On the empirical scaling of running time of c2d for solving random instances with mu = 1.0

Empirical Scaling Analyzer

7th February 2022

1 Introduction

This is the automatically generated report on the empirical scaling of the running time of c2d for solving random instances with mu = 1.0.

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. 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 1001 bootstrap samples per instance set and 1001 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.

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 running times; we say that a scaling model prediction is weakly consistent with the observed data if the bootstrap confidence interval for

n	10	11	12	13
# instances	213	368	499	579
mean	0.389	0.431	0.482	0.5948
Q(0.1)	0.3406	0.37	0.4005	0.4605
Q(0.25)	0.3601	0.39	0.4303	0.5001
median	0.3801	0.42	0.47	0.5599
Q(0.75)	0.4098	0.4501	0.5199	0.6302
Q(0.9)	0.4396	0.4993	0.5797	0.7101

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

n	14	15	16
# instances	651	667	644
mean	0.7437	1.043	1.578
Q(0.1)	0.5505	0.7101	0.9398
Q(0.25)	0.6198	0.8002	1.16
median	0.7099	0.98	1.47
Q(0.75)	0.83	1.22	1.82
Q(0.9)	0.9692	1.46	2.4

n	17	18
# instances	620	416
mean	2.325	3.394
Q(0.1)	1.37	1.821
Q(0.25)	1.71	2.29
median	2.14	3.09
Q(0.75)	2.7	4.09
Q(0.9)	3.48	5.249

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

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 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 c2d algorithm solving 5100 instances of different sizes . We split the running times into two categories, *support* or *training* instances ($n \leq 13$) and *challenge* or *test* instances ($n > 13$) with 1801 and 3299 instances, respectively. The details of the dataset can be found in Tables 1 and 2.

		Model	Support loss	Challenge loss
c2d	Exp. Model	$\mathbf{0.1302142 \times 1.115612^n}$	**57.745**	**2153.2**
	Poly. Model	$0.02637194 \times n^{1.169902}$	58.682	2297.3

Table 3: Fitted models of the medians of the 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.

Figure 1: Fitted models of the medians of the running times. The models correspond to predictions for the medians of the running times of c2d solving the set of random instances with mu = 1.0 with $7 \leq n \leq 13$ variables, and are challenged by the medians of the running times of $13 < n \leq 21$ variables.

4 Empirical Scaling of Solver Performance

We first fitted our parametric scaling models to the medians of the running times of c2d, 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 c2d 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 is concluded that the Exp model tends to under-estimate the data, and the Poly model tends to under-estimate the data (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, ; and we say a model tends to under-estimate the data if $> 10\%$ of the confidence intervals for predictions on challenge instance sizes are disjoint from the confidence intervals for observed running time data and $\geq 90\%$ of the predicted intervals are below or are consistent with the observed intervals.

5 Conclusion

In this report, we presented an empirical analysis of the scaling behaviour of c2d on random instances with mu = 1.0. We found the Exp model tends to under-estimate the data, and the Poly model tends to under-estimate the data.

Figure 2: Residues of the fitted models of the medians of the running times.

Solver	Model	Confidence interval of a	Confidence interval of b
c2d	Exp.	$[0.1200647, 0.1375862]$	$[1.109363, 1.122611]$
	Poly.	$[0.02093711, 0.02931343]$	$[1.128618, 1.260531]$

Table 4: 95% bootstrap intervals of model parameters for the medians of the running times

Solver	Model	Support Loss	Challenge Loss
c2d	Exp.	$[\mathbf{50.798}, \mathbf{68.913}]$	$[\mathbf{2130.4}, \mathbf{2202}]$
	Poly.	$[51.684, 69.766]$	$[2308.8, 2366.1]$

:

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

Solver	n	Predicted confidence intervals	Observed median run-time	
		Exp. model	Point estimates	Confidence intervals
c2d	10	$[\mathbf{0.3817}, \mathbf{0.3901}]$	0.3801	$[0.3799, 0.3899]$

Solver	n	Predicted confidence intervals	Observed median run-time	
		Poly. model	Point estimates	Confidence intervals
c2d	10	$[\mathbf{0.3827}, \mathbf{0.3946}]$	0.3801	$[0.3799, 0.3899]$

Table 6: 95% bootstrap confidence intervals for the medians of the running time predictions and observed running times on random instances with mu = 1.0. 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 (*).

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	Observed median run-time	
		Exp. model	Point estimates	Confidence intervals

Solver	n	Predicted confidence intervals	Observed median run-time	
		Poly. model	Point estimates	Confidence intervals

Table 7: 95% bootstrap confidence intervals for the medians of the running time predictions and observed running times on random instances with mu = 1.0. 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 (*).