Experimental Research in Evolutionary Computation

T. Bartz-Beielstein and M. Preuß

Lehrstuhl für Algorithm Engineering und Systemanalyse
Universität Dortmund

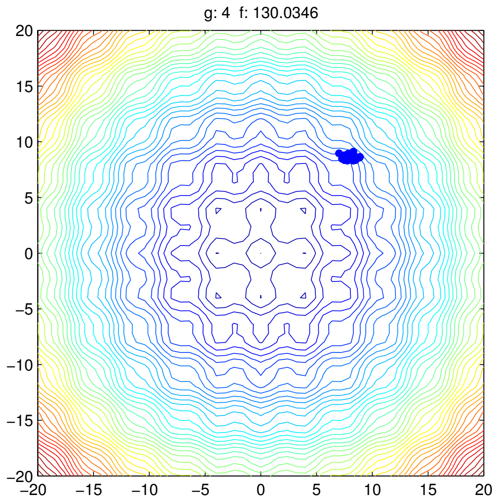
June 25, 2005

Outline

1. Motivation
   - Computer Experiments
   - Existing Approaches
2. The New Experimentalism—Results
   - Sequential Parameter Optimization
   - Example (tuning): Distillation facility
   - Example (comparison): PSO variants
   - Difference Detection and the p-Value

Particle swarm optimization. Simplified and idealized.

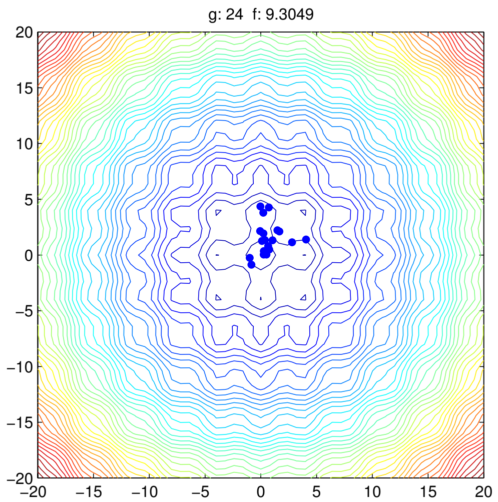
Stage 1: Soon after initialization.



Motivation Computer Experiments

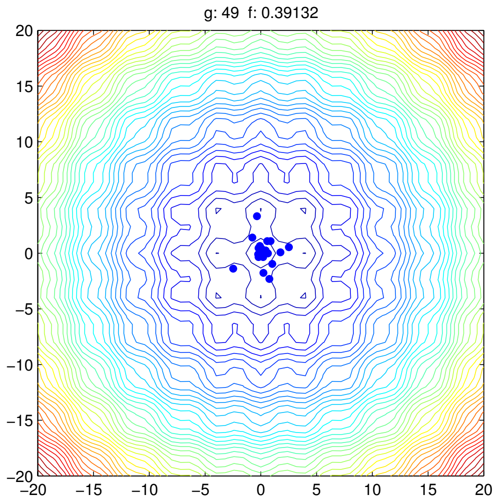
Particle swarm optimization. Simplified and idealized.

Stage 2: Search space exploration.



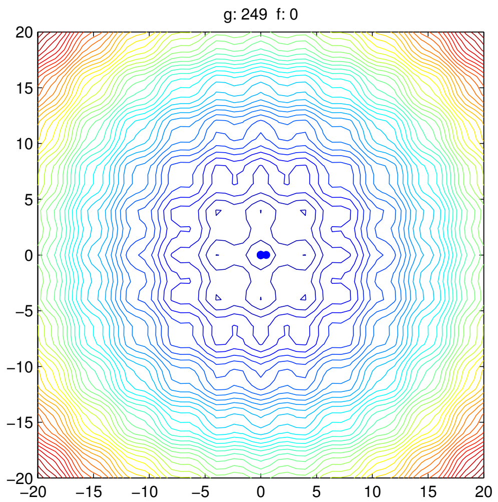
Particle swarm optimization. Simplified and idealized.

Stage 3: Detecting the optimum.



Particle swarm optimization. Simplified and idealized.

Stage 4: The search is finished.



PSO converges very quickly.

