An experimental study of adaptive capping in irace

Leslie Pérez Cáceres[1], Manuel López-Ibáñez[2], Holger Hoos[3], and Thomas Stützle[1]

[1] IRIDIA, Université Libre de Bruxelles, Brussels, Belgium
{leslie.perez.caceres,stuetzle}@ulb.ac.be

[2] Alliance Manchester Business School, University of Manchester, UK
manuel.lopez-ibanez@manchester.ac.uk

[3] Computer Science Department, University of British Columbia, Vancouver, Canada
hoos@cs.ubc.cs

Abstract. The irace package is a widely used for automatic algorithm configuration and implements various iterated racing procedures. The original irace was designed for the optimisation of the solution quality reached within a given running time, a situation frequently arising when configuring algorithms such as stochastic local search procedures. However, when applied to configuration scenarios that involve minimising the running time of a given target algorithm, irace falls short of reaching the performance of other general-purpose configuration approaches, since it tends to spend too much time evaluating poor configurations. In this article, we improve the efficacy of irace in running time minimisation by integrating an adaptive capping mechanism into irace, inspired by the one used by ParamILS. We demonstrate that the resulting irace$_{\text{cap}}$ reaches performance levels competitive with those of state-of-the-art algorithm configurators that have been designed to perform well on running time minimisation scenarios. We also investigate the behaviour of irace$_{\text{cap}}$ in detail and contrast different ways of integrating adaptive capping.

A Experimental evaluation alternatives

The experiments performed with irace in Section 5 use a non-penalized evaluation, that is, timed out executions are not penalized by a factor in the evaluation. As discussed in Section 6, the PAR10 evaluation is commonly applied to configure exact algorithms. This section gives the results of the experiments performed in Section 5 (using PAR1 in the training), with a PAR10 and PAR100 evaluation on the testing.

Table A.1 gives the performance of the experiments that compare irace$_{\text{cap}}$ and irace using the non-penalized (PAR1) evaluation and using PAR10 and PAR100 on the testing. Additionally, Figure A.1 shows the results of irace$_{\text{cap}}$ and irace with PAR1 and PAR10 testing evaluation. Since the PARX evaluation increments the differences between non timed out and timed out evaluations, the results change depending of the scenario. The best means are maintained for all the scenarios reflecting that irace produces more time outs than irace$_{\text{cap}}$. For the Regions 100 scenario, depending on the size of the penalty there will be a significant difference between the results of irace$_{\text{cap}}$ and irace. This is an indication that the size of the penalty used is important and probably scenario dependent.

	Regions 100		Regions 200		Corlat		Lingeling		Spear	
PAR1 evaluation										
	irace$_{cap}$	irace	irace$_{cap}$	irace	irace$_{cap}$	irace	irace$_{cap}$	irace	irace$_{cap}$	irace
q25	0.327	0.374	9.487	10.983	8.616	13.526	42.379	44.274	3.028	4.776
mean	**0.338**	0.395	**10.498**	13.231	**11.899**	15.935	*45.501*	46.923	**4.116**	13.068
median	0.332	0.401	10.469	12.871	9.688	14.911	44.453	47.034	3.765	14.617
q75	0.34	0.413	10.75	14.256	13.941	18.436	48.996	49.738	4.242	19.993
sd	0.018	0.033	1.335	2.908	5.645	4.325	3.799	3.658	1.848	8.092
sd/mean	0.054	0.082	0.127	0.22	0.474	0.271	0.083	0.078	0.449	0.619
p-value	5.7e-06		0.0001049		0.0055809		0.2942524		0.0002613	
PAR10 evaluation										
	irace$_{cap}$	irace	irace$_{cap}$	irace	irace$_{cap}$	irace	irace$_{cap}$	irace	irace$_{cap}$	irace
q25	0.336	0.411	9.487	10.983	9.652	25.01	240.209	243.003	3.565	21.98
mean	**0.374**	0.433	**10.768**	13.636	**30.663**	48.465	*271.599*	280.617	**14.844**	102.915
median	0.371	0.428	10.469	13.259	21.836	49.279	263.347	289.227	12.421	80.301
q75	0.385	0.447	10.75	14.984	31.585	74.975	299.313	317.201	15.621	189.578
sd	0.049	0.042	2.305	3.514	34.684	28.861	33.462	44.603	16.352	85.331
sd/mean	0.13	0.096	0.214	0.258	1.131	0.595	0.123	0.159	1.102	0.829
p-value	0.0005856		0.0001049		0.0120792		0.5216732		0.0003223	
PAR100 evaluation										
	irace$_{cap}$	irace	irace$_{cap}$	irace	irace$_{cap}$	irace	irace$_{cap}$	irace	irace$_{cap}$	irace
q25	0.336	0.411	9.487	10.983	29.902	133.01	2207.096	2209.891	3.565	200.788
mean	*0.734*	0.816	**13.468**	17.686	**218.313**	373.815	*2533.52*	2618.531	**122.128**	1001.425
median	0.821	0.451	10.469	13.259	143.336	386.779	2453.744	2703.134	101.825	750.831
q75	0.835	1.328	10.75	15.143	224.475	648.725	2802.624	2999.32	127.376	1888.253
sd	0.421	0.522	14.178	12.605	326.961	279.7	332.22	456.108	163.096	860.637
sd/mean	0.574	0.641	1.053	0.713	1.498	0.748	0.131	0.174	1.335	0.859
p-value	0.4304333		0.0031528		0.0136166		0.5216732		0.0003223	
%timeout	0.08	0.085	0.01	0.015	0.695	1.205	8.377	8.659	0.397	3.328

Table A.1: Statistics of the mean PAR10 and PAR100 execution time and percentage of timed out evaluations of 20 executions of irace$_{cap}$ and irace over the test set. Wilcoxon test p-values (significance 0.05). Significantly better results in bold and best mean in cursive.

