Cosmological parameter estimation from CMB data

An optimization problem

Student

Pranati Dalvi

(*Physics Department, Fergusson College, Pune (India)*))

Supervisor

Dr. Jayanti Prasad

(*CMS, Savitribai Phule Pune University, Pune (India)*)

A thesis submitted in fulfilment of the requirements for the degree of Masters in Physics (M.Sc) in Physics

Physics Department

Fergusson College Pune

Feb 2016

Contents

Abstract

In Astronomy and Astrophysics a large amount of data is collected from ground and space based instruments and this data needs to be reduced preserving the information contained in the data. One of the most common ways to achieve this has been to fit the observed data with a theoretical model which leads to finding the global maximum of a function (called "optimization function" or the "cost function") in a multi-dimensional parameter space (chi-square minimization etc.). In general, optimization problems are challenging because they involves search in very high dimensional space and in many cases there are local maxima also present. There is no single optimization technique which will work for all the problems and so one needs to explore a set of optimization techniques in order to find the most efficient technique for the problem one has.

In the present work we study a set of optimization techniques in the context of cosmological parameter estimation from cosmic microwave background (CMB) data. We use publicly available data from the Wilkinson Microwave Anisotropy Probe (WMAP) nine year relase for our exercise. The summary of the work is as follows.

In Chapter 1 we introduce the problem of optimization and discuss the challenges of multi-dimensional optimization. We also introduce Bayesian parameter estimation in this Chapter. In Chapter 2 we discuss stochastic optimization techniques and present the results for a set of toy/simulated problems. In Chapter 3 We introduce CMB data analysis and give the back-

ground of the cosmological parameter estimation. In Chapter 4 we present a new technique for cosmological parameter estimation and compare that with Particle Swarm Optimization method.

Chapter 1

Optimization

1.1 Introduction

Optimization problems are very common in social, natural and engineering sciences and can be defined as finding a solution out of many solutions which is the best, in the sense minimizes/maximizes a user defined "cost" or "optimization" function. One of the common examples of it is the traveling salesman problem - finding a route which will minimize the distance traveled by a salesman when visiting a group of cities, with or without any constraint.

Solving of a set of M linear equations with N independent variables also can be cast as an optimization problem. In other words solving,

$$y = Ax, \tag{1.1}$$

is equivalent to finding the minimum of the quadratic function $f(x)$ defined as

$$f(x) = \frac{1}{2}x^T Ax - y^T x + c. \tag{1.2}$$

Optimization is the determination of a variable, for which the given function holds a maximum or a minimum value. At this particular value of the function, it's derivative becomes zero. Thus if we have a function f(x), the optimization problem may be written as

$$f'(x) = 0 = y - Ax. \tag{1.3}$$

Further, the second derivative of that function will tell if the point is a maximum or a minimum. The point is said to be it's maximum when the second derivative is negative, and when the second derivative is positive, it is said to be a minimum. Thus, when

$$f''(x) > \qquad 0 : Minima \tag{1.4}$$

$$f''(x) < \qquad 0 : Maxima \tag{1.5}$$

Chi-square minimization by which we try to find a set of theoretical parameters given by a N-dimensional vector x from a set of data points given by a M-dimensional vector y is an example of optimization. In this case case we minimize (with respect to x) the Chi-square which is defined as:

$$\chi^2(x) = \sum_{i=1}^{M} \left(\frac{y_i - y^{th}(x)}{\sigma_i} \right)^2. \tag{1.6}$$

Here we assume that every data points y_i includes Gaussian noise also which has zero mean and variance σ_i and y^{th} is the theoretical model we are trying to fit data, which may be linear or non-linear. For the case of fitting a line:

$$y = ax + b, \tag{1.7}$$

we can write:

$$y = \begin{pmatrix} y_1 \\ y_2 \\ .. \\ .. \; y_M \end{pmatrix}; x = \begin{pmatrix} a \\ b \end{pmatrix}; A = \begin{pmatrix} x_1 & 1 \\ y_2 & 1 \\ .. & \\ .. \; x_M & 1 \end{pmatrix}. \tag{1.8}$$

For the case of one dimension optimization problem can be casted as the root finding problem, if the derivative of the function is changing sign between two points x_1 and x_2 then there must be a point $x_1 < x < x_2$ at

which $f'(x) = 0$. Bisection as is explained in Fig (1.1) is one of common methods of root finding.

In multidimensional optimization problems root finding methods do not guarantee success due to complexity involved, although there are some methods like Golden section [1] which sometime can work. One of the ways to classify optimization methods is whether they need to find the derivative of the optimization function or not. Newton's like methods need derivatives so are not very useful for the problem in which either we can compute the derivative of the function or it is computationally very expensive.

Optimization methods which do not need computing derivatives can be further divided into deterministic and stochastic. In general, deterministic methods like Downhill simplex methods are Geometrical in nature. On the other hand stochastic methods like Markov Chain Monte Carlo (MCMC) and Artificial Intelligence (AI) inspired methods like Artificial Neural Network (ANN), Genetic Algorithms (GA), Particle Swarm Optimization (PSO), Ant Colony Optimization(ACP) and Opinion Dynamics (OD) are based on random walks.

1.2 Bayesian Analysis

Bayesian analysis is one of the most common methods to test theoretical models, represented by a set of parameters against data. Bayesian theorem is based on the following proposition.

Conditional probability $P(A|B)$ of an event A given that event B has already occurred is given by:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \tag{1.9}$$

similarly we can also write:

$$P(B|A) = \frac{P(B \cap A)}{P(A)}. \tag{1.10}$$

Since $P(A \cap B) = P(B \cap A)$ so we can write:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{1.11}$$

If we identify A with the parameters θ of the theoretical model we want to constrain and B with the data d then we can write :

$$\boxed{P(\theta|d) = \frac{P(d|\theta)P(\theta)}{P(d)}.} \tag{1.12}$$

This expression is called the Bayes' theorem. In the above equation various terms are as follows:
$P(\theta|d)$: Posterior probability
$P(d|\theta)$: Likelihood
$P(\theta)$: Prior probability
$P(d) = \int d\theta P(d|\theta)P(\theta)$:Evidence.

Bayesian method uses probability to express belief in a statement about unknown quantities, and uses subjective probability or prior $P(\theta)$.