- Experimental setup:
  - 4 test functions: Sphere, Rosenbrock, Rastrigin, Griewangk.
  - Initialization: Asymmetrically.
  - Termination: Maximum number of generations.
  - PSO Parameter: Default.
- Results: In table form.
- Conclusion: “Under all the testing cases, the PSO always converges very quickly.”

Table: Mean fitness values for the Rosenbrock function.

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

Scientific goals?

- Why is astronomy considered scientific—and astrology not?

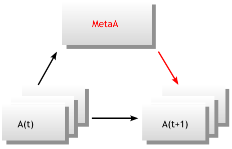
- And what about experimental research in EC?

1. Analyze: Important factors?
2. Compare: Different algorithms. Demonstrate.
3. Conclude: Explain. Understand.
4. Improve: Effectivity. Efficiency.

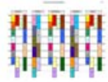
Similarities and differences to existing approaches.

- Agriculture, industry: Design of Experiments (DoE).

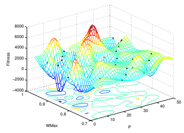

- Evolutionary algorithms: Meta-algorithms.



- Algorithm engineering: Rosenberg Study (ANOVA).



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

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.
6. Statistical model and prediction (DACE). Evaluation and visualization.
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.

Heuristic for stochastically disturbed function values.

1. Latin hypercube design: Relatively many starting points, small number of evaluations.
2. Sequential enhancement, guided by DACE model.
3. Expected improvement: Compromise between optimization (min Y) and model exactness (min MSE).
4. Budget-concept: Best search point are re-evaluated.
5. Fairness: Evaluate new condidates 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

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.
  - Stochstic process.
  - DACE stochastic process model.
- Until now: DACE for deterministic functions, e.g. [Santner et al., 2003].
- New: DACE for stochastic functions.

Weighted distances

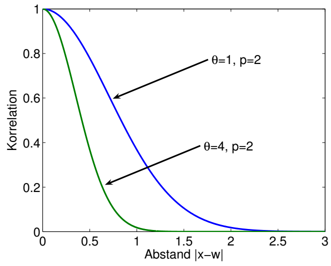
Design and Analysis of Computer Experiments (DACE)

- Euklidean distance: all variables weighted equally.
- DACE uses weighting

$$d(x,w) = \sum_{h=1}^{k} \theta_h |x_h - w_h|^{p_h},$$

$\Longrightarrow$ Correlation of errors in x and w:

$$\text{Corr}\left(\epsilon(x), \epsilon(w)\right) = \exp\left(-d(x,w)\right).$$



- Meaning, activity of x_h : θ_h.
- "Smoothness": p_h.

- Active: Even small distances produce large differences of function values (low correlation).

Expected model improvement

Design and Analysis of Computer Experiments (DACE)

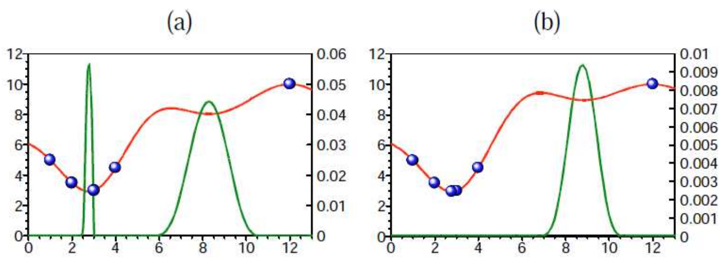
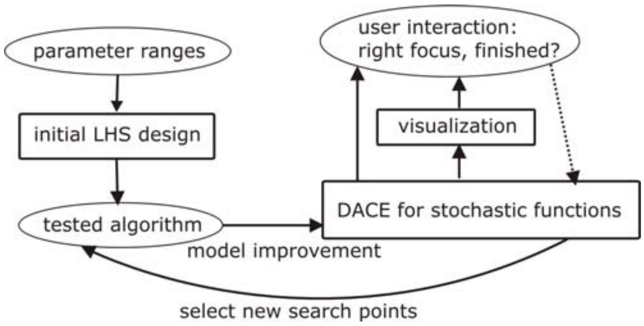


Figure: Axis labels left: function value, right: expected improvement. Source: [Jones et al., 1998].