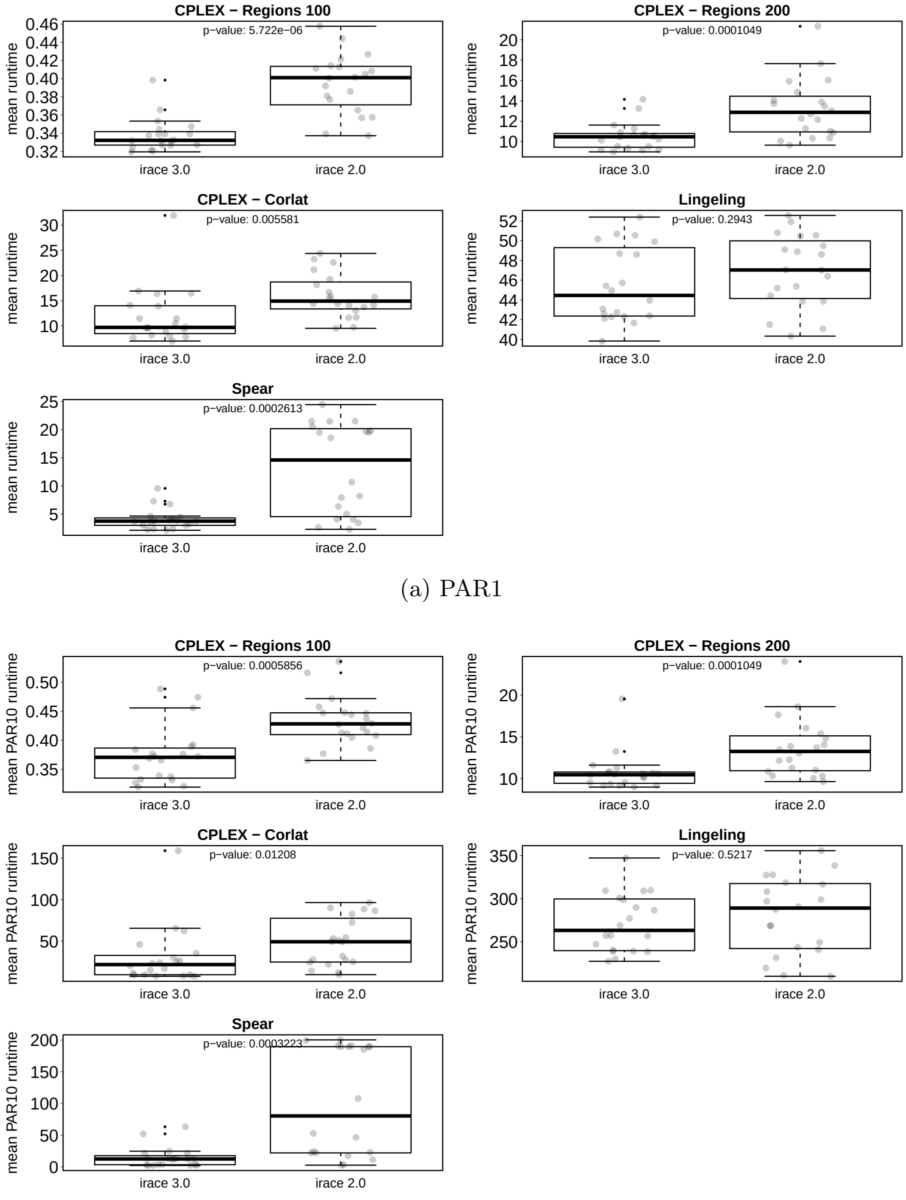
In the following we give the results of all the experiments performed applying PAR1, PAR10 and PAR100 in the test set evaluation. Table A.2 and Figure A.3 gives the results of irace$_{cap}$ with initial instance shuffling disabled (*no shuf.*) and *default* irace$_{cap}$ (shuffling enabled). The results are only significantly different for the Regions 200, Lingeling and Spear scenarios and this is generally mantained across the evaluations. For Regions 200 and Lingeling disabling the shuffling obtains the best results, while for Spear the shuffling seems to very important to obtain good performance. The Corlat scenario shows significant differences only when using a high penalty (PAR100).

Table A.3 and Figure A.4 gives the results of irace$_{cap}$ setting the confidence of the statistical test as 0.95 (default) and 0.75. Additionally, Table A.4 gives the statistics of the execution of irace$_{cap}$ using th two confidence settings.

Table A.5 and Figure A.6 compare the results of irace$_{cap}$ (*default*) and a modified version in which the statistical test is disabled.

	Regions 100		Regions 200		Corlat		Lingeling		Spear	
	default	*no shuf.*	*default*	*no shuf.*	*default*	*no shuf.*	*default*	*no shuf.*	*default*	*no shuf.*
q25	0.327	0.32	9.487	8.88	8.616	8.473	42.379	41.51	3.028	6.495
mean	0.338	*0.333*	10.498	**9.301**	11.899	*9.977*	45.501	**42.987**	**4.116**	22.263
median	0.332	0.325	10.469	9.233	9.688	9.557	44.453	42.31	3.765	13.872
q75	0.34	0.339	10.75	9.615	13.941	11.452	48.996	44.515	4.242	26.97
sd	0.018	0.02	1.335	0.71	5.645	2.373	3.799	2.044	1.848	21.853
sd/mean	0.054	0.059	0.127	0.076	0.474	0.238	0.083	0.048	0.449	0.982
p-value	0.3117943		0.0001335		0.4304333		0.0239506		3.8e-06	
PAR10										
	default	*no shuf.*	*default*	*no shuf.*	*default*	*no shuf.*	*default*	*no shuf.*	*default*	*no shuf.*
q25	0.336	0.324	9.487	8.88	9.652	10.969	240.209	229.802	3.565	15.433
mean	0.374	*0.369*	10.768	**9.436**	30.663	*16.457*	271.599	**246.293**	**14.844**	169.323
median	0.371	0.368	10.469	9.233	21.836	13.326	263.347	241.073	12.421	85.219
q75	0.385	0.384	10.75	9.853	31.585	17.498	299.313	252.407	15.621	243.758
sd	0.049	0.042	2.305	0.968	34.684	10.926	33.462	18.444	16.352	196.364
sd/mean	0.13	0.115	0.214	0.103	1.131	0.664	0.123	0.075	1.102	1.16
p-value	0.9854355		0.0023251		0.0582581		0.0120792		3.62e-05	
PAR100										
	default	*no shuf.*	*default*	*no shuf.*	*default*	*no shuf.*	*default*	*no shuf.*	*default*	*no shuf.*
q25	0.336	0.324	9.487	8.88	29.902	10.969	2207.096	2107.286	3.565	104.837
mean	0.734	*0.729*	13.468	**10.786**	218.313	**81.257**	2533.52	**2280.233**	**122.128**	1640.018
median	0.821	0.815	10.469	9.233	143.336	53.84	2453.744	2207.96	101.825	800.451
q75	0.835	0.826	10.75	9.853	224.475	79.117	2802.624	2331.05	127.376	2411.804
sd	0.421	0.384	14.178	6.735	326.961	104.062	332.22	187.222	163.096	1943.123
sd/mean	0.574	0.526	1.053	0.624	1.498	1.281	0.131	0.082	1.335	1.185
p-value	0.9854355		0.0036545		0.0484409		0.0120792		3.62e-05	
%timeout	0.08	0.08	0.01	0.005	0.695	0.24	8.377	7.533	0.397	5.447

Table A.2: Statistics of the mean PAR10 and PAR100 execution time and percentage of timed out instances of 20 executions of $\text{irace}_{\text{cap}}$ with default settings (*default*) and a version with the initial instance shuffling disabled (*no shuf.*). Wilcoxon test p-values (significance 0.05). Significantly better results in bold and best mean in cursive.



(a) PAR1

(b) PAR10

Fig. A.1: Mean PAR1 and PAR10 performance of 20 executions of irace and $\text{irace}_{\text{cap}}$ across the test set. Wilcoxon test (significance 0.05) p-values on each plot.

