1 Definition of subgroup parameters

Let's define the averaging operator for some function $f(E)$ over some energy interval ΔE as follows:

$$\langle f(E)\rangle_{\Delta E} = \frac{\int\limits_{\Delta E} f(E)\phi(E)\,\mathrm{d}E}{\int\limits_{\Delta E} \phi(E)\,\mathrm{d}E}, \tag{1}$$

where $\phi(E)$ is neutron spectrum. Thus $\langle f(E)\rangle_{\Delta E}$ is the average value of function $f(E)$ in ΔE from the transport equation point of view. In the following the subscript ΔE will be omitted if the energy interval is clear out of the context.

The theory of subgroup parameters is described in [1], and somewhat more detailed in [2]. The definition of subgroup parameters begins with the splitting of the energy interval ΔE called *group* into intervals of monotonicity of cross section of interaction between neutron and nuclei $\sigma(E)$ with respect to neutron energy E. After the splitting integrals in equation (1) for some integrable function $f(\sigma(E))$ are replaced with an integral over the cross section:

$$\langle f(\sigma(E))\rangle = \int\limits_0^\infty f(\sigma)p(\sigma)\,\mathrm{d}\sigma, \tag{2}$$

where

$$p(\sigma) = \frac{\sum\limits_i \phi(E_i(\sigma))\left|\frac{\mathrm{d}E}{\mathrm{d}\sigma}\right|}{\int\limits_{\Delta E} \phi(E)\,\mathrm{d}E}, \tag{3}$$

$E_i(\sigma)$ is the inverse of $\sigma(E)$ in cross section monotonicity interval i (see fig.1). The sum is calculated over all intervals.

$p(\sigma)$ is the probability for neutron to be of such energy that its cross section is equal to σ under assumption that its energy is inside ΔE. The subgroup approximation is based on the expansion of $p(\sigma)$ into finite series of δ-functions:

$$p(\sigma) = \sum_{k=1}^{K} a_k \delta(\sigma - \sigma_k). \tag{4}$$

It is equalent to approximation of the correspondent probability function with histogram made of steps of a_k height and of σ_k width. a_k and σ_k are referred to as *subgroup parameters*. The former are referred to as *subgroup probabilities*[1], the latter—as *average subgroup cross sections*. a_k is considered to be equal to the probability for neutron to be in subgroup k under assumption that it is in the correspondent group. However, it is hard to define the term "subgroup". It is tempting to define subgroup as a set of energy intervals ΔE_k complying with the following

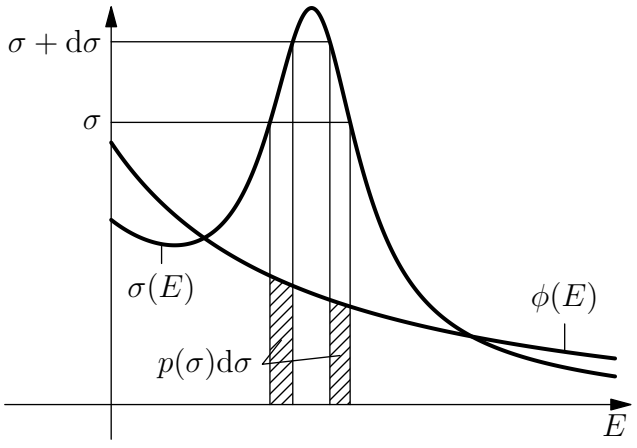

Figure 1: $p(\sigma)$ is the probability for neutron scattered into the current group having such energy E that $\sigma(E) \in [\sigma; \sigma + \mathrm{d}\sigma]$

[1] Also *subgroup shares* in Russian literature

conditions:

$$a_k = \frac{\int\limits_{\Delta E_k} \phi(E)\,\mathrm{d}E}{\int\limits_{\Delta E} \phi(E)\,\mathrm{d}E} \tag{5}$$

$$\sigma_k = \frac{\int\limits_{\Delta E_k} \sigma(E)\phi(E)\,\mathrm{d}E}{\int\limits_{\Delta E_k} \phi(E)\,\mathrm{d}E}, \tag{6}$$

$$\bigcup_{k=1}^{K} \Delta E_k = \Delta E, \tag{7}$$

$$\Delta E_k \cap \Delta E_n = \varnothing,\ k \neq n. \tag{8}$$

These equations define *arranged subgroup parameters*, that is, subgroup parameters bijectively mapped to a set of non-intersecting energy intervals completely covering the group. Unfortunately, the matter of consistency of equations (5)–(8) as well as the choice of the way to arrange subgroups remains open. In the following the work with subgroup parameters will be based on equation (4), so there is no need in the definition of subgroup.

