Adv. Studies Theor. Phys., Vol. 8, 2014, no. 1, 1 - 9
HIKARI Ltd, www.m-hikari.com
http://dx.doi.org/10.12988/astp.2014.311129

Modeling Hadronic Collisions Using Genetic Programming Approach

Esraa El-Khateeb, Amr Radi*, Salah Yaseen El-Bakry

Physics Department, Faculty of Science, Ain Shams University
Abbassia 11566, Cairo, Egypt
*British University in Egypt, and CMS Egyptian team leader

Mahmoud Yaseen El-Bakry

Physics department, Faculty of Education
Ain Shams University, Roxy, Cairo, Egypt

Abstract

New technique, Genetic Programming, is presented for modeling total cross section of both pp and $\bar{p}p$ collisions from low to high energy regions. Recent total cross section data are taken from Particle Data Group and LHC collaboration. The model seems to fit the experimental data well.

Keywords: Hadronic collisions; total cross section; pp collisions; $\bar{p}p$ collisions; genetic programming

1 Introduction

It is a well known fact that high energy total hadronic cross sections grow with rising center of mass energy. Increase of the pp total cross section with energy has been discovered since the first results of the Intersecting Storage Ring (ISR) at CERN in the 70s [1, 2]. The CERN $S\bar{p}pS$ Collider found this rising valid for the $\bar{p}p$ total cross section as well [9], and this was also confirmed at Fermilab accelerator [10].

From different hadronic processes, the pp and $\bar{p}p$ are of special interest. This is because they accumulate the most precise and energy broader set of data. Besides, the pp set is the only one including the cosmic ray informations on total cross section.

Recently, TOTEM collaboration at LHC announced that they could use a luminosity-independent method to measure the pp total collision cross-section at $\sqrt{s} = 8\ TeV$. A total pp cross-section of $(101.7 \pm 2.9)\ mb$ has been determined[3]. They were also announced in a previous work that they could determine the total pp cross-section at $\sqrt{s} = 7\ TeV$ as $(98.3 \pm 3.0)\ mb$[4, 5]. On the other hand, data on $\bar{p}p$ total scattering cross section extends up to $\sim 1.8\ TeV$. At higher energy intervals, data for pp collisions , up to$\sqrt{s} \sim 40\ TeV$, may be inferred from Cosmic Ray experiments. However, some disagreements exist among different experiments. These discrepancies are mainly a result of the strong model-dependence of the relation between the basic hadron-hadron cross section and the hadronic cross section in air. The latter determines the attenuation length of hadrons in the atmosphere, which is usually measured in different ways, and depends strongly on the rate of energy dissipation of the primary proton into the electromagnetic shower observed in the experiment; such a cascade is simulated by different Monte Carlo techniques implying additional discrepancies between different experiments, [8, 12, 14, 15].

However, the actual energy dependence of the total hadronic cross section is still an open question of intense theoretical interest. Variety of models, theoretical, empirical and semi-empirical has been established to study the subject. Recently, Genetic programming (GP) has been one of researchers interests in modeling of high energy physics as well as in different fields (see for example [6]-[13]). Genetic programming is one of a number of machine learning techniques in which a computer program is given the elements of possible solutions to the problem. This technique, through a feedback mechanism, attempts to discover the best solution, a function, to the problem at hand, based on the programmers definition of success. The Genetic programming framework creates a program which consists of a series of linked nodes. Each node takes a number of arguments and supplies a single return value. There are two general types of nodes: functions (or operators) and terminals (variables and constants) [18]. The series of linked nodes can be represented as a tree where the leaves of the tree represent terminals and operators reside at the forks of the tree. The tree elements are called nodes. The functions (F) have one or more inputs and produce a single output value. These provide the internal nodes in expression trees. The terminals (T) represent external inputs, constants and zero argument functions.

The aim of this work is to use the Genatic Programming (GP) technique, to discover the functions that describe and interpolate accelerator pp and $\bar{p}p$ total cross section data and to extrapolate to cosmic ray available data. The data base analyzed and compiled by the Particle Data Group (PDG) has become a standard reference and a corresponding readable files are available [7]. In section II modeling with GP is discussed, while the proposed GP is given in section III. Results are given in section IV and conclusions are presented in section V.

