Homework 1: Traveling Salesman Problem

Rebecca Roskill
rcr2144

MECS 4510
Prof. Hod Lipson

Submitted: Tuesday, October 4th, 2021
Grace hours used: 0
Grace hours remaining: 96

1 Methods

1.1 Datasets: TSP 1 and TSP 2

1.2 Random Search

1.3 Random Mutation Hill Climber

1.4 Genetic Algorithm

1.1 Datasets: TSP 1 and TSP 2

One dataset was provided for the assignment, and one dataset was produced for testing purposes:

TSP 1 The dataset provided for the assignment consisted of 1000 points (x, y) that fall within x=(0, 1) and y=(0, 1). The points are clustered in four square formations.

TSP 2 The dataset produced for testing consisted of 100 points (x, y) that fall within x=[-1, 1]. and y=[-1, 1]. The points are uniformly distributed along the path of a circle with radius 1, centered at the origin.

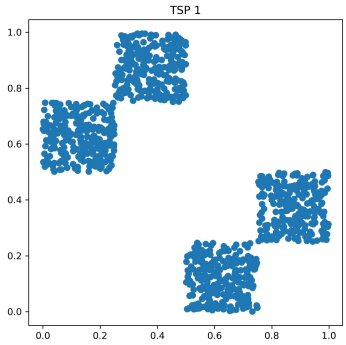

Figure 1.1a: TSP 1 dataset.

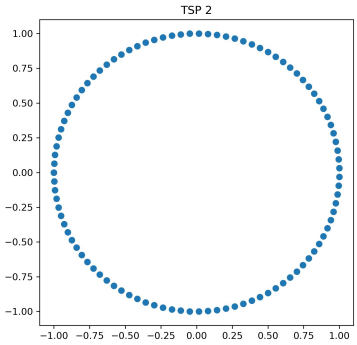

Figure 1.1b: TSP 2 dataset.

1.2 Random Search

Each trial run by *run_random_search* repeats the following process, maintaining the best path found (according to either longest or shortest path, depending on the experiment), as well as the current path as a measure of the learning curve, or "fitness":

1. The *uniform* function implemented in the Python *random* library was used to randomly initialize a priority weight [0, 1] for each point in the dataset.
2. The *assess_input* function was used to order the points by priority, simulating the process of visiting each location, and calculate the euclidean distance between each sequential pair of points.

1.3 Random Mutation Hill Climber

The *uniform* function implemented in the Python *random* library was used to randomly initialize a priority weight [0, 1] for each point in the dataset. Then, each trial run by *run_rmhc_search* repeats the following process, maintaining the best path found (according to either longest or shortest path, depending on the experiment), as well as the current path as a measure of the learning curve, or "fitness":

1. Using *get_random_neighbor,* one "neighbor" of the current ordering is found by swapping two points.
2. The *assess_input* function was used to order the neighbor's points by priority and calculate the euclidean distance between each sequential pair of points.
3. If the neighbor's path outperforms the current "best" path, the neighbor becomes the starting point for the next iteration. Otherwise, we keep the current ordering.

1.4 Genetic Algorithm

The *uniform* function implemented in the Python *random* library was used to randomly initialize a population of *n_processes* inputs, consisting of priority weights [0, 1] for each point in the dataset.

An optional step was performed in some versions of the algorithm (see *1.4.1*, *1.4.2*, *1.4.3*) to order the genes on the chromosome according to approximate regions. These regions were determined using the *KMeans* clustering implementation in the Scikit *sklearn* Python library. k=[1, 9] were tested sequentially until the improvement yielded by incrementing k was <10%.

Then, each trial run by *run_ga_search* repeats the following process, maintaining the best path found (according to either longest or shortest path, depending on the experiment), as well as the current path as a measure of the learning curve, or "fitness":

1. The *assess_input* function was used to order each input in the population's points by priority and calculate the euclidean distance between each sequential pair of points. The population was then sorted by this path distance, and the top-scoring p fraction was selected to reproduce.

