CERN 68-15
1 May 1968

ORGANISATION EUROPÉENNE POUR LA RECHERCHE NUCLÉAIRE
CERN EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH

MONTE CARLO PHASE SPACE

F. James

Lectures given in the Academic
Training Programme of CERN

G E N E V A
1968

Genève - Service d'Information scientifique - H/314 - 2000 - mai 1968

CONTENTS

1. INTRODUCTION

It will be assumed from the outset that the kind of problem we are likely to want to solve is not accessible to direct (analytic) calculation, since such a calculation, where it is possible, is nearly always far superior to a Monte Carlo determination. As we will see later, direct integration of most interesting phase-space problems is exceedingly complicated and has been achieved only for a small class of relatively simple problems. We therefore will begin directly with a general treatment of Monte Carlo techniques, gradually specializing our study by taking as working examples first the one-dimensional integral, then the multidimensional integral and finally the phase-space integral.

* * *

PART I

THE MONTE CARLO METHOD

2. WHAT IS MONTE CARLO? TWO EXAMPLES

We begin with a very elementary treatment of the problem of defining clearly what we mean by Monte Carlo since this may avoid some fundamental misunderstandings, which often arise in discussing details of actual calculations. In particular we start by oversimplifying the question: we define two kinds of problems (which are unfortunately not orthogonal and in actual practice we encounter problems which can be considered from either point of view):

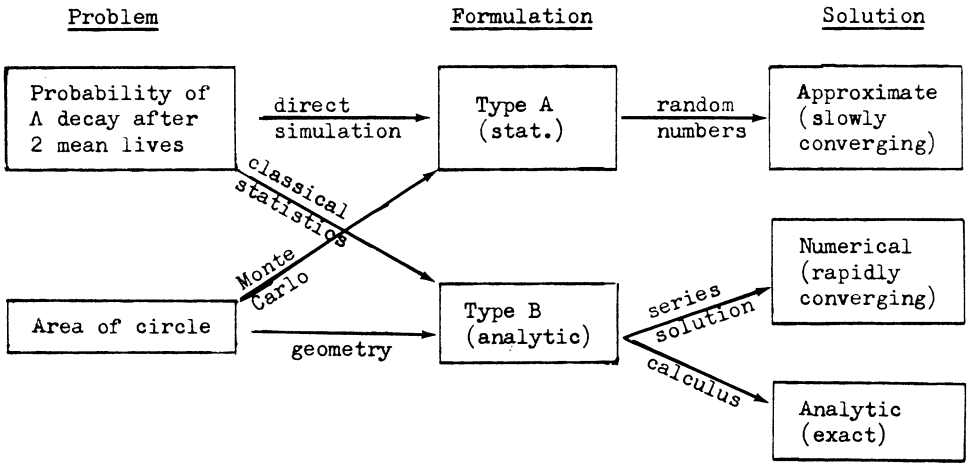
Type A - statistical, probabilistic, involving random fluctuation;

Type B - analytic, exact, classical problems.

As an example of a Type A problem we consider what is the probability of observing a Λ^0 decay after two mean lives, and what fluctuation we might expect.

As an example of a Type B problem, we ask what is the area of a circle of unit radius.

We are accustomed to solving both problems analytically, although it is quickly seen that they may both be solved using random processes also. These relationships may be visualized with the following diagram:



As everyone knows, the area of the unit circle is just = π. This can be considered as an exactly known number, or as a number which can be determined to any desired degree of accuracy by a rapidly converging series expression. On the other hand this same problem could be formulated as a Monte Carlo problem by noticing that the area of the unit circle is just four times the probability that a point chosen at random within a 2 x 2 square will also fall within the inscribed circle. This formulation has the unfortunate property that the solution converges only as $N^{-1/2}$, where N is the number of random points chosen (meaning that to gain one more significant figure you must generate 100 times as many points). As we will see, it is possible to reduce the magnitude of this error by using more sophisticated techniques, but the convergence as $N^{-1/2}$ is impossible to avoid as long as we use truly random numbers.

Now let us look at the Type A (statistical) problem concerning Λ^0 decay. It is clear that the problem can be formulated as a Type B (analytic) problem since the answer is just the integral of the Λ-decay curve from two lifetimes to infinity and is equal to e^{-2}. The expected fluctuations about this number can be calculated using the well-known technique based on the binomial expansion. So far there seems to be no advantage in going to a Monte Carlo formulation since we can formulate even statistical problems in such a way as to obtain precise results by analytic means. But just for fun let us apply Monte Carlo to the Λ decay problem. If we divide our time scale into small intervals dt, then the probability of a decay is just the negative of

$$dN = -\lambda N \, dt,$$

where λ is the decay constant ($\lambda = 1/\tau$) and N is the number of Λ^0 not yet decayed. Since this formula is exact only in the limit that N does not change over the time interval considered, we must choose dt small enough so that $\lambda N \, dt \ll 1$. Then, starting with N_0 particles in the first time interval, we choose a random number between 0 and 1. If this number is larger than $\lambda N \, dt$, we say that a decay did <u>not</u> occur in this time interval; if it is smaller than $\lambda N \, dt$, we reduce N by one before going on to the next interval. (This general Monte Carlo technique is valid because the probability that a random number be smaller than $\lambda N \, dt$ is just $\lambda N \, dt$: this is what we mean by a random number uniformly distributed between zero and one.) We then repeat this process for each successive time interval until there are no more Λ^0 left to decay. This method of following the actual physical process in order to study it is called <u>direct simulation</u> and is the most obvious Monte Carlo technique. It can be seen that this typical example has two clear disadvantages:

i) We are forced to make a compromise between calculation speed (large dt) and accuracy (small dt).

ii) Our numerical results will contain inherent statistical fluctuations which will go away only as $\sqrt{N}$.

However, things are not so bad as they seem because:

i) The statistical fluctuations will anyway not be worse than those that we see experimentally, and in fact we may wish to study these fluctuations.

ii) Since we have directly simulated the desired process, we may study many aspects of it at the same time, whereas an analytic treatment would probably require a separate calculation for each desired result.

iii) We may easily introduce complicating effects such as the resolution time of the counters used, as well as their geometric efficiency, etc.

3. MONTE CARLO AS INTEGRATION

Both of the examples treated above can be regarded as problems of integration:

$$\text{area of circle} = 4 \cdot \int_0^1 \sqrt{(1 - x^2)}\, dx$$

$$\text{probability } \Lambda^0 \text{ decay} = \int_2^\infty e^{-\lambda t}\, d(\lambda t) \ .$$

Indeed, it has been noted that all Monte Carlo problems are essentially integrations, for the following reason. All Monte Carlo calculations involve the determination of a number or a set of numbers (call these the vector $\vec{F}$) using an ensemble of random numbers r_1, r_2, r_3 ... r_N. We have then determined a vector valued function of random numbers:

$$\vec{F}(r_1, r_2 \ \ldots \ r_n) \ ,$$

where n may be of the order of several thousand, for example. This is an unbiased estimator (albeit a rather bad estimator) of the n-dimensional integral:

$$\vec{I} = \int_0^1 \int_0^1 \int_0^1 \int_0^1 \cdots \int_0^1 \vec{F}(x_1, x_2, \ \ldots \ x_n)\, dx_1\, dx_2 \ \ldots \ dx_n \ .$$

Of course only a mathematician could think of such a far-fetched idea as estimating an integral over many-dimensional space by evaluating the function at one random point in the space. But if we think of the inverse -- that a more precise evaluation of the integral $\vec{I}$ would be a more precise determination of what we are looking for, $\vec{F}$ -- it begins to look like a more reasonable idea. Indeed it is well to emphasize this point now in order to avoid a confusion which often arises in discussions of phase-space problems. In such multiple integration problems we will usually perform some of the integration classically (analytically) to prepare the problem for the application of Monte Carlo method. This must not be confused with the Monte Carlo itself which constitutes an effective integration over the remaining variables. The fact that phase-space problems will usually be formulated as direct simulation should not obscure the nature of the problem which could alternatively be regarded as an integration over a region of phase space.

We now go on to consider numerical integration in more detail.

4. NUMERICAL INTEGRATION

The simplest method for numerical integration (not Monte Carlo) is the trapezoid rule, which consists of dividing the required interval into n bands and approximating the integral over each band by the area of the trapezoid inscribed under (or over) the curve. This reduces to estimating the integral by taking the average value of the function as determined from n equally-spaced points and multiplying by the total interval. For large n, we can think of the function as being expressed by a Taylor's series expansion about each of the n points; then the constant terms and the first derivative terms will be integrated exactly by the trapezoid rule, and the largest contribution to the error will come from the second derivative (constant curvature) terms. This error will be proportional to the sagittas of the curve segments over each band, and these sagittas will each be proportional to the square of the distance between successive points where the function is evaluated. Therefore, if the function is evaluated at n equally spaced points, the error on the integral will be proportional to $1/n^2$ for large n. In fact, if we push the method a little farther we can obtain even faster convergence. Consider an estimate of an integral found by using the trapezoid rule at n points:

$$T_n = I + \varphi(n^{-2}) + \varphi(n^{-4}) + \ldots ,$$

where I is the "true" value of the integral and $\varphi(n^{-2})$ means an error of the order of $1/n^2$, or this may be considered to be the integral of the third term in the Taylor's series expansion mentioned above. Now if we make another estimate based on 2n points, we will have:

$$T_{2n} = I + \frac{1}{4}\varphi(n^{-2}) + \frac{1}{16}\varphi(n^{-4}) + \ldots .$$

We now consider the following linear combination:

$$T' = \frac{4T_{2n} - T_n}{3} = I - \frac{1}{4}\varphi(n^{-4}) + \ldots .$$

Now the term in $1/n^2$ has dropped out, and we see that this method will converge as $1/n^4$. This process can then be extended in an obvious way in order to eliminate the $1/n^4$ term, etc.

But rather than extend this method to its limit, we can use the fully optimized formula due to Gauss, which in fact represents the limit of this kind of method in the sense that the n-point Gauss formula is exact for all polynomials up to order $2n - 1$.