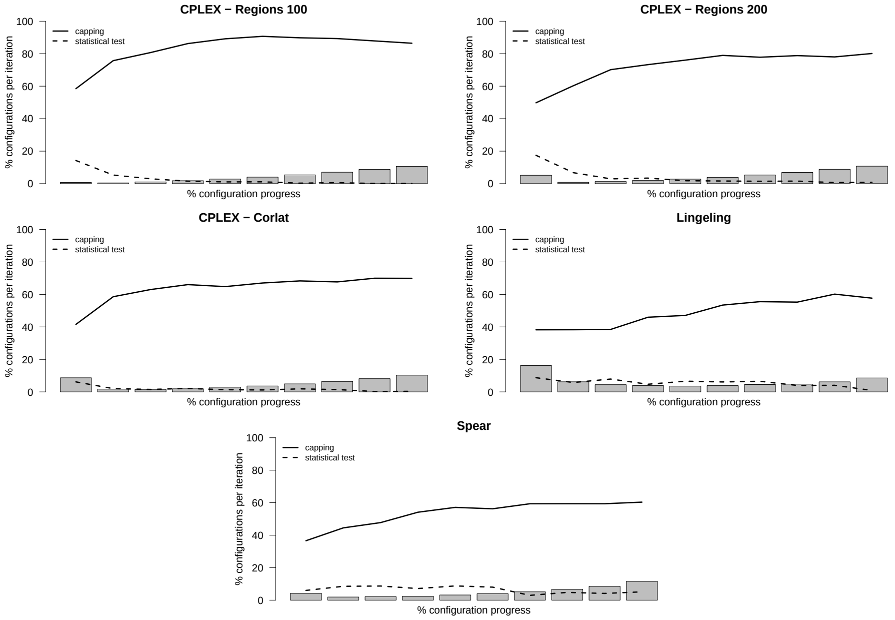
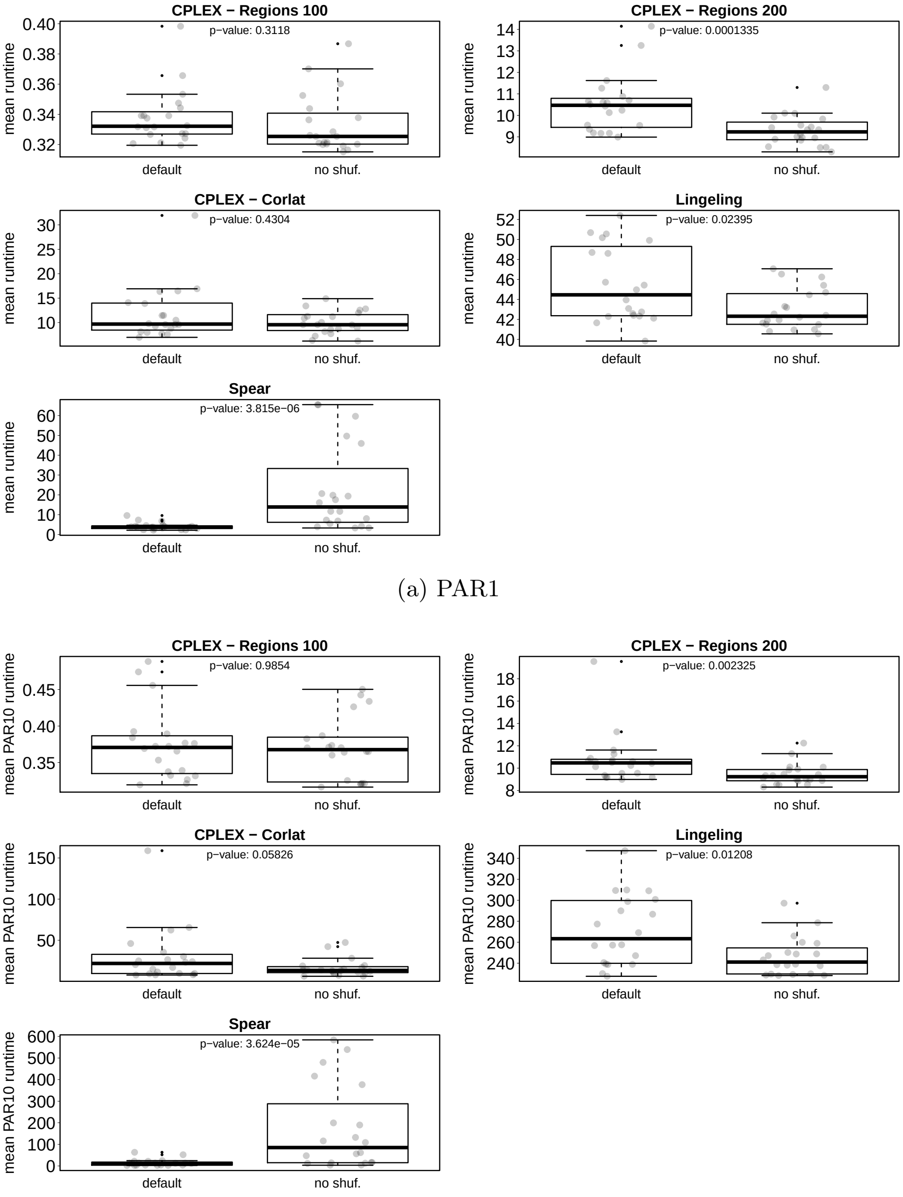


Fig. A.2: Mean percentage of configurations selected for elimination by the capping procedure and the statistical test (solid and dashed lines respectively), and mean percentage of initial configurations that become elite configurations at the end of the iteration (bars). Means obtained across 20 executions of $\text{irace}_{\text{cap}}$.



(a) PAR1

(b) PAR10

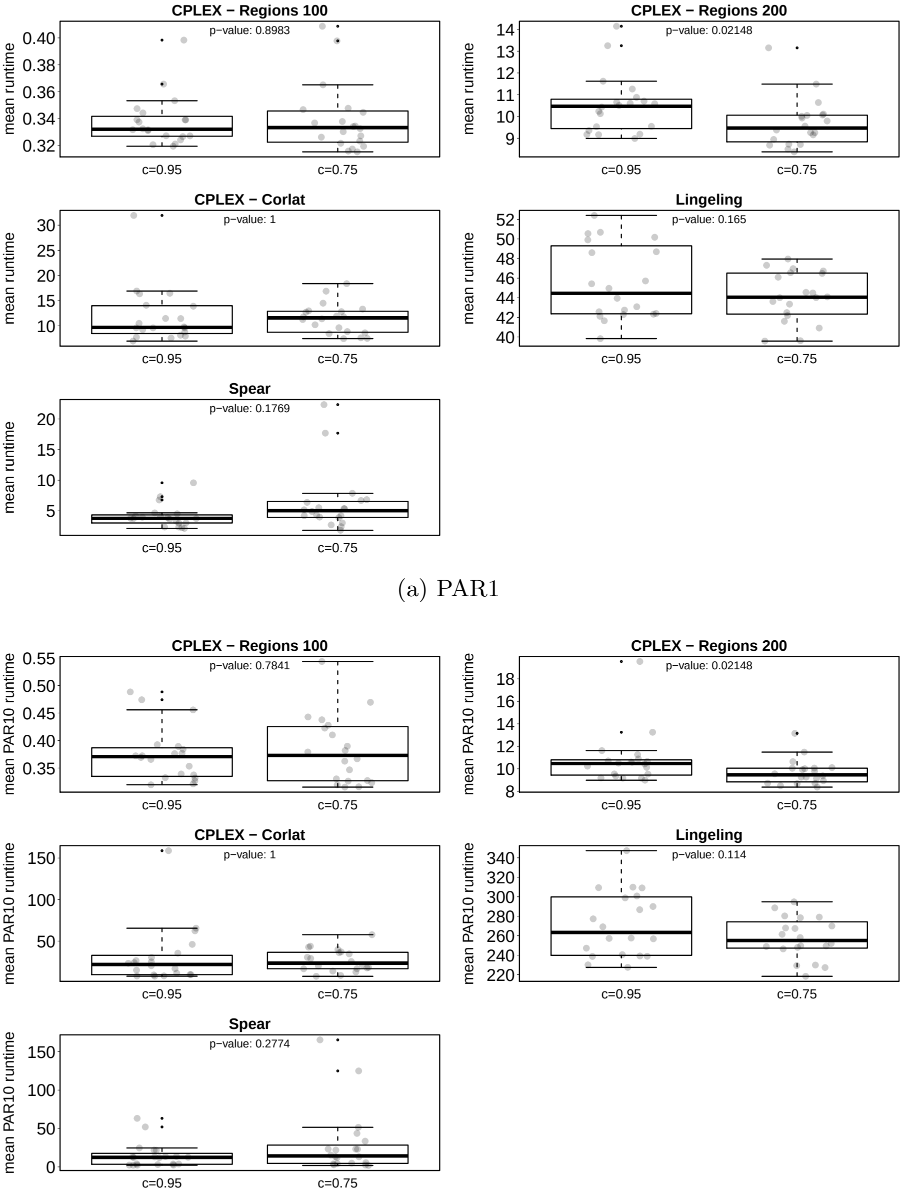
Fig. A.3: Mean PAR1 and PAR10 performance of 20 executions of irace$_{\mathsf{cap}}$ with initial instance shuffling enabled (*default*) and disabled (*no shuf.*) across the test set. Wilcoxon test (significance 0.05) p-values on each plot.