…

1.4 Genetic Algorithm (cont.)

2. Using *reproduce,* 80% of the population undergoes mutation and 20% of the population undergoes recombination. This stage repeats to multiply the parents by $1/p$, thus maintaining the same population size for the offspring.

   Mutation – Five types of mutation were implemented:
   i. Random swap: The priority values of two random genes are swapped.
   ii. Random flip: The priority, x, of one random gene is flipped to $x–1$.
   iii. Intra-region swap: The priority values of two random genes within a single region are swapped.
   iv. Cross-region swap: The priority values of two random genes in different regions are swapped.
   v. Hybrid: 50% of the population undergoes "intra-region swap" mutation, while 50% undergoes "cross-region swap" mutation

   Recombination – Four types of recombination were implemented:
   i. Single-point: The first parent and the second parent undergo crossover at a single random point.
   ii. Intra-region: The child is the first parent, with one region replaced a single-point crossover between the first parent and the second parent in that region.
   iii. Cross regions: The child is the first parent, with one region replaced by the second parent's values in that region.
   iv. Hybrid: 50% of the population undergoes "intra-region" recombination, while 50% undergoes "cross regions" recombination.
3. The offspring become the future generation for the next iteration.

1.4.1 Genetic Algorithm: No linkage

- The chromosome is not ordered in a randomized fashion.
- Mutations are random swaps anywhere on the chromosome
- Recombinations are a single-point crossover at a random point, anywhere on the chromosome.

1.4.2 Genetic Algorithm: Reverse linkage → hybrid linkage → tight linkage

- The chromosome is ordered according to the region numbers yielded by k-means clustering. Genes within these regions are ordered in a randomized fashion.
- In the first stage, "reverse linkage," mutations are swaps between regions and crossovers are two-point: they return a copy of the first parent, with one entire region replaced by the second parent's.
- In the second stage, "hybrid linkage," 50% of the population undergoes reverse linkage reproduction, while 50% undergoes tight linkage reproduction.
- In the third and final stage, "tight linkage," mutations are swaps within regions and crossovers are are two-point: they return a copy of the first parent, with part of one region replaced by the second parent's.

1.4.3 Genetic Algorithm: Reverse linkage → hybrid linkage

- The chromosome is ordered as in *1.4.2*.
- In the first stage, "reverse linkage," mutations are swaps between regions and crossovers are two-point: they return a copy of the first parent, with one entire region replaced by the second parent's.
- In the second stage, "hybrid linkage," 50% of the population undergoes reverse linkage reproduction, while 50% undergoes tight linkage reproduction (described above in *1.4.2*).

2 Results

2.1 Performance across algorithms by dataset

2.2 Shortest and longest paths found

2.3 Search animations

2.1 Performance across algorithms by dataset

2.1.1 TSP 1

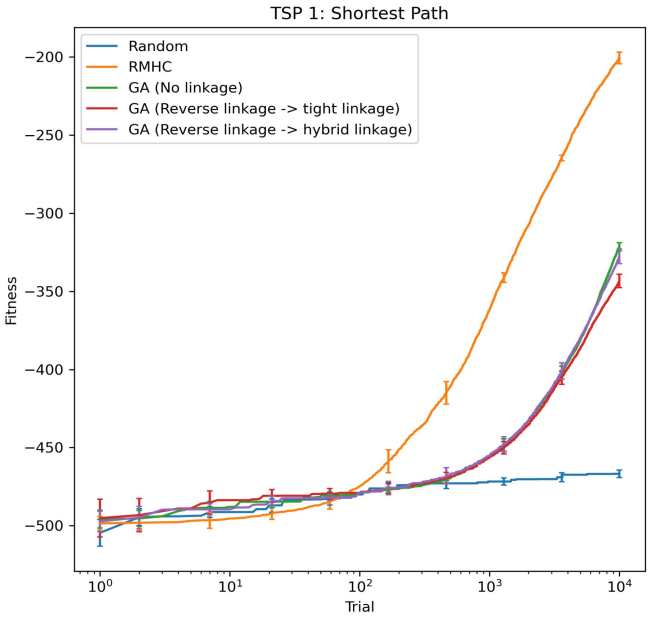

Figure 2.1.1a: Fitness performance across algorithms, searching for the shortest path in TSP 1.