We remind the reader here that Monte Carlo integration using standard random numbers converges always as $\sqrt{n}$. The following table may therefore be established comparing the dependence of errors on number of points for different methods of numerical estimation of a one-dimensional integral: in order of increasingly fast convergence:

Monte Carlo	$\sigma \sim n^{-1/2}$
Trapezoid	$\sigma \sim n^{-2}$
"Second-order rule"	$\sigma \sim n^{-4}$
m-point Gauss rule	$\sigma \sim n^{-2m+1}$.

This makes the Monte Carlo method look pretty bad, even if we notice that it is the only one of these for which the error estimate does not depend on any assumed "smooth" behaviour of the function.

However, the future for Monte Carlo looks considerably brighter when we go to multi-dimensional integration. If we modify the above formulas to hold for integration over K dimensions, we see that the Monte Carlo expression remains unchanged, whereas for the others the exponent must be divided by K. This means that Monte Carlo converges faster than the trapezoid rule in five or more dimensions, and when we go above five dimensions the non-Monte Carlo methods begin to break down completely because of the outrageous number of points required. As we have seen, these error formulas are only valid for large n (that is, small spacing between points) which certainly means at least 10 points per dimension, for example. But 10 points in 5 dimensions means already 10^5 points in the whole space, and this kind of number must be considered as the limit of the possibilities of today's computers. We remark here that a complete description of an ℓ-particle final state requires $3\ell - 4$ parameters, so that for 4-body phase-space we are already operating in 8 dimensions (although some of these, such as the azimuthal orientation of the event about the beam direction, may be of no interest).

It seems at first surprising that choosing points randomly gives better convergence for multidimensional integrals than choosing the points uniformly or according to some carefully "optimized" rules. This question is discussed extensively in the literature (see bibliography) and lies beyond the scope of these lectures. But the situation can be summarized as follows: if points are chosen intelligently in many dimensions, we can in fact obtain faster convergence than with Monte Carlo. However, rules which are optimum in one dimension, when extended straightforwardly to many dimensions are no longer optimal, and rapidly become worse than Monte Carlo. Some rules are known which are nearly optimal for certain dimensionalities, but the general problem is far from being solved. Probably the best progress to data has come from the discovery of a class of rules (discussed below in Section 8.3 and called quasi-random numbers) which can be shown to have better asymptotic properties than truly random numbers in any number of dimensions.

5. VARIANCE-REDUCING TECHNIQUES

Having justified the use of Monte Carlo and having introduced the main features of this method, we are ready to develop these ideas more precisely and more systematically by considering optimization of Monte Carlo calculations. This always means reducing the statistical error inherent in this method. Since this error is proportional to a quantity known in mathematics jargon as the variance, optimization is called variance-reducing. We have already seen an example of optimization when we considered the n-point trapezoidal rule for numerical integration. In that case we found that a certain linear combination of the n-point and 2n-point estimates was better than the 2n-point estimate alone. A large and growing number of similar techniques exists also for reducing the error in Monte Carlo calculations. Rather than to mention all of these techniques, I will discuss only those which I have found to be useful in Monte Carlo phase-space programmes.

5.1 Definition of statistical terms

Before considering the reduction of variance, we should clearly define what we mean by variance, since some confusion concerning notation exists in the literature.

We define the expectation of a function $g(y)$ as

$$\mathcal{E}g(y) = \int yg(y)\, dy ,$$

this is true if the y are a priori uniformly distributed, or have equal probabilities of occurring. If they are not uniformly distributed then,

$$\mathcal{E}g(y) = \int g(y)\, dF(y)$$

where F is the a priori distribution function of the y. The expectation of g is also known as the mean of g. We define the variance of g as:

$$\text{var}\{g(y)\} = \mathcal{E}\{[g(y) - \mathcal{E}g(y)]^2\} .$$

The variance is a measure of the dispersion of g about its mean. The standard deviation σ is defined as the square root of the variance. If we denote the mean by μ, we have:

$$\sigma = \sqrt{\text{var}}\{g(y)\} = \left[\int (yg(y) - \mu)^2\, dy\right]^{1/2} .$$

When we put a bar over a symbol, that will denote an estimate (for example, a Monte Carlo estimate) of the quantity. Suppose that an estimate $\bar{\mu}$ of μ exists. Then we define its standard error δ such that the probability

$$P(\bar{\mu} - \delta < \mu < \bar{\mu} + \delta) = 68\% .$$

In fact, an unbiased estimator $\bar{\mu}$ of μ is just the (unweighted) average of n random observations of the function $g(y)$, that is, with the different y_i chosen at random according to their a priori distribution.

$$\bar{\mu} = \frac{1}{n} \sum_i y_i g(y_i) .$$

The standard error on this estimate can be found by using the Central Limit Theorem. This theorem states that if a large number of random variables r_i are drawn from any distributions having means g_i and standard deviations σ_i^2, then the sum $\Sigma_i r_i$ is distributed normally with a mean equal to Σg_i and standard deviation equal to $\Sigma\sigma_i^2$. This powerful theorem gives us immediately the standard error on $\bar{\mu}$ which is

$$\delta = \sigma/\sqrt{n} .$$

This expression is exact, but since σ involves an integration over y, which is even more complicated than the integration necessary to find μ, we will usually have to settle for

an estimate of δ also. An unbiased estimator of the variance of g is

$$\overline{\text{var}}\ (g) = \frac{(\Sigma_i\ g^2\ (y_i) - n\bar{\mu}^2)}{n - 1} \quad .$$

The square root of this variance is normally taken as the estimate of the standard deviation $\bar{\sigma}$, even though this is now not an unbiased estimate (a function of an unbiased estimate is not in general an unbiased estimate of the function). This then allows us an estimate of the standard error δ:

$$\bar{\delta} = \frac{\bar{\sigma}}{\sqrt{n}} = \sqrt{\frac{\overline{\text{var}}\ (g)}{n}} \quad .$$

Since σ is a constant, $\bar{\delta}$ and δ diminish only as the square root of n, a result quoted earlier. Before going on, we will need one more result from probability theory. Suppose that our function g is a step-function, that is it takes on only the values 0 or 1. This is then the function of interest when considering hit-or-miss Monte Carlo as in our example earlier of the area of the circle. Then an unbiased estimator of g is the ratio of the number of successful trials ($g = 1$) to the total number

$$\bar{\mu} = n_1 / n \quad .$$

The standard error of this estimate is

$$\bar{\delta} = \sqrt{\bar{\mu}(1 - \bar{\mu})/n} \quad .$$

Since the following will concern only Monte Carlo integration (that is, estimates of the means of functions $\bar{\mu}$) we are primarily concerned with reducing the error

$$\delta = \sigma/\sqrt{n} \quad .$$

If we use truly random or pseudo-random numbers, there is nothing we can do about the $\sqrt{n}$. (We will discuss later the possibility of using so-called "quasi-random" or "uniformly-distributed" numbers to improve the convergence) and we can only work on reducing the variance σ of our sample.

5.2 Hit-or-miss Monte Carlo

It was this most primitive kind of Monte Carlo which was widely used in the early days of computers and which led to the wide discredit of random-sampling methods as grossly inefficient and unsophisticated. In our hit-or-miss calculation of the area of the circle, we were actually evaluating the two-dimensional integral

$$I = \int_0^1 \int_0^1 \epsilon(1 - x^2 - y^2)\ dx\ dy$$

where the function ϵ is defined as:

$$\epsilon(g) = 0 \text{ for } g < 0$$
$$\epsilon(g) = 1 \text{ for } g > 0 \ .$$

The variance for this method, by straightforward application of the formula given above, is

$$\frac{\pi}{4}(1 - \frac{\pi}{4}) \approx \frac{3}{16} \ .$$

We have stated, in Section 5.1, the error formula for hit-or-miss Monte Carlo integration as

$$\delta = \sqrt{\frac{\mu(1-\mu)}{n}}$$

where $\mu = n_1/n$ is the probability of a "hit". It can be seen that this error becomes vanishingly small under two conditions: either as $\mu \to 1$ (all hits) or as $\mu \to 0$ (all misses). If μ is close to one, this really means that the error is small, but if μ is very small, it is more meaningful to look at the percentage error:

$$\frac{\delta}{\mu} = \sqrt{\frac{\mu(1-\mu)}{\mu^2 n}}$$

$$= \sqrt{\frac{1}{\mu n} - \frac{1}{n}}$$

$$\approx \sqrt{\frac{1}{\mu n}} = \sqrt{\frac{1}{n_1}} \ .$$

Since μ is small, the first term in the radical on the second line above is much larger than the second term, and it approaches infinity as $\mu \to 0$, with the final error depending essentially only on the number of "hits". In fact we may state a general principle that in hit-or-miss Monte Carlo, the (relative) error is minimized if the probability of a hit is maximized. Moreover, if the probability of a hit is very small, additional errors may result from certain unfortunate properties of pseudo-random numbers [see Section 8.2 (ii)].

5.3 Crude Monte Carlo

If we perform one of the integrations in the above problem analytically, we turn it into a one dimensional integral:

$$I = \int_0^1 \sqrt{1 - x^2} \, dx \ .$$

In doing this, we have passed from hit-or-miss Monte Carlo to crude Monte Carlo because we have replaced a Monte Carlo estimate of a function by its known value. Clearly we must have reduced the variance by this technique. In fact the variance is now:

$$\int \left(\sqrt{1 - x^2} - \frac{\pi}{4} \right)^2 dx \approx \frac{1}{24} \ .$$

Translated into phase-space terms, this usually means that it is advantageous to use weighted events rather than to get rid of the weights by using a Monte Carlo technique.

