Machine Learning in Physics and Astronomy: Particles Collision Simulation

Reza Shokrzad
s1056369
(Dated: August 8, 2022)

The project includes two main phases sampling data with the Monte Carlo (MC) approach, then generating fake data with a generative machine learning model. This generating task is to simulate a physics event in which two sub-atomic electrons collide with each other and bread two new other particles. I sample 100000 data with MC by considering physics-related rules. Afterward, I train a VAE model to generate fake data. The result indicates that we can have more data without doing any practical experiments in a realistic situation like what happens in the Large Hadron Collider (LHC). This experiment is so important for physicians to infer phenomena like particle collision based on more data.

I. PROBLEM DESCRIPTION

A. Particle Collisions

The objective of the assignment is to simulate a collision of two subatomic particles (an electron e^- and a positron e^+) and interpret the possible outcomes. To simplify the process, the electron and positron are mass-less and have the same energy, for instance 200 GeV. Also, after the collision, they turn into two new particles x_1 and x_2, which have the same mass (e.g. 114 GeV). So, the event is like $e^+e^- \rightarrow x_1x_2$ that is shown in figure 1. By considering $c = 1$, it is also supposed to simulate what happens to the energy and momentum of generated particles.

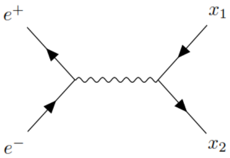


FIG. 1: Electron-positron Collision

The first step of task is to produce 100000 instances of x_1 and x_2 concluded from the event, including their features and final states with degeneracy in the entries (E, p_x, p_y, p_z, m). The Large Hadron Collider (LHC) has a gigantic detector installed in the intersection of circles to record the momentum-energy values. The device estimates the energy and angles of new particles not accurately. Thus, it is assumed that it measures the energy and angels by 10% and 5% deviation, respectively.

B. Monte Carlo Sampling

The Monte Carlo experiment is a repeated arbitrary sampling to create numerical outcomes. The method is widely used in physics-related problems to generate draws from a probability distribution. Since the process is law-based with determined formula (equation 1) and boundaries, the approach is practical in particle collision simulation.

$$E^2 = M^2 + P^2, \tag{1}$$

where E, M, and P stand for energy, mass, and momentum of a particle, respectively.

C. Generative Model

Variational Autoencoders (VAEs) are associated with autoencoder models with two main sections, encoder, and decoder. Figure 2 shows the architecture of VAEs, which is fed by a parametrized distribution as the input data and the encoder and decoder blocks are trained jointly with the objective of minimization of the reconstruction error.

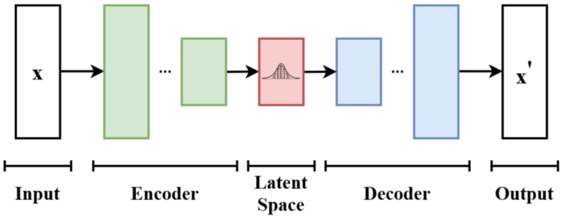


FIG. 2: The architecture of VAE models.

The model estimates the input data distribution in the encoder section to make the latent space, then samples from this space to reconstruct some new data with the same distribution.

II. METHODOLOGY

The simulation phase has two different parts. First, Monte Carlo sampling is applied to generate 100000 instances which take around 10 minutes in practice. Then, a generative model like VAE is fed by made dataset to simulate a distinct dataset with the same features. Last, the performance of simulators will be compared. As it is assumed that initial particles are mass-less and their energy is similar at the level of 200

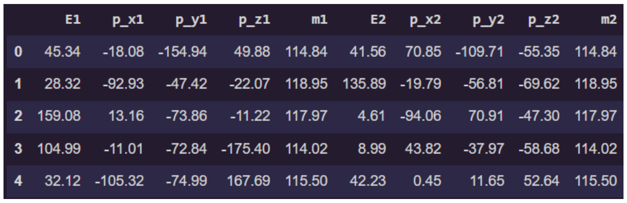
GeV, simulation of new parameters is started by MC sampling from a uniform distribution for momentum. Afterward, energy is calculated using the equation 2 to keep the system conserved. In the equation, E_i refers to the initial energy of electron and positron, P_1 and P_2 are the total momentum of new particles and E_r is the rest of the energy that is split into two parts of E_1 and E_2, randomly, later.

