On the empirical scaling of running time of Lingeling for solving the csmacdp0neg HWMCC08 circuit bounded model checking SAT instances

Empirical Scaling Analyzer

7th February 2022

1 Introduction

This is the automatically generated report on the empirical scaling of the running time of Lingeling for solving the csmacdp0neg HWMCC08 circuit bounded model checking SAT instances.

2 Methodology

For our scaling analysis, we considered the following parametric models:

- $Exp\left[a,b\right](n) = a \times b^n$ (2-parameter Exp)
- $Poly\left[a,b\right](n) = a \times n^b$ (2-parameter Poly)

Note that the approach could be easily extended to other scaling models. For fitting parametric scaling models to observed data, we used iteratively re-weighted linear least squres to perform quantile regression. Since this method works best for linear models, we used log transformations to convert non-linear models into linear models. This transformation biases the fitted models to more heavily favour the smaller training instance sizes, so we used a heuristic error correction term to compensate. Preliminary studies using this fitting method when applied to simulated running time datasets with known scaling properties show that using the heuristic error correction term improves the quality of the fitted models and allows the procedure to fit scaling models consistent with the true, underlying scaling of the data.

The fitted models correspond to performance predictions for the empirical scaling of the median of the distribution of running times. Since Lingeling is a randomized algorithm, we analyzed the per-instance median of 10 independent runs for each instance. This means that the model predictions correspond to medians of per-instance medians. Similarly, the running time statistics reported throughout this report are statistics of per-instance medians. To assess the fit of a given scaling model to observed data, we used the mean absolute error as a loss function.

Closely following [1], we computed 95.0% bootstrap confidence intervals for the performance predictions obtained from our scaling models, based on 51 bootstrap samples per instance set and 51 automatically fitted variants of each scaling model. To extend this idea, we calculated training and challenge losses for each of the fitted models' predictions and the corresponding bootstrap samples of the observed data. We used these bootstrap sample losses to calculate median and 95% confidence intervals of the support and challenge losses for each model. Since this analysis was performed on per-instance medians, we also computed these statistics on nested, per-instance bootstrap samples, by

n	221109	451076	681043
# instances	65.66	67.77	48.65
mean	17.99	57.15	151.4
Q(0.1)	12.11	34.48	97.23
Q(0.25)	14.06	41.24	111.5
median	16.89	54.75	135.7
Q(0.75)	20.53	68.99	177.9
Q(0.9)	26.52	87.85	208.7

Table 1: Details of the running time dataset used as support data for model fitting. The reported statistics are of the per-instance median running times. The "# of instances" is the sum of the weights of the instances used to calculate these statistics.

first computing medians for 21 bootstrap samples for each instance and then randomly selecting one of the per-instance medians when needed.

We calculated the observed point estimates for the medians of the data by fitting a linear model to local data with Guassian weights, and then recording the observed statistic as the prediction from the linear model as the mid-point of the local data. In the following, we say that a scaling model prediction is in-consistent with observed data if the bootstrap confidence interval for the observed data is disjoint from the bootstrap confidence interval for the predicted median of per-instance median running times; we say that a scaling model prediction is weakly consistent with the observed data if the bootstrap confidence interval for the prediction overlaps with the bootstrap confidence interval for the observed data; and, we say that a scaling model is strongly consistent with observed data, if the bootstrap confidence interval for the observed median of per-instance medians is fully contained within the bootstrap confidence interval for predicted running times. Also, we define the residue of a model at a given size as the observed point estimate less the predicted value using the fitted running time scaling model (fitted to the set of training data).

3 Dataset Description

The dataset contains running times of the Lingeling algorithm solving 500 instances of different sizes with 10 independent runs per instance. We split the running times into two categories, *support* or *training* instances ($n \leq 757881$) and *challenge* or *test* instances ($n > 757881$) with 151 and 349 instances, respectively. The details of the dataset can be found in Tables 1 and 2.

4 Empirical Scaling of Solver Performance

We first fitted our parametric scaling models to the medians of the per-instance median running times of Lingeling, as described in Section 2. The models were fitted using the training instance data and later challenged with the test instance data. This resulted in the models, shown along with losses on support and challenge data, shown in Table 3. In addition, we illustrate the fitted models of Lingeling in Figure 1, and the residues for the models in Figure 2.