The elements of the sample space might also be interpreted as hypotheses, i.e., statements that are either true or false. After repetition of an experiment, conclusions can be drawn regarding how true it is. One can interpret this as

$$P(theory|data) \propto P(data|theory)P(theory) \tag{1.13}$$

where *theory* represents some hypothesis and *data* is the outcome of the experiment. Here $P(theory)$ is the *prior probability* for the theory, $P(data|theory)$ is the *likelihood* and $P(theory|data)$ is the *posterior distribution*

Bayesian Analysis includes the following steps:

1. Formulate a probability model for the data.

2. Decide a *prior distribution* (A Prior Distribution is a function which represents the current state of knowledge or uncertainty in the values of the unknown model parameters before the data are observed).

3. Construct a Likelihood function. (A function based on the data and the probability model. It is then combined with prior distribution to determine *posterior distribution*).

4. Describe a Posterior Distribution and calculate quantities of interest.

Thus, Bayesian Analysis is used to obtain posterior distribution of model parameters. The posterior distribution can be said to be weighted average of knowledge of parameters before data is observed (Prior Distribution) and information about parameters contained in observed data (Likelihood Function). A likelihood function expresses how likely the observed distribution is for a given model. For the Gaussian case we can write,

$$\mathcal{L} = \prod_{i}^{N} \left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i - y_{avg})^2}{2\sigma^2}}\right) \tag{1.14}$$

Also, it is often required to compare an observed distribution with a simulated model. In such cases, an analysis might be designed to extract some physical parameter from the simulation which best fits the data. Here, the chi square test is commonly utilized to test the goodness of fit. Chi square expresses the deviation of the observed data from the fit. It can also be used for finding the optimal values of the parameters by minimizing the chi square. For data points wth Gaussian errors, we can write chi square function as

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - y_{avg})^2}{\sigma^2} \tag{1.15}$$

Here, the y_i are the observed data, y_{avg} is the average of the observed data and σ^2 is the variance observed in the data.

Thus, minimizing chi square is nearly same as maximizing likelihood. From equation (1.14), it is understood that likelihood function $\mathcal{L}$ is nothing but the product of the probabilities over N trials of an experiment. The

probability of occurrence of an event in each trial is less than 1 and the product of N trials would yield a very small number. So, we define a logarithmic scale and convert the product to sum, such that

$$M = \ln \mathcal{L} \tag{1.16}$$

Therefore,

$$\ln \mathcal{L} = \ln \frac{1}{\sigma\sqrt{2\pi}} + \sum \frac{(y_i - y_{avg})^2}{2\sigma^2} \tag{1.17}$$

Thus, we can see that,

$$\chi^2 = 2 \ln \mathcal{L} + \ln \frac{1}{\sigma\sqrt{2\pi}}$$

In general,

$$\chi^2 = 2 \ln \mathcal{L} + constant \tag{1.18}$$

Now, for a linear problem, Y = AX + B, where Y and B are M dimensional vectors, X is N dimensional vector and A is a $M \times N$ matrix, the χ^2 is defined as:

$$\chi^2 = (Y - AX)^T C_N^{-1} (Y - AX) \tag{1.19}$$

where, $C_N =< BB^T >$ is a noise covariance matrix. This can be explained as: If we have a series of measurements, giving N data points, each consisting of (x_i, y_i, σ_i), where σ_i is the standard deviation of the measurement y_i, and we expect the relationship between x and y to be given by the expression

$$y = ax + b \tag{1.20}$$

then we are interested in finding a and b. Here,

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - ax_i - b)^2}{\sigma_i^2} \tag{1.21}$$

The minimum χ^2 occurs at $\frac{\partial \chi^2}{\partial a} = 0$ and $\frac{\partial \chi^2}{\partial b} = 0$. Thus, after solving these partial derivatives, we estimate b and a. Further, for Y = AX + B, if we

write

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_N \end{pmatrix}$$

and

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ . & . \\ . & . \\ 1 & x_N \end{pmatrix}$$

and

$$A = \begin{pmatrix} a \\ b \end{pmatrix}$$

and let the noise be uncorrelated. The correlation matrix is

$$C_N = \begin{pmatrix} \sigma_i^2 & 0 & 0 & . & . & . & . & . & 0 \\ 0 & \sigma_i^2 & . & . & . & . & . & . & 0 \\ . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . \\ 0 & . & . & . & . & . & . & . & \sigma_i^2 \end{pmatrix}$$

The maximum likelihood is given by:

$$\mathcal{L} = [(A^T C_N^{-1} A)^{-1} A^T C_N^{-1}] Y$$

Thus, by simplifying the above multiplication of matrices, we find the parameters A and B

1.3 The Chi Square Distribution

In probability theory and statistics, the chi-square distribution, having k degrees of freedom is the distribution of a sum of the squares of k independent standard normal random variables. A standard normal variable is the one which has zero mean and unit variance and follows the Gaussian Distribution. The probability density function is given by

$$f(x) = \frac{1}{2^{k/2}\Gamma(n/2)} x^{(n/2)-1} e^{-x/2} \tag{1.22}$$

Thus as k becomes significantly large, the Chi-Square distribution tends to the Gaussian distribution. The statistic is given by

$$\chi^2 = \sum_{i=1}^{k} \frac{(x_i - \mu)^2}{\sigma^2} \tag{1.23}$$

where x_i are the data points and μ is the expectation value. This expression however, does not represent the χ^2 distribution, but is the Test Statistic. We also say that χ^2 is a statistic that characterizes the dispersion of the observed frequencies from the expected frequencies.

This distribution has many applications like the Hypothesis testing or Testing the goodness of fit.

Test of a hypothesis is a rule that states for which data values x the hypothesis is rejected. There exists a significance level α of the test,which is the probability of rejection H_0 when it is tru,e and a critical region for the data which is the set of values which would cause us to reject H_0. Thus, if the data are observed in the critical region, H_0 is rejected. Thus, after carrying out the test, decisions are made depending upon the level of significance.

The idea behind the chi square test of goodness of fit is used to check whether the sample comes from the claimed distribution. Two values of frequencies are involved in this test, the observed one and the expected one. If the observed frequency is closed to the claimed frequency, then the square of the deviations will be small. If this is the case, then the claim is accepted.