	Regions 100		Regions 200		Corlat		Lingeling		Spear	
	$c = 0.95$	$c = 0.75$	$c = 0.95$	$c = 0.75$	$c = 0.95$	$c = 0.75$	$c = 0.95$	$c = 0.75$	$c = 0.95$	$c = 0.75$
q25	0.327	0.323	9.487	8.889	8.616	8.795	42.379	42.411	3.028	3.959
mean	*0.338*	0.339	10.498	**9.689**	11.899	*11.402*	45.501	*44.13*	*4.116*	6.233
median	0.332	0.333	10.469	9.468	9.688	11.576	44.453	44.051	3.765	5.03
q75	0.34	0.345	10.75	10.054	13.941	12.859	48.996	46.503	4.242	6.459
sd	0.018	0.025	1.335	1.127	5.645	3.004	3.799	2.517	1.848	5.016
sd/mean	0.054	0.074	0.127	0.116	0.474	0.264	0.083	0.057	0.449	0.805
p-value	0.8983173		0.0214844		1		0.164957		0.1768532	
PAR10										
	$c = 0.95$	$c = 0.75$	$c = 0.95$	$c = 0.75$	$c = 0.95$	$c = 0.75$	$c = 0.95$	$c = 0.75$	$c = 0.95$	$c = 0.75$
q25	0.336	0.327	9.487	8.889	9.652	16.749	240.209	247.634	3.565	4.722
mean	*0.374*	0.382	10.768	**9.689**	30.663	*26.52*	271.599	*257.269*	*14.844*	29.476
median	0.371	0.373	10.469	9.468	21.836	23.426	263.347	255.065	12.421	14.279
q75	0.385	0.424	10.75	10.054	31.585	36.3	299.313	272.092	15.621	25.812
sd	0.049	0.062	2.305	1.127	34.684	13.44	33.462	21.287	16.352	42.419
sd/mean	0.13	0.161	0.214	0.116	1.131	0.507	0.123	0.083	1.102	1.439
p-value	0.7841263		0.0214844		1		0.113987		0.2773552	
PAR100										
	$c = 0.95$	$c = 0.75$	$c = 0.95$	$c = 0.75$	$c = 0.95$	$c = 0.75$	$c = 0.95$	$c = 0.75$	$c = 0.95$	$c = 0.75$
q25	0.336	0.327	9.487	8.889	29.902	70.749	2207.096	2303.926	3.565	4.722
mean	*0.734*	0.809	13.468	**9.689**	218.313	*177.72*	2533.52	*2389.554*	*122.128*	261.927
median	0.821	0.823	10.469	9.468	143.336	143.364	2453.744	2356.058	101.825	103.683
q75	0.835	1	10.75	10.054	224.475	279.3	2802.624	2529.542	127.376	226.971
sd	0.421	0.507	14.178	1.127	326.961	126.265	332.22	214.239	163.096	417.699
sd/mean	0.574	0.627	1.053	0.116	1.498	0.71	0.131	0.09	1.335	1.595
p-value	0.7561665		0.0214844		0.9854355		0.113987		0.2773552	
%timeout	0.08	0.095	0.01	0	0.695	0.56	8.377	7.897	0.397	0.861

Table A.3: Statistics of the mean PAR10 and PAR100 execution time performance and percentage of timed out instances of 20 executions of irace$_{\text{cap}}$ using the statistical test with confidence $\{0.95, 0.75\}$. Wilcoxon test p-values (significance 0.05). Significantly better results in bold and best mean in cursive.

	Regions 100		Regions 200		Corlat		Lingeling		Spear	
confidence	0.95	0.75	0.95	0.75	0.95	0.75	0.95	0.75	0.95	0.75
iterations	253.5	260.3	85.8	100.0	68.7	60.6	27.4	25.7	67.0	64.9
instances	258.6	264.3	91.1	104.0	75.1	65.3	35.5	31.5	83.2	74.3
candidates	27914	32239	5191	6200	5318	6348	2595	2754	11193	14155
elites	1.10	1.03	1.23	1.15	1.82	1.41	3.28	2.24	2.26	1.92
executions	30604	34251	6779	7715	8873	9798	5219	5502	28039	27923

Table A.4: Statistics of 20 executions of irace$_{\text{cap}}$ with statistical test confidence set to $\{0.95, 0.75\}$. Mean number of iterations performed (iterations), average number of instances used in the evaluation (instances), mean overall configurations sampled (candidates), mean elite configurations per iteration (elites) and mean total executions (executions).



(a) PAR1

(b) PAR10

Fig. A.4: Mean PAR1 and PAR10 performance of 20 executions across the test set of $\mathsf{irace}_{\mathsf{cap}}$ setting the statistical test confidence as $\{0.95, 0.75\}$. Wilcoxon test (significance 0.05) p-values on each plot.

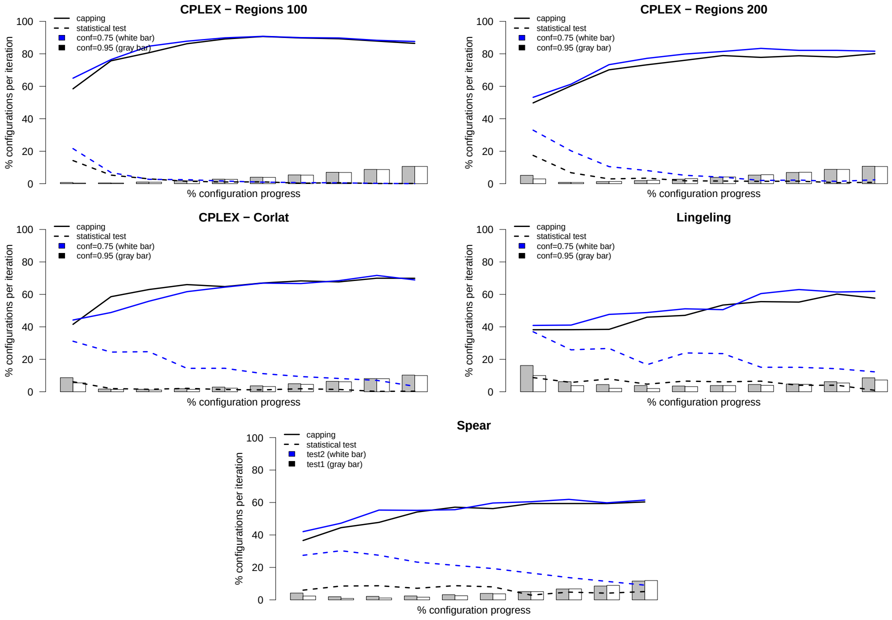
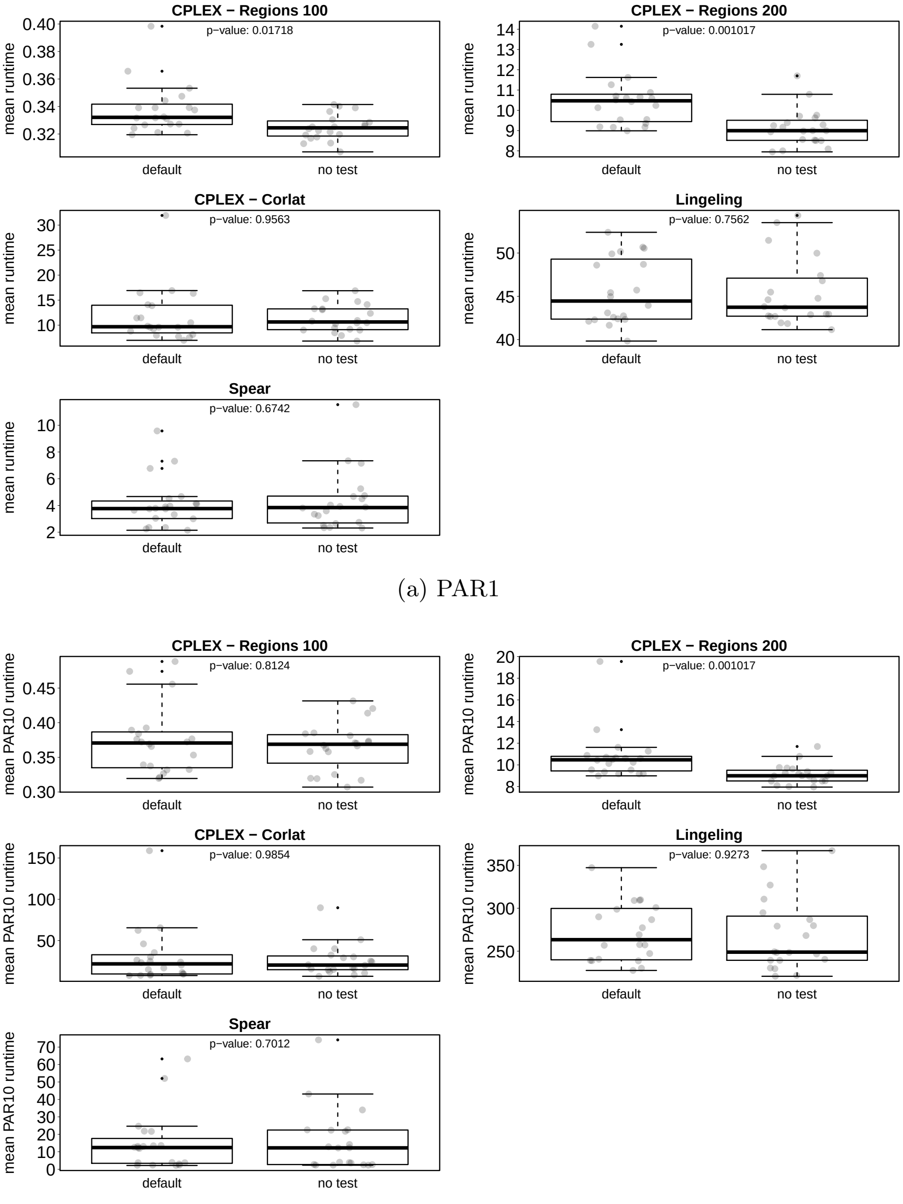


