Multiobjective Optimization Based Approach for Truth Discovery

by

Karan Jain

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved April 2019 by the
Graduate Supervisory Committee:

Guoliang Xue, Chair
Arunabha Sen
Mohamed Sarwat

ARIZONA STATE UNIVERSITY

May 2019

ABSTRACT

There are many applications where the truth is unknown. The truth values are guessed by different sources. The values of different properties can be obtained from various sources. These will lead to the disagreement in sources. An important task is to obtain the truth from these sometimes contradictory sources. In the extension of computing the truth, the reliability of sources needs to be computed. There are models which compute the precision values. In those earlier models Banerjee *et al.* (2005) Dong and Naumann (2009) Kasneci *et al.* (2011) Li *et al.* (2012) Marian and Wu (2011) Zhao and Han (2012) Zhao *et al.* (2012), multiple properties are modeled individually. In one of the existing works, the heterogeneous properties are modeled in a joined way. In that work, the framework i.e. Conflict Resolution on Heterogeneous Data (CRH) framework is based on the single objective optimization. Due to the single objective optimization and non-convex optimization problem, only one local optimal solution is found. As this is a non-convex optimization problem, the optimal point depends upon the initial point. This single objective optimization problem is converted into a multi-objective optimization problem. Due to the multi-objective optimization problem, the Pareto optimal points are computed. In an extension of that, the single objective optimization problem is solved with numerous initial points. The above two approaches are used for finding the solution better than the solution obtained in the CRH with median as the initial point for the continuous variables and majority voting as the initial point for the categorical variables. In the experiments, the solution, coming from the CRH, lies in the Pareto optimal points of the multi-objective optimization and the solution coming from the CRH is the optimum solution in these experiments.

ACKNOWLEDGMENT

I would like to thank Dr. Guoliang Xue for his guidance and support. This research was supported in part by NSF grant 1704092. The information reported here does not reflect the position or the policy of the federal government.

TABLE OF CONTENTS

Page

LIST OF TABLES iv

LIST OF FIGURES v

CHAPTER

1 INTRODUCTION 1

1.1 Objective 1

1.2 Motivation 1

2 BACKGROUND LITERATURE 3

3 METHODOLOGY 5

3.1 Problem Formulation 5

3.2 CRH Framework 8

3.3 Two Approaches 9

3.3.1 Single Objective Optimization With Multi-Start (SOOWMS) 9

3.3.2 Non-dominated Sorting Genetic Algorithm II (NSGA) 11

4 EXPERIMENTS AND RESULTS 13

4.1 Dataset Information 13

4.2 Small Dataset 14

4.2.1 Single Objective Optimization With Multi-Start (SOOWMS) 14

4.2.2 Non-dominated Sorting Genetic Algorithm II (NSGA) 31

4.3 Large Dataset 46

5 CONCLUSION 68

REFERENCES 69

LIST OF TABLES

Table Page

4.1 Statistics of SOOWMS. 31

LIST OF FIGURES

Figure Page

3.1 Visualization of Non Convex Optimization Problem. 10

3.2 Flow Chart of Non-dominated Sorting Genetic Algorithm II (NSGA). . . 12

4.1 SOOWMS on the Problem Involving 2 Variables Having Small Range. . 15

4.2 Plot of Function Value with Respect to x_1 and x_2. 16

4.3 SOOWMS on the Problem Involving 2 Variables with a Large Range. . . 17

4.4 Plot of Function Value with Respect to x_1 and x_2. 18

4.5 SOOWMS on the Problem Involving 4 Variables with a Small Range. . . 20

4.6 SOOWMS on the Problem Involving 4 Variables with a Large Range and a Function Change. 22

4.7 SOOWMS on the Problem Involving 8 Variables with a Small Range. . . 24

4.8 SOOWMS on the Problem Involving 8 Variables with a Large Range. . . 26

4.9 SOOWMS on the Problem Involving 10 Variables with a Small Range. . 28

4.10 SOOWMS on the Problem Involving 10 Variables with a Large Range and a Function Change. 30

4.11 NSGA on the Problem Involving 2 Variables Having a Small Range. . . . 32

4.12 NSGA on the Problem Involving 2 Variables Having a Large Range. . . . 34

4.13 NSGA on the Problem Involving 4 Variables Having a Small Range. . . . 36

4.14 NSGA on the Problem Involving 4 Variables Having a Large Range. . . . 38

4.15 NSGA on the Problem Involving 8 Variables with a Small Range. 40

4.16 NSGA on the Problem Involving 8 Variables Having a Large Range. . . . 42

4.17 NSGA on the Problem Involving 10 Variables Having a Small Range . . 44

4.18 NSGA on the Problem Involving 10 Variables Having a Large Range and a Function Change. 46

4.19 NSGA on the Problem Having 14 Homogeneous Variables. 49

Figure Page

4.20 NSGA on the Problem Having 28 Homogeneous Variables. 52

4.21 NSGA on the Problem Having 38 Homogeneous Variables. 55

4.22 NSGA on the Problem Having 62 Homogeneous Variables. 58

4.23 NSGA on the Problem Having 14 Heterogeneous Variables. 61

4.24 NSGA on the Problem Having 28 Heterogeneous Variables. 64

4.25 NSGA on the Problem Having 38 Heterogeneous Variables. 67

Chapter 1

INTRODUCTION

In this section, the objective of the thesis is provided, which is followed by the motivation of the thesis.

1.1 Objective

The purpose of the thesis is to express a single objective optimization problem in Truth Discovery into a multi-objective optimization problem. This is accompanied by an analysis of the results of both single objective optimization problem as well as its similar multi-objective optimization problem with the different approaches. Throughout an investigation, either the result of the single optimization problem is present in the Pareto points of the multi-objective optimization problem Srinivas and Deb (1994) Deb *et al.* (2002) or a better solution is present in the Pareto points of the multi-objective optimization problem.

1.2 Motivation

Due to the growth of big data, the companies are collecting data from different origins including business activities, social media, etc. With the rise of big data, there are wide variations of data values. The information of sufferers can be found from different hospitals. The weather information of different cities can be recorded by different laboratories. The information sent by the satellites can be inconsistent. Due to the malfunctioning of the machines, error in communications, intentional variations, etc., the values of different sources can be contradictoryLi *et al.* (2014). The reliance on unreliable sources can lead to a huge loss in terms of money and data.

For example, while getting an account information of the customer in banks, if the value of the account number is incorrect, the money can be transferred to a wrong person. One can contact a wrong person by an incorrect phone number. One can get the wrong arrival date of a train in a station. One can make crazy business decisions. So, there is a requirement of finding the correct values as well as the source reliability of different sources for the future use. In an extension to this, these days, there is a lot of incorrect information present on the Web.

Chapter 2

BACKGROUND LITERATURE

There are many case scenarios where the truth is hidden. For example, in the case of weather forecasting, an average temperature for the day is predicted by different agencies. These values can be different for different agencies. There is a need to compute the truth value of the average temperature for the day. The values of different properties can be obtained from various sources. These will lead to disagreement in the sources. The important task is to obtain the truth from these clashing sources. In addition, to compute the truth, the reliability of the sources needs to be computed. This is very important to compute the reliability of the sources. As in the real world, there are always sources of different reliability. It is always useful to compute the reliability of the sources for future use. As in the case of an above example, computing the reliability of the source is always good for future use. In the database area, resolving conflicts, in case of data integration, have been studied in detail Dong and Naumann (2009) Bleiholder and Naumann (2006) Jiang (2012). There are some approaches that have been proposed to induce the truth in case of disputes. In the case of categorical data, the most commonly used method is Majority Voting. The value having a maximum number of appearances will be taken as truth. In the case of continuous data, it is the median method. These approaches essentially deal with a single data type. In addition to that, it is assuming that all sources have equal reliability measure. In the real world, there are many objects of heterogeneous properties. For example, in banks, there are many attributes for a customer like age, gender, salary, etc. In the above example, gender is of a type categorical and salary is of a type continuous. In the case of a weather forecast, the day can be described

in terms of high temperature, low temperature, weather condition, etc. In the above example, the high temperature is of a type continuous and weather condition is of a type categorical. However, it is not easy to unify the data of different properties in one model. Existing works model multiple properties separately Banerjee *et al.* (2005) Dong and Naumann (2009) Kasneci *et al.* (2011) Li *et al.* (2012) Marian and Wu (2011) Zhao and Han (2012) Zhao *et al.* (2012). The sources can behave differently with different properties for the object. As in the case of categorical data, it is either right or wrong and in case of continuous data, it is a distance from the true value. For example, in the case of categorical data, if the truth value is White and the value other than White will be equal in terms of distance from the true value. In case of continuous data, if the truth value is 60F and the observation having value 61F is closer to true value as compared to the observation having 66F. In one of the current works, it deals with data of many types i.e. heterogeneous data. In that work, the multiple heterogeneous properties are modeled in a joined way, but it is calculating only one solutionLi *et al.* (2014). In that work, the optimization framework is a single objective optimization. Due to the single objective optimization, a local optimal solution is obtained Boyd and Vandenberghe (2004). As the problem which is solved in Li *et al.* (2014) is a non-convex optimization problem, there can be many local optimal points. It may happen that the solution found in Li *et al.* (2014) is a local optimal point. There is a need to devise another method for computing another global optimal point if exist.

Chapter 3

METHODOLOGY

In this section, firstly, the single objective optimization problem introduced in Li *et al.* (2014) is expressed into the multi-objective optimization problem. It is followed by the description of the CRH framework. After that, two methods which are used to find a more optimal solution than the solution Li *et al.* (2014) are provided. One of the methods is Single Objective Optimization With Multi-Start and another method is Non-dominated Sorting Genetic Algorithm II (NSGA).

3.1 Problem Formulation

The single objective optimization problem, introduced in Conflict Resolution on Heterogeneous Data (CRH) framework Li *et al.* (2014), is converted into the multi-objective optimization problem. In the paper Li *et al.* (2014), there are K sources and M heterogenous properties. The problem in Li *et al.* (2014) has the following form:

$$\begin{aligned} \underset{X^*,W}{\text{minimize}} \quad & f(X^*, W) = \sum_{k=1}^{K} w_k \sum_{i=1}^{N} \sum_{m=1}^{M} d_m(x_{im}^*, x_{im}^k) \\ \text{subject to} \quad & \xi(W) = \sum_{k=1}^{K} e^{-w_k} = 1, \\ & W \geq 0. \end{aligned} \tag{3.1}$$

In problem (3.1), W is a weight vector having K elements. x_{im}^k is a value of i^{th} object corresponding to m^{th} property given by k^{th} source. N is the number of data objects. $d_m(*,*)$ is a loss function. $\xi(W)$ is a regularization function whose value is equal to 1. W corresponds to the reliability of the sources. The regularization

function is used to constrain the values of W. In problem (3.1), X^* is defined below:

$$X^* = \begin{bmatrix} x^*_{11} & x^*_{12} & \cdots & x^*_{1M} \\ x^*_{21} & x^*_{22} & \cdots & x^*_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ x^*_{N1} & x^*_{N2} & \cdots & x^*_{NM} \end{bmatrix}. \tag{3.2}$$

In equation (3.2), x^*_{ij} is truth value of i^{th} object corresponding to j^{th} property. In equation (3.2), X^* can be written as a vector having N*M elements as below:

$$X^* = \begin{bmatrix} x^*_{11} & x^*_{12} & x^*_{13} & \ldots & x^*_{N*M-1} & x^*_{NM} \end{bmatrix}. \tag{3.3}$$

In problem (3.1), W is defined below:

$$W = \begin{bmatrix} w_1 & w_2 & w_3 & \ldots & w_K \end{bmatrix}. \tag{3.4}$$

w_i is a weight value corresponding to i^{th} source.

The problem (3.1) is transformed into the multi-objective optimization problem as follows:

$$\begin{aligned} \underset{X^*,W}{\text{minimize}} \quad & (f^1(X^*,W), f^2(X^*,W), \ldots, f^M(X^*,W)) \\ \text{subject to} \quad & \xi(W) = \sum_{k=1}^{K} e^{-w_k} = 1, \\ & W \geq 0. \end{aligned} \tag{3.5}$$

In the multi-objective problem (3.5), W is a vector having K elements. In the problem (3.5), $f^i(X^*, W)$ is defined as follows:

$$f^i(X^*, W) = \sum_{k=1}^{K} w_k \sum_{j=1}^{N} d_M(x^*_{jM}, x^k_{jM}),\ i = 1, \ldots, M. \tag{3.6}$$

The loss function, used in the multi-objective optimization problem, is selected according to the problem and properties. On categorical data, one of the most commonly used loss function is the 0-1 loss in which an error is incurred if the observation is different from the truth. If the m^{th} property is categorical, the formula of the 0-1 loss, where x^*_{im} is truth and x^k_{im} is observation, is as follows:

$$d(x^*_{im}, x^k_{im}) = \begin{cases} 1 & if\ x^*_{im} \neq x^k_{im}, \quad (3.7) \\ 0 & if\ x^*_{im} = x^k_{im}. \quad (3.8) \end{cases}$$

On continuous data, there are many loss functions. One of the loss functions is the normalized absolute deviation. If the m^{th} property is continuous, the formula of the normalized absolute deviation, where x^*_{im} is truth and x^k_{im}, x^1_{im} and x^K_{im} are observations, is as follows:

$$d(x^*_{im}, x^k_{im}) = \frac{|x^*_{im} - x^k_{im}|}{\sigma(x^1_{im}, \ldots, x^K_{im})}. \tag{3.9}$$

In equation (3.9), $\sigma(x^1_{im}, \ldots, x^K_{im})$ is a standard deviation.

In case of continuous data, another loss function is the normalized squared loss. If the m^{th} property is continuous, the formula of the normalized squared loss, where x^*_{im} is truth and x^k_{im}, x^1_{im} and x^K_{im} are observations, is as follows:

$$d(x_{im}^*, x_{im}^k) = \frac{(x_{im}^* - x_{im}^k)^2}{\sigma(x_{im}^1, \ldots, x_{im}^K)}. \tag{3.10}$$

In equation (3.10), $\sigma(x_{im}^1, \ldots, x_{im}^K)$ is a standard deviation. This notation is used in subsequent chapters.

The solution of the problem (3.5) is a set of Pareto points. The set of Pareto points is a collection of the Pareto optimal points. The point P is said to be Pareto-optimal if no solution of problem (3.5) dominates P. Here, P is an (M*N+K)-dimensional vector where M is a number of heterogeneous properties and N is the number of objects and K is the number of sources. Vector u=$(u_1, u_2, \ldots, u_{M*N+K})$ dominates Vector v=$(v_1, v_2, \ldots, v_{M*N+K})$ if u is better than v with respect to one objective and not worse than with respect to all other objectives.

3.2 CRH Framework

Algorithm 1 CRH Framework.

Input : Data from K sources: X^1, X^2, ..., X^K.
Output :Truth X^*=$\{x_{im}^*\}_{i=1,m=1}^{N,M}$, source weights W =$\{w_1, w_2, \ldots, w_K\}$.

1: Initialize the truths X^*;
2: **while** Convergence criterion is not satisfied **do**
3: Update source weights W while minimizing the equation corresponding to the problem (3.1) and keeping X^* constant;
4: Update truth X^* while minimizing the equation corresponding to the problem (3.1) and keeping W constant;
return X^* and W.

The algorithm of the CRH framework is given in an Algorithm 1Li *et al.* (2014). X^* is started with an initial point. The values of W and X^* are updated according to the block coordinate descent approachBertsekas (2006). The output of the algorithm is the value of X^* and W for which the equation corresponding to problem (3.1) is

minimized. As, problem (3.1) is a non-convex optimization problem, the optimum value of $f(X^*, W)$ depends upon the initial point.

3.3 Two Approaches

The two approaches are used for finding the solution, more optimal than the solution found from the Conflict Resolution on Heterogeneous Data framework Li *et al.* (2014). These are as follows:

3.3.1 Single Objective Optimization With Multi-Start (SOOWMS)

The problem (3.1) is a non-convex optimization problem. The value of the function in search space can be visualized as shown in figure 3.1.

In the figure, if an initial point is A or B, then an optimal point will be C. If an initial point is D, then an optimal point will be E. This is evident from the figure that E is a local optimal point but C is the global optimal point. The different optimal points can be obtained from the different initial points. In this method, the CRH method has been applied with the different initial points. The problem is of form as below:

$$\begin{aligned} \underset{X^*,W}{\text{minimize}} \quad & f(X^*, W) = \sum_{k=1}^{K} w_k \sum_{i=1}^{N} \sum_{m=1}^{M} d_m(x_{im}^*, x_{im}^k) \\ \text{subject to} \quad & \xi(W) = \sum_{k=1}^{K} e^{-w_k} = 1, \\ & W \geq 0. \end{aligned} \tag{3.11}$$

Please note that the data in these experiments is continuous and there is no categorical data. As there is a continuous data, the normalized squared loss is used as a loss function. If the m^{th} property is continuous, the formula of the normalized squared loss, where x_{im}^* is truth and x_{im}^k, x_{im}^1 and x_{im}^K are observations, is as follows:

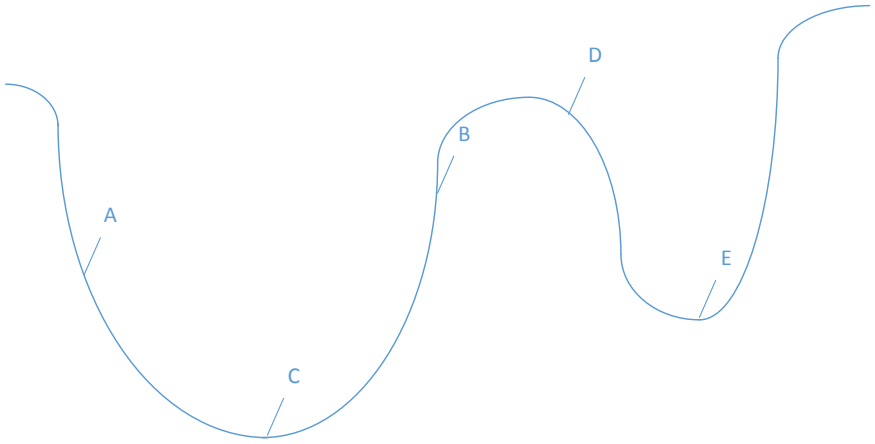

Figure 3.1: Visualization of Non Convex Optimization Problem.

$$d(x^*_{im}, x^k_{im}) = \frac{(x^*_{im} - x^k_{im})^2}{\sigma(x^1_{im}, \ldots, x^K_{im})}. \tag{3.12}$$

In equation (3.12), $\sigma(x^1_{im}, \ldots, x^K_{im})$ is a standard deviation. It has been used in subsequent chapters. In each experiment, there is a value of ϵ in each dimension which is used to select the initial points. The ϵ value of a dimension is minimum distance between two initial points along the dimension. The value of ϵ is selected differently in every dimension and in every experiment. If the value of the ϵ is increased, the number of the initial points is reduced. If the value of the ϵ is decreased, the number

of the initial points is increased. The solutions obtained with many different initial points is compared with the solution obtained by applying the CRH method with an initial point as the median Li *et al.* (2014). As this is a time taking approach, it has been applied to small data set only.

3.3.2 Non-dominated Sorting Genetic Algorithm II (NSGA)

Non-dominated Sorting Genetic Algorithm II is a multi-objective genetic algorithm. This algorithm is fast and elitist Srinivas and Deb (1994) Deb *et al.* (2002) Agrawal *et al.* (1995). This approach is applied to both small and large data set. The description of the NSGA is as follows:

Description The population of size N is initialized randomly. Once the population is initialized, the population is classified and sorted based on non-domination into each front. The first front being completely non-dominant in the current candidates. The first front dominates second front and so on. The rank value of 1 is assigned to members of the first front and rank value of 2 is assigned to the second front and so on. The solution having less rank value is preferred as compared to the solution having a higher rank value. In addition to the rank value, there is another measure which is said to be crowding distance. The crowding distance of Point P is calculated as the average distance of two points on either side of P along each of the objectives. If both solutions belong to the same front, then the solution that is having higher crowding distance is selected first. The child population of size N is created by binary tournament selection on parent population based on rank and crowding distance and simulated binary crossover and mutation operation. In order to preserve an elitism, the population of parent and child is combined into the resultant population of size

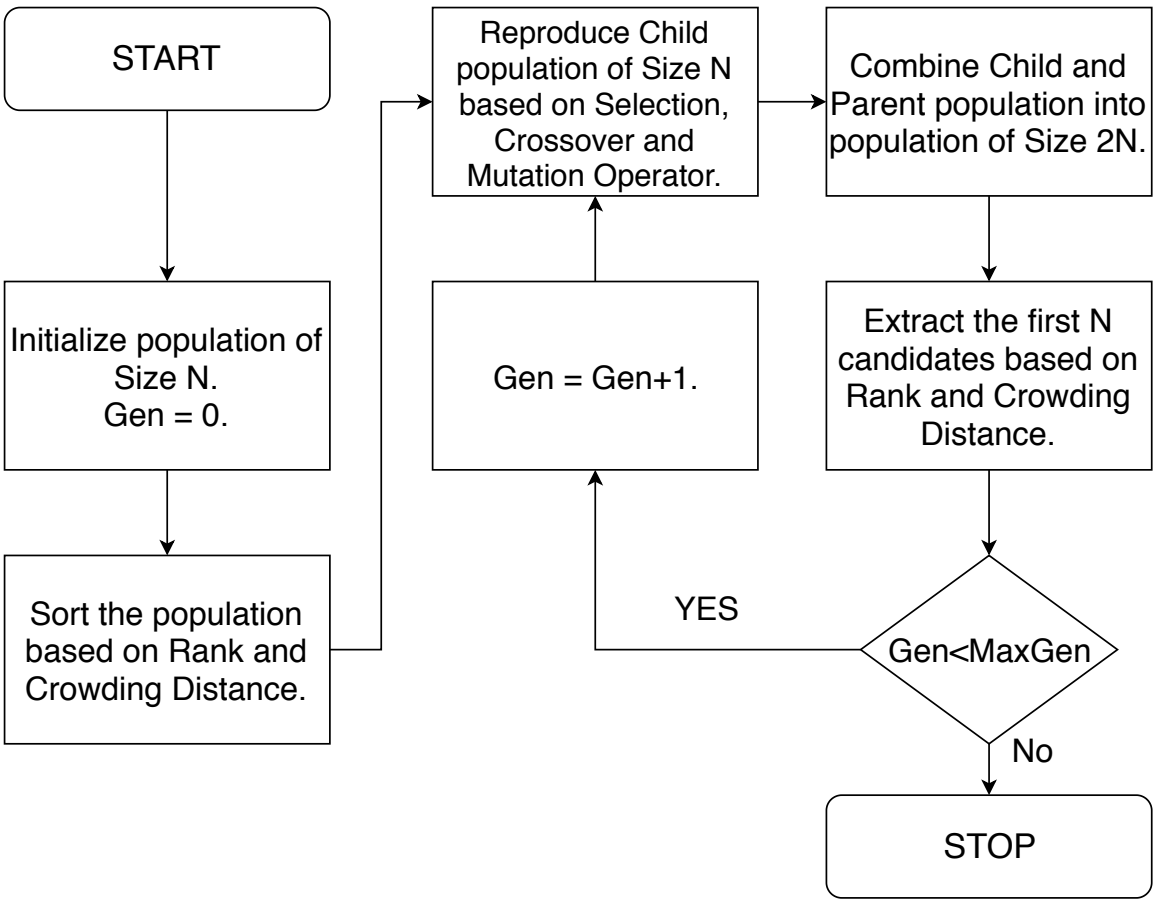

Figure 3.2: Flow Chart of Non-dominated Sorting Genetic Algorithm II (NSGA).

