ML4HEP 2026 TIFR Mumbai— Tutorial 2

1 Accept-Reject Sampling

Standard data science libraries (for example `scipy.stats`) provide built-in generators for many common distributions. In realistic scientific applications, however, one often encounters complicated probability distributions that must be handled numerically. So we will look one particle method here.

Consider the probability density function

$$f(x) = Ax^2$$

defined on the interval

$$0 \leq x \leq 20$$

where A is a normalization constant.

1. Find the normalization constant A:
   (a) analytically,
   (b) numerically.
2. Derive the cumulative distribution function (CDF) corresponding to $f(x)$.
3. Using the Inverse Transform Sampling technique:
   (a) derive the transformation equation,
   (b) generate random numbers distributed according to $f(x)$,
   (c) plot the resulting histogram and the corresponding analytical curve.
4. Using the Accept-Reject Sampling method:
   (a) choose an appropriate proposal distribution,
   (b) explain your acceptance criterion,
   (c) generate random numbers distributed according to $f(x)$,
   (d) plot the resulting histogram and the corresponding analytical curve.

2 Central Limit Theorem in Particle Detection

A detector measures the energy deposited by muons passing through it. Suppose the energy deposited by a single muon is a random variable E following an exponential distribution with mean

$$\mu = 5 \text{ GeV}$$

The detector records the average energy deposited by $N = 100$ independent muons during a fixed time interval. Define the sample mean energy as

$$\bar{E} = \frac{1}{N} \sum_{i=1}^{N} E_i$$

1. What are the mean and standard deviation of the single-muon energy distribution E?
2. What are the mean and standard deviation of the sample mean distribution $\bar{E}$?
3. According to the Central Limit Theorem:
   (a) what is the approximate distribution of $\bar{E}$?
   (b) what are its parameters?
4. Simulate the experiment assuming E follows an exponential distribution with mean 5 GeV.
5. Generate a histogram of the simulated sample means $\bar{E}$.
6. Overlay the theoretical Gaussian prediction from the Central Limit Theorem on top of the histogram.
7. Estimate numerically the probability that

$$4.8 \text{ GeV} < \bar{E} < 5.2 \text{ GeV}$$

8. Discuss:
   (a) why the CLT works even though the parent distribution is non-Gaussian,
   (b) how the width of the sample-mean distribution changes with N,
   (c) what happens for smaller values of N.

3 Custom Statistical Modeling & The Central Limit Theorem

Consider the continuous random variable X defined on the finite interval

$$x \in [0, 10]$$

with probability density kernel

$$q(x) = C \left(\sin^2(x) \, (10x - x^2) + 0.1 \right)$$

where C is a normalization constant.

1. Determine the normalization constant C such that $q(x)$ becomes a valid probability density function.
2. Compute numerically:
   (a) the true mean,
   (b) the true variance,
   (c) the standard deviation.
3. Plot the analytical probability density function $q(x)$.

4. Using the Rejection Sampling technique:
   (a) construct a sampling engine for $q(x)$,
   (b) generate 1000 random samples,
   (c) clearly explain the proposal distribution and acceptance criterion used.
5. Plot a normalized histogram of the accepted samples and overlay the analytical curve $q(x)$ on top of the histogram.
6. Compute the acceptance ratio of your rejection sampling algorithm and discuss possible strategies to improve the acceptance ratio.
7. Central Limit Theorem Study:

   For each sample size

$$n \in \{2, 5, 30\}$$

   perform 500 independent trials:

   (a) In each trial, draw n random samples from your rejection sampling engine.
   (b) Compute the arithmetic mean

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

   (c) Store the value of $\bar{X}$.
8. For each value of n:
   (a) plot the distribution of sample means,
   (b) overlay the theoretical Gaussian prediction

$$\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

   predicted by the Central Limit Theorem.
9. Discuss:
   (a) how the distribution evolves as n increases,
   (b) convergence toward Gaussianity,
   (c) the physical importance of the Central Limit Theorem in data analysis and detector physics.