1.4 Stochastic Optimization

Stochastic optimization is an area which comprises modeling and methodology for solving optimization problems taking the uncertainties into account. It is nothing but a random search of the root with some prior knowledge of the distribution. This method assumes that if a sufficient number of samples are conducted, the optimum will be eventually located.

These methods are somewhat intuitive methods, wherein the best possible value of the objective function is found by random movements in search space. The search is terminated after a certain criteria are fulfilled like the desired number of iterations etc. One such technique which is commonly employed for the process is the Monte Carlo Method. This method uses random numbers in scientific computing. The main idea is to generate a Markov chain whose limiting distribution is equal to the desired distribution. A Markov Chain is a stochastic model describing a sequence of possible events in which the probability of each event depends only on the state attained in the previous event.

If we have a model with parameter λ and the data y, we can evaluate the probability density $P(\lambda) \equiv P(\lambda|y)$ as

$$P(\lambda) = P(y|\lambda)\frac{P(\lambda)}{P(y)} \tag{1.24}$$

where usually the distribution of the data $y = (y_i, ..., y_N)$ usually follows Gaussian Distribution. Once the prior distribution is fixed, we can evaluate the probability of the parameter λ for the given data y.

1.5 Artificial Intelligence

Artificial intelligence, or AI, is the field that studies the synthesis and analysis of computational agents that act intelligently. A computational agent is someone whose decisions about its actions can be explained in terms of

computation. As far as optimization is considered, it helps to extract characteristic information and pattern recognition. Some of the optimization techniques based on this concept are:

1. Stimulated Annealing: This is a random search method, which makes its potential solutions to move around the search space. Firstly, it generates a random solution and calculate its cost using some pre defined cost function. Then, a random neighboring solution is generated and it's cost is calculated. The two costs are then compared. If $c_{new} < c_{old}$: move to the new solution. If $c_{new} > c_{old}$: maybe move to the new solution. Above steps are then repeated until an acceptable solution is found or the stopping criteria is reached.

2. Swarm Intelligence: It is artificial intelligence based on the collective behavior of self-organized systems. Such systems are made up of a population of simple agents interacting locally with one another, and with their environment. Each member of the system is considered to be a potential solution, and these particles move in search space in search of the solution. As in case of stimulated annealing, the new position of particle is evaluated in terms of cost function, here, the position is updated by considering the velocity of neighbors, and the local best, and ultimately, global best is reached.

3. Genetic Algorithm: Genetic Algorithm (GA) is a search algorithm based on Charles Darwin's theory of 'The survival of the fittest'. The idea behind this is that the individuals in a population compete for resources and mates. Those individuals most successful in each 'competition' will produce more offspring than those individuals that perform poorly. Genes from 'good' individuals propagate throughout the population so that two good parents will sometimes produce offspring that are better than either parent. Thus each successive generation will

become more suited to their environment. This idea is converted into a standard algorithm to solve the optimization problems.

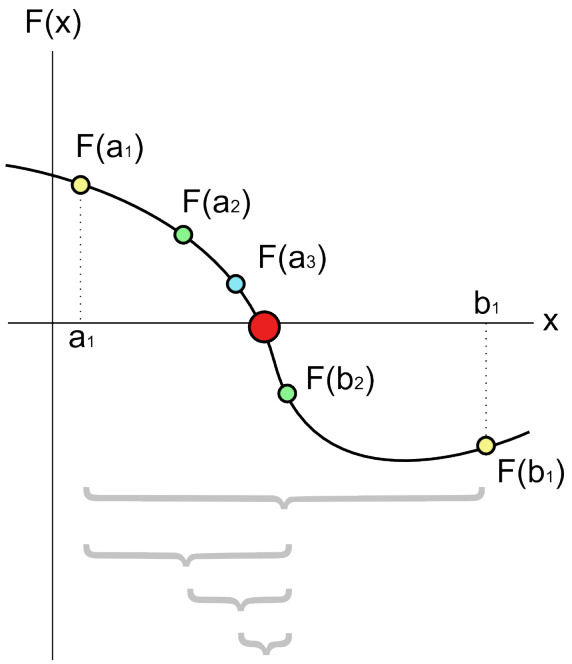


Figure 1.1: The bisection method is the simplest method for finding the root of a one-dimensional continuous function on a closed interval, and is based on the intermediate value theorem on continuous functions. It works by continuously reducing the search range into two halves. It assumes that the function has opposite signs at the two ends of the given range, and thus finds the point at which the function changes the sign, i.e, the root. This method is iterative. It accepts the more probable half, and repeats the procedure, till the zero is found out.

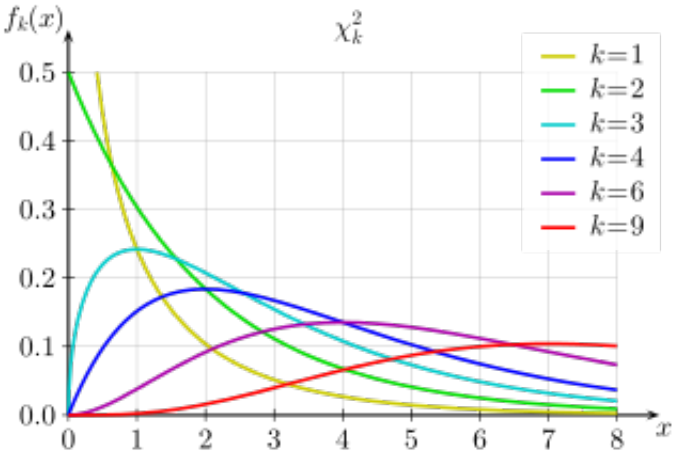


Figure 1.2: The Chi Square Distribution

Chapter 2

Parameter Estimation

2.1 Introduction

Parameter estimation is nothing but finding the value of a population attribute from sample statistics. These parameters are fixed and not known. They are the factors that define the population. Usually, one doesn't have the knowledge of the entire population, so the parameters are estimated with only the knowledge of sample. In general, this is achieved by solving an optimization problem in which a function connects the variables of the problem and the functional part of the model that contains unknown parameters. The unknown parameters are treated as variables to be solved for in the optimization while the data being treated as coefficients of the function. While estimating the parameters, one must take into consideration the identifiability and limits of parameters and whether the model under consideration is linear or non-linear. Also, a large number of data points is a must.