The substitution of (4) into (2) provides the following simple expression to calculate function average through subgroup parameters:

$$\langle f(\sigma) \rangle = \sum_{k=1}^{K} a_k f(\sigma_k). \tag{9}$$

The function $p(\sigma)$ may be expanded into series of δ functions with different coefficients. The choice of the set of coefficients influences on the accuracy of equation (9). Therefore in order to completely define subgroup parameters, additional criteria have to be provided, e.g. criteria maximizing the accuracy of some values computed through subgroup parameters. In reactor physics the natural choice of these values are *self-shielded cross sections*:

$$f_{t,0}(z) = \frac{\left\langle \frac{\sigma}{\sigma+z} \right\rangle}{\left\langle \frac{1}{\sigma+z} \right\rangle}, \tag{10}$$

$$f_{t,1}(z) = \frac{\left\langle \frac{1}{\sigma+z} \right\rangle}{\left\langle \frac{1}{(\sigma+z)^2} \right\rangle} - z, \tag{11}$$

where z is the dilution cross section. Self-shielded cross sections may be expressed through subgroup parameters by means of equation (9):

$$\tilde{f}_{t,0}(z) = \frac{\sum\limits_{k=1}^{K} \frac{a_k \sigma_k}{\sigma_k+z}}{\sum\limits_{k=1}^{K} \frac{a_k}{\sigma_k+z}} \tag{12}$$

$$\tilde{f}_{t,1}(z) = \frac{\sum\limits_{k=1}^{K} \frac{a_k}{\sigma_k+z}}{\sum\limits_{k=1}^{K} \frac{a_k}{(\sigma_k+z)^2}} - z. \tag{13}$$

Let's define *error in y with respect to x* as follows:

$$\delta(x, y) = \left| \frac{y}{x} - 1 \right|. \tag{14}$$

In other words, if x is some value and y is its approximation computed through subgroup parameters, $\delta(x, y)$ provides a measure for an error in y. The *error in function $g(z)$ with respect to function $f(z)$* (*residual*) is defined as follows:

$$\delta[f, g] = \sup_{z \in \mathcal{Z}} \delta(f(z), g(z)). \tag{15}$$

In the following a finite set of points will be used as $\mathcal{Z}$:

$$\mathcal{Z} = \mathcal{Z}_N = \{z_i,\ i = 1, \ldots, N\}. \tag{16}$$

With the help of residual the additional criteria on subgroup parameters is defined as the problem of minimization of functions $\delta[f_{t,0}, \tilde{f}_{t,0}]$ and $\delta[f_{t,1}, \delta f_{t,1}]$. In practice, it is more convinient to set the required accuracy ε and search for subgroup parameters complying with the following inequalities:

$$\begin{cases} \delta[f_{t,0}, \tilde{f}_{t,0}] < \varepsilon, \\ \delta[f_{t,1}, \tilde{f}_{t,1}] < \varepsilon. \end{cases} \tag{17}$$

In addition to a_k and σ_k, *partial subgroup parameters* $\sigma_{x,k}$ are required to compute partial self-shielded cross sections of reaction x. If the term "subgroup" is defined via (5)–(8), there parameters can be defined in the same way as σ_k:

$$\sigma_{x,k} = \frac{\int\limits_{\Delta E_k} \sigma_x(E)\phi(E)\,\mathrm{d}E}{\int\limits_{\Delta E_k} \phi(E)\,\mathrm{d}E}. \tag{18}$$

However in this work partial subgroup parameters are defined as such $\sigma_{x,k}$ that

$$\delta[f_{x,0}, \tilde{f}_{x,0}] < \varepsilon, \tag{19}$$

where

$$f_{x,0}(z) = \frac{\left\langle \frac{\sigma_x}{\sigma+z} \right\rangle}{\left\langle \frac{1}{\sigma+z} \right\rangle}, \tag{20}$$

$$\tilde{f}_{x,0}(z) = \frac{\sum\limits_{k=1}^{K} \frac{a_k \sigma_{x,k}}{\sigma_k+z}}{\sum\limits_{k=1}^{K} \frac{a_k}{\sigma_k+z}}. \tag{21}$$

Despite that the term subgroup was not strictly defined, it still has to be considered. Under assumption that there exists some partitioning of group energy interval into subgroups, subgroup parameters can be assigned the physical meaning of the probability for neutron to scatter into the subgroup and average (total and partial) subgroup cross sections. Under some approximations these subgroup parameters can be used in existing group programs given that subgroup parameters are put in order and used in place of group average cross sections. In these cases the following properties of subgroup parameters become important. They would have these properties if they were defined via the term subgroup.