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

Data flow and user interaction



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

How to use it?

- Possible use cases:

Table: Categorization based on algorithm/problem knowledge

	Problem well-known	Problem unknown
Algorithm well-known	Comparison	Tuning
Algorithm unknown	Tuning	Tuning

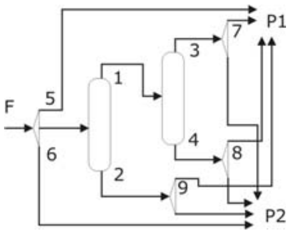
- Two examples:
  1. How to determine an improved parameter setting for an evolution strategy (parameter tuning).
  2. How to compare the performance of two algorithms (comparison).

Example problem (real-world): Distillation facility.

Parameter tuning.

Task: Design a non-sharp separation sequence.

- Separate 3-component feed into two different mixtures.
- 9 real-valued variables control columns and stream dividers.
- 18 (17 hidden) constraints, discretized penalties.



- Shortcut simulator checks physical validity of generated layouts.
- Commercial simulator (slow) evaluates valid layouts.
- Even with shortcut simulator, only few ($\approx 10^5$) evals possible.

The New Experimentalism—Results | Example (tuning): Distillation facility

Pre-experimental planning.

Parameter tuning.

- First tests with default (μ, κ, λ)-ES reveals:
  It is hard to reach valid solutions at all.
- ρ measure $< 10^{-5}$
- Further tests give evidence for non-convex valid search space.
- Manual tuning results in success rates $p(\text{valid}) < 0.1$.

Table: Parameter settings for manual tuning.

Parameter name	tried range
Population size μ	10-20
Maximum age κ	1-20
Selection pressure λ/μ	1-5
Learning rate τ	0.05-0.2

Scientific/Statistical theses.

Parameter tuning.

- Scientific thesis: $\exists$ a parameter set that leads to high success rates (for reaching valid solutions)
- Statistical thesis: $SR(\text{SPO-tuned}) >> SR(\text{man-tuned})$
- In the following: Commercial simulator switched off, finding good quality solutions deferred to a second step.

Experimental design.

Parameter tuning.

- Problem design (mainly fixed):
  - Maximum number of evaluations $\Leftarrow$ 10000 (5-10 mins).
  - Performance measure $\Leftarrow$ MBF
    (penalized invalid solutions always worse than valid ones)
- Algorithm design:

Table: Parameter ranges for SPO (intervals enlarged now).

Parameter name	min	max
Population size μ	10	100
Maximum age κ	1	50
Selection pressure λ/μ	1	10
Learning rate τ	0	1

- Experimental goal: detect parameter region that minimizes the MBF (fitness of invalid points $\geq 10^6$).

Experiments.

Parameter tuning.

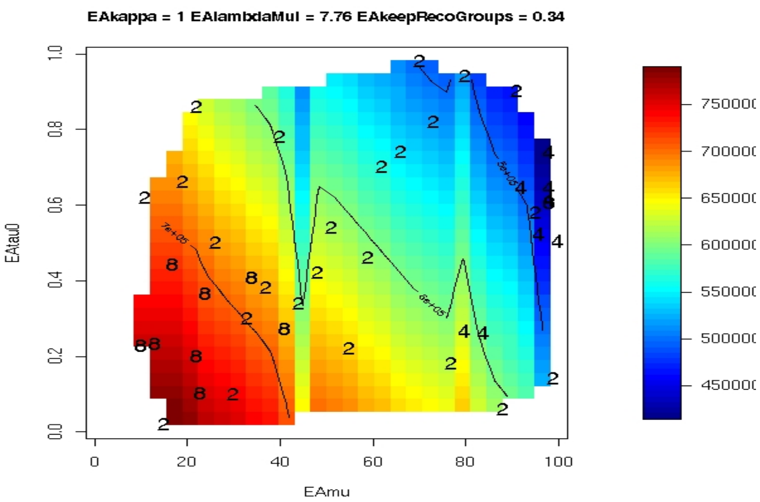
- Test function is costly, try to minimize number of runs.
- 25 initial design points (LHD).
- Initial configurations repeated $r = 2$ times.
- Model enlarged with 1 best, 4 expected best, 4 model improvement points per step.