5.4 Stratified sampling

It seems intuitively obvious that the random fluctuations in our Monte Carlo integrations would be reduced if we could somehow assure that points were more evenly distributed over the space. One clear way of doing this is to divide the space up into equal cells and to choose an equal number of points randomly within each cell. Indeed this usually will improve the variance, but in actual practice the situation is somewhat complicated by the following considerations:

i) If we are working in many dimensions the division of the space into equal hypercubes becomes unwieldy since the minimum number of such cubes is 2^k for a k-dimensional space.

ii) If we write down the expression for the variance in the case of stratified sampling (it is simply a complicated sum of squares of integrals and integrals of squares) we see that the amount gained -- or in fact whether we gain at all -- depends on the behaviour of the function, on the way we choose the cells, and on the way we distribute points among the different cells.

iii) Rather than choosing equal-sized cells, a better way is to choose them so that the variation of the function is about the same in each cell. Indeed we will gain if the differences between mean values of the function in the different cells are greater than the variations within the cells.

iv) Given the division into cells, one should choose the number of points in each cell in the following way: in the j^{th} cell choose

$$n_j^2 \propto \left[\Delta j \int_{\Delta_j} g^2(\vec{x})\, d\vec{x} - \left\{ \int_{\Delta_j} g(\vec{x})\, d\vec{x} \right\}^2 \right]$$

where Δ_j represents the volume of the j^{th} cell in the k-dimensional space.

v) If nothing is known a priori about the function the only natural way to choose the cells is to divide the space equally. It can be shown that if we do this, and if we choose equal numbers of points for each cell, we cannot lose by such a stratification (that is, the variance cannot be larger than for the unstratified estimate). In the worst case, where the mean value of the function within each cell is the same, the variance is not changed by the stratification. The next question is how many cells should we choose. Since we have seen that we cannot lose by subdividing a cell, we should choose as many cells as possible. However, if there is only one point per cell, we no longer have a reliable error estimate since we have no estimate of the variance within a cell. But we have an upper limit on the error, namely the error calculated as though the space had not been stratified.

5.5 Importance sampling

Another intuitive idea which comes to anyone who has done much Monte Carlo work is that it is a waste of time to play around in regions of the space where the contributions to the integral are small -- we should choose more points in regions where the contributions (function values) are larger. In phase-space terms, suppose we are integrating a matrix element which strongly favours low momentum transfers to the proton. Then all the time spent outside this low momentum transfer region may be lost if when we get a low transfer event it gives a contribution many times larger than all the rest together. In such a case it can be seen that the final error will essentially depend only on the number of points chosen within the important region. An obvious corollary to this principle is that we should spend no time in regions where the function is strictly zero (corresponding to momentum transfer cut-offs, for example). In actual practice the exclusion of such regions may be difficult because they are awkwardly shaped or disconnected.

We can make these ideas more precise by considering a simple integration

$$I = \int \sin \vartheta \, d\vartheta \, .$$

In Monte Carlo terms, this means we choose random ϑ, evenly distributed over the desired interval, and take the sum of the corresponding sines. But we could rewrite this as

$$I = \int \sin \vartheta \cdot \frac{\sin \vartheta \, d\vartheta}{\sin \vartheta} = \int \sin \vartheta \cdot \frac{d \cos \vartheta}{\sin \vartheta} \, .$$

What we have done in Monte Carlo terms is to decide to choose randomly $\cos \vartheta$ instead of ϑ, which means that we must then weight the corresponding function values by $1/\sin \vartheta$, which is the inverse of the Jacobian of the transformation from $\cos \vartheta \to \vartheta$. When we do this, we see that the weighted function value, $\sin \vartheta / \sin \vartheta$ is now constant, hence the variance of our estimate will be zero: no error. This is of course because we have already done the integration analytically in recognizing that $\sin \vartheta \, d\vartheta = d \cos \vartheta$.

In actual practice, of course, we cannot integrate these functions analytically, or else we would not need Monte Carlo. But it is already a great help in reducing the variance if we can find some integrable function which is similar in form to the desired function. Then applications of the above rule will lead to a nearly constant weighted function, and hence a small variance. This is called importance sampling since it corresponds to choosing a density of points such that more points are chosen in regions where the original (unweighted) function is larger. This is developed further in Section 11.3 below.

6. LIMITS OF INTEGRATION

One of the fundamental advantages of the Monte Carlo method is that it can handle easily problems in which the limits of integration are awkward and interdependent. However, we must be careful about how these limits are chosen, as the following example illustrates:

$$I = \int_{x=0}^{1} \int_{y=0}^{x} g(x,y) \, dy \, dx \, .$$

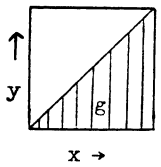
It is tempting to solve this problem in the following way:

a) choose a random number (x_i) between 0 and 1,

b) choose another random number (y_i) between 0 and x_i,

c) take the sum of $g(x_i,y_i)$, repeating (a) and (b).

A simple graphical representation of what we have done here shows that it gives the wrong answer:



While it is true that this procedure would yield points only in the allowed region (the lower triangle), it would give the same number of points along each vertical line. This gives a much higher density of points on the left-hand side than the right-hand side.

A correct way to integrate the above problem is:

a) choose a random number (x_i) between 0 and 1;

b) choose another random number (y_i) between 0 and 1;

c) if $y_i > x_i$, reject the point;

d) take the sum of $g(x_i,y_i)$ over the remaining points.

This method, although correct, has the disadvantage of using only half the points generated. That is, it is equivalent to integrating over the whole square, but considering the function value is equal to zero in the upper triangle. A better way to handle this problem is the following:

a) choose two independent random numbers, r_1, r_2;

b) set x_i = larger of r_1, r_2;

c) set y_i = smaller of r_1, r_2;

d) sum up $g(x_i,y_i)$ as before.

Graphically, this is equivalent to choosing points randomly over the square, then folding the square about the diagonal so that all points fall into the lower triangle. It is clear that this now results in a constant density of points without rejecting any points.

6.1 Rejecting points

If a function is zero over a certain region, this always results in an increase of its variance, as can be seen in the following way: consider the standard deviation (square root of the variance) of the function calculated only over the non-zero region, σ_{nz}, and the standard deviation calculated over the whole space, σ_{tot}. Then the error in the Monte Carlo integral of the function can be considered as

$$\delta = \frac{\sigma_{nz}}{\sqrt{n_1}}$$

where n_1 is the number of hits (non-zero function values). The same error could also be calculated by

$$\delta = \frac{\sigma_{tot}}{\sqrt{n}}$$

where n now includes both hits and misses. Setting the two errors equal, we have

$$\frac{\sigma^2_{tot}}{\sigma^2_{nz}} = \frac{n}{n_1} > 1 \ .$$

In the example of Section 6 above (integration over a triangle) a particular trick was presented, which would reduce the error by a factor of $\sqrt{2}$ since it would result in twice as many useful points. However, if we think of the calculation on a large computer, we see that if all points in the upper triangle were rejected, this would amount to very little extra time if the time required for the test (is x_i greater than y_i?) is short compared with the time necessary to evaluate the function. That is, we are still wasting our time in the upper triangle, but we may only be wasting a negligible amount of time, even though we are throwing away half the points.

7. INTEGRATING DELTA FUNCTIONS

Most well-behaved functions can be integrated in a satisfactory manner by straightforwardly applying the Monte Carlo method and using standard variance-reducing techniques. There is however one class of functions which occurs in many physical problems and especially in the phase-space expression, which stubbornly resists all variance-reducing techniques: namely the delta function, which is everywhere zero except for one infinitely sharp peak. We first calculate the variance of such a function. Consider a variable width Gaussian function

$$G(\alpha,x) = \frac{\alpha}{\sqrt{\pi}} e^{-\alpha^2 x^2}$$

where x is considered as the independent variable and α is a constant parameter which gives the sharpness of the peak. It has been normalized so that

$$\int_{-\infty}^{\infty} G(\alpha,x)dx = 1 \quad \text{for all } \alpha \ .$$

In the limit as $\alpha \to \infty$, this function G takes on the properties of a delta function. We may therefore calculate the variance of G as a function of α, then take the limit of this variance as $\alpha \to \infty$. The mean value of the function is clearly zero since its integral is unity over an infinite interval. (If one takes a finite interval one gets the same result for the variance in the limit $\alpha \to \infty$). One then has for the variance

$$\text{var}\ \{G(\alpha,x)\} = \int_{-\infty}^{\infty} G^2(\alpha,x)\ dx$$

$$= \int_{-\infty}^{\infty} \frac{\alpha^2}{\pi}\, e^{-2\alpha^2 x^2}\ dx$$

$$= \frac{\alpha}{\sqrt{2\pi}}\ .$$

Therefore we can say that the variance of a Gaussian is proportional to the sharpness of the peaking, as expressed by the parameter α, and in the limit of a delta function this variance is infinite. The rule can easily be extended to Breit-Wigner functions, which have somewhat different shapes but are generally single peaks and will also have variances which become very large as the peak becomes very narrow. Since the functions we will integrate in phase-space problems will very often be Breit-Wigner shapes, we must recognize the trouble they will give in the case of very narrow resonances.

Returning now to the delta function, there is a simpler way of seeing that its variance is infinite. As we have seen before, when a function is zero over a certain region, the error on its integral can be considered to depend only on the number of non-zero points chosen. But since the probability of making a "hit" on a delta function is zero, we will always have an infinite error. (See Section 5.2).

Historically, the delta function has been handled by approximating it by a narrow peak of non-zero width. This means introducing an approximation into the theory which is essentially different from the approximation resulting from the statistical fluctuations of Monte Carlo results. The "wide delta-function" approximation can lead to completely unphysical answers unless the results are interpreted carefully. For example, this approximation in phase-space calculations means allowing momentum and energy conservation to be violated by a small amount in each event.

But a much more satisfactory way of dealing with delta functions is to get rid of them analytically. Delta functions are usually defined by their properties under integration, and in fact their integration is trivial if the variables of integration are separable. For example, we have

$$\iint f(x,y)\delta(y-y_0)\ dx\ dy = \int f(x,y_0)\ dx$$

(where δ is the delta function and should not be confused with our earlier use of this symbol to denote the error on a Monte Carlo estimate.) We see that the effect of the delta function is to reduce the dimension of the space by one. In this case it changes an integral over the (x,y) plane into an integral along the line $y = y_0$.

