Authors' response to the review of Manuscript 2095167

Manuscript 2095167, "Epistasis-based Basis Estimation Method for Simplifying the Problem Space of an Evolutionary Search in Binary Representation"

We appreciate the knowledgeable and helpful comments that we have taken carefully and did our best to comply with. And this is how we have responded to each of the referee's comments.

Authors' response to the comments from Referee 1

- The authors never provide a clear definition for what they deem as a difficult or easy problem/search space.

  Answer: We agree with the referee that we do not provide a definition for difficult or easy problem/search space. Reflecting the referee's comments, we added a sentence related defintion which difficult of problem/search space to the introduction section as follows:

  Epistasis has the advantage that it is possible to measure the extent of nonlinearity only with fitness function. In this paper, we define the difficulty of the problem or problem search space as the nonlinearity level of gene expression. Also, we use epistasis as a measure for the difficulty of the problem.

- It seems that difficulty is presumed to be captured by epistasis alone, but this is not the case. A much longer discussion should be included about difficulty in this context, the role that epistatic effects may have in determining problem difficulty and the limitations it brings as well, including for the particular measure used in this manuscript.

  Answer: Since we did not provide a clear definition of problem difficulty, we think that the referee commented that estimating the difficulty with epistasis alone may be a trouble. So we think that the referee advised us that we need more explanation of the difficulty. For that reason, the answer to the trouble will be resolved by providing a clear definition of the difficulty.

  We agree with the referee that the discussion on the role and limitation of epistasis is needed. We added a description of the role of epistasis to the beginning of the difficulty definition. We also clearly explained the limit in the end of Section 2.3 as follows:

  Note that nonlinearity may be misleading due to approximation error by solution sampling. It hinders to find the proper basis for the target problem. The target problem may be transformed into a more complex problem through a basis transformation. That is, the basis transformation can rather prevent a GA from efficiently finding the solution.

- Often the authors refer to "better", "significantly", "good", "very long", "size of the problem", etc but do not qualify their statement with a numerical assessment as well.

  Answer: That is a good point. We confirmed that we did not express the result numerically in Section 6.2, and we added two sentences to the fourth paragraph as follows:

  a) Note that when n is 50, it was over 2 hours.

  b) In particular, when n is 20, the number of optima found in 'Original' is 30, and the numbers of optima found in 'Epistasis-sq' and 'Epistasis-cu' are 64 and 33, respectively.

- S6.2 4th paragraph ("In general, good results...") is particularly very confusing and vague.

  Answer: We agree with the referee. We have made the sentence precise as follows:

  In Table 3, 'Meta' found opimal solutions more frequently than the other methods.

- The tables and figures should have more informative captions that allow the reader to better understand what is being presented.

  Answer: We agree with the referee. We tried to put enough explanation in each table and figure caption.

  Caption of Table 3:

  Results of each of the best solutions obtained by conducting the GA experiments 100 times on an instance of the variant-onemax problem. ('# of optima' is the number of optima found during 100 experiments, 'Average' is the average of 100 best solutions, and 'SD' is the standard deviation of 100 best solutions. Q_1, Q_2, and Q_3 are the first, second, and third quartiles, respectively. 'Time' is the sum of the time to search for the basis and that for the GA experiments.)

  Caption of Figure 3:

  A box plot of each of the best solutions obtained by conducting the GA experiment 100 times on an instance of the variant-onemax problem.

  Caption of Table 4:

  Results of each of the best solutions obtained by conducting the GA experiments 100 times on an instance of the NK-landscape problem. ('Best' is the best fitness among solutions found in 100 experiments, 'Average' is the average of 100 best solutions, and 'SD' is the standard deviation of 100 best solutions. Q_1, Q_2, and Q_3 are the first, second and third quartiles, respectively. 'Time' is the sum of the time to search for the basis and that for the GA experiments.)

  Caption of Figure 4:

  A box plot of each of the best solutions obtained by conducting the GA experiment 100 times on an instance of the NK-landscape problem.

- Many of the results focus on best found solution from the 100 trials. However, why is the median or distribution of final results not included? It would seem quite important to provide such data since the expected performance is more what should be ... expected... and gets more to the question of "are we likely to get a better outcome if we use a change of basis approach"?