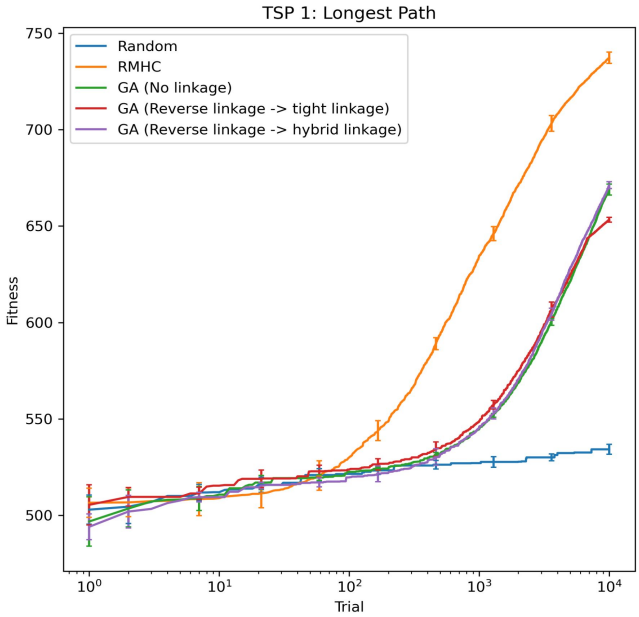

Figure 2.1.1b: Fitness performance across algorithms, searching for the longest path in TSP 1.

2.1.2 TSP 2

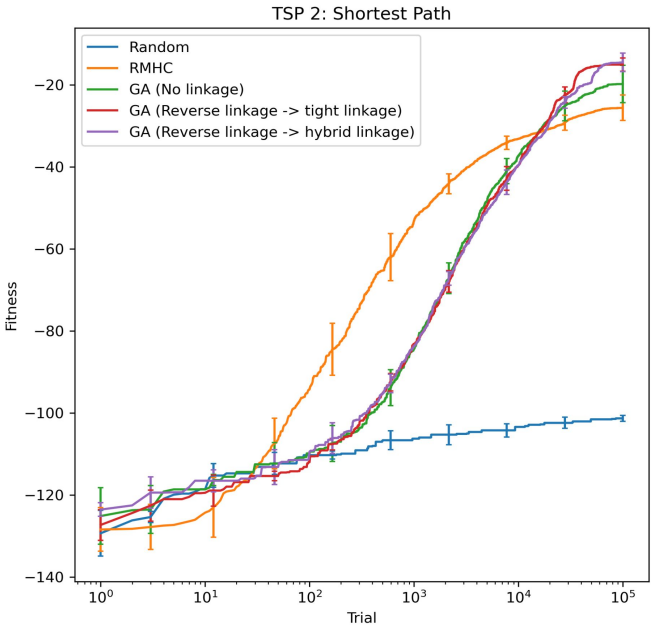

Figure 2.1.2a: Fitness performance across algorithms, searching for the shortest path in TSP 2.

