Performance Assessment in Optimization

Anne Auger

RandOpt Team INRIA and CMAP, Ecole Polytechnique anne.auger@inria.fr

General Problem

Evaluate the performance of optimization algorithms

Compare the performance of different algorithms

understand strength and weaknesses of algorithms

help in design of new algorithms

General Problem (cont.)

Algorithms are in general too complicated to be evaluated theoretically on the wide range of problems/ difficulties one is interested to solve

need to do some benchmarking, i.e. evaluate empirically on test functions the performance of an optimizer

run the optimizer several times independently on a set of benchmark function

display some statistical measures of performance

Test functions

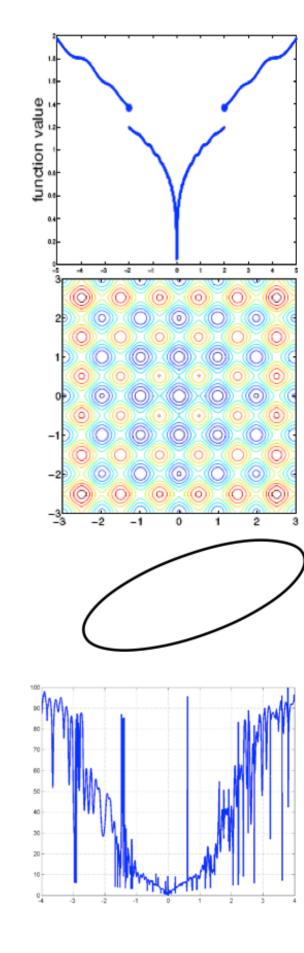
Many real world problems share common difficulties:

- \rightarrow non separability (correlations between variables)
- \rightarrow ill-conditioned (certain direction steeper than others),
- \rightarrow ruggedness (noise, ...),
- \rightarrow multi-modality
- \rightarrow non-convexity

Ideally an optimizer should cope with all of them

function testbed:

should "reflect reality": should model typical difficulties one is willing to solve mainly non-convex and non-separable scalable with the search space dimension not too easy to solve, but yet comprehensible



State-of-the-art Test Suite

Black-Box Optimization Benchmarking test suite noiseless / noisy testbed

http://coco.gforge.inria.fr/doku.php?id=start

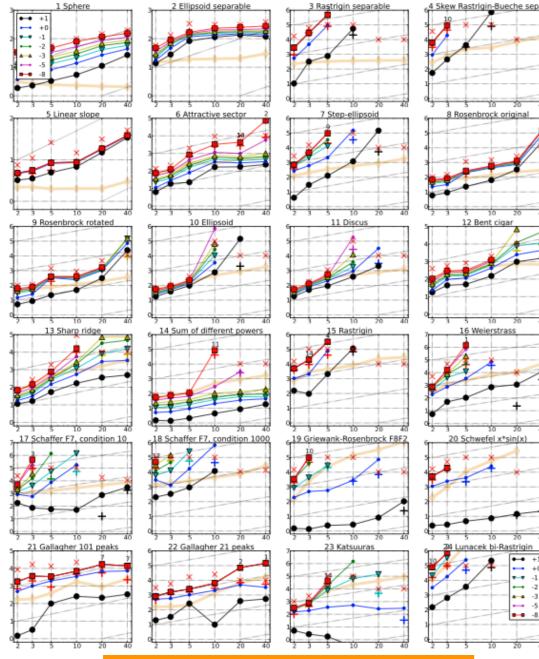


Figure 1: Expected nu of *f*-evaluations to real *f*-evaluations in any t are $\Delta f = 10^{\{1,0,-1,-2,-3\}}$ thick line with diamon mean linear scaling, share

noiseless testbed

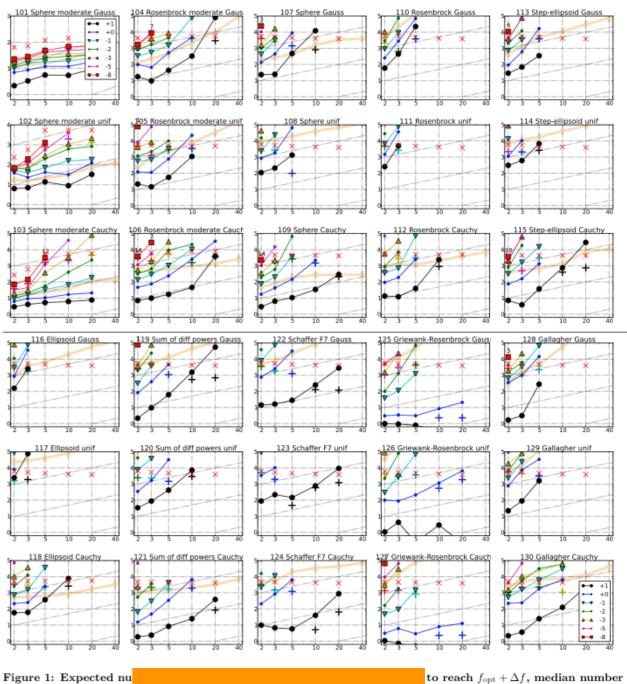


Figure 1: Expected nu of *f*-evaluations to reac *f*-evaluations in any tr are $\Delta f = 10^{(1,0,-1,-2,-3,-1)}$

thick line with diamone

Ded once (+) and maximum number of ¹⁰ values versus dimension. Shown mber of successful trials. The light 2009 for $\Delta f = 10^{-8}$. Horizontal lines

mean linear scaling, slanted grid lines depict quadratic scaling.

Performance measure

CPU time (to reach a given target) drawbacks: depend on the implementation, on the language, on the machine

time is spent on code optimization instead of science Testing heuristics, we have it all wrong, J.N. Hooker, 1995 Journal of Heuristics

Prefer "absolute" value: # of function evaluations to reach a given target

assumptions: internal cost of the algorithm negligible or measured independently

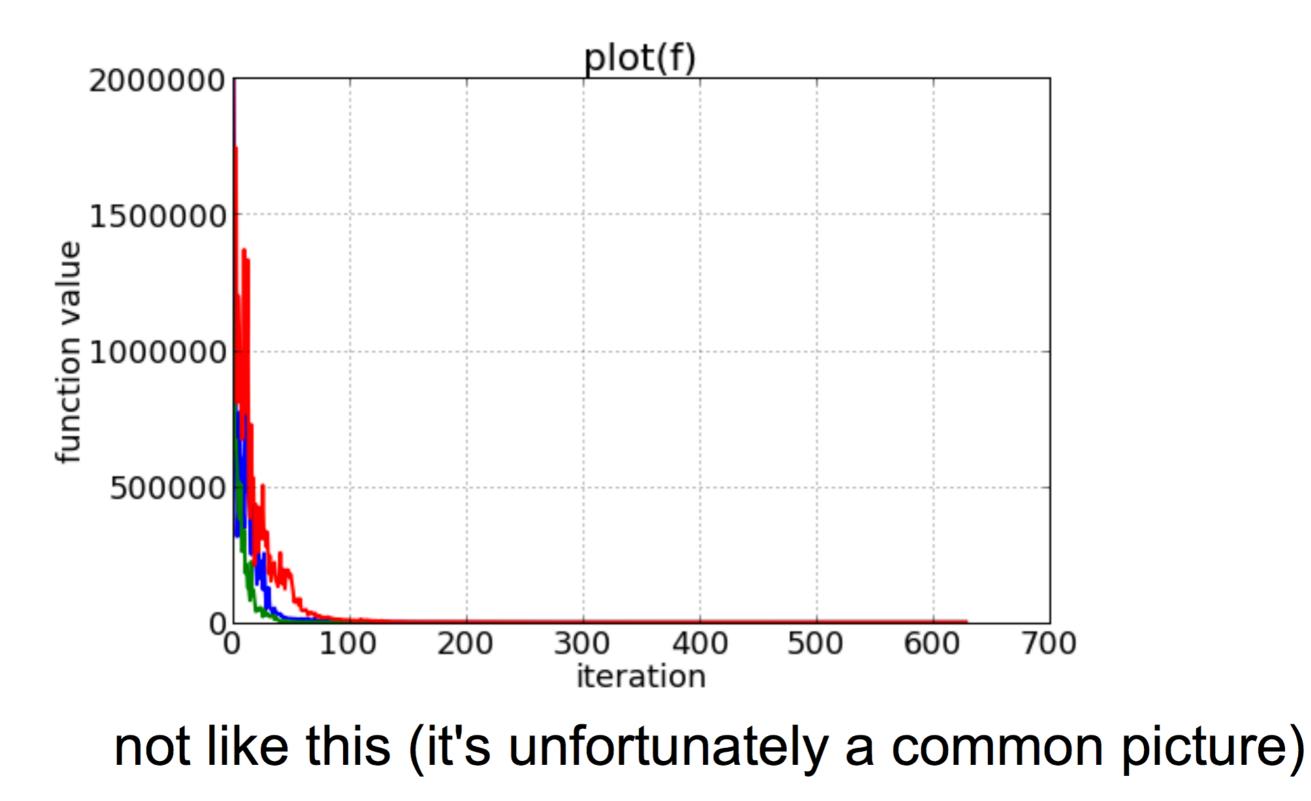
Performance measure

empirically

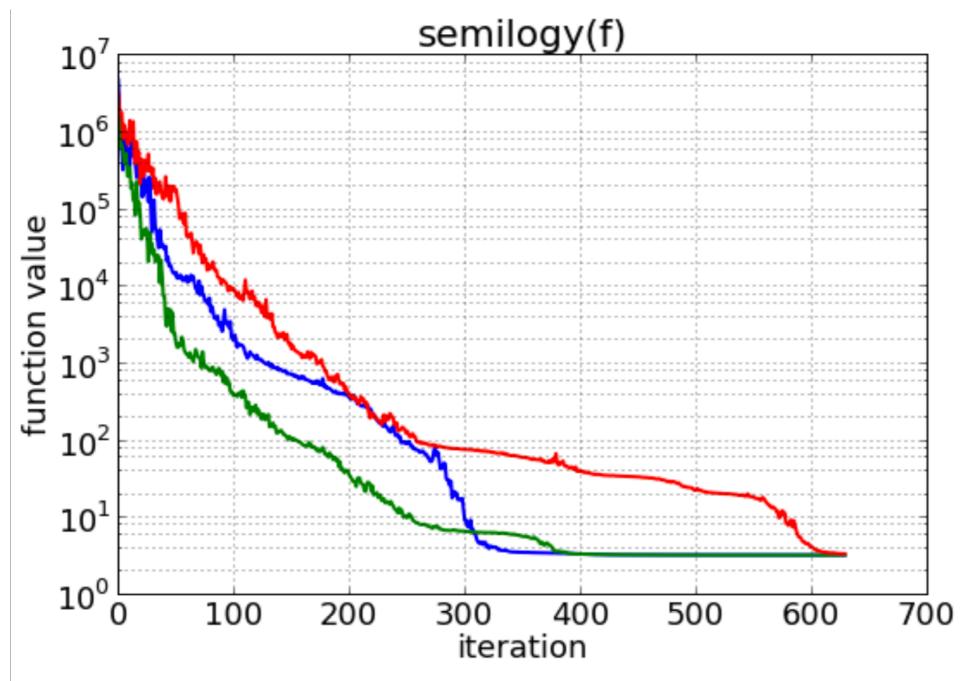
convergence graphs is all we have to start with

the right presentation cannot be overestimated: details are important!