Table: First entries of result file after initial design + 3 iterations.

κ	μ	λ/μ	τ	*recGrp*	*r*	*conf*	MBF	std.dev.
2	44	7.03	0.34	0.02	2	14	3.2306E+05	1.7635E+04
1	98	7.7576	0.6045	0.3425	8	27	3.2516E+05	3.7345E+04
32	33	9.406	0.3	0.94	2	16	3.2704E+05	301
22	91	6.238	0.9	0.42	2	13	3.3018E+05	3.0601E+04
16	100	5.342	0.5035	0.1695	4	32	3.3048E+05	1.8736E+04
42	95	3.466	0.58	0.22	2	21	3.3644E+05	3.2716E+04
29	55	3.862	0.22	0.26	2	10	3.3916E+05	3.668E+04
12	70	8.614	0.98	0.46	2	22	3.6124E+05	2.9507E+04
1	96	5.8865	0.5215	0.4075	4	26	3.7467E+05	2.6731E+04
28	84	1.09	0.26	0.14	4	23	4.8457E+05	3.1373E+05
19	41	9.8763	0.2724	0.5165	8	39	4.969E+05	3.0772E+05

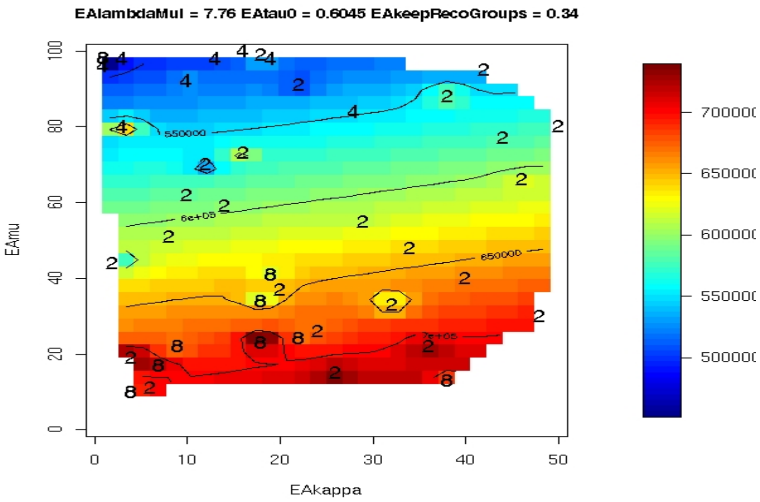
Evaluation and visualization.

Parameter tuning.



Evaluation and visualization.

Parameter tuning.



Acceptance/rejection of the statistical hypothesis.

Parameter tuning.

- We chose configuration 27 ($\mu = 98, \kappa = 1, \lambda/\mu = 7.76, \tau = 0.6$) because it performs well (2nd) and is stable under 8 repeats.
- Verify result: 40 new runs, measure SR.
- SR $\approx$ 65%, significantly better than 10%, no t-test needed.

Objective interpretation of the results.

Parameter tuning.

- (First) task fulfilled: ES parameters for high SR found.
- Better performance may be possible: μ value at upper limit.
- Parameters κ and *recGrp* have little influence.
- Possible explanations:
  - Increased population size induces higher (needed?) diversity.
  - Large selection pressure and high learning rates lead to fast reaction when lesser constrained search points are found.

■

Pre-experimental planning.

Comparison.

- Experiments to avoid floor and ceiling effects:
  - Run length distributions: How many runs were completed successfully after $t_{\max}$ iterations?
  - Varying the starting points: How many runs were completed successfully after $t_{\max}$ iterations from different starting points?
  - Varying the problem dimension: How many runs were completed successfully after $t_{\max}$ iterations for different problem dimensions?
- Here (different to Mike's example): Specify the problem design.

Table: Comparison. Problem design.

n	$t_{\max}$	d	Init	Term	x_l	x_u	Perf
50	2500	10	I-4	T-3	15	30	PM-3

Scientific claim.
Comparison.

- Consider the experimental setup from [Shi and Eberhart, 1998].