  Answer: That is a good point. We think that there was a lack of distribution data for expected performance. We have changed Tables 3 and 4 with more statistical information as follows:

Table 3: Results of each of the best solutions obtained by conducting the GA experiments 100 times on an instance of the variant-onemax problem. ('# of optima' is the number of optima found during 100 experiments, 'Average' is the average of 100 best solutions, and 'SD' is the standard deviation of 100 best solutions. Q_1, Q_2, and Q_3 are the first, second, and third quartiles, respectively. 'Time' is the sum of the time to search for the basis and that for the GA experiments.)

n	Type	# of optima	Average	SD	Q_1	Q_2	Q_3	Time (mm:ss)*
20	Original	30	0.945	0.0452	0.900	0.950	1.000	0:44
	Meta	66	0.980	0.0302	0.950	1.000	1.000	3:07
	Epistasis-sq	64	0.982	0.0241	0.950	1.000	1.000	1:01
	Epistasis-cu	33	0.964	0.2760	0.950	0.950	1.000	3:11
30	Original	31	0.963	0.0329	0.930	0.970	1.000	1:09
	Meta	82	0.993	0.0155	1.000	1.000	1.000	12:15
	Epistasis-sq	47	0.979	0.0216	0.967	0.967	1.000	3:49
	Epistasis-cu	40	0.979	0.0187	0.967	0.967	1.000	7:03
50	Original	0	0.931	0.0257	0.920	0.940	0.940	2:58
	Meta	0	0.939	0.0240	0.920	0.940	0.960	136:46
	Epistasis-sq	2	0.934	0.0272	0.920	0.940	0.945	7:48
	Epistasis-cu	0	0.927	0.0272	0.900	0.920	0.940	67:59

* On Intel (R) Core TM i7-6850K CPU @ 3.60GHz

Table 4: Results of each of the best solutions obtained by conducting the GA experiments 100 times on an instance of the NK-landscape problem. ('Best' is the best fitness among solutions found in 100 experiments, 'Average' is the average of 100 best solutions, and 'SD' is the standard deviation of 100 best solutions. Q_1, Q_2, and Q_3 are the first, second, and third quartiles, respectively. 'Time' is the sum of the time to search for the basis and that for the GA experiments.)

N, K	Type	Best	Average	SD	Q_1	Q_2	Q_3	Time (mm:ss)*
20, 3	Original	0.817	0.8135	0.0085	0.8170	0.8170	0.8170	1:02
	Meta	0.825	0.8226	0.0057	0.8250	0.8250	0.8250	5:52
	Epistasis	0.825	0.8200	0.0056	0.8170	0.8170	0.8250	1:32
20, 5	Original	0.761	0.7449	0.0157	0.7400	0.7405	0.7610	1:03
	Meta	0.761	0.7533	0.0131	0.7470	0.7610	0.7610	5:39
	Epistasis	0.761	0.7505	0.0109	0.7460	0.7470	0.7610	1:40
20, 10	Original	0.779	0.7306	0.0253	0.7020	0.7335	0.7520	1:10
	Meta	0.785	0.7572	0.0155	0.7660	0.7550	0.7660	7:13
	Epistasis	0.785	0.7558	0.0136	0.7460	0.7530	0.7653	2:16
30, 3	Original	0.776	0.7687	0.1373	0.7740	0.7760	0.7760	2:06
	Meta	0.776	0.7719	0.0109	0.7760	0.7760	0.7760	5:39
	Epistasis	0.776	0.7718	0.0090	0.7740	0.7760	0.7760	1:40
30, 5	Original	0.795	0.7725	0.0125	0.7638	0.7740	0.7870	2:06
	Meta	0.795	0.7661	0.0170	0.7540	0.7710	0.7770	32:28
	Epistasis	0.795	0.7706	0.0136	0.7623	0.7730	0.7830	2:50
30, 10	Original	0.779	0.7349	0.0181	0.7260	0.7310	0.7443	2:06
	Meta	0.805	0.7391	0.0179	0.7310	0.7370	0.7470	49:47
	Epistasis	0.796	0.7366	0.0198	0.7220	0.7335	0.7960	3:48
30, 20	Original	0.750	0.7039	0.0152	0.6938	0.7010	0.7113	2:51
	Meta	0.762	0.7181	0.0163	0.7070	0.7155	0.7243	49:47
	Epistasis	0.770	0.7220	0.0133	0.7120	0.7200	0.7300	3:48
50, 3	Original	0.776	0.7576	0.0102	0.7515	0.7590	0.7640	5:31
	Meta	0.776	0.7599	0.0119	0.7530	0.7585	0.7730	220:14
	Epistasis	0.776	0.7578	0.0096	0.7508	0.7590	0.7630	6:34

* On Intel (R) Core TM i7-6850K CPU @ 3.60GHz

- Table 5 and 6 show data for increasing n and before/after. This gap seems to be closing as n increases, and the authors should comment on this, and perhaps explain/prove some convergence (if one exists).