- Positiveness: $a_k \geq 0$, $\sigma_k \geq 0$, $\sigma_{x,k} \geq 0$.
- Probability norm compliance: $\sum\limits_{k=1}^{k} a_k = 1$.
- Total cross section norm compliance: $\sum\limits_{k=1}^{K} a_k \sigma_k = \langle \sigma \rangle$.

- Partial cross sections norm compliance: $\sum\limits_{k=1}^{K} a_k \sigma_{x,k} = \langle \sigma_x \rangle$.
- Partial cross sections norm compliance: $\sum\limits_{x} \sigma_{x,k} = \sigma_k$.

To conclude, here is the final definition of subgroup parameters:

Definition 1. Subgroup parameters is the set of numbers a_k, σ_k and $\sigma_{x,k}$, $k = 1, \ldots, K$, $x \in \mathcal{X}$, satisfying the following requirements:

$$a_k \geq 0,\ \sigma_k \geq 0,\ \sigma_{x,k} \geq 0,\ \forall k,\ x, \tag{22}$$

$$\sum_{k=1}^{K} a_k = 1, \tag{23}$$

$$\sum_{k=1}^{K} a_k \sigma_k = \langle \sigma \rangle\,, \tag{24}$$

$$\sum_{k=1}^{K} a_k \sigma_{x,k} = \langle \sigma_x \rangle\,, \tag{25}$$

$$\sum_{x \in \mathcal{X}} \sigma_{x,k} = \sigma_k, \tag{26}$$

$$\delta[f_{t,0}, \tilde{f}_{t,0}] < \varepsilon, \tag{27}$$

$$\delta[f_{t,1}, \tilde{f}_{t,1}] < \varepsilon, \tag{28}$$

$$\delta[f_{x,0}, \tilde{f}_{x,0}] < \varepsilon\ \forall x \in \mathcal{X}, \tag{29}$$

where $f_{t,0}$, $f_{t,1}$, $f_{x,0}$ are self-shielded cross sections (10), (11) and (20), $\tilde{f}_{t,0}$, $\tilde{f}_{t,1}$ and $\tilde{f}_{x,0}$ are self-shielded cross sections expressed through subgroup parameters (12), (13) and (21), $\delta[f, g]$ is the residual (15), ε is the accuracy required.

2 Computation of subgroup parameters

2.1 The method of Padé approximants

2.1.1 Multipoint Padé approximation

Rational function is the ratio of two polynomes. If numerator polynome is of power M and denominator polynome is of power N, the rational function is usually denoted with symbol $[M/N]$ [3]. Padé approximation is the approximation of some function with rational function. The way to construct Padé approximation through first few coefficients of Taylor series of the function being approximated computed in several points is discussed in this section.

Consider inequality (27). Under definitions given in section 1, it is equivalent to the following approximation:

$$\frac{\left\langle \frac{\sigma}{\sigma+z} \right\rangle}{\left\langle \frac{1}{\sigma+z} \right\rangle} \approx \frac{\sum\limits_{k=1}^{K} \frac{a_k \sigma_k}{\sigma_k + z}}{\sum\limits_{k=1}^{K} \frac{a_k}{\sigma_k + z}}, \quad z \in \mathcal{Z}. \tag{30}$$

Obviously, these two ratios are equal when their numerators and denominators are equal (although it's better to control the accuracy via (30)). Together with inequalities (28) and (29), it produces the following

system of equations:

$$\left\langle \frac{1}{\sigma + z} \right\rangle \approx \sum_{k=1}^{K} \frac{a_k}{\sigma_k + z}, \tag{31}$$

$$\left\langle \frac{1}{(\sigma + z)^2} \right\rangle \approx \sum_{k=1}^{K} \frac{a_k}{(\sigma_k + z)^2}, \tag{32}$$

$$\left\langle \frac{\sigma_x}{\sigma + z} \right\rangle \approx \sum_{k=1}^{K} \frac{a_k \sigma_{x,k}}{\sigma_k + z}, \tag{33}$$

$$z \in \mathcal{Z}.$$

(the equation for numerators in (30) can be reduced to the first of these equations by adding and subtracting $\left\langle \frac{z}{\sigma+z} \right\rangle$ in the left side and $\sum\limits_{k=1}^{K} \frac{a_k z}{\sigma_k + z}$ in the right side).