Fig. A.5: Mean percentage of configurations selected for elimination by the capping procedure and the statistical test (full and dashed lines respectively), and mean percentage of iterations in which the statistical test selected configurations that were not selected by the capping procedure (bars). Means obtained across 20 executions of $\text{irace}_{\text{cap}}$ setting the confidence of the statistical test to $\{0.75, 0.95\}$.

	Regions 100		Regions 200		Corlat		Lingeling		Spear	
	default	*no test*	*default*	*no test*	*default*	*no test*	*default*	*no test*	*default*	*no test*
q25	0.327	0.319	9.487	8.522	8.616	9.143	42.379	42.717	3.028	2.715
mean	0.338	**0.325**	10.498	**9.136**	11.899	*11.291*	45.501	*45.377*	*4.116*	4.295
median	0.332	0.324	10.469	8.996	9.688	10.641	44.453	43.734	3.765	3.841
q75	0.34	0.329	10.75	9.449	13.941	13.249	48.996	46.945	4.242	4.686
sd	0.018	0.009	1.335	0.91	5.645	2.706	3.799	3.972	1.848	2.227
sd/mean	0.054	0.029	0.127	0.1	0.474	0.24	0.083	0.088	0.449	0.518
p-value	0.0171814		0.0010166		0.9563293		0.7561665		0.6742229	
PAR10										
	default	*no test*	*default*	*no test*	*default*	*no test*	*default*	*no test*	*default*	*no test*
q25	0.336	0.35	9.487	8.522	9.652	14.947	240.209	239.32	3.565	2.715
mean	0.374	*0.365*	10.768	**9.136**	30.663	*26.004*	271.599	*268.795*	*14.844*	15.917
median	0.371	0.369	10.469	8.996	21.836	20.404	263.347	248.848	12.421	12.232
q75	0.385	0.382	10.75	9.449	31.585	30.87	299.313	288.788	15.621	22.418
sd	0.049	0.034	2.305	0.91	34.684	19.027	33.462	42.418	16.352	18.034
sd/mean	0.13	0.094	0.214	0.1	1.131	0.732	0.123	0.158	1.102	1.133
p-value	0.812355		0.0010166		0.9854355		0.9272785		0.7011814	
PAR100										
	default	*no test*	*default*	*no test*	*default*	*no test*	*default*	*no test*	*default*	*no test*
q25	0.336	0.687	9.487	8.522	29.902	61.954	2207.096	2206.207	3.565	2.715
mean	*0.734*	0.77	13.468	**9.136**	218.313	*173.154*	2533.52	*2503.894*	*122.128*	132.142
median	0.821	0.819	10.469	8.996	143.336	86.393	2453.744	2305.139	101.825	101.636
q75	0.835	0.832	10.75	9.449	224.475	206.37	2802.624	2725.047	127.376	201.226
sd	0.421	0.322	14.178	0.91	326.961	192.013	332.22	429.392	163.096	177.16
sd/mean	0.574	0.418	1.053	0.1	1.498	1.109	0.131	0.171	1.335	1.341
p-value	0.9563293		0.0010166		0.9563293		0.9272785		0.7011814	
%timeout	0.08	0.09	0.01	0	0.695	0.545	8.377	8.278	0.397	0.43

Table A.5: Statistics of the mean PAR10 and PAR100 execution time and percentage timed out instances of 20 executions of $\mathsf{irace}_{\mathsf{cap}}$ with default settings (*default*) and a version with the statistical disabled (*no test*). Wilcoxon test p-values (significance 0.05). Significantly better results in bold and best mean in cursive.



(a) PAR1

(b) PAR10

Fig. A.6: Mean PAR1 and PAR10 performance of 20 executions across the test set of irace$_{\text{cap}}$ with the statistical test enabled (*default*) and disabled (*no test*). Wilcoxon test (significance 0.05) p-values on each plot.

	Regions 100		Regions 200		Corlat		Lingeling		Spear	
	default	*log*	*default*	*log*	*default*	*log*	*default*	*log*	*default*	*log*
q25	0.327	0.327	9.487	8.781	8.616	8.9	42.379	41.779	3.028	3.67
mean	*0.338*	0.332	10.498	**9.535**	*11.899*	12.072	45.501	*44.1*	**4.116**	6.478
median	0.332	0.333	10.469	9.258	9.688	11.653	44.453	43.492	3.765	4.845
q75	0.34	0.337	10.75	10.109	13.941	14.857	48.996	44.481	4.242	5.971
sd	0.018	0.009	1.335	0.973	5.645	3.664	3.799	3.218	1.848	5.501
sd/mean	0.054	0.027	0.127	0.102	0.474	0.303	0.083	0.073	0.449	0.849
p-value	0.2773552		0.0153122		0.6476555		0.3682766		0.0266418	
PAR10										
	default	*log*	*default*	*log*	*default*	*log*	*default*	*log*	*default*	*log*
q25	0.336	0.356	9.487	8.781	9.652	14.108	240.209	239.368	3.565	4.388
mean	*0.374*	0.395	10.768	**9.67**	*30.663*	32.724	271.599	*258.578*	*14.844*	31.51
median	0.371	0.384	10.469	9.445	21.836	28.466	263.347	248.477	12.421	13.768
q75	0.385	0.43	10.75	10.474	31.585	39.75	299.313	266.573	15.621	26.19
sd	0.049	0.055	2.305	1.055	34.684	25.782	33.462	31.786	16.352	51.954
sd/mean	0.13	0.139	0.214	0.109	1.131	0.788	0.123	0.123	1.102	1.649
p-value	0.3299828		0.0171814		0.6215134		0.2454872		0.113987	
PAR100										
	default	*log*	*default*	*log*	*default*	*log*	*default*	*log*	*default*	*log*
q25	0.336	0.693	9.487	8.781	29.902	68.108	2207.096	2206.256	3.565	4.388
mean	*0.734*	1.025	13.468	*11.02*	*218.313*	239.274	2533.52	*2404.274*	*122.128*	281.841
median	0.821	0.834	10.469	9.445	143.336	163.466	2453.744	2304.768	101.825	103.172
q75	0.835	1.33	10.75	10.474	224.475	321.158	2802.624	2501.672	127.376	227.349
sd	0.421	0.568	14.178	6.548	326.961	254.166	332.22	321.811	163.096	517.495
sd/mean	0.574	0.554	1.053	0.594	1.498	1.062	0.131	0.134	1.335	1.836
p-value	0.2773552		0.0973072		0.7011814		0.2454872		0.113987	
%timeout	0.08	0.14	0.01	0.005	0.695	0.765	8.377	7.947	0.397	0.927