Table: Comparison. Problem design. I-4 denotes non-uniform random starts. The algorithm terminates, if the ressources are exhausted (T-3).

n	$t_{\max}$	d	Init	Term	x_l	x_u	Perf
50	10,000 – 60,000	10 – 30	I-4	T-3	15	30	PM-3

Example

Claim: The PSO constriction variant (PSO_C) outperforms the PSO inertia weight variant (PSO) on the Rosenbrock function.

Experimental design.
Comparison.

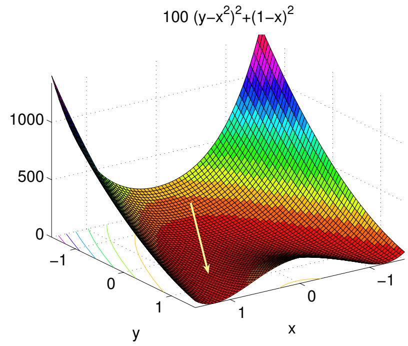
- Experimental design combines algorithm and problem designs.
- Experimental goal:
  - Determine improved algorithm designs for both algorithms for a given problem design.
  - Compare both algorithms based on the improved designs.

Table: Comparison. Algorithm design.

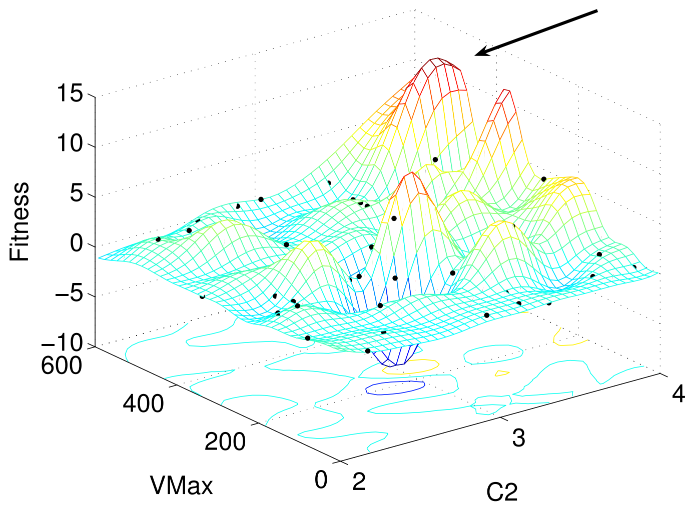
Design	s	c_1	c_2	$w_{\max}$	w_{scale}	$w_{\text{iterScale}}$	$v_{\max}$
$x_{\text{PSO}}^{(l)}$	5	1.0	1.0	0.7	0.2	0.5	10
$x_{\text{PSO}}^{(u)}$	100	2.5	2.5	0.99	0.5	1	750
x_{PSO}^{*}	21	2.25	1.75	0.789	0.283	0.94	11.05

Experiments.
Comparison.

- Test function not very costly.
- Many experiments.
- Each run configuration repeated $r = 5$ times.
- 80 design points (LHD).
- Max. 2000 runs to determine improved algorithm designs.
- Actual experiment requires a few minutes only.

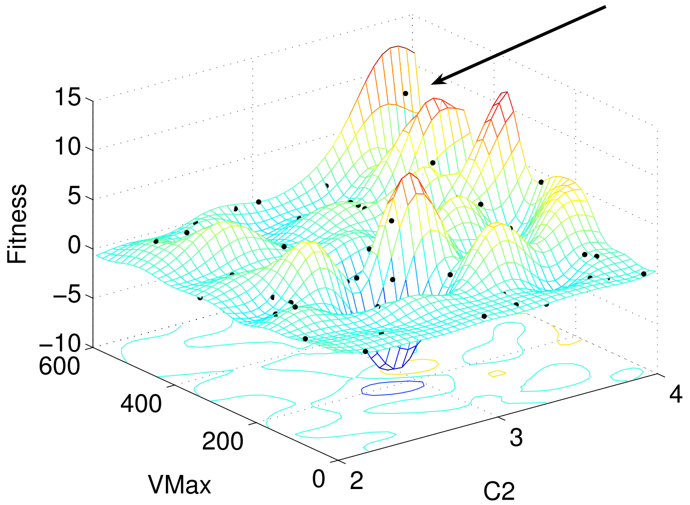


Evaluation and visualization.
DACE. Particle swarm optimization.