Note that in the second equation both sides are derivatives of the correspondent sides of the first equation. Let

$$f(z) = \left\langle \frac{1}{\sigma + z} \right\rangle. \tag{34}$$

Then the first two approximal equations written above define Padé approximation $[(K-1)/K]$ for $f(z)$ by its first two coefficients of Taylor series in points $z \in \mathcal{Z}$:

$$f(z) = \left(\left\langle \frac{1}{\sigma + z} \right\rangle + \left\langle \frac{1}{(\sigma + z)^2} \right\rangle (z - z_i) + \mathcal{O}(z - z_i) \right) \Bigg|_{z_i \in \mathcal{Z}_{K-1}} = \sum_{k=1}^{K} \frac{a_k}{\sigma + z}. \tag{35}$$

Equations (23) and (24) can be expressed through coeffiecients of $f(z)$ decomposition into series with negative powers in the neighborhood of zero:

$$\sum_{k=1}^{K} a_k = 1 = \frac{\mathrm{d}}{\mathrm{d}t} f\left(\frac{1}{t}\right) \Bigg|_{t=0}, \tag{36}$$

$$\sum_{k=1}^{K} a_k \sigma_k = \langle \sigma \rangle = -\frac{1}{2} \frac{\mathrm{d}^2}{\mathrm{d}t^2} f\left(\frac{1}{t}\right) \Bigg|_{t=0}. \tag{37}$$

There exist methods to obtain Padé approximation that shares with the function being approximated coefficients of its decomposition into series with both positive and negative powers. These methods are considered in detail in [4]. They are based on recursive formulas to add a new point or increase the number of fixed coefficients of decomposition into series with either positive or negative powers in the existing point. Considered in this work are Padé approximations sharing with the function being approximated the first two coefficients of its decomposition into series with negative powers in the neighborhood of zero and the first two coefficients of its decomposition into series with positive powers in several points. Hence, only formulas required to obtain such a function will be derived here, in particular formula to add a new point that takes into account values of both the function and its first derivative, and the formula to compute the starting Padé approximation that takes into account the first two coefficients in the decomposition into series with negative powers.

Let

$$f_n(z) \equiv \frac{P_n(z)}{Q_n(z)} = c_{k,0} + c_{k,1}(z - z_k) + \mathcal{O}(z - z_k), \quad k = 1, \ldots n \tag{38}$$

—Padé approximation $[n/(n+1)]$ with first two coefficients of its decomposition into Taylor series in the neighborhood of n points $z_k \in \mathcal{Z}_n$ being equal to $c_{k,0}$ and $c_{k,1}$. Let c_{-1} and c_{-2} be the first two coefficients of its decomposition into series with negative powers in the neighborhood of zero:

$$f_n(z) = c_{-1}z^{-1} + c_{-2}z^{-2} + \mathcal{O}(z^{-2}). \tag{39}$$

Let $f_{n-1}(z)$ and $f_n(z)$ be two Padé approximations such that $\mathcal{Z}_{n-1} \subset \mathcal{Z}_n$. Then for any α and β

$$f(z) \equiv \frac{(z+\alpha)P_n(z) + \beta(z-z_n)^2 P_{n-1}(z)}{(z+\alpha)Q_n(z) + \beta(z-z_n)^2 Q_{n-1}(z)} = c_{k,0} + c_{k,1}(z-z_k) + \mathcal{O}(z-z_k) \quad \forall z_k \in \mathcal{Z}_n \tag{40}$$

and

$$f(z) = c_{-1}z^{-1} + c_{-2}z^{-2} + \mathcal{O}(z^{-2}), \tag{41}$$

that is, $f(z)$ is the Padé approximation $[(n+1)/n]$ that describes the function being aproximated as well as $f_n(z)$ do. To proof that one can multiply both sides of these equations with the denominator of $f(z)$ and subtruct them. For equation (40) this results in

$$(z+\alpha)(P_n(z)-(c_{k,0}+c_{k,1}(z-z_k))Q_n(z))+\beta(z-z_n)^2(P_{n-1}(z)-(c_{k,0}+c_{k,1}(z-z_k))Q_{n-1}(z)) = \mathcal{O}(z-z_k). \tag{42}$$

It is true because by the definition of the Padé approximation (38) $P_n(z) - (c_{k,0} + c_{k,1}(z-z_k))Q_n(z) = \mathcal{O}(z-z_k)$ for all k and $P_{n-1}(z)-(c_{k,0}+c_{k,1}(z-z_k))Q_{n-1}(z) = \mathcal{O}(z-z_k)$ for $k < n$ and $(z-z_n)^2 = \mathcal{O}(z-z_k)$ when $k = n$. Same is for the second equation.