Table A.6: Statistics of the mean PAR10 and PAR100 execution time and percentage timed out instances of 20 executions of $\text{irace}_{\text{cap}}$ with default settings (*default*) and a version using log-tranformed evaluations for the statistical test (*log*). Wilcoxon test p-values (significance 0.05). Significantly better results in bold and best mean in cursive.

Table A.6 and Figure A.7 compare the results of $\text{irace}_{\text{cap}}$ (*default*) and a modified version in which the log transformed evaluations is used for the statistical test (*log*).

Table A.9 and Figure A.8 compare the results of $\text{irace}_{\text{cap}}$ setting the initial new instances evaluated in each race as $t^{new} = \{0, 1, 5\}$.

A.1 Evaluations for configuration in irace

Table A.7 compared mean performance of the $\text{irace}_{\text{cap}}$ executions that use the PAR10 evaluation in the configuration process (training) and evaluates them on the test set using PAR1 and PAR10.

A.2 PARX evaluation of the tuners

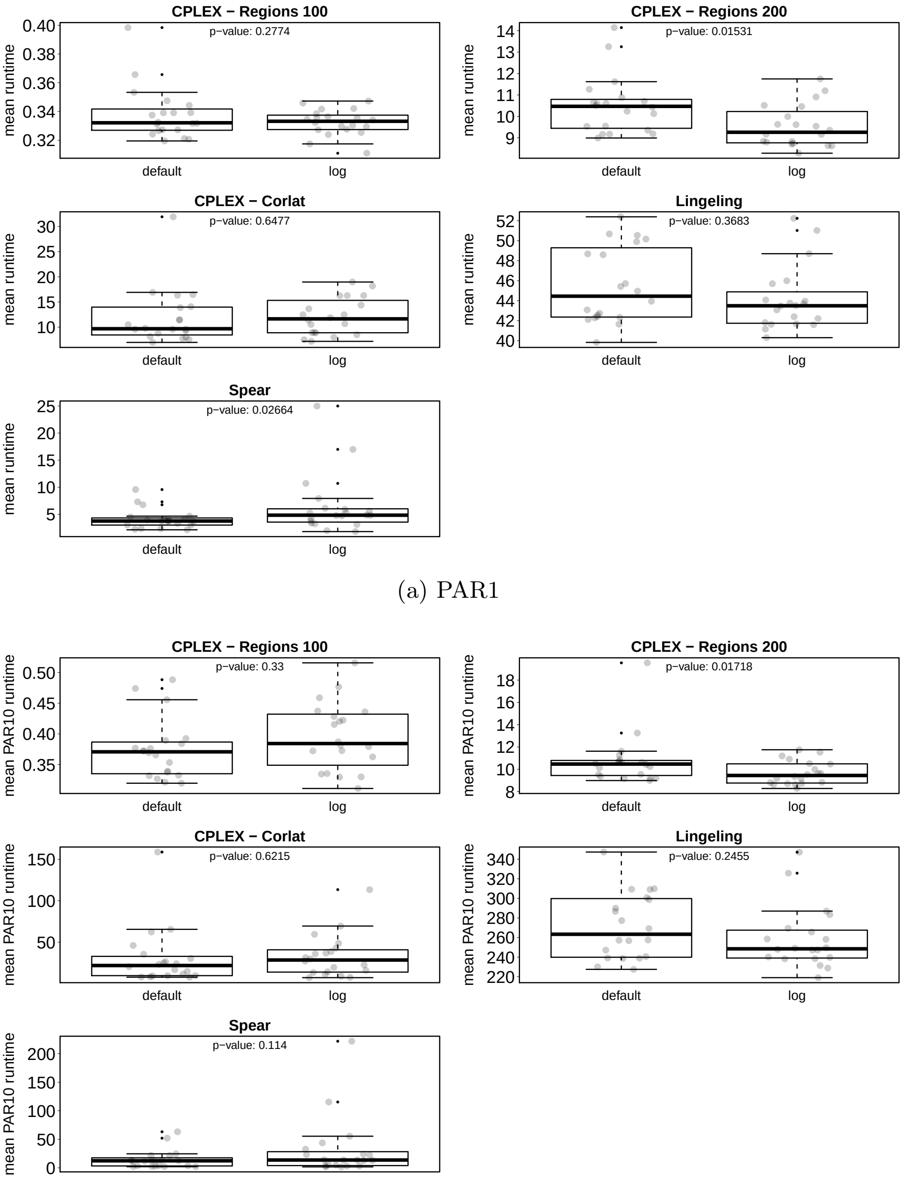
In Section 6, we evaluated the peformance of $\text{irace}_{\text{cap}}$, SMAC and ParamILS using PAR10 as evaluation during the configuration process. Table A.8 give the performance of the configurators using PAR10 in the configuation process and PAR1 evaluation and Figure A.9 compares the performance of the configurators using PAR10 in the configuration process and PAR10 in the evaluation.

	Regions 100		Regions 200		Corlat		Lingeling		Spear	
$\text{irace}_{\text{cap}}$	PAR1	PAR10	PAR1	PAR10	PAR1	PAR10	PAR1	PAR10	PAR1	PAR10
q25	0.327	0.317	9.487	8.854	8.616	8.447	42.379	42.246	3.028	3.695
mean	0.338	*0.329*	10.498	*9.791*	11.899	*11.237*	45.501	*44.705*	*4.116*	5.862
median	0.332	0.329	10.469	9.349	9.688	11.522	44.453	44.376	3.765	4.758
q75	0.34	0.338	10.75	10.177	13.941	12.783	48.996	45.975	4.242	6.787
sd	0.018	0.013	1.335	1.364	5.645	2.739	3.799	3.227	1.848	3.212
sd/mean	0.054	0.038	0.127	0.139	0.474	0.244	0.083	0.072	0.449	0.548
p-value	0.1536465		0.089695		0.8694878		0.4980087		0.0973072	
PAR10 (testing)										
$\text{irace}_{\text{cap}}$	PAR1	PAR10	PAR1	PAR10	PAR1	PAR10	PAR1	PAR10	PAR1	PAR10
q25	0.336	0.32	9.487	8.854	9.652	12.24	240.209	244.313	3.565	5.666
mean	0.374	*0.372*	10.768	*9.926*	30.663	*27.974*	271.599	*263.651*	*14.844*	23.741
median	0.371	0.365	10.469	9.349	21.836	26.763	263.347	259.768	12.421	22.3
q75	0.385	0.395	10.75	10.533	31.585	33.902	299.313	271.119	15.621	25.872
sd	0.049	0.057	2.305	1.459	34.684	19.426	33.462	31.736	16.352	21.512
sd/mean	0.13	0.154	0.214	0.147	1.131	0.694	0.123	0.12	1.102	0.906
p-value	0.9854355		0.2024498		0.8983173		0.3883762		0.1768532	
%timeout	0.08	0.095	0.01	0.005	0.695	0.62	8.377	8.113	0.397	0.662