$$E_r = \sqrt{2 \times (E_i^2 - M^2) - P_1^2 - P_2^2}. \tag{2}$$

At the second stage, converting Cartesian to Spherical coordinate, smearing the angels of θ and ϕ there, then inverse converse to Cartesian was done to shift 5% of the momentum. So, angels are smeared by a Gaussian distribution with ($\mu = 1, \sigma = 0.05$). Additionally, energy is shifted with a G($\mu = 1, \sigma = 0.1$) setup that leads to 10% error. Afterward, mass is recalculated to keep the equation 3 conserved for having the right dataset. M_r is the rest of the mass after smearing that should be halved for x_1, x_2. Also, E_{rs} stands for the rest of the energy after smearing.

$$M_r = \sqrt{2 \times (E_i^2) - (E_{rs}^2 + P_1^2 + P_2^2)/2}. \tag{3}$$

Followingly, the dataset with 10 features would look like Figure 3 which shows the head and description of the data frame.



	E1	p_x1	p_y1	p_z1	m1	E2	p_x2	p_y2	p_z2	m2
0	45.34	-18.08	-154.94	49.88	114.84	41.56	70.85	-109.71	-55.35	114.84
1	28.32	-92.93	-47.42	-22.07	118.95	135.89	-19.79	-56.81	-69.62	118.95
2	159.08	13.16	-73.86	-11.22	117.97	4.61	-94.06	70.91	-47.30	117.97
3	104.99	-11.01	-72.84	-175.40	114.02	8.99	43.82	-37.97	-58.68	114.02
4	32.12	-105.32	-74.99	167.69	115.50	42.23	0.45	11.65	52.64	115.50

FIG. 3: Generated dataset by Monte Carlo Sampling approach

Since providing events of particle collisions is a costly process, in reality, physicians are trying to find alternatives to have more data for events without doing more practical experiments. So, by having a real dataset, generating fake data with similar distribution to real ones can be a solution for this concern. In this regard, machine learning could answer this need. Among generative models, two of them, called GANs and VAE, are more popular due to their promising performance. In this project, VAE is the one that is applied in the second phase of the task. Since the number of features is so few to use a deep convolution network, the selected setup for encoder and decoder is a fully connected shallow one. The loss consists of two parts, including MSE, which computes the difference between real and fake values, added by the Kullback-Leiber Divergence term for discrete probability distributions (equation 4).

$$L = MSE + D_{KL}(P||Q) = MSE + \sum_{x \in X} P(x) log(\frac{P(x)}{Q(x)}), \tag{4}$$

where P(x) and Q(x) are the distribution of real and fake data. Besides, the experimental setup is Adam optimizer, learning rate equals 1×10^{-3}, the number of epochs 300, the dimension of the layers of the encoder are 50 and 20, and these dimensions for the decoder are 20 and 50. I receive a better result on the latent space 2dimensional. As merit of checking the relativistic relation between real and fake setup I apply the Kullback-Leiber Divergence function that is estimate the difference between two distributions.

III. RESULTS AND DISCUSSION

In this section, we first discuss the distributions of generated parameters by the Monte Carlo approach. Then we will compare them to the outcomes of the generative model. I considered that the system has $2 * 200$ GeV of energy. To divide this energy into energy, mass, and momentum of new particles, I started with sampling from three uniform distributions in the range [-400, 400] for each component of momentum by imposing the Monte Carlo approach. Due to sampling uniformly for elements of $\vec{P}$, their distributions should be Normal by a mean around zero (Figure 4). Because the probability of generated components that can fulfill the equation 1 is higher when all values are nearer to zero. To clarify, once one of the components is sampled nearer the ±400, the chance of staying the system conserved would decrease.

