M&C 2017 - International Conference on Mathematics & Computational Methods Applied to Nuclear Science & Engineering, Jeju, Korea, April 16-20, 2017, on USB (2017)

Equation Section (Next)Adaptive Tally Bin Structure for Problem Analysis and Surface Source Sampling

David Legrady, Zoltan Boroczki, Ildiko Papp

Budapest University of Technology, Műegyetem rkp. 9., H-1111 Budapest, Hungary, legrady@reak.bme.hu

Abstract – *A a data-adaptive regular paving method is presented for density estimation regarding particle distributions. The algorithm halves each phase space segment sequentially in each dimension creating an unevenly segmented but regular mesh. Division target functions of smallest L_2norm and equal relative variance have been studied. The results were applied to multidimensional MCNP tallying using PTRAC files and surface source sampling.*

I. INTRODUCTION

Monte Carlo nuclear particle transport calculations produce a set of samples with phase space dimensions of spatial location, solid angle, energy and time. For sophisticated problems we may add more phase space variables: number of collisions, starting location, parent history number etc. When we wish to analyze our Monte Carlo calculation in terms of phase-space dependent distributions to understand particle pathways, to visualize response flow or to optimize variance reduction a multidimensional bin structure is needed. Without detailed knowledge of the system we may overpartition some unpopulated part of the phase space and will get a few to none hits per bin. In this paper we propose an adaptive tally binning using regular paving using samples generated by MCNP [1]. Further, the created tallying system is used for surface source sampling.

The problem at hand is similar to Probability Density Estimation (PDE) with a long history of development [2]. Some techniques have found their way to the nuclear field like PDE with continuous functions [3], using kernel density estimators [4] or even adaptive binning [5]. We would like to contribute to this ongoing effort by optimizing for the most information that statistic allows for when a sample set has already been generated in order to obtain useable MC estimates in each bin. A further development regards the use of such binned quantities as surface or volumetric tallies.

II. THEORY AND ALGORITHMS

1. Density estimation using Monte Carlo samples

Adaptive binning with variable bin boundaries face its greatest challenge by multidimensionality: variable boundaries in at least two dimensions would lead to a paving problem where a small step toward generality would mean a way bigger step towards unproductive sophistication. To ease the treatment we have settled for a tree structure where parent bins are divided into two daughters along bin boundaries that are set along hyperplanes perpendicular to each other and each level of branching happens in the same dimension.

To establish nomenclature let us model the Monte Carlo quantity estimation mathematically by considering an $f(P)$ probability density function (*pdf*) with argument P standing for a set of phase space variables, let us denote a pay-off or detector function as $g(P)$ and an I result of a definite integral on the whole phase space domain that is the outcome of the Monte Carlo calculation:

$$I = \int f\left(P\right) g\left(P\right) dP \tag{1.1}$$

With P_i samples following the $f(P)$ distribution we can estimate I by

$$I \cong \frac{1}{N} \sum_{i=1}^{N} g\left(P_i\right) \tag{1.2}$$

We may attribute a w_i weight for each coordinate sample (P'_i, w_i) as long as the following holds:

$$I \cong \frac{1}{N} \sum_{i=1}^{N} g\left(P_i^{'}\right) w_i \tag{1.3}$$

With this definition the restriction on f to be a pdf can be relaxed as long as we can formulate our problem as Eq.(1.1) and a random variable–weight set that fulfills Eq.(1.3).

The r^2 relative variance may be estimated [1] by

$$r^2 = \frac{\sum_{i=1}^{N} \left[w_i g\left(P'_i\right) \right]^2}{\left[\sum_{i=1}^{N} w_i g\left(P'_i\right) \right]^2} - \frac{1}{N} \tag{1.4}$$