Displaying 3 runs (three trials)

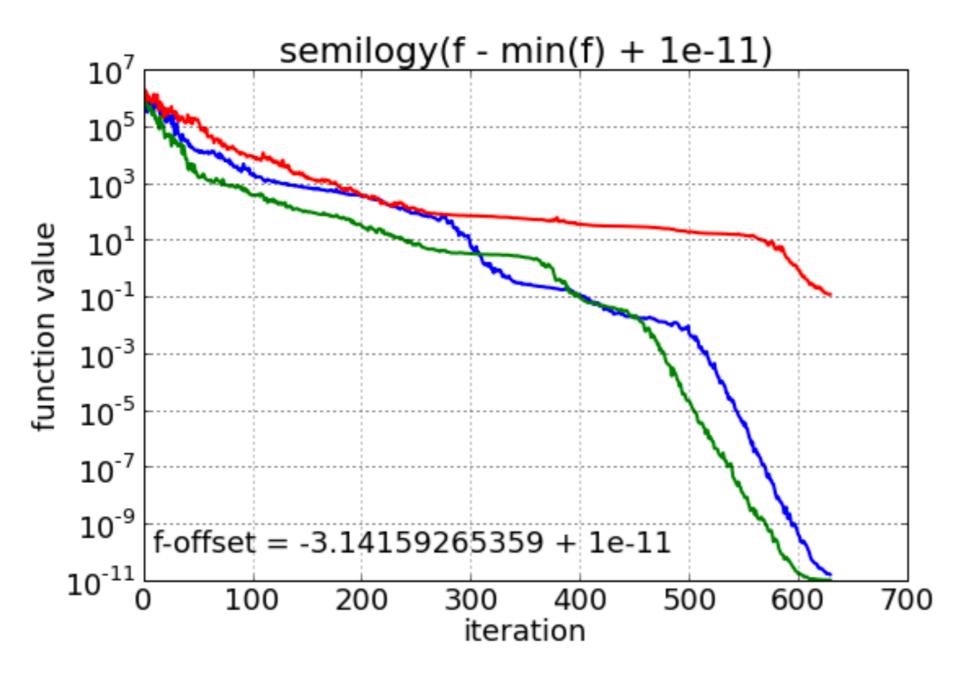


Displaying 3 runs (three trials)



better like this (shown are the same data), caveat: fails with negative f-values

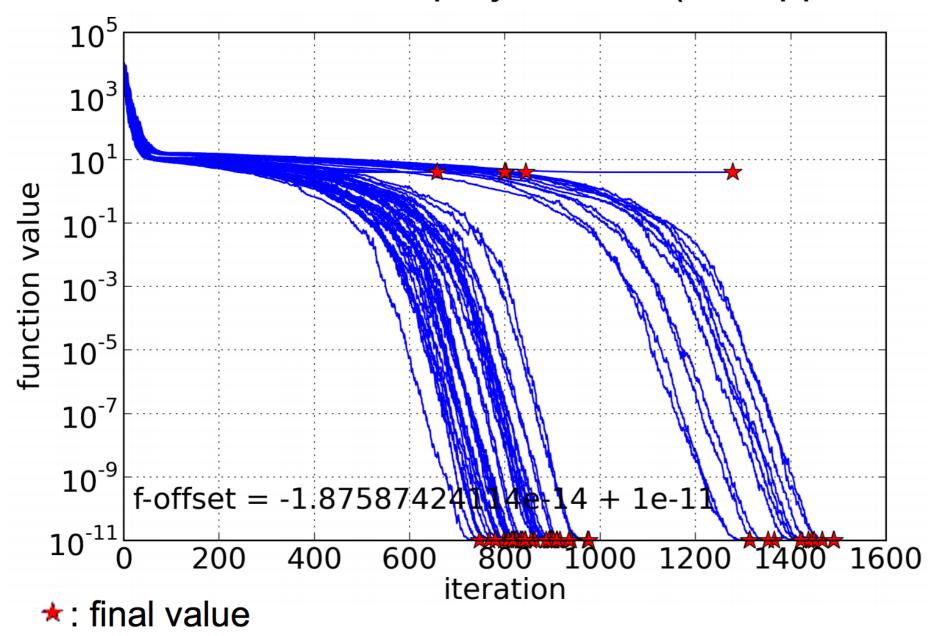
Displaying 3 runs (three trials)



even better like this: subtract minimum value over all runs

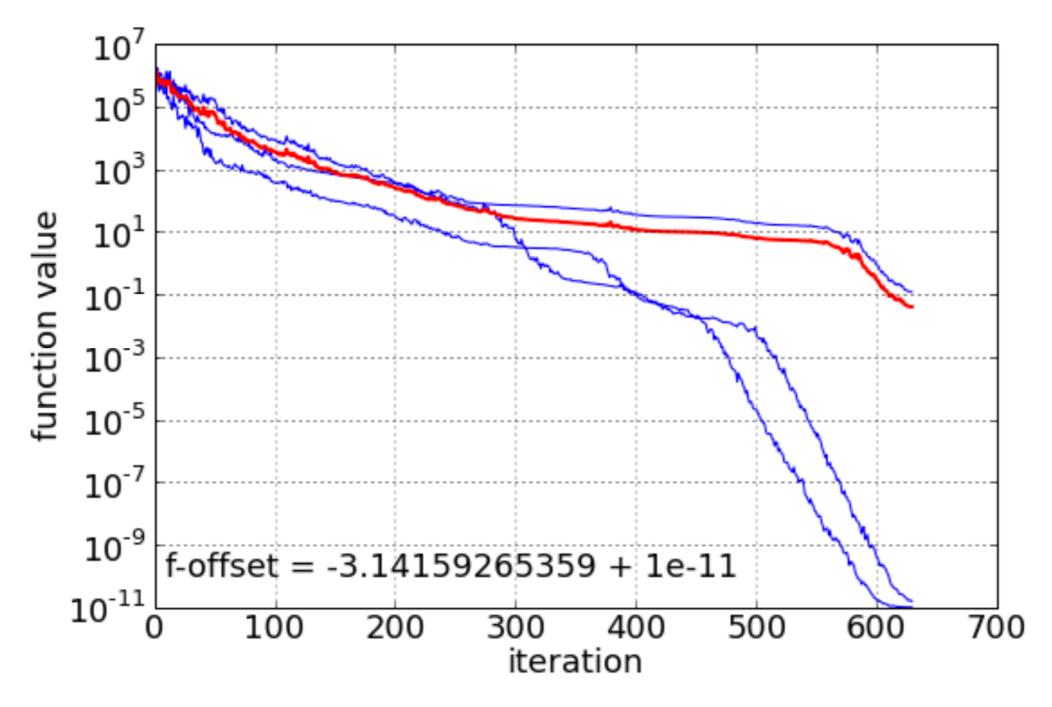
Displaying 51 runs

don't hesitate to display all data (the appendix is your friend)



observation: three different "modes", which would be difficult to represent or recover in single statistics

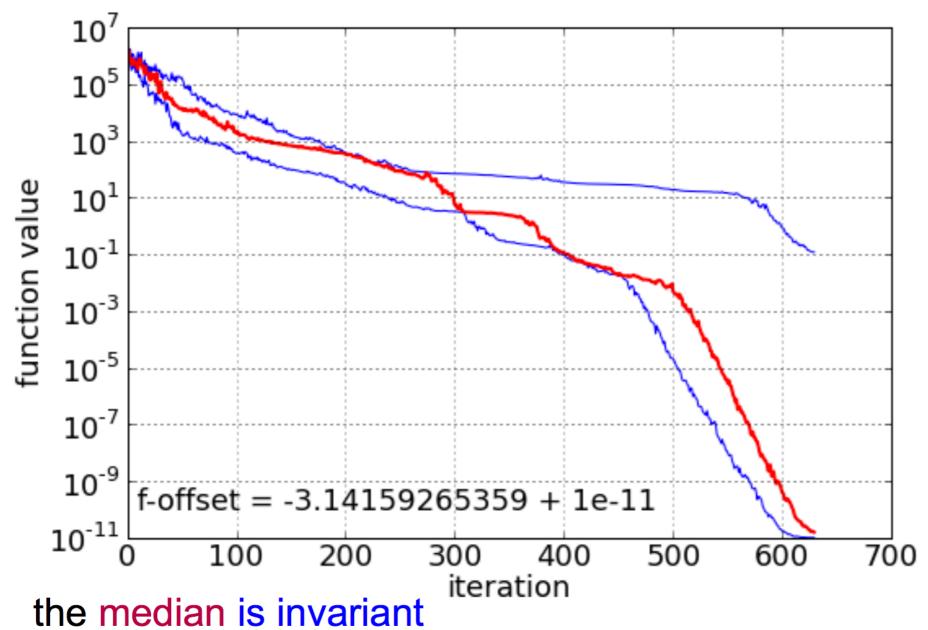
Which Statistics?



mean/average function value

tends to emphasize large values

Which Statistics?



- unique for uneven number of data
- independent of log-scale, offset...

```
median(log(data))=log(median(data))
```

same when taken over x- or y-direction

"Algorithm A is 10/100 times faster than Algorithm B to solve this type of problems"

"Algorithm A is 10/100 times faster than Algorithm B to solve this type of problems"

quantitative measures

"Algorithm A is 10/100 times faster than Algorithm B to solve this type of problems"

quantitative measures

As opposed to

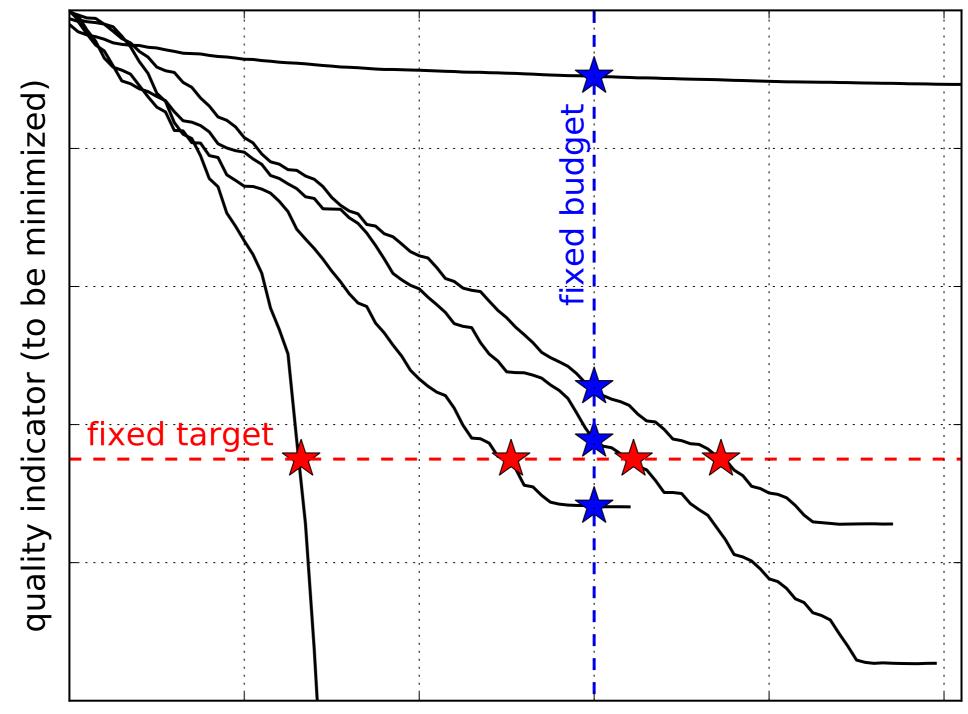
F.	EFWA vs EFWA-NG		
	EFWA	EFWA-NG	<i>p</i> -value
f_1	-1.3999E+03	-1.3999E+03	2.316E-03
f_2	6.8926E+05	6.5258E+05	4.256E-01
f_3	7.7586E+07	6.4974E+07	8.956E-01
f_4	-1.0989E+03	-1.0989E+03	7.858E-01
f_5	-9.9992E+02	-9.9992E+02	4.290E-02
f_6	-8.5073E+02	-8.4462E+02	1.654E-01
	C 8 C 8 (1 1 0 8	(000 1 T 00	

displayed: mean f-value after 3.10^5 f-evals (51 runs) bold: statistically significant concluded: "EFWA significantly better than EFWA-NG"

