-box functions.
Journal of Global Optimization, 13:455–492, 1998.

D. G. Mayo.
An objective theory of statistical testing.
Synthese, 57:297–340, 1983.

D. C. Montgomery.
Design and Analysis of Experiments.
Wiley, New York NY, 5th edition, 2001.

T. J. Santner, B. J. Williams, and W. I. Notz.
The Design and Analysis of Computer Experiments.
Springer, Berlin, Heidelberg, New York, 2003.