In a Monte Carlo transport we obtain samples (P_i, w_i) by a complicated simulation process where $f(P)$ (a quantity proportional to a density that describes particle population e.g. the particle flux) is not known explicitly but in many instances we would very much like to reconstruct it. Monte Carlo may estimate definite integrals as in Eq.(1.1) where the set of pay-off functions can be chosen according to our needs. We may use kernel functions [4] for such purpose or a set of orthogonal functions $g_j(P)$ such that:

$$\int g_j\left(P\right) g_k\left(P\right) dP = c_{jk} \delta_{jk} \tag{1.5}$$

with c_{jk} a known normalization constant and δ_{jk} the Kronecker delta. If a function may be expanded into a series of $g_j(P)$:

$$f(P) \approx \sum_{j=1}^{K} d_j g_j(P) \tag{1.6}$$

the coefficients d_j can be calculated as

$$d_j = \int g_j(P) f(P) dP / c_{jj} \tag{1.7}$$

and can be therefore estimated by a Monte Carlo calculation.

We can choose for example $g_j(P)$ to be Legendre polynomials [3] or as it is most often done as a disjoint set of intervals completely covering the support of *f*, with a constant scoring function:

$$g_j(P) = \frac{\Pi_j(P)}{\sqrt{\Delta_j}} \tag{1.8}$$

with Π_j characteristic function (*1* if *P* falls in the j^{th} interval, *0* otherwise) and Δ_j the size of the domain. By this the estimation is simply the average value of function on a certain domain.

For the best estimate of *f* formulating an optimal strategy seems within reach if we operate in one dimension, however already for two dimensional problems the need for paving the whole support of *f* adds serious complications to the otherwise already challenging task. In the nuclear field most commonly the phase space dimensions are segmented independently resulting in a structured mesh with some bins obtaining many Monte Carlo scores and some obtaining very little.

As very commonly done in pdf estimation in the field of mathematical statistics we may partition our phase space along perpendicular hyperplanes in a recurring succession of the phase space dimensions always halving the bins only along the dimension in turn.

To illustrate this process let us choose an optimization strategy of creating bins with equal number of samples falling into each in a two dimensional setup. The algorithm is as follows: take all the samples and divide them into two groups by finding the median of the group along the first dimension. Take the two groups and divide each independently by finding the median along the second dimension. This process can be continued restarting with the first dimension until the desired number of bins is reached. After the binning is done the corresponding tally estimate is formed by adding up the contributions to that phase space cell and division by the cell size.

Fig. 1 shows such a process. 5 independent two dimensional Gaussian distributions were set at the beginning of the calculation each had randomly selected centers and spreads. From theses distributions altogether 10^6 random samples were drawn and binned adaptively into more and more bins (succession goes top to bottom then along columns, only some images of the process are shown) reaching 32x32 bins (divisions in respective dimensions) at the end.

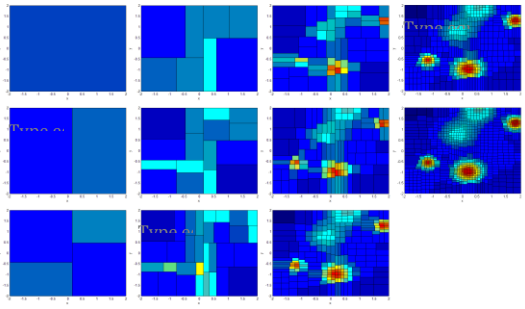

Fig. 1. Adaptive histogram of 5 independent Gaussians using 10^6 samples, with increasing number of bins. Color scale indicates relative density.

The effort leading to [5] does an adaptive binning but the optimization criterion is stated rather than derived from a preset goal and even the underlying algorithm is not thoroughly explained. We attempt here to set optimization goals first and derive the belonging optimization scheme next.

2. Optimizing for the best L2 norm for tallying application

If our aim is to recreate a density of interest to the highest precision allowed by the data we can try to minimize the L_2 distance of the estimate $\hat{f}$ to the actual function *f*.

$$L_2 = \int \left[f(P) - \hat{f}(P) \right]^2 dP \tag{1.9}$$

The L_2 residual is a sum of the deterministic error and stochastic error.

