Summary of the Titanic

The RMS Titanic, a British passenger liner, is infamous for its tragic sinking on April 15, 1912, during its maiden voyage from Southampton to New York City. Deemed "unsinkable" before its first journey, the Titanic collided with an iceberg in the North Atlantic Ocean, leading to the deaths of more than 1,500 passengers and crew members out of approximately 2,224 on board. This disaster is one of the deadliest maritime tragedies in history and highlighted the severe lack of lifeboats and safety measures on passenger liners. The sinking of the Titanic had a lasting impact on maritime regulations, leading to significant improvements in safety protocols

Math Theory Behind the Project

Monte Carlo

Monte Carlo methods use repeated random sampling to obtain numerical results. They rely on the law of large numbers, predicting that averages from many trials are close to the expected value. In this project, we use Monte Carlo simulations to generate synthetic data resembling the Titanic dataset's statistics.

Monte Carlo Algorithm

The general algorithm for a Monte Carlo simulation can be described as follows:

1. **Define a Domain of Possible Inputs:** Determine a range of possible input values.
2. **Generate Random Inputs:** Randomly generate inputs from the defined domain using a probability distribution that covers the domain.
3. **Perform a Deterministic Computation:** For each random input, compute the deterministic function that models the system or process.
4. **Aggregate the Results:** Collect the results of individual computations.
5. **Analyze the Results:** Analyze the aggregated results to estimate a final outcome. This could involve calculating the mean, variance, or other statistical measures.

This approach is particularly useful for systems or processes that are too complex to model analytically. The accuracy of the results typically increases with the number of trials.

Box-Muller

The Box-Muller transform generates independent, normally distributed random numbers. It's essential for simulating natural phenomena and statistical models assuming a normal distribution. The formulae are:

$$Z_0 = \sqrt{-2\ln U_1}\cos(2\pi U_2), \quad Z_1 = \sqrt{-2\ln U_1}\sin(2\pi U_2)$$

where U_1, U_2 are uniform random variables. For log-normal distributions, we use $Y = e^{Z_0}$ or $Y = e^{Z_1}$.

Variables

Feature	Description
Survived	1 = Survived, 0 = Not Survived
Pclass	1 = Upper, 2 = Middle, 3 = Lower Class
Name	Passenger's Name
Sex	Passenger's Gender
Age	Passenger's Age
SibSp	# Siblings/Spouse Aboard
Parch	# Parents/Children Aboard
Ticket	Ticket Number
Fare	Passenger Fare
Embarked	C = Cherbourg, Q = Queenstown, S = Southampton

Simplified Data Creation Process

1. **Initialization of GPU and CUDA Environment**
   - Initialize curand states on GPU for random number generation.
2. **Memory Allocation for Results on GPU**
   - Allocate GPU memory for continuous (Float) and categorical (Int) variables.
3. **Defining Distribution Parameters**
   - Set parameters for lognormal and normal distributions (mean and standard deviations).
   - Set Parameters for Binary and Ternary Categorical (ratios).

Data Generation Using CUDA Kernels

   - Generate lognormal data for age and fare, normal data for sibsp and parch.
   - Generate Binary Categorical for Survived and sex, Ternary Categorical for Pclass and Embarked.

4. **Synchronization of CUDA Threads**
   - Ensure all CUDA threads complete tasks before proceeding.
5. **Copying Results Back to Host**

1. `nvcc -o data data_creation.cu` (**Note: It is critical to set the output file name to 'data'**)
2. `python3 data_modeling.py` (This script tests different models with both initial and synthetic datasets of any size)
3. `python3 fun.py` (Utilizes logistic regression to predict survival on the Titanic based on various criteria)
4. `python3 fun_gpu.py` (Designed for testing on synthetic datasets of any desired size using GPU acceleration)
5. Maximum n is: 366,090,240