2N. Out of this resultant population, first N solutions, having less rank and in case of the tie, having higher crowding distance are selected. These N solutions act as a parent for the next iteration. While implementing NSGA, the link [1] has been referred. The flow chart of NSGA has been given in figure 3.2. The solution of the NSGA is the set of Pareto points.

[1]https://www.mathworks.com/matlabcentral/fileexchange/49806-matlab-code-for-constrained-nsga-ii-dr-s-baskar-s-tamilselvi-and-p-r-varshini

Chapter 4

EXPERIMENTS AND RESULTS

In this section, the information about the dataset is provided. After that, the experiments that are performed on small data using both approaches (SOOWMS and NSGA) are presented. After that, the experiments that are conducted on large data set using NSGA are presented. In the case of a large data set, the data is of both heterogeneous and homogeneous type.

4.1 Dataset Information

For a large data set, Weather Forecast Data Set has been used. As it contains heterogeneous types of properties, it is an adequate data set for testing. The data set is available at the link [1] Li *et al.* (2014). The data is crawled from the three types of platforms: Wunderground [2] , HAM weather [3] , and World Weather Online [4] . The data of three properties are crawled: high temperature, low temperature and weather condition for the day. Of these three properties, the first two are continuous and the last is categorical.

[1] https://cse.buffalo.edu/ jing/software.htm

[2] http://www.wunderground.com

[3] http://www.hamweather.com

[4] http://www.worldweatheronline.com

4.2 Small Dataset

4.2.1 Single Objective Optimization With Multi-Start (SOOWMS)

a) First experiment having two variables with a small range:- In this experiment, there are two variables. The problem solved in this experiment is as follows:

$$\begin{aligned} \underset{x_1,x_2,w_1,w_2,w_3}{\text{minimize}} \quad & f_2(x_1, x_2, w_1, w_2, w_3) \\ \text{subject to} \quad & e^{-w_1} + e^{-w_2} + e^{-w_3} = 1, \\ & w_i \geq 0. \end{aligned} \tag{4.1}$$

In problem (4.1), $f_2(x_1, x_2, w_1, w_2, w_3)$ is defined as follows:

$$\begin{aligned} f_2(x_1, x_2, w_1, w_2, w_3) \quad &= \frac{1}{\sigma(1,3,6)} * [w_1 * (x_1 - 3)^2 + w_2 * (x_1 - 1)^2 \\ &+ w_3 * (x_1 - 6)^2] \\ &+ \frac{1}{\sigma(1,4,5)} * [w_1 * (x_2 - 4)^2 + w_2 * (x_2 - 5)^2 \\ &+ w_3 * (x_2 - 1)^2]. \end{aligned} \tag{4.2}$$

ϵ_1 is a minimum distance between two initial points along x_1 dimension and ϵ_2 is a minimum distance between two initial points along x_2 dimension. The value of the ϵ_1 is 0.001 and the value of the ϵ_2 is 0.01. The number of initial points, in this case, is calculated as follows:

$$(\frac{6-1}{0.001} + 1) * (\frac{5-1}{0.01} + 1) = 2005401. \tag{4.3}$$

The number of initial points is 2005401. After applying the CRH method with

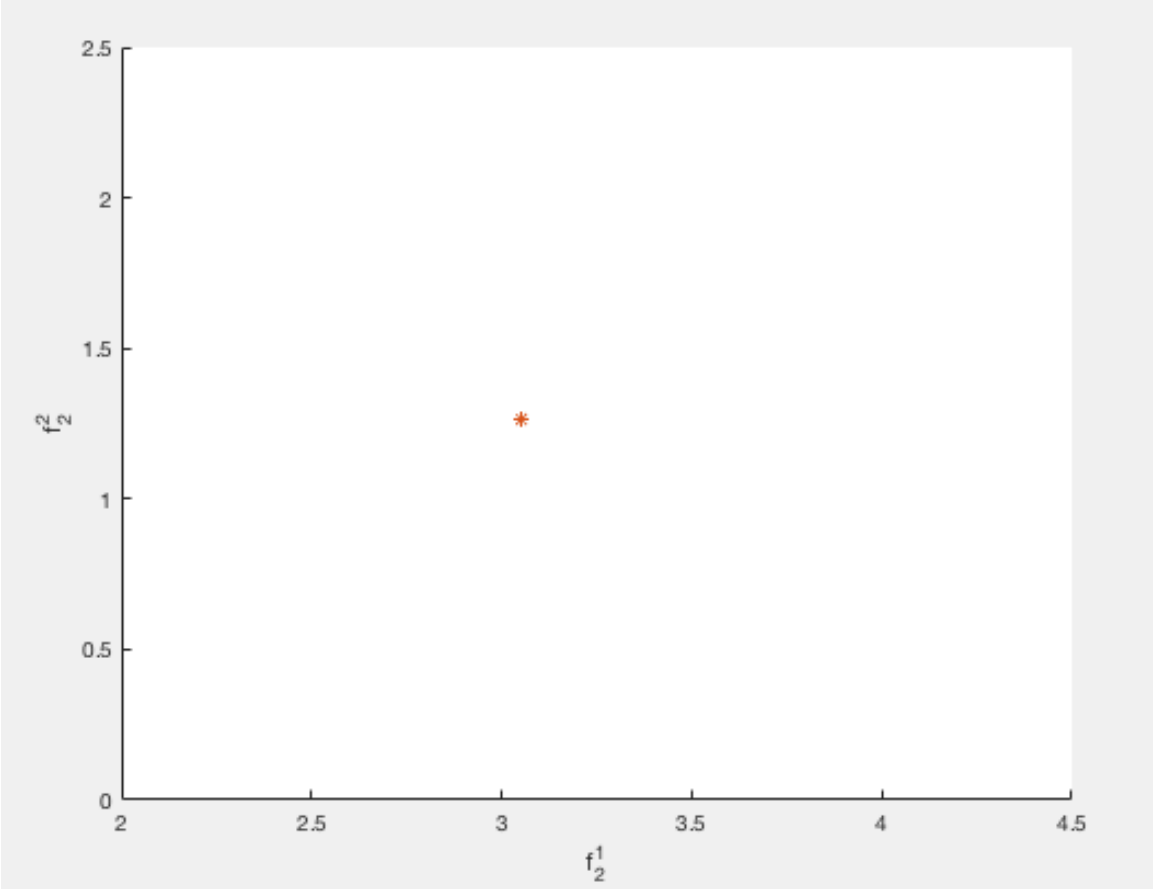

Figure 4.1: SOOWMS on the Problem Involving 2 Variables Having Small Range.

2005401 different initial points, only one optimal solution is found. The figure 4.1 shows the plot of f_2^1 and f_2^2. These functions f_2^1 and f_2^2 are defined in equations (4.26) and (4.27). In figure 4.1, the point shows the solution obtained from the SOOWMS and the solution obtained from the CRH method with the median as an initial point Li *et al.* (2014). The figure shows that both points are the same.

For given x_1 and x_2, the values of w_1, w_2 and w_3 are computed Li *et al.* (2014) while minimizing f_2 in problem (4.1). The minimum f_2 values are computed for given x_1 and x_2. The figure 4.2 shows the plot of minimum function value, x_1 and x_2. In figure 4.2, Min(f_2) is the minimum value of f_2 for given x_1 and x_2. As it is evident from the figure, the problem corresponding to problem (4.1) is a non-convex

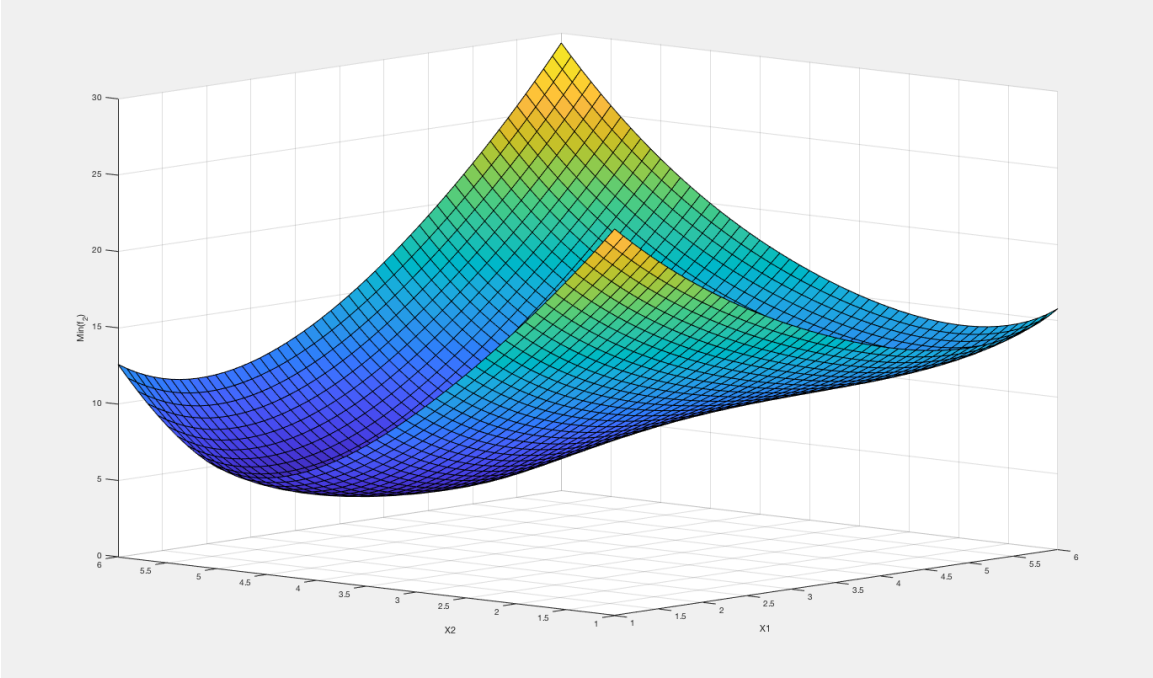

Figure 4.2: Plot of Function Value with Respect to x_1 and x_2.

optimization problem with one minimal point.

b) Second experiment involving two variables with a large range:- In this experiment, there are two variables. The problem solved in this experiment is as follows:

$$\begin{aligned} \underset{x_1,x_2,w_1,w_2,w_3}{\text{minimize}} \quad & g_2(x_1, x_2, w_1, w_2, w_3) \\ \text{subject to} \quad & e^{-w_1} + e^{-w_2} + e^{-w_3} = 1, \\ & w_i \geq 0. \end{aligned} \tag{4.4}$$

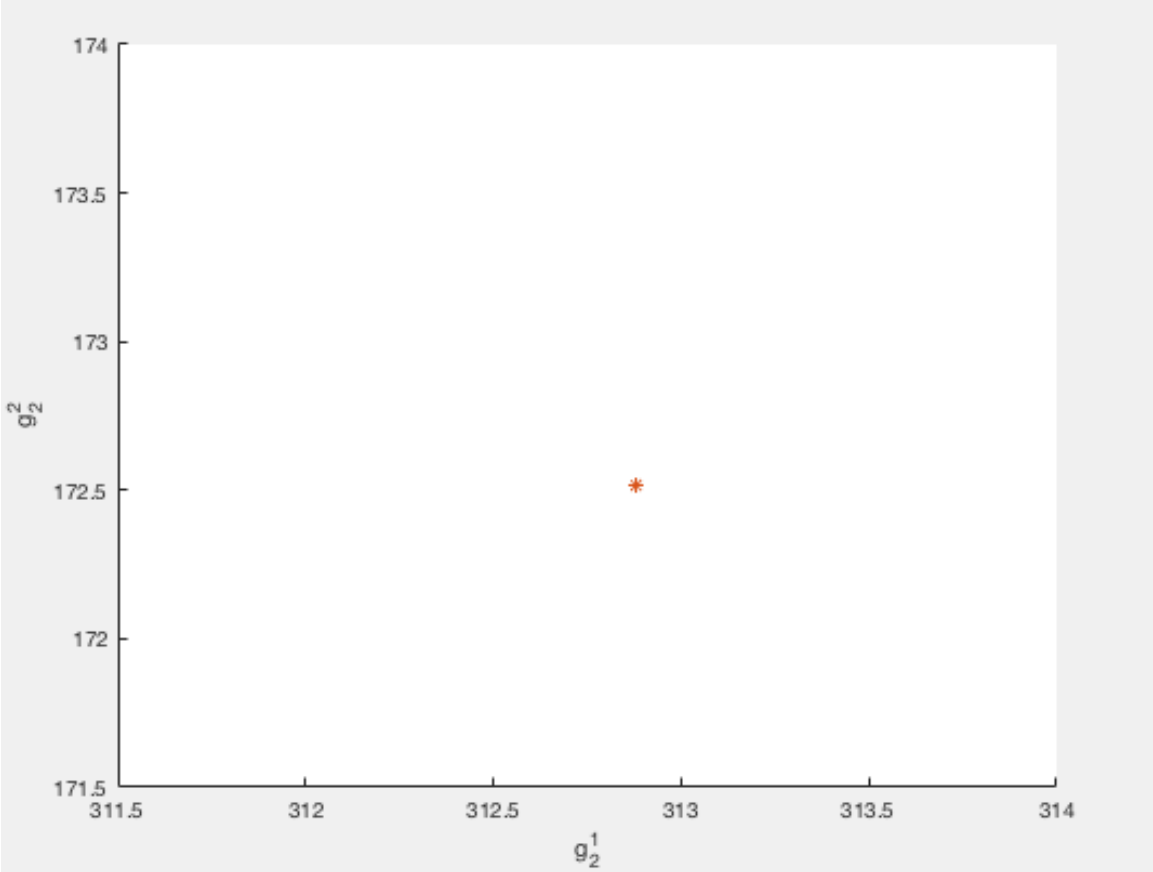

Figure 4.3: SOOWMS on the Problem Involving 2 Variables with a Large Range.

In problem (4.4), $g_2(x_1, x_2, w_1, w_2, w_3)$ is defined as follows:

$$\begin{aligned} g_2(x_1, x_2, w_1, w_2, w_3) \quad &= \frac{1}{\sigma(100, 300, 500)} * [w_1 * (x_1 - 300)^2 \\ &+ w_2 * (x_1 - 100)^2 + w_3 * (x_1 - 500)^2] \\ &+ \frac{1}{\sigma(400, 500, 600)} * [w_1 * (x_2 - 400)^2 + w_2 * (x_2 - 600)^2 \\ &+ w_3 * (x_2 - 500)^2]. \end{aligned} \tag{4.5}$$

ϵ_1 is a minimum distance between two initial points along x_1 dimension and ϵ_2 is a minimum distance between two initial points along x_2 dimension. The value of the ϵ_1 is 0.5 and the value of the ϵ_2 is 0.5. The number of initial points, in this case, is

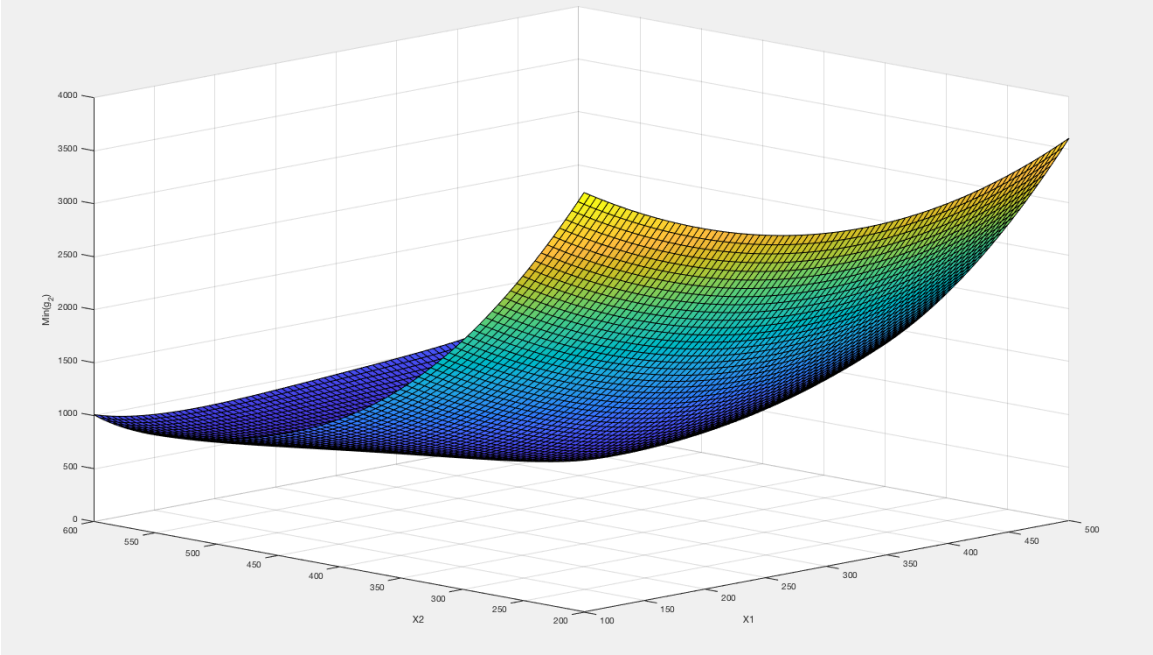

Figure 4.4: Plot of Function Value with Respect to x_1 and x_2.

calculated as follows:

$$(\frac{500-100}{0.5}+1)*(\frac{600-400}{0.5}+1)=321201. \tag{4.6}$$

The number of initial points is 321201. After applying the CRH method with 321201 different initial points, only one optimal solution is found. The figure 4.3 shows the plot of g_2^1 and g_2^2. These functions g_2^1 and g_2^2 are defined in equations (4.29) and (4.30). In figure 4.3, the point shows the solution obtained from the SOOWMS and the solution obtained from the CRH method with the median as an initial point. As evident from the figure, both points are the same.

For given x_1 and x_2, the values of w_1, w_2 and w_3 are computed Li *et al.* (2014) while minimizing g_2 in problem (4.4). The minimum g_2 values are computed for given x_1 and x_2. The figure 4.4 shows the plot of minimum function value, x_1 and x_2. In figure 4.4, Min(g_2) is the minimum value of g_2 for a given x_1 and x_2. As it is evident

from the figure, the problem (4.4) is a non-convex optimization with one minimal point.

c) Third experiment involving four variables with a small range:- In this experiment, there are four variables. The problem solved in this experiment is as follows:

$$\begin{aligned} \underset{x_1,\ldots,x_4,w_1,w_2,w_3}{\text{minimize}} \quad & f_4(x_1,\ldots,x_4,w_1,w_2,w_3) \\ \text{subject to} \quad & e^{-w_1}+e^{-w_2}+e^{-w_3}=1, \\ & w_i \geq 0. \end{aligned} \tag{4.7}$$

In problem (4.7), $f_4(x_1,\ldots,x_4,w_1,w_2,w_3)$ is defined as follows:

$$\begin{aligned} f_4(x_1,\ldots,x_4,w_1,w_2,w_3) \quad &= \frac{1}{\sigma(1,4,6)} * [w_1 * (x_1-1)^2 + w_2 * (x_1-4)^2 \\ &+ w_3 * (x_1-6)^2] \\ &+ \frac{1}{\sigma(1,2,4)} * [w_1 * (x_3-2)^2 + w_2 * (x_3-4)^2 \\ &+ w_3 * (x_3-1)^2] \\ &+ \frac{1}{\sigma(3,5,7)} * [w_1 * (x_2-3)^2 + w_2 * (x_2-5)^2 \\ &+ w_3 * (x_2-7)^2] \\ &+ \frac{1}{\sigma(1,5)} * [w_1 * (x_4-1)^2 + w_2 * (x_4-5)^2]. \end{aligned} \tag{4.8}$$

ϵ is minimum distance between two initial points along any dimension. The value of the ϵ is 0.1. The number of initial points is calculated as follows:

$$(\frac{6-1}{0.1}+1) * (\frac{7-3}{0.1}+1) * (\frac{4-1}{0.1}+1) * (\frac{5-1}{0.1}+1) = 2657661. \tag{4.9}$$

The number of initial points is 2657661. After applying the CRH method with

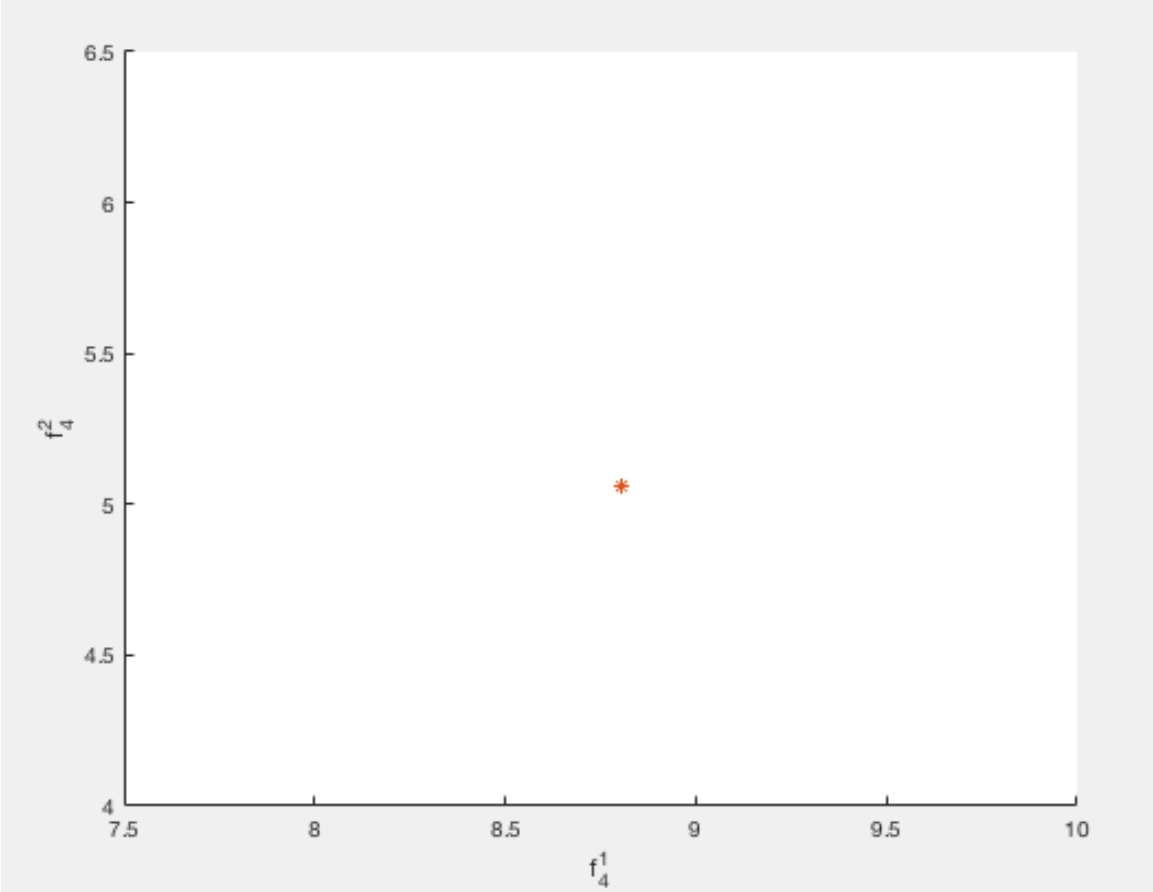

Figure 4.5: SOOWMS on the Problem Involving 4 Variables with a Small Range.