Now let us choose optimal division of a certain interval with respect to the L_2 error. Let our interval length be Δ to be subdivided into Δ_1 and Δ_2 with function estimates d_1 and d_2. The D^2_{Stat} stochastic error on this interval reads:

$$D^2_{Stat}\left(\hat{f}\right) = r^2(d_1)\left[\frac{d_1}{\Delta_1}\right]^2 + r^2(d_2)\left[\frac{d_2}{\Delta_2}\right]^2 \tag{1.10}$$

To facilitate the simple interpretation of this formula for finding the optimum, let us take an analog case, when *M* number of particles is started, out of which *N* has hit our interval and after the subdivision N_1 and N_2. Omitting the *1/M* term and using Eq.(1.4), the total variance will be proportional to

$$D^2_{Stat}\left(\hat{f}\right) \propto \frac{N_1}{\Delta_1^2} + \frac{N - N_1}{(\Delta - \Delta_1)^2} \tag{1.11}$$

If the density function is such that N_1 grows at least linearly with Δ_1 the optimum is at Δ_1=0, otherwise an optimum falling within the interval exists. Also, if we only

have a low number of scores in a bin, the error estimate becomes very unstable and the minimization fails.

The disretization error is harder to estimate but a linear approximation may be given and can be used for optimization of the segmenting. Given a segment we would like to subdivide, and with that a single dimension selected we can expand *f* into Legendre polynomials. The zeroth order term does not give any discretization error, the linear term is the first to consider. We can transform the coordinates onto an interval of (-Δ/2,Δ/2) and calculate

$$d_1^{Legendre} \cong \frac{1}{N}\sum_{i=1}^{N} P_i w_i \tag{1.12}$$

thereby obtaining an estimate of the linear component of the density in question since Eq. (1.12) is the estimate of the coefficient of the firs Legendre expansion term. Integrating the L_2 error caused by the linear component on the interval, the deterministic D^2_{Det} error is proportional to

$$D^2_{Det} \propto \left(d_1^{Legendre}\right)^2 \frac{2}{3}\Delta^3 \tag{1.13}$$

The total L_2 error can be thus approximated by combining Eq.(1.10), Eq.(1.12) and Eq. (1.13) and do a numerical optimization for the best division. Note that if *N* is small the stochastic error contributes to the total error the most and may yield a degenerate division point at one of the interval edges. Also in this case the deterministic error estimate becomes suspect. For testing the method we have chosen 100 points in an interval and calculated the estimated total L_2 error and selected the best value, posing an obvious burden on the calculation time.

3. Optimizing for the best relative variance for surface source applications

Dividing a Monte Carlo calculation into two disjoint geometrical parts is often done when a series of calculation is to be performed while altering the setup in a closed geometrical region. This is usually achieved by writing the raw particle data crossing a surface separating the two geometrical parts. This may also be done by calculating the particle population distribution at the surface and sampling this distribution as a source term for the rest of the geometry. For both cases we may formulate the mathematical model for the second calculation as follows [6]:

$$I = \int \underline{n}\phi(P)\phi^{+}(P)\,dA \tag{1.14}$$

with ϕ the angular flux, ϕ^+ the adjoint function, dA denoting a surface integral on all variables and $\underline{n}$ the surface normal. The calculation starting from the surface source would give formally a sample of the adjoint function and finally a contribution to the estimate of *I*.

For a surface source with source density estimated on disjoint bins we can formulate the estimator using the discrete flux estimate on the surface as:

$$I \cong \sum_{i=1}^{K}\left(\underline{n}\phi_i\right)\phi_i^{+} \tag{1.15}$$

with ($\underline{n}\,\phi_i$) the discrete surface source density estimates and ϕ_i^+ their respective contribution to the final estimate.

The variance estimator can be expressed as follows:

$$D^2_{Stat}(I) = \sum_{i=1}^{K} r^2\left(\underline{n}\phi_i\right)\phi_i^{+2} + r^2\left(\phi_i^{+}\right)\underline{n}\phi_i^{2} + \\ + r^2\left(\underline{n}\phi_i\right) r^2\left(\phi_i^{+}\right)\phi_i^{+2}\underline{n}\phi_i^{2} \tag{1.16}$$

Let us disregard the double product of the relative variances and focus on the first two terms. The term $r^2(\phi_i^+)$ stands for the variance associated with the simulation staring from the i[th] surface source bin with as many samples as we wish to start and by that this term can be decreased without theoretical limit regardless how we optimized our surface source. The error term $r^2(\underline{n}\phi_i)\phi_i^+$ however should yield a criterion for optimal binning. At the time of creating the surface source we have no information on the adjoint function thus the best option we have is to create an even set of relative variances in every bin.

