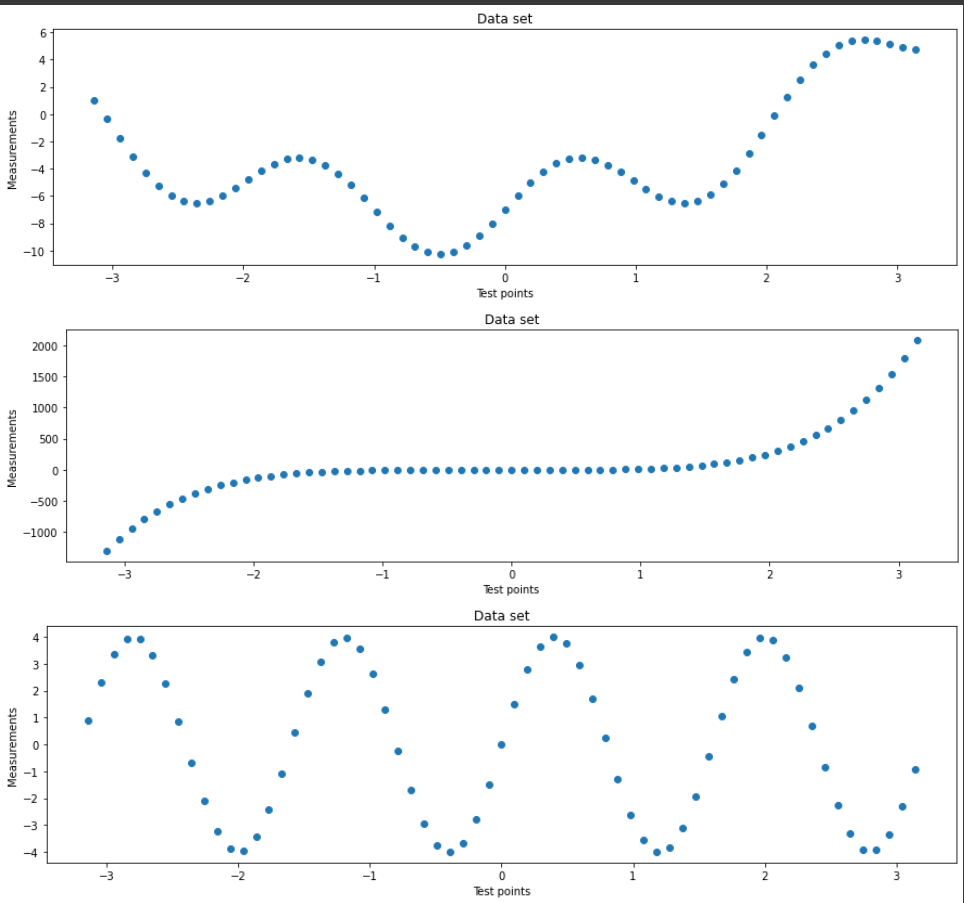
## Genetic Programming

The two problems function I have selected to create symbolic regression data sets are:

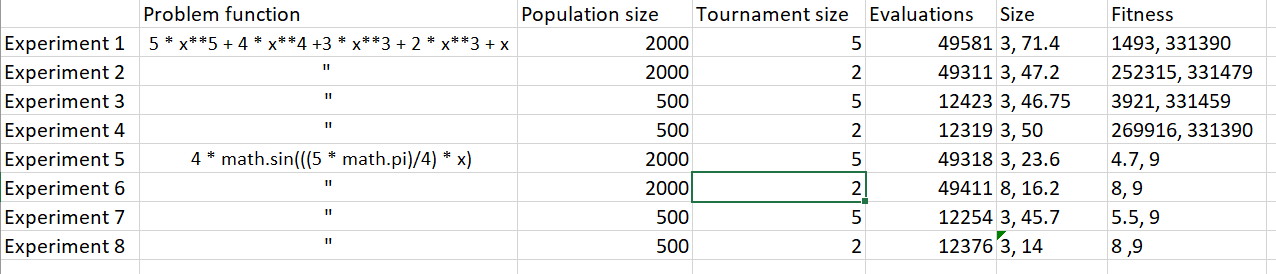
𝑝1 (𝑥) = 5 ∙ 𝑥 5 + 4 ∙ 𝑥 4 + 3 ∙ 𝑥 3 + 2 ∙ 𝑥 2 + 𝑥

𝑝2 (𝑥) = 4 ∙ sin ((5 ∙ 𝜋 /4) ∙ 𝑥)

The graph diagram below displays the graph created using the problem functions. The first diagram is from the function p1, and the second diagram is from the function p2.