2 Modeling Using Genetic Programming

GP, evolves a population of computer programs, which are possible solution to a given optimization problem , using the Darwinian principle of survival of the fittest. It uses biologically inspired operations like reproduction, crossover and mutation. Each program or individual on the population is generally represented as a tree composed of functions $(*, +)$ and terminals (x, y) appropriate to the problem domain. For example, Fig. 1 shows the representation of the function $+(*(x, y), *(x, *(x, y)))i.e.((x*(x*y)) + (x*y))$. To read trees in this fashion, one resolves the sub-trees in a bottom-up fashion, where $F = (*, +)$ and $T = (x, y)$. The set of functions and set of terminals/inputs must satisfy the closure and sufficiency properties. The closure property demands that the function set is well defined and closed for any combination of arguments that it may encounter. On the other hand, the sufficiency property requires that the set of functions and the set of terminals be able to express a solution of problem. The function set may contain standard arithmetic operators, mathematical functions, logical operators, and domain-specific functions. The terminal set usually consists of feature variables and constants. Each individual in the population is assigned a fitness value, which quantifies how well it performs in the problem environment. The fitness value is computed by a problem dependent fitness function

A typical implementation of GP (i.e. the process of determining the best (or nearly best) solution to a problem in GP) involves the following steps:
1) GP begins with a randomly generated initial population of solutions.
2) A fitness value is assigned to each solution of the populations.
3) A genetic operator is selected probabilistically:
Case i) If it is the reproduction operator, then an individual is selected (we use fitness proportion-based selection) from the current population and it is copied into the new population. Reproduction replicates the principle of natural selection and survival of the fittest.

Case ii) If it is the crossover operator, then two individuals are selected. We use tournament selection where number of individuals is taken randomly from the current population, and out of these, the best two individuals (in terms of fitness value) are chosen for the crossover operation. Then, we randomly select a sub tree from each of the selected individuals and interchange these two sub-

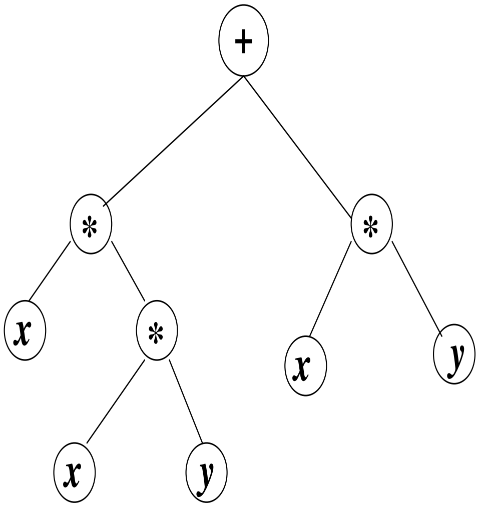

Figure 1: Tree representation of the equation $+(*(x,y),*(x,*(x,y)))i.e.((x*(x*y))+(x*y))$.

trees. These two offspring are included in the new population. Crossover plays a vital role in the evolutionary process.

Case iii) if the selected operator is mutation, then a solution is (randomly) selected. Now, a sub-tree of the selected individual is randomly selected and replaced by a new randomly generated sub-tree. This mutated solution is allowed to survive in the new population. Mutation maintains diversity.
4) Continue step 3), until the new population gets solutions. This completes one generation.
5)Unlike genetic algorithm [16], GP will not converge.
So, steps 2)-4) are repeated till a desired solution (may be 100% correct solution) is achieved. Otherwise, terminate the GP operation after a predefined number of generations.

3 The Proposed Genetic Programming

Our approach is to use the experimental data pp and $\bar{p}p$ total cross sections to produce calculated total cross sections for each of them. The center of mass energy is used as input variable to find the suitable function $\sigma_t(\sqrt{s})$, that describes the experimental data. Our representation, the fitness function, is calculated as a negative value of the total absolute performance error of the discovered function on the experimental data set, i.e. a lower error must correspond to a higher fitness. The total performance error can be defined for all the experimental data $(i = 1, ..., n)$ set as:

$$E = \sum_{j=1}^{n} |X_j - Y_j|^2 \tag{1}$$

Where X_j represent the experimental data for element j and Y_j represent the calculated data for element j. The running process stops when the error E is reduced to an acceptable level (0.00001).
To find $\sigma_t(\sqrt{s})$ for pp and $\bar{p}p$ GP was run for 800 generations with a maximum population size of 1000. The operators (and selection probability) were: crossover with probability 0.9 and mutation with probability 0.01. The function set is $(+, -, *, /, \ln)$, and the terminal set is random constant from 0 to 10, the incident center of mass energy. the "full". initialization method was used with an initial maximum depth of 27, and tournament selection with a tournament size of 8. The GP was run until the fitness function is reduced to an acceptable level (0.00001). The discovered function has been tested to associate the input patterns to the target output patterns using the error function.