Estimating parameters in a cognitive model is similar to estimating regression coefficients. However, Regression models belong to the general linear class, whereas cognitive models are usually nonlinear. Linear models have a property that the average of the predictions from two different sets of parameters is equal to the prediction produced by the average of the two

sets of parameters. Nonlinear models do not satisfy this property.

2.2 Regression

Regression is an approach for modeling the relationship between a scalar dependent variable y and one or more variables x. Linear regression consists of finding the best-fitting straight line through the points. The best-fitting line is called a regression line. A best fit line is the one which minimizes the variance of the points. Frequently we have an estimate of the uncertainty in the data points Y, the "error bar". A fit would be considered satisfactory if it goes through the points within the error bars. The formula for a regression line is:

$$Y = aX + b \tag{2.1}$$

where a is the slope and b is the y intercept. These are the parameters of the data.

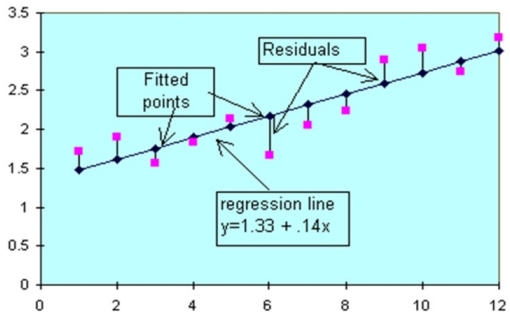


Figure 2.1: Regression Line

Linear models are fairly simple and solving the problem of regression is similar to that of solving χ^2 minimization or maximizing likelihood $\mathcal{L}$, which is explained earlier.

2.3 Least Square Fitting To A Polynomial

Often it is required to fit a higher order polynomial than a straight line, as for example in the plot below where the fit is a parabola. By defining the function

$$f(x) = \sum_{i=0}^{N} (y_i - \sum_{\alpha} a_\alpha x_i^\alpha)^2 \tag{2.2}$$

and following the same procedure, we can find the best fit parameters.

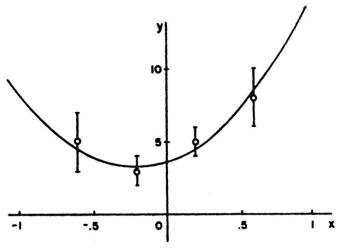


Figure 2.2: Least Square Fit

2.4 Non Linear Fitting

As mentioned earlier, most of the models of our concern are non linear, and a need of different estimators arises. Because analytic methods of least-squares fitting cannot be used for nonlinear fitting problems, approximation methods must be considered and searches of parameter space should be made.

In some cases, non-linear function is converted into a linear one by taking an logarithmic scale. For example,

$$y = ae^{-bx} \tag{2.3}$$

is rewritten as

$$\ln y = \ln a - bx \tag{2.4}$$

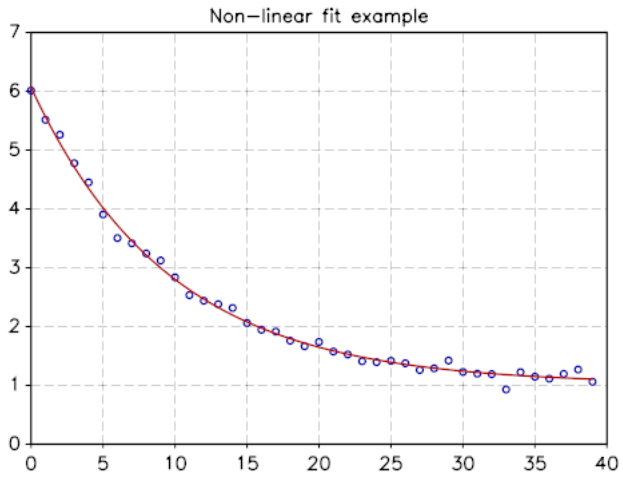


Figure 2.3: Non Linear Fitting

and then solved by the method of χ^2 minimization. Many of the non linear functions face the problem of flat minima or local minima. To overcome these problems, smarter, non-exhaustive methods are needed. Some such methods like Genetic Algorithm, Stimulated Annealing, Simplex etc. have already been discussed. It is to be noted that the parameters have certain constraints and the search space has to be constrained to values that fall within theoretical boundaries while estimating their values.

2.5 Noise and Measurement Errors

In science, the word error means the uncertainty. No measurement of any sort is complete without a consideration of this inherent error. Error is defined by Webster as the difference between an observed or calculated value and the true value. True value is usually not known, but the approximate value is guessed. Different experiments deal with different aspect of errors.

There are two main types of errors associated with an experimental re-

sult. They are referred to as **precision** and **accuracy**. The precision is usually related to the random error distribution associated with a particular experiment. The accuracy is related to the existence of systematic errors, for example, the incorrect calibration. The main objective of a good experiment is to improve both precision and accuracy.

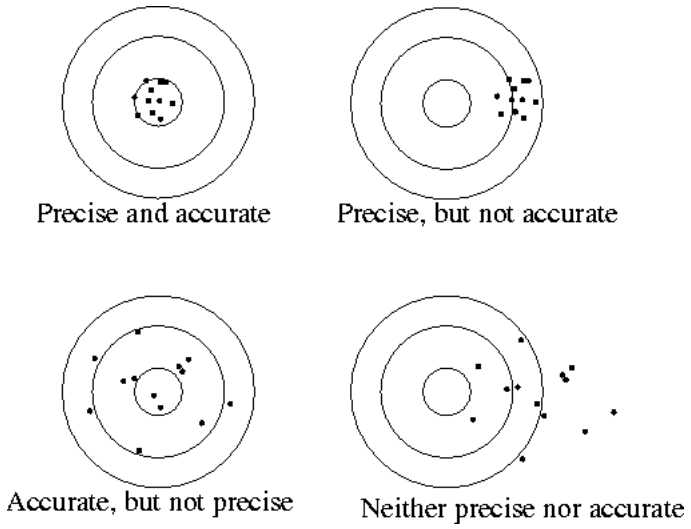


Figure 2.4: Accuracy v/s Precision

The sources of errors

Major sources of errors in any experiment are the Systematic Errors and the Random Errors. Random Error refers to the spread in the values of a physical quantity from one measurement of the quantity to the next, caused by random fluctuations in the measured value. This type of error affects the precision of the experiment. The uncertainty in any measured quantity has the same dimensions as the measured quantity. Such errors could be minimized by using statistical techniques. On the other hand, Systematic Error refers to an error which is present for every measurement of a given quantity;

it may be caused by a bias on the part of the experimenter, a miscalibrated or even faulty measuring instrument, etc. Systematic errors affect the accuracy of the experiment. They are repeatable (and therefore predictable), and are assumed to be time invariant. They can be characterized during the calibration process and mathematically reduced during measurements. They are never completely removed. There are always some residual errors due to limitations in the calibration process.