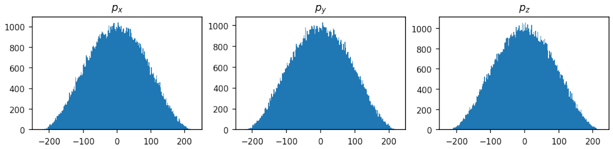


FIG. 4: Momentum samples distribution obtained from Monte Carlo approach

Figure 5 shows the energy distribution of x_1, x_2 before smearing. Since initial energy would convert to generated particles' mass, momentum, and energy, it is clear that the most possible collision outcome has energy less than 200, even the value would be closer to zero because of momentum share.

It can be expected that imposing a 10% error on energy would not change the original distribution form considerably. Apart from the extremes, Figure 6 approves that energy distribution would not be changed after shifting its values. It can be interpretable that without a significant change in the mean, the beginning and end of the interval are influenced more. More data is focused on the beginning because energy can not be negative, while fewer data at the tail but with larger value.

Figure 7 shows the distribution of ϕ in the left and θ in the right before smearing.

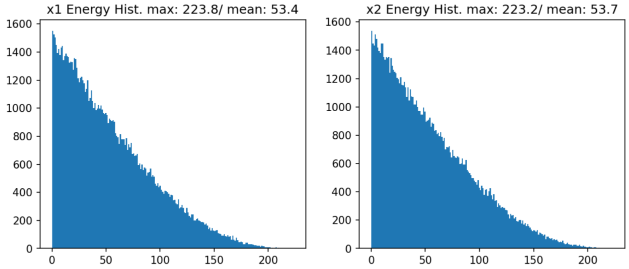


FIG. 5: energy distribution of x_1, x_2 before smearing

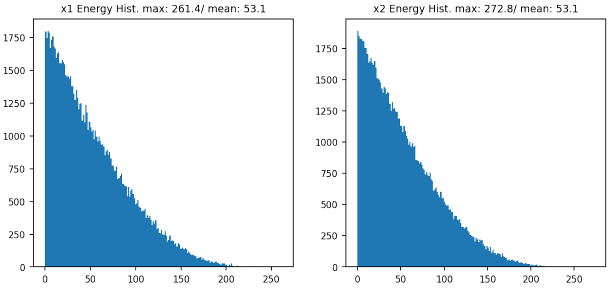


FIG. 6: energy distribution of x_1, x_2 after smearing

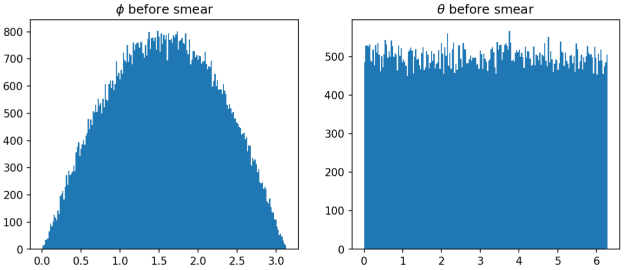


FIG. 7: Distribution of ϕ (left) and θ (right) before smearing.

Since the distributions of momentum components are similarly Gaussian, and we are trimming the values above π and below zero, it is natural that the distribution of ϕ would be Gaussian without tails. According to the definition of ϕ in spherical coordinate, all values are in the interval $[0, \pi]$. As the distribution of $\vec{P}$ components are almost the same, equation 5 conveys the fact that the portion of nominator is higher than the denominator about twice. Thus we can see most values are around $\frac{\pi}{2}$ where $\sqrt{p_x^2 + p_y^2}$ is greater than p_z in more possible cases. Based on the same reason, it is less probable to see zero or π.

$$\phi = \arctan \frac{\sqrt{p_x^2 + p_y^2}}{p_z}. \tag{5}$$

The equation 6 is the θ formula in spherical system. Due to the p_x and p_y distribution, every possible value can be expected for the fraction in front of arctan uniformly. Thus, the theta distribution should be uniform.

$$\theta = \arctan \frac{p_y}{p_x}. \tag{6}$$

The distribution shifts for these angles are negligible by smearing (Figure 8). This phenomenon happens after adding errors based on a normal distribution. There is a small difference in the extreme values in theta distribution that is proved by the fact that higher values of 2π should be trimmed to their equivalent above zero. This occurrence for phi is not detectable from the diagram because its extremes have small frequencies.