4 Results and Discussion

The final discovered function σ_t for describing the pp total cross section at low and high energies is

$$\sigma_t^{pp}(\sqrt{s}) = \frac{Z}{f_1 - f_2 + 0.94209} + \frac{\sqrt{s}}{2.252914 * (f_3 + u * f_4)} + [u * (\ln(u) + 3.801376)] \quad (2)$$

where $u = 6.6758$ and

$$Z = \ln\left(\sqrt{s}\right) - 10 + \ln\left(\frac{10}{s - 10 - \ln(0.66758)} - \ln\left(0.76524 + \frac{0.79057}{\ln(\ln(\sqrt{s}))} + \frac{10}{\sqrt{s}} + 10\right)\right) \quad (3)$$

while

$$f_1 = \ln\left(\ln\left(\sqrt{s}\right) - 0.52427\right) - \frac{\sqrt{s}}{f_5/f_6} \quad (4)$$

$$f_2 = \ln\left(\ln\left(\sqrt{s}\right) * \left(0.20971 - \sqrt{s}\right)\right) \quad (5)$$

$$f_3 = \frac{f_7}{\ln(\sqrt{s}) - \sqrt{s}/f_8 - f_9} + \frac{\sqrt{s}}{21.3679 * \ln(\sqrt{s})} \quad (6)$$

$$f_4 = \ln(u) - \ln(1.06528) \quad (7)$$

$$f_5 = -\left[1.309879 * \left(\frac{0.9938}{s * (0.32376 - \sqrt{s})}\right)\right] - 10 \quad (8)$$

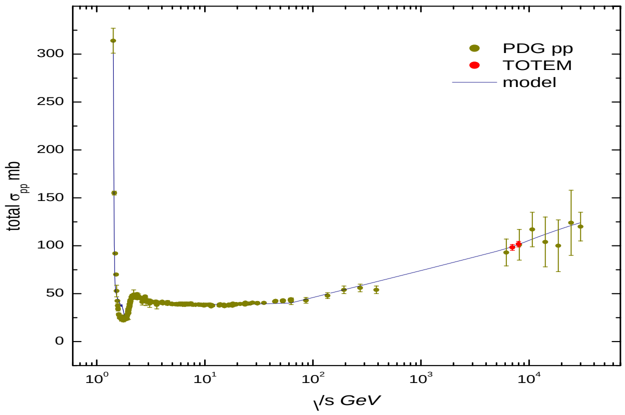

Figure 2: A comparison between the experimental data and the model for *pp* total cross section. TOTEM recent measures are shown in red.

$$f_6 = \ln\left(\sqrt{s}\right) * \left(0.20971 - \sqrt{s}\right) \tag{9}$$

$$f_7 = \ln\left(\frac{f_{10}}{0.92544}\right) - 10.235 \tag{10}$$

$$f_8 = \frac{-9.53557}{\ln(u) * (0.20971 - \sqrt{s})} \tag{11}$$

$$f_9 = \frac{0.2234}{\ln(\sqrt{s}) * 21.3679} - 0.94209 \tag{12}$$

$$f_{10} = \left(\sqrt{s} - 0.23495\right) * \left(\frac{0.76524 - \sqrt{s}}{\ln(10) - 10}\right) \tag{13}$$

Fig.(2) shows a comparison between the model and the experimental data. The function of $\bar{p}p$ is

$$\sigma_t^{\bar{p}p}(\sqrt{s}) = \frac{a * \sqrt{s} * ln(\sqrt{s})}{(b * \sqrt{s} + c)} + \frac{d}{ln(e - \sqrt{s})} \tag{14}$$

where
$a = 3.4055$, $b = 0.3975$, $c = -0.4659$, $d = 48.4646$, $e = 0.2872$

Fig.(3) shows how much the calculated data match the experimental one.

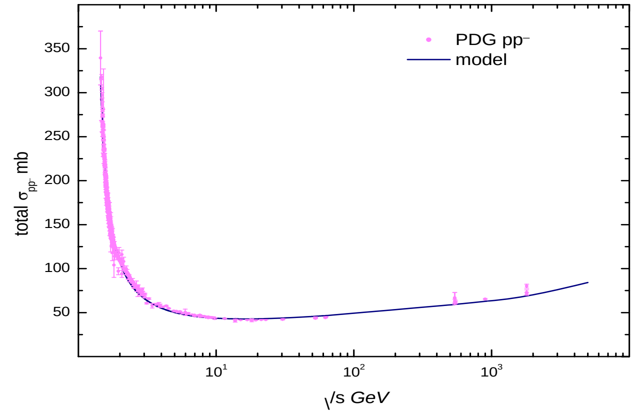

Figure 3: The experimental data compared to the model for $\bar{p}p$ total collision cross section.

In order to measure the goodness of fit of our model we use a combination of the Pearson correlation coefficient[20, 22] r^2, and the Mean Absolute Deviation[11], MAD. the Pearson correlation coefficient r^2 is used to measure to what extent the model predicts data obey the trend of the experimental one, and is given by

$$r^2 = 1 - \frac{\sum_i \left(\left(y_{mod_i} - y_{obs_i}\right)/err\right)^2}{\sum_i \left(\left(y_{obs_i} - \bar{y}_{obs}\right)/err\right)^2} \tag{15}$$

The Mean Absolute Deviation, MAD, is a conceptually easier understand measure of the deviation of the model predictions from exact data location. It is the mean of the absolute value of the deviation between each model prediction and its corresponding data point.

$$MAD = \frac{\sum_i \left|y_{mod_i} - y_{obs_i}\right|}{N} \tag{16}$$

Table(1) shows a summary for the values of these variables.

