



**Universiteit  
Leiden**  
The Netherlands

## **Sampling strategies in automated algorithm configuration**

Anastacio, M.I.A.

### **Citation**

Anastacio, M. I. A. (2026, June 23). *Sampling strategies in automated algorithm configuration*. Retrieved from <https://hdl.handle.net/1887/4307419>

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/4307419>

**Note:** To cite this publication please use the final published version (if applicable).

# List of Figures

3.1	Critical difference diagram of the average score ranks of the configurators. Configurators linked with a black line were not found statistically different. . . . .	38
3.2	Shapley values of the configurators in an oracle portfolio . . . . .	40
3.3	PAR10 values on the train and test set for a subset of our scenarios . . . . .	42
4.1	Cumulative distribution functions for PAR10 scores (in CPU seconds, $x$ -axis) over independent configurator runs on the testing set. . . . .	65
5.1	Accuracy over median running time. $y$ -axis: percentage over all ordered pairs of algorithms in the dataset. $x$ -axis: the time spent running the new algorithm. . . . .	88
5.2	Accuracy over running time used. $y$ -axis: percentage over all ordered pairs of algorithms in the dataset. $x$ -axis: time spent running the new algorithm. . . . .	89
5.3	Accuracy over running time used for the full and reduced SAT20 datasets. $y$ -axis: percentage over all ordered algorithms' pairs in the dataset. $x$ -axis: time spent running the new algorithm. . . . .	90
6.1	Mean accuracy of the Wilcoxon test ( $p < 0.05$ ) on which among $\omega_{\text{ch}}$ and $\omega_{\text{inc}}$ performs best <i>vs</i> the percentage of time spent on evaluations (100% means that all instances of $\mathcal{I}_{\text{known}}$ have been run) . . . . .	108
6.2	Area under the curve of the mean accuracy of the Wilcoxon test ( $p < 0.05$ ) on which among $\omega_{\text{ch}}$ and $\omega_{\text{inc}}$ performs best against the time spent on evaluations. . . . .	109

## List of Figures

---

6.3	Time used (in seconds) before deciding that one configuration is better than the other based on a Wilcoxon test ( $\alpha = 0.05$ ) or reaching a maximum of 10 instance selected . . . . .	110
-----	---	-----

# List of Tables

2.1	List of scenario characteristics . . . . .	20
2.2	Benchmark instance sets characteristics. . . . .	21
2.3	Characteristics of the target algorithms. . . . .	23
2.4	Characteristics of the scenarios. . . . .	25
2.5	Statistics on scenario characteristics . . . . .	26
3.1	Median expected performance of configurators on continuous search space, the best values are underlined and the values statistically tied with the best are boldfaced . . . . .	37
3.2	Correlation between the normalised performance of the configurators and the scenario features . . . . .	43
3.3	Median expected performance of configurators on CPLEX scenarios, the best value for each configurator on a scenario is underlined . . . .	44
4.1	Results for SMAC (left) and irace (right) for default and random initial configurations; median PAR10 (in CPU sec) for Boolean satisfiability (SAT) and mixed integer programming (MIP), 10-fold cross-validated error rate for ML; best results are underlined, while boldface indicates results that are statistically tied to the best, according to a one-sided Mann-Whitney test ( $\alpha = 0.05$ ). . . . .	54
4.2	Target algorithms parameter space description; The total number of numerical parameters; the number of parameters reduced by our techniques and, in parentheses, the number of these conditionally dependent on at least one other parameter. . . . .	56

**List of Tables**

---

4.3 Results for SMAC; median PAR10 (in CPU sec) for SAT, MIP and Planning, 10-fold cross-validated error rate for ML; best results are underlined, while boldface indicates results that are statistically tied to the best, according to a one-sided Mann-Whitney test ( $\alpha = 0.05$ ). . . . 57

4.4 Results for irace and GGA++; median PAR10 (in CPU sec) for SAT, MIP and Planning, 10-fold cross-validated error rate for ML; best results are underlined, while boldface indicates results that are statistically tied to the best, according to a one-sided Mann-Whitney test ( $\alpha = 0.05$ ). 58

4.5 Number of parameters of the best known configuration (testing set) that take a value outside the reduced ranges (using R1 and R2, respectively) 60

4.6 Results for SMAC, SMACPS and irace; median PAR10 (in CPU sec); best results are underlined, while boldface indicates results that are statistically tied to the best, according to a one-sided Mann-Whitney test ( $\alpha = 0.05$ ). Right columns highlight the result of the pairwise comparison between SMACPS and the two baselines; ✓ if it is better, ✗ if not; parentheses indicate that the difference is not statistically significant. . . . . 64

5.1 Characteristics of the used datasets . . . . . 82

5.2 Median log likelihood of Maximum Likelihood Estimation for Levy and Cauchy distributions over the instances of each dataset. The highest likelihood for each dataset is shown in boldface. . . . . 85

6.1 Benchmark instance sets characteristics. . . . . 103

6.2 Median time in seconds for each method over every tested prior data, with lowest medians boldfaced (statistical significance according to a permutation test with  $\alpha = 0.05$ ). . . . . 110

6.3 Median performance of SMAC-IS with the selection methods random (rand), variance-based (var) and discrimination-based (disc) at both phases. Boldfaced values are better than those for vanilla SMAC. The lowest median is underlined. All underlined medians are significantly different from others based on a Mann-Whitney U-test ( $\alpha = 0.05$ ). . . 113

6.4 Comparison of SMAC and SMAC-W, respectively, without and with a Wilcoxon test to decide whether a challenger configuration should be kept longer. . . . . 114

# List of Algorithms

2.1	ParamILS . . . . .	14
2.2	GGA++ . . . . .	15
2.3	irace (Based on Algorithm 1 from <a href="#">Birattari et al. (2010)</a> ) . . . . .	16
2.4	SMAC . . . . .	18
2.5	GPS . . . . .	19
4.1	SMACPS . . . . .	62
5.1	PSEAS solving strategy . . . . .	75
5.2	Correcting timeouts for a sample $(t_{I,A})_{A \in \mathcal{A}}$ . . . . .	84
6.1	Intensification for one challenger (based on SMAC <a href="#">Hutter et al. (2011b)</a> ) . . . . .	99



# List of Symbols

$A$  The algorithm to configure

$T_{\text{cut}}$  Target algorithm cutoff time

$\Omega$  The set of valid configurations

$\omega$  A valid configuration from  $\Omega$

$\omega_{\text{inc}}$  The current best known configuration

$\omega_{\text{ch}}$  A challenger configuration to evaluate

$\mathcal{P}$  A parameter from the configuration space  $\Omega$

$\mathcal{D}$  The domain of possible values of a parameter

$\mathcal{I}$  A set of problem instances

$I$  A single instance

$\mathbf{f}$  The feature values of an instance  $I$

$rt$  The running time of an algorithm applied with a specific configuration  $\omega$  on a specific instance  $I$

$\mathcal{M}$  Performance of the algorithm at hand, applied with a specific configuration  $\omega$  on a set of instances  $\mathcal{I}$  (typically  $\mathcal{M} = rt$ )

$\mathcal{C}$  A set of configurators

$C$  A configurator

## LIST OF SYMBOLS

---

$\mathcal{K}$  A set of configuration scenarios

$K$  A configuration scenario

$C_{thres}$  a confidence threshold to take a decision