Now consider a problem in which the variables are not separated.

$$I = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y)\ \delta(R^2 - x^2 - y^2)\ dx\ dy.$$

By inspection we see that this is just the integral of $f(x,y)$ over a circle of radius R centred at the origin. We therefore transform to polar coordinates:

$$x = r \cos \vartheta$$

$$y = r \sin \vartheta$$

$$dx\ dy = \frac{\partial(x,y)}{\partial(r,\vartheta)}\ dr\ d\vartheta = r\ dr\ d\vartheta$$

where the variables are now separable:

$$I = \int_{r=-\infty}^{\infty} \int_{\vartheta=0}^{2\pi} f(r \cos \vartheta,\ r \sin \vartheta)\ \delta(R^2 - r^2) r\ dr\ d\vartheta$$

$$= \int_{\vartheta=0}^{2\pi} f(R \cos \vartheta,\ R \sin \vartheta)\ R\ d\vartheta\ .$$

8. RANDOM NUMBERS

We will distinguish three different types of sequences of random numbers, namely:

i) truly random numbers are those which are chosen in such a way that at any given point in the sequence all numbers are equally probable and independent of the preceding numbers.

ii) pseudo-random numbers are those which are generated according to an arithmetic prescription so that each number depends on the preceding one, but in such a way that any finite sequence (up to a certain maximum length) satisfies (nearly) the same statistical tests as a truly random sequence.

iii) quasi-random numbers are generated according to an arithmetic prescription which results in certain strong correlations between the numbers in any short sequence, but in such a way that certain asymptotic properties of the distributions are more advantageous than the corresponding properties of truly random or pseudo-random sequences.

We have seen that any Monte Carlo calculation involves the use of a sequence of "random" numbers; in fact this may be considered as the definition of a Monte Carlo calculation. The choice of a proper sequence of "random" numbers is therefore central to the method, and, in spite of great progress in the field, this choice is unfortunately not at all trivial. We shall discuss in more detail the three main classes of "random" sequences:

8.1 Truly random numbers

A sequence of truly random numbers is unpredictable and unreproducible. Such a sequence can only be generated by a random physical process, for example beta decay. Although such a process yields, in principle, the most random numbers possible, the technical problems involved in eliminating biases in detection equipment turn out to be very great. In addition it is hard to generate truly random numbers fast enough and with great enough accuracy to be of use in modern computers. Faced with these difficulties, people have largely abandoned trying to generate truly random numbers and the technology has not advanced significantly in recent years.

8.2 Pseudo-random numbers

It is in the field of pseudo-random numbers that recent progress has been the greatest, but the situation is still far from ideal. The great advantage with pseudo-random numbers is that calculations may be checked or compared by regenerating the identical sequence. Also, the pseudo-random sequences used for large computers are very fast and require very little storage and no special hardware (a slight exception is the generators which can be initialized by the computer's internal clock, thereby allowing "random" entry points if independent sequences are desired). However, pseudo-random generators are plagued by certain fundamental problems:

i) Since each number is generated from the preceding one, the recurrence of one number results in the recurrence of an entire sequence. Each generator is therefore characterized by a certain period and most of the progress in finding better generators consists in finding generators with longer periods. Nowadays one can generally achieve the maximum period (limited by the word length in the computer) but for machines like the CDC 6600 or 6400 this period is so long that there is no point in going to the limit.

ii) Again since each number is generated from the preceding number by a fixed formula, if any number ξ_j is arbitrarily close to some preceding number ξ_i, then ξ_{j+1} will be arbitrarily close to ξ_{i+1}. Exactly what is meant by arbitrarily close depends on the generator and some are better than others in this respect, but in general this means that it is difficult to avoid having some serial correlation in a pseudo-random generator. Such a correlation is exceedingly dangerous in the case of a sharply peaked weighting function (discussed in Sections 5.2 and 7) since this essentially imposes the condition that the corresponding pseudo-random number must be in a very narrow range in order to get a good event. If the following number is then correlated with it, spurious results will be produced.

iii) In order to show that a particular pseudo-random generator is acceptable, it must be submitted to an infinite number of statistical tests, which is clearly impossible. In practice the generators are submitted to increasingly more complicated tests until one finds the lowest order correlation existing (i.e. the simplest test which is failed) at which time it is said that the principal correlations (unrandomness) are understood. It is generally agreed that it is better to use a generator whose correlations are understood than an unknown one which may be better. Generally good pseudo-random generators exist in the libraries of all large computer centres, but they are not always suitable for all Monte Carlo problems.

8.3 Quasi-random generators

The principle of quasi-random generators is that certain correlations are harmless (depending of course on the problem) and indeed they can sometimes be helpful. In fact the development of suitable quasi-random generators probably constitutes the most important recent progress in phase-space integration, since it is only in this way that one can improve on the $n^{-1/2}$ convergence of Monte Carlo estimates.

For an integration in k dimensions we require sets of k random numbers, each set giving rise to a point in the space. The k numbers within each set must be uncorrelated, for if any number in a given set depends on another number in the same set, this restricts the possible combinations, making certain regions of the space inaccessible. However, certain correlations between successive sets of numbers do not bother us. For example, if each point is chosen close to the preceding point (an obvious non-randomness), this does not matter as long as the final sample is evenly distributed over the space. This is equivalent to saying that the order in which the points are chosen is of no importance since this does not normally enter into such calculations.

This leads us to study uniformly-distributed numbers, a subset of quasi-random numbers which have the property that the density of points is more uniform than that of truly random numbers. There is a very strong serial correlation between successive numbers produced by the same generator, so we need in general k different generators for an integration in k dimensions. A point in phase space is then chosen by taking one "random" number from each of the k generators.

Two different types of quasi-random generators have been studied in detail and seem well-suited to Monte Carlo integration problems. Each type has as many different generators as there are prime numbers. Integrals calculated with these generators converge faster than 1/n, a considerable improvement over pseudo-random numbers.

i) The Richtmyer formula gives, for the i^{th} "random" number of the j^{th} generator:

$$r_{ij} = i\ S_j, \text{ modulo } 1.$$

where S_j is the square root of the j^{th} prime number. This means that the difference between successive numbers of the same generator is always S_j or $S_j - 1$. This also leads unfortunately to strong short-term correlations (i.e. correlations which go away as $n \to \infty$) between corresponding numbers produced by different generators, which makes this method a dangerous one in practice.

ii) The Van der Corput formula consists of expressing the integers in a system of base P, then reversing the digits, preceding them by a point, and interpreting the resulting numbers as fractions in a system of base P. P is any prime number. For example, in the binary system (P = 2):

decimal	binary	binary fraction	decimal fraction
n = 1	1	0.1	0.5
2	10	0.01	0.25
3	11	0.11	0.75
4	100	0.001	0.125
5	101	0.101	0.625
6	110	0.011	0.375
7	111	0.111	0.875
8	1000	0.0001	0.0625

Generators of this form can be made fast, and in general seem to produce much less serious correlations than those of the Richtmyer generators, but with similarly good convergence properties.

BIBLIOGRAPHY FOR PART I

Some good general books on statistics:

H. Cramer, Mathematical methods of statistics (Princeton Univ. Press, 1946).
E. Sverdrup (North-Holland, Amsterdam, 1967), Laws and chance variations (2 vol. transl. from Norwegian).
S. Wilks, Mathematical statistics (Wiley, N.J., 1962).

General Monte Carlo and random numbers, with bibliography:

J.M. Hammersley and D.C. Handscomb, Monte Carlo methods (Methuen, London, 1964).

Quasi-random numbers (Van der Corput method):

J.H. Halton, Numerische Mathematik 2, 84 (1960).

Monte Carlo applied to a certain class of phase space problems:

F. Cerulus and R. Hagedorn, Nuovo Cimento 9, 646 (1958).

Numerical integration:

C.B. Haselgrove, Math.Comp. 15, 323 (1960).
A.H. Stroud and D. Secrest, Gaussian quadrature formulas (Prentice Hall, N.J., 1966).
J.N. Lyness, Math.Comp. 19, 625 (1965).

PART II

MONTE CARLO APPLIED TO THE PHASE-SPACE INTEGRAL

9. THE PHASE-SPACE INTEGRAL

We define the n-body phase space integral as:

$$R_n = \int^{4n} \delta^4 \left(\underline{P} - \sum_{j=1}^{n} p_j \right) \prod_{i=1}^{n} \delta(p_i^2 - m_i^2)\, d^4 p_i \qquad (9.1)$$

where

$\underline{P}$ is the total four-vector of the n-body system

p_i are the four-vectors of the individual particles

m_i are the masses of the particles.

If we then consider any kinematic parameter (a momentum, angle, etc.) of the system, its spectrum will be given by

$$f(\alpha) = \frac{d}{d\alpha} \left(|\text{M.E.}|^2 \cdot R_n \right) \qquad (9.2)$$

where M.E. is the matrix element describing the interactions between the particles. If there is no interaction between the outgoing particles, M.E. = 1, and we say that all spectra are given by phase-space alone. Of course, if the expression for R_n is wrong, we can always fix things up by adjusting M.E. to fit the data, but this whole analysis has meaning only if R_n describes the major features of the spectra correctly. In particular it is supposed to describe exactly the features determined by statistical and kinematic factors, which means:

i) statistical factor is the density of states: one expects a higher probability of finding an event in a "region" where there are more available states, that is, there is more phase space. This comes into the expression in the product of $d^4 p_i$, which, with the energy conservation delta function, can be expressed in spherical polar coordinates as

$$\delta(p_i^2 - m_i^2)\, d^4 p_i = \frac{p_i^2}{E_i}\, dp_i\, d\cos\vartheta_i\, d\varphi_i \qquad (9.3)$$

(for derivation, see Hagedorn[1], pages 89 and 45) where p_i is the absolute value of the three-momentum. The above equation expresses the fact that when the momentum of a particle is greater, there are more ways for it to arrange itself: more available states.

ii) kinematical factor is expressed by the delta functions which assure that momentum and energy are conserved. The correlations between particle momenta which follow from four-momentum conservation are considered as fundamentally different from those which arise from the particles' interactions (M.E.).

It has been remarked that if we wish really to include all statistical and kinematic factors into R_n, then we should put in delta functions expressing all the known conservation laws, such as (for the strong interactions) conservation of leptons, baryons, isotopic spin, parity, and angular momentum. Most of these we take account of implicitly because we do not calculate phase space for unallowed reactions. However, angular momentum poses a special problem because it leads to truly kinematic effects, which are difficult to handle for the following reasons:

i) one does not generally know what angular momentum states are present in the initial state;

ii) it is not generally possible to find variables in which to express both linear and angular momentum in a separable way.