2657661 different initial points, only one optimal solution is found. The figure 4.5 shows the plot of f_4^1 and f_4^2. These functions f_4^1 and f_4^2 are defined in equations (4.32) and (4.33). In figure 4.5, the point shows the solution obtained from the SOOWMS and the solution obtained from the CRH method with the median as an initial point Li *et al.* (2014). The figure demonstrates that both points are the same.

d) Fourth experiment involving four variables with a large range and a function change:- In this experiment, there are four variables. The problem solved in this experiment is as follows:

$$\begin{aligned} \underset{x_1,\ldots,x_4,w_1,w_2,w_3}{\text{minimize}} \quad & g_4(x_1,\ldots,x_4,w_1,w_2,w_3) \\ \text{subject to} \quad & e^{-w_1}+e^{-w_2}+e^{-w_3}=1, \\ & w_i \geq 0. \end{aligned} \tag{4.10}$$

In problem (4.10), $g_4(x_1,\ldots,x_4,w_1,w_2,w_3)$ is defined as follows:

$$\begin{aligned} g_4(x_1,\ldots,x_4,w_1,w_2,w_3) \quad &= \frac{1}{\sigma(1,46,50)} * [w_1 * (x_1-1)^2 + w_2 * (x_1-46)^2 \\ &+ w_3 * (x_1-50)^2] \\ &+ \frac{1}{\sigma(2,15,22)} * [w_1 * (x_3-2)^2 + w_2 * (x_3-15)^2 \\ &+ w_3 * (x_3-22)^2] \\ &+ \frac{1}{\sigma(3,23,33)} * [w_1 * (x_2-3)^2 + w_2 * (x_2-23)^2 \\ &+ w_3 * (x_2-33)^2] \\ &+ \frac{1}{\sigma(4,18,44)} * [w_1 * (x_4-4)^2 + w_2 * (x_4-18)^2 \\ &+ w_3 * (x_4-44)^2]. \end{aligned} \tag{4.11}$$

ϵ is a minimum distance between two initial points along any dimension. The value of the ϵ is 1. The number of initial points is calculated as follows:

$$(\frac{50-1}{1}+1) * (\frac{33-3}{1}+1) * (\frac{22-2}{1}+1) * (\frac{44-4}{1}+1) = 1334550. \tag{4.12}$$

The number of initial points is 1334550. After applying the CRH method with 1334550 different initial points, only one optimal solution is found. The figure 4.6 shows the plot of g_4^1 and g_4^2. These functions g_4^1 and g_4^2 are defined in equations (4.35) and (4.36). In figure 4.6, the point shows the solution obtained from the SOOWMS

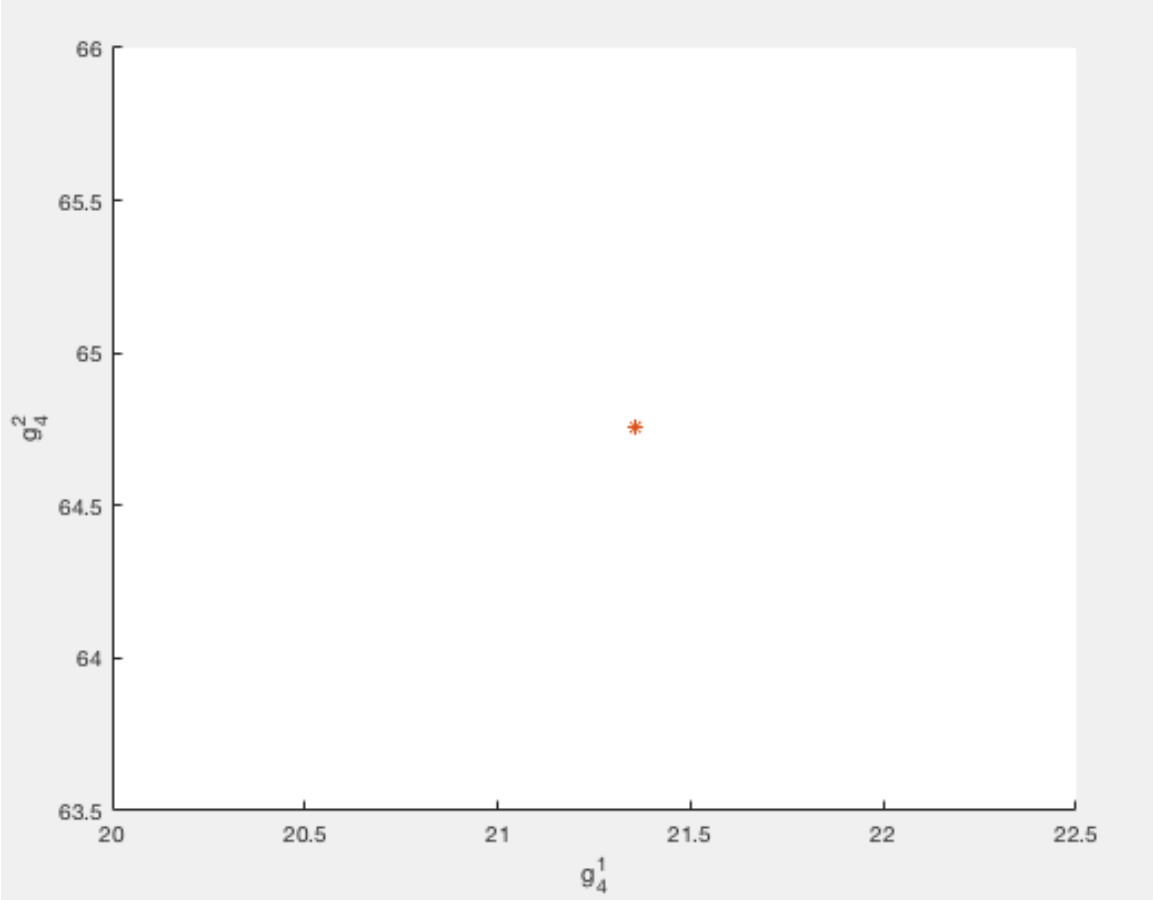

Figure 4.6: SOOWMS on the Problem Involving 4 Variables with a Large Range and a Function Change.

and the solution obtained from the CRH method with the median as an initial point Li *et al.* (2014). According to the figure, both points are the same.

e) Fifth experiment involving eight variables with a small range:- In this experiment, there are eight variables. The problem solved in this experiment is as follows:

$$\begin{aligned} \underset{x_1,\ldots,x_8,w_1,w_2,w_3}{\text{minimize}} \quad & f_8(x_1,\ldots,x_8,w_1,w_2,w_3) \\ \text{subject to} \quad & e^{-w_1}+e^{-w_2}+e^{-w_3}=1, \\ & w_i \geq 0. \end{aligned} \tag{4.13}$$

In problem (4.13), $f_8(x_1, \ldots, x_8, w_1, w_2, w_3)$ is defined as follows:

$$\begin{aligned} f_8(x_1, \ldots, x_8, w_1, w_2, w_3) &= \frac{1}{\sigma(1,2)} * [w_1 * (x_1 - 1)^2 + w_2 * (x_1 - 2)^2] \\ &+ \frac{1}{\sigma(4, 4.3, 5)} * [w_1 * (x_3 - 4)^2 + w_2 * (x_3 - 4.3)^2 \\ &+ w_3 * (x_3 - 5)^2] \\ &+ \frac{1}{\sigma(0.1, 0.5)} * [w_1 * (x_5 - 0.1)^2 + w_3 * (x_5 - 0.5)^2] \\ &+ \frac{1}{\sigma(6, 6.5)} * [w_1 * (x_7 - 6)^2 + w_2 * (x_7 - 6.5)^2] \\ &+ \frac{1}{\sigma(6, 6.5)} * [w_2 * (x_2 - 6)^2 + w_3 * (x_2 - 6.5)^2] \\ &+ \frac{1}{\sigma(7, 7.7)} * [w_1 * (x_4 - 7)^2 + w_2 * (x_4 - 7.7)^2] \\ &+ \frac{1}{\sigma(8, 8.5)} * [w_1 * (x_6 - 8.5)^2 + w_3 * (x_6 - 8)^2] \\ &+ \frac{1}{\sigma(9, 9.5, 9.7)} * [w_1 * (x_8 - 9)^2 + w_2 * (x_8 - 9.5)^2 \\ &+ w_3 * (x_8 - 9.7)^2]. \end{aligned} \tag{4.14}$$

ϵ is a minimum distance between two initial points along any dimension. The value of the ϵ is 0.1. The number of initial points, in this case, is calculated as follows:

$$\begin{aligned} &(\frac{2-1}{0.1} + 1) * (\frac{6.5-6}{0.1} + 1) * (\frac{5-4}{0.1} + 1) * (\frac{7.7-7}{0.1} + 1) \\ &* (\frac{0.5-0.1}{0.1} + 1) * (\frac{8.5-8}{0.1} + 1) * (\frac{6.5-6}{0.1} + 1) * (\frac{9.7-9}{0.1} + 1) = 8363520. \end{aligned} \tag{4.15}$$

The number of initial points is 8363520. After applying the CRH method with 8363520 different initial points, only one optimal solution is found. The figure 4.7 shows the plot of f_8^1 and f_8^2. These functions f_8^1 and f_8^2 are defined in equations (4.38)

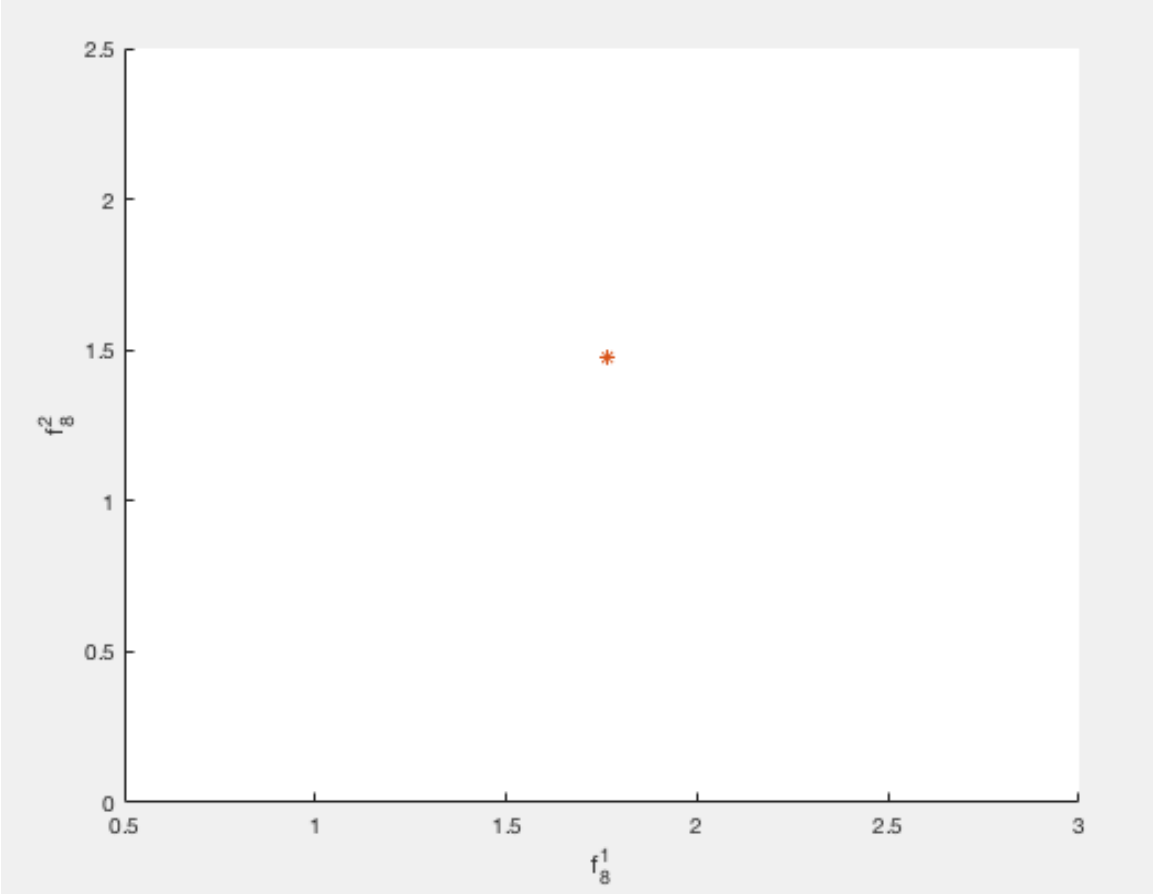

Figure 4.7: SOOWMS on the Problem Involving 8 Variables with a Small Range.

and (4.39). In figure 4.7, the point shows the solution obtained from the SOOWMS and the solution obtained from the CRH method with the median as an initial point. In the figure, both points are the same.

f) Sixth experiment involving eight variables with a large range:- In this experiment, there are eight variables. The problem solved in this experiment is as follows:

$$\begin{aligned} \underset{x_1,\ldots,x_8,w_1,w_2,w_3}{\text{minimize}} \quad & g_8(x_1,\ldots,x_8,w_1,w_2,w_3) \\ \text{subject to} \quad & e^{-w_1}+e^{-w_2}+e^{-w_3}=1, \\ & w_i \geq 0. \end{aligned} \tag{4.16}$$

In problem (4.16), $g_8(x_1, \ldots, x_8, w_1, w_2, w_3)$ is defined as follows:

$$\begin{aligned} g_8(x_1, \ldots, x_8, w_1, w_2, w_3) \quad &= \frac{1}{\sigma(1,4)} * [w_1 * (x_1 - 1)^2 + w_2 * (x_1 - 4)^2] \\ &+ \frac{1}{\sigma(4,6,7)} * [w_1 * (x_3 - 4)^2 + w_2 * (x_3 - 7)^2 \\ &+ w_3 * (x_3 - 6)^2] \\ &+ \frac{1}{\sigma(1,4)} * [w_1 * (x_5 - 1)^2 + w_3 * (x_5 - 4)^2] \\ &+ \frac{1}{\sigma(6,10)} * [w_1 * (x_7 - 6)^2 + w_2 * (x_7 - 10)^2] \\ &+ \frac{1}{\sigma(5,10)} * [w_2 * (x_2 - 5)^2 + w_3 * (x_2 - 10)^2] \\ &+ \frac{1}{\sigma(7,10)} * [w_1 * (x_4 - 7)^2 + w_2 * (x_4 - 10)^2] \\ &+ \frac{1}{\sigma(13,20)} * [w_1 * (x_6 - 20)^2 + w_3 * (x_6 - 13)^2] \\ &+ \frac{1}{\sigma(10,14,20)} * [w_1 * (x_8 - 10)^2 + w_2 * (x_8 - 14)^2 \\ &+ w_3 * (x_8 - 20)^2]. \end{aligned} \tag{4.17}$$

ϵ is a minimum distance between two initial points along any dimension. The value of the ϵ is 1. The number of initial points, in this case, is calculated as follows:

$$\begin{aligned} &(\frac{4-1}{1} + 1) * (\frac{10-5}{1} + 1) * (\frac{7-4}{1} + 1) * (\frac{10-7}{1} + 1) \\ &* (\frac{4-1}{1} + 1) * (\frac{20-13}{1} + 1) * (\frac{10-6}{1} + 1) * (\frac{20-10}{1} + 1) = 675840. \end{aligned} \tag{4.18}$$

The number of initial points is 675840. After applying the CRH method with 675840 different initial points, only one optimal solution is found. The figure 4.8 shows the plot of g_8^1 and g_8^2. These functions g_8^1 and g_8^2 are defined in equations (4.41) and (4.42). In figure 4.8, the point shows the solution obtained from the SOOWMS

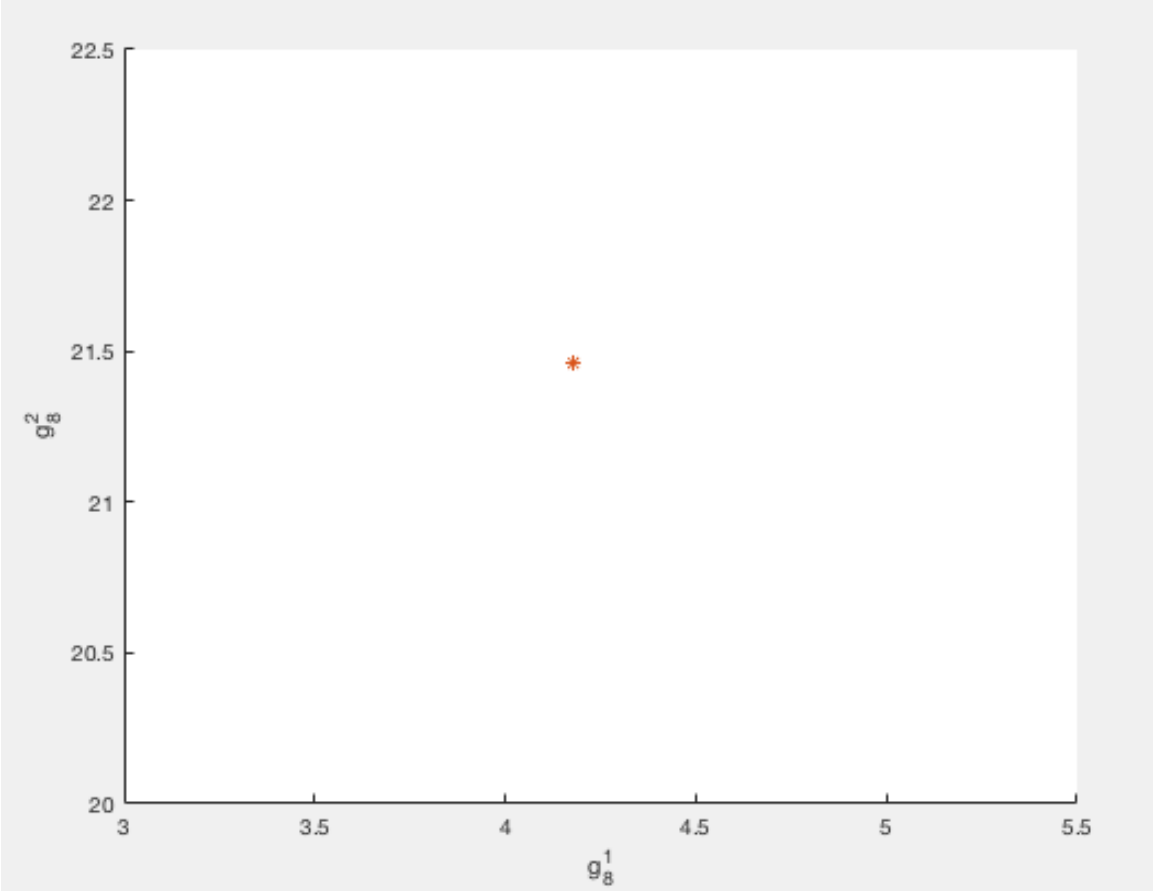

Figure 4.8: SOOWMS on the Problem Involving 8 Variables with a Large Range.

and the solution obtained from the CRH method with the median as an initial point Li *et al.* (2014). The figure demonstrates that both points are the same.

g) Seventh experiment involving ten variables with a small range:- In this experiment, there are ten variables. The problem solved in this experiment is as follows:

$$\begin{aligned} \underset{x_1,\ldots,x_{10},w_1,w_2,w_3}{\text{minimize}} \quad & f_{10}(x_1,\ldots,x_{10},w_1,w_2,w_3) \\ \text{subject to} \quad & e^{-w_1}+e^{-w_2}+e^{-w_3}=1, \\ & w_i \geq 0. \end{aligned} \tag{4.19}$$