Source: Dynamic search in fireworks algorithm, Shaoqiu Zheng, Andreas Janecek, Junzhi Li and Ying Tan CEC 2014

a performance measure should be quantitative, with a ratio scale well-interpretable with a meaning relevant in the "real world" simple

Fixed Cost versus Fixed Budget Collecting Data



number of function evaluations

Fixed Cost versus Fixed Budget Collecting Data

Collect for a given target (several target), the number of function evaluations needed to reach a target

Repeat several times:

if algorithms are stochastic, never draw a conclusion from a single run

if deterministic algorithm, repeat by changing (randomly) the initial conditions

Displaying Performance

ECDF

Average RunTime (ART)

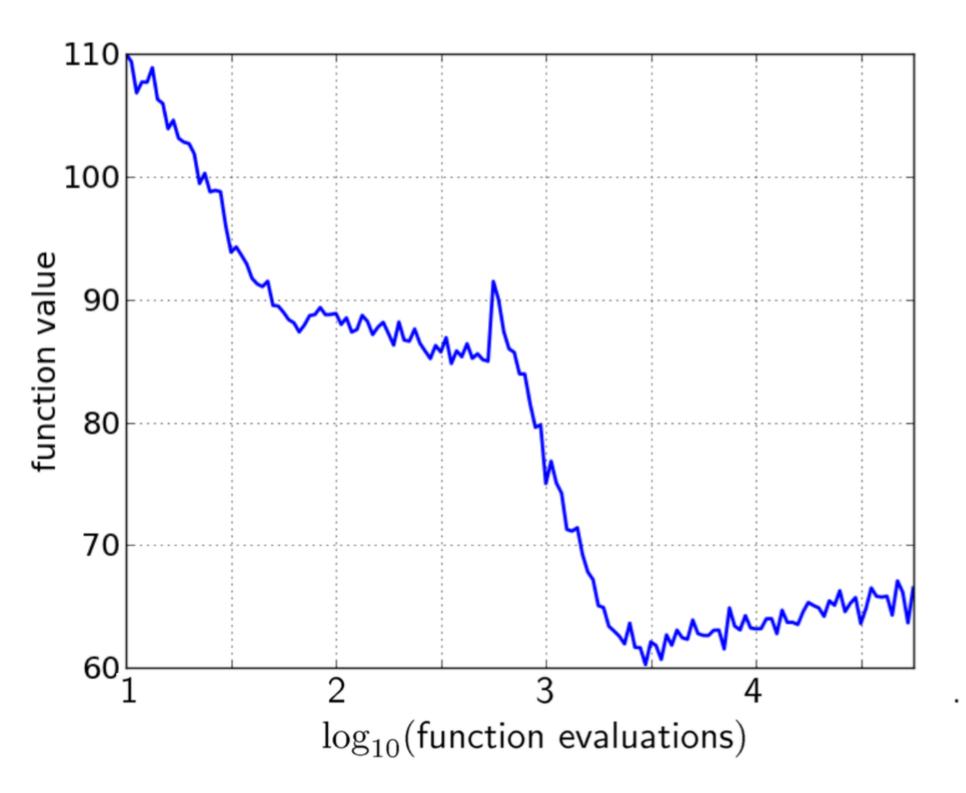
Displaying Performance ECDF

Empirical Cumulative Distribution (ECDF) of Runtime

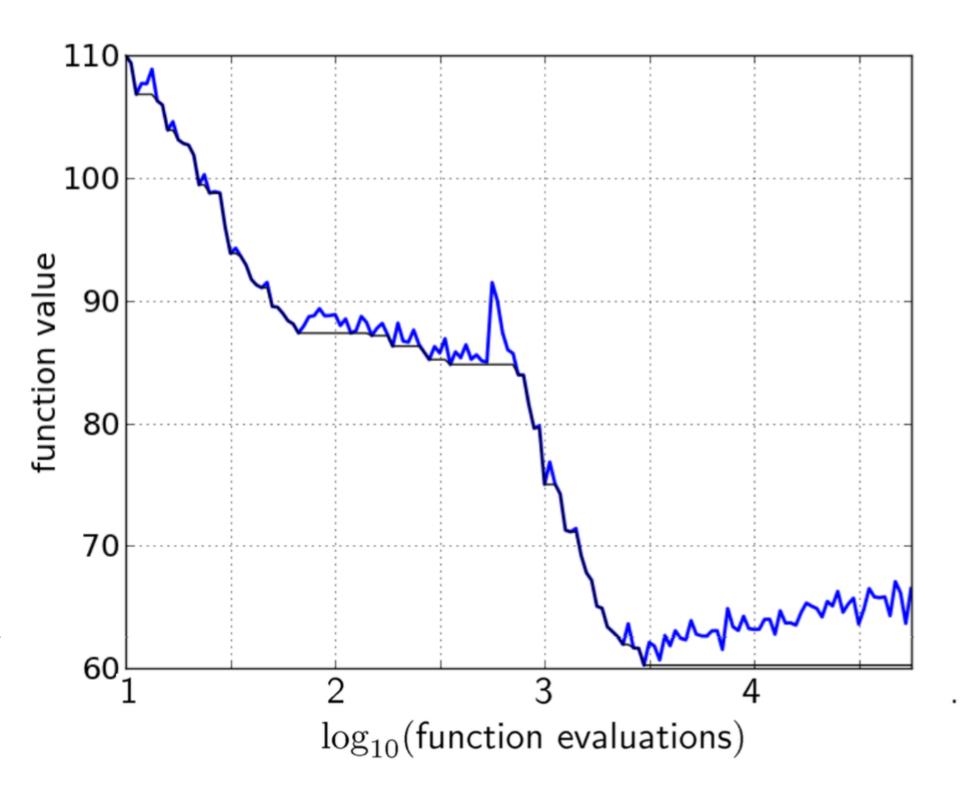
also known as data profile

[Moré, Wild, Benchmarking Derivative-Free Optimization Algorithms, SIOPT 2009]

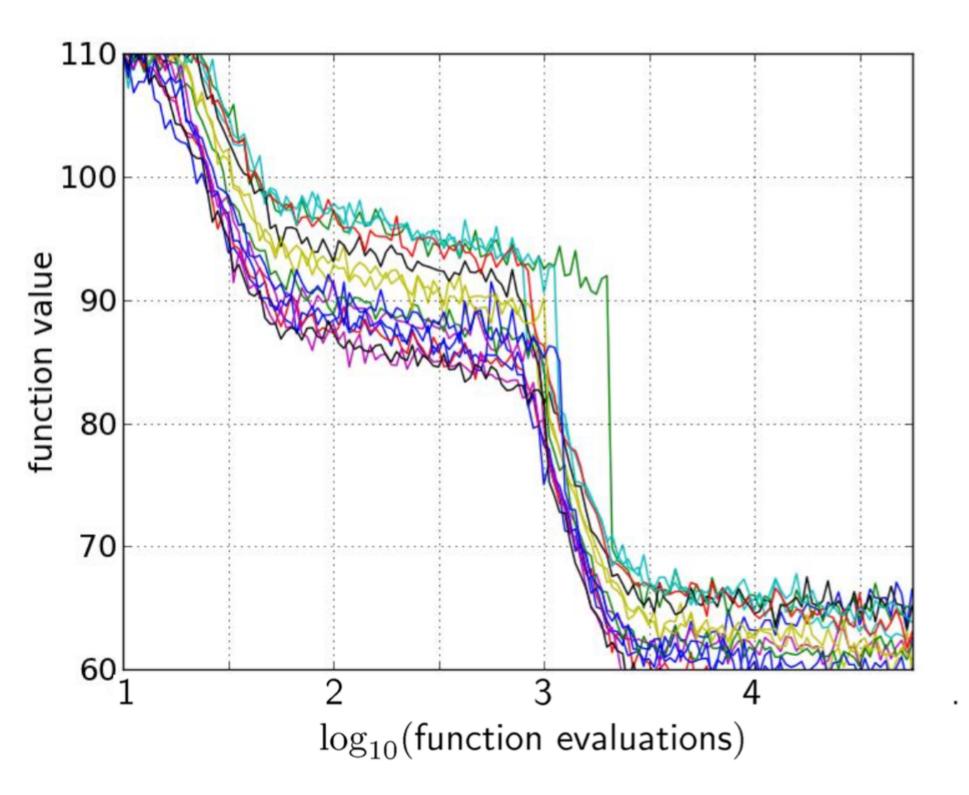
A Convergence Graph.