Arbitrary parameters α and β can be picked so that function $f(z)$ and its first derivative in the neighborhood of z_n have the same values as those of function being approximated, that is, $f(z) \equiv f_{n+1}(z)$. The correspondent equation is:

$$\frac{(z+\alpha)P_n(z) + \beta(z-z_n)^2 P_{n-1}(z)}{(z+\alpha)Q_n(z) + \beta(z-z_n)^2 Q_{n-1}(z)} = c_{n+1,0} + c_{n+1,1}(z-z_{n+1}) + \mathcal{O}(z-z_{n+1}). \tag{43}$$

Let $\tilde{z} = z - z_{n+1}$ and let

$$P_n(z) = \sum_{k=0}^{n} \tilde{p}_{n,k}\tilde{z}^k, \tag{44}$$

$$Q_n(z) = \sum_{k=0}^{n+1} \tilde{q}_{n,k}\tilde{z}^k. \tag{45}$$

Multiplication of both sides of equation (43) with the denominator and subtraction of one from the other produce the following equations defining the coefficients of the polynome acquired:

$$\begin{cases} \alpha d_{00} + \beta d_{01} = 0, \\ \alpha d_{10} + \beta d_{11} = -d_{00}, \end{cases} \tag{46}$$

where

$$\begin{aligned}
d_{00} &= \tilde{p}_{n,0} - c_{n+1,0}\tilde{q}_{n,0}, \\
d_{01} &= (z_{n+1}-z_n)^2(\tilde{p}_{n-1,0} - c_{n+1,0}\tilde{q}_{n-1,0}), \\
d_{10} &= \tilde{p}_{n,1} - c_{n+1,0}\tilde{q}_{n,1} - c_{n+1,1}\tilde{q}_{n,0}, \\
d_{11} &= (z_{n+1}-z_n)(2(\tilde{p}_{n-1,0} - c_{n+1,0}\tilde{q}_{n-1,0}) + (z_{n+1}-z_n)(\tilde{p}_{n-1,1} - c_{n+1,0}\tilde{q}_{n-1,1} - c_{n+1,1}\tilde{q}_{n-1,0})).
\end{aligned}$$

The solution of this system of equations is

$$\alpha = -\frac{d_{00}d_{01}}{\begin{vmatrix} d_{00} & d_{01} \\ d_{10} & d_{11} \end{vmatrix}}, \tag{47}$$

$$\beta = -\frac{d_{00}^2}{\begin{vmatrix} d_{00} & d_{01} \\ d_{10} & d_{11} \end{vmatrix}}. \tag{48}$$

Let's find the starting functions $f_0(z)$ and $f_1(z)$. The zeroeth function has to be of the form

$$f_0(z) = \frac{p_{0,0}}{q_{0,0} + z} \tag{49}$$

and comply to equations

$$\frac{\mathrm{d}}{\mathrm{d}t} f_0\left(\frac{1}{t}\right)\bigg|_{t=0} = c_{-1}, \tag{50}$$

$$\frac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}t^2} f_0\left(\frac{1}{t}\right)\bigg|_{t=0} = c_{-2}. \tag{51}$$

Hence

$$f_0(z) = \frac{c_{-1}}{z - \frac{c_{-2}}{c_{-1}}}. \tag{52}$$

The first function has to be of the form

$$f_1(z) = \frac{p_{1,0} + p_{1,1}z}{q_{1,0} + q_{1,1}z + z^2} \tag{53}$$

and comply to two additional equations:

$$f_1(z_0) = c_{0,0} \tag{54}$$

$$\frac{\mathrm{d}}{\mathrm{d}z} f_1(z)\bigg|_{z=z_0} = c_{0,1}. \tag{55}$$

The solution of the corresponding system of equations with respect to Padé approximation coefficients produces the following result:

$$f_1(z) = \frac{c_{0,0}(c_{-1}^2 + c_{0,0}c_{-2}) + c_{-1}(c_{0,0}^2 + c_{-1}c_{0,1})(z - z_0)}{c_{-1}^2 + c_{0,0}c_{-2} + (c_{0,0}c_{-1} - c_{0,1}c_{-2})(z - z_0) + (c_{0,0}^2 + c_{-1}c_{0,1})(z - z_0)^2}. \tag{56}$$

2.1.2 Computation of subgroup parameters

Given Padé aproximation $f(z)$ such that (35)–(37) hold, subgroup parameters a_k and σ_k can be computed from its polar form:

$$f(z) \equiv \frac{\sum\limits_{k=0}^{K-1} p_k z^k}{\sum\limits_{k=0}^{K} q_k z^k} \equiv \sum_{k=1}^{K} \frac{a_k}{\sigma_k + z}. \tag{57}$$

From this equivalence it follows that $-\sigma_k$ are roots of polynome $\sum\limits_{k=0}^{K} q_k z^k$, and a_k are the solution of the system of linear equations

$$\sum_{n=1}^{K} r_{n,k} a_k = p_k, \quad k = 1, \ldots, K, , \tag{58}$$

where

$$r_{n,k-1} = q_{k-1} + z_n q_k,$$
$$r_{n,K} = q_K.$$

Given subgroup parameters a_k and σ_k, partial subgroup parameters $\sigma_{x,k}$ can be obtained from the system of linear equations (33) with z running across the set of dilution cross sections used during computation of a_k and σ_k, i.e. $z \in \mathcal{Z}_{K-1}$. To make the system complete, partial cross section norm (25) also has to be considered. However, one can use the linearity of equation (33) and derive relatively simple equations for $\sigma_{x,k}$ by means of least-squares method. This case is also more convenient to take equations (26) into account.