In problem (4.19), $f_{10}(x_1,\ldots,x_{10},w_1,w_2,w_3)$ is defined as follows:

$$
\begin{aligned}
f_{10}(x_1,\ldots,x_{10},w_1,w_2,w_3) \quad &= \frac{1}{\sigma(1,2)} * [w_1 * (x_1-1)^2 + w_2 * (x_1-2)^2] \\
&+ \frac{1}{\sigma(4,4.4,5)} * [w_1 * (x_3-4)^2 + w_2 * (x_3-4.4)^2 \\
&+ w_3 * (x_3-5)^2] \\
&+ \frac{1}{\sigma(0.1,0.7)} * [w_1 * (x_5-0.1)^2 + w_3 * (x_5-0.7)^2] \\
&+ \frac{1}{\sigma(2,2.4)} * [w_1 * (x_7-2)^2 + w_2 * (x_7-2.4)^2] \\
&+ \frac{1}{\sigma(3,3.4,3.8)} * [w_1 * (x_9-3)^2 + w_2 * (x_9-3.4)^2 \\
&+ w_3 * (x_9-3.8)^2] \\
&+ \frac{1}{\sigma(6,6.6)} * [w_2 * (x_2-6)^2 + w_3 * (x_2-6.6)^2] \\
&+ \frac{1}{\sigma(7,7.6)} * [w_1 * (x_4-7)^2 + w_2 * (x_4-7.6)^2] \\
&+ \frac{1}{\sigma(8,8.8)} * [w_2 * (x_6-8)^2 + w_3 * (x_6-8.8)^2] \\
&+ \frac{1}{\sigma(9,9.8)} * [w_1 * (x_8-9)^2 + w_2 * (x_8-9.8)^2 \\
&+ \frac{1}{\sigma(5,5.6,5.8)} * [w_1 * (x_{10}-5)^2 + w_2 * (x_{10}-5.6)^2 \\
&+ w_3 * (x_{10}-5.8)^2].
\end{aligned}
\tag{4.20}
$$

ϵ is a minimum distance between two initial points along any dimension. The value of the ϵ is 0.2. The number of initial points, in this case, is calculated as follows:

$$
\begin{aligned}
&(\frac{2-1}{0.2}+1) * (\frac{6.6-6}{0.2}+1) * (\frac{5-4}{0.2}+1) * (\frac{7.6-7}{0.2}+1) \\
&* (\frac{0.7-0.1}{0.2}+1) * (\frac{8.8-8}{0.2}+1) * (\frac{2.4-2}{0.2}+1) * (\frac{9.8-9}{0.2}+1) \\
&* (\frac{3.8-3}{0.2}+1) * (\frac{5.8-5}{0.2}+1) = 4320000.
\end{aligned}
\tag{4.21}
$$

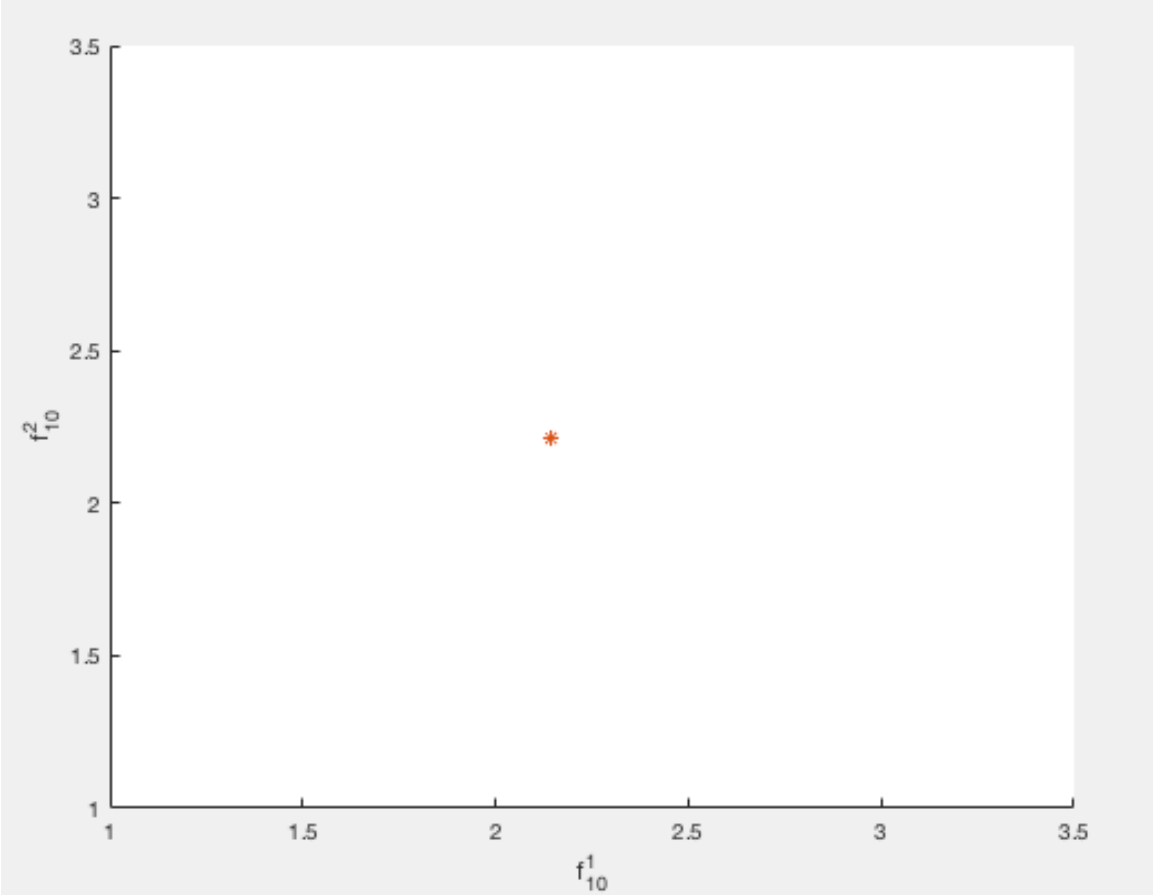

Figure 4.9: SOOWMS on the Problem Involving 10 Variables with a Small Range.

The number of initial points is 4320000. After applying the CRH method with 4320000 different initial points, only one optimal solution is found. The figure 4.9 shows the plot of f_{10}^1 and f_{10}^2. These functions f_{10}^1 and f_{10}^2 are defined in equations (4.44) and (4.45). In figure 4.9, the point shows the solution obtained from the SOOWMS and the solution obtained from the CRH method with the median as an initial point Li *et al.* (2014). Both points are the same in the figure.

h) Eighth experiment involving ten variables with a large range and a function change:- In this experiment, there are ten variables. The problem solved in this experiment is as follows:

$$\begin{aligned} \underset{x_1,\ldots,x_{10},w_1,w_2,w_3}{\text{minimize}} \quad & g_{10}(x_1,\ldots,x_{10},w_1,w_2,w_3) \\ \text{subject to} \quad & e^{-w_1}+e^{-w_2}+e^{-w_3}=1, \\ & w_i \geq 0. \end{aligned} \tag{4.22}$$

In problem (4.19), $g_{10}(x_1,\ldots,x_{10},w_1,w_2,w_3)$ is defined as follows:

$$\begin{aligned} g_{10}(x_1,\ldots,x_{10},w_1,w_2,w_3) \quad &= \frac{1}{\sigma(1,5)} * [w_1 * (x_1-1)^2 + w_2 * (x_1-5)^2] \\ &+ \frac{1}{\sigma(4,6,8)} * [w_1 * (x_3-4)^2 + w_2 * (x_3-6)^2 \\ &+ w_3 * (x_3-8)^2] \\ &+ \frac{1}{\sigma(6,9)} * [w_1 * (x_5-6)^2 + w_3 * (x_5-9)^2] \\ &+ \frac{1}{\sigma(2,4)} * [w_2 * (x_7-4)^2 + w_3 * (x_7-2)^2] \\ &+ \frac{1}{\sigma(3,5,7)} * [w_1 * (x_9-3)^2 + w_2 * (x_9-5)^2 \\ &+ w_3 * (x_9-7)^2] \\ &+ \frac{1}{\sigma(5,7,8)} * [w_1 * (x_2-5)^2 + w_2 * (x_2-7)^2 \\ &+ w_3 * (x_2-8)^2] \\ &+ \frac{1}{\sigma(7,10)} * [w_1 * (x_4-7)^2 + w_2 * (x_4-10)^2] \\ &+ \frac{1}{\sigma(8,10)} * [w_2 * (x_6-10)^2 + w_3 * (x_6-8)^2] \\ &+ \frac{1}{\sigma(10,12,13)} * [w_1 * (x_8-10)^2 + w_2 * (x_8-12)^2 \\ &+ w_3 * (x_8-13)^2] \\ &+ \frac{1}{\sigma(10,14)} * [w_1 * (x_{10}-10)^2 + w_3 * (x_{10}-14)^2]. \end{aligned} \tag{4.23}$$

ϵ is a minimum distance between two initial points along any dimension. The

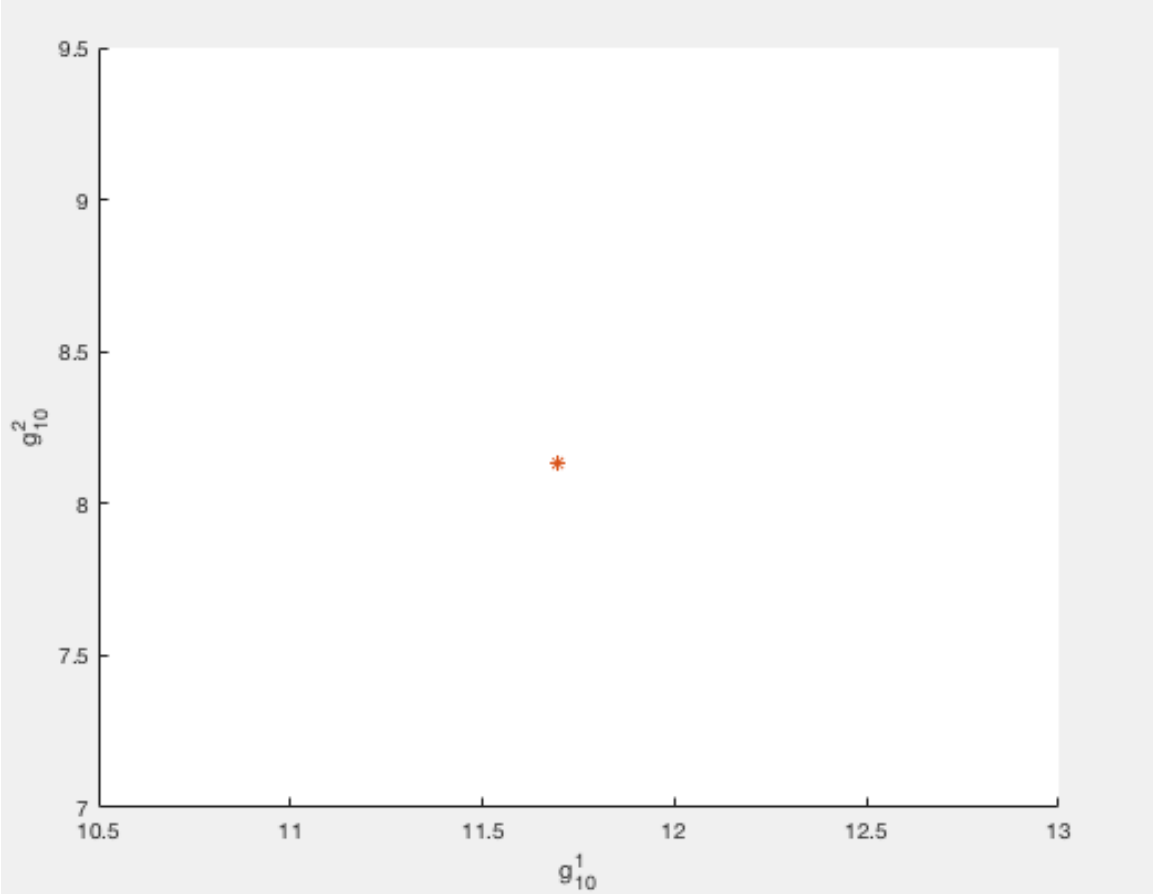

Figure 4.10: SOOWMS on the Problem Involving 10 Variables with a Large Range and a Function Change.

value of the ϵ is 1. The number of initial points, in this case, is calculated as follows:

$$\begin{aligned}&(\frac{5-1}{1}+1)*(\frac{8-5}{1}+1)*(\frac{8-4}{1}+1)*(\frac{10-7}{1}+1)\\&*(\frac{9-6}{1}+1)*(\frac{10-8}{1}+1)*(\frac{4-2}{1}+1)*(\frac{13-10}{1}+1)\\&*(\frac{7-5}{1}+1)*(\frac{14-10}{1}+1)=1440000.\end{aligned} \tag{4.24}$$

The number of initial points is 1440000. After applying the CRH method with 1440000 different initial points, only one optimal solution is found. The figure 4.10 shows the plot of g_{10}^1 and g_{10}^2. These functions g_{10}^1 and g_{10}^2 are defined in equations (4.47) and (4.48). In figure 4.10, the point shows the solution obtained from the

SOOWMS and the solution obtained from the CRH method with the median as an initial point Li *et al.* (2014). In this experiment, both points are the same.

The summary of the above experiments is as follows:

Table 4.1: Statistics of SOOWMS.

Function	**Number of Initial Points**	ϵ	**Number of Solutions**
f_2	2005401	ϵ_1=0.001 ϵ_2=0.01	1
f_4	2657661	ϵ=0.1	1
f_8	8363520	ϵ=0.1	1
f_{10}	4320000	ϵ=0.2	1
g_2	321201	ϵ=0.5	1
g_4	1334550	ϵ=1	1
g_8	675840	ϵ=1	1
g_{10}	1440000	ϵ=1	1

4.2.2 Non-dominated Sorting Genetic Algorithm II (NSGA)

a) First experiment involving two variables with a small range:- In this experiment, there are two variables. The problem (4.1) is transformed into the multi-objectiive optimization problem as follows:

$$\begin{aligned} \underset{x_1,x_2,w_1,w_2,w_3}{\text{minimize}} \quad & (f_2^1(x_1,x_2,w_1,w_2,w_3), f_2^2(x_1,x_2,w_1,w_2,w_3)) \\ \text{subject to} \quad & e^{-w_1}+e^{-w_2}+e^{-w_3}=1, \\ & w_i \geq 0. \end{aligned} \tag{4.25}$$

In problem (4.25), $f_2^1(x_1,x_2,w_1,w_2,w_3)$ and $f_2^2(x_1,x_2,w_1,w_2,w_3)$ are defined as follows:

$$\begin{aligned} f_2^1(x_1,x_2,w_1,w_2,w_3) \quad &= \frac{1}{\sigma(1,3,6)} * [w_1 * (x_1-3)^2 + w_2 * (x_1-1)^2 \\ &\quad + w_3 * (x_1-6)^2]. \end{aligned} \tag{4.26}$$

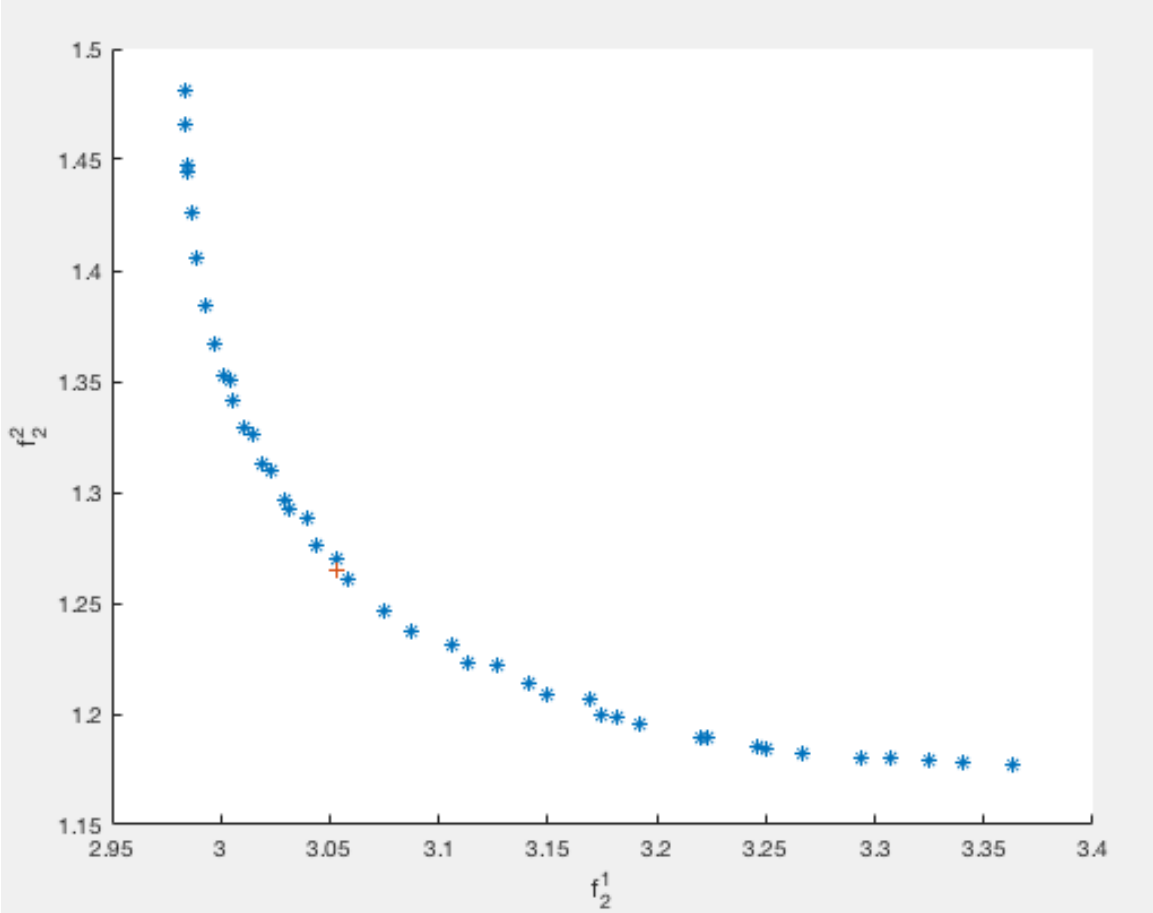

Figure 4.11: NSGA on the Problem Involving 2 Variables Having a Small Range.

$$f_2^2(x_1, x_2, w_1, w_2, w_3) = \frac{1}{\sigma(1,4,5)} * [w_1 * (x_2 - 4)^2 + w_2 * (x_2 - 5)^2 + w_3 * (x_2 - 1)^2]. \tag{4.27}$$

After the NSGA is run with a population size of 50 for solving the problem (4.25), the Pareto optimal points are computed. In figure 4.11, the blue points denote the Pareto optimal points and the red point denotes the solution obtained by the CRH having the median as an initial point. As demonstrated by the figure, the solution obtained by the CRH lies in the Pareto front of the NSGA. The solution, coming

from the CRH with the median as an initial point, is optimal in this experiment.

b) Second experiment involving two variables with a large range:- In this experiment, there are two variables. The problem (4.4) is transformed into the multi-objectiive optimization problem as follows:

$$\begin{aligned} \underset{x_1,x_2,w_1,w_2,w_3}{\text{minimize}} \quad & (g_2^1(x_1,x_2,w_1,w_2,w_3), g_2^2(x_1,x_2,w_1,w_2,w_3)) \\ \text{subject to} \quad & e^{-w_1}+e^{-w_2}+e^{-w_3}=1, \\ & w_i \geq 0. \end{aligned} \tag{4.28}$$

In problem (4.28), $g_2^1(x_1,x_2,w_1,w_2,w_3)$ and $g_2^2(x_1,x_2,w_1,w_2,w_3)$ are defined as follows:

$$\begin{aligned} g_2^1(x_1,x_2,w_1,w_2,w_3) &= \frac{1}{\sigma(100,300,500)} * [w_1*(x_1-300)^2 + w_2*(x_1-100)^2 \\ &\quad + w_3*(x_1-500)^2]. \end{aligned} \tag{4.29}$$

$$\begin{aligned} g_2^2(x_1,x_2,w_1,w_2,w_3) &= \frac{1}{\sigma(400,500,600)} * [w_1*(x_2-400)^2 + w_2*(x_2-600)^2 \\ &\quad + w_3*(x_2-500)^2]. \end{aligned} \tag{4.30}$$

After the NSGA is run with a population size of 50 for solving the problem (4.28), the Pareto optimal points are obtained. In figure 4.12, the blue points represent the Pareto front and the red point denotes the solution obtained by the CRH having the median as an initial point. As evident from the figure, the solution obtained by the CRH lies on the Pareto frontier of the NSGA. The solution, getting from the CRH

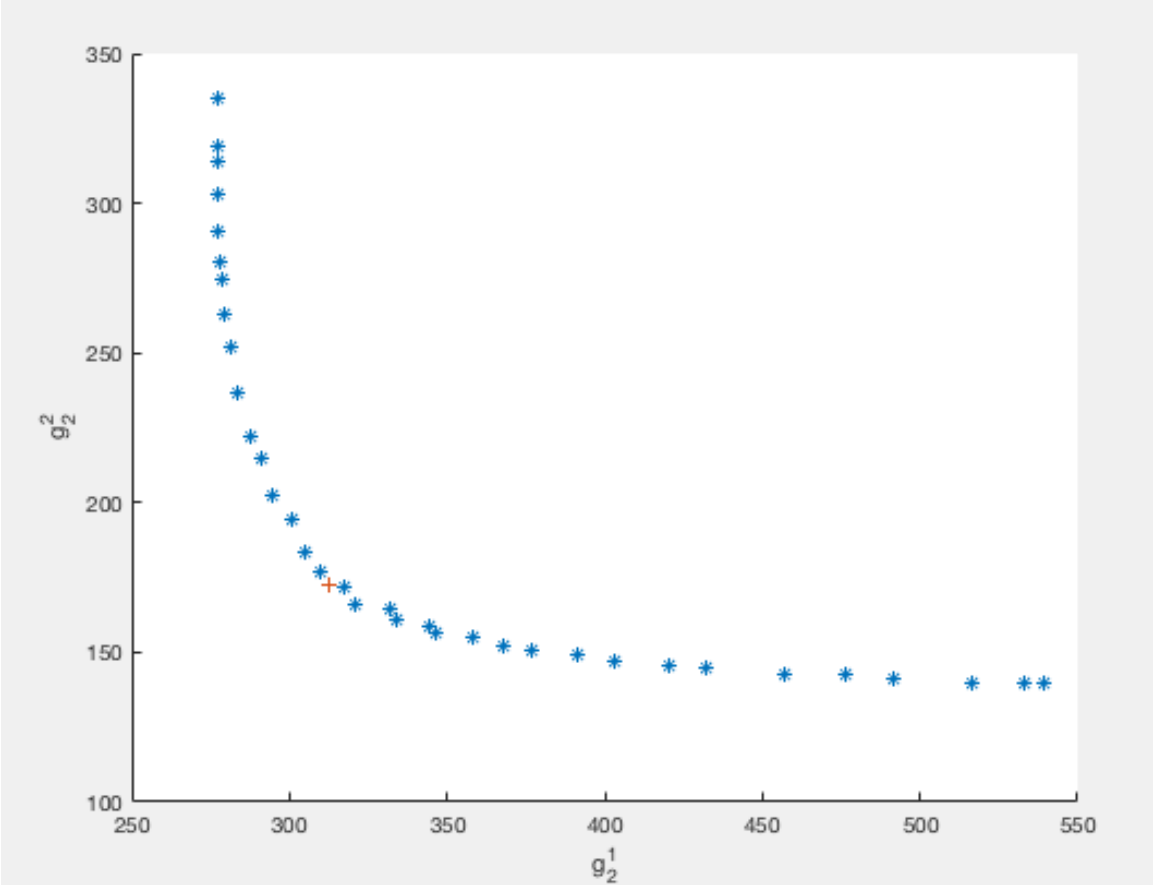

Figure 4.12: NSGA on the Problem Involving 2 Variables Having a Large Range.

