Mathematical Problems of Computer Science 44, 138--144, 2015.

The Optimal Approach for Kolmogorov-Smirnov Test Calculation in High Dimensional Space

Norayr Z. Akopov and Narek H. Martirosyan

Yerevan Physics Institute
e-mail: -narek@mail.yerphi.am

Abstract

Numerical estimation of the Kolmogorov-Smirnov discrepancy D_N in high dimensional space is an extremely time and memory consuming problem. New approach with the minimal bin number, which essentially reduces the time and memory requirements, to perform the D_N tests in two and more dimensional space is discussed.

Keywords: Statistics, Kolmogorov–Smirnov test, Goodness-of-Fit tests, PRNG.

1. Introduction

One of the most important tasks nowadays is the numerical experiments on super computers with the modeling of the sophisticated physics system. The Monte Carlo method is used for solving the problems where the high dimensional integration is involved. To use the Monte Carlo method for analysis of the high energy physics experimental and theoretical problems we have to solve the problem of the quality of the pseudo-random generators, which should have a strong statistical feature, also a large period of sequences and a high speed of generating of the pseudo-random number.

The Kolmogorov-Smirnov (KS) test is used to compare how well the empirical cumulative distribution function (ECDF) of a sample fits the cumulative distribution function (CDF) of the reference distribution by computing the maximum distance D_{max} between these two functions. The KS test is applied to exactly continuous data, and for dimensions $d \geq 2$ we have to compute the distance on infinite points then take a maximum, because when $d \geq 2$ for CDF and ECDF we have d-dimensional surfaces. The KS test is widely used as a powerful statistical test to check the quality of different pseudo-random number generators (PRNGs).

Several algorithms for computing the two-dimensional KS test were proposed [1-3]. We propose another approach for 2 and higher dimensions. Our method is based on discretization of the space introducing a binning technique applied to continuous multidimensional data. This technique does not correspond to the usual KS test principle, but we show that it is possible to compute D_{max} very precisely with the quite computationally efficient algorithm taking minimal bin number.

Here our KS test results are obtained using the well-tested uniform Mersenne Twister PRNG [4], which is included in CERN library [5].

In the proposed paper we describe a new technique, which allows to reduce essentially the number of the bins and correspondingly decrease the needed time and memory used.

2. Formalism

In one-dimensional (1D) case with N random numbers to check if they come from uniform distribution in the (0, 1) range, for which the CDF is equal to:

$$F(x) = x,$$

we need to compute D_N. Usual (unbinned) approach [6- 8] in 1D case looks as follows:

- N numbers should be sorted in increasing order $\{x_1, x_2, \dots, x_N\}, x_i \leq x_{i+1}$
- then D_N is computed in the following way defining ECDF as:

$$F_N(x) = \frac{k}{N}, x = x_1, x_2, \dots, x_N,$$

where k is the number of random points out of N with $x_1, x_2, \dots, x_k \leq x$

$$D_N = \max_{0 \leq x \leq 1} |(F_N(x) - x)| = \max_{1 \leq i \leq N} \left\{ \left| \frac{i}{N} - x_i \right|, \left| \frac{i-1}{N} - x_i \right| \right\}.$$

The unbinned approach can be also applied for 2D case following the algorithm described in [1], which has been used for the performed studies.

In the 1D binned approach the (0, 1) interval is divided into n bins, then N random numbers are distributed in n bins according to their values. The ECDF is again defined as:

$$F_N(x) = \frac{k}{N}.$$

The difference is that here x is already the bin edge: $x = \frac{i}{n}, i = 1,2, \dots, n$. For example, if $n = 10$, we have the bins edges like: $\{0.1, 0.2, \dots, 0.9,1\}$. Therefore, the advantage of the method is that if N is quite large then ECDF can be defined with a relatively small number of points ($n < N$).

$$D_N = \max_{0 \leq x \leq 1} |(F_N(x) - x)| = \max_{1 \leq i \leq n} |F_N(x_i) - x_i| = \max_{1 \leq i \leq n} \left| \frac{\sum_{k=1}^{i} Y_k}{N} - \frac{i}{n} \right|.$$

In 1D case there exists the theoretical distribution function for the value of $K_N = \sqrt{N}\, D_N$ (also the mean value and dispersion for K_N), which is used to make a decision to accept or reject the hypothesis concerning the uniform distribution of the tested sample of random numbers.
This binned method can be generalized for dimensions 2, 3 and higher. In two-dimensional case for $x_1 \le X_1, x_2 \le X_2$ the corresponding expression for D_N is:

$$D_N = \max_{\substack{0\le x_1\le 1\\ 0\le x_2\le 1}} |(F_N(x_1,x_2) - x_1x_2)| = \max_{\substack{1\le i\le n\\ 1\le j\le n}} \left| \frac{\sum_{k=1}^{i}\sum_{m=1}^{j} Y_{km}}{N} - \frac{ij}{n^2} \right|.$$

where Y_{km} is the number of observations that actually do fall into category $\{k, m\}$, N is the count of all observations, *n* is the number of bins used along the *x* and *y*.
In d-dimension case of $x_1 \le X_1, x_2 \le X_2, \dots\dots, x_d \le X_d$ the expression for KS test is as follows:

$$D_N = \max_{\substack{0\le x_1\le 1\\ 0\le x_2\le 1\\ \dots\\ 0\le x_d\le 1}} |(F_N(x_1,x_2,\dots,x_d) - x_1x_{2\dots}x_d)| = \max_{\substack{1\le i_1\le n\\ 1\le i_2\le n\\ \dots\\ 1\le i_d\le n}} \left| \frac{\sum_{k_1=1}^{i_1}\sum_{k_2=1}^{i_2}\dots\sum_{k_d}^{i_d} Y_{k_1k_2\dots k_d}}{N} - \frac{i_1 i_2 \dots i_d}{n^d} \right|.$$

With any of schemes to calculate the average K_N described above (binned and unbinned) we can estimate also the statistical uncertainties ΔK_N making sampling for calculated values of K_N and K_N^2 with the certain number *M* of the used samples. The mean value for K_N averaging over *M* samples is given by:

$$< K_N >= \frac{1}{M}\sum_{i=1}^{M} K_N^i,$$

and mean squared K_N is defined as:

$$< K_N^2 >= \frac{1}{M}\sum_{i=1}^{M} (K_N^i)^2,$$

then we can calculate $\sigma_{K_N}^2$ as:

$$\sigma_{K_N}^2 =< K_N^2 > -< K_N >^2.$$

And finally, based on the central limiting theorem we can calculate the statistical uncertainty ΔK_N as:

$$\Delta K_N = \frac{1}{\sqrt{M}}\sqrt{\sigma_{K_N}^2}.$$

3. Results and Discussion

The first set of the obtained results is related to the unbinned approach, where in 1D case we can compare not only the obtained average $< K_N >$ with the predicted theoretical value of 0.8687311605.. [7], but also the observed and theoretical (TH) distributions for K_N (see Fig. 1).

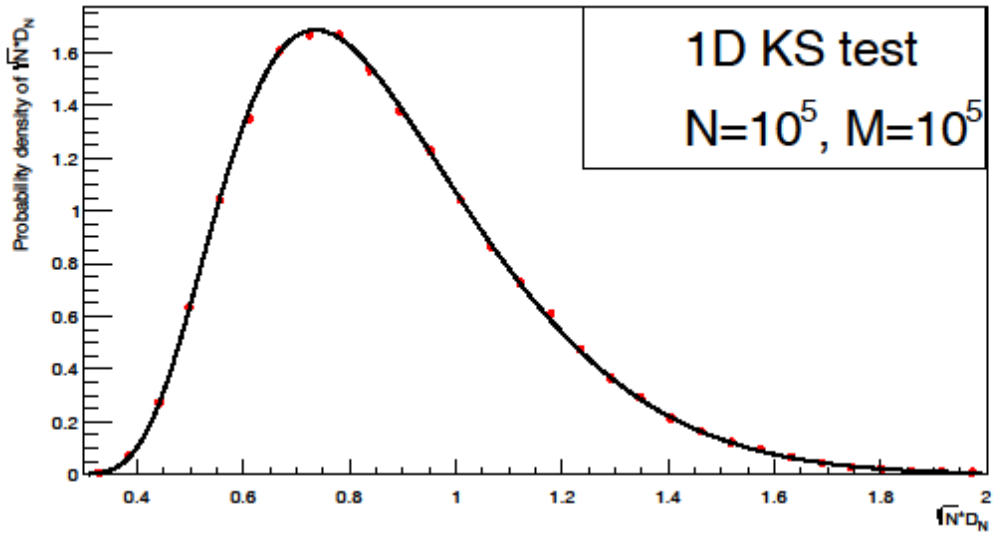

Fig 1. TH (curve) and observed (red points) for K_N distributions.

On Fig. 2 one can see the observed distribution obtained for 2D case, which is systematically shifted in respect to 1D TH distribution.

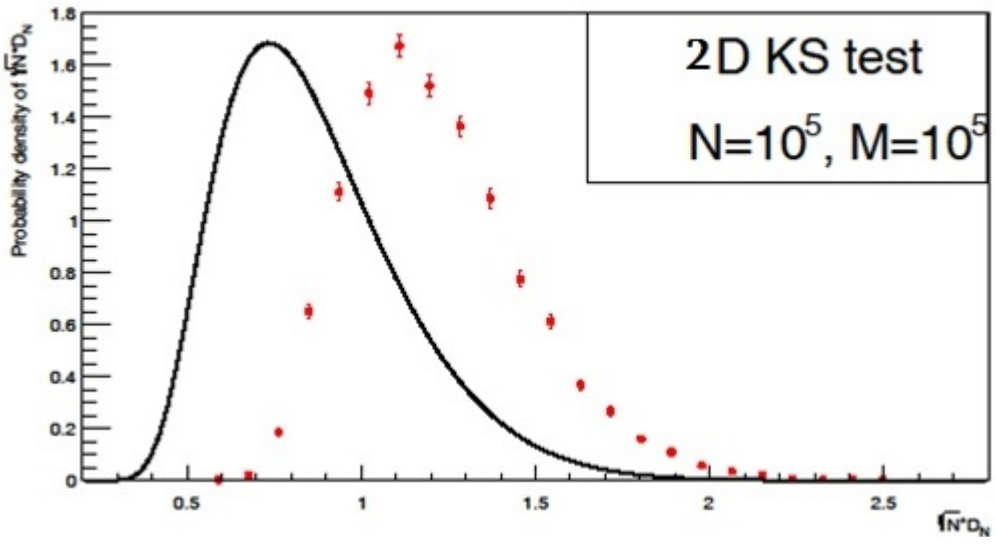

Fig 2. TH distribution (curve) for 1D and observed distribution (red points) for 2D cases.

Then the next set of results, which is related to the binned method, is shown in Figs. 3-4. In Fig.3 one can see that for $N = 10^5$ the number of $n = 10^4$ bins is enough to get the exact distribution which was the goal of this paper.

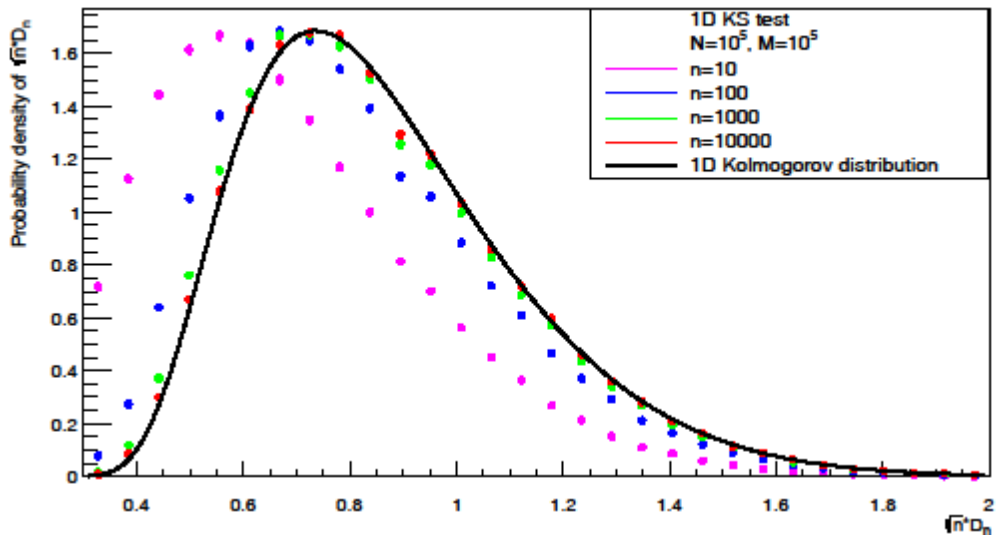

Fig 3. 1D case: TH distribution (curve) and binned distributions (colored points).

In Fig. 4 the $< K_N >$ dependence on the number of bins is presented for 1, 2 and 3D cases. Here one can recognize the clear dependence of the obtained values for $< K_N >$ on the number of the bins (n) used. Taking into account the obtained saturating shape on the n for such dependences we performed the fit with the following fitting function:

$$f_{fit}^{\vec{p}}(x) = p_1 + p_2 \exp(-p_3 x).$$

The meaning of the parameter p_1 corresponds to the saturation level, which can be achieved with quite large value of n. Although, using very high values for the bins number it is impossible in practice to calculate the K_N in the sense of the needed CPU time and memory, especially in case of higher (>4) dimensional space. That is why the idea to use a limited number of points over the n, then estimate the saturation parameter (p_1), and then make a correction for the average value ofK_N, estimated with e.g., $n=10$, introducing the correction factor: $C_{10} = p_1 - < K_N(n = 10) >$, seems to be very interesting and effective in order to realize the procedure of the $< K_N >$ estimation in case of high dimensional space. Also the exponential coefficient (p_3), which regulates the speed of convergence to the saturation level, can be used to check the quality of different PRNGs.

It should be noted, that the used fitting function is quite stable in respect to the variation of the fitting points used. In 3D case with the total number of points (n_{total}=11), this variation was 11, 9, 7, 5. In 2D case with the n_{total}=19: 19, 15, 11, 7, and in 1D case with n_{total}=75: 75, 55, 35, 15. This is an important feature of the functional form used for a fit, because in case of essentially higher dimensions one can compute with the binned method a limited number of points, perform the fit and estimate the saturation level. In further studies the authors are going to provide the $<K_N>$ dependence as a function of the space dimension, which is very interesting due to the lack of knowledge on TH distribution for high dimension space.

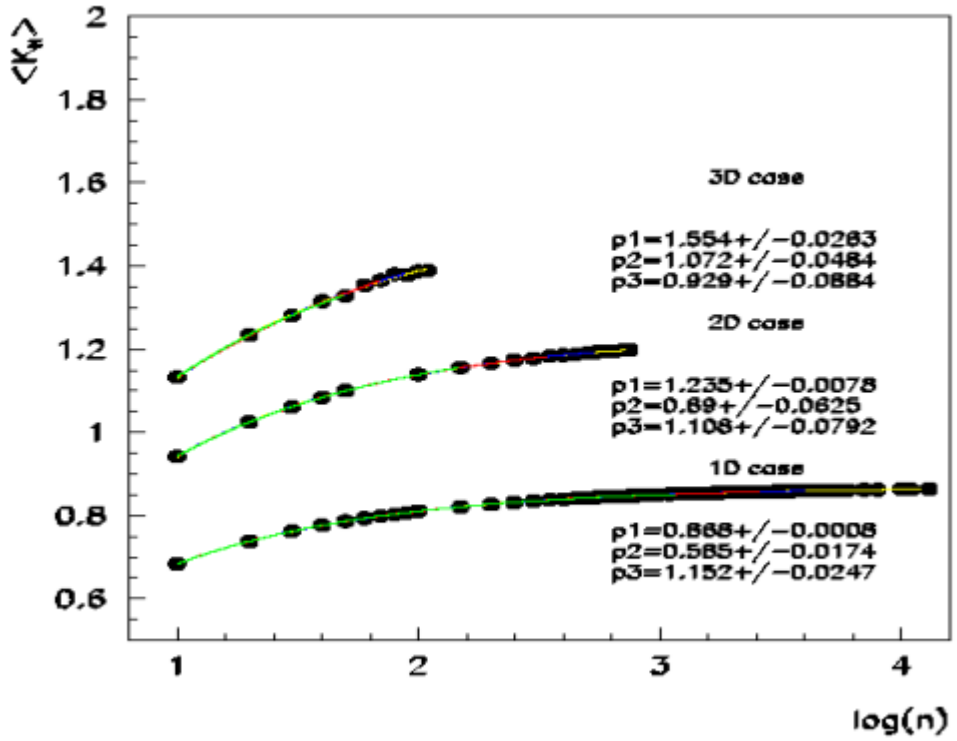

Fig. 4. Average K_N dependence on the number of bins (n). The 1, 2 and 3D cases are presented with the fit parameters for each case. Different colors along the curves correspond to the variation of the number of fitting points used (see explanation in the text).

4. Conclusion

The optimization of the Kolmogorov-Smirnov discrepancy D_N calculation with the binned method for high dimensional spaces by means of reducing the used bin number is performed. Developed approach allows to extend the estimation of D_N to very high dimensional spaces with reasonable requirements to the CPU time and memory, and, thus, to provide very important tests of the PRNGs, which are used to apply the Monte Carlo method for solving of essentially high dimensional physics problems.

References

[1] A. Justel, D. Pena and R. Zamar, “A multivariate Kolmogorov-Smirnov test of goodness of Fit”, *Statistics & Probability Letters* 35, pp. 251-259, 1997.

[2] J. A. Peacock, “Two-dimensional goodness-of-fit testing in astronomy”, *Monthly Notices Royal Astronomy Society*, vol. 202, pp. 615-627, 1983.

[3] G. Fasano and A. Franceschini, “A multidimensional version of the Kolmogorov-Smirnov Test”, *Monthly Notices Royal Astronomy Society*, vol. 225, pp. 155-170, 1987.

[4] M. Matsumoto and T. Nishimura, “Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator”, *ACM Transactions on Modeling and Computer Simulation*, vol. 8, pp. 3-30, 1998.
[5] [Online]. Available: https://cern.ch/cernlib
[6] M. A. Stephens, “EDF statistics for goodness of fit and some comparisons”, *Journal of the American Statistical Association (American Statistical Association)*, vol. 69, pp. 730–737, 1974.
[7] G. Marsaglia, W. W. Tsang and J. Wang, “Evaluating Kolmogorov’s distribution”, *Journal of Statistical Software*, vol. 8, pp. 1-4, 2003.
[8] D. Knuth, *The Art of Computer Programming*, V.2, Addison-Wesley Publishing Company, 1981.

Submitted 30.07.2015, accepted 19.11.2015

Օպտիմալ մոտեցում Կոլմոգորով-Սմիրնով թեստի հաշվարկի համար բարձր չափողականությամբ տարածությունում

Ն. Ակոպով և Ն. Մարտիրոսյան

Ամփոփում

Կոլմոգորով-Սմիրնով թեստի D_N թվային գնահատականը բարձր չափողականությամբ տարածությունում բավականին ժամանակատար և հիշողություն պահանջող խնդիր է: Առաջարկվող հոդվածում քննարկվում է նվազագույն բինների քանակով նոր մոտեցումը, որը էապես նվազեցնում է ժամանակի և հիշողության պահանջները D_N մեծությունը 2 և ավելի բարձր չափողականությամբ տարածությունում հաշվելու համար:

Оптимальный подход для численных реализаций теста Колмогорова-Смирнова в пространствах высокой размерности

Н. Акопов и Н. Мартиросян

Аннотация

Численные оценки дискрепанса Колмогорова-Смирнова D_N в пространствах высокой размерности являются существенной проблемой в плане необходимых ресурсов памяти и времени. В статье обсуждается новый подход для вычисления величины D_N в пространствах двух и более высоких размерностей с минимальным числом разбиений, который существенно снижает требования к памяти и времени.