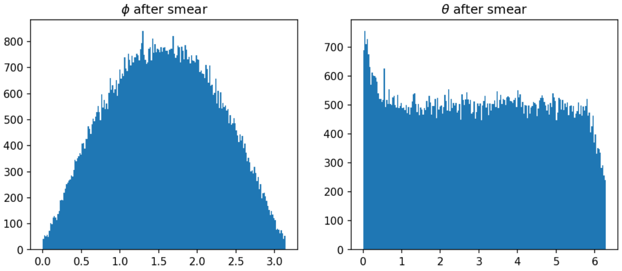


FIG. 8: Distribution of ϕ (left) and θ (right) after smearing.

We have considered both particles' mass constant, but after smearing angels and energy mass is recalculated to conserve the equation $M^2 = E^2 + P^2$. Figure 9 shows the distribution of new values for mass that is normal with low variance and mean about 114 GeV.

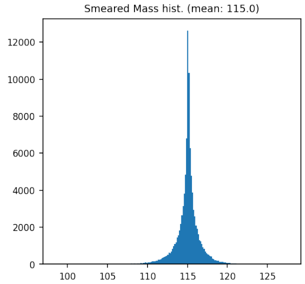


FIG. 9: Recalculated mass distribution after smearing energy and momentum (angels)

To analyze more detailed, the relativistic dispersion relation between energy and the total momentum is plotted (shown in Figure 10). The relation looks rational because for the low values of energy every possible amount of momentum can occur, while for the larger values of it, momentum weight can not be high. Figure 11 shows the fact that the correlation between the real and fake (VAE-generated) data is high.

To be more specific, Figure 12 compares the distribution of spherical angels of real and fake data for

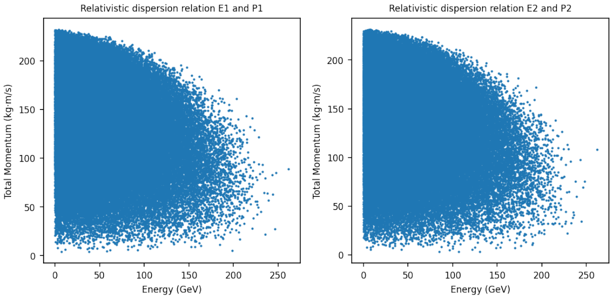


FIG. 10: The relativistic dispersion relation between energy and the total momentum that shows for larger values of the former, the latter should be low.

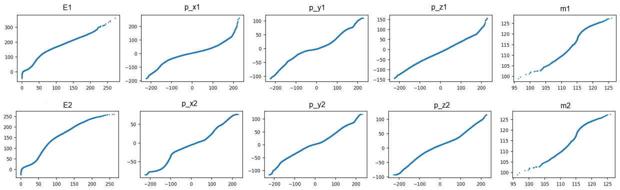


FIG. 11: The diagram shows the relation between each component of real and fake data for both new particles. The x-axis is for real and y-axis is for fake data.

both particles. Also, this comparison for energy and mass is shown in figure 13.

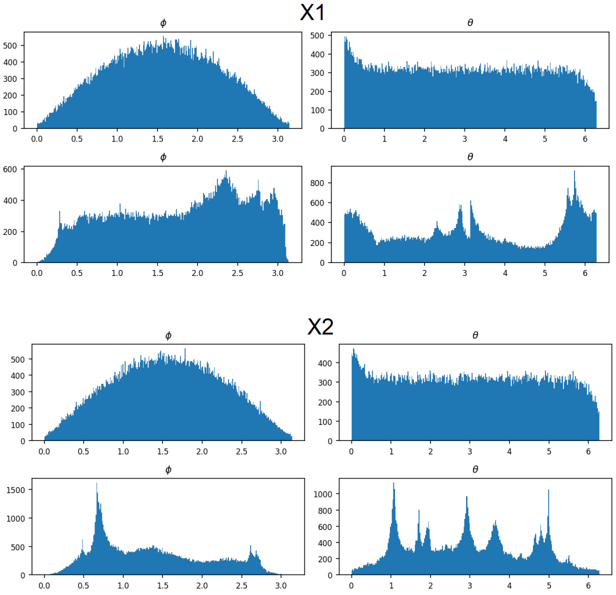


FIG. 12: The diagrams in the first and third rows are related to real spherical angels distribution of particles X1 and X2, and the ones in rows two and four are angels of fake data