The general problem of determining kinematic effects due to angular momentum conservation is not solved. It is customary to treat the problem in the following way:

i) for two-body (elastic scattering) problems, the conservation of linear momentum is trivial and one can concentrate on expressing the angular momentum states directly; this is usually accomplished by a development in spherical harmonics which are eigenstates of the angular momentum operator;

ii) for three-body problems the complications increase enormously, but this has recently been solved by Zemach[2];

iii) for n-body problems, $n > 3$, the possibility of an exact solution appears hopeless, and one usually either neglects angular momentum conservation or considers that the effects are to be included not in phase space (R_n) but in the matrix element (M.E.).

For further details on the physical justification of the use of the phase space integral, we refer the reader to the excellent review article by Kretschmar[3], and we concentrate here on the various methods which have been used to calculate this integral.

9.1 The first ideas. Fermi (1950)

In his original paper[4], Fermi bases his theory on an analogy with classical thermodynamics and considers that the particles are in a short-lived state of thermodynamic equilibrium. By assuming a certain interaction volume, the theory predicts particle multiplicities and cross-sections as a function of incident energy. Nowadays, however, this is considered to be the least reliable part of the theory, and it is mostly used to find momentum spectra and correlations within a given reaction. (However, it should be mentioned that in testing SU_3 predictions of cross-section ratios, the phase space factor is often important.) Fermi realized that the evaluation of R_n would pose calculation problems and proposed the following approximations:

i) nucleons are non-relativistic: $E = p^2/2m$

ii) pions are ultra-relativistic: $E = p$.

By getting rid of the square root in the delta function, this allowed some calculations, but the approximation is really too crude. This is a general problem of kinematics in the region 100 MeV to 5 GeV: we are forced to carry out the full relativistic treatment because our particles are neither non-relativistic nor ultra-relativistic.

9.2 Possibilities of exact calculation. Block (1956)

Block[5] has done a systematic treatment of the phase-space integral from the point of view of exact analytic integration, and his paper indicates at what point (four particles) the calculations become prohibitively complex. This is of course done with a constant matrix element.

9.3 The Monte Carlo idea. Kopylov (1958)

Although Monte Carlo was already in wide use for nuclear physics problems, especially in connection with shielding and nuclear bomb development, it was probably Kopylov[6] who first formally proposed Monte Carlo as a means of solving the Fermi phase space problem. Having no computers at his disposal he nevertheless was able to produce by hand a set of two hundred random events (which he called a list of stars) from which he formed desired distributions. However, efficient Monte Carlo methods had to await further analytic simplification.

9.4 The recurrence relation. Srivastava and Sudarshan (1958)

A fundamental difficulty with the original Fermi formulation (which we have not written here) is the fact that it was not written in a Lorentz invariant way (three-momenta were used instead of four-momenta). It was Srivastava and Sudarshan[7] who produced the covariant formulation given above [Eq. (9.1)] which allows one to express R_n in any frame of reference. This in turn allows one to express the $(n+1)$-body phase space in terms of the n-body phase space, each one being calculated in its own centre of mass. We write, as in Eq. (9.1):

$$R_n(P; m_1 \ldots m_n) = \int \delta^4 \Big(\sum_{i=1}^{n} p_i - P \Big) \prod_{j=1}^{n} \delta(p_j^2 - m_j^2)\, d^4 p_j, \tag{9.4}$$

where we have written explicitly the quantities upon which R_n depends. Actually it depends only on the length of the four-vector P (i.e. the total energy in the centre of mass, or invariant rest mass) from Lorentz invariance. If we rewrite the argument of the first delta function and split off the last factor in the indicated product, we obtain:

$$R_n(P; m_1 \ldots m_n) = \int \Big\{ \int \delta^4 \Big(\sum_{i=1}^{n-1} p_i - (P - p_n) \Big) \prod_{j=1}^{n-1} \delta(p_j^2 - m_j^2)\, d^4 p_j \Big\} \delta(p_n^2 - m_n^2)\, d^4 p_n. \tag{9.5}$$

It is seen by inspection that the factor in curly brackets is just $R_{n-1}(P - p_n; m_1 \ldots m_{n-1})$. Now using Eq. (9.3), we have the important relation

$$R_n(P; m_1 \ldots m_n) = \int R_{n-1}\,(P - p_n; m_1 \ldots m_{n-1})\,\frac{d^3p_n}{2E_n}\,. \qquad (9.6)$$

At first glance, this looks similar to the expressions used by Block[5] and Cerulus and Hagedorn[8]; however, one must realize that the R_{n-1} in Eq. (9.6) is Lorentz-invariant, which means that the integration indicated is only over one parameter, and that if repeated use is made of this relation, each R_i may be the same algebraic function since it may be evaluated in its own rest frame. In other work up to 1958, each R_i had to be evaluated separately in the over-all centre of mass.

9.5 The T-generator programs

Equation (9.6) can be used as a basis for a class of Monte Carlo event generators which I will call the T-generators (T for kinetic energy). They were the first reasonably efficient generators to be used and we mention them here because they are still in widespread use, but we do not go into great detail since they are not the best generators for general problems. Most of the early work was published only in internal reports, so it is not clear who is the originator of this method, but descriptions of it can be found in the works of Lynch[9] and of Goldhaber et al.[10].

Equation (9.3) is applied to Eq. (9.1) to get rid of the mass delta functions, and the result is:

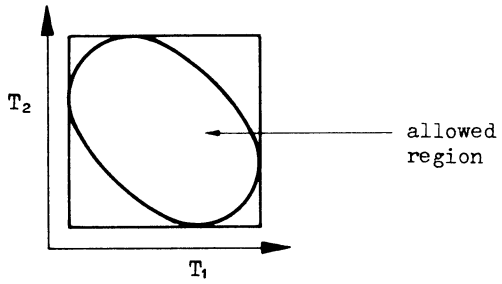
$$R_n = \int \frac{d^3p_1\; d^3p_2 \ldots d^3p_n\; \delta^3\,(\Sigma_i \vec{p}_i)\; \delta(E_o - \Sigma_i E_i)}{E_1 E_2 \ldots E_n}\,. \qquad (9.7)$$

Let us see what happens if we try to apply Monte Carlo immediately using this expression. We would first calculate the minimum and maximum kinematically allowed values for each component of each momentum and choose them all randomly between these two limits, except for the last four components, which are "determined" by the integrations over the delta functions. That is, we see if there are four numbers which, when added to the 3n-4 values already chosen, will conserve energy and momentum. If no such set exists (which will usually be the case) we choose another set of 3n-4 random momentum components. When we find a set which lands in the physical region, we use it, and give it a weight equal to the inverse of the product of the particle energies [the denominator of Eq. (9.7)]. In practice such a method would be extremely inefficient because, in choosing the momentum components at random between $-p_{max}$ and $+p_{max}$, we would often get momenta whose absolute values were greater than p_{max}, meaning that the event would automatically fail the conservation laws. One therefore transforms Eq. (9.7) into a hybrid polar coordinate system, where each set of angles ($\cos\vartheta$ and φ) is chosen in a different rest frame, corresponding to repeated applications of the recurrence relation Eq. (9.6). (These angles are the same as those of Section 9.6 which will be discussed in more detail.) A further transformation from momentum to kinetic energy is usually made (using $p\,dp = T\,dT$) and one arrives at an expression of the form

$$\int \left(\frac{p_1 p_2 \cdots p_{n-2}}{\sum_{i=1}^{n-2} p_i} \right) dT_1\, dT_2 \cdots dT_{n-1}\, d\varphi_1\, d\varphi_2 \cdots d\varphi_{n-1} \times d\cos\vartheta_1\, d\cos\vartheta_2 \cdots d\cos\vartheta_{n-2}\,, \tag{9.8}$$

where the remaining four variables (T_n, φ_n, $\cos\vartheta_{n-1}$, $\cos\vartheta_n$) are chosen (if possible) in an obvious way to balance energy and momentum.

For the three-body case, this method can be visualized quite easily. Choosing T_1 and T_2 corresponds to choosing a random point on the Dalitz plot of the three-body decay, for which the physical region looks something like the figure below:



Although the exact shape of the Dalitz plot boundary depends on the total energy and on the three particle masses, it generally covers a little more than half of the circumscribed rectangle within which the points are chosen, so that something less than half of the points must be rejected. The three remaining coordinates are φ_1, φ_2, and $\cos\vartheta_1$, which correspond to a complete orientation of the plane of the event. We could, for example, have chosen instead the Euler angles familiar to all students of rigid body mechanics.

Unfortunately for the T-generators, the efficiency goes down drastically with increasing numbers of particles in the final state, and it becomes quite impossible to use this method for more than five or six particles.

9.6 The M-generator programs

The recurrence relation (9.6) which was used as a starting point to get to the T-generators, can be used also to find a more efficient method, which I will call the M-generator method since it works essentially with invariant masses as kinematic coordinates. Again it is not clear who was the discoverer of this technique, but it seems to have been developed independently by Kopylov[11] in the Soviet Union and by Raubold and Lynch (unpublished) at CERN. Lynch incorporated this method into his phase space program OWL for the CERN 7090 library. The same generator is used for the program FOWL (CERN Program Library).