with the median as an initial point, is optimal in this test.

c) Third experiment involving four variables with a small range:- In this experiment, there are four variables. The problem (4.7) is transformed into the multi-objectiive optimization problem as follows:

$$\begin{aligned} \underset{x_1,x_2,w_1,w_2,w_3}{\text{minimize}} \quad & (f_4^1(x_1,\ldots,x_4,w_1,w_2,w_3), f_4^2(x_1,\ldots,x_4,w_1,w_2,w_3)) \\ \text{subject to} \quad & e^{-w_1}+e^{-w_2}+e^{-w_3}=1, \\ & w_i \geq 0. \end{aligned} \tag{4.31}$$

In problem (4.32), $f_4^1(x_1, \ldots, x_4, w_1, w_2, w_3)$ and $f_4^2(x_1, \ldots, x_4, w_1, w_2, w_3)$ are defined as follows:

$$\begin{aligned} f_4^1(x_1, \ldots, x_4, w_1, w_2, w_3) &= \frac{1}{\sigma(1,4,6)} * [w_1 * (x_1 - 1)^2 + w_2 * (x_1 - 4)^2 \\ &+ w_3 * (x_1 - 6)^2] \\ &+ \frac{1}{\sigma(1,2,4)} * [w_1 * (x_3 - 2)^2 + w_2 * (x_3 - 4)^2 \\ &+ w_3 * (x_3 - 1)^2]. \end{aligned} \tag{4.32}$$

$$\begin{aligned} f_4^2(x_1, \ldots, x_4, w_1, w_2, w_3) &= \frac{1}{\sigma(3,5,7)} * [w_1 * (x_2 - 3)^2 + w_2 * (x_2 - 5)^2 \\ &+ w_3 * (x_2 - 7)^2] \\ &+ \frac{1}{\sigma(1,5)} * [w_1 * (x_4 - 1)^2 + w_2 * (x_4 - 5)^2]. \end{aligned} \tag{4.33}$$

After the NSGA is run with a population size of 50 for solving the problem (4.31), the Pareto optimal points are computed. In figure 4.13, the blue points denote the Pareto optimal points and the red point denotes the solution obtained by the CRH having the median as an initial point. As demonstrated by the figure, the solution obtained by the CRH lies on the Pareto frontier of the NSGA. The solution, getting from the CRH with the median as an initial point, is optimal in this test.

d) Fourth experiment involving four variables with a large range and a function change:- In this experiment, there are four variables. The problem (4.10) is transformed into the multi-objectiive optimization problem as follows:

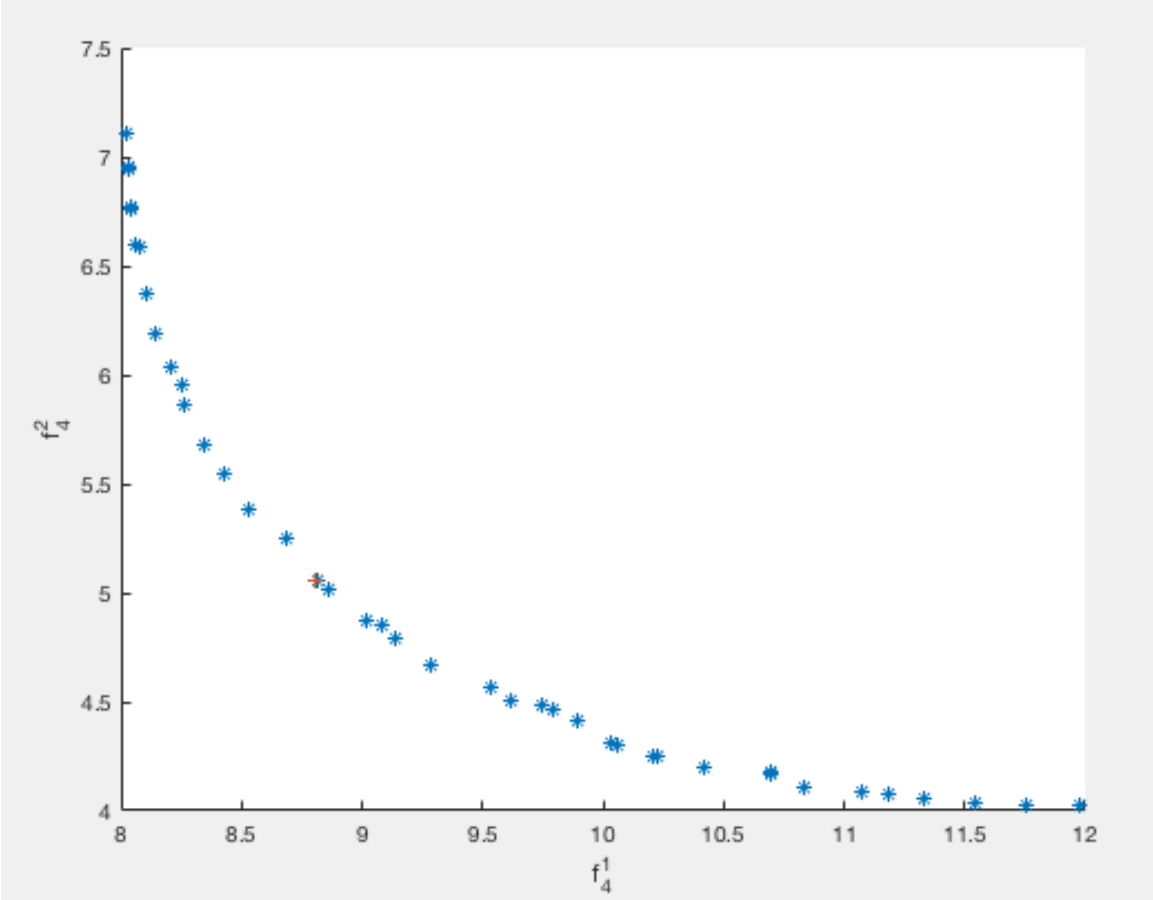

Figure 4.13: NSGA on the Problem Involving 4 Variables Having a Small Range.

$$\begin{aligned} \underset{x_1,x_2,w_1,w_2,w_3}{\text{minimize}} \quad & (g_4^1(x_1,\ldots,x_4,w_1,w_2,w_3), g_4^2(x_1,\ldots,x_4,w_1,w_2,w_3)) \\ \text{subject to} \quad & e^{-w_1}+e^{-w_2}+e^{-w_3}=1, \\ & w_i \geq 0. \end{aligned} \tag{4.34}$$

In problem (4.34), $g_4^1(x_1,\ldots,x_4,w_1,w_2,w_3)$ and $g_4^2(x_1,\ldots,x_4,w_1,w_2,w_3)$ are defined as follows:

$$
\begin{aligned}
g_4^1(x_1,\ldots,x_4,w_1,w_2,w_3) &= \frac{1}{\sigma(1,46,50)} * [w_1 * (x_1-1)^2 + w_2 * (x_1-46)^2 \\
&\quad + w_3 * (x_1-50)^2] \\
&\quad + \frac{1}{\sigma(2,15,22)} * [w_1 * (x_3-2)^2 + w_2 * (x_3-15)^2 \\
&\quad + w_3 * (x_3-22)^2].
\end{aligned} \tag{4.35}
$$

$$
\begin{aligned}
g_4^2(x_1,\ldots,x_4,w_1,w_2,w_3) &= \frac{1}{\sigma(3,23,33)} * [w_1 * (x_2-3)^2 + w_2 * (x_2-23)^2 \\
&\quad + w_3 * (x_2-33)^2] \\
&\quad + \frac{1}{\sigma(4,18,44)} * [w_1 * (x_4-4)^2 + w_2 * (x_4-18)^2 \\
&\quad + w_3 * (x_4-44)^2].
\end{aligned} \tag{4.36}
$$

After the NSGA is run with a population size of 50 for solving the problem (4.34), the Pareto optimal points are computed. In figure 4.14, the blue points are the Pareto optimal points and the red point is the solution obtained by the CRH having the median as an initial point. As apparent from the figure, the solution obtained by the CRH lies on the Pareto frontier of the NSGA. The solution, coming from the CRH with the median as an initial point, is optimal in this experiment.

e) Fifth experiment involving eight variables with a small range:- In this experiment, there are eight variables. The problem (4.13) is transformed into the multi-objectiive optimization problem as follows:

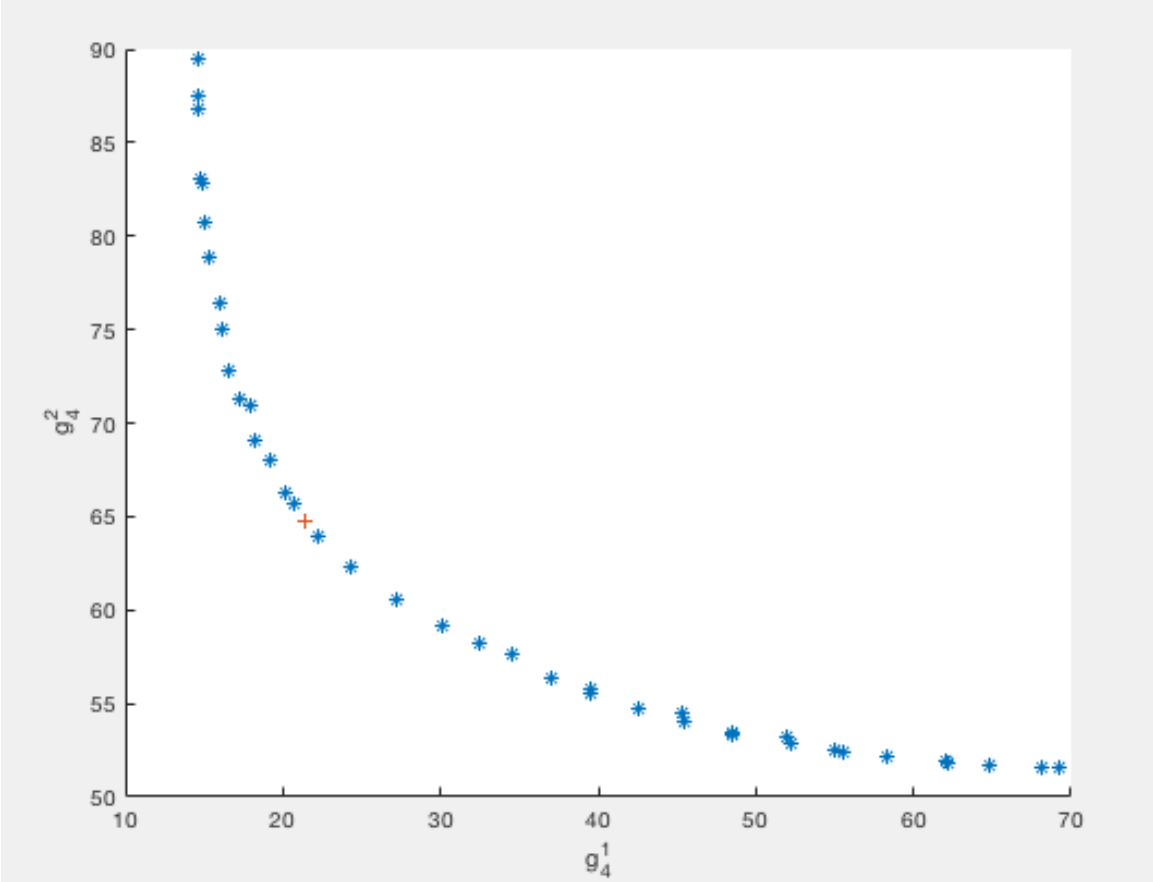

Figure 4.14: NSGA on the Problem Involving 4 Variables Having a Large Range.

$$\begin{aligned} \underset{x_1,\ldots,x_8,w_1,w_2,w_3}{\text{minimize}} \quad & (f_8^1(x_1,\ldots,x_8,w_1,w_2,w_3), f_8^2(x_1,\ldots,x_8,w_1,w_2,w_3)) \\ \text{subject to} \quad & e^{-w_1} + e^{-w_2} + e^{-w_3} = 1, \\ & w_i \geq 0. \end{aligned} \tag{4.37}$$

In problem (4.37), $f_8^1(x_1,\ldots,x_8,w_1,w_2,w_3)$ and $f_8^2(x_1,\ldots,x_8,w_1,w_2,w_3)$ are defined as follows:

$$
\begin{aligned}
f_8^1(x_1,\ldots,x_8,w_1,w_2,w_3) \quad &= \frac{1}{\sigma(1,2)} * [w_1 * (x_1-1)^2 + w_2 * (x_1-2)^2] \\
&+ \frac{1}{\sigma(4,4.3,5)} * [w_1 * (x_3-4)^2 + w_2 * (x_3-4.3)^2 \\
&+ w_3 * (x_3-5)^2] \\
&+ \frac{1}{\sigma(0.1,0.5)} * [w_1 * (x_5-0.1)^2 + w_3 * (x_5-0.5)^2] \\
&+ \frac{1}{\sigma(6,6.5)} * [w_1 * (x_7-6)^2 + w_2 * (x_7-6.5)^2].
\end{aligned} \tag{4.38}
$$

$$
\begin{aligned}
f_8^2(x_1,\ldots,x_8,w_1,w_2,w_3) \quad &= \frac{1}{\sigma(6,6.5)} * [w_2 * (x_2-6)^2 + w_3 * (x_2-6.5)^2] \\
&+ \frac{1}{\sigma(7,7.7)} * [w_1 * (x_4-7)^2 + w_2 * (x_4-7.7)^2] \\
&+ \frac{1}{\sigma(8,8.5)} * [w_1 * (x_6-8.5)^2 + w_3 * (x_6-8)^2] \\
&+ \frac{1}{\sigma(9,9.5,9.7)} * [w_1 * (x_8-9)^2 + w_2 * (x_8-9.5)^2 \\
&+ w_3 * (x_8-9.7)^2].
\end{aligned} \tag{4.39}
$$

After the NSGA is run with a population size of 50 for solving the problem (4.37), the Pareto optimal points are computed. In figure 4.15, the blue points represent the Pareto frontier and the red point denotes the solution obtained by the CRH having the median as an initial point. As evident from the figure, the solution obtained by the CRH lies on the Pareto frontier of the NSGA. The solution, coming from the CRH with the median as an initial point, is optimal in this experiment.

f) Sixth experiment involving eight variables with a large range:- In this experiment, there are eight variables. The problem (4.16) is transformed into the

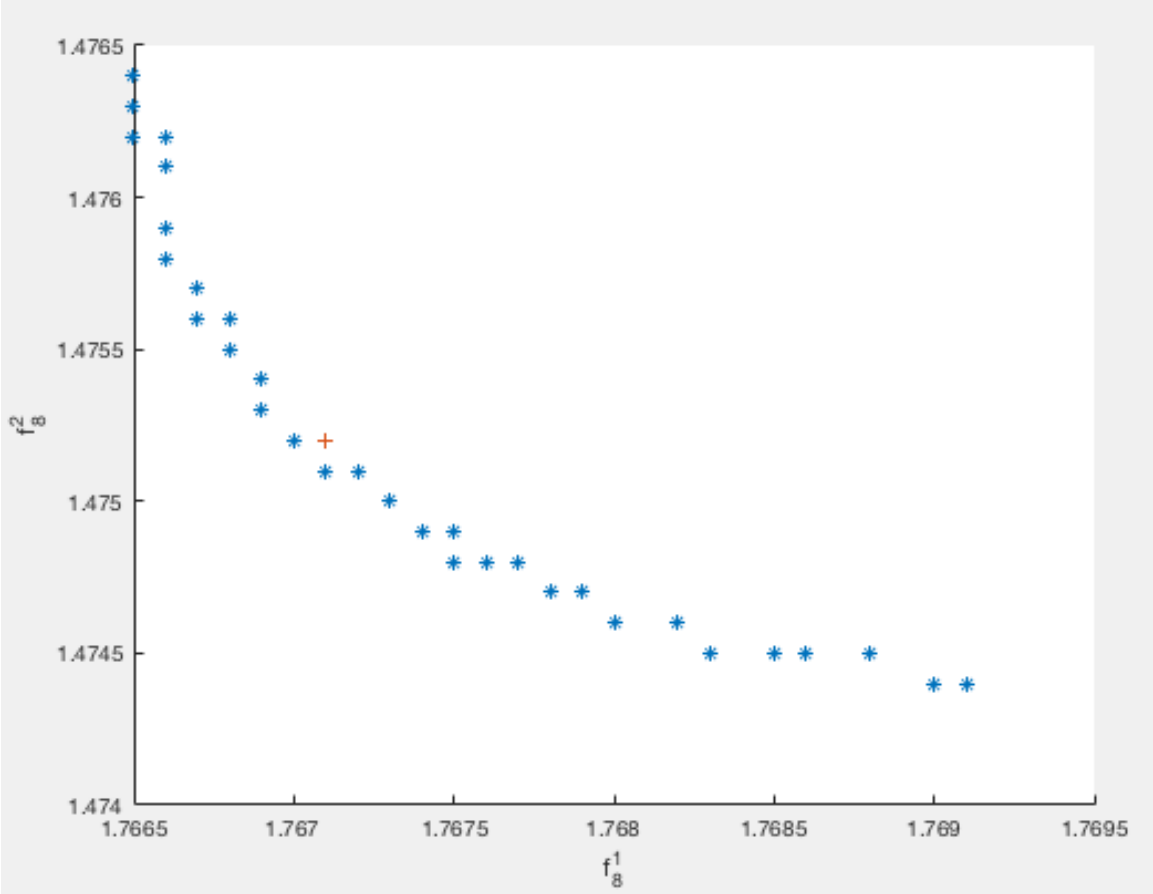

Figure 4.15: NSGA on the Problem Involving 8 Variables with a Small Range.

multi-objectiive optimization problem as follows:

$$\begin{aligned} \underset{x_1,\ldots,x_8,w_1,w_2,w_3}{\text{minimize}} \quad & (g_8^1(x_1,\ldots,x_8,w_1,w_2,w_3), g_8^2(x_1,\ldots,x_8,w_1,w_2,w_3)) \\ \text{subject to} \quad & e^{-w_1}+e^{-w_2}+e^{-w_3}=1, \\ & w_i \geq 0. \end{aligned} \tag{4.40}$$

In problem (4.40), $g_8^1(x_1,\ldots,x_8,w_1,w_2,w_3)$ and $g_8^2(x_1,\ldots,x_8,w_1,w_2,w_3)$ are defined as follows:

$$
\begin{aligned}
g_8^1(x_1,\ldots,x_8,w_1,w_2,w_3) \quad &= \frac{1}{\sigma(1,4)} * [w_1 * (x_1-1)^2 + w_2 * (x_1-4)^2] \\
&+ \frac{1}{\sigma(4,6,7)} * [w_1 * (x_3-4)^2 + w_2 * (x_3-7)^2 \\
&+ w_3 * (x_3-6)^2] \\
&+ \frac{1}{\sigma(1,4)} * [w_1 * (x_5-1)^2 + w_3 * (x_5-4)^2] \\
&+ \frac{1}{\sigma(6,10)} * [w_1 * (x_7-6)^2 + w_2 * (x_7-10)^2].
\end{aligned} \tag{4.41}
$$

$$
\begin{aligned}
g_8^2(x_1,\ldots,x_8,w_1,w_2,w_3) \quad &= \frac{1}{\sigma(5,10)} * [w_2 * (x_2-5)^2 + w_3 * (x_2-10)^2] \\
&+ \frac{1}{\sigma(7,10)} * [w_1 * (x_4-7)^2 + w_2 * (x_4-10)^2] \\
&+ \frac{1}{\sigma(13,20)} * [w_1 * (x_6-20)^2 + w_3 * (x_6-13)^2] \\
&+ \frac{1}{\sigma(10,14,20)} * [w_1 * (x_8-10)^2 + w_2 * (x_8-14)^2 \\
&+ w_3 * (x_8-20)^2].
\end{aligned} \tag{4.42}
$$

After the NSGA is run with a population size of 50 for solving the problem (4.40), the Pareto optimal points are obtained. In figure 4.16, the blue points denote the Pareto optimal points and the red point denotes the solution obtained by the CRH having the median as an initial point. As demonstrated by the figure, the solution obtained by the CRH lies on the Pareto frontier of the NSGA. The solution, getting from the CRH with the median as an initial point, is optimal in this test.

g) Seventh experiment involving ten variables with a small range:- In this experiment, there are ten variables. The problem (4.19) is transformed into the