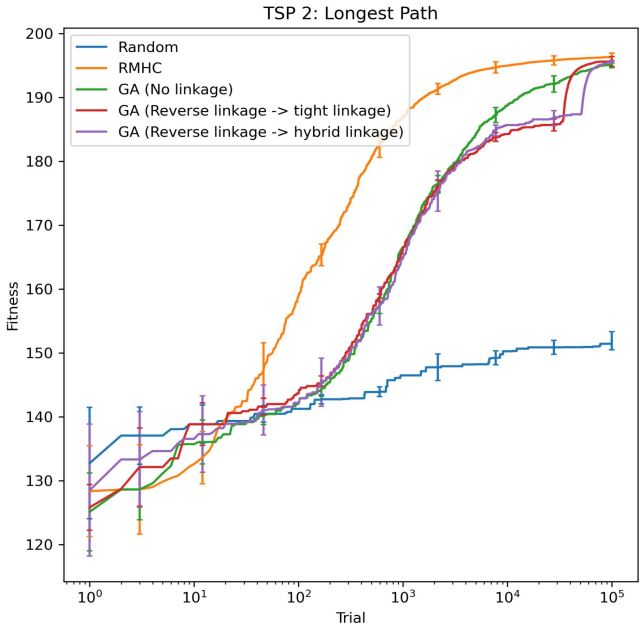

Figure 2.1.2b: Fitness performance across algorithms, searching for the longest path in TSP 2.

2.2 Shortest and longest paths found by algorithm

2.2.1 Random Search

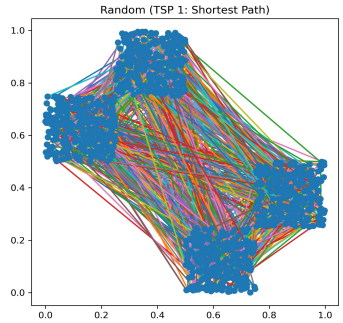

Figure 2.2.1a (left): Shortest path found by Random Search algorithm on TSP 1 dataset in 10^4 iterations; Path length = 465.26

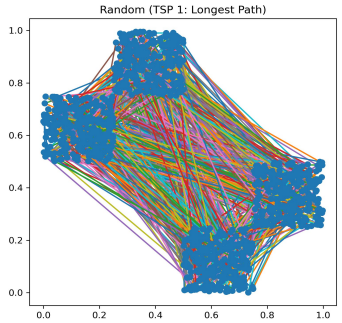

Figure 2.2.1b (left): Longest path found by Random Search algorithm on TSP 1 dataset in 10^4 iterations; Path length = 521.58

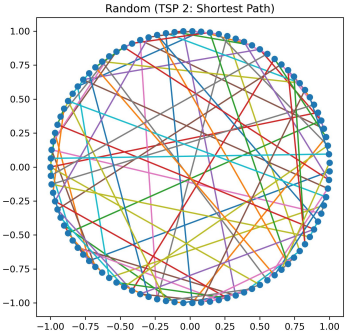

Figure 2.2.1c (left): Shortest path found by Random Search algorithm on TSP 2 dataset in 10^5 iterations; Path length = 99.17

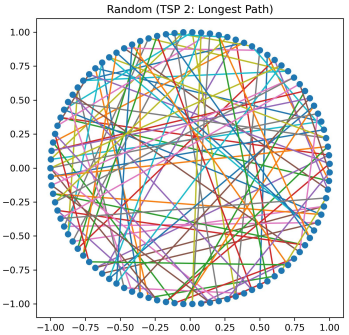

Figure 2.2.1d (left): Longest path found by Random Search algorithm on TSP 2 dataset in 10^5 iterations; Path length = 153.79

2.2.2 RMHC Search

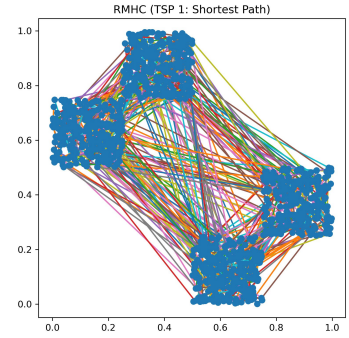

Figure 2.2.2a (left): Shortest path found by RMHC Search algorithm on TSP 1 dataset in 10^4 iterations; Path length = 194.92