Table 1: Statistical Calculations.

Collision	r^2	MAD
pp	0.915	2.4
$\bar{p}p$	0.993	5.69

We can see from the table that the the GP approach gives results which can match the experimental data to a good extent from low to ultra high values of energy, and this occurs for the trained and untrained observations of the total cross sections of both pp and $\bar{p}p$.

5 Conclusions

Genetic programming (GP) method is one of a number of machine learning techniques in which a computer program is given the elements of possible solution to the problem (in our case center of mass energy) and attempts, through a feedback mechanism, to discover the best function (in our case the total cross section) to the problem at hand, based on the programmers definition of success. The program consists of a series of linked nodes which can be represented as a tree. Each node takes a number of arguments and supplies a single returned value. There are two general types of nodes (or tree elements): functions(or operators) such as $(*, +, -, /, exp, log, \ln, sin, cos, sqr)$ and terminals (constants and variables) such as random constant from 0 to 10, the energy . The GP model seeks to imitate the biological processes of evolution, treating a tree or program as an organism. Through natural selection and reproduction over a number of generations, the fitness of a population of organisms is improved.

Finally, the present work presents a new technique for modeling the total cross sections of pp and $\bar{p}p$ based on GP technique.The discovered functions show a good match to the experimental data. Moreover, the discovered functions are capable of predicting experimental data for the total cross sections that are not used in the training session.

References

[1] U. Amaldi et al., *Phys. Lett.,* **B44** (1973), 11.

[2] S.R. Amendolia et al., *Phys. Lett.,* **B44** (1973), 119.

[3] G. Antchev et al., TOTEM-2012-005; CERN-PH-EP-2012-354, (2013).

[4] G. Antchev et al., *Europhys. Lett.,* **96** (2011), 21002.

[5] G. Antchev et al., (TOTEM Collaboration), *Europhys. Lett.* **101** (2013), 21004.

[6] W. Banzhaf and P. Nordin and R. Keller and F. Francone, *Genetic Programming An Introduction: On the Automatic Evolution of Computer Programs and Its Applications*, Morgan Kaufmann Publishers, Inc. San Francisco, California,1998.

[7] J. Beringer et al. (Particle Data Group), *Phys.Rev. D* **86**, 010001 (2012), The full data sets are available at http://pdg.lbl.gov/2013/hadronic-xsections.

[8] M.M. Block, *Phys. Rev. D,* **36** (2006), 1350.

[9] M. Bozzo et al., *Phys. Lett.,* **B147** (1984), 392.

[10] A.S. Carroll et al., *Phys. Lett.,* **61** (1976), 303.

[11] C.D. Schunn and D. Wallach *Evaluating Goodness of Fit in Comparison of Model to Data,* electronic version (2005), www.lrdc.pitt.edu/schunn/gof.

[12] R. Engel, T.K. Gaisser,P. Lipari and T. Stanev, *Phys. Rev. D,* **58** (1998), 014019.

[13] M Y El-Bakry and A Radi, Int. J. Mod. Phys C, **17** (2006), issue 8.

[14] T.K. Gaisser, *Nucl. Phys.* **12** (1990), 172.

[15] T.K. Gaisser, U.P. Sukhatme and G.B. Yodh, *Phys. Rev. D,* **36** (1987), 1350.

[16] D.E. Goldberg, *Genetic Algorithm in Search, Optimization and Machine Learning*, Addison-Wesley, 1989.

[17] J. Koza, M. Keane, M. Streeter, W. Mydlowec, J. Yu, *Genetic Programming IV: Routine Human-Competitive Machine Intelligence*, Guido Lanza - Computers 2005.

[18] R. Poli, *Genetic Programming for Feature Detection and Image Segmentation,*"Dec" School of Computer Science, University of Birmingham,1995.

[19] R. Riolo, *Genetic Programming Theory and Practice*, Bill Worzel Computers, 2003.

[20] J.L. Rodgers and W.A. Nicewander, *The American Statisticians,* **42** (1988), no. 1.

[21] K. Sharman and A. Alcazar and Y. Li, *Evolving Signal Processing Algorithms by Genetic Programming, First International Conference on Genetic Algorithms in Engineering Systems: Innovations and Applications* (GALESIA'95), 473-480, 1995.

[22] S.M. Stigler, *Statistical Science,* **4** (1989), no. 2.

Received: November 5, 2013