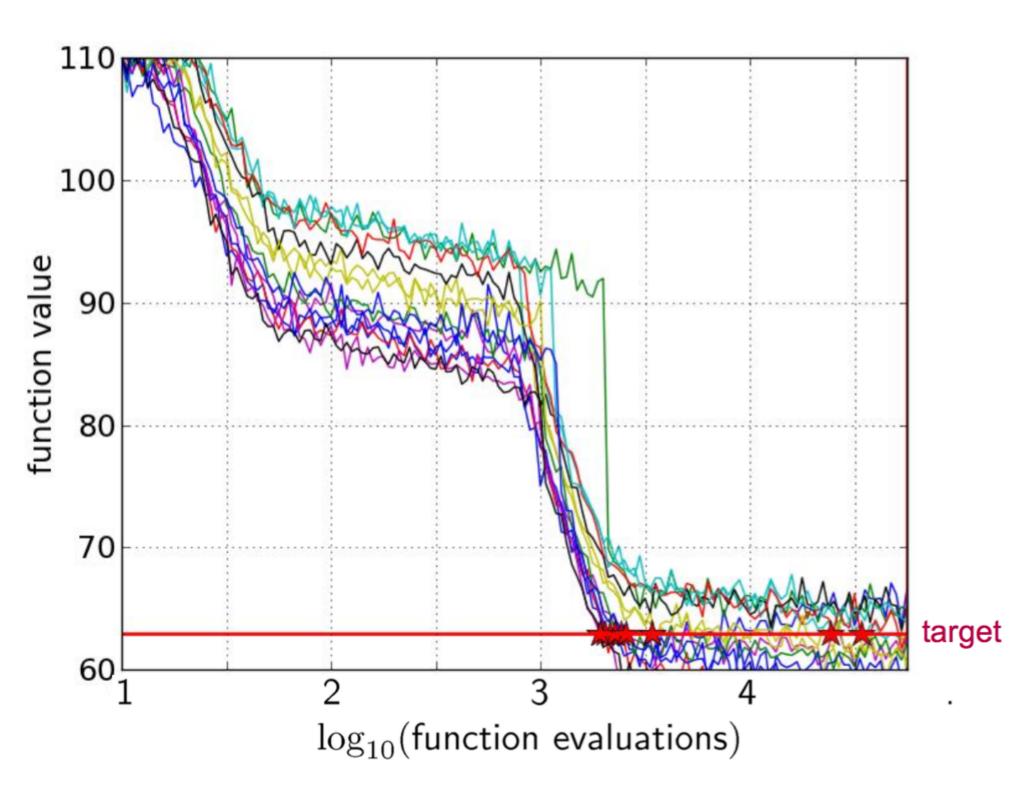
First Hitting Time is Monotonous



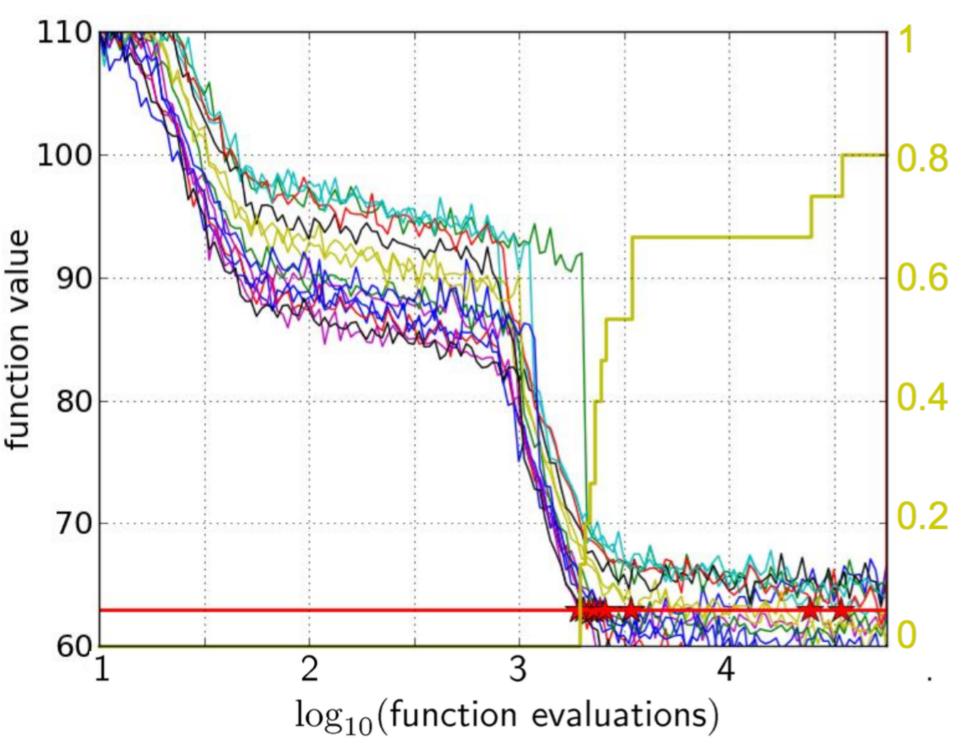
15 Runs



15 Runs ≤ 15 Runtime Data Points



Empirical Cumulative Distribution

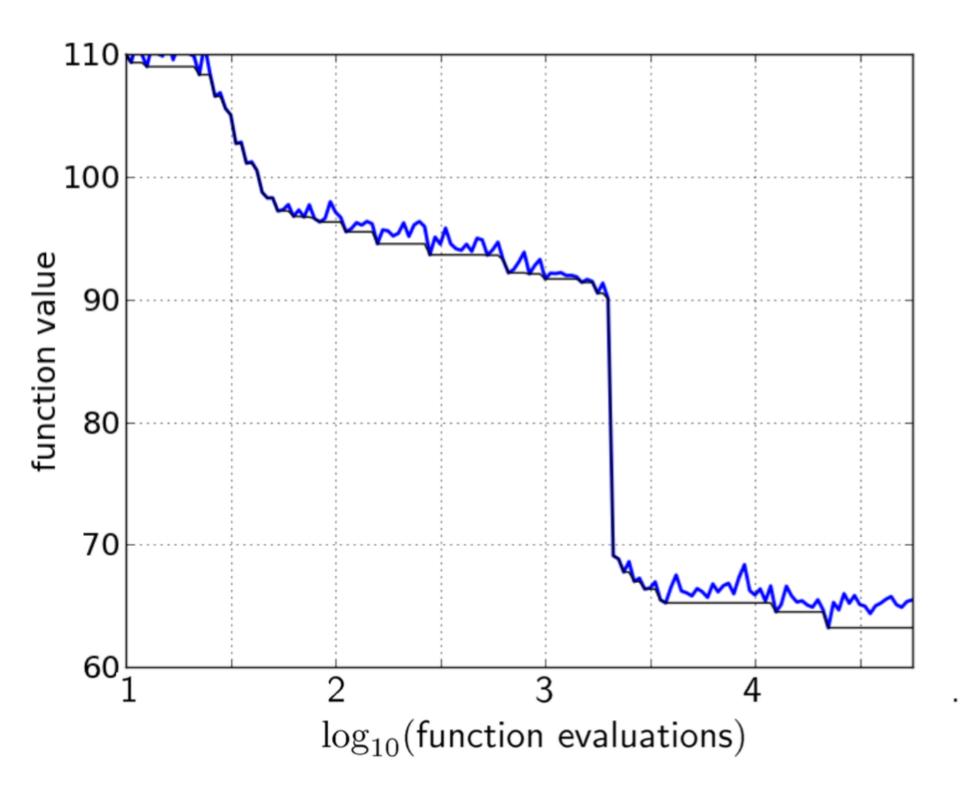


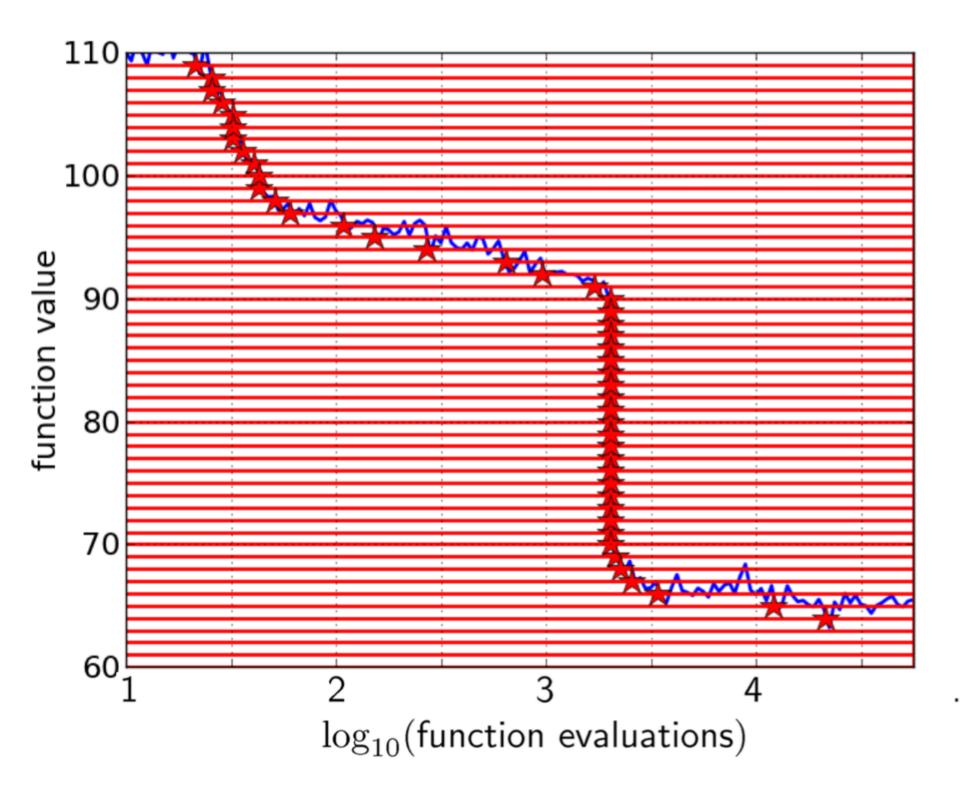
the ECDF of run lengths to reach the target

- has for each data point a vertical step of constant size
- displays for each x-value (budget) the count of observations to the left (first hitting times)

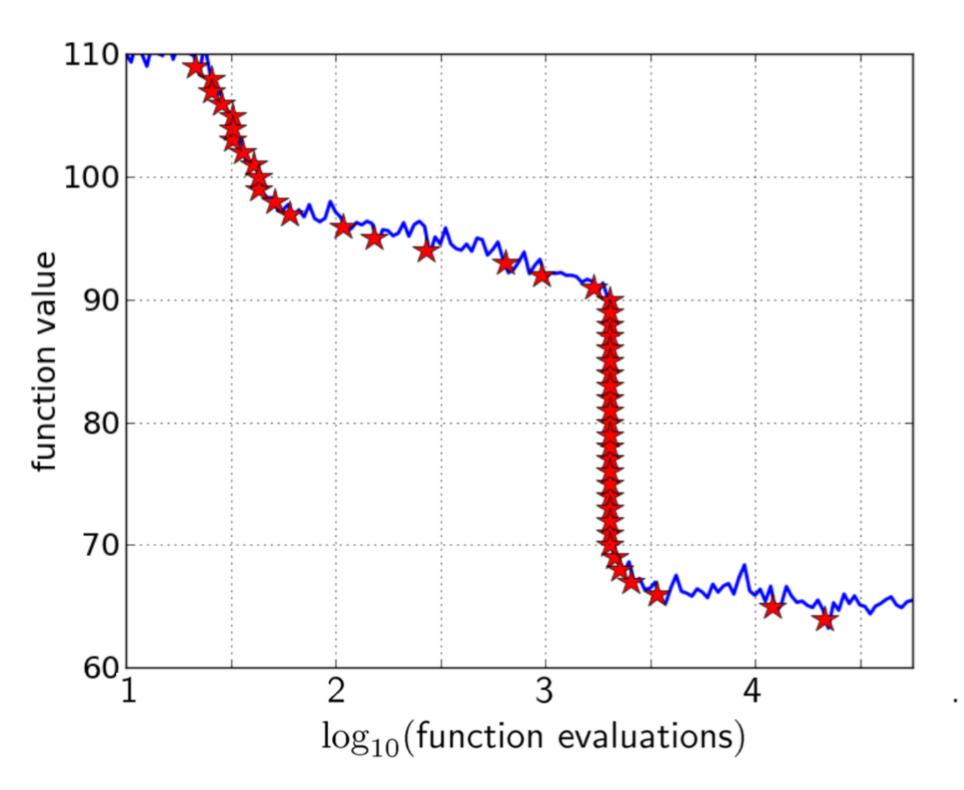
e.g. 60% of the runs need between 2000 and 4000 evaluations 80% of the runs reached the target

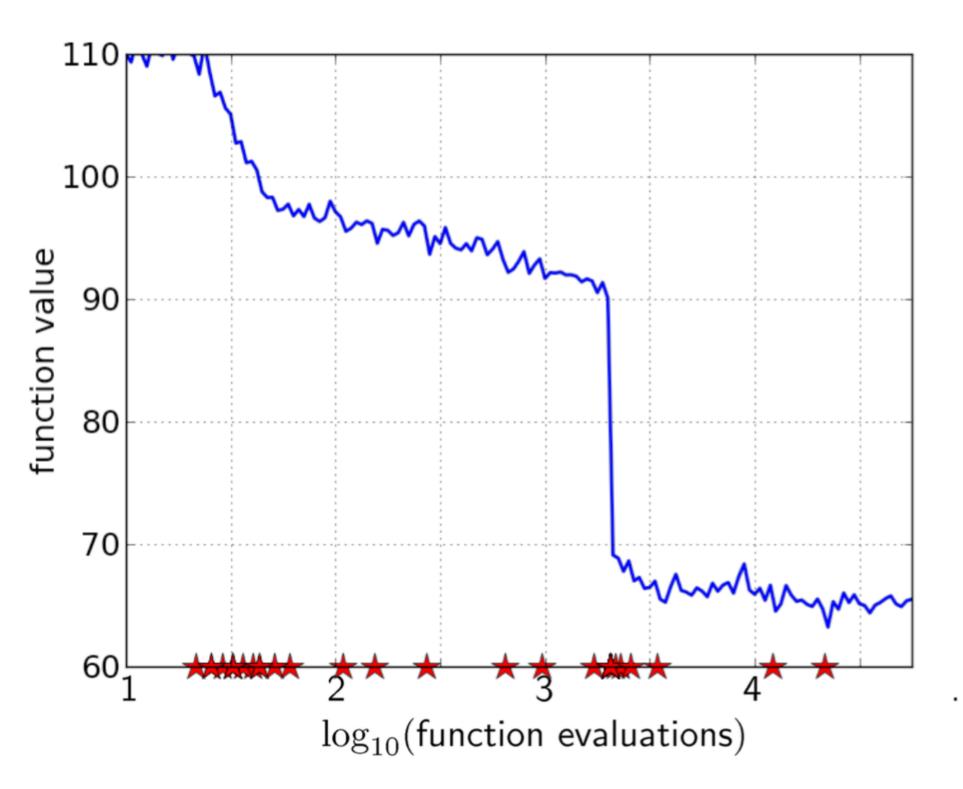
Reconstructing a single run via ECDF...