Significant Figures

The precision of an experimental result is implied by the number of digits recorded in the result. An uncertainty should not be stated with too much precision. The last significant figure in any stated answer should usually be of the same order of magnitude (in the same decimal position) as the uncertainty. However, the number of significant figures used in the calculation of the uncertainty should generally be kept with one more significant figure than the appropriate number of significant figures in order to reduce the inaccuracies introduced by rounding off numbers. After the calculations, the final answer should be rounded off to remove the extra figure.

Noise

Noise is any unwanted signal in the data. For example, while observing any astronomical object from ground using a photometer, the data received is not only from the source and it is composed of fluctuations due to the atmosphere. This noise usually follows Gaussian Distribution, and hence can be treated statistically. Hence, errors are also a type of noise in the actual data that is required.

Chapter 3

Particle Swarm Optimization

3.1 Introduction

Particle swarm optimization (PSO) is a population based stochastic optimization technique developed by Dr. Eberhart and Dr. Kennedy in 1995, inspired by social behavior of bird flocking or fish schooling. It is based on social relationships established among simple individuals in a group.

Every individual in the group is called as "particle". In PSO every particle holds the knowledge of its position, and every particle moves in the search space to find the optimal position. It is the fitness value of that particle. The best position achieved by every particle is also stored in its memory. This value is known as "pbest". Another best value that is tracked by the particle is the best value obtained so far by any particle in its neighborhood. This is called lbest. After having the knowledge of lbest, every particle updates its position and velocity and follows lbest. The comparison of different values of lbest and communication between the particles helps them to reach the final optimal value, known as the "gbest". PSO has many advantages, some of them being:

1. PSO is easy to implement and there are few parameters to adjust.

2. Particles update themselves with the internal velocity. They also have a memory, which is important to the algorithm.

3. PSO takes real numbers as particles. Some other methods involve conversion of numbers into binary form.

3.2 The Algorithm

As mentioned above, PSO simulates the behavior of bird flocking. PSO is initialized with a group of random particles and then it searches for optimum. In every iteration, each particle is updated by following two "best" values. The position is updated according to the Flight Formula:

$$x(t+1) = x(t) + v(t+1) \tag{3.1}$$

There are two approaches to update the velocity:

1. PSO with inertia weight:

   $$v_i(t+1) = w*v_i(t)+c_1*rand()*(x_{pbesti})+c_2*rand()*(x_{gbesti}-x_i) \tag{3.2}$$

   where w is the inertia weight, x_{pbesti} is the particles pbest, x_{gbesti} is the position of lbest, c_1 and c_2 are the acceleration constants (user-defined) which control the influence of memory of the particle and position of the lbest elements respectively, rand() is a function that generates a uniform-distributed random real number between 0 and 1.

2. PSO with constriction factor: The constriction factor k is included in the velocity update formula as follows:

   $$v_i(t+1) = k*[v_i(t)+c_1*rand()*(x_{pbesti}-x_i)+c_2*rand()*(x_{pbesti}-x_i)] \tag{3.3}$$

   where the constriction factor k is calculated by means of the acceleration constants c_1 and c_2

The Algrorithm can be written as:

1. Generate a swarm of random particles.
2. Calculate fitness value for every particle.
3. If the new fitness value is better than the best fitness value (pbest) in history, set present value as new pbest.
4. DO
   (a) Find lbest
   (b) For each particle, update its position with the flight formula. Calculate its new velocity.
   (c) Evaluate the fitness of the new position of each particle
   (d) Update the pbest (memory) value of each particle.
5. Until maximum iterations or minimum error criteria is not attained, do step 4.
6. Find gbest.

Thus, the parameters required for the PSO are: The number of particles, dimension of particles, range of particles, v_{max} (the maximum change one particle can take during one iteration), constants c_1 and c_2 usually which are usually equal to 2, the stop condition (the maximum number of iterations the PSO execute and the minimum error requirement) and inertia weight.

3.3 Test Functions

Any algorithm proposed for solving optimization problems should yield results similar to those obtained earlier, in addition to its ability to converge faster. There are a number of functions which are commonly used to test

this. Below is the list of some functions used in this project. 30 Dimensional space is considered.

1. **Dejong :** This function is also known as Spherical Function. It is continuous, strictly convex and unimodal. It is defined as:

$$f(x) = \sum_{i=1}^{n} x_i^2 \tag{3.4}$$

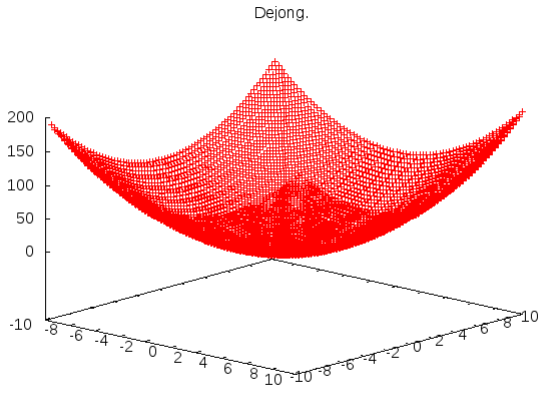


Figure 3.1: Dejong Function

2. **Rosenbrock:** It is defined as:

$$f(x) = \sum_{i=0}^{n-1} 100(x_{i+1} - x_i^2)^2 + (x_i - 1)^2 \tag{3.5}$$

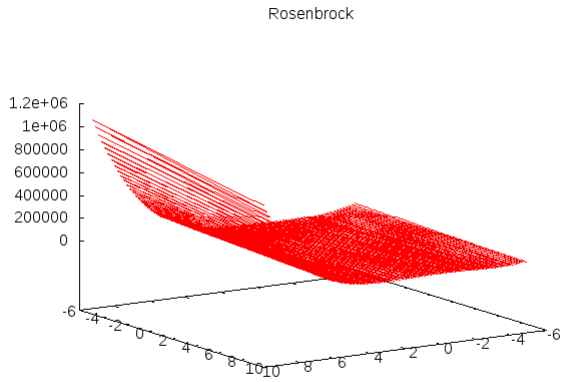


Figure 4.2: Rosenbrock Function

3. **Rastrigin:** This function is a highly multimodal function, and the degree of multimodality increases with the dimension of the problem. It is defined as:

$$f(x) = \sum_{i=1}^{n} (x_i^2 - 10cos(2\pi x_i) + 10) \tag{3.6}$$

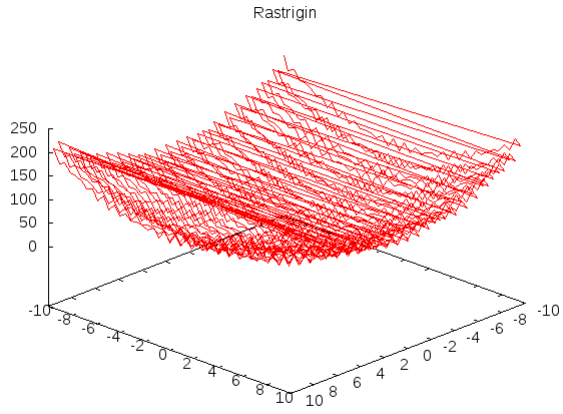


Figure 4.3: Rastrigin Function

4. **Griewank:** This is also a highly modal function and has several local minima. It is defined as:

$$f(x) = \frac{1}{4000}\sum_{i=0}^{n-1} x_i^2 + \sum_{i=0}^{n-1} cos(\frac{x_i}{\sqrt{i+1}}) + 1 \tag{3.7}$$

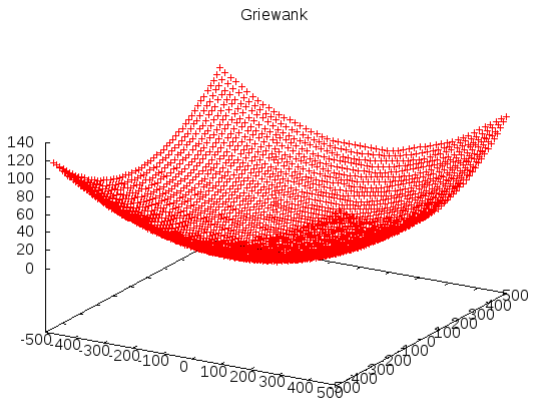


Figure 4.4: Griewank Function

5. **Schwefel:** This function too, is multimodal. It is defined as:

$$f(x) = -\sum_{i=1}^{n} x_i sin\sqrt{\mid x_i \mid} \tag{3.8}$$

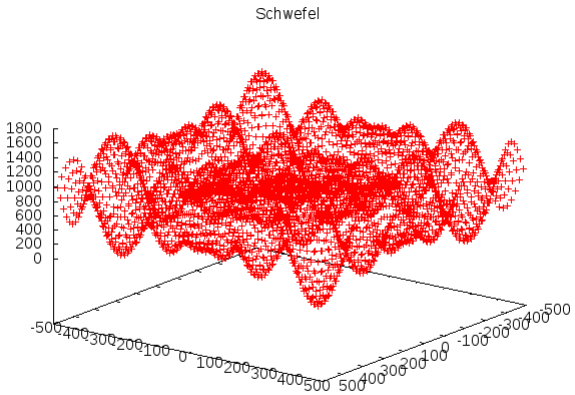


Figure 4.5: Schwefel Function

6. **Ackley:** This is also a multimodal function, defined as:

$$f(x) = 20 + e^{-20} exp(-0.2\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}) - exp(\frac{1}{n}\sum_{i=1} ncos(2\pi x_i)) \quad (3.9)$$

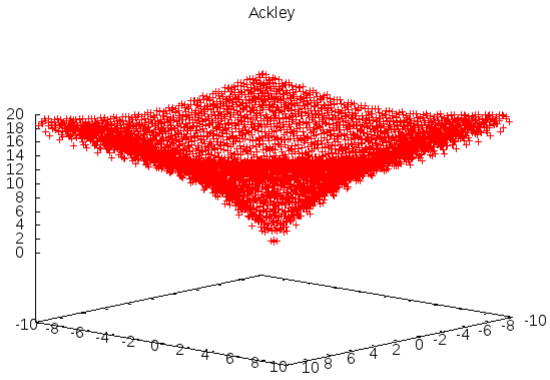


Figure 4.6: Ackley Function

7. **Levy:** It is a multimodal function. It is defined as:

$$f(x) = sin^2(\pi w_1) + \sum_{i=1}^{n-1}(w_i - 1)^2[1 + 10sin^2(\pi w_i + 1)] + (w_n - 1)^2[1 + sin^2(2\pi w_n)] \tag{3.10}$$

where, $w_i = 1 + \frac{x_i - 1}{4}$, for all i = 1,2,....,n

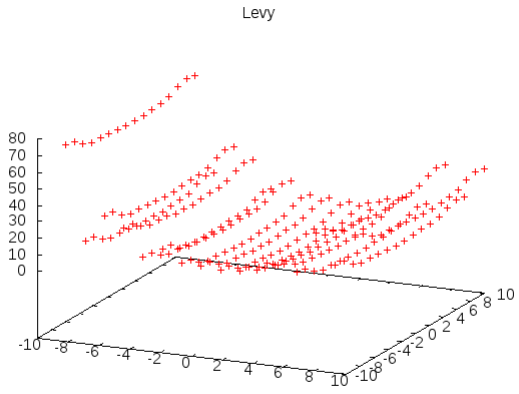


Figure 4.7: Levy Function

8. **Zakharov:** This function does not have local minima, and has only one global minima. It is defined as:

$$f(x) = \sum_{i=1}^{n} x_i^2 + (\sum_{i=1}^{n} 0.5ix_i)^2 + (\sum_{i=1}^{n} 0.5ix_i)^4 \tag{3.11}$$