Here are the equations referred to in the previous paragraph:

$$\sum_{k=1}^{K} \frac{a_k \sigma_{x,k}}{\sigma_k + z} = \left\langle \frac{\sigma_x}{\sigma + z} \right\rangle, \quad z \in \mathcal{Z}, \tag{33}$$

$$\sum_{k=1}^{K} a_k \sigma_{x,k} = \langle \sigma_x \rangle, x \in \mathcal{X}, , \tag{25}$$

$$\sum_{x \in \mathcal{X}} \sigma_{x,k} = \sigma_k, k = 1, \ldots, K. \tag{26}$$

Let m be some whole number between 1 and K (inclusive). From (25), $\sigma_{x,m}$ can be expressed as follows:

$$\sigma_{x,m} = \frac{1}{a_m} \left(\langle \sigma_x \rangle - \sum_{k \neq m} a_k \sigma_{x,k} \right). \tag{59}$$

Let u be some reaction in $\mathcal{X}$. Then

$$\sigma_{u,k} = \sigma_k - \sum_{x \neq u} \sigma_{x,k} \tag{60}$$

The system of equations {(25), (26)} is linearly dependent under assumption that $\sum\limits_{k=1}^{K} a_k \sigma_k = \langle \sigma \rangle = \sum\limits_{x \in \mathcal{X}} \langle \sigma_x \rangle$. Hence both expressions for $\sigma_{u,m}$ are equivalent.

As it is usual for the least squares method, let residual be

$$\delta^2 = \sum_{z \in \mathcal{Z}} \sum_{x \in \mathcal{X}} \left(\sum_{k=1}^{K} \frac{a_k \sigma_{x,k}}{\sigma_k + z} - \left\langle \frac{\sigma_x}{\sigma + z} \right\rangle \right)^2. \tag{61}$$

The exlusion of $\sigma_{x,m}$ and $\sigma_{u,k}$ leads to

$$\delta^2 = \sum_z \left(\sum_{x \neq u} \left(\sum_{k \neq m} a_k \sigma_{x,k} \left(\frac{1}{\sigma_k + z} - \frac{1}{\sigma_m + z} \right) + \frac{\langle \sigma_x \rangle}{\sigma_m + z} - \left\langle \frac{\sigma_x}{\sigma + z} \right\rangle \right)^2 + \right.$$
$$\left. + \left(\sum_{x \neq u} \sum_{k \neq m} a_k \sigma_{x,k} \left(\frac{1}{\sigma_k + z} - \frac{1}{\sigma_m + z} \right) + \left\langle \frac{\sigma}{\sigma + z} \right\rangle - \left\langle \frac{\sigma_x}{\sigma + z} \right\rangle + \frac{\langle \sigma_u \rangle - \langle \sigma \rangle}{\sigma_m + z} \right)^2 \right). \tag{62}$$

Having set partial derivatives of the residual with respect to $\sigma_{v,n}$ equal to zero and having made simple transformation, the following system of equations is obtained:

$$\sum_{x\neq u}(1+\delta_{x,v})\sum_{k\neq m}a_k\sigma_{x,k}\sum_z\left(\frac{1}{\sigma_k+z}-\frac{1}{\sigma_m+z}\right)\left(\frac{1}{\sigma_n+z}-\frac{1}{\sigma_m+z}\right)=$$
$$=\sum_z\left(\left\langle\frac{\sigma}{\sigma+z}\right\rangle-\left\langle\frac{\sigma_u}{\sigma+z}\right\rangle+\left\langle\frac{\sigma_v}{\sigma+z}\right\rangle-\frac{\langle\sigma\rangle-\langle\sigma_u\rangle+\langle\sigma_v\rangle}{\sigma_m+z}\right)\left(\frac{1}{\sigma_n+z}-\frac{1}{\sigma_m+z}\right),$$
$$n=1,\ldots,m-1,m+1,\ldots,K,\quad v\in\mathcal{X}-\{u\},\quad(63)$$

where $\delta_{x,v}$ is the Kronecker symbol, it is equal to one if $x=v$, otherwise it is equal to zero. $\sigma_{x,k}$ are obtained by solving that system.