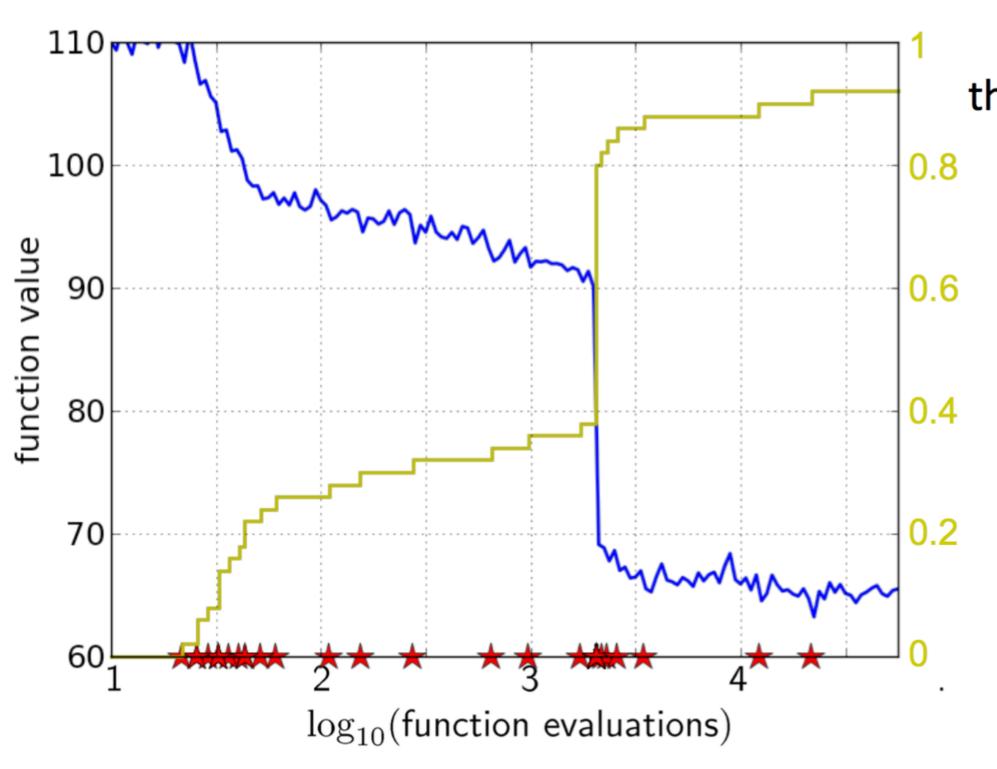




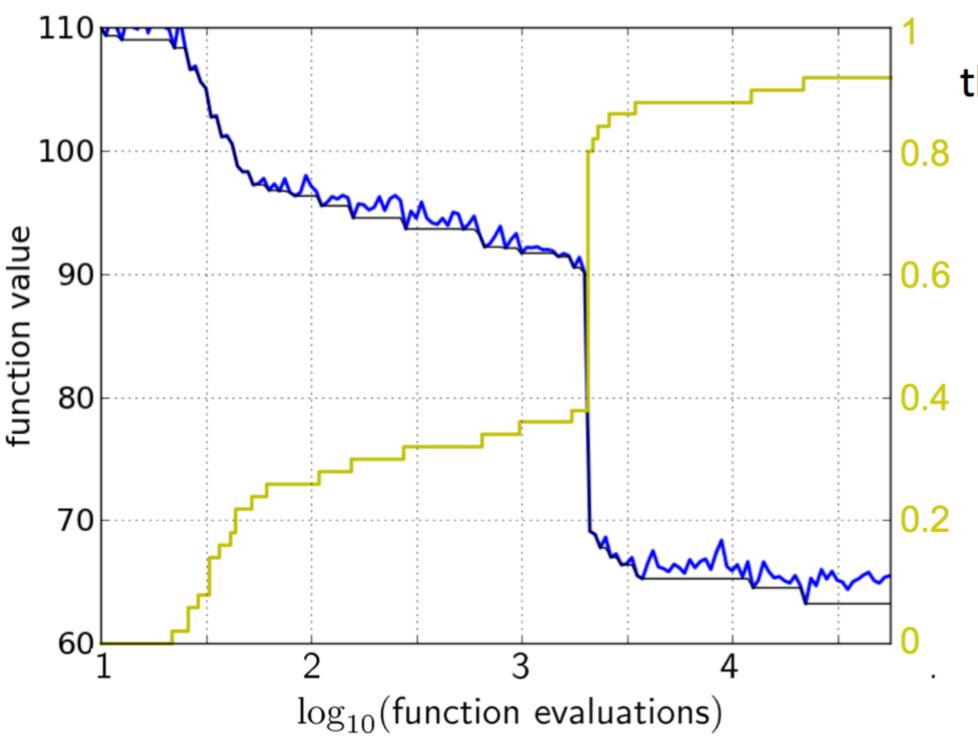
50 equally spaced targets



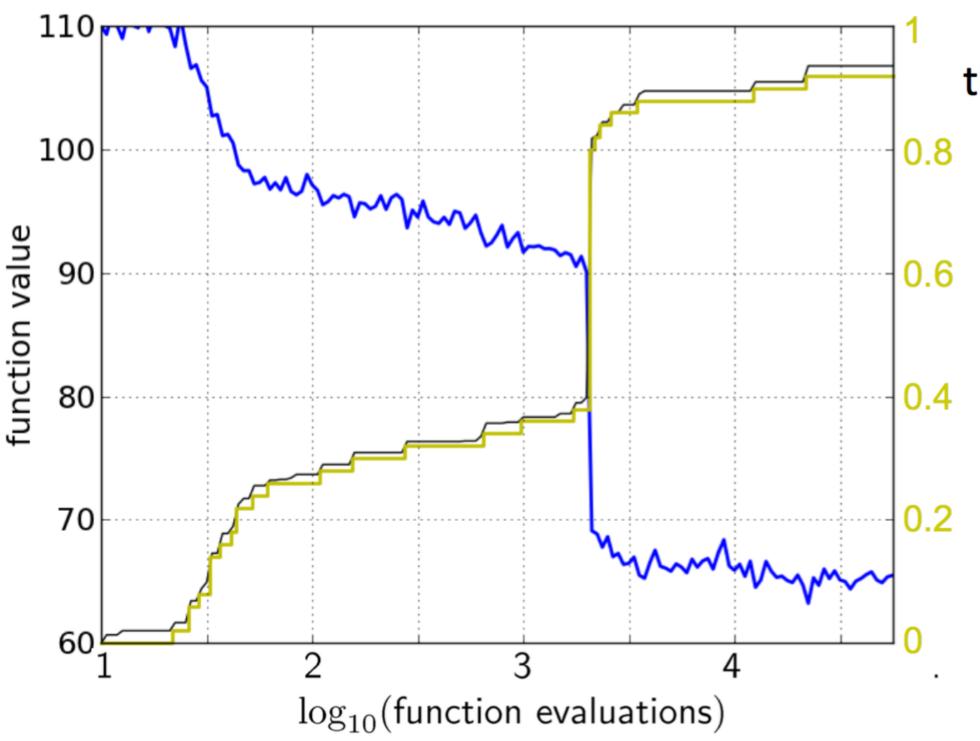




the empirical CDF makes a step for each star, is monotonous and displays for each budget the fraction of targets achieved within the budget



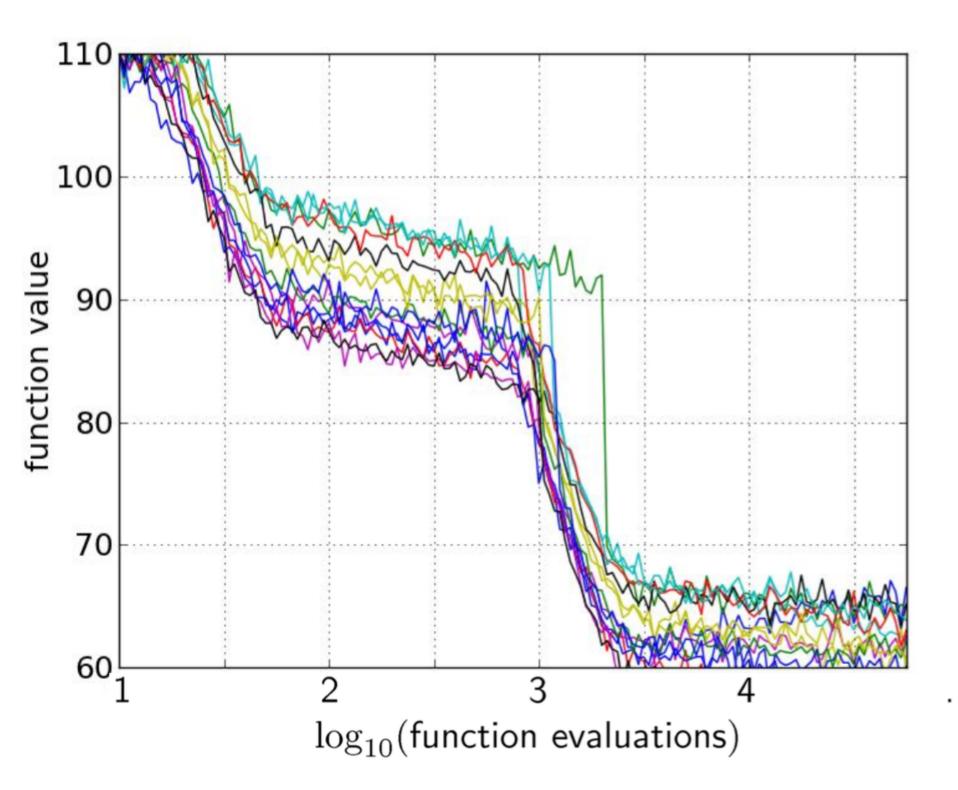
the ECDF recovers the monotonous graph, discretised and flipped



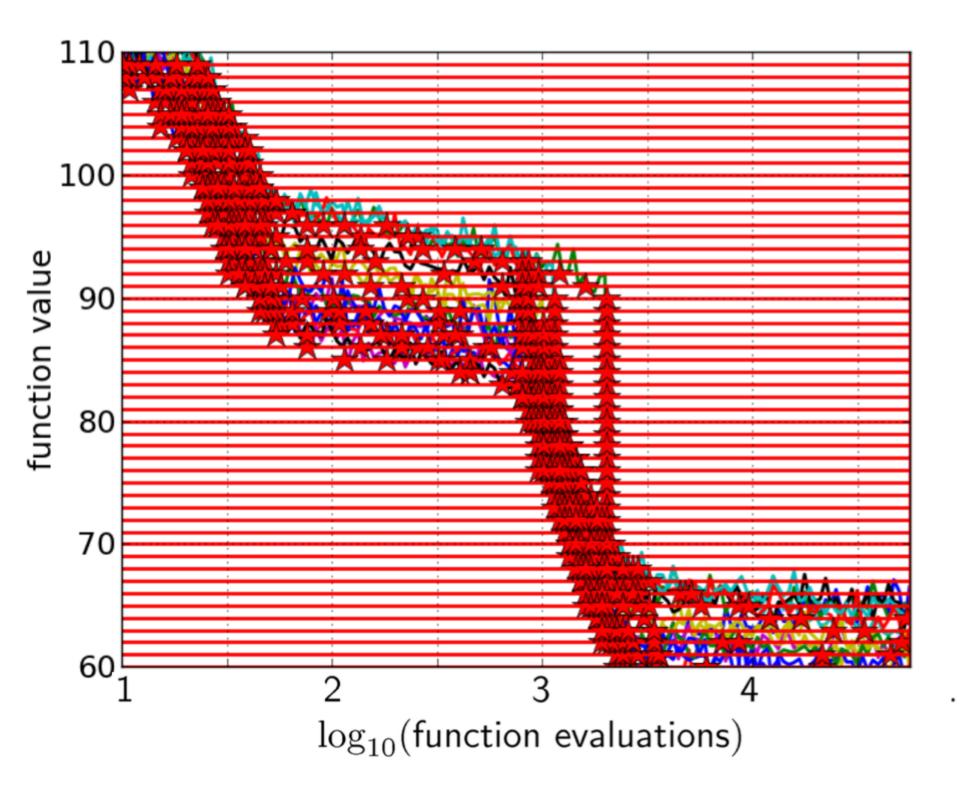
the ECDF recovers the monotonous graph, discretised and flipped

Aggregation ...