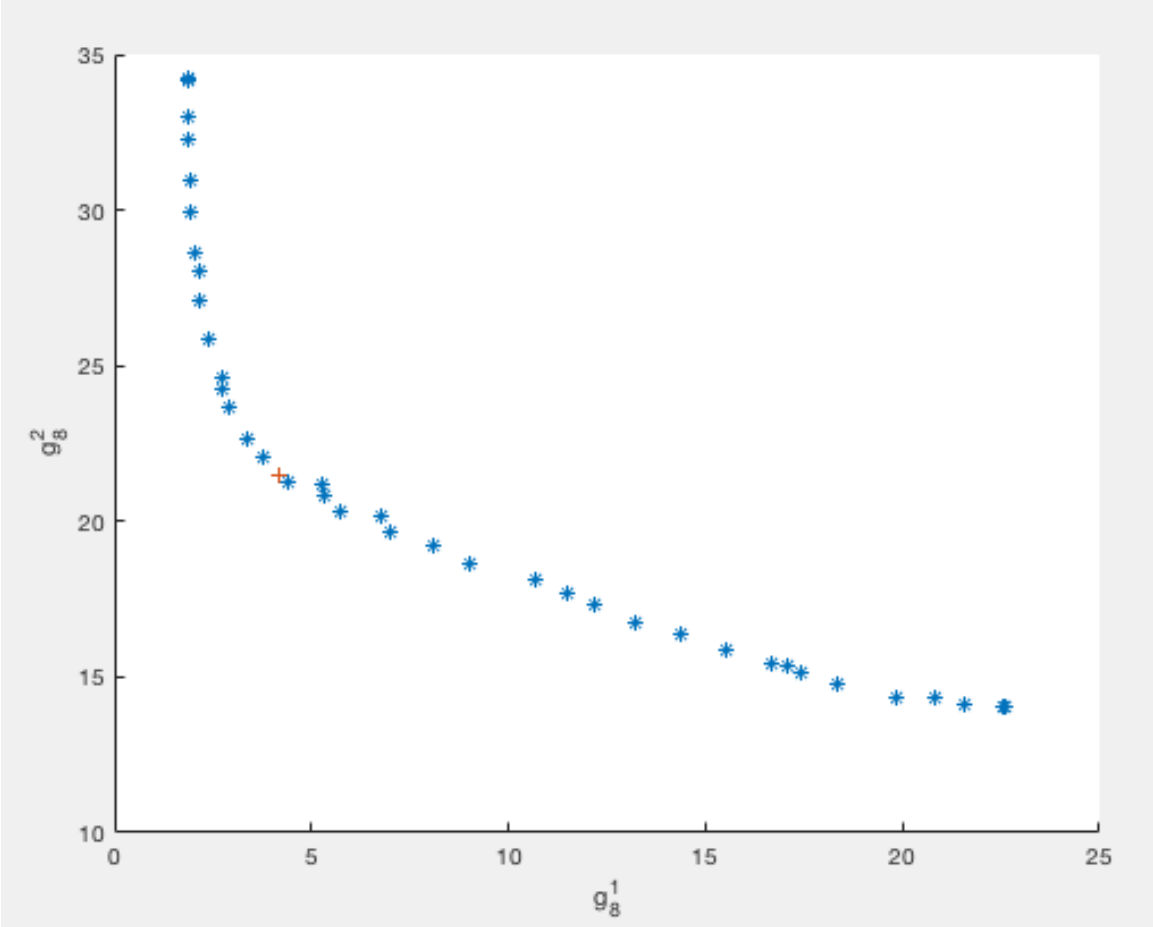

Figure 4.16: NSGA on the Problem Involving 8 Variables Having a Large Range.

multi-objectiive optimization problem as follows:

$$\begin{aligned} \underset{x_1,\ldots,x_{10},w_1,w_2,w_3}{\text{minimize}} \quad & (f_{10}^1(x_1,\ldots,x_{10},w_1,w_2,w_3), f_{10}^2(x_1,\ldots,x_{10},w_1,w_2,w_3)) \\ \text{subject to} \quad & e^{-w_1}+e^{-w_2}+e^{-w_3}=1, \\ & w_i \geq 0. \end{aligned} \tag{4.43}$$

In problem (4.43), $f_{10}^1(x_1,\ldots,x_{10},w_1,w_2,w_3)$ and $f_{10}^2(x_1,\ldots,x_{10},w_1,w_2,w_3)$ are defined as follows:

$$
\begin{aligned}
f_{10}^1(x_1,\ldots,x_{10},w_1,w_2,w_3) \quad &= \frac{1}{\sigma(1,2)} * [w_1 * (x_1-1)^2 + w_2 * (x_1-2)^2] \\
&+ \frac{1}{\sigma(4,4.4,5)} * [w_1 * (x_3-4)^2 + w_2 * (x_3-4.4)^2 \\
&+ w_3 * (x_3-5)^2] \\
&+ \frac{1}{\sigma(0.1,0.7)} * [w_1 * (x_5-0.1)^2 + w_3 * (x_5-0.7)^2] \\
&+ \frac{1}{\sigma(2,2.4)} * [w_1 * (x_7-2)^2 + w_2 * (x_7-2.4)^2] \\
&+ \frac{1}{\sigma(3,3.4,3.8)} * [w_1 * (x_9-3)^2 + w_2 * (x_9-3.4)^2 \\
&+ w_3 * (x_9-3.8)^2].
\end{aligned} \tag{4.44}
$$

$$
\begin{aligned}
f_{10}^2(x_1,\ldots,x_{10},w_1,w_2,w_3) \quad &= \frac{1}{\sigma(6,6.6)} * [w_2 * (x_2-6)^2 + w_3 * (x_2-6.6)^2] \\
&+ \frac{1}{\sigma(7,7.6)} * [w_1 * (x_4-7)^2 + w_2 * (x_4-7.6)^2] \\
&+ \frac{1}{\sigma(8,8.8)} * [w_2 * (x_6-8)^2 + w_3 * (x_6-8.8)^2] \\
&+ \frac{1}{\sigma(9,9.8)} * [w_1 * (x_8-9)^2 + w_2 * (x_8-9.8)^2] \\
&+ \frac{1}{\sigma(5,5.6,5.8)} * [w_1 * (x_{10}-5)^2 + w_2 * (x_{10}-5.6)^2 \\
&+ w_3 * (x_{10}-5.8)^2].
\end{aligned} \tag{4.45}
$$

After the NSGA is run with a population size of 50 for solving the problem (4.43), the Pareto optimal points are computed. In figure 4.17, the blue points are the Pareto optimal points and the red point is the solution obtained by the CRH having the median as an initial point. As apparent from the figure, the solution obtained by the CRH lies on the Pareto frontier of the NSGA. The solution, coming from the CRH

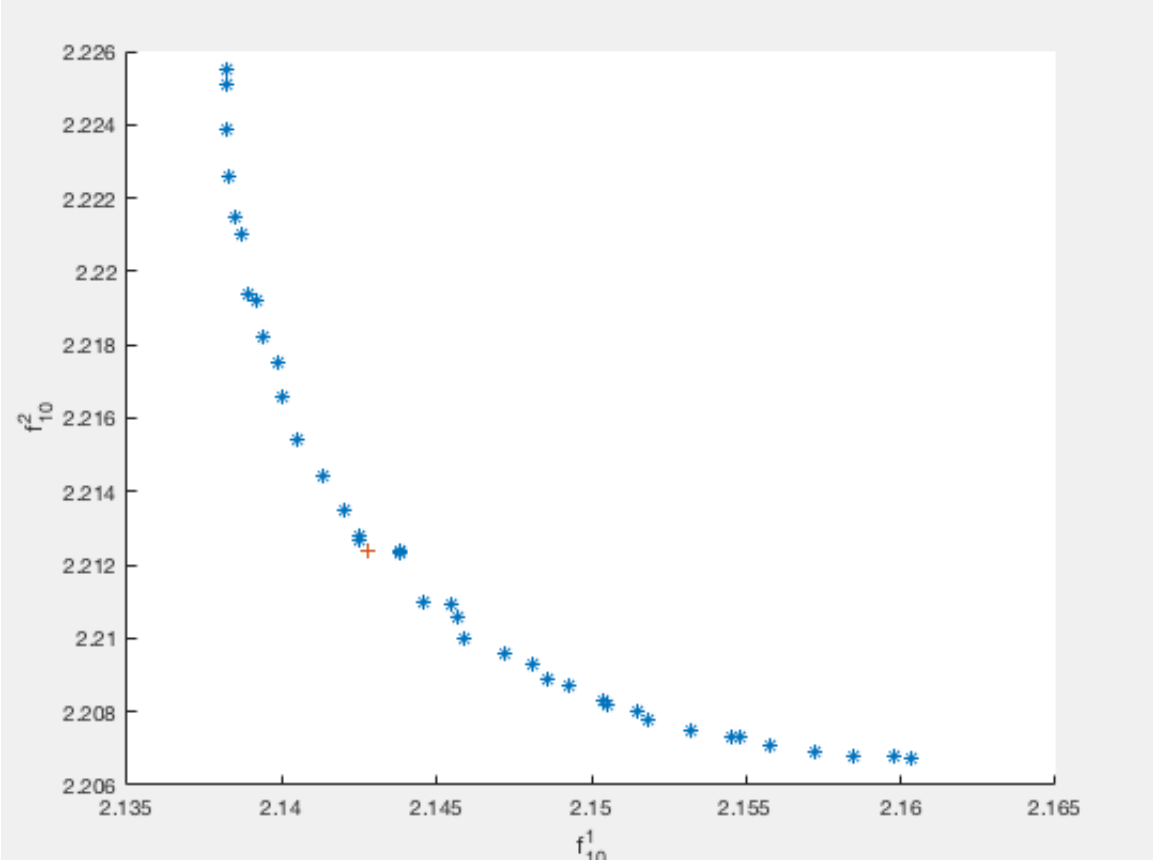

Figure 4.17: NSGA on the Problem Involving 10 Variables Having a Small Range

with the median as an initial point, is optimal in this experiment.

h) Eighth experiment involving ten variables with a large range and a function change:- In this experiment, there are ten variables. The problem (4.22) is transformed into the multi-objectiive optimization problem as follows:

$$\begin{aligned} \underset{x_1,\ldots,x_{10},w_1,w_2,w_3}{\text{minimize}} \quad & (g_{10}^1(x_1,\ldots,x_{10},w_1,w_2,w_3), g_{10}^2(x_1,\ldots,x_{10},w_1,w_2,w_3)) \\ \text{subject to} \quad & e^{-w_1}+e^{-w_2}+e^{-w_3}=1, \\ & w_i \geq 0. \end{aligned} \tag{4.46}$$

In problem (4.46), $g_{10}^1(x_1,\ldots,x_{10},w_1,w_2,w_3)$ and $g_{10}^2(x_1,\ldots,x_{10},w_1,w_2,w_3)$ are defined as follows:

$$\begin{aligned} g_{10}^1(x_1,\ldots,x_{10},w_1,w_2,w_3) &= \frac{1}{\sigma(1,5)} * [w_1 * (x_1-1)^2 + w_2 * (x_1-5)^2] \\ &+ \frac{1}{\sigma(4,6,8)} * [w_1 * (x_3-4)^2 + w_2 * (x_3-6)^2 \\ &+ w_3 * (x_3-8)^2] \\ &+ \frac{1}{\sigma(6,9)} * [w_1 * (x_5-6)^2 + w_3 * (x_5-9)^2] \\ &+ \frac{1}{\sigma(2,4)} * [w_2 * (x_7-4)^2 + w_3 * (x_7-2)^2] \\ &+ \frac{1}{\sigma(3,5,7)} * [w_1 * (x_9-3)^2 + w_2 * (x_9-5)^2 \\ &+ w_3 * (x_9-7)^2]. \end{aligned} \tag{4.47}$$

$$\begin{aligned} g_{10}^2(x_1,\ldots,x_{10},w_1,w_2,w_3) &= \frac{1}{\sigma(5,7,8)} * [w_1 * (x_2-5)^2 + w_2 * (x_2-7)^2 \\ &+ w_3 * (x_2-8)^2] \\ &+ \frac{1}{\sigma(7,10)} * [w_1 * (x_4-7)^2 + w_2 * (x_4-10)^2] \\ &+ \frac{1}{\sigma(8,10)} * [w_2 * (x_6-10)^2 + w_3 * (x_6-8)^2] \\ &+ \frac{1}{\sigma(10,12,13)} * [w_1 * (x_8-10)^2 + w_2 * (x_8-12)^2 \\ &+ w_3 * (x_8-13)^2] \\ &+ \frac{1}{\sigma(10,14)} * [w_1 * (x_{10}-10)^2 + w_3 * (x_{10}-14)^2]. \end{aligned} \tag{4.48}$$

After the NSGA is run with a population size of 50 for solving the problem (4.46), the Pareto optimal points are computed. In figure 4.18, the blue points represent the

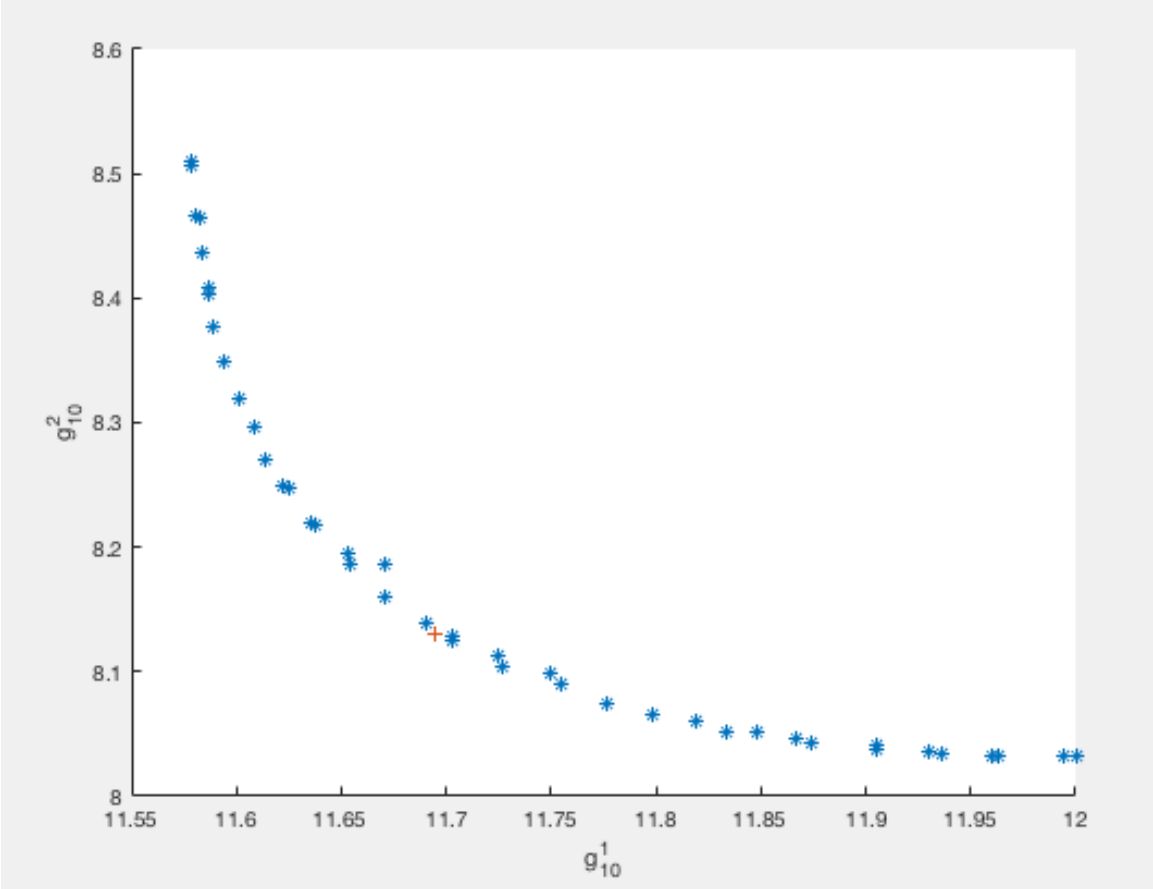

Figure 4.18: NSGA on the Problem Involving 10 Variables Having a Large Range and a Function Change.

Pareto frontier and the red point denotes the solution obtained by the CRH having the median as an initial point. As evident from the figure, the solution obtained by the CRH lies on the Pareto frontier of the NSGA. The solution, getting from the CRH with the median as an initial point, is optimal in this test.

4.3 Large Dataset

a) First experiment involving 14 homogeneous variables:- In this experiment, it contains data corresponding to two properties. There are 14 variables in

which all of them are of continuous type. The loss function used corresponding to continuous variables is the normalized squared loss. The problem solved in this experiment, of form (3.1), is below:

$$\begin{aligned} \underset{X^*,W}{\text{minimize}} \quad & j_{14}(X^*, W) = \sum_{k=1}^{9} w_k \sum_{i=1}^{7} \sum_{m=1}^{2} d_m(x_{im}^*, x_{im}^k) \\ \text{subject to} \quad & \xi(W) = \sum_{k=1}^{9} e^{-w_k} = 1, \\ & W \geq 0. \end{aligned} \tag{4.49}$$

In problem (4.49), W is a weight vector having 9 elements. x_{im}^k is a value of i^{th} object corresponding to m^{th} property given by k^{th} source. The number of data objects is 7. The number of different properties is 2. $d_1(*,*)$ is the normalized squared loss and $d_2(*,*)$ is the normalized squared loss. $\xi(W)$ is a regularization function whose value is equal to 1. W corresponds to the reliability of the sources. The values of W are constrained by the regularization function. In problem (4.49), X^* is defined as follows:

$$X^* = \begin{bmatrix} x_{11}^* & x_{12}^* \\ x_{21}^* & x_{22}^* \\ x_{31}^* & x_{32}^* \\ x_{41}^* & x_{42}^* \\ x_{51}^* & x_{52}^* \\ x_{61}^* & x_{62}^* \\ x_{71}^* & x_{72}^* \end{bmatrix}. \tag{4.50}$$

In equation (4.50), x_{ij}^* is truth value of i^{th} object corresponding to j^{th} property.

In problem (4.49), W is defined as follows:

$$W = \begin{bmatrix} w_1 & w_2 & w_3 & w_4 & w_5 & w_6 & w_7 & w_8 & w_9 \end{bmatrix}. \tag{4.51}$$

w_i is a weight value corresponding to i^{th} source.

The problem (4.49) is transformed into the multi-objective optimization problem as follows:

$$\begin{aligned} \underset{X^*,W}{\text{minimize}} \quad & (j_{14}^1(X^*,W), j_{14}^2(X^*,W)) \\ \text{subject to} \quad & \xi(W) = \sum_{k=1}^{9} e^{-w_k} = 1, \\ & W \geq 0. \end{aligned} \tag{4.52}$$

In problem (4.52), $j_{14}^1(X^*,W)$ and $j_{14}^2(X^*,W)$ are defined as follows:

$$j_{14}^1(X^*,W) = \sum_{k=1}^{9} w_k \sum_{i=1}^{7} d_1(x_{i1}^*, x_{i1}^k). \tag{4.53}$$

$$j_{14}^2(X^*,W) = \sum_{k=1}^{9} w_k \sum_{i=1}^{7} d_2(x_{i2}^*, x_{i2}^k). \tag{4.54}$$

In problem (4.52), the definition of the variables is the same as in problem (4.49). In the dataset, if the value of a property of an object is not present, then the value of the property of the object is ignored in problem (4.52). It is applied to all experiments.

The NSGA is run for solving the multi-objective problem (4.52) and the CRH is run for solving the single-objective problem (4.49) on same data set. In the NSGA, the population size is 100. The figure 4.19 shows the Pareto optimal points of the NSGA and the solution from the CRH. In figure 4.19, j_{14}^1 is the function value corresponding

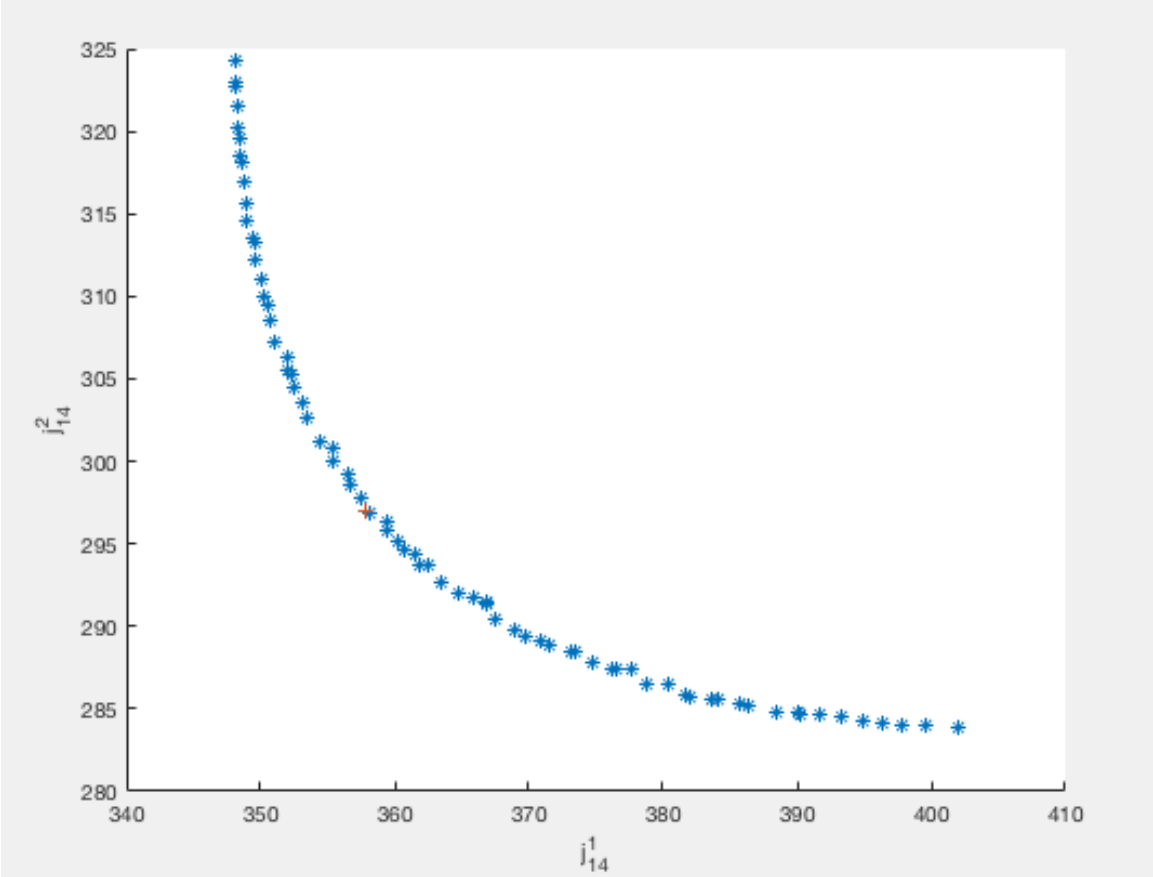

Figure 4.19: NSGA on the Problem Having 14 Homogeneous Variables.

to seven continuous variables and j_{14}^2 is the function value corresponding to another seven continuous variables. The blue points represent the Pareto optimal points and the red point represents the solution coming from the CRH. As evident from the figure, the solution, coming from CRH, lies on the Pareto frontier of the NSGA. The solution, getting from the CRH with the median as an initial point, is optimal in this test.

b) Second experiment involving 28 homogeneous variables:- In this experiment, it contains data corresponding to two properties. There are 28 variables in which all of them are of continuous type. The loss function used corresponding to continuous variables is the normalized squared loss. The problem solved in this

experiment, of form (3.1), is below:

$$\begin{aligned} \underset{X^*,W}{\text{minimize}} \quad & j_{28}(X^*, W) = \sum_{k=1}^{9} w_k \sum_{i=1}^{14} \sum_{m=1}^{2} d_m(x_{im}^*, x_{im}^k) \\ \text{subject to} \quad & \xi(W) = \sum_{k=1}^{9} e^{-w_k} = 1, \\ & W \geq 0. \end{aligned} \tag{4.55}$$

In problem (4.55), W is a weight vector having 9 elements. x_{im}^k is a value of i^{th} object corresponding to m^{th} property given by k^{th} source. The number of data objects is 14. The number of different properties is 2. $d_1(*,*)$ is the normalized squared loss and $d_2(*,*)$ is the normalized squared loss. $\xi(W)$ is a regularization function whose value is equal to 1. W corresponds to the reliability of the sources. The regularization function is used to constrain the values of W. In problem (4.55), X^* is defined as follows:

$$X^* = \begin{bmatrix} x_{11}^* & x_{12}^* \\ x_{21}^* & x_{22}^* \\ x_{31}^* & x_{32}^* \\ \vdots & \vdots \\ x_{N1}^* & x_{N2}^* \end{bmatrix}. \tag{4.56}$$

In equation (4.56), x_{ij}^* is truth value of i^{th} object corresponding to j^{th} property. The value of N is 14. In problem (4.55), W is defined as follows:

$$W = \begin{bmatrix} w_1 & w_2 & w_3 & w_4 & w_5 & w_6 & w_7 & w_8 & w_9 \end{bmatrix}. \tag{4.57}$$

w_i is a weight value corresponding to i^{th} source.