Evaluation and visualization.

DACE. Particle swarm optimization.



Acceptance/rejection of the statistical hypothesis.

Comparison.

Table: Results on the Rosenbrock function. NMS is a Nelder-Mead simplex algorithm, QN denotes a Quasi-Newton strategy. 50 repeats.

Design	Mean	Median	StD	Min	Max
$x^{(0)}_{PSO}$	1.84e+03	592.13	3.1e+03	64.64	18519
x^{*}_{PSO}	39.70	9.44	5.38	0.79	254.19
$x^{(0)}_{PSOC}$	162.02	58.51	378.08	4.55	2.62e+03
x^{*}_{PSOC}	116.91	37.65	165.90	0.83	647.91
$x^{(0)}_{NMS}$	9.07e+03	1.14e+03	2.50e+04	153.05	154966
x^{*}_{NMS}	112.92	109.26	22.13	79.79	173.04
QN	5.46e-11	5.79e-11	8.62e-12	1.62e-11	6.20e-11

- Is PSO really better than PSO_C?

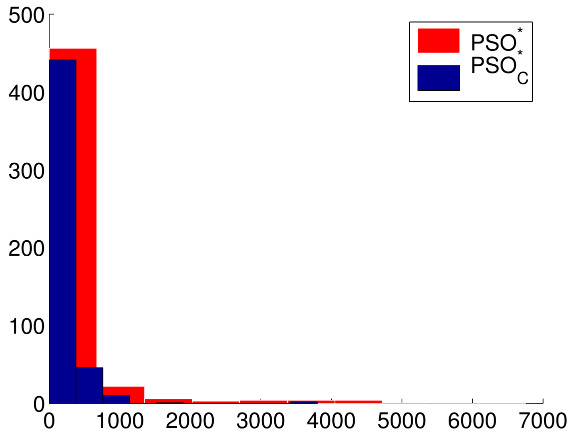
Objective interpretation of the results.

Comparison. A closer look at the data.

- Here: 500 repeats.
- PSO: Mean Y1 = 287.15
- PSO_C: Mean Y2 = 150.90
- p value: p = 6.7035e-05
- 95 % confidence interval: ci = [69.67 202.84].
- t test: reject the null hypothesis.
- Now: PSO_C outperforms PSO?

Objective interpretation of the results.

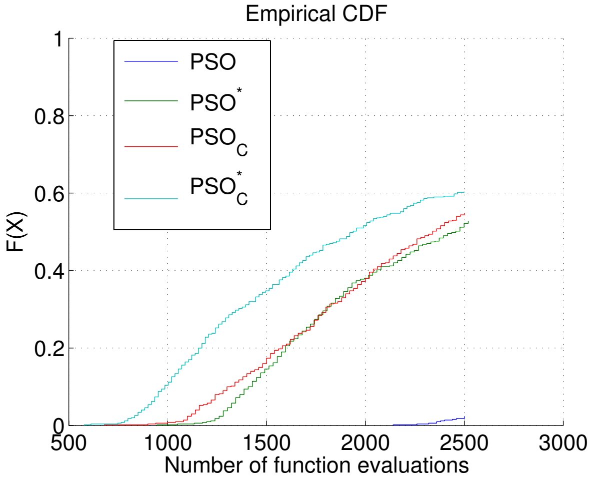
Comparison. Histogram.



- Histograms indicate:
  - High variance in the data.

Objective interpretation of the results.

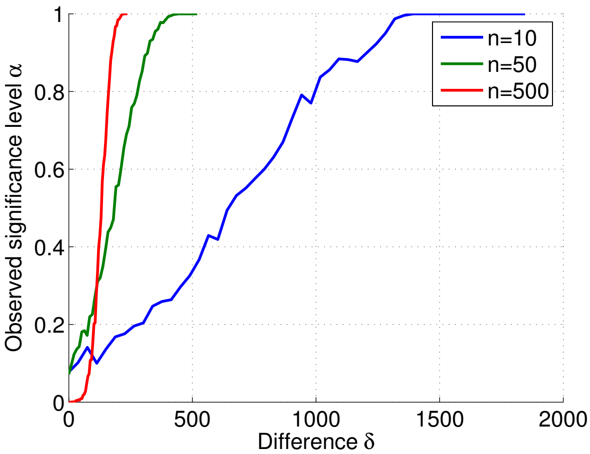
Comparison. Run-length distribution.