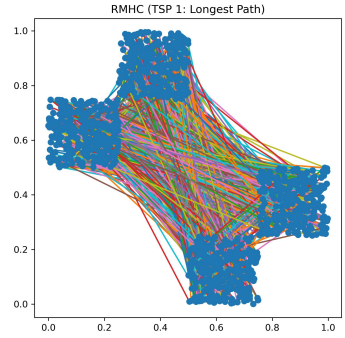

Figure 2.2.2b (left): Longest path found by RMHC Search algorithm on TSP 1 dataset in 10^4 iterations; Path length = 741.24

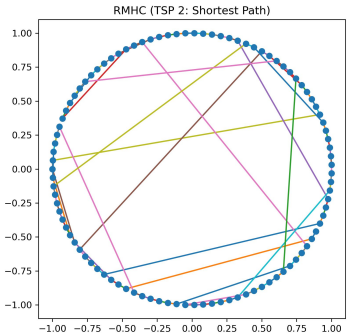

Figure 2.2.2c (left): Shortest path found by RMHC Search algorithm on TSP 2 dataset in 10^5 iterations; Path length = 26.96

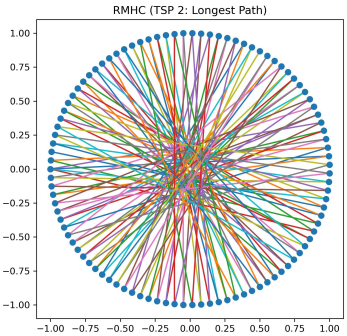

Figure 2.2.2d (left): Longest path found by RMHC Search algorithm on TSP 2 dataset in 10^5 iterations; Path length = 197.04

2.2.3 GA (No linkage) Search

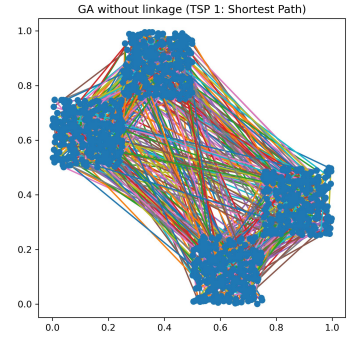

Figure 2.2.3a (left): Shortest path found by RMHC Search algorithm on TSP 1 dataset in 10^4 iterations; Path length = 322.94

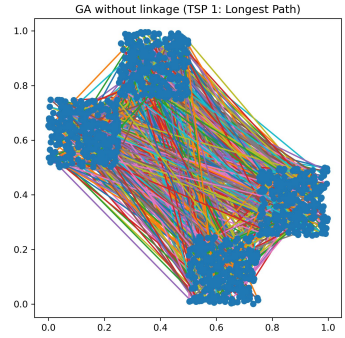

Figure 2.2.3b (left): Longest path found by GA (No linkage) Search algorithm on TSP 1 dataset in 10^4 iterations; Path length = 670.83

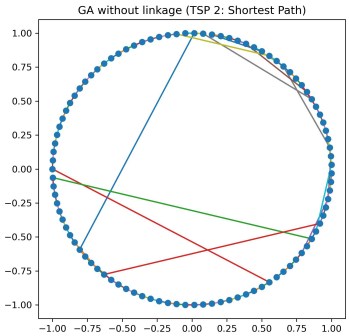

Figure 2.2.3c (left): Shortest path found by GA (No linkage) Search algorithm on TSP 2 dataset in 10^5 iterations; Path length = 18.14

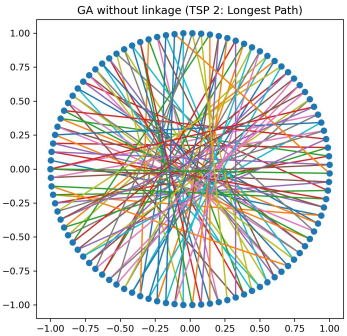

Figure 2.2.3d (left): Longest path found by GA (No linkage) Search algorithm on TSP 2 dataset in 10^5 iterations; Path length = 194.53