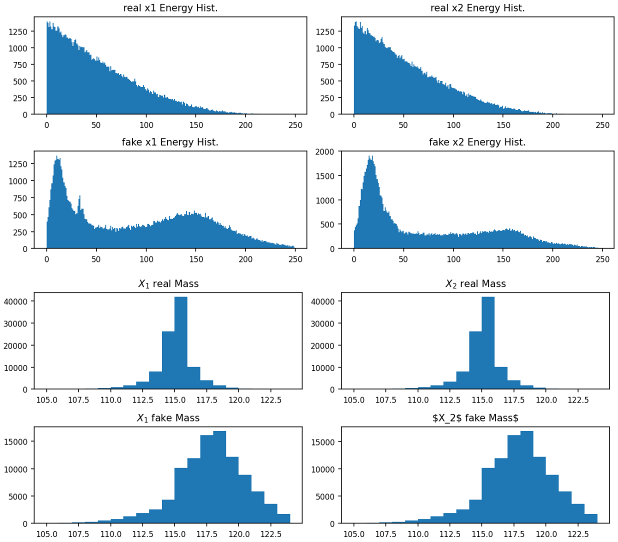


FIG. 13: The diagrams in the first row are related real energy of particles X1 and X2, followed by their fake ones in the second row. The third and fourth row of the figure indicates the mass distribution of particles in the case of real and fake, respectively.

Notably, table I indicates the result of KLD comparison setup between real and fake data for each component of new particles.

Particle	**energy**	ϕ	θ	**mass**
X_1	0.44	0.14	0.09	1.46
X_2	0.31	0.14	0.15	1.46

TABLE I: The KL divergence between two distribution of real and fake data

IV. CONCLUSION

The challenges in producing data with Monte Carlo sampling were mostly related allocating values to features order that I addressed it by starting to sample from three uniform distribution in range [-400, 400] for three components of momentum. As we were supposed to consider the value of mass constant (e.g. 114 Gev), after determining the momentum by Monte Carlo method, energy of particles were calculated easily. The other challenge was setting for ϕ and θ that should stay in ranges $[0, \pi]$ and $[0, 2\pi]$, so I trimmed them by related conditions. Besides, after smearing I should keep the mentioned ranges for angels that got out of range. In the phase of generating a fake model by VAEs, I faced more challenges related training VAE model. First of all, I did a wide range of experiments regarding the capacity of the network from the shallowest to different deeper models. Consequently, I could receive an acceptable result, but not promising, with one of the shallow ones with only one layer. Based on my intuition, I believe VAEs are appropriate models for generating data from the input

Gaussian distribution like the MNIST dataset. While our dataset includes parameters with skewed and non-Gaussian for example for the energy.

Appendix A: Appendixes

Figure 14 shows the latent space of VAE for enrgy as an example.

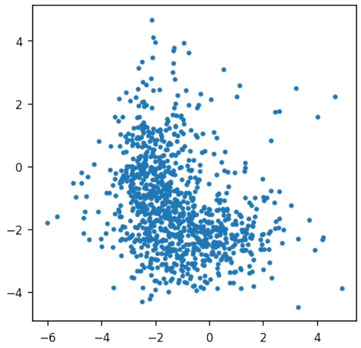


FIG. 14: Latent space of energy for particle X1 in VAE that is generated after training of the encoder.

Figure 15 is the trend of decreasing train and validation loss during the learning of model.

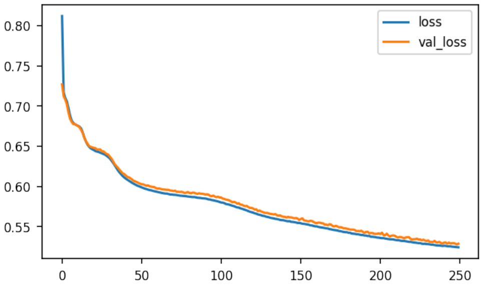


FIG. 15: Train and validation loss trends