- RLD indicate:
  - PSO_C performs slightly better than PSO.

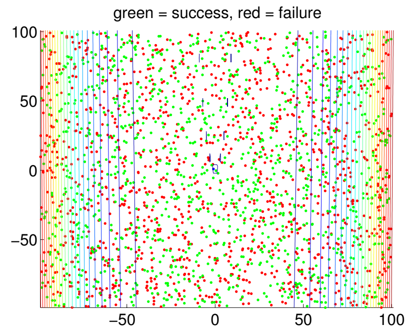
Objective interpretation of the results.

Comparison. OSL plots.



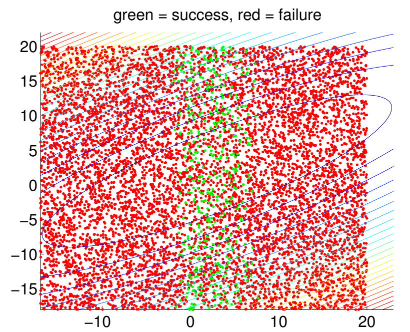
- OSL plots indicate:
  - Difference depends on the number of experiments.

Objective interpretation of the results.
Comparison. Starting points.



- Starting points indicate no structure.

Objective interpretation of the results.
Comparison. Starting points.



- $y = 3 + (x_1 - 1.5x_2)^2 + (x_2 - 2)^2$.
- $x^* = [3 \quad 2]$.
- $f^* = 3$.
- 10,000 starting points.
- Starting points indicate structure.

Objective interpretation of the results.

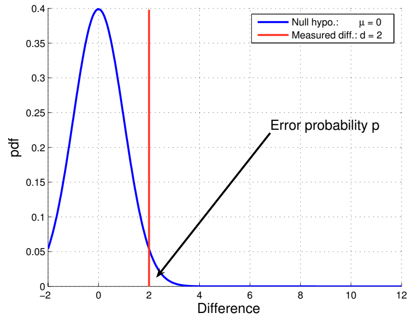
Comparison.

- Analysis reveals:
  - Experimental setup has to be modified.
  - t tests, confidence intervals, or p values alone are not sufficient.
  - Vary factors that influence the results (and statistics of these results).
  - Rosenbrock function is not well suited to compare the performance of algorithms, because it causes high variance.
  - Both PSO variants perform very poorly on the Rosenbrock function.
  - Other performance measures necessary, e.g. best result from 5 runs.
  - Good comparisons can pose new questions, they can be regarded as starting points for further investigations.

■

Statistical hypothesis.

- Run algorithm A and B n times.
- Two result vectors: y_A and y_B that contain the best function values.
- Difference $d_{AB} = y_A - y_B$.
- $\mathrm{Var}(d_{AB}) \neq 0$.
- Null hypothesis: " There is no difference in means $\mu = 0$."
- Error probability.



Error statistics.

- p value: Probability, that the observed (or larger) effect occurs, under the assumption that the null hypothesis is true.
- Small p values $\Rightarrow$ improbable that the observed effect occurs under the null hypothesis.
- Convention: p value ≤ 0.05 statistically significant.

Definition (p value)

$P(H \text{ true} \mid \text{result})$ or $P(\text{result} \mid H \text{ true})$?

- p value is not the probability that the null hypothesis H is true. The null hypothesis is either true or wrong.

Quotes from recent publications.

Example (Problematic.)

. . . are compared using the null-hypothesis $H0 : FX = FY$ and the one-sided alternative $H1 : FX < FY$. Only if the probability of the null-hypothesis $P(H0)$ is at most 0.01, it is rejected and the alternative hypothesis is accepted.