Keeping the same relative variance in each bin would also mean in the analog case keeping the amount of samples contributing to a bin estimate constant along all bins. For an analog case the algorithm is as follows: take all the samples and divide them into two groups by finding the median of the group along the first dimension. Take the two groups and divide each independently by finding the median along the second dimension. This process can be continued along each following dimension going through the available dimensions repeatedly until the desired number of bins is reached. After the binning is done the corresponding tally estimate is formed by adding up the contributions to that phase space cell and division by the cell size with an appropriate surface flux estimator.

III. RESULTS

1. Analysis of variance as function of number of bins

The number of bins is optimal for a given sample set if the discretization error of the underlying (probability density) function given as the squared difference to the piecewise constant approximation equals to the variance of the statistical estimator. For a well behaving smooth function the discretization error decreases with the length of the interval and the statistical variance decreases by the number of samples falling into a bin in an analog case strictly proportional to the number of bins. Thus an optimum may exist as a trade-off.

For demonstrating this behavior we have analyzed analog samples generated from a simple one dimensional Gaussian with sample numbers 2^9, 2^{10}, 2^{11}, 2^{12}, and 2^{13}. The optimization scheme was the equal relative variance criterion. The depth of the binary tree was increased and the relative squared deviation of the estimated piecewise

constant function recorded. This procedure was repeated 20 times and the relative deviation was calculated as the average of the deviations over the 20 sets. Results are shown in fig.2.

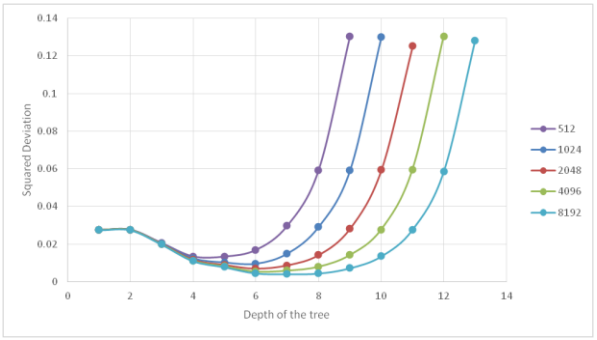

Fig. 2. Relative deviations with different number of samples and tree depth

Optimal sample number per bin was found in this case at 32. We aim at bins where the expected value can be determined and as such this estimate is sound. Note that as a rule of thumb [7] 10% relative error is cited under which the estimate may be expected. Also note that analogue scores would fulfill this criterion by having 100 scores per bin. For regular, equidistant non-adaptive binning such requirement could not be fulfilled in each bin only for flat distributions or vast number of samples.

2. Comparing optimization strategies for a known function

We have compared three segmentation strategies for samples generated from a 2D Gaussian distribution: first an equidistant binning (the "Struct"), second the constant relative variance binning (the "Median") and last the L_2 optimized method (the "min(L2)"). The resulting paving can be seen in Fig. 3 for one quarter of the space.

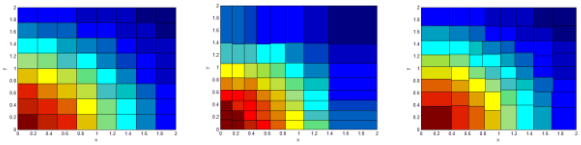

Fig. 3. Bin structures for different optimization strategies from left to right: "Struct", "Median", "min(L2)"

The "Median" strategy results in the best resolution at high function values, "min(L2)" strategy gives smaller divisions at lower function values.

We can compare the total deviation from the analytical function as can be seen in Fig.4. We have taken 10^5 samples from a 2D Gaussian centered at (0,0) and showed the L_2 norm with increasing number of divisions per dimensions (n). The same type of behavior can be seen as in Fig.2.