The table below displays the result obtained from 4 different experiments. The experiments have 5 different column, tournament size, population size, evaluations, size, and fitness respectively. The GP experiments were executed 10 times each, the results from evaluation, size fitness are obtained by calculating the average between the 10 experiments.

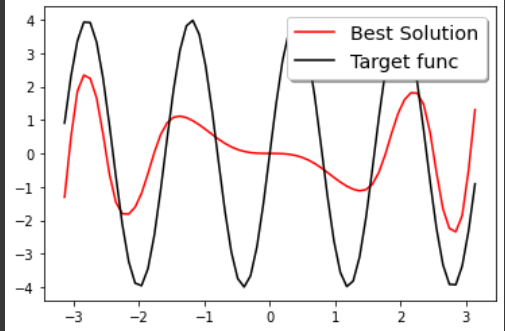
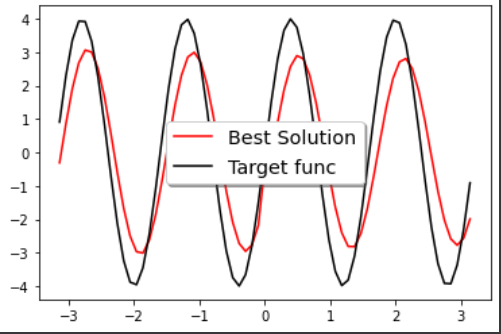


After obtaining the results we observed that the GP program changed in behaviour as the problem, population size and tournament size changed. As you can see on the graph above, as the tournament size increased the total number of evaluations increased, similarly as the tournament size decreased the total number of evaluations decreased. Furthermore, the change in population size changed the fitness score and the size.

In terms of problem function, for the second problem function, the tournament size had the most impact, where regardless of the population size the function still produced better accuracy then a tournament size of 2, similarly for the first problem the accuracy of the GP increased as the tournament size increased to 5.

Although we had 10 tests for each experiment, we cannot fully give a fixed results due to the variations of results the GP outputs.

The diagram below displays a line graph which were generated using the second problem function with a tournament size of 2 and the population size of 2000. As you can see the diagram on the right is far more accurate than the diagram on the right

The variations of results are likely due to the combination of randomness, the large search space size, random mutations, and the size of initial population having a possibility to converge the fitness peak.

I have used the fitness score and speed to identify the best parameter configuration for the two problems, smaller the fitness score, better the parameter configuration. However, speed is more important if the result is approximately the same. After choosing the optimal configuration for each problem I have tested it by running it 10 times. From that test, we found out that the problem function 𝑝2 (𝑥) = 4 ∙ sin ((5 ∙ 𝜋 /4) ∙ 𝑥) was not consistent with the statistics I obtained in the first batch of the run, however problem function 𝑝1 (𝑥) = 5 ∙ 𝑥 5 + 4 ∙ 𝑥 4 + 3 ∙ 𝑥 3 + 2 ∙ 𝑥 2 + 𝑥 Stayed consistent. The result from p1 is likely due to the outliers produced by mutations, with more mutations, the more different the output is going to be. We can solve this issue with more testing, for example instead of running it for 10 times, we run it for 30 times, the bigger the sample size, there will be less chance of mutations happening.

We can conclude from the information above that the GP is not sustainable to solve the problem 𝑝2 (𝑥) = 4 ∙ sin ((5 ∙ 𝜋 /4) ∙ 𝑥), due to the chance of GP outputting the best solution with great accuracy is relatively low. However, GP is sustainable to solve the problem 𝑝1 (𝑥) = 5 ∙ 𝑥 5 + 4 ∙ 𝑥 4 + 3 ∙ 𝑥 3 + 2 ∙ 𝑥 2 + 𝑥, due to the accuracy of the result being consistent.