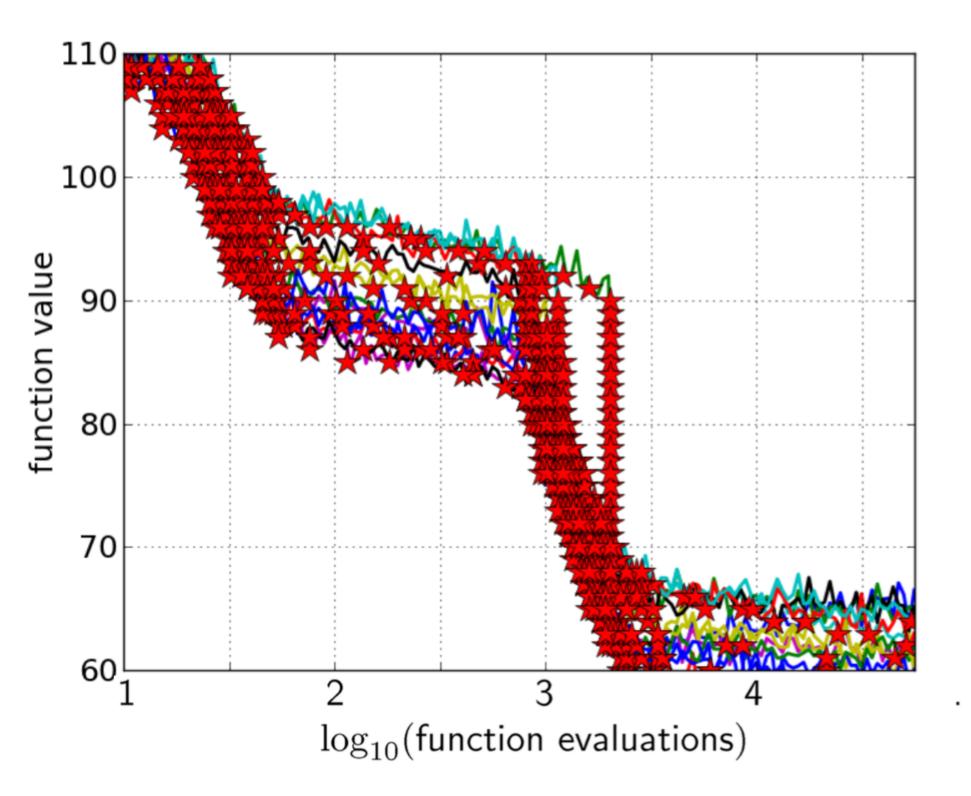
Aggregation



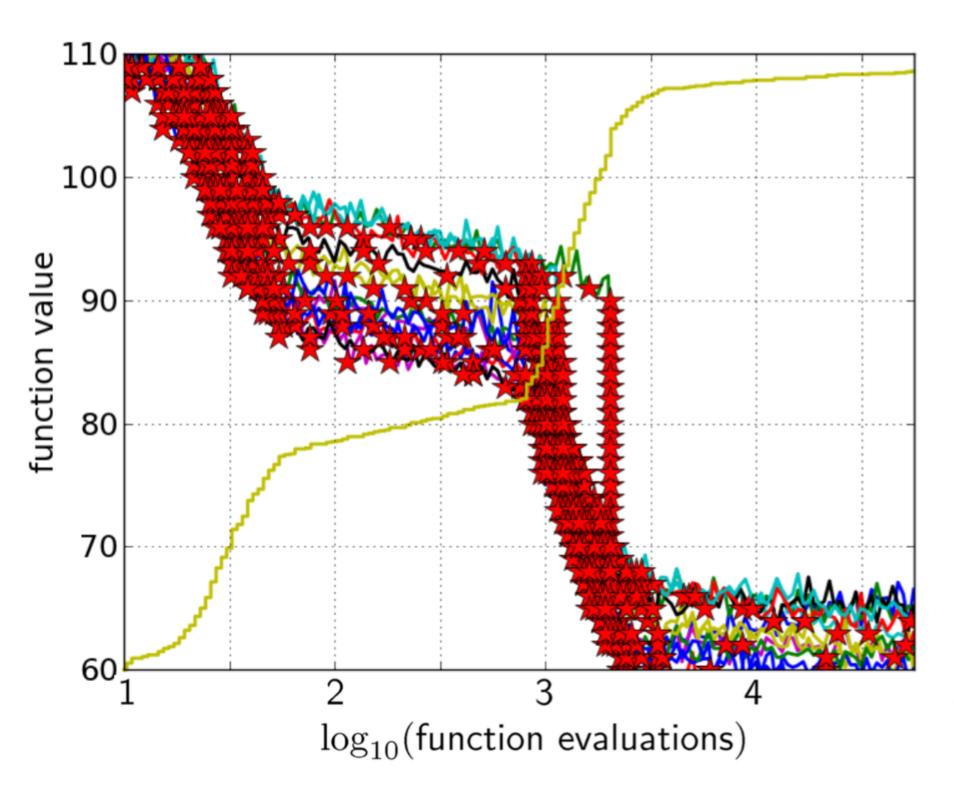
15 runs



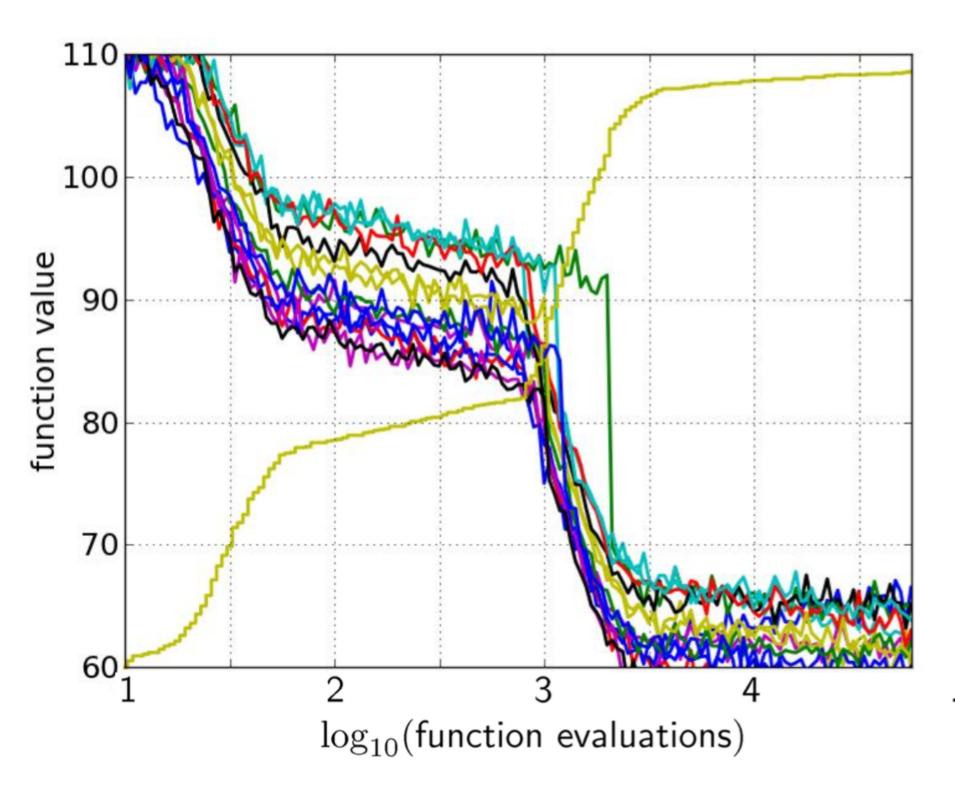
15 runs 50 targets



15 runs 50 targets



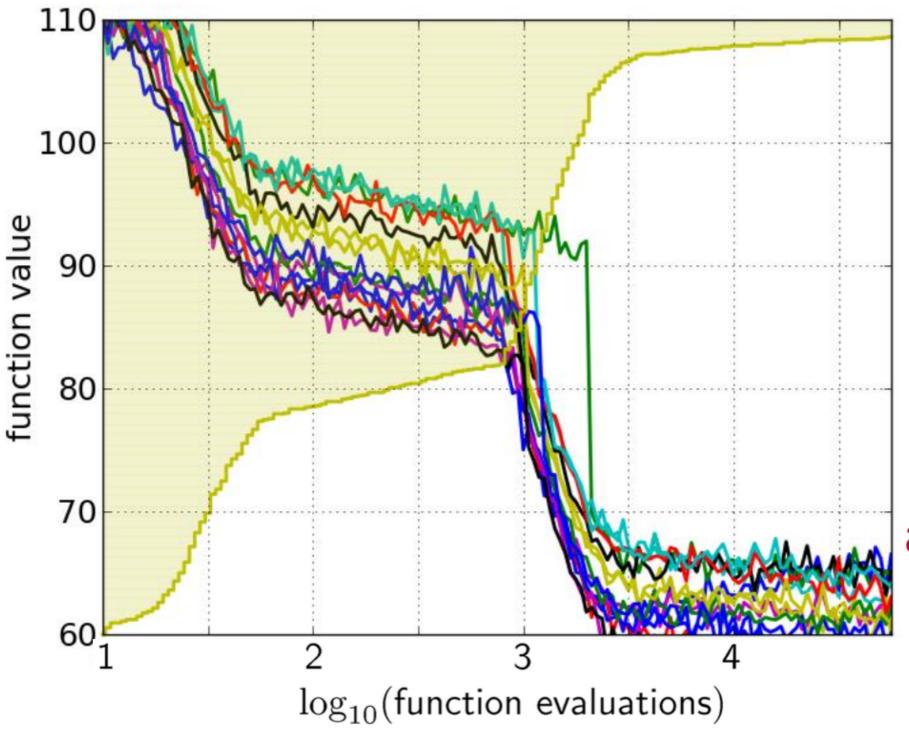
15 runs 50 targets ECDF with 750 steps



50 targets from 15 runs

...integrated in a single graph

Interpretation



50 targets from 15 runs integrated in a single graph

area over the ECDF curve