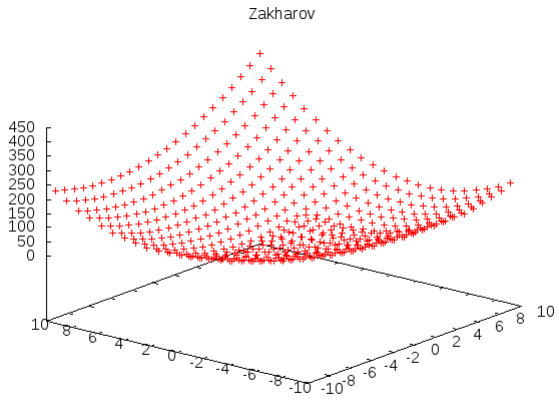


Figure 4.8: Zakharov Function

9. **Styblinski - Tang:** It is defined as:

$$f(x) = \frac{1}{2}\sum_{i=1}^{n}(x_i^4 - 16x_i^2 + 5x_i) \tag{3.12}$$

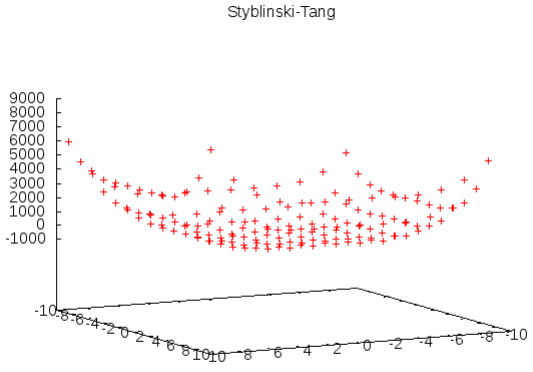


Figure 4.9: Styblinski-Tang Function

10. **Alpine:** This function has several local extrema and only one global extremum. It is defined as:

$$f(x) = \prod_{i=1}^{n} sin(x_i) \sqrt{\prod_{i=1}^{n} x_i} \tag{3.13}$$

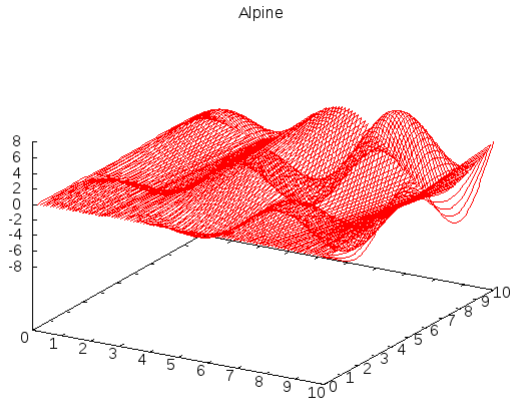


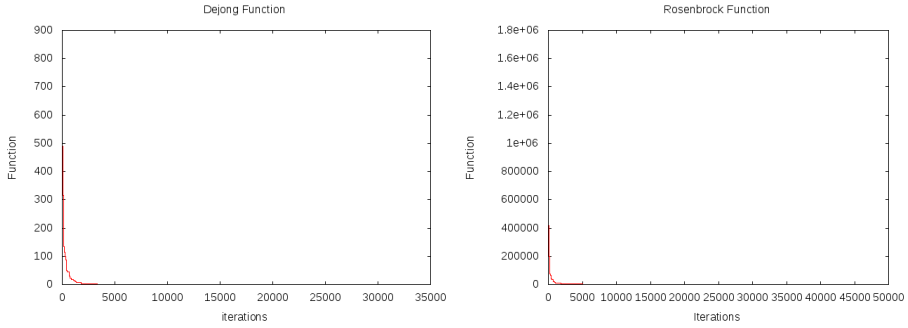
Figure 4.10: Alpine Function

Sr. No.	Function	Equation	Dimensions
1	Dejong	$f(x) = \sum_{i=1}^{n} x_i^2$	30
2	Rosenbrock	$f(x) = \sum_{i=0}^{n-1} 100(x_{i+1} - x_i^2)^2 + (x_i - 1)^2$	30
3	Rastrigin	$f(x) = \sum_{i=1}^{n}(x_i^2 - 10cos(2\pi x_i) + 10)$	30
4	Griewank	$f(x) = \frac{1}{4000}\sum_{i=0}^{n-1} x_i^2 + \sum_{i=0}^{n-1} cos(\frac{x_i}{\sqrt{i+1}}) + 1$	30
5	Schwefel	$f(x) = -\sum_{i=1}^{n} x_i sin\sqrt{\mid x_i \mid}$	30
6	Ackley	$f(x) = -\sum_{i=1}^{n} x_i sin\sqrt{\mid x_i \mid}$	30
7	Levy	$f(x) = sin^2(\pi w_1) + \sum_{i=1}^{n-1}(w_i - 1)^2[1 + 10sin^2(\pi w_i + 1)] + (w_n - 1)^2[1 + sin^2(2\pi w_n)]$	30
8	Zakharov	$f(x) = \sum_{i=1}^{n} x_i^2 + (\sum_{i=1}^{n} 0.5ix_i)^2 + (\sum_{i=1}^{n} 0.5ix_i)^4$	30
9	Styblinski Tang	$f(x) = \frac{1}{2}\sum_{i=1}^{n}(x_i^4 - 16x_i^2 + 5x_i)$	30
10	Alpine	$f(x) = \prod_{i=1}^{n} sin(x_i)\sqrt{\prod_{i=1}^{n} x_i}$	30

Table 4.1: Test Functions

3.4 PSO Results

The number of particles considered is 30 and the number of iterations is 50,000. For this set of parameters, PSO yielded the following results:



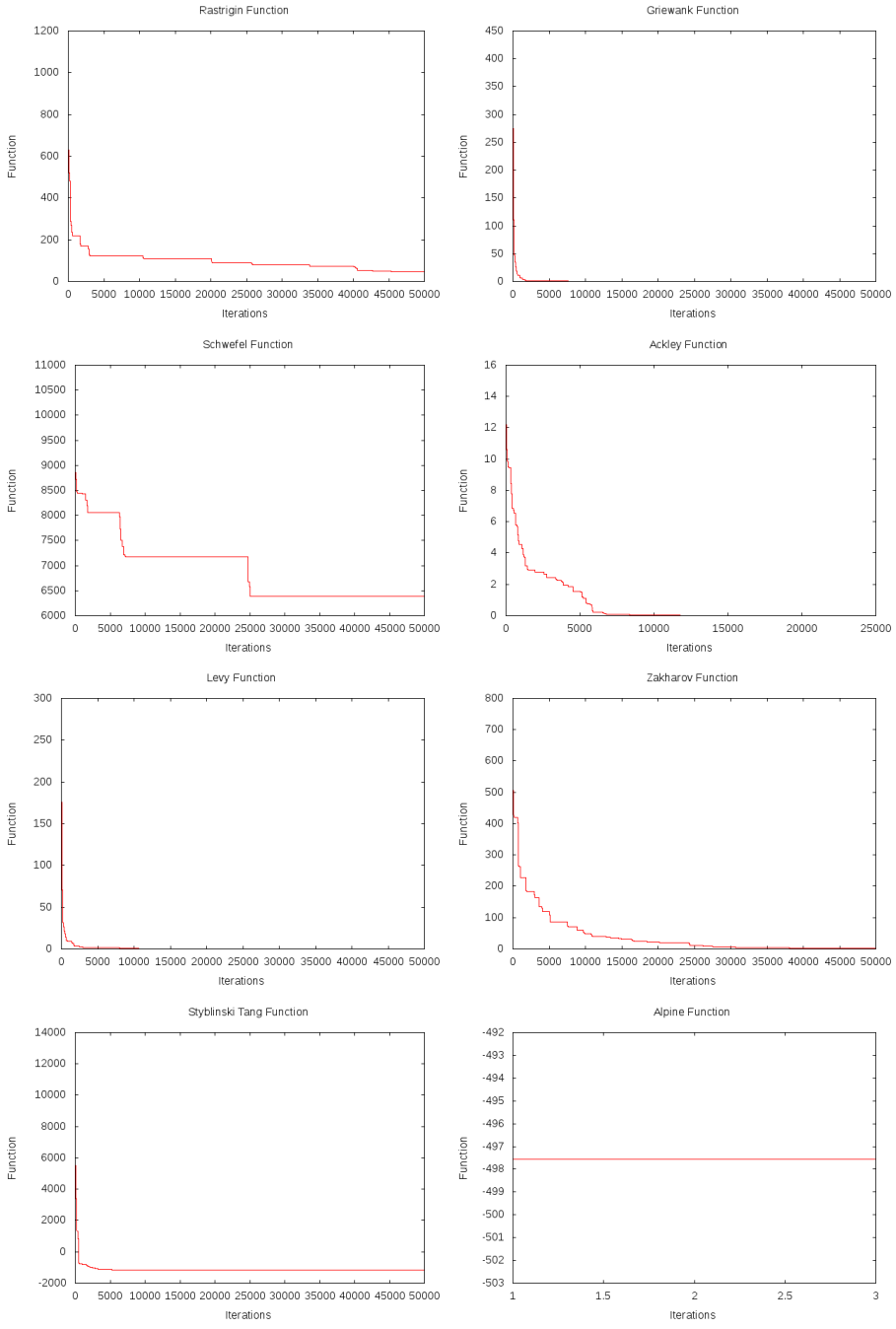


Figure: 4.11: PSO results for 30 particles;

50,000 iterations; $c_1 = c_2 = 2.0, w = 0.72$;
for the functions mentioned in the table respectively.

For the set of parameters: particles = 30, iterations = 50,000, $c_1 = c_2 = 2.0$, w = 0.72, PSO yielded the following results:

Sr. No.	**Function**	**Convergence after no. of steps**
1	Dejong	2083
2	Rosenbrock	5000
3	Rastrigin	> 50000
4	Griewank	7083
5	Schwefel	> 50000
6	Ackley	12000
7	Levy	10400
8	Zakharov	> 50000
9	Styblinski Tang	> 50000
10	Alpine	3

Table 4.2: PSO Results

According to Off-the-Shelf PSO by Anthony Carlisle and Gerry Dozier, [3], the recommended settings for good PSO Results are: a population size of 30, with $c_1 = 2.8$ and $c_2 = 1.3$. The neighborhood size for a swarm is the number of neighboring particles that influence a particular particles movement. According to the initial works of Kennedy and Eberhart, a small neighbourhood is better at avoiding local minima, and that a global neighborhood converges faster. However, Anthony Carlisle and Gerry Dozier in their work have shown that neighbourhood of size of the population, i.e the global neighbourhood is a better choice as it requires lesser amount of work to be done for same results. As far as updating the velocity of the particles

is concerned, for problems where the position of the particle is considered, when a particle reaches x_{max}, set its velocity to zero. If x_{max} is not enforced, v_{max} need not be enforced, either. [3]

Chapter 4

Cosmic Microwvae Background

Chapter 5

CMB parameter estimation

Bibliography

[1] Press, William H.; Flannery, Brian P.; Teukolsky, Saul A., *Numerical recipes. The art of scientific computing*, Cambridge: University Press (1986).

[2] Bevington, Philip R., *Data reduction and error analysis for the physical sciences*, New York: McGraw-Hill (1969).

[3] Anthony Carlisle and Gerry Dozier, An Off-The-Shelf PSO, PSO Workshop, IUPUI, 2001

[4] Data Reduction and Error Analysis for the Physical Sciences (Third Edition). - Philip R. Bevington and D. Keith Robinson

[5] Computational Physics: An Introduction - R. C. Verma

[6] Press-Tseukolsky-Vetterling-Flannery: Numerical Recipes, 3rd Edition
http://www.it.uom.gr/teaching/linearalgebra/NumericalRecipiesInC/c10-4.pdf

[7] mat.iitm.ac.in/home/sryedida/public_html/caimna/transcendental/bracketing%20methods/bisection/bisection.html

[8] http://katrinaeg.com/simulated-annealing.html

[9] http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol1/hmw/article1.html#introduction

[10] http://www.swarmintelligence.org/tutorials.php