The problem (4.55) is transformed into the multi-objective optimization problem as follows:

$$\begin{aligned} \underset{X^*,W}{\text{minimize}} \quad & (j_{28}^1(X^*,W), j_{28}^2(X^*,W)) \\ \text{subject to} \quad & \xi(W) = \sum_{k=1}^{9} e^{-w_k} = 1, \\ & W \geq 0. \end{aligned} \tag{4.58}$$

In problem (4.58), $j_{28}^1(X^*,W)$ and $j_{28}^2(X^*,W)$ are defined as follows:

$$j_{28}^1(X^*,W) = \sum_{k=1}^{9} w_k \sum_{i=1}^{14} d_1(x_{i1}^*, x_{i1}^k). \tag{4.59}$$

$$j_{28}^2(X^*,W) = \sum_{k=1}^{9} w_k \sum_{i=1}^{14} d_2(x_{i2}^*, x_{i2}^k). \tag{4.60}$$

In problem (4.58), the definition of variables is the same as in problem (4.55).

The NSGA is run for solving the multi-objective problem (4.58) and the CRH is run for solving the single-objective (4.55) on same data set. In the NSGA, the population size is 100. The figure 4.20 shows the Pareto optimal points of the NSGA and the solution from the CRH. In figure 4.20, j_{28}^1 is the function value corresponding to fourteen continuous variables and j_{28}^2 is the function value corresponding to another fourteen continuous variables. The blue points represent the Pareto optimal points and the red point represents the solution coming from the CRH. In the figure, the solution, coming from the CRH, lies on the Pareto frontier of the NSGA. The solution, getting from the CRH with the median as an initial point, is optimal in this test.

c) Third experiment involving 38 homogeneous variables:- In this experiment, it contains data corresponding to two properties. There are 38 variables

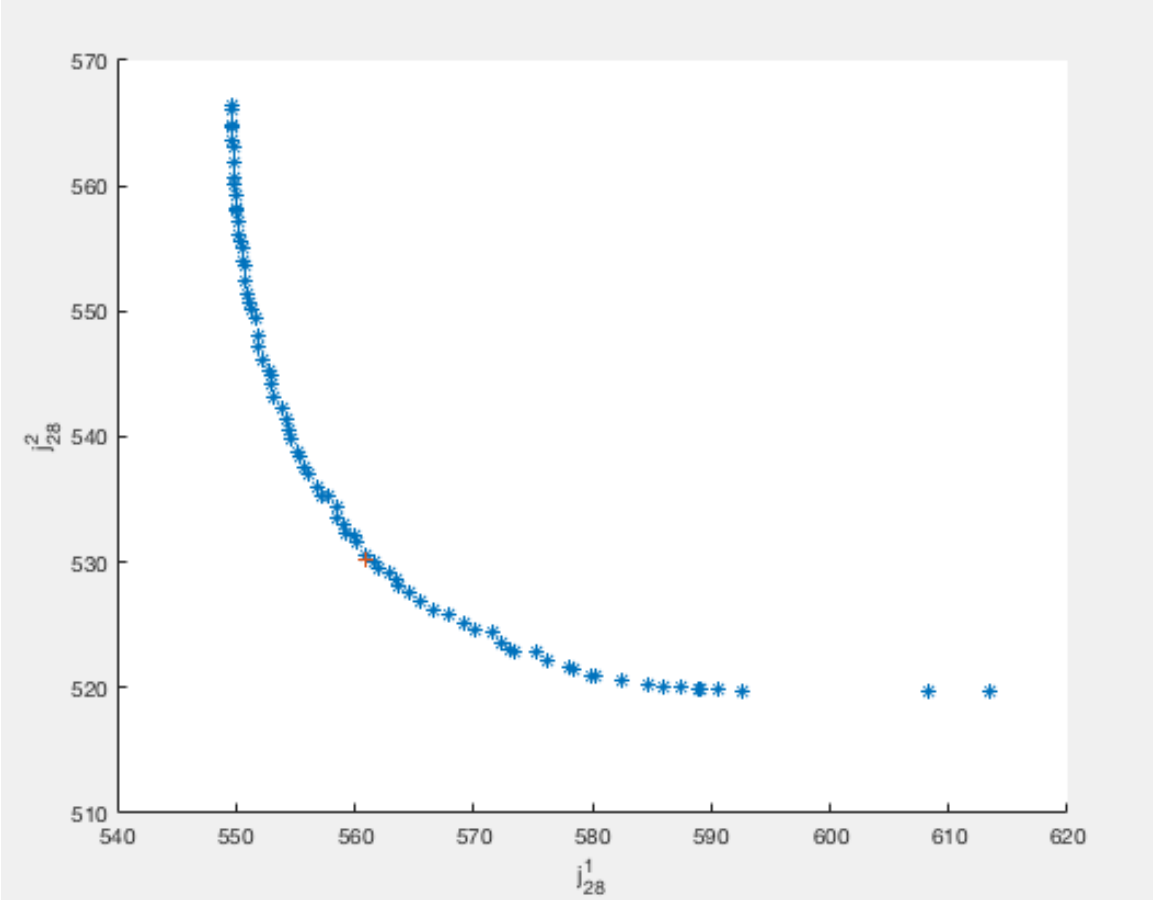

Figure 4.20: NSGA on the Problem Having 28 Homogeneous Variables.

in which all of them are of continuous type. The loss function used corresponding to continuous variables is the normalized squared loss. The problem solved in this experiment, of form (3.1), is below:

$$\begin{aligned} \underset{X^*,W}{\text{minimize}} \quad & j_{38}(X^*, W) = \sum_{k=1}^{9} w_k \sum_{i=1}^{19} \sum_{m=1}^{2} d_m(x^*_{im}, x^k_{im}) \\ \text{subject to} \quad & \xi(W) = \sum_{k=1}^{9} e^{-w_k} = 1, \\ & W \geq 0. \end{aligned} \tag{4.61}$$

In problem (4.61), W is a weight vector having 9 elements. x^k_{im} is a value of

i^{th} object corresponding to m^{th} property given by k^{th} source. The number of data objects is 19. The number of different properties is 2. $d_1(*,*)$ is the normalized squared loss and $d_2(*,*)$ is the normalized squared loss. $\xi(W)$ is a regularization function whose value is equal to 1. W corresponds to the reliability of the sources. The regularization function is used to constrain the values of W. In problem (4.61), X^* is defined as follows:

$$X^* = \begin{bmatrix} x_{11}^* & x_{12}^* \\ x_{21}^* & x_{22}^* \\ x_{31}^* & x_{32}^* \\ \vdots & \vdots \\ x_{N1}^* & x_{N2}^* \end{bmatrix}. \tag{4.62}$$

In equation (4.62), x_{ij}^* is truth value of i^{th} object corresponding to j^{th} property. The value of N is 19. In problem (4.61), W is defined as follows:

$$W = \begin{bmatrix} w_1 & w_2 & w_3 & w_4 & w_5 & w_6 & w_7 & w_8 & w_9 \end{bmatrix}. \tag{4.63}$$

w_i is a weight value corresponding to i^{th} source.

The single objective problem (4.61) is transformed into the multi-objective opti-

mization problem as follows:

$$\begin{aligned} \underset{X^*,W}{\text{minimize}} \quad & (j_{38}^1(X^*,W), j_{38}^2(X^*,W)) \\ \text{subject to} \quad & \xi(W) = \sum_{k=1}^{9} e^{-w_k} = 1, \\ & W \geq 0. \end{aligned} \tag{4.64}$$

In problem (4.64), $j_{38}^1(X^*,W)$ and $j_{38}^2(X^*,W)$ are defined as follows:

$$j_{38}^1(X^*,W) \quad = \sum_{k=1}^{9} w_k \sum_{i=1}^{19} d_1(x_{i1}^*, x_{i1}^k). \tag{4.65}$$

$$j_{38}^2(X^*,W) \quad = \sum_{k=1}^{9} w_k \sum_{i=1}^{19} d_2(x_{i2}^*, x_{i2}^k). \tag{4.66}$$

In problem (4.64), the definition of variables is the same as in problem (4.61).

The NSGA is run for solving the problem (4.64) and the CRH is run for solving the problem (4.61) on same data set. In the NSGA, the population size is 150. The figure 4.21 shows the Pareto optimal points of the NSGA and the solution from the CRH. In figure 4.21, j_{38}^1 is the function value corresponding to nineteen continuous variables and j_{38}^2 is the function value corresponding to another nineteen continuous variables. The blue points represent the Pareto optimal points and the red point represents the solution coming from the CRH. The figure shows that the solution, coming from the CRH, lies on the Pareto frontier of the NSGA. The solution, coming from the CRH with the median as an initial point, is optimal in this experiment.

d) Fourth experiment involving 62 homogeneous variables:- In this experiment, it contains data corresponding to two properties. There are 62 variables in which all of them are of continuous type. The loss function used corresponding to

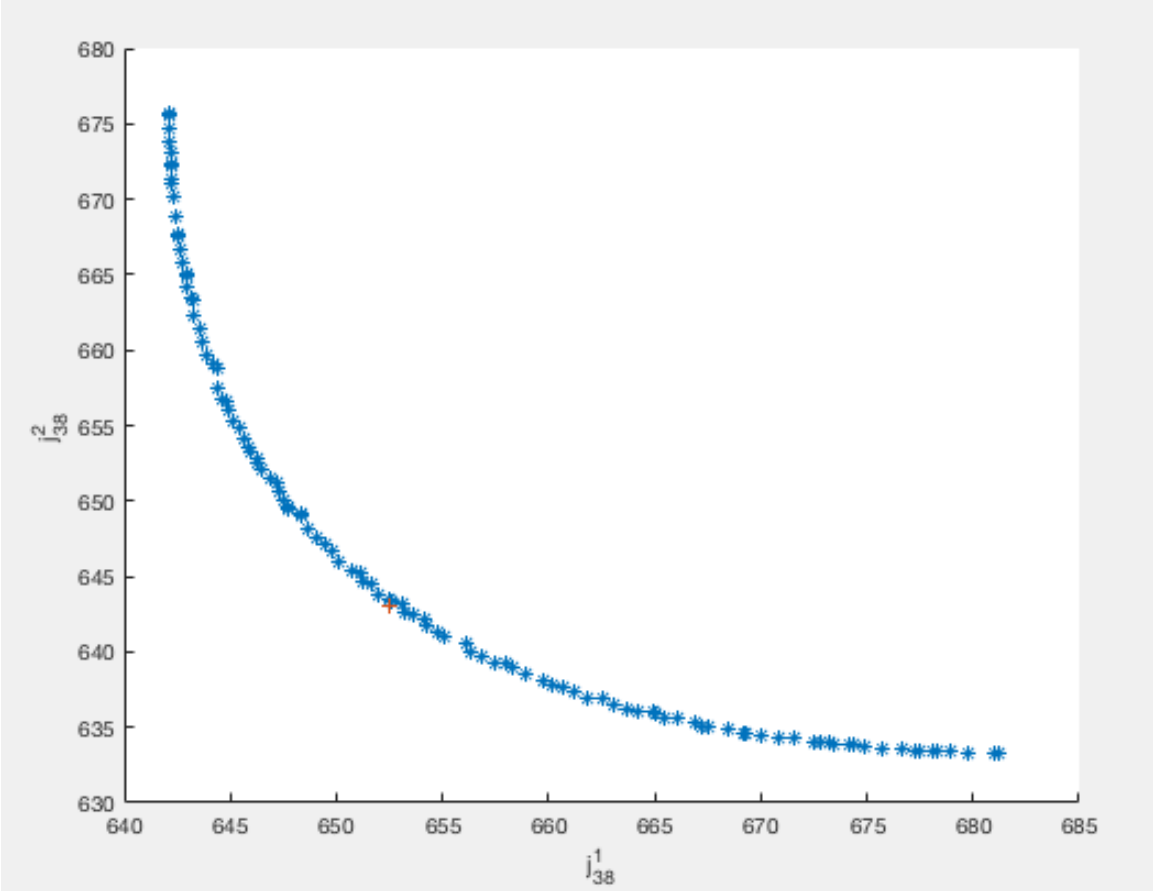

Figure 4.21: NSGA on the Problem Having 38 Homogeneous Variables.

the continuous variables is the normalized squared loss. The problem solved in this experiment, of form (3.1), is below:

$$\begin{aligned} \underset{X^*,W}{\text{minimize}} \quad & j_{62}(X^*, W) = \sum_{k=1}^{9} w_k \sum_{i=1}^{31} \sum_{m=1}^{2} d_m(x^*_{im}, x^k_{im}) \\ \text{subject to} \quad & \xi(W) = \sum_{k=1}^{9} e^{-w_k} = 1, \\ & W \geq 0. \end{aligned} \tag{4.67}$$

In problem (4.67), W is a weight vector having 9 elements. x^k_{im} is a value of i^{th} object corresponding to m^{th} property given by k^{th} source. The number of data objects is 31. The number of different properties is 2. $d_1(*,*)$ is the normalized squared loss

and $d_2(*,*)$ is the normalized squared loss. $\xi(W)$ is a regularization function whose value is equal to 1. W corresponds to the reliability of the sources. The regularization function is used to constrain the values of W. In single objective problem (4.67), X^* is defined as follows:

$$X^* = \begin{bmatrix} x^*_{11} & x^*_{12} \\ x^*_{21} & x^*_{22} \\ x^*_{31} & x^*_{32} \\ \vdots & \vdots \\ x^*_{N1} & x^*_{N2} \end{bmatrix}. \tag{4.68}$$

In equation (4.68), x^*_{ij} is truth value of i^{th} object corresponding to j^{th} property. The value of N is 31. In problem (4.67), W is defined as follows:

$$W = \begin{bmatrix} w_1 & w_2 & w_3 & w_4 & w_5 & w_6 & w_7 & w_8 & w_9 \end{bmatrix}. \tag{4.69}$$

w_i is a weight value corresponding to i^{th} source.

The problem (4.67) is transformed into the multi-objective optimization problem as follows:

$$\begin{aligned} \underset{X^*,W}{\text{minimize}} \quad & (j^1_{62}(X^*,W), j^2_{62}(X^*,W)) \\ \text{subject to} \quad & \xi(W) = \sum_{k=1}^{9} e^{-w_k} = 1, \\ & W \geq 0. \end{aligned} \tag{4.70}$$

In problem (4.70), $j^1_{62}(X^*,W)$ and $j^2_{62}(X^*,W)$ are defined as follows:

$$j_{62}^{1}(X^*, W) = \sum_{k=1}^{9} w_k \sum_{i=1}^{31} d_1(x_{i1}^*, x_{i1}^k). \tag{4.71}$$

$$j_{62}^{2}(X^*, W) = \sum_{k=1}^{9} w_k \sum_{i=1}^{31} d_2(x_{i2}^*, x_{i2}^k). \tag{4.72}$$

In problem (4.70), the definition of variables is the same as in problem (4.67).

The NSGA is run for solving the problem (4.70) and the CRH is run for solving the problem (4.67) on same data set. In the NSGA, the population size is 200. The figure 4.22 shows the Pareto optimal points of the NSGA and the solution from the CRH. In figure 4.22, j_{62}^1 is the function value corresponding to thirty one continuous variables and j_{62}^2 is the function value corresponding to another thirty one continuous variables. The blue points represent the Pareto optimal points and the red point represents the solution coming from the CRH. As evident from the figure, the solution, coming from the CRH, lies on the Pareto frontier of the NSGA. The solution, coming from the CRH with the median as an initial point, is optimal in this experiment.

e) Fifth experiment involving 14 heterogeneous variables:- In this experiment, it contains data corresponding to two properties. One property is of a categorical type and another property is of a continuous type. There are 14 variables in which 7 variables are of the categorical type and 7 variables are of the continuous type. The loss function used corresponding to the categorical variables is the 0-1 loss and the loss function used corresponding to the continuous variables is the normalized squared loss. The problem solved in this experiment, of form (3.1), is below:

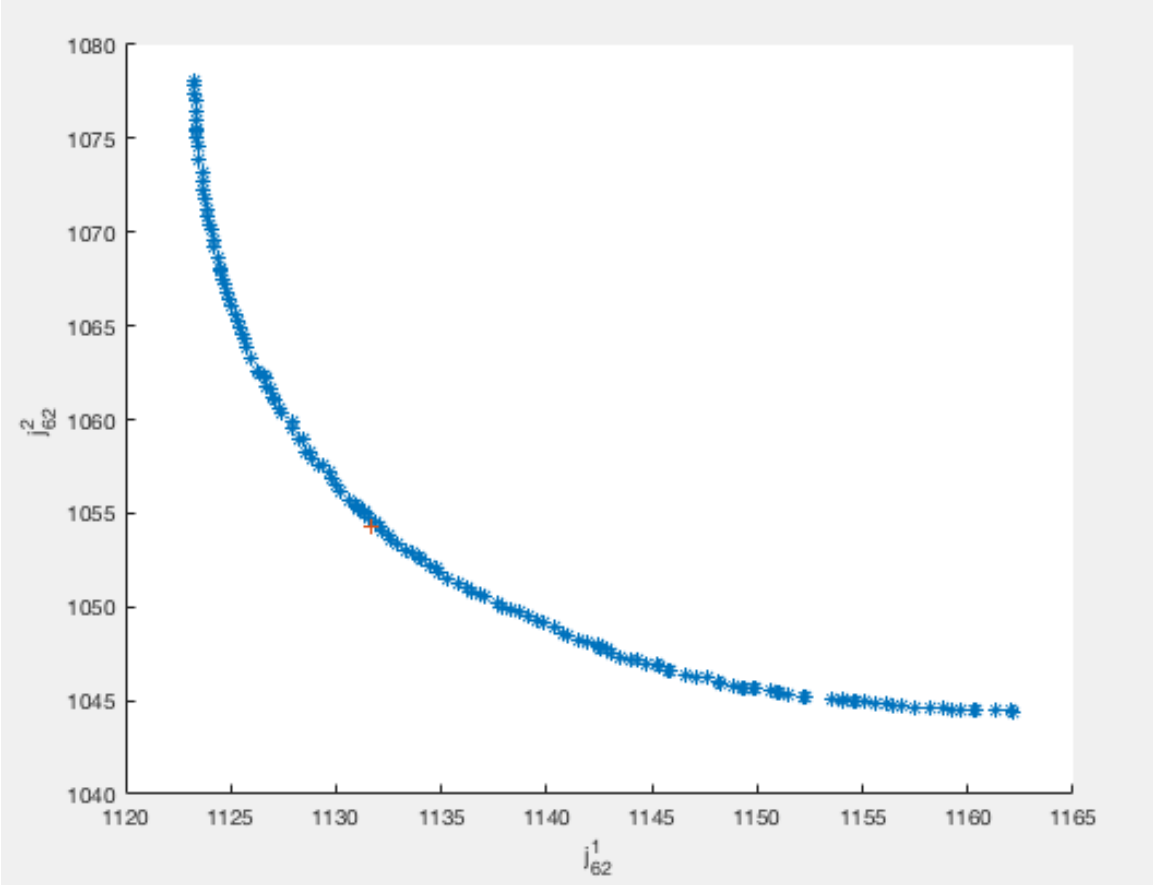

Figure 4.22: NSGA on the Problem Having 62 Homogeneous Variables.

$$\begin{aligned} \underset{X^*,W}{\text{minimize}} \quad & h_{14}(X^*, W) = \sum_{k=1}^{9} w_k \sum_{i=1}^{7} \sum_{m=1}^{2} d_m(x^*_{im}, x^k_{im}) \\ \text{subject to} \quad & \xi(W) = \sum_{k=1}^{9} e^{-w_k} = 1, \\ & W \geq 0. \end{aligned} \tag{4.73}$$

In problem (4.73), W is a weight vector having 9 elements. x^k_{im} is a value of i^{th} object corresponding to m^{th} property given by k^{th} source. The number of data objects is 7. The number of different properties is 2. $d_1(*,*)$ is the 0-1 loss and $d_2(*,*)$ is the normalized squared loss. $\xi(W)$ is a regularization function whose value is equal to 1. W corresponds to the reliability of the sources. The regularization function is

used to constrain the values of W. In the problem, X^* is defined as follows:

$$X^* = \begin{bmatrix} x_{11}^* & x_{12}^* \\ x_{21}^* & x_{22}^* \\ x_{31}^* & x_{32}^* \\ x_{41}^* & x_{42}^* \\ x_{51}^* & x_{52}^* \\ x_{61}^* & x_{62}^* \\ x_{71}^* & x_{72}^* \end{bmatrix}. \tag{4.74}$$

In equation (4.74), x_{ij}^* is the truth value of i^{th} object corresponding to j^{th} property. In problem (4.73), W is defined as follows:

$$W = \begin{bmatrix} w_1 & w_2 & w_3 & w_4 & w_5 & w_6 & w_7 & w_8 & w_9 \end{bmatrix}. \tag{4.75}$$

w_i is a weight value corresponding to i^{th} source.