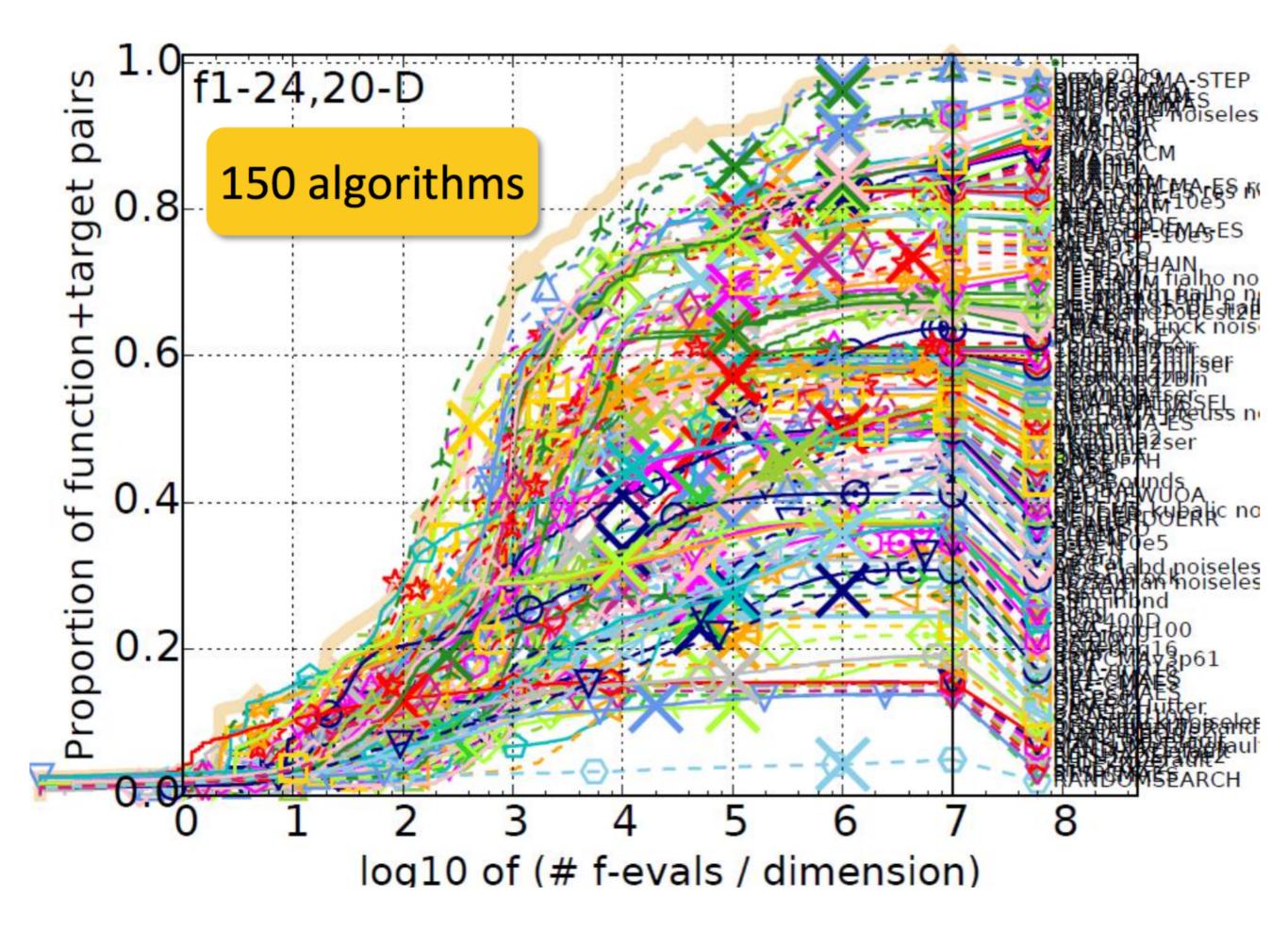
average log runtime (or geometric avg. runtime) over all targets (difficult and easy) and all runs

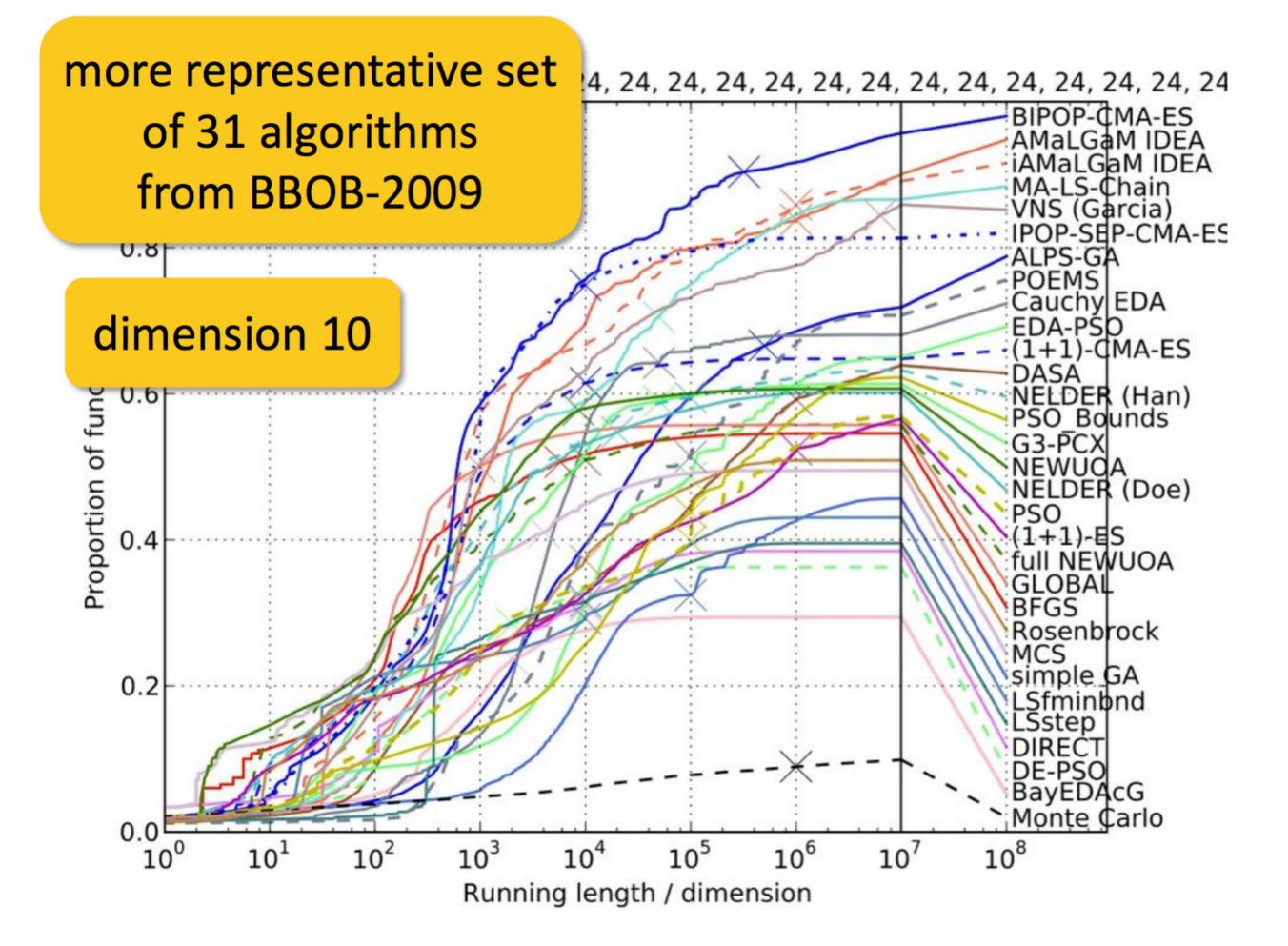
over runs

over test functions

over targets

not over dimension





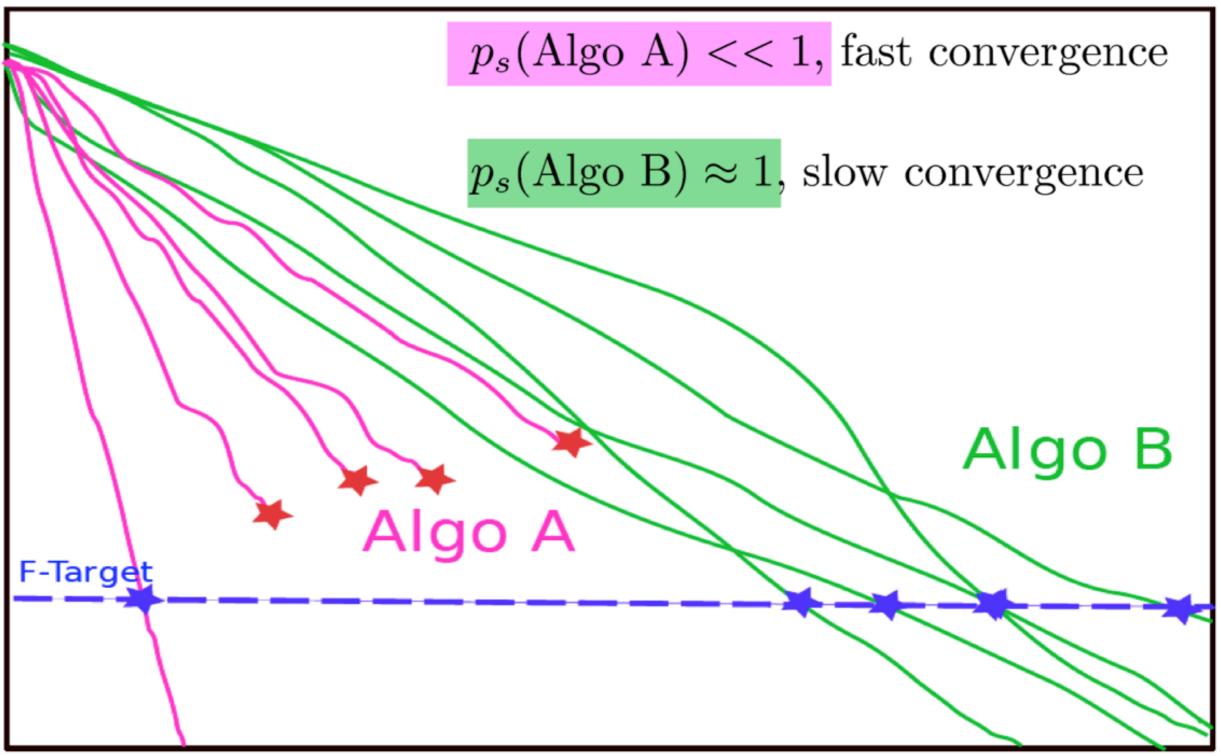
Displaying Performance

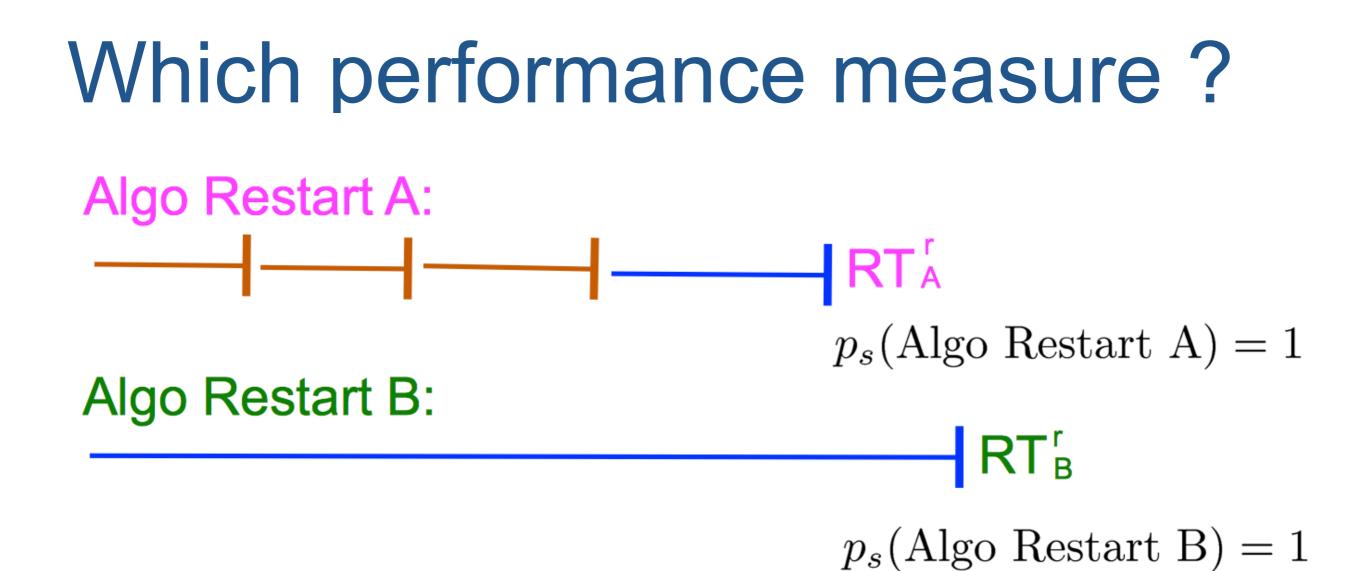
ECDF

Average RunTime (ART)

Which performance measure ?

to compare the two following scenario?





