The Energy Test

MULTIVARIATE BINNING-FREE AND NONPARAMETRIC GOODNESS-OF-FIT TEST

Goodness of fit tests

Such a test of a given model describes how well it fits a set of observations. Usually, we are interested in the confirmation or negation of the chosen probability distribution function (p.d.f.).

We can use them to check whether :

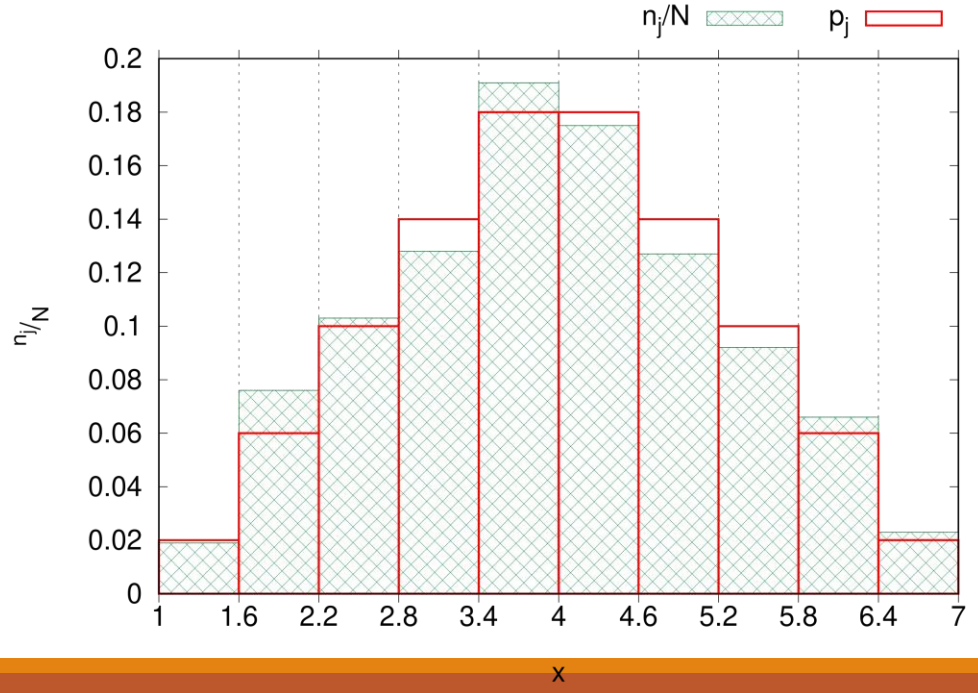
- our statistical model agree with empirical data
- two samples are drawn from identical distributions
- given sample follow the specified distribution

There exist a big varariety of the most popular are:

- Kolmogorov–Smirnov (require estimation of parameters)
- Chi squared (require arbitrary binning)

Why Energy Test can be better?

1. It is nonparametric, so there is no need to use pre-prepared tables.
2. It is binning free (binning always is arbitrary).
3. It is multivariate and scales well with further dimensions.



Chi-square test for a triangular random number generator

n_j - number of drawn numbers in j-th interval
N – number of randomly generated values
p_j - the theoretical probability of drawing a number in the j-th interval

Test statistic

$$T = \frac{1}{2}\frac{1}{n(n-1)}\sum_{i\neq j}^{n}\psi_{ij} + \frac{1}{2}\frac{1}{\bar{n}(\bar{n}-1)}\sum_{i\neq j}^{\bar{n}}\psi_{ij} - \frac{1}{n\bar{n}}\sum_{i\neq j}^{n,\bar{n}}\psi_{ij}$$

- Terms are normalized
- ψ_{ij} weighting function

How can we adjust your test?

First of all , we must choose metric e. g. Euclidean distance:

- $d_{ij}^2 = \sum_{axes,k}(x_{i,k} - x_{j,k})^2$

Now we can consider different weighting functions:

- $\psi_{ij} = \frac{1}{|d_{ij}|}$
- $\psi_{ij} = \text{-}ln|d_{ij} + \epsilon|$
- $\psi_{ij} = e^{\frac{d_{ij}^2}{2\sigma^2}}$

Significance level

Now we need to evaluate given T in order to confirm/deny null hypothesis. We do it by the p-value which is the probability of obtaining test results at least as extreme as the results actually observed, under the assumption that the null hypothesis is correct.

P-value will be different for each entry data and to determine it we perform permutation test.

We assign randomly data entries to each sample and perform Energy Test on a big number of such small chunks. Then we place T from full analyze and see where it belongs against other T-ies.

Usecase at CERN

At LHCb Energy Test is used to find CP-violation. CP (charge conjugation parity symmetry) is a physics law that sometimes does not apply and this indicates a room to formulate new more precise theory than the Standard Model. However such cases are extremely rare and hidden in a big number of „normal” events.

In the experiment, decays can be described by 2 dimensions in case of 2-body decay or 5 dimensions in 3-body decay that is why standard statistical tests are not the best option.

We compare empirical measurements witch theoretical p.d.f and when we spot a p-value smaller than that associated with 5σ we can tell that we observed CP-violation and can further investigate this specific data sample.