The problem (4.73) is transformed into the multi-objective optimization problem as follows:

$$\begin{aligned} \underset{X^*,W}{\text{minimize}} \quad & (h_{14}^1(X^*, W), h_{14}^2(X^*, W)) \\ \text{subject to} \quad & \xi(W) = \sum_{k=1}^{9} e^{-w_k} = 1, \\ & W \geq 0. \end{aligned} \tag{4.76}$$

In problem (4.76), $h_{14}^1(X^*, W)$ and $h_{14}^2(X^*, W)$ are defined as follows:

$$h_{14}^1(X^*, W) = \sum_{k=1}^{9} w_k \sum_{i=1}^{7} d_1(x_{i1}^*, x_{i1}^k). \tag{4.77}$$

$$h_{14}^2(X^*, W) = \sum_{k=1}^{9} w_k \sum_{i=1}^{7} d_2(x_{i2}^*, x_{i2}^k). \tag{4.78}$$

In problem (4.76), the definition of the variables is same as in problem (4.73).

The NSGA is run for solving the multi-objective problem (4.76) and the CRH is run for solving the single objective problem (4.73) on same data set. In the NSGA, the population size is 100 and the number of generations is 5000. The figure 4.23 shows the pareto-optimal points of the NSGA and the solution from the CRH. In figure 4.23, h_{14}^1 is the function value corresponding to the categorical variables and h_{14}^2 is the function value corresponding to the continuous variables. The blue points represent the Pareto optimal points and the red point represents the solution coming from the CRH. As evident from the figure, the solution, coming from the CRH, lies on the Pareto frontier of the NSGA. The solution, getting from the CRH with the median as an initial point for the continuous variables and majority voting as an initial point for the categorical variables, is optimal in this test.

f) Sixth experiment involving 28 heterogeneous variables:- In this experiment, it contains data corresponding to two properties. One property is of a categorical type and another property is of a continuous type. There are 28 variables in which 14 variables are of the categorical type and 14 variables are of the continuous type. The loss function used corresponding to the categorical variables is the 0-1 loss and the loss function used corresponding to the continuous variables is the normalized squared loss. The problem solved in this experiment, of form (3.1), is below:

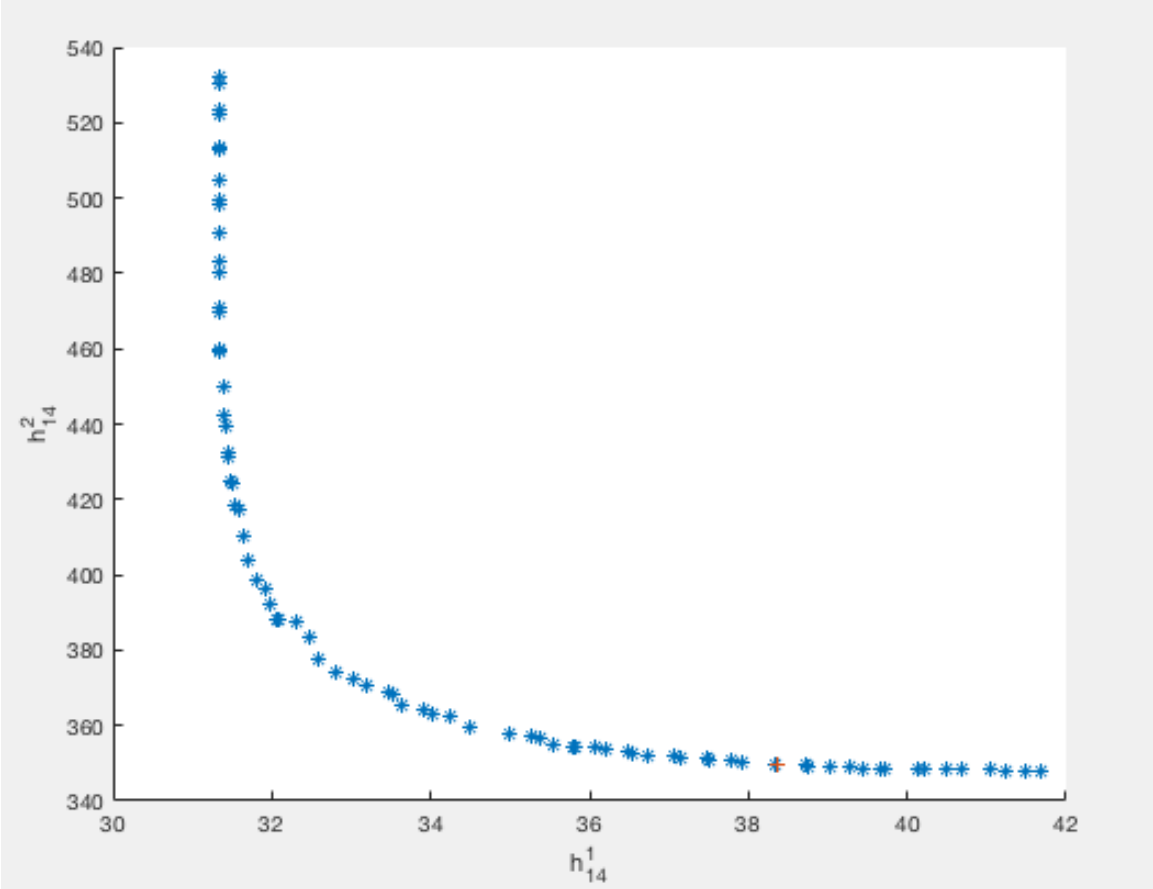

Figure 4.23: NSGA on the Problem Having 14 Heterogeneous Variables.

$$\begin{aligned} \underset{X^*,W}{\text{minimize}} \quad & h_{28}(X^*, W) = \sum_{k=1}^{9} w_k \sum_{i=1}^{14} \sum_{m=1}^{2} d_m(x^*_{im}, x^k_{im}) \\ \text{subject to} \quad & \xi(W) = \sum_{k=1}^{9} e^{-w_k} = 1, \\ & W \geq 0. \end{aligned} \tag{4.79}$$

In problem (4.79), W is a weight vector having 9 elements. x^k_{im} is a value of i^{th} object corresponding to m^{th} property given by k^{th} source. The number of data objects is 14. The number of different properties is 2. $d_1(*,*)$ is the 0-1 loss and $d_2(*,*)$ is the normalized squared loss. $\xi(W)$ is a regularization function whose value is equal to 1. W corresponds to the reliability of the sources. The regularization function is

used to constrain the values of W. In problem (4.79), X^* is defined as follows:

$$X^* = \begin{bmatrix} x^*_{11} & x^*_{12} \\ x^*_{21} & x^*_{22} \\ x^*_{31} & x^*_{32} \\ \vdots & \vdots \\ x^*_{N1} & x^*_{N2} \end{bmatrix}. \tag{4.80}$$

In equation (4.80), x^*_{ij} is the truth value of i^{th} object corresponding to j^{th} property and N = 14. In problem (4.79), W is defined as follows:

$$W = \begin{bmatrix} w_1 & w_2 & w_3 & w_4 & w_5 & w_6 & w_7 & w_8 & w_9 \end{bmatrix}. \tag{4.81}$$

w_i is a weight value corresponding to i^{th} source.

The single objective problem (4.79) is transformed into the multi-objective optimization problem as follows:

$$\begin{aligned} \underset{X^*,W}{\text{minimize}} \quad & (h^1_{28}(X^*, W), h^2_{28}(X^*, W)) \\ \text{subject to} \quad & \xi(W) = \sum_{k=1}^{9} e^{-w_k} = 1, \\ & W \geq 0. \end{aligned} \tag{4.82}$$

In problem (4.82), $h^1_{28}(X^*, W)$ and $h^2_{28}(X^*, W)$ are defined as follows:

$$h^1_{28}(X^*, W) \quad = \sum_{k=1}^{9} w_k \sum_{i=1}^{14} d_1(x^*_{i1}, x^k_{i1}). \tag{4.83}$$

$$h_{28}^2(X^*, W) \quad = \sum_{k=1}^{9} w_k \sum_{i=1}^{14} d_2(x_{i2}^*, x_{i2}^k). \tag{4.84}$$

In problem (4.82), the definition of the variables is same as in problem (4.79).

The NSGA is run for solving the multi-objective problem (4.82) and the CRH is run for solving the single objective problem (4.79) on same data set. In the NSGA, the population size is 200. The figure 4.24 shows the Pareto optimal points of the NSGA and the solution from the CRH. In figure 4.24, h_{28}^1 is the function value corresponding to the categorical variables and h_{28}^2 is the function value corresponding to the continuous variables. The blue points represent the Pareto optimal points and the red point represents the solution coming from the CRH. The figure shows that the solution, coming from the CRH, lies on the Pareto frontier of the NSGA. The solution, getting from the CRH with the median as an initial point for the continuous variables and majority voting as an initial point for the categorical variables, is optimal in this test.

g) Seventh experiment involving 38 heterogeneous variables:- In this experiment, it contains the data corresponding to two properties. One property is of a categorical type and another property is of a continuous type. There are 38 variables in which 19 variables are of the categorical type and 19 variables are of the continuous type. The loss function used corresponding to the categorical variables is the 0-1 loss and the loss function used corresponding to the continuous variables is the normalized squared loss. The problem solved in this experiment, of form (3.1), is below:

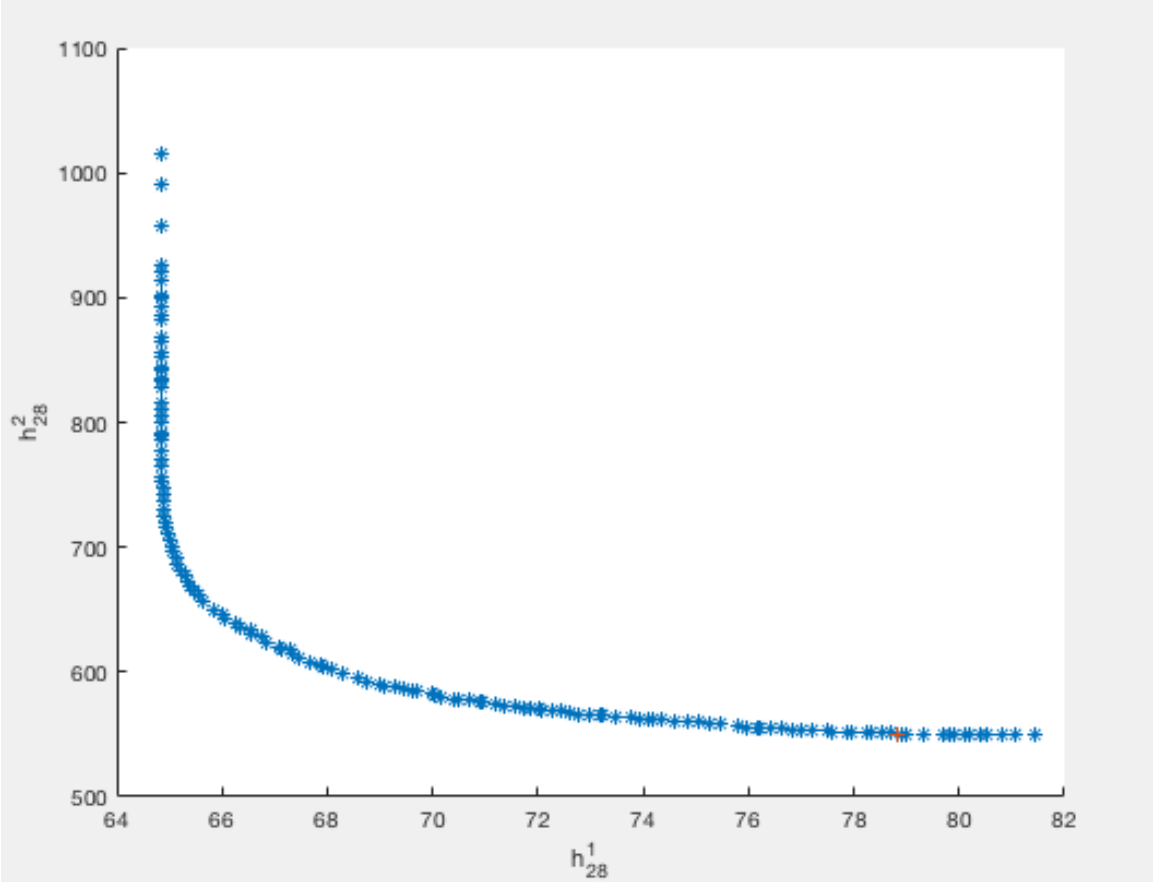

Figure 4.24: NSGA on the Problem Having 28 Heterogeneous Variables.

$$\begin{aligned} \underset{X^*,W}{\text{minimize}} \quad & h_{38}(X^*, W) = \sum_{k=1}^{9} w_k \sum_{i=1}^{19} \sum_{m=1}^{2} d_m(x^*_{im}, x^k_{im}) \\ \text{subject to} \quad & \xi(W) = \sum_{k=1}^{9} e^{-w_k} = 1, \\ & W \geq 0. \end{aligned} \tag{4.85}$$

In problem (4.85), W is a weight vector having 9 elements. x^k_{im} is a value of i^{th} object corresponding to m^{th} property given by k^{th} source. The number of data objects is 19. The number of different properties is 2. $d_1(*,*)$ is the 0-1 loss and $d_2(*,*)$ is the normalized squared loss. $\xi(W)$ is a regularization function whose value is equal to 1. W corresponds to the reliability of the sources. The values of W are constrained

by the regularization function. In the problem, X^* is defined as follows:

$$X^* = \begin{bmatrix} x^*_{11} & x^*_{12} \\ x^*_{21} & x^*_{22} \\ x^*_{31} & x^*_{32} \\ \vdots & \vdots \\ x^*_{N1} & x^*_{N2} \end{bmatrix}. \tag{4.86}$$

In equation (4.86), x^*_{ij} is truth value of i^{th} object corresponding to j^{th} property and N = 19. In problem (4.85), W is defined as follows:

$$W = \begin{bmatrix} w_1 & w_2 & w_3 & w_4 & w_5 & w_6 & w_7 & w_8 & w_9 \end{bmatrix}. \tag{4.87}$$

w_i is a weight value corresponding to i^{th} source.

The problem (4.85) is transformed into the multi-objective optimization problem as follows:

$$\begin{aligned} \underset{X^*,W}{\text{minimize}} \quad & (h^1_{38}(X^*, W), h^2_{38}(X^*, W)) \\ \text{subject to} \quad & \xi(W) = \sum_{k=1}^{9} e^{-w_k} = 1, \\ & W \geq 0. \end{aligned} \tag{4.88}$$

In problem (4.88), $h^1_{38}(X^*, W)$ and $h^2_{38}(X^*, W)$ are defined as follows:

$$h^1_{38}(X^*, W) \quad = \sum_{k=1}^{9} w_k \sum_{i=1}^{19} d_1(x^*_{i1}, x^k_{i1}). \tag{4.89}$$

$$h_{38}^2(X^*, W) \quad = \sum_{k=1}^{9} w_k \sum_{i=1}^{19} d_2(x_{i2}^*, x_{i2}^k). \tag{4.90}$$

In problem (4.88), the definition of variables is same as in problem (4.85).

The NSGA is run for solving the multi-objective problem (4.88) and the CRH is run for solving the single-objective problem (4.85) on same data set. In the NSGA, the population size is 150. The figure 4.25 shows the Pareto optimal points of the NSGA and the solution from the CRH. In figure 4.25, h_{38}^1 is the function value corresponding to the categorical variables and h_{38}^2 is the function value corresponding to the continuous variables. The blue points represent the Pareto optimal points and the red point represents the solution coming from the CRH. As evident from the figure, the solution, coming from the CRH, lies on the Pareto frontier of the NSGA. The solution, getting from the CRH with the median as an initial point for the continuous variables and majority voting as an initial point for the categorical variables, is optimal in this test.

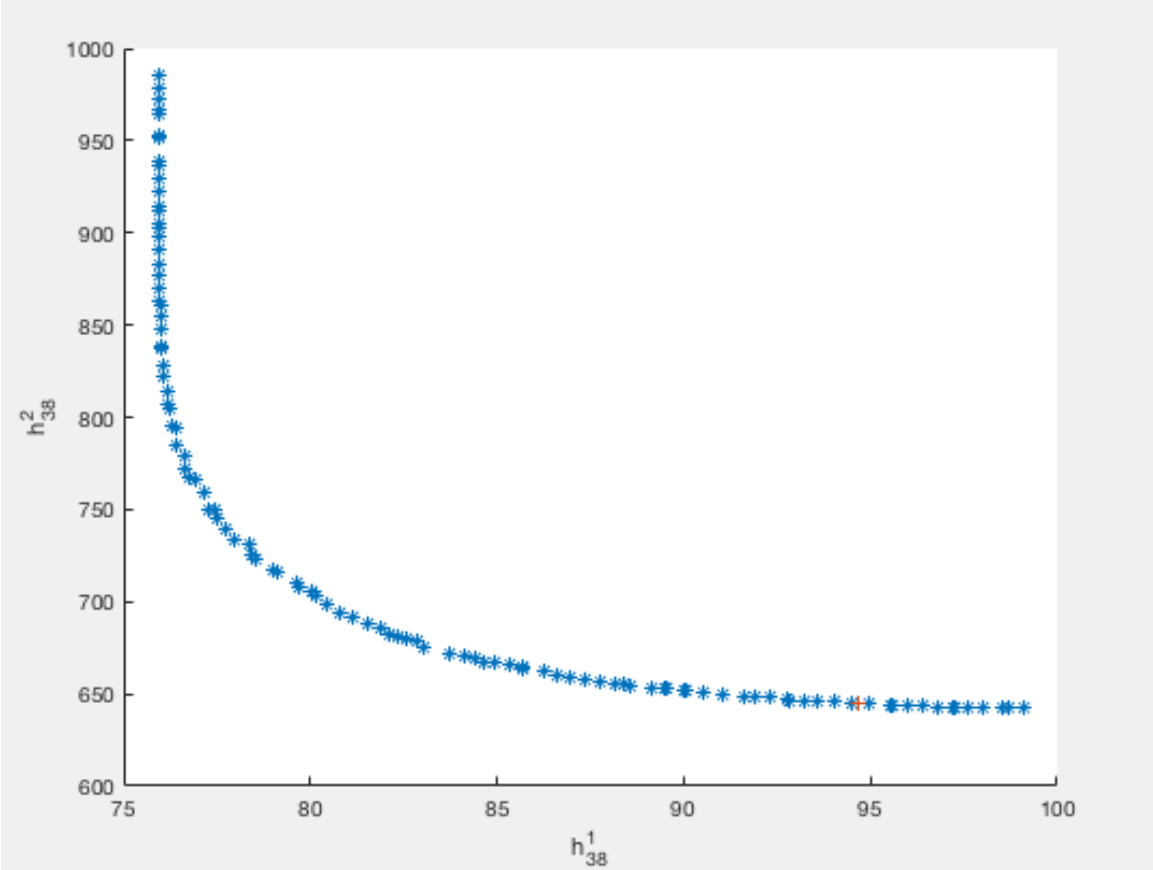

Figure 4.25: NSGA on the Problem Having 38 Heterogeneous Variables.

Chapter 5

CONCLUSION

To extract the useful knowledge from the data while keeping in mind the non-uniform reliability of the sources, the single objective optimization introduced in Li *et al.* (2014), is formulated into the multi-objective optimization problem. In one approach (SOOWMS), the CRH method Li *et al.* (2014) is run with many different initial points on different test cases. The solution coming from these experiments is the same as the solution coming from the CRH method with the median as an initial point for continuous variables. In another approach (NSGA), the NSGA method Deb *et al.* (2002) is run with data set (heterogeneous or homogeneous). These experiments give an output of the Pareto frontier. The solution coming from the CRH method, with the median as an initial point for continuous variables and majority voting as an initial point for categorical variables, lies on the Pareto frontier of the NSGA. The solution getting from the CRH method, with the median as an initial point for continuous variables and majority voting as an initial point for categorical variables, is optimal in these experiments.

REFERENCES

Agrawal, R. B., K. Deb and R. B. Agrawal, "Simulated binary crossover for continuous search space", Complex systems **9**, 2, 115–148 (1995).

Banerjee, A., S. Merugu, I. S. Dhillon and J. Ghosh, "Clustering with bregman divergences", Journal of machine learning research **6**, Oct, 1705–1749 (2005).

Bertsekas, D., "Nonlinear programming, athena scientific, 1999", REFER ENCIAS BIBLIOGR AFICAS **89** (2006).

Bleiholder, J. and F. Naumann, *Conflict handling strategies in an integrated information system* (Humboldt-Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät , 2006).

Boyd, S. and L. Vandenberghe, *Convex optimization* (Cambridge university press, 2004).

Deb, K., A. Pratap, S. Agarwal and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: Nsga-ii", IEEE Transactions on Evolutionary Computation **6**, 2, 182–197 (2002).

Dong, X. L. and F. Naumann, "Data fusion: resolving data conflicts for integration", Proceedings of the VLDB Endowment **2**, 2, 1654–1655 (2009).

Jiang, Z., "A decision-theoretic framework for numerical attribute value reconciliation", IEEE Transactions on Knowledge and Data Engineering **24**, 7, 1153–1169 (2012).

Kasneci, G., J. V. Gael, D. Stern and T. Graepel, "Cobayes: bayesian knowledge corroboration with assessors of unknown areas of expertise", in "Proceedings of the fourth ACM international conference on Web search and data mining", pp. 465–474 (ACM, 2011).

Li, Q., Y. Li, J. Gao, B. Zhao, W. Fan and J. Han, "Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation", in "Proceedings of the 2014 ACM SIGMOD international conference on Management of data", pp. 1187–1198 (ACM, 2014).

Li, X., X. L. Dong, K. Lyons, W. Meng and D. Srivastava, "Truth finding on the deep web: Is the problem solved?", Proceedings of the VLDB Endowment **6**, 2, 97–108 (2012).

Marian, A. and M. Wu, "Corroborating information from web sources.", IEEE Data Eng. Bull. **34**, 3, 11–17 (2011).

Srinivas, N. and K. Deb, "Muiltiobjective optimization using nondominated sorting in genetic algorithms", Evolutionary computation **2**, 3, 221–248 (1994).

Zhao, B. and J. Han, "A probabilistic model for estimating real-valued truth from conflicting sources", Proc. of QDB (2012).

Zhao, B., B. I. Rubinstein, J. Gemmell and J. Han, "A bayesian approach to discovering truth from conflicting sources for data integration", Proceedings of the VLDB Endowment **5**, 6, 550–561 (2012).