Table A.7: Statistics of the mean PAR1 and PAR10 execution time and percentage timed out instances of 20 executions of $\text{irace}_{\text{cap}}$ using PAR1 and PAR10 evaluations. Wilcoxon test p-values (significance 0.05). Significantly better results in bold and best mean in cursive.

	q25	mean	median	q75	sd	sd/mean	%timeout
	Regions 100 p-value: 0.0531693						
paramILS	0.311	*0.322*	0.32	0.33	0.012	0.038	0.13
SMAC	0.418	0.458	0.458	0.487	0.049	0.107	0.045
$\text{irace}_{\text{cap}}$	0.317	0.329	0.329	0.338	0.013	0.038	0.095
	Regions 200 p-value: 0.0399895						
paramILS	9.409	11.521	9.883	13.606	3.382	0.294	0.005
SMAC	14.205	16.702	16.452	18.395	4.063	0.243	0.045
$\text{irace}_{\text{cap}}$	8.854	**9.791**	9.349	10.177	1.364	0.139	0.005
	Corlat p-value: 5.7e-06						
paramILS	13.611	32.788	16.889	22.587	38.84	1.185	5.945
SMAC	17.11	18.712	18.962	20.059	3.367	0.18	1.005
$\text{irace}_{\text{cap}}$	8.447	**11.237**	11.522	12.783	2.739	0.244	0.62
	Lingeling p-value: 0.0048599						
paramILS	44.611	48.903	48.611	51.636	5.638	0.115	9.023
SMAC	45.657	47.434	47.37	48.515	2.604	0.055	8.758
$\text{irace}_{\text{cap}}$	42.246	**44.705**	44.376	45.975	3.227	0.072	8.113
	Spear p-value: 9.5e-06						
paramILS	3.037	12.09	4.368	7.387	19.986	1.653	2.815
SMAC	1.6	**2.075**	1.746	2.371	0.741	0.357	0.05
$\text{irace}_{\text{cap}}$	3.695	5.862	4.758	6.787	3.212	0.548	0.662

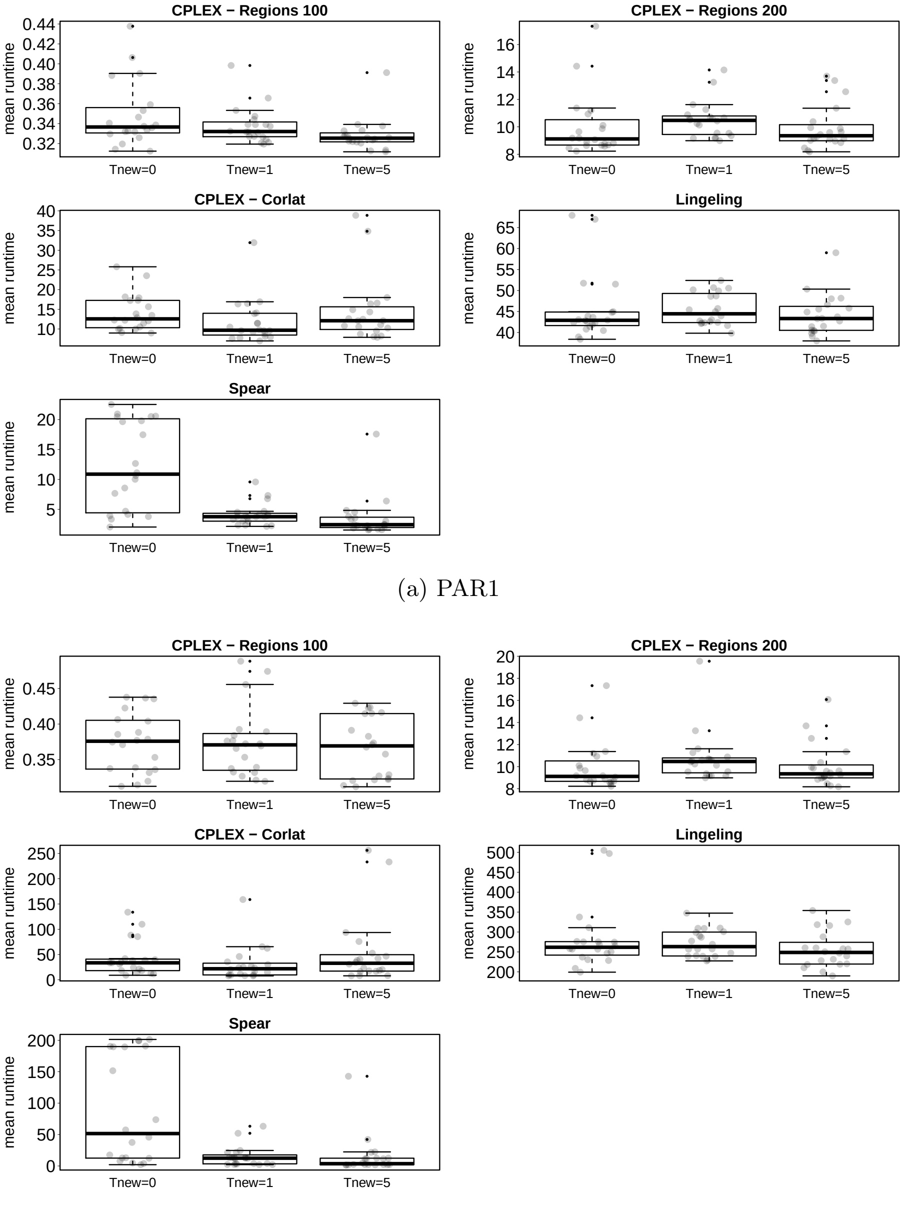
Table A.8: Statistics of the mean PAR1 performance and percentage of timed out instances of 20 executions of $\text{irace}_{\text{cap}}$, SMAC and ParamILS using default settings and PAR10 evaluation. Wilcoxon test p-values (significance 0.05). Significantly better results in bold and best mean in cursive.



(a) PAR1

(b) PAR10

Fig. A.7: Mean PAR1 and PAR10 performance of 20 executions across the test set of default irace$_{\mathsf{cap}}$ (*default*) and a version using log-tranformed evaluations for the statistical test (*log*). Wilcoxon test (significance 0.05) p-values on each plot.



(a) PAR1

(b) PAR10

Fig. A.8: Mean PAR1 and PAR10 performance across the test set of 20 configurations obtained by irace$_{\mathsf{cap}}$ using $T^{new} = \{0, 1, 5\}$. Wilcoxon test (significance 0.05) p-values on each plot.

