NUMERICAL INTEGRATIONS

In most cases of any interest the equations of stellar structure cannot be integrated analytically. Instead, the integrations have to be performed numerically. With fast and inexpensive computers this is and easy thing to do. You may consult a book *NUMERICAL RECIPES* by W. Press et al. for many practical and accurate numerical integration techniques. Here, a very rudimentary description will be provided. In many cases this may be all you will need.

1) One equation, explicit method

Consider a first order, ordinary differential equation

$$\frac{dy}{dx} = f\left(x, y\right), \tag{ni.1a}$$

with the initial condition

$$y = y_0 \qquad \text{at} \quad x = x_0. \tag{ni.1b}$$

The simplest numerical integration technique is as follows. Choose an integration step Δx. Calculate the derivative at the starting point, $f(x_0, y_0)$, and calculate the change in variables x and y over the integration step:

$$x_1 = x_0 + \Delta x, \tag{ni.2}$$

$$y_1 = y_0 + f\left(x_0, y_0\right) \times \Delta x.$$

We may repeat this procedure as many times as we wish. Suppose we already know the values of x and y after "k" integration steps, i.e. we know x_k and y_k. We may calculate x_{k+1} and y_{k+1} according to

$$x_{k+1} = x_k + \Delta x, \tag{ni.3}$$

$$y_{k+1} = y_k + f\left(x_k, y_k\right) \times \Delta x.$$

This is the simplest integration scheme possible. It is not very accurate but if the step size Δx is made sufficiently small then we may achieve as high accuracy as we wish, at least in principle. In practice this will not be very efficient, as the number of steps may be prohibitively large if they are very small. The accuracy may be estimated as follows. The second of the two equations (ni.3) would be exact if the derivative $f(x, y)$ were calculated at the proper point between x_k and x_{k+1}. However, we do not know the location of this point; instead we expand the solution in a Taylor series at the point (x_k, y_k), and we retain the first two terms only. The largest term neglected is of the order $(\Delta x)^2$. A more accurate technique should allow for that term. This may be done in many ways. The simplest one is the second order Runge - Kutta method, according to which we calculate the derivative $f(x, y)$ twice while calculating one integration step. First, we make half a step:

$$x_{k+1/2} = x_k + 0.5 \times \Delta x, \tag{ni.4a}$$

$$y_{k+1/2} = y_k + 0.5 \times f\left(x_k, y_k\right) \times \Delta x.$$

Next, we calculate the derivative at the middle of the step, and we make the full step:

$$x_{k+1} = x_k + \Delta x, \tag{ni.4b}$$

$$y_{k+1} = y_k + f\left(x_{k+1/2}, y_{k+1/2}\right) \times \Delta x.$$

In this method the error is of the order $(\Delta x)^3$, i.e. it is much smaller than in the first technique.

There are many other, more accurate methods. One of them is the fourth order Runge - Kutta method, in which the derivative $f(x, y)$ is calculated 4 times at every integration step, and error is of the order $(\Delta x)^5$. There are also many ways to estimate the accuracy of any method. The simplest way is to make the integrations with two different step sizes. The difference in the results is approximately equal to the error. If you want to have more accurate results, you may have to use a smaller integration step Δx, or a higher order integration scheme.
Sometimes the results are numerically unstable, and taking smaller step size does not help. This usually happens when the derivative $f(x, y)$ is a difference of two large and nearly equal terms. This may physically mean that the two terms describe opposite reaction rates, like ionization and recombination, and the two are nearly equal, i.e. we are close to equilibrium. If some parameter, e.g. temperature changes slowly, the equilibrium changes slowly as well, while the two reactions may proceed very rapidly in two opposite directions. It does not make any sense to use a very small time step in this case, yet a long time step combined with integration scheme like the one given with equation (ni.4) may be unstable. In this technique we calculate the derivatives at the **beginning** and/or in the **middle** of the step. This is called an **explicit** method. It turns out that instead we may have to use an **implicit** method, in which we calculate the derivative at the **end** of the step.