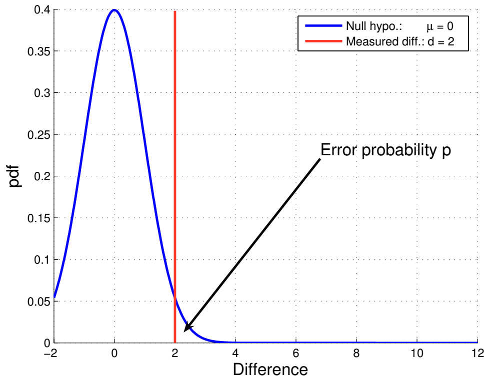
Example (Good.)

The test $H_0 : \beta = 1$ versus $H_1 : \beta < 0$ has $p < 0.001$. This provides some evidence that the empirical relative complexity coefficient is . . .. However, the model implies that . . ..

How to detect differences?

Null hypothesis true. Prob., that a difference *d* (or larger) occurs: 0,0228.

- Experimenter assumes a difference. But: *p* value too large. Therefore: further experimemts.
- Experimenter demonstrates a difference. Now: *p* value small enough. Therefore: no further experiments necessary.

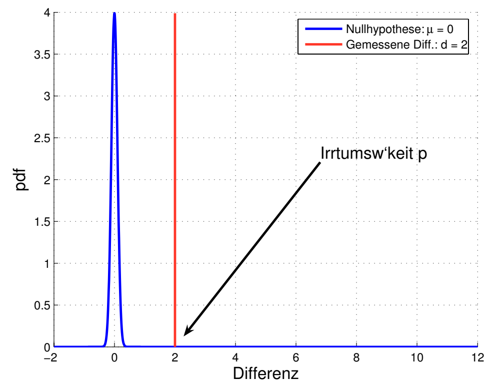


The New Experimentalism—Results | Difference Detection and the p-Value

How to produce differences?

Null hypothesis true. Prob., that a difference *d* (or larger) occurs: 0.

- Experimenter assumes a difference. But: *p* value too large. Therefore: further experimemts.
- Experimenter demonstrates a difference. Now: *p* value small enough. Therefore: no further experiments necessary.



Idea: Control the variables that influence the *p* value.

- Statistical controversies: *p* value and hypothesis testing.

- How to avoid this arbitrariness?
- Control the variation ⇒ Observed significance.
- OSL plots.
- Dynamic analysis.
- Furthermore: Vary problem dimension, instances, starting points etc. ■

Benefits of this approach.

- Combination and improvement of classical and modern statistical techniques such as:
  - Design of Experiments.
  - Regression trees.
  - Design and Analysis of Computer Experiments.
- Based on the new experimentalism, an influential trend in the philosophy of science:
  - Learning from error.
  - Statistical idea: Not avoiding, but controlling error.
  - Offers extensions and new interpretations of the Popperian view.

The philosophy of science seems to be in a state of flux, and the possibilities opened up by the new experimentalists seem to offer genuine hope for a recovery of some of the solid intuitions of the past about the objectivity of science, but in the context of a much more detailed and articulate understanding of actual scientific practice.

—Robert Ackermann, 1989.

Further literature I

- Ackermann, R. (1989).
  The new experimentalism.
  Brit. J. Phil. Sci., 40:185–190.
- Bartz-Beielstein, T. (2005).
  New Experimentalism Applied to Evolutionary Computation.
  PhD thesis, University of Dortmund.
- Bartz-Beielstein, T., Parsopoulos, K. E., and Vrahatis, M. N. (2004).
  Design and analysis of optimization algorithms using computational statistics.
  Applied Numerical Analysis & Computational Mathematics (ANACM), 1(2):413–433.

Further literature II

- Jones, D., Schonlau, M., and Welch, W. (1998).
  Efficient global optimization of expensive black-box functions.
  Journal of Global Optimization, 13:455–492.
- Mayo, D. G. (1996).
  Error and the Growth of Experimental Knowledge.
  The University of Chicago Press.
- Santner, T., Williams, B., and Notz, W. (2003).
  The Design and Analysis of Computer Experiments.
  Springer, Berlin.
- Shi, Y. and Eberhart, R. (1998).
  Parameter selection in particle swarm optimization.
  In Porto, V., Saravanan, N., Waagen, D., and Eiben, A., editors, *Evolutionary Programming*, volume VII, pages 591–600. Springer.