Obviously, subgroup parameters computed by means of the method described depend on the choice of the set of dilution cross sections $\mathcal{Z}_{K-1}$ used in (35). Let's call these cross sections *fitting dilution cross sections*. It could be undesirable to use as fitting dilution cross sections all points the accuracy of subgroup approximation is checked in. Indeed, the number of subgroups obtained via this method is one more than the number of fitting dilution cross sections, while it is possible to use only one subgroup to describe non-resonant group with almost perfect accuracy. Hence the matter of choice of fitting dilution cross sections. In order to solve it let's consider some properties of Padé approximants.

Padé approximation of continuous function constructed with respect to several points, perhaps with regard to function derivatives, possess interpolation properties. That is, it allows one to compute approximate value of the function being approximated in between fitting points as long as it does not have a pole between these points. For dilution cross sections, the interval $[0;\infty)$ is of interest; there won't be any poles in it if all roots of the polynome in the denominator of Padé approximation (57) are negative, i.e. if $\sigma_k>0$.

As well as any other smooth interpolating function, Padé approximation may dramatically differ from the function being approximated if the latter has some properties badly described by rational approximation and/or the fitting points were chosed poorly. In the case of Padé approximation it results in the appearance of poles between fitting points. The pole may also appear in the case when the function, conversely, well agrees with the rational approximation. It happens when the addition of the next fitting point results in the construction of approximation in the form of $\frac{(Az+B)P_N(z)}{(A'z+B')Q_M(z)}$, where A and A' and B and B' marginally differ from each other. Due to this small difference binomials do not reduce, and the approximation get a so-called "noise doublet", a pole and a zero close to each other: the approximation equals to zero when $z=-B/A$ and to infinity when $z=-B'/A'$. The appearance of the noise doublet is characterized by small values of a_k in subgroup parameters. The source of noise doublets are, among other things, computational errors, so it is rather hard to struggle with them.

The properties of Padé approximants are described in further details in [4, ch. 5]. The algorithm used to find the most optimal fitting dilution cross sections is also presented there, but for completeness, it will also be discussed here.

The search for most optimal fitting dilution cross sections starts with the choice of $K-1$ cross section from the set of dilution cross sections $\mathcal{Z}$ (16). Let's designate the set of these cross sections with $\tilde{\mathcal{Z}}$. $\tilde{\mathcal{Z}}$ is sorted with some order and one of the cross sections in it is varied, for instance the first one. Being varied, the cross section takes all the values from $\mathcal{Z}$ except those present in $\tilde{\mathcal{Z}}$, and for every value the residual is calculated:

$$\delta=\max_{\substack{z\in\tilde{\mathcal{Z}}\\x\in\mathcal{X}}}(\delta[f_{t,0},\tilde{f}_{t,0}],\delta[f_{t,1},\tilde{f}_{t,1}],\delta[f_{x,0},\tilde{f}_{x,0}]).\quad(64)$$

When the variation is finished, the value corresponding to the smallest residual is written into the end of $\tilde{\mathcal{Z}}$, the varied first element is dumped and the whole procedure is repeated from start. The algorithm ends if either such $\tilde{\mathcal{Z}}$ was found that the correspondend residual is small enough, or the set $\tilde{\mathcal{Z}}$ did not change

after all its elements had been varied. In the latter case one can to increase K by one and try again, until either the satisfactory solution is found or K becomes equal to the number of elements in $\mathcal{Z}$.

A few important properties of the algorithm should be mentioned. First, the algorithm provides no guarantees to find the global minimum of the residual, only some local one. Second, the algorithm assumes the unicity of the Padé approximation constructed with respect to several points. While it is true in theory, in practice computational errors can influence on the construction dramatically, up to appearance of positive poles dependent on the order of choice of fitting points. Third, it can happen that the residual for total cross section is satisfactory small, while it is not true for the residual for partial cross sections. With ε being small, the addition of another fitting point may result in appearance of the noise doublet.

In spite of all its shortcomings, the algorithm provides adequate compromise between calculation speed and the quality of the result obtained.

2.2 Residual minimization[2]

The problem of finding subgroup parameters satisfying conditions (22)–(29) can be considered as the problem of finding the minimum of multidimensional function. In this case the function is the residual:

$$\delta = \max_{\substack{z\in\mathcal{Z}\\ x\in\mathcal{X}}}(\delta[f_{t,0}, \tilde{f}_{t,0}], \delta[f_{t,1}, \tilde{f}_{t,1}], \delta[f_{x,0}, \tilde{f}_{x,0}]). \tag{64}$$