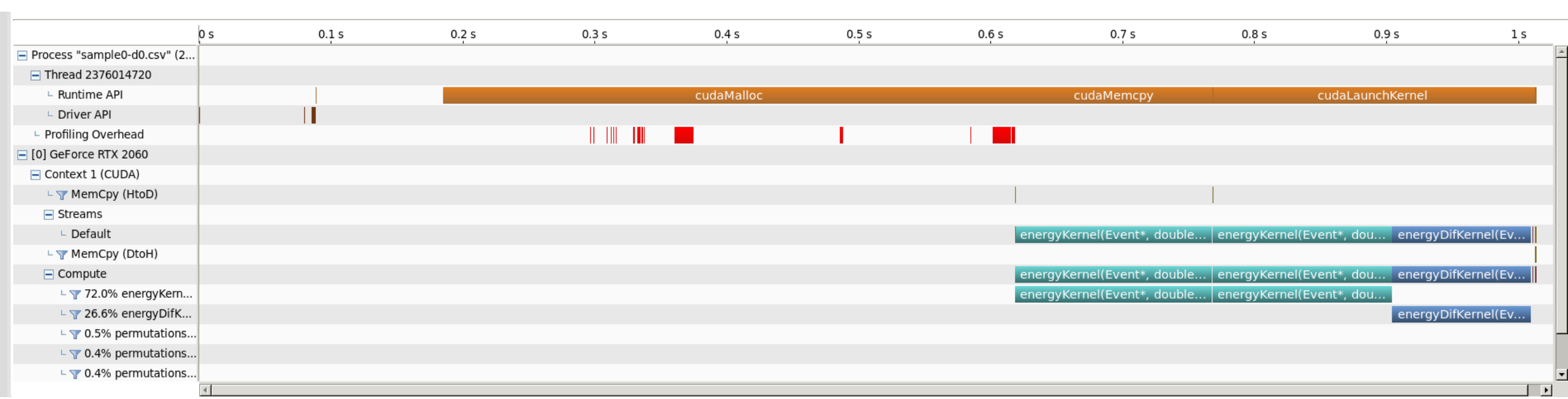
CUDA

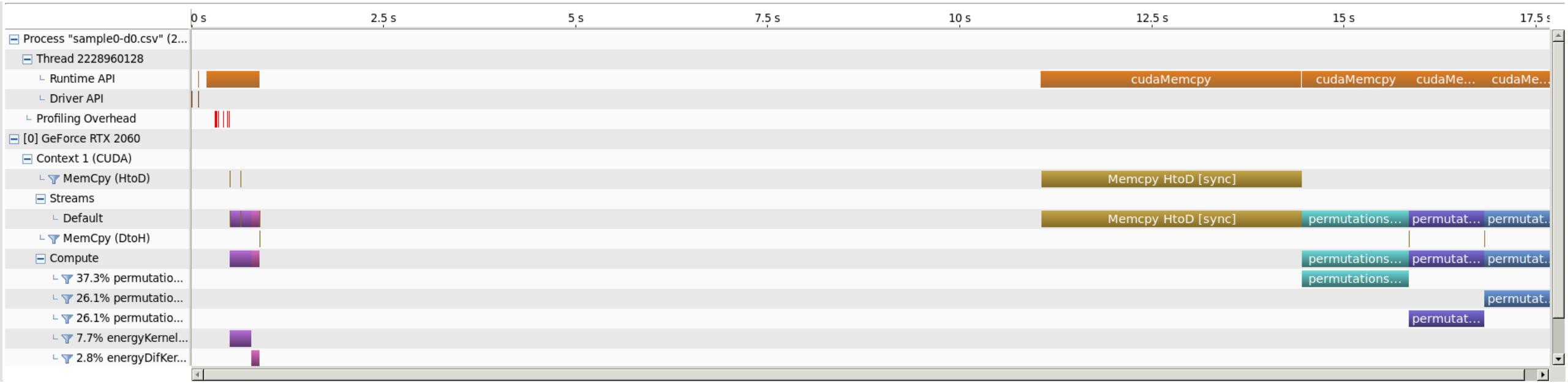
- CUDA is a language extension that allows making computations on GPU instead of on CPU.
- Computations on GPU are massively parallel (thousands of threads concurrently).
- If calculations can be efficiently divided into independent subparts there is a possibility to gains significant performance bust.
- Moving data is costly in comparison to making calculations so GPU can outperform CPU only on large enough datasets. GPU scale better than CPU with the growth of data input size.

Energy Test with CUDA acceleration

Number of permutations	CPU [s]	GPU[s]
1000	0.511	1.040
10000	0.921	1.030
100000	5.038	2.700
900000	45.319	17.7

One permutation 64 elements.


0 s
0.1 s
0.2 s
0.3 s
0.4 s
0.5 s
0.6 s
0.7 s
0.8 s
0.9 s
1 s
Process "sample0-d0.csv" (2...
Thread 2376014720
Runtime API
Driver API
Profiling Overhead
[0] GeForce RTX 2060
Context 1 (CUDA)
MemCpy (HtoD)
Streams
Default
MemCpy (DtoH)
Compute
72.0% energyKern...
26.6% energyDifK...
0.5% permutations...
0.4% permutations...
0.4% permutations...
cudaMalloc
cudaMemcpy
cudaLaunchKernel
energyKernel(Event*, double...
energyKernel(Event*, dou...
energyDifKernel(Ev...


0 s
2.5 s
5 s
7.5 s
10 s
12.5 s
15 s
17.5 s
Process "sample0-d0.csv" (2...
Thread 2228960128
Runtime API
Driver API
Profiling Overhead
[0] GeForce RTX 2060
Context 1 (CUDA)
MemCpy (HtoD)
Streams
Default
MemCpy (DtoH)
Compute
37.3% permutatio...
26.1% permutatio...
26.1% permutatio...
7.7% energyKernel...
2.8% energyDifKer...
cudaMemcpy
cudaMemcpy
cudaMe...
cudaMe...
Memcpy HtoD [sync]
Memcpy HtoD [sync]
permutations...
permutat...
permutat.
permutations...
permutat...
permutat.
permutations...
permutat.
permutat...

Thank for listening!

Questions?

Bibliografia:

- arXiv:hep-ex/0203010

Wiktor Żychowicz