2.3 Search animations

RMHC

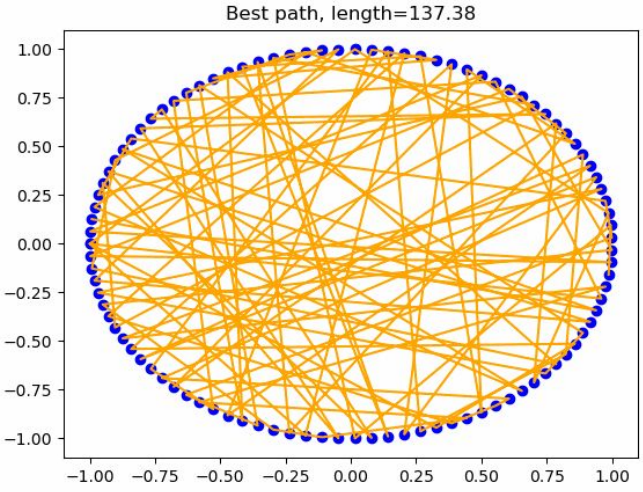

Figure 2.3a: RMHC on sparse circle dataset (10^5 iterations)

GA (No linkage)

Mutation=80% (swap); Recombination=20% (single-point)
No linkage, random ordering

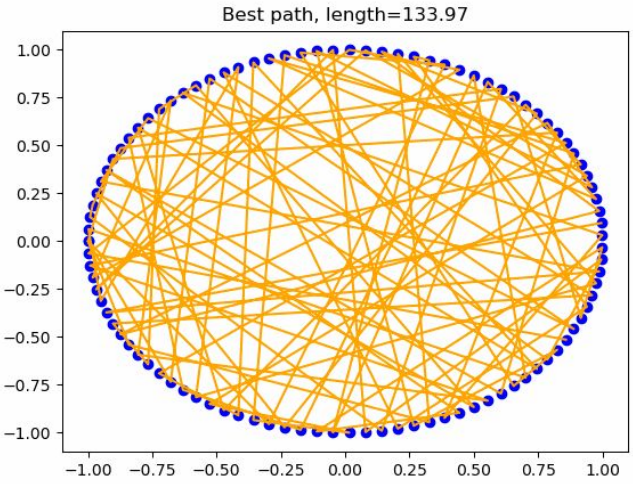

Figure 2.3b: GA (No linkage) on sparse circle dataset (10^5 iterations)

3 Analysis

3.1 Comparison of Random, RMHC, and GA performance

3.2 Effects of mutation-recombination ratio on GA performance

3.3 Effects of linkage on GA performance

3.1 Comparison of Random, RMHC, and GA performance

3.1.1 TSP 1

The shortest-path learning curves achieved by the Random, RMHC, and GA search algorithms demonstrate that RMHC and GA both become favorable to Random within about 10^2 iterations (Figure *2.1.1a*). After this point, Random plateaus to an almost non-improving state, while RMHC and GA maintain super-log improvement through a point around 10^3 iterations. At that point, an inflection point occurs where RMHC begins to slow to sub-log improvement, while GA continues to achieve log or super-log improvement.

This section of my study was significantly hindered by the runtime performance of Python, which I will take into account in future assignments. While interesting differences between the algorithms can be observed within the first 10^4 iterations, the long-term behavior revealed when leveraging a smaller dataset to execute more iterations (see *3.1.2*) cannot be observed. It therefore cannot be concluded from this study that GA will overtake RMHC, although RMHC seems to be plateauing to sub-log improvement in later iterations, while GA maintains approximately log improvement (Figure *2.1.1a*).

3.1.2 TSP 2

Using this smaller dataset to execute more iterations, the long-term behavior of RMHC and GA for shortest-path search can be observed. These results demonstrate the superiority of GA beyond about 10^4 iterations (Figure *2.1.2a*). It is, however, worth noting that GA also seems to plateau shortly after overtaking RMHC.