A different form of recurrence relation can be derived, starting from the original expression (9.1), in which we take full advantage of the Lorentz invariance of the expression. We first express the four-momentum conservation delta function as an integral over two delta functions.

$$\delta^4\Big(\sum_{j=1}^{n} p_j - P\Big) = \int \delta^4\Big(P - P_\ell - \sum_{j=\ell+1}^{n} p_j\Big)\ \delta^4\Big(P_\ell - \sum_{j=1}^{\ell} p_\ell\Big)\ d^4P_\ell \ . \tag{9.9}$$

The above relation can be seen to be true since it is of the form

$$\begin{aligned}\delta(x - y) &= \int \delta[(y - c) - z]\ \delta[z - (x - c)]\ dz \\ &= \delta[(x - c) - (y - c)] \ .\end{aligned} \tag{9.10}$$

Applying Eq. (9.9) to Eq. (9.1) we obtain:

$$\begin{aligned}R_n(P;\ m_1 \ \ldots\ m_n) = \int \delta^4 \Big(P - P_\ell - \sum_{j=\ell+1}^{n} p_j\Big) \prod_{j=\ell+1}^{n} \delta(p_j^2 - m_j^2)\ d^4p_j \\ \times \int \delta^4 \Big(P_i - \sum_{j=1}^{\ell} p_\ell\Big) \prod_{j=1}^{\ell} \delta(p_j^2 - m_j^2)\ d^4p_j\ d^4P_\ell \ .\end{aligned} \tag{9.11}$$

While P_ℓ was introduced in Eq. (9.9) as a dummy integration variable, we see that it is beginning to take on the role implied by its name: the four momentum of the ℓ-body subsystem within the n-body system. However, the integral is not yet divided into exactly corresponding factors, since there is one delta function missing in the d^4P_ℓ integration, namely $\delta(P_\ell^2 - M_\ell^2)$. Following Hagedorn[1], we introduce the factor 1 of the form

$$1 = \int_0^\infty \delta(P_\ell^2 - M_\ell^2)\ dM_\ell^2 \tag{9.12}$$

which gives us the desired expression:

$$\begin{aligned}R_n(P;\ m_1 \ \ldots\ m_n) = \int_0^\infty \Big\{ \int \delta^4 \Big(P - P_\ell - \sum_{\ell+1}^{n} p_j\Big) \prod_{\ell+1}^{n} \delta(p_j^2 - m_j^2) \\ \times \delta(P_\ell^2 - M_\ell^2)\ d^4p_j\ d^4P_\ell \int \delta^4\Big(P_\ell - \sum_{1}^{\ell} p_j\Big) \\ \times \prod_{1}^{\ell} \delta(p_j^2 - m_j^2)\ d^4p_j \Big\}\ dM_\ell^2\end{aligned} \tag{9.13}$$

which can be rewritten as the important "splitting" relation

$$R_n(P;\ m_1 \ \cdots\ m_n) = \int_0^\infty R_{n-\ell+1}\ (P;\ M_\ell, m_{\ell+1} \ \cdots\ m_n)\ R_\ell(P_\ell;\ m_1 \ \cdots\ m_\ell)\ dM_\ell^2 \ . \qquad (9.14)$$

From Eq. (9.12) we see that M_ℓ^2 is the square of the four momentum of the ℓ particles, which is just their invariant rest mass. The limits of integration for M_ℓ^2 are:

$$\left(\sum_{i=1}^{\ell} m_\ell\right)^2 \le M_\ell^2 \le \left(M_n - \sum_{i=\ell+1}^{n} m_i\right)^2 \ . \qquad (9.15)$$

Repeated applications of the splitting relation (9.14), starting with $\ell = 2$, lead us to the recurrence relation:

$$R_n = \int dM_{n-1}^2 \ \cdots \int dM_2^2 \prod_{i=1}^{n-1} R_2(M_{i+1};\ m_i, m_{i+1}) \qquad (9.16)$$

where the limits of integration of Eq. (9.16) are discussed below. Here R_2, the invariant two-body phase space factor, is[1)]

$$R\ (M_{i+1};\ M_i, m_{i+1}) = \frac{2\pi}{M_{i+1}} \sqrt{M_{i+1} + \left(\frac{M_i^2 - m_{i+1}^2}{M_{i+1}}\right)^2 - 2(M_i^2 + m_{i+1}^2)} \ . \qquad (9.17)$$

Transforming from dM^2 to $2M\ dM$ and rearranging, we get the final form used by the M-generators:

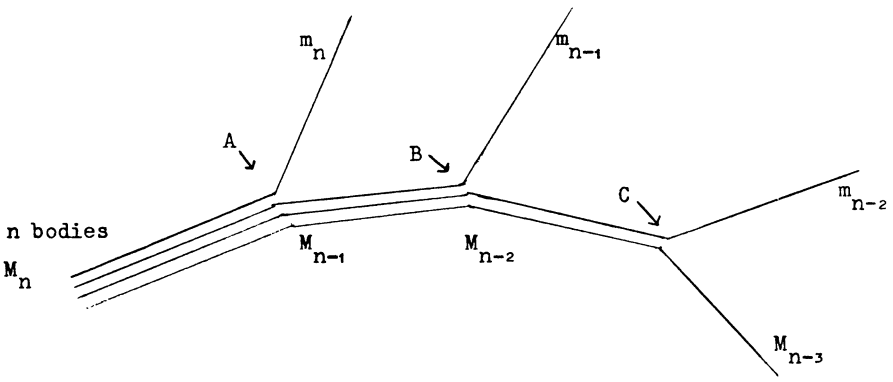
$$R_n = \frac{1}{m_1} \int\!\!\int \prod_{i=1}^{n-1} \left\{ 2\ M_i\ R_2\ (M_{i+1};\ M_i, m_{i+1}) \right\} dM_{n-1} \ \cdots\ dM_2 \ . \qquad (9.18)$$

There remain two principal problems to solve before this formula can be used for Monte Carlo. These are:

i) What are the limits of integration on M? Can they be chosen in an efficient way?

ii) We should have 3n-4 variables of integration, but we have only n-2. What happened to the other 2n-2?

The answer to these two problems can be seen more easily if we represent pictorially the splitting of the n-body problem into (n-2) two-body cases.



(9.19)

Equation (9.18) tells us that the problem can be treated as though it were the sequential decay shown in Eq. (9.19). Namely, we need apply only the two-body phase-space factor at the first vertex (A), but this must be integrated over all possible masses M_{n-1}. Then we apply the two-body factor at B, integrating now over all possible masses M_{n-2} and all possible M_{n-1}, since the "initial state energy" at B is now variable also. Similarly for the other "vertices".

Now the physical "limits of integration" of the M in Eq. (9.18) are just the limits which will make each two-body vertex exothermic. That is, for each M_j we must have

$$M_{j-1} + m_j < M_j < M_{j+1} - m_{j+1} \ , \tag{9.20}$$

and at first glance we seem to be back in the position we were in with the T-generators, since the limits of the j^{th} integral are not constant but depend on the other integrals. We have seen in Section 6 with integration over a triangle that we cannot choose the limits of one integral to be dependent on the choice of variable for the preceding integral in such a situation. The M_j must be chosen independently, which means they must be chosen to satisfy only the less restrictive condition:

$$\sum_{i=1}^{j} m_i < M_j < M_n - \sum_{i=j+1}^{n} m_i \tag{9.21}$$

where M_n is the total energy of the system in its rest frame, so that only constants enter into the choice of M_j. But we know that the event will be non-physical unless the M_j also satisfy the more restrictive condition (9.20). The usefulness of the method lies in the fact that this can be assured in a simple way. Consider the invariant masses M_j chosen according to the weaker conditions (9.21):

$$m_j = r_j \left(M_n - \sum_{i=j+1}^{n} m_i \right) + \sum_{i=1}^{j} m_j$$

where r_j is a random number between zero and one. Then it happens that the necessary and sufficient condition that the M_j chosen in this way will also satisfy the more restrictive limits (9.20) is just:

$$0 < r_1 < \ldots < r_j < r_{j+1} < \ldots < r_{n-2} < 1 ,$$

that is, the random numbers must be in ascending order. Just as in the case of the triangle integration of Section 6, our random numbers may be ordered without biasing the result, as long as they were originally chosen independently. Or else we may simply throw away each set which is not ordered. By ordering the random numbers we finally achieve what we have been looking for: an unbiased event generator with every event falling in the physical region.

There is still the problem of the missing variables, but this is easily solved by again considering the process of diagram (9.19). At each two-body vertex we have the energy variation but nothing else. There clearly remain two angular variables in order to complete the description of each vertex. If we take each pair of angles to be defined in the rest frame of that vertex, then we need only choose them isotropic in $\cos\vartheta$ and φ as we know that two-body phase space is isotropic in angular space. However, since this is only true for a two-body decay in its centre of mass, we need the Lorentz invariance of R_2 (M_j, m_j, m_{j-1}) which is a function only of invariant quantities. To obtain a final description of the event in the over-all centre of mass, we must successively Lorentz-transform each momentum into the rest frame of the group of particles preceding it.

10. PHASE SPACE SPECTRA AND MATRIX ELEMENTS

We have seen in Section 9 how points may be chosen randomly in phase space. The question of what to do with these points is rather obvious, and this was discussed briefly earlier. But it is perhaps useful here to state more fully the ways in which these points or events may be used since this really demonstrates the power of the Monte Carlo method.

10.1 Integration versus direct simulation

Let us return for a minute to the analytic (non-Monte Carlo) statement of the phase-space problem. We are usually looking for one of two types of answers:

i) The absolute value of a phase-space integral: this allows us to calculate, for example

$$\frac{\Gamma(\bar{p}p \to \pi\pi\pi)}{\Gamma(\bar{p}p \to \pi\pi\pi\pi)} = \frac{R_3(1.876;\ \pi,\pi,\pi)}{R_4(1.876;\ \pi,\pi,\pi,\pi)} .$$

This says that if antiproton annihilations at rest follow pure phase space, the ratio of 3π events to 4π events will be given by the above expression. This is just the ratio of two integrals, and we have seen how R_n can be evaluated by Monte Carlo integration. (In fact, the expressions of Section 9 are not complete for this purpose since we have dropped some constants. See Section 12.1).

ii) The spectrum, or distribution of events with respect to some kinematic parameter. This we have given as Eq. (9.2) which has a misleadingly simple form. To imply the complexity of calculation involved, we should rather express it as:

$$f(\alpha) = \int f(\alpha,\vec{\beta})\ d\vec{\beta} = \frac{d}{d\alpha} \int |M.E.|^2\ dR_n \qquad (10.1)$$

where α is one kinematic parameter among the whole set necessary to describe an event, and the $\vec{\beta}$ are the rest of the parameters, all orthogonal to α. Analytically then, the problem would be not only to calculate R_n but also to express it in such a system of variables that the integration and differentiation can be performed. The matrix element must of course be expressible in the same variables, and for nearly all interesting problems, this is impossible. With Monte Carlo, however, we have not to express the integrand in any special way; we need only know its value at each point. The integration is then straightforward. However we have not yet seen how to do a differentiation. In fact we do it approximately by noting that:

$$\lim_{\Delta\alpha \to 0} \frac{1}{\Delta\alpha} \int_{\Delta\alpha} f(\alpha)\ d\alpha = f(\alpha)\ . \qquad (10.2)$$

That is, we will divide the total range of α into bins and calculate the integral of $f(\alpha)$ over each bin. An "event" contributes to the integral if it has a value of α in the desired range. If the bins are small enough, the value of the integral over each bin will approach the function value by Eq. (10.2). Even if the bins are not small, we still have something we can compare directly with experiment since this is in fact how experimental spectra are obtained from individual observations.