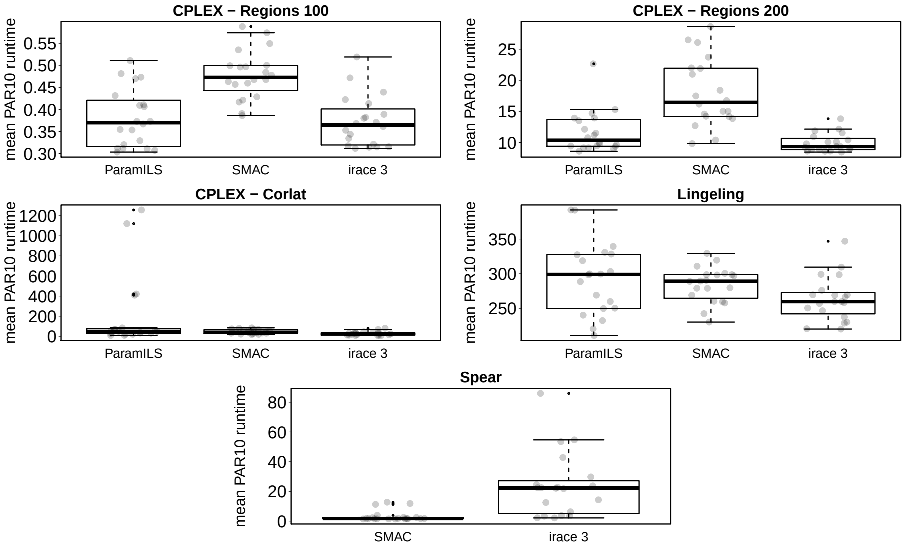


Fig. A.9: Mean PAR10 performance across the test set of 20 configurations obtained by paramILS, SMAC and $\mathsf{irace}_{\mathsf{cap}}$ using default settings and PAR10 evaluation.

	Regions 100			Regions 200			Corlat			Lingeling			Spear		
T^{new}	0	1	5	0	1	5	0	1	5	0	1	5	0	1	5
q25	0.331	0.327	0.322	8.679	9.487	9	10.426	8.616	10.053	41.773	42.379	40.536	4.55	3.028	2.004
mean	0.348	0.338	*0.328*	9.997	10.498	*9.91*	14.122	**11.899**	14.409	45.673	45.501	*44.132*	12.226	4.116	*3.57*
median	0.337	0.332	0.325	9.118	10.469	9.351	12.569	9.688	12.101	42.901	44.453	43.323	10.877	3.765	2.433
q75	0.355	0.34	0.33	10.315	10.75	10.048	17.25	13.941	15.265	44.838	48.996	46.016	19.967	4.242	3.612
sd	0.033	0.018	0.017	2.262	1.335	1.606	4.663	5.645	8.233	8.153	3.799	4.831	7.314	1.848	3.53
sd/mean	0.094	0.054	0.05	0.226	0.127	0.162	0.33	0.474	0.571	0.179	0.083	0.109	0.598	0.449	0.989
p-value	0.0531693		0.9272785		0.0484409		0.089695		0.0582581						
PAR10															
T^{new}	0	1	5	0	1	5	0	1	5	0	1	5	0	1	5
q25	0.337	0.336	0.323	8.679	9.487	9	18.189	9.652	17.419	244.471	240.209	219.856	12.616	3.565	2.004
mean	0.373	0.374	*0.367*	*9.997*	10.768	10.045	41.929	*30.663*	54.094	282.94	271.599	*254.589*	90.002	*14.844*	15.639
median	0.376	0.371	0.369	9.118	10.469	9.351	33.574	21.836	32.758	261.709	263.347	248.729	51.567	12.421	3.644
q75	0.405	0.385	0.415	10.315	10.75	10.048	40.379	31.585	48.214	275.619	299.313	267.197	189.824	15.621	12.338
sd	0.042	0.049	0.043	2.262	2.305	1.982	34.742	34.684	68.97	80.841	33.462	44.907	85.405	16.352	31.623
sd/mean	0.112	0.13	0.117	0.226	0.214	0.197	0.829	1.131	1.275	0.286	0.123	0.176	0.949	1.102	2.022
p-value	0.6742229		0.9272785		0.0695801		0.1230927		0.3488102						
PAR100															
T^{new}	0	1	5	0	1	5	0	1	5	0	1	5	0	1	5
q25	0.337	0.336	0.323	8.679	9.487	9	99.189	29.902	98.419	2278.411	2207.096	2007.935	102.02	3.565	2.004
mean	*0.621*	0.734	0.749	*9.997*	13.468	11.395	320.029	*218.313*	450.994	2656.615	2533.52	*2360.053*	867.817	*122.128*	136.334
median	0.422	0.821	0.813	9.118	10.469	9.351	221.859	143.336	237.085	2452.107	2453.744	2305.02	453.885	101.825	3.644
q75	0.83	0.835	1.315	10.315	10.75	10.048	317.129	224.475	385.714	2600.122	2802.624	2479.945	1888.499	127.376	101.742
sd	0.337	0.421	0.434	2.262	14.178	7.584	339.419	326.961	678.57	809.732	332.22	449.377	868.615	163.096	313.255
sd/mean	0.542	0.574	0.579	0.226	1.053	0.666	1.061	1.498	1.505	0.305	0.131	0.19	1.001	1.335	2.298
p-value	0.5458755		0.8983173		0.0758514		0.1230927		0.3488102						
%timeout	0.055	0.08	0.085	0	0.01	0.005	1.03	0.695	1.47	8.791	8.377	7.798	2.881	0.397	0.447

Table A.9: Statistics of the mean performance, mean PAR10 performance, and percentage timed out instances of 20 executions of $\mathsf{irace}_{\mathsf{cap}}$ with default settings ($t^{new} = 1$) and two versions setting the new instances executed at the begining of the race as $T^{new} = \{0, 5\}$. Wilcoxon test p-values (significance 0.05). Significantly better results in bold and best mean in cursive.

Given that the size of the penalty is arbitary and is dependent of the scenario, we compare the mean performance obtained by the configurators using different PARX penalties $\{1, \sqrt{10}, 10, 10 \cdot \sqrt{10}, 100, \}$. Figure A.10 compares the mean of the results using different PARX evaluations.

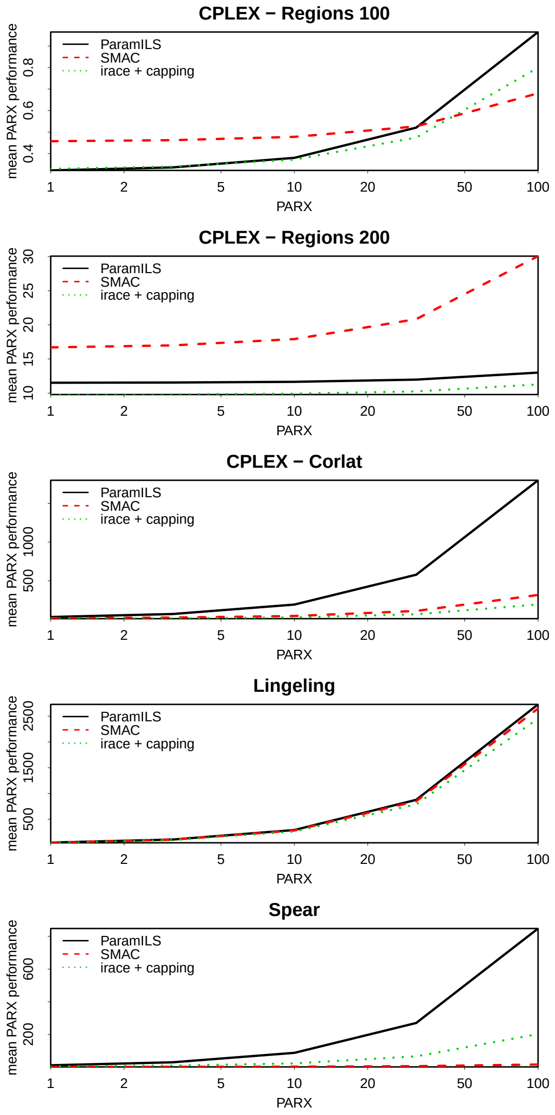


Fig. A.10: Mean PARX performance across the test set of 20 configurations obtained by paramILS, SMAC and $\text{irace}_{\text{cap}}$ using default settings and PAR10 evaluation.