  Answer: It does not seem to decrease the *gap* as n increases. For the helping to understand the *gap*, we added decrease rates to Tables 5 and 6 as follows:

Table 5: Epistasis of the original and modified basis sampling in the variant-onemax problem.

n	Sampling size	Epistasis		
		Before	After	Decrease rate (%)*
20	square	4.46	3.23	27.6
	cubic	4.35	3.83	12.0
30	square	4.57	3.20	30.0
	cubic	5.00	3.72	25.6
50	square	9.27	7.53	18.8
	cubic	9.69	8.93	7.8

* Decrease rate = 100 × (Before − After) /Before

Table 6: Epistasis of the original and modified basis sampling in the NK-landscape problem.

N, K	Epistasis		
	Before	After	Decrease rate (%)*
20, 3	$3.17e^{-3}$	$2.25e^{-3}$	29.0
20, 5	$3.16e^{-3}$	$2.90e^{-3}$	8.2
20, 10	$4.28e^{-3}$	$3.82e^{-3}$	10.7
30, 3	$1.85e^{-3}$	$1.60e^{-3}$	13.5
30, 5	$2.61e^{-3}$	$2.37e^{-3}$	9.2
30, 10	$2.68e^{-3}$	$2.39e^{-3}$	10.8
30, 20	$2.78e^{-3}$	$2.50e^{-3}$	10.1
50, 3	$1.13e^{-3}$	$9.32e^{-4}$	17.5

* Decrease rate = 100 × (Before − After) /Before

Authors' response to the comments from Referee 3

- This paper proposes a technique for changing the basis of the underlying problem representation, thereby reducing the epistasis. In essence, this allows a more challenging problem to be transformed into an easier problem, thereby allowing a more effective search, by way of genetic algorithm, to be carried out. Overall, the paper is well written and easy to follow. In my opinion, this is an excellently written paper that is clear in its methodology and the results are sound.

  Answer: Thanks a lot for the good summaray of the paper. We did our best to improve the paper based on the reviewer's constructive comments.

- Perhaps it would be beneficial to state that the proposed process aims to transform a non-separable problem to a separable problem. This terminology may target a wider audience and gives a much clearer indication of the overall effect that chaning the basis can have.

  Answer: Thanks for the constructive comments. Reflecting the referee's comments, we added a sentence of the statement to the introduction section as follows:

  Our intention in this study is that a non-separable problem can be transformed into a separable problem by performing an appropriate basis transformation. Such an altered environment enables GA to search space effectively.

- Page 3 - it is good to include a header paragraph before the first subsection. Thus, you should include a paragraph at the beginning of Section 3 that describes what the overall purpose of Section 3 is.

  Answer: Thanks for the good comment. Reflecting the referee's comments, we added a paragraph to the beginning of Section 3 as follows:

  This section presents a GA that performs an effective search through a change of basis. Before presenting the GA, we introduce the related terminologies and theories of change of basis in binary representation. Next, we apply the change of basis in the onemax problem to show how the problem actually transformed. In addition, a methodology for evaluating solutions in the transformed problem will be described. Finally, we propose a GA that effectively searches solutions through applying the change of basis. On the other hand, searching for an appropriate basis will be covered in Sections 4 and 5.

- Algorithm 1 - if P' and O' are used to generate the new population, should there not be a process in this algorithm that translates $P' \rightarrow P$? Otherwise, how is fitness calculated using the new representation? Perhaps Step 6 should read 'the process from Step 2 onward' rather than 'Step 3 onward'? Also, would it not make sense to return the population in the original basis, rather than the new basis? Otherwise, the GA is returning a solution with a different encoding than it was initially set to optimize. I find that Algorithm 1 should be edited for clarity as it provides the foundation for the remainder of the work and any uncertainty about the process stemming from here would likely detract from the overall impact of the proposed technique.

  Answer: As the referee said, the process from $P' \rightarrow P$ must be at the end. Reflecting the referee's comments, we changed Step 6 of Algorithm 1 as follows:

Step 6: *the process from Step 3 onward is repeated as many times as there are generations. When the number of generations has been exceeded, then we return P' whereby the basis B is changed to the standard basis B_s.*

Thank you again.