Expected Running Time (restart algo)

 $ERT = E[RT^r] = \frac{1-p_s}{p_s} E[RT_{\text{unsuccessful}} + E[RT_{\text{successful}}]]$

Estimator for ERT

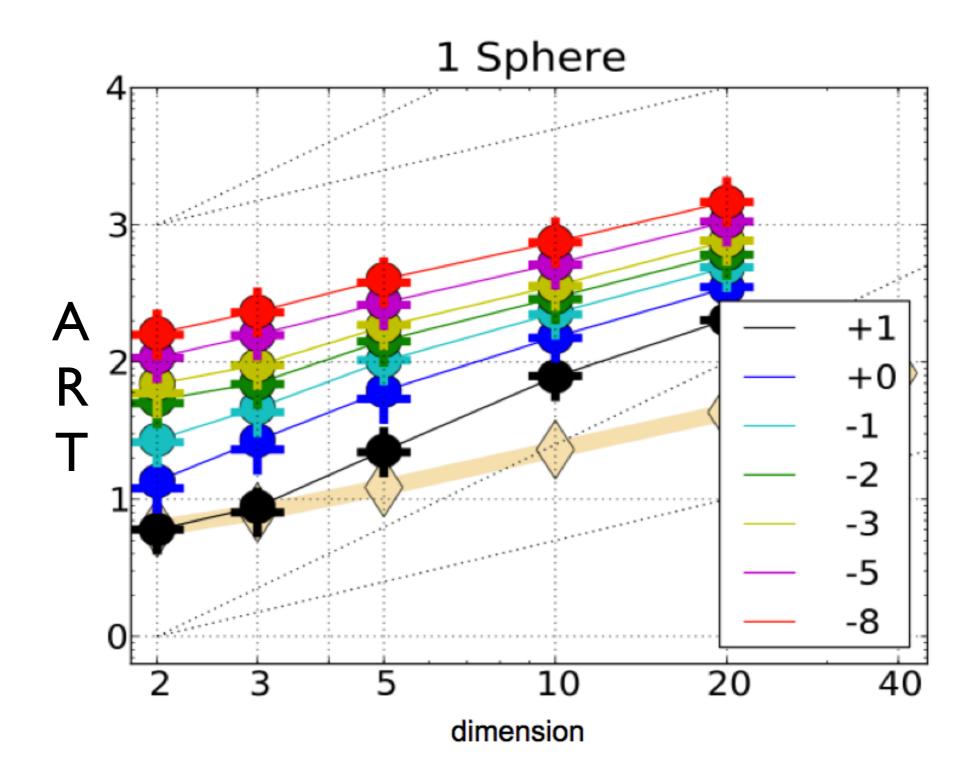
$$\widehat{p_s} = \frac{\#\text{succ}}{\#\text{Runs}}$$

 $\widehat{RT}_{\text{unsucc}} = \text{Average Evals of unsuccessful runs}$

 $\widehat{RT}_{\text{succ}} = \text{Average Evals of successful runs}$

$$ART = \frac{\#Evals}{\#success}$$

Example: scaling behavior



ART on f1 of a variant of CMA-ES – linear scaling

Automatizing the benchmarking COCO platform

COCO platform - COmparing Continuous Optimizers

https://github.com/numbbo/coco

umbbo / coco ● Unwatch + 10 ★ Unstar 9 ♀ Fork • Code ● Issues 111 Pull requests 1 + Pulse ▲ Graphs ♦ Settings • code ● Issues 111 Pull requests 1 + Pulse ▲ Graphs ♦ Settings • code ● Issues 111 Pull requests 1 + Pulse ▲ Graphs ♦ Settings • code ● G.931 commits ● 11 branches ● 15 releases 13 contributors • anch: master • New pull request New file Upload files Find file HTTPS • https://github.com/numble ● Download ZI • takohansen Merge pull request #720 from numbbo/development ■ Latest commit bceaeb2 5 days ag code-postprocessing Stop condition fixed. 6 days ag code-postprocessing Stop condition fixed. 6 days ag clang-format raising an error in bbob2009_logger.c when best_value is NULL. Plus s a year ag hogignore raising an error in bbob2009_logger.c when best_value is NULL. Plus s a year ag AUTHORS minor a month ag LICENSE Create LICENSE 2 months ag	bo/coco: Numerical × +							
This repository Search Pull requests Issues Gist Image: Control of the search of the s	Hub, Inc. (US) https://github.com	/numbbo/coco	C	C Search	☆自			
umbbo / coco Issues 111 P Code Issues 111 P Pull requests 1 Pullse Incident and the pull request 1 New file Incid	ed 🛞 Getting Started 🌠 algorit	hms [COmparin 🔾 numbbo/numbbo · Gi						
Code Issues mini Pull requests 1 Pulse Graphs Settings nerrical Black-Box Optimization Benchmarking Framework http://coco.gforge.inria.fr/ — Edit Image: Code Code 11 branches 15 releases 13 contributors anch: master - New pull request New file Upload files Find file HTTPS - https://github.com/numble Image: Code-experiments nach: master - New pull request #720 from numbbo/development Image: Code-experiments/build/python/cython/interface.c 5 days age code-experiments modified: code-experiments/build/python/cython/interface.c 5 days age 6 days age code-osptprocessing Stop condition fixed. 6 days age docs docs/coco-doc edit 7 days age howtos Update release-howto.md 20 days age clang-format raising an error in bbob2009_logger.c when best_value is NULL. Plus s a year age hygignore raising an error in bbob2009_logger.c when best_value is NULL. Plus s a year age AUTHORS minor a month age LICENSE Create LICENSE 2 months age	This repository Search	Pull requests	s Issues Gist		1	++ 🔛		
Inerical Black-Box Optimization Benchmarking Framework http://coco.gforge.inria.fr/ — Edit Image: Code approximation of the state of	numbbo / coco			⊙ Unwatch -	10 ★ Unstar 9	¥ Fork 12		
Inerical Black-Box Optimization Benchmarking Framework http://coco.gforge.inria.fr/ — Edit Image: Code approximation of the state of								
	<> Code () Issues 111	🖒 Pull requests 👔 🔸 Pulse 📊 Graph	ns 🔅 Settings					
	Numerical Black Box Opti	mization Benchmarking Framework http://co	co aforae inria	fr/ Edit				
Anch: master <	autiencal black-box opti	hization benchinarking Framework http://co	co.giorge.inna.					
Anch: master <								
nikohansen Merge pull request #720 from numbbo/development Latest commit bcea@b2 5 days ag code-experiments modified: code-experiments/build/python/cython/interface.c 5 days ag code-postprocessing Stop condition fixed. 6 days ag docs docs/coco-doc edit 7 days ag howtos Update release-howto.md 20 days ag .clang-format raising an error in bbob2009_logger.c when best_value is NULL. Plus s a year ag .hgignore minor a month ag LICENSE Create LICENSE Create LICENSE	(7) 6,931 commits	∦2 11 branches	°) 15	releases	13 contribut	tors		
nikohansen Merge pull request #720 from numbbo/development Latest commit bcea@b2 5 days ag code-experiments modified: code-experiments/build/python/cython/interface.c 5 days ag code-postprocessing Stop condition fixed. 6 days ag docs docs/coco-doc edit 7 days ag howtos Update release-howto.md 20 days ag .clang-format raising an error in bbob2009_logger.c when best_value is NULL. Plus s a year ag .hgignore minor a month ag LICENSE Create LICENSE Create LICENSE								
nikohansen Merge pull request #720 from numbbo/development Latest commit bcea@b2 5 days ag code-experiments modified: code-experiments/build/python/cython/interface.c 5 days ag code-postprocessing Stop condition fixed. 6 days ag docs docs/coco-doc edit 7 days ag howtos Update release-howto.md 20 days ag .clang-format raising an error in bbob2009_logger.c when best_value is NULL. Plus s a year ag .hgignore minor a month ag LICENSE Create LICENSE Create LICENSE	Development	New Circ. History Circ. The				D		
code-experimentsmodified: code-experiments/build/python/cython/interface.c5 days agcode-postprocessingStop condition fixed.6 days agdocsdocs/coco-doc edit7 days aghowtosUpdate release-howto.md20 days ag.clang-formatraising an error in bbob2009_logger.c when best_value is NULL. Plus sa year agAUTHORSminora month agLICENSECreate LICENSE2 months ag	Branch: master - New p	Il request New file Opload files Find	file HTTPS+	https://github.com	i/numbt 💽 📡	Download ZIP		
code-experimentsmodified: code-experiments/build/python/cython/interface.c5 days agcode-postprocessingStop condition fixed.6 days agdocsdocs/coco-doc edit7 days aghowtosUpdate release-howto.md20 days ag.clang-formatraising an error in bbob2009_logger.c when best_value is NULL. Plus sa year agAUTHORSminora month agLICENSECreate LICENSE2 months ag	🐯 nikohansen Merge pull re	equest #720 from numbbo/development			Latest commit bcea	062 5 days ago		
code-postprocessingStop condition fixed.6 days agdocsdocs/coco-doc edit7 days aghowtosUpdate release-howto.md20 days ag.clang-formatraising an error in bbob2009_logger.c when best_value is NULL. Plus sa year ag.hgignoreraising an error in bbob2009_logger.c when best_value is NULL. Plus sa year agAUTHORSminora month agLICENSECreate LICENSE2 months ag	code-experiments		n/interface.c			5 days ago		
docsdocs/coco-doc edit7 days agehowtosUpdate release-howto.md20 days age.clang-formatraising an error in bbob2009_logger.c when best_value is NULL. Plus sa year age.hgignoreraising an error in bbob2009_logger.c when best_value is NULL. Plus sa year ageAUTHORSminora month ageLICENSECreate LICENSE2 months age	code-postprocessing					6 days ago		
.clang-formatraising an error in bbob2009_logger.c when best_value is NULL. Plus sa year ag.hgignoreraising an error in bbob2009_logger.c when best_value is NULL. Plus sa year agAUTHORSminora month agLICENSECreate LICENSE2 months ag	docs	s docs/coco-doc edit			7 days ago			
Ingignore raising an error in bbob2009_logger.c when best_value is NULL. Plus s a year ag AUTHORS minor a month ag LICENSE Create LICENSE 2 months ag	howtos	Update release-howto.md				20 days ago		
AUTHORS minor a month ag LICENSE Create LICENSE 2 months ag	.clang-format	raising an error in bbob2009_logger.c when best_value is NULL. Plus s				a year ago		
LICENSE Create LICENSE 2 months ag	.hgignore	raising an error in bbob2009_logger.c when bes	raising an error in bbob2009_logger.c when best_value is NULL. Plus s			a year ago		
	AUTHORS	minor				a month ago		
		Create LICENSE				2 months ago		
README.md Update README.md 10 days ag								