The best L_2 norm deviation is given by the "min(L2)" method as to be expected. For this curve data points vanish at the minimum deviation point as after these divisions the algorithm started producing zero or very small sized cells with high fluctuations in the norm therefore we omitted them from the graph.

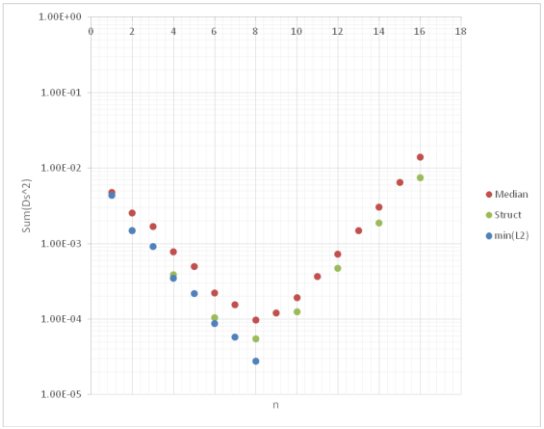

Fig. 4. Total deviations for the different optimization strategies: "Struct", "Median", "min(L2)" with increasing number of divisions per dimension

Somewhat surprisingly the "Struct" equidistant binning produced better total deviation figures than the "Median" strategy.

3. Implementation in MCNP6 for tallying: verification against analytical calculations

The first step of the implementation of the method for use with MCNP-generated data was coding a particle log (PTRAC file) processor in C++ that is capable of binning the MCNP data and printing any 2D plot along any hyperplane. This step did not require any modification of MCNP. The developed code MOnte Carlo SegmentING (MOCSING) is capable of reading the MCNP6 generated binary PTRAC files. User input may provide the order of dimensions in which the segmentation should proceed, together with the level of the tree structure. The produced data involves nodes in a tree structure; each node includes the total weight in the bin the upper boundary and pointers to the daughter nodes. Thus memory need is approximately triple of the needs for a structured mesh.

For checking the validity of the codes an analytically verifiable model was run with MCNP with a point source of 1 MeV photons and an aluminum plate of 5 cm thickness located 30 cm from the source, the plate had a 4 cm diameter circular hole and only unscattered events were recorded. Numerical and analytical results matched within statistics.

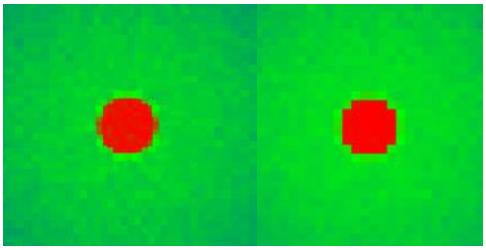

Fig. 5. Transmission through a circular whole. Adaptively binned (left) and regularly binned tally (right)

Fig. 5 shows the comparison of the images obtained by constant relative variance aimed adaptive and equidistantly spaced binning on the plate bounding plane further from the source. Color scale indicates net current values, with red color for higher and green color for lower values.

4. Implementation in MCNP6 for tallying: verification of the binning algorithm

Our next step was to verify the PTRAC processor MOCSING against the MCNP tallying. MCNP does not give a simple solution for a 2D structured mesh tallying in spatial dimensions. We have chosen therefore an energy-angle regular mesh for particles crossing a surface. The geometry consisted of a monodirectional gamma point source of 0.9MeV was directed towards a 1cm lead sphere. Only those particles were registered that scatter only once in the lead sphere and reach an almost infinite plane placed on the other side of the sphere, thereby a strong correlation of energy and cosine with the surface normal was expected. We have compared the results given by MCNP and our regular ("Struct") binning based on PTRAC events. The resulting 2D distribution can be seen on Fig. 6