2) One equation, implicit method

Here is a simple example of an **implicit** numerical integration. First, we calculate the derivative at the beginning of the step, and estimate the values of the variables at the end of the step:

$$x_{k+1} = x_k + \Delta x, \tag{ni.5}$$

$$y_{k+1,old} = y_k + f\left(x_k, y_k\right) \times \Delta x.$$

Of course, x_{k+1} is just what we want, but $y_{k+1,old}$ is only an estimate. We want y_{k+1} to satisfy the following equation:

$$y_{k+1} = y_k + f\left(x_{k+1}, y_{k+1}\right) \times \Delta x. \tag{ni.6}$$

Notice, that the quantity we are looking for, y_{k+1}, is on the left hand as well as on the right hand side of the equation (ni.6) . However, if the function $f(x, y)$ is nonlinear, then in general we cannot solve this equation analytically. Instead, we shall use the numerical Newton - Raphson technique. We shall write our equation in a form

$$F\left(y_{k+1}\right) \equiv y_{k+1} - y_k - f\left(x_{k+1}, y_{k+1}\right) \times \Delta x = 0. \tag{ni.7}$$

Our first approximation is $y_{k+1} \approx y_{k+1,old}$. We put it into the equation (ni.7) , and of course we shall find that $F(y_{k+1,old})$ is not equal to zero. Therefore, we expand it in a Taylor series and retain the first two terms to obtain

$$F\left(y_{k+1,old}\right) + \left(\frac{\partial F}{\partial y}\right)_{k+1,old} \delta y = 0, \tag{ni.8}$$

and we may find a new, corrected value of y_{k+1} according to

$$y_{k+1,new} = y_{k+1,old} + \delta y = y_{k+1,old} - F\left(y_{k+1,old}\right) / \left(\frac{\partial F}{\partial y}\right)_{k+1,old}, \tag{ni.9}$$

where

$$\left(\frac{\partial F}{\partial y}\right)_{k+1,old} = 1 - \left(\frac{\partial f}{\partial y}\right)_{k+1,old} \Delta x. \tag{ni.10}$$

The new value of $y_{k+1,new}$ may be used to calculate the value of $F(y)$, and to see how close it is to zero. In general, we may want to repeat the iteration as may times as necessary, to make $F = 0$ to the desired accuracy. Once this is accomplished, the integration step number "k+1" has been completed. The implicit integration scheme is complicated, slow, and only first order, i.e. the errors are of the order $(\Delta x)^2$. It should not be used, unless the explicit method does not work.

3) Many equations, explicit method

In general, we may have more then one ordinary differential equation, or we may have higher order differential equations. Any high order equation may be always replaced by a set of first order equations, so we shall restrict ourselves to this case only. Imagine, that we have n first order, non-linear differential equations, which we shall write in a form

$$\frac{dx_i}{dx_1} = f_i\left(x_j\right), \qquad i = 2,3,4,....n,n+1, \qquad j = 1,2,3,....n,n+1, \tag{ni.11}$$

$$f_i\left(x_j\right) \equiv f_i\left(x_1, x_2, x_3, ..., x_n, x_n + 1\right).$$

Notice, that there are n+1 equations (ni.11), because we treat the independent variable x_1 the same way as the n dependent variables, $x_2,\ x_2,\ \ldots\ , x_n,\ x_{n+1}$; this is convenient from numerical point of view.
We may want to choose the integration step so, that none of the variables x_i varies by more than $\Delta_{i,max}$. The selection of the step size can be made at the beginning of every integration step by calculating all the derivatives, and calculating Δx_1 according to

$$\Delta x_1 \le \Delta_{1,max}, \tag{ni.12}$$

$$\Delta x_i = f_i \times \Delta x_1 \le \Delta_{i,max}, \qquad i = 2,3,4,....n,n+1,$$

i.e.

$$\Delta x_1 = min\left[\Delta_{1,max}, \frac{\Delta_{i,max}}{f_i}\right], \qquad i = 2,3,4,....n,n+1. \tag{ni.13}$$

Now, that the integration step has been selected, we may use the second order Runge - Kutta method to make an integration step. We shall use a subscript "k" to indicate the values of all variables at the beginning of the step, a subscript "k+1/2" for the middle of the step, and a subscript "k+1" for the end of the step. It will be convenient to define

$$f_1 \equiv \frac{dx_1}{dx_1} = 1. \tag{ni.14}$$

We make an integration step as follows:

$$x_{i,k+1/2} = x_{i,k} + 0.5 \times f_{i,k} \times \Delta x_1, \qquad i = 1,2,3,....n,n+1, \tag{ni.15}$$

$$x_{i,k+1} = x_{i,k} + f_{i,k+1/2} \times \Delta x_1, \qquad i = 1,2,3,....n,n+1, \tag{ni.16}$$

where

$$f_{i,k} = f_i\left(x_{j,k}\right), \qquad f_{i,k+1/2} = f_i\left(x_{j,k+1/2}\right), \qquad i,j = 1,2,3,....n,n+1. \tag{ni.17}$$

In this way the integration step has been completed and we may proceed to the selection of the next step size, according to equation (ni.13) , with the subscript "k" replaced by "k+1".

4) Many equations, implicit method.

If we have to use an implicit method then the selection of