It should therefore be clear that the end result of all our phase space calculations is the value of an integral or a set of integrals. However, we do this by generating points in phase space, "or events", and so we may alternatively consider that we are directly simulating an experiment. The only problem is that our events are not all equal, but come with intrinsic weights. However, if this bothers us, the weights can be eliminated (see Section 11.2).

10.2 Jacobian of the transformation to mass space

If our Monte Carlo calculation is considered as direct simulation, it is proper to ask what is the meaning of the weight of the event in this context. From Section 9, we have made a transformation from momentum space to mass space in order to gain efficiency and avoid delta functions. Since the density of points is not the same in these two spaces, we must weight the events inversely as the Jacobian from momentum space to mass space. Since a Jacobian is the determinant of a derivative matrix, it would seem that we could have avoided all the derivation of Eq. (9.6) simply by calculating some derivatives. The answer is that because of the delta functions we do not have a proper Jacobian here, but something equivalent in the sense that if

$$\int f(\vec{\alpha})\ d\vec{\alpha} = \int f(\vec{\alpha}(\vec{\beta}))\ J(\vec{\beta})\ d\vec{\beta}\ , \qquad (10.3)$$

then J is the Jacobian

$$\left| \frac{\partial(\vec{\alpha})}{\partial(\vec{\beta})} \right|$$

However we have had to make use of the integral relation (10.3) since the derivatives of delta functions do not exist and since the dimensionality of our integration has changed (see Section 7).

10.3 Distortions due to experimental procedure

It is now clear that M.E. can be any finite-valued function which is defined everywhere. It can be as complicated as we please, it can be discontinuous, and it can be expressed in any variables. Then if M.E. expresses correctly the particle interactions in nature, we can calculate the resulting spectra. However, the observed spectra are not always the same as the spectra occurring in nature because of experimental difficulties and inobservabilities. But these problems can easily be solved in a Monte Carlo calculation simply by applying the same biases and selection criteria to the artificial events as were applied to the real ones. For example it often occurs that there are two or more indistinguishable particles in a reaction, but that they play different rôles in the interaction mechanism. Such a case would be the reaction $\pi^+ p \to \pi^+ p \pi^+ \pi^-$ where one π^+ may form a vector meson with the π^- where the other might form a nucleon isobar with the proton. Experimentally of course there is no way of knowing, in a given case, which is which, and the physicist is obliged to put each event into his histograms twice, once for each π^+. This results in a $\pi\pi$ invariant mass spectrum which shows the $\pi\pi$ interaction highly distorted by the presence of the "wrong" π^+ in half the combinations. On the other hand, if the physicist tries to guess which π^+ interacted more strongly with the π^- by using his knowledge of the mass dependence of the interaction, he will sometimes be wrong and will therefore distort the spectrum in another way. But all hope is not lost, for although he can never obtain the "pure" spectrum that he wants, he can always evaluate the distortion he has caused with his treatment of the data by applying the same distortion to the Monte Carlo events (where he knows which π^+ interacts with the π^- because he wrote the matrix element himself).

11. ERRORS AND VARIANCE-REDUCING

Now that we have a general method of attacking phase-space problems of the greatest complexity, it is well to look at the practical side of the calculation, for we often find functions which, although in principle integrable, have in fact so large a variance that one never converges on a sensible answer. This happens in fact for an unfortunately large class of problems including highly peripheral collisions, narrow resonances, multiple resonances, and some beta decays. Before considering these cases in detail, we obtain an expression for the statistical error on our calculated spectra.

11.1 Error on the integral over a bin

Let w_i be the weight for the i^{th} event if the event falls in the bin under consideration, and zero if the event is outside the bin. Then, from Section 5.1, our estimate for the integral over the bin is

$$W = \frac{1}{N} \sum_{i=1}^{N} w_i \ , \tag{11.1}$$

where we drop the factor corresponding to the bin width since we will take bins of equal width and we are only interested in the relative height in each bin. The estimate for the error on W is

$$\bar{\delta} = \sqrt{\frac{\bar{\sigma}^2}{N}} = \sqrt{\frac{\Sigma_i (w_i - W)^2}{N(N-1)}}$$

$$= \sqrt{\frac{\Sigma_i (w_i)^2 - NW^2}{N(N-1)}}$$

$$= \sqrt{\frac{1}{N(N-1)} \left[\Sigma (w_i)^2 - \frac{1}{N} (\Sigma w_i)^2 \right]} \ . \tag{11.2}$$

But since most of the w_i are zero (they fall in another bin), the second term in square brackets in Eq. (11.2) is much smaller than the first and can be neglected. (It makes the error slightly smaller and represents the constraint due to normalization). If in addition we say that $N^2 >> N$, we have a common factor $1/N$ in both W and $\bar{\delta}$ which is the same for all bins and therefore represents only another common scale factor which does not interest us. Thus we end up with the final contents of one bin as

$$W = \sum_i w_i \pm \sqrt{\sum_i w_i^2} \ . \tag{11.3}$$

This is an important formula and some of its properties should be stressed. First of all we notice that the relative error is minimized when all the weights are equal. (This gives the well-known Poisson error = $\sqrt{N}$.) A related property is that if one weight is much larger than all the rest, the error on W is 100%.

11.2 Weighted versus unweighted events

In order for the Monte Carlo method to work, all weights must be finite. [In some formulations, singular weighting functions may occur. These must be avoided by using the importance sampling technique. See Section 11.3.] There must then exist a maximum weight that an event can take on. If this maximum is known (call it W_m) then unweighted events may be generated from weighted events by the following technique: for each event, choose a random number r between zero and W_m. If the weight of the event is smaller than r, reject the event; otherwise keep the event and give it a weight of one. This means that the probability of keeping an event will be equal to its weight. In this way we produce events

which all have the same weight, which we know minimizes the error for a given sample size, but in doing so we have decreased the sample size. The net effect, of course, is to increase the relative error since it corresponds to going from crude Monte Carlo (Section 5.3) to hit-or-miss Monte Carlo (Section 5.2). However, this technique may be useful on long calculations where the event generator time is short compared with the time to complete the treatment of the event. In such a case one does not want to waste calculating time on an event which will not be worth much in the histograms, so it is better to go back and generate a better one.

This brings us again to the problem of finding the maximum weight. This has been solved in closed form for the T-generators in the case of four, five, and six bodies.

For the M-generator, the weight of an event, from Section 9, is proportional to

$$WT(M_n,m_k) = \prod_{i=2}^{n=1} \left\{ M_i R_2 (M_{i+1};\ M_i, m_{i+1}) \right\} \quad R_2(M_2;\ m_1, m_2) \ . \tag{11.4}$$

The factors within the products may be regrouped as follows:

$$WT = \frac{1}{M_1} \prod_{i=1}^{n-1} \left\{ M_i R_2 (M_{i+1};\ M_i, m_{i+1}) \right\} \tag{11.5}$$

$$= \frac{1}{M_n} \prod_{i=1}^{n-1} \left\{ M_{i+1} R_2 (M_{i+1};\ M_i, m_{i+1}) \right\} . \tag{11.6}$$

For a given reaction with a given centre of mass energy, each of these factors with the product sign takes on its maximum value for

$$\begin{aligned} \text{maximum} \quad M_{i+1} &= M_n - \sum_{j=i+2}^{n} m_j \\ \text{and minimum} \quad M_i &= \sum_{j=1}^{i} m_j \ . \end{aligned} \tag{11.7}$$

This gives therefore a method for calculating the maximum weight. However, it is not very efficient because in reality the maximum weight is smaller than this since we really have the constraint that for each factor in Eq. (11.6) the M_{i+1} must be equal to the M_i of the next factor, which would not be true under conditions (11.7). It is therefore suggested

that for short runs one could use conditions (11.7) to calculate the maximum weight, but for longer runs it is preferable to look at the distribution of weights directly in order to determine the maximum value for a particular configuration.

11.3 Variance-reducing by importance sampling

In Section 5.5 we have seen a rather trivial example of importance sampling. In fact this technique is so important as to merit rather detailed study. We start by defining three classes of functions for which importance sampling is useful (if not necessary):

i) functions which are null over a large region of phase space; this is usually due to geometrical detection efficiency or simply a decision to study only a certain class of events;

ii) functions with sharp peaks (highly peripheral events or narrow resonances, for instance);

iii) functions with singularities; we have seen in Section 7 that there are functions which have finite integrals but are not integrable by Monte Carlo since they have infinite variances.

Each of the above cases is normally handled differently, but there is a general technique which may help to diagnose the sampling problem for all cases. This consists essentially in seeing how the random numbers were used. Since each n-body event requires $3n - 4$ random numbers for its specification, we can make $3n-4$ distributions, or histograms, showing the random numbers that were used. Of course, since each random number is drawn evenly between zero and one, we will find that each distribution is flat. But if we weight the distributions in the same way that the phase space spectra were weighted, we will find which ranges of random numbers give the most important contributions to the phase-space integral. If our formulation of the problem was an efficient one, then all the $3n-4$ distributions will be nearly flat even after weighting. If, however, one of the above three conditions holds, this probably will show up in a very uneven weighted random number distribution. Although this is a very easy method, it is not always efficient because the trouble may be due not to just the behaviour of one variable but to a particular combination of variables which does not show up in our distributions. This latter case is much more serious since it necessitates a complete reformulation of the problem (see Section 11.6).