3.2 Effects of mutation-recombination ratio on GA performance

At constant population (*n_processes*=100) and linkage technique ("No linkage"), the ratio between mutation and recombination in reproduction was varied between 0.2:0.8, 0.5:0.5, 0.8:0.2, and 0.9:0.1. Based on the results (long-term behavior best visualized in Figure *3.2b*), 0.8:0.2 was used in the primary experiment to compare algorithms.

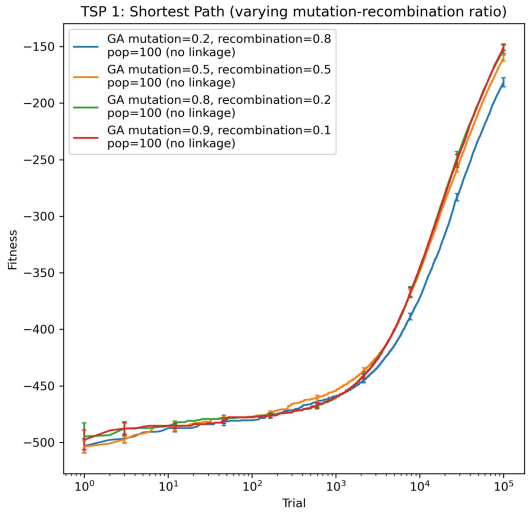

Figure 3.2a: Fitness performance across algorithms, searching for the shortest path in TSP 1.

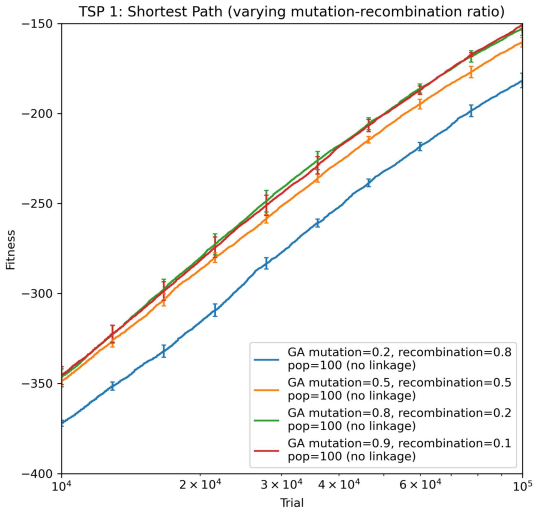

Figure 3.2b: Fitness performance across algorithms, searching for the longest path in TSP 1.

3.2 Effects of linkage on GA performance

The various versions of the GA algorithm (described in *1.4*) experimented with techniques for linking genes on the chromosome and mutating and recombining the population in accordance with these linkages. The results were inconclusive as to the best linkage technique for searching for the shortest path in TSP 1 over 10^4 iterations, as both "No linkage" and "Reverse linkage → hybrid linkage" compete for the best GA in this setting (Figure *2.1.1a*). However, when the techniques were applied to search for the shortest path in TSP 2 over 10^5 iterations, both "Reverse linkage → hybrid linkage" and "Reverse linkage → tight linkage" demonstrated improvements over "No linkage" (Figure *2.1.2a*).

Some intuition for why this might be the case: The algorithms undergo an adequate number of iterations to form similar priorities within regions, by mutating and recombining such that genes can swap regions. Then, when the points are adequately ordered by region, both algorithms increase the opportunity for mutations within the regions themselves, so as to order the path between clustered points more favorably. This intuition is supported by the jumps achieved by the algorithms where they change mutation/combination types (near $3x10^4$ for "Reverse linkage → hybrid linkage" and near $5x10^4$ for "Reverse linkage → tight linkage"). The jumps are much greater in the TSP 2 experiment with more iterations, indicating that the TSP 1 experiment possibly had yet to reach a point where it would have as much to gain from switching mutation/combination types.