Its arguments are subgroup parameters a_k, σ_k and $\sigma_{x,k}$. The feasible space is constrained with equations (22)–(26):

$$a_k \geq 0,\ \sigma_k \geq 0,\ \sigma_{x,k} \geq 0,\ \forall k,\ x, \tag{22}$$

$$\sum_{k=1}^{K} a_k = 1, \tag{23}$$

$$\sum_{k=1}^{K} a_k\sigma_k = \langle\sigma\rangle\,, \tag{24}$$

$$\sum_{k=1}^{K} a_k\sigma_{x,k} = \langle\sigma_x\rangle\,, \tag{25}$$

$$\sum_{x} \sigma_{x,k} = \sigma_k, \tag{26}$$

For convenience, let $s_k = \frac{a_k\sigma_k}{\langle\sigma\rangle}$, $s_{x,k} = \frac{a_k\sigma_{x,k}}{\langle\sigma_x\rangle}$. For some $m\colon 1 \leq m \leq k$

$$\begin{aligned} a_m &= 1 - \sum_{k\neq m} a_k,\\ s_m &= 1 - \sum_{k\neq m} s_k,\\ s_{x,m} &= 1 - \sum_{k\neq m} s_{x,k}. \end{aligned} \tag{65}$$

[2] Unfortunately, when I was working on this subject I lacked the knowledge of constrained multidimensional minimization methods, so the following is a rather naive theory. It is included here for completeness because it is implemented in the `subgroups` program.

Left sides of these equations have to be positive, hence the following system of inequalities:

$$\begin{cases} a_k \geq 0, \\ \sum\limits_{k \neq m} a_k \leq 1, \\ s_k \geq 0, \\ \sum\limits_{k \neq m} s_k \leq 1, \\ s_{x,k} \geq 0, \\ \sum\limits_{k \neq m} s_{x,k} \leq 1. \end{cases} \tag{66}$$

In n-dimensional space inequalities

$$\begin{cases} x_k \geq 0,\ k = 1, \ldots, n, \\ \sum\limits_{k=1}^{n} x_k \leq 1 \end{cases} \tag{67}$$

define standard simplex—a multidimensional rectangular triangle with sides coincident with orths. Consequently, if N is the number of partial cross sections, the solution has to be inside the $(K-1) \cdot (N+2)$-dimensional body which is the direct sum of $N+2$ $K-1$-dimensional standard simplexes. Constraints (22)–(25) are taken into account by bijective mapping of this body into the whole space of the same dimension and solving the minimization problem in this space. Since the body is the direct sum of standard simplexes, in order to construct bijective mapping it is enough to construct bijective mapping of one simplex to the whole space of the same dimension and apply it to every simplex.

Given bijective mapping $T(x, a)\colon [0; a] \to [-\infty; \infty]$ for some $a > 0$, bijective mapping of the standard simplex can be constructed as follows. Consider the line parallel to axis number m and passing through a point $\vec{x}$. The simplex cuts an interval such that its end points coordinates are

$$(x_1, \ldots, x_{m-1}, 0, x_{m+1}, \ldots, x_n) \text{ and } (x_1, \ldots, x_{m-1}, 1 - \sum_{\substack{k=1 \\ k \neq m}}^{n} x_k, x_{m+1}, \ldots, x_n).$$

The desired mapping is the application of T to every such interval for every point $\vec{x}$ inside the simplex and every axis m. Coordinates of the mapped point are:

$$\xi_m = T(x_m, 1 - \sum_{\substack{k=1 \\ k \neq m}}^{n} x_k),\ m = 1, \ldots, n. \tag{68}$$

As for the mapping T, it can be, for instance, cotangent:

$$T(x, a) = \cot \frac{\pi x}{a}. \tag{69}$$

Then the coordinates of the mapped point are

$$\xi_m = \cot \frac{\pi x_m}{1 - \sum\limits_{\substack{k=1 \\ k \neq m}}^{n} x_k},\ m = 1, \ldots, n, \tag{70}$$

and the inverse of T is obtained by solving the system of linear equations

$$\sum_{\substack{k=1\\k\neq m}}^{n} x_k + \frac{\pi}{\operatorname{acot}\xi_m} x_m = 1,\ m = 1, \ldots, n. \tag{71}$$

References

[1] M. Nikolaev, B. Ryazanov, M. Savoskin, A. Tsibulia. Multigroup approximation in neutron transport theory. Moscow, Energoatomizdat, 1984. (in Russian).

[2] E. Zhemchugov. The development of the algorithm for computation of subgroup parameters of neutron cross sections. Graduation thesis. Obninsk, Obninsk State University for Nuclear Power Engineering, 2008. (in Russian).

[3] G. Baker, P. Graves-Morris. Padé approximants. AW, 1981. ISBN 0201135124

[4] V. Vinogradov, E. Guy, N. Rabotnov. Analythical data approximation in nuclear and neutron physics. Moscow. Energoatomizdat, 1987. (in Russian).