If the trouble can be seen directly from the weighted random number distributions, it can be solved according to the case as follows:

i) For the function which is zero over a large interval, we simply do not choose random numbers in that interval. This is easy once the interval has been found by the above technique.

ii) For the function which is everywhere non-zero and finite, but is simply unevenly distributed, an integrable function g must be found which has the same general features as the observed distribution f so that $f(r)/g(r) \approx 1$ for all r. Then a transformation is made:

$$f(r)\, dr \rightarrow f(r)\, \frac{dG(r)}{g(r)} \quad ,$$

where $G(r) = \int_0^r g(r')\,dr'$. For example, we could have approximated the circle of Section 5.3 by one loop of a cosine curve

$$\sqrt{1 - x^2} \approx \cos\left(\frac{\pi x}{2}\right) \quad \text{for } 0 < x < 1 ,$$

thereby transforming our integral:

$$\int_0^1 \sqrt{1 - x^2}\,dx = \int_0^{2/\pi} \sqrt{1 - x^2}\;\frac{d(\frac{2}{\pi}\sin\frac{\pi x}{2})}{\cos\frac{\pi x}{2}} .$$

The procedure is then the following. We choose an r between 0 and $2/\pi$. We then calculate

$$x = \frac{2}{\pi}\arcsin\frac{\pi r}{2} .$$

For each x that we find in this way, we take the function value

$$\frac{\sqrt{1 - x^2}}{\cos(\pi x/2)}$$

which, as we have seen, has a small variance. The error on the integral is thus reduced.

iii) For functions which actually become infinite the technique is the same as the preceding one, except that now we must be sure that our approximate function removes the singularity, which usually means that we must know something about the behaviour of the functions.

11.4 Importance sampling for resonances

One common problem which can be solved very efficiently is that of the production of a narrow resonance. As we have seen in Section 7, this would give rise to a large error if we did not apply a variance-reducing technique.

By a resonance I mean any Breit-Wigner function of an invariant mass. Since the M-generator chooses invariant masses directly, this technique is a natural one for this generator. However, not <u>all</u> invariant masses are specified explicitly as basic variables, so care must be taken in the ordering of particles. In particular, random number one corresponds to M_{12}, two corresponds to M_{123}, etc. The nice feature of this method is that the simplest Breit-Wigner function is exactly integrable:

$$\int \frac{\Gamma/2}{(M - M_0)^2 + (\Gamma/2)^2}\,dM = \arctan\left(\frac{M - M_0}{\Gamma}\right) .$$

This method will then be 100% efficient if a simple Breit-Wigner is used, and only slightly less efficient if it is desired to reweight the event by a more complicated form, the random

number selection still being done by use of the simple integrable Breit-Wigner. One word of caution: if importance sampling is used on the random numbers corresponding to invariant masses, we can no longer use the trick of ordering the numbers in order to fall always in the physical region (see the end of Section 9) since the different numbers are not all drawn from the same distribution. In general, however, not much time is wasted in throwing away the combinations which are not in order.

11.5 Importance sampling for peripheral collisions

For the M-generator for n bodies, the first n-2 random numbers correspond to effective masses. The next two numbers correspond to the cos ϑ and φ of particle 1 in the rest system of particles 1 and 2. The next two random numbers correspond to the cos ϑ and the φ of particles 1 and 2 in the rest system of particles 1, 2 and 3. And so forth. Therefore, for example, if one is going to select on momentum transfer to a given particle, that particle should be arranged last since there will then be two random numbers which correspond to the c.m.s cos ϑ and φ of all the other particles with respect to that one.

11.6 More difficult cases requiring reformulation

As we have seen, the non-uniformity of our weighting function may not show up in the variables we have used to express the phase-space problem. If this function has a very large variance, it will necessitate the use of a different generator in order to apply variance-reducing techniques. This in fact happens for some weak decays (for example K_{e3} decay) where the matrix element or form factor, which is well know, is very non-uniform and is usually expressed in variables different from the ones we have been using in this course. However, since K_{e3} decay is only a three-body process, it can be reformulated in many ways with very little effort.

12. THE ABSOLUTE VALUE OF R_n

In Sections 9, 10, and 11 we have been concerned principally with getting phase space spectra without worrying about the normalization. An exception was Section 10.1 where we saw that R_n could be used to predict branching ratios or relative decay rates or cross-sections. It is understood that the relevant matrix elements must be included in the evaluation of the integrals, as well as any Clebsch-Gordan coefficients from SU_3, isotopic spin conservation, etc. It is often true, however, that the dominant factor in such determinations is that coming from pure phase-space considerations. Three different cases may be considered:

i) The ratio of the total rates for two interactions involving the same number of particles. In this case the formulation of Section 9 is sufficient since all constants which have been neglected will be anyway the same for both rates. There is, however, an extra factor to be calculated, corresponding to the interval size in M-space.

ii) The ratio of the total rates for two interactions involving different numbers of particles. This case is more delicate since it depends on a choice for the interaction volume which is essentially a free parameter in the theory but which should be of the order of L^n, where L is a pion Compton wavelength. This constant can in principle be determined experimentally for one ratio and then be used to predict others.

iii) The calculation of absolute cross-sections. This case requires the determination of still another constant parameter of the theory, in addition to the one mentioned above.

The treatment of the various parameters of the Fermi theory as absolute constants independent of the reactions is not considered entirely reliable, which makes the determination of absolute numbers a rather risky affair. The situation is discussed at length in the literature (see for example Kretschmar's review[3]).

12.1 Cross-sections and energy variation

If we consider the phase space available for any given reaction as a function of the total initial energy, it is clear that the number of possible states will increase rapidly with increasing energy, especially for many-particle final states. The pure phase-space prediction for the behaviour of cross-sections with energy is therefore a monotonic increasing function, with rapidly increasing slope for many-particle states since there are many ways for the extra energy to be distributed among the outgoing particles. In fact the observed behaviour of cross-sections is not very different from this prediction if the energy is near the reaction threshold. But far from threshold, experiments show that the available states are not evenly filled and a better approximation is to consider the cross-section to be constant with energy, with the largest variations coming from resonance production and peripheral or Regge-pole dominance.

The formulation of Section 9 can be used directly in the case where the initial energy M_n is variable (for example, the beam momentum may not be well determined), but this will give generally larger weights to events with higher beam momentum in accordance with the Fermi theory. If it is desired that this energy-dependence be suppressed in order to obtain constant cross-sections, each weight should be divided by $W_m(M_n)$, which is the maximum weight possible for the total energy of the event (see Section 11.2). This is in fact the standard procedure followed in the CERN programs OWL and FOWL, with the approximation of Section 11.2 being used for the maximum weight.

12.2 The deuterium problem I

It often occurs that the target particle in an experiment is not a simple hydrogen nucleus (proton) but is a more complicated nucleus, of which we take deuterium as the simplest example. In this case, when the beam energy is large compared with the binding energy of the nucleus (a hundred MeV is sufficient) the interaction can be considered to take place on one of the nucleons alone, leaving the rest of the nucleus (called the spectator) untouched. However, the situation is complicated by the fact that the nucleons are not at rest inside the nucleus, but have a certain momentum distribution, known as the Fermi momentum, which for the case of deuterium is given by the Hulthén wave function. The problem of finding the resulting phase-space distributions can be solved in two different ways:

i) The target can be considered as a single nucleon moving with Fermi momentum distribution in random directions with respect to the beam.

ii) The target can be considered as the entire nucleus, at rest in the laboratory, with the spectator recoil being given by the Fermi momentum distribution.

The second method will be discussed in Section 13.1. The first method is generally more efficient and can be applied in a straightforward manner without special variance-reducing techniques. For each event a target nucleon momentum is chosen according to the known momentum distribution at random angles with respect to the beam direction and the new centre-of-mass total energy is calculated. (The full relativistic treatment must be used since a small Fermi momentum can give rise to a large difference in total energy.) The resultant momentum vector of the centre of mass with respect to the laboratory should be stored so that the event, after being generated in the centre of mass, may be transformed properly into the laboratory system if desired. If the spectator recoil momentum is required, it is simply equal and opposite to the target nucleon momentum in the laboratory, and is independent of all other features of the event.

13. IMPOSING KNOWN SPECTRA

There sometimes arise problems which cannot be expressed in the general framework of Eq. (9.2). This usually happens because the knowledge we have about the reaction is not the knowledge of a matrix element or particle interaction, but is rather ad hoc experimental knowledge of certain features of the reaction which already include the phase-space factor in an unknown manner. In these cases, the formalism of Eq. (9.2) can be used only if the separation of our knowledge can be made into "phase-space part" and "interaction part", because if we simply put in the experimentally found spectrum as a matrix element, we will get out this spectrum altered by the phase space density. So in general the phase-space density must be determined beforehand in a calculation without matrix element, then a matrix element must be constructed which is the desired distribution divided by the phase-space distribution, in order that the final result after weighting by phase space will be the desired distribution. This then assures us of having a matrix element which reproduces at least one spectrum correctly and we can investigate the behaviour of other spectra under the assumption that this matrix element is the only important one. The theoretical justification for such a procedure will depend on the particular case and will in general be difficult to establish, so that one must be careful about the conclusions which are to be drawn from such an analysis. But considering the current state of understanding of strong interactions, such a procedure is probably as valid as any other.

13.1 The deuterium problem II

One case in which the above method can be well justified is in imposing on deuterium interactions the known spectator momentum distribution as described above in Section 12.2 (ii). The essential problem here is a practical one, namely that this method produces spectator momenta according to phase space, then reweights them in order to impose the known spectator distribution, which is a very inefficient process because of the small amount of phase space occupied by the desired spectator distribution. In order to render this method efficient, careful use must be made of variance-reducing techniques (Section 11.3).

REFERENCES FOR PART II

1) R. Hagedorn, Relativistic kinematics (Benjamin, N.Y., 1964).

2) C. Zemach, Phys.Rev. 133, B1201 (1964).

3) M. Kretschmar, Ann.Rev.Nucl.Sci. 11, 1 (1961).

4) E. Fermi, Progr.Theoret.Phys. 5, 570 (1950).

5) M. Block, Phys.Rev. 101, 796 (1956).

6) G. Kopylov, Zh.Eksperim. i Teor.Fiz. 35, 1426 (1958), translation Soviet Physics JETP 8, 996 (1959).

7) P.P. Srivastava and G. Sudarshan, Phys.Rev. 110, 765 (1958).

8) F. Cerulus and R. Hagedorn, Nuovo Cimento 9, 646 (1958).

9) G. Lynch, CERN TC Program Library, Section FAKE, B001, or the original Berkeley report UCRL 10335.

10) G. Goldhaber, S. Goldhaber, W. Lee and A. Pais, Phys.Rev. 120, 300 (1960).

11) G. Kopylov, Zh.Eksperim. i Teor.Fiz. 39, 1091 (1960), translation Soviet Physics JETP 12, 761 (1961).