But how much confidence should we have in these models? Are the losses small enough that we should accept them? To answer this question, we assessed the fitted models using the bootstrap approach outlined in Section 2. Table 4 shows the bootstrap intervals of the model parameters, Table 5 shows the bootstrap intervals of the model prediction losses, and Table 6 contains the bootstrap intervals for the support data. Challenging the models with extrapolation, as shown in Table 7, it

n	911010	1140977	1370944	1600911
# instances	58.78	67.77	67.77	67.91
mean	301.4	479.7	692.1	958.5
Q(0.1)	188	237.7	434.7	676.4
Q(0.25)	218.3	336.3	543.2	755.1
median	305.5	509	704.4	990.1
Q(0.75)	374.8	632.4	879.8	1123
Q(0.9)	425.1	661.2	986.7	1398

n	1830878	2060845	2290812	2520779
# instances	67.91	67.91	66.73	39.14
mean	1356	2157	2837	3297
Q(0.1)	1356	1649	2198	2745
Q(0.25)	1356	2118	2574	3023
median	1356	2166	2897	3420
Q(0.75)	1356	2568	3152	3597
Q(0.9)	1356	2771	3355	3731

Table 2: Details of the running time dataset used as challenge data for model fitting. The reported statistics are of the per-instance median running times. The "# of instances" is the sum of the weights of the instances used to calculate these statistics.

		Model	Support loss	Challenge loss
Lingeling	Exp. Model	$4.120941 \times 1.000005^{n}$	1047.5	4.9353×10^{7}
	Poly. Model	$\mathbf{8.185079 \times 10^{-11} \times n^{2.089114}}$	**1099.6**	**98958**

Table 3: Fitted models of the medians of the per-instance median running times and loss values (in CPU sec). The models yielding the most accurate predictions (as per losses on challenge data) are shown in boldface.

is concluded that the Exp model over-estimates the data, and the Poly model fits the data very well (as also illustrated in Figure 1). We base these statements on an analysis of the fraction of predicted bootstrap intervals that are strongly consistent, weakly consistent and disjoint from the observed bootstrap intervals for the challenge data. To provide stronger emphasis for the largest instance sizes, we also consider these fractions for the largest half of the challenge instance sizes. To be precise, we say a model over-estimates the data if $\geq 70\%$ of the confidence intervals for predictions on all challenge instance sizes or $\geq 70\%$ of those on the larger half of the challenge sizes are above the observed intervals; and we say a model predicts very well if $\geq 90\%$ of the predictions for challenge sizes are strongly consistent, or $\geq 90\%$ of the predictions for the larger half of the challenge sizes are strongly consistent and $\geq 90\%$ of all of the predictions for all challenge sizes are weakly consistent.

Figure 1: Fitted models of the medians of the per-instance median running times. The models correspond to predictions for the medians of the per-instance median running times of Lingeling solving the set of the csmacdp0neg HWMCC08 circuit bounded model checking SAT instances with $200 \leq n \leq 757881$ variables, and are challenged by the medians of the per-instance median running times of $757881 < n \leq 2545110$ variables.

Figure 2: Residues of the fitted models of the medians of the per-instance median running times.

Solver	Model	Confidence interval of a	Confidence interval of b
Lingeling	Exp.	$[3.410004, 5.023061]$	$[1.000005, 1.000006]$
	Poly.	$[4.787953 \times 10^{-13}, 3.29137 \times 10^{-9}]$	$[1.80878, 2.490952]$

Table 4: 95% bootstrap intervals of model parameters for the medians of the per-instance median running times

Solver	Model	Support Loss	Challenge Loss
Lingeling	Exp.	$[1027.4, 1560]$	$[1.827 \times 10^7, 1.1063 \times 10^8]$
	Poly.	$[\mathbf{1034.5}, \mathbf{1626.2}]$	$[\mathbf{1.1245 \times 10^5}, \mathbf{1.6624 \times 10^5}]$
		$\vdots$	

Table 5: 95% bootstrap confidence intervals of model prediction losses for the medians of the per-instance median running times.The model with the smallest lower bound is shown in boldface, as well as any models with overlapping intervals.

5 Conclusion

In this report, we presented an empirical analysis of the scaling behaviour of Lingeling on the csmacdp0neg HWMCC08 circuit bounded model checking SAT instances. We found the Exp model over-estimates the data, and the Poly model fits the data very well.

References

[1] Yasha Pushak and Holger H. Hoos. Advanced Statistical Analysis of Empirical Performance Scaling. in *Proceedings of the 22nd Genetic and Evolutionary Computation Conference*, (GECCO 2020), pages 236–244, 2020.

[2] Yasha Pushak, Zongxu Mu and Holger H. Hoos. Empirical scaling analyzer: An automated system for empirical analysis of performance scaling. *AI Communications* – to appear, 2020.

[3] Jérémie Dubois-Lacoste, Holger H. Hoos, and Thomas Stützle. On the empirical scaling behaviour of state-of-the-art local search algorithms for the Euclidean TSP. In *Proceedings of the 17th Genetic and Evolutionary Computation Conference*, (GECCO 2015), pages 377–384, 2015.

[4] Holger H. Hoos. A bootstrap approach to analysing the scaling of empirical run-time data with problem size. Technical report, Technical Report TR-2009-16, Department of Computer Science, University of British Columbia, 2009.

[5] Holger H. Hoos and Thomas Stützle. On the empirical scaling of run-time for finding optimal solutions to the travelling salesman problem. *European Journal of Operational Research*, 238(1):87–94, 2014.

Solver	n	Predicted confidence intervals Exp. model	Observed median run-time Point estimates	Observed median run-time Confidence intervals
Lingeling	221109	$[11.88, 14.45]$	16.89	$[15.75, 18.78]$
	313095	$[19.86, 22.88]$	27.81	$[26.34, 29.4]$
	405081	$[32.05, 36.47]$	42.86	$[38.7, 49.8]$
	497067	$[\mathbf{50.92}, \mathbf{60.39}]$	66.91	$[58.92, 76.08]$
	589053	$[\mathbf{81.25}, \mathbf{99.89}]$	96.2	$[84.19, 108.8]$
	681039	$[\mathbf{125.6}, \mathbf{169.3}]$	135.7	$[118, 169.7]$
Solver	n	Predicted confidence intervals Poly. model	Observed median run-time Point estimates	Observed median run-time Confidence intervals
Lingeling	221109	$[9.416, 15.3]$	16.89	$[15.75, 18.78]$
	313095	$[\mathbf{21.41}, \mathbf{29.85}]$*	27.81	$[26.34, 29.4]$
	405081	$[\mathbf{39.35}, \mathbf{49.29}]$	42.86	$[38.7, 49.8]$
	497067	$[\mathbf{61.19}, \mathbf{74.15}]$	66.91	$[58.92, 76.08]$
	589053	$[\mathbf{83.54}, \mathbf{110}]$*	96.2	$[84.19, 108.8]$
	681039	$[\mathbf{109}, \mathbf{154.8}]$	135.7	$[118, 169.7]$

Table 6: 95% bootstrap confidence intervals for the medians of the per-instance median running time predictions and observed running times on the csmacdp0neg HWMCC08 circuit bounded model checking SAT instances. The instance sizes shown here are those used for fitting the models. Bootstrap intervals on predictions that are weakly consistent with the observed point estimates are shown in boldface and those that are strongly consistent are marked by asterisks (*).

Solver	n	Predicted confidence intervals Exp. model	Observed median run-time Point estimates	Observed median run-time Confidence intervals
Lingeling	773025	$[\mathbf{195.9}, \mathbf{284.5}]$	159.9	$[123.9, 222.5]$
	865011	$[307.4, 476.9]$	253	$[229.8, 287.2]$
	956997	$[475.8, 799.4]$	354.8	$[306.2, 382.6]$
	1048983	$[736.4, 1340]$	435.3	$[378.4, 466]$
	1140969	$[1140, 2247]$	509	$[442.7, 547]$
	1232955	$[1764, 3766]$	566.8	$[509.7, 647.6]$
	1324941	$[2730, 6313]$	657.3	$[591, 755.7]$
	1416927	$[4225, 1.058 \times 10^4]$	766.8	$[711.8, 862.4]$
	1508913	$[6540, 1.774 \times 10^4]$	890.5	$[815, 998.7]$
	1600899	$[1.012 \times 10^4, 2.974 \times 10^4]$	990.1	$[895.6, 1093]$
	1692885	$[1.567 \times 10^4, 4.985 \times 10^4]$	1103	$[1008, 1236]$
	1784871	$[2.425 \times 10^4, 8.357 \times 10^4]$	1243	$[1106, 1382]$
	1876857	$[3.753 \times 10^4, 1.401 \times 10^5]$	1522	$[1315, 1604]$
	1968843	$[5.808 \times 10^4, 2.348 \times 10^5]$	1814	$[1646, 1927]$
	2060829	$[8.99 \times 10^4, 3.937 \times 10^5]$	2166	$[2013, 2335]$
	2152815	$[1.391 \times 10^5, 6.599 \times 10^5]$	2534	$[2350, 2638]$
	2244801	$[2.154 \times 10^5, 1.106 \times 10^6]$	2785	$[2627, 2842]$
	2336787	$[3.333 \times 10^5, 1.854 \times 10^6]$	3005	$[2873, 3121]$
	2428773	$[5.159 \times 10^5, 3.109 \times 10^6]$	3221	$[3033, 3371]$
	2520759	$[7.984 \times 10^5, 5.211 \times 10^6]$	3420	$[3201, 3574]$

Solver	n	Predicted confidence intervals Poly. model	Observed median run-time Point estimates	Observed median run-time Confidence intervals
Lingeling	773025	$[\mathbf{137.5}, \mathbf{211.1}]$	159.9	$[123.9, 222.5]$
	865011	$[\mathbf{169}, \mathbf{285.8}]$	253	$[229.8, 287.2]$
	956997	$[\mathbf{204}, \mathbf{375.3}]$	354.8	$[306.2, 382.6]$
	1048983	$[\mathbf{242}, \mathbf{476}]$*	435.3	$[378.4, 466]$
	1140969	$[\mathbf{283.1}, \mathbf{586.9}]$*	509	$[442.7, 547]$
	1232955	$[\mathbf{327.1}, \mathbf{711.9}]$*	566.8	$[509.7, 647.6]$
	1324941	$[\mathbf{374}, \mathbf{851.6}]$*	657.3	$[591, 755.7]$
	1416927	$[\mathbf{423.8}, \mathbf{1007}]$*	766.8	$[711.8, 862.4]$
	1508913	$[\mathbf{474.1}, \mathbf{1177}]$*	890.5	$[815, 998.7]$
	1600899	$[\mathbf{526.4}, \mathbf{1364}]$*	990.1	$[895.6, 1093]$
	1692885	$[\mathbf{581}, \mathbf{1568}]$*	1103	$[1008, 1236]$
	1784871	$[\mathbf{638}, \mathbf{1789}]$*	1243	$[1106, 1382]$
	1876857	$[\mathbf{699.6}, \mathbf{2028}]$*	1522	$[1315, 1604]$
	1968843	$[\mathbf{763.8}, \mathbf{2284}]$*	1814	$[1646, 1927]$
	2060829	$[\mathbf{830.5}, \mathbf{2559}]$*	2166	$[2013, 2335]$
	2152815	$[\mathbf{899.7}, \mathbf{2853}]$*	2534	$[2350, 2638]$
	2244801	$[\mathbf{971.5}, \mathbf{3167}]$*	2785	$[2627, 2842]$
	2336787	$[\mathbf{1046}, \mathbf{3500}]$*	3005	$[2873, 3121]$
	2428773	$[\mathbf{1122}, \mathbf{3853}]$*	3221	$[3033, 3371]$
	2520759	$[\mathbf{1202}, \mathbf{4227}]$*	3420	$[3201, 3574]$

Table 7: 95% bootstrap confidence intervals for the medians of the per-instance median running time predictions and observed running times on the csmacdp0neg HWMCC08 circuit bounded model checking SAT instances. The instance sizes shown here are larger than those used for fitting the models. Bootstrap intervals on predictions that are weakly consistent with the observed data are shown in boldface and those that are strongly consistent are marked by asterisks (*).