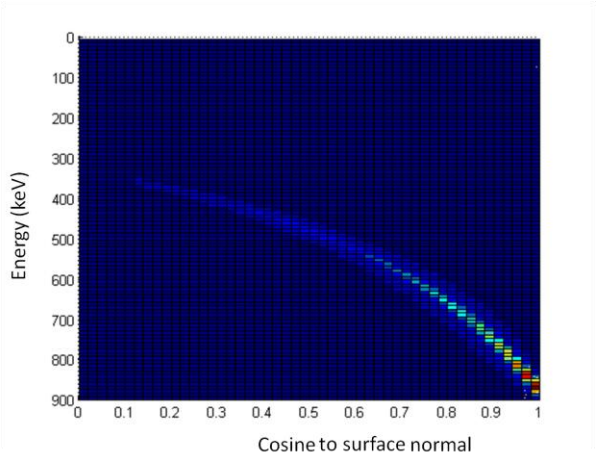

Fig. 6. Energy-Angle distribution of single scattered photons

. Numerical values of the PTRAC binning and the MCNP output matched to the last digit. With this verification we were able to move on to comparing distributions using only our PTRAC processor code.

5. Implementation in MCNP6 for surface source sampling and an illustration

MCNP has a very well established system for using surface sources by printing out the trajectories crossing a future surface source. The code gives opportunity to resample this file if the subsequent run for the second part of the problem requires more particles.

Using this surface source files may challenge the user because of their prohibitively large size for complex problems and the lack of information on the sampling quality of the second calculation: it may easily happen that important directions may have a low number of source particles and they get heavily resampled in the second run. Our motivation to modify MCNP to read source density files tried to address these two issues, the adaptive tree structure offers a compact representation of the surface source file going down to all the information the data set contained and may be resampled without fear of heavy correlations.

To achieve this goal the first step was modifying MCNP to be able to read the binned file on a subsequent run and sample the data as a surface or volumetric source. The SOURCE subroutine of MCNP was modified to read and use the density file produced by MOCSING. The surface density file was created with the equal relative variance scheme in mind. A single random number is enough to select the sampled bin in the binary tree. When a bin is found, in every dimension the rest of the coordinates are sampled uniformly in the segment. The modifications do not influence parallel processing capabilities of the code.

For illustration of the use of surface density files we have built a complex geometry for a transmission problem. The problem geometry can be seen in Fig.7.

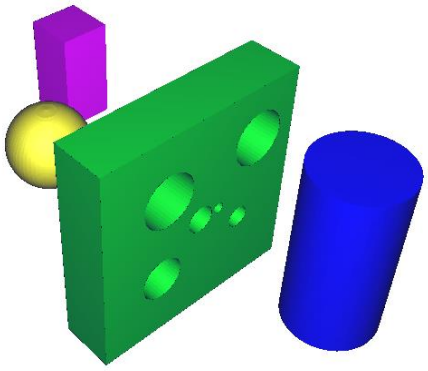

Fig. 7. MCNP geometry for surface source sampling of the density file

2.2 MeV photons were started from a surface perpendicular to the surface normal. Various shapes of material cells were filled with lead, aluminum, copper and graphite. A surface has been created after the photons pass the aluminum cylinder to write out the surface source. In a subsequent run the surface density file is sampled and transmission through the rest of the geometry is recorded at the geometry edge. This result is compared to a transmission without applying any surface source in the process. The resulting transmission spatial distributions are compared in Fig. 8 and Fig. 9.

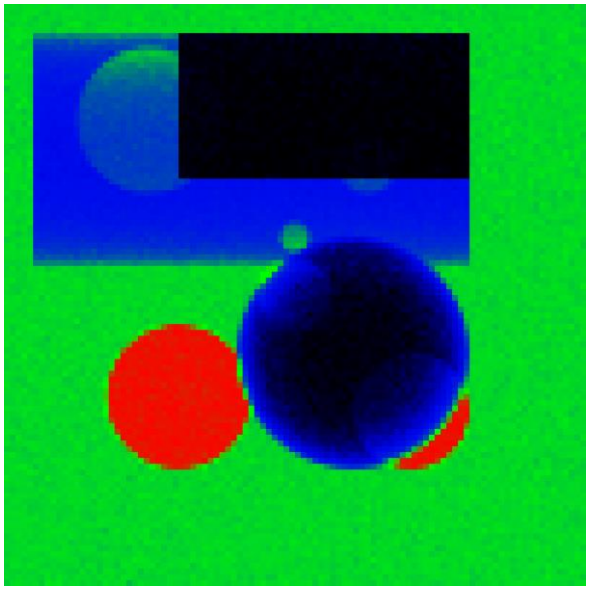

Fig. 8.Spatial distribution of photons without using surface source

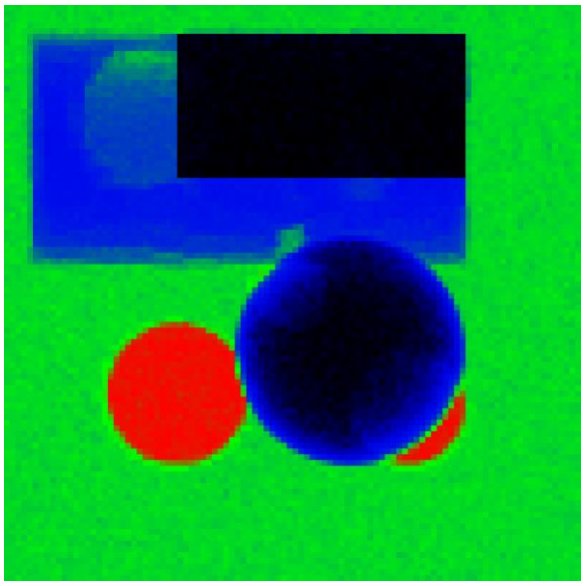

Fig. 9.Spatial distribution of photons using surface source from a density file

The results show excellent agreement, tough note that the problem was very heavily forward-peaked and we applied no segmenting in the direction perpendicular to the surface normal.

IV. CONCLUSIONS

We have analyzed adaptive density estimation of quantities arising from Monte Carlo event generation. Two optimization strategies were devised and their results compared. An event log file processor code was written for

binning raw event files produced by MCNP6. The equal relative variance scheme was put into practice in a modified version of MCNP6 where particles were started from the surface density file.

The work presented here is preliminary in a sense that further analysis could reveal the practical usefulness of these techniques in actual cases. More importantly the technique proposed here for surface sampling is inherently biased because of the discretisation error.

Further work may include a partitioning scheme where the order of the dimensions is also determined by the algorithm and instead of a strict pre-given order the code should decide which segment should be further subdivided. Further application may involve forward-adjoint coupling , volumetric source sampling and weight window generation.

ACKNOWLEDGMENTS

This work has been carried out in the frame of VKSZ_14-1-2015-0021 Hungarian project 260 supported by the National Research, Development and Innovation Fund

REFERENCES

1. T. GOORLEY, et al,. "Initial MCNP6 Release Overview", *Nuclear Technology*, **180**, (2012).

2. D. W. SCOTT and S. R. SAIN, *Multi-dimensional density estimation*, in. *Handbook of Statistics vol.23 Data Mining and Computational Statistics*, C. R. RAO and E.J. WEGMANN. KRAMDEN, Ed., Elsevier Amsterdam (2004).

3. D.P. GRIESHEIMER, W.R. MARTIN, "Monte Carlo Based Angular Flux Distribution with Orthogonal Function Expansion," *Trans. Am. Nucl. Soc.* 89 (2003).

4. K. BANARJEE and W.R. MARTIN, "Kernel Density Estimation Method for Monte Carlo Tallies withUnbounded Variance," *Trans. Am. Nucl. Soc.* 101, 430-432, Washington, DC (November 2009).

5. D. DANNHEIM, A. VOIGT, K.-J. GRAHN, P. SPECKMAYER AND T. CARLI, „PDE-Foam—A probability density estimation method using self-adapting phase-space binning," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, pp. 717-727, 2009

6. D LEGRADY, J E HOOGENBOOM, J L KLOOSTERMAN, " The Time Dependent Monte Carlo Midway Method for Application to Borehole Logging", *M&C 2001, Salt Lake City, Utah, USA (*2001*)*

7. X-5 Monte Carlo Team, "MCNP - Version 5, Vol. I: Overview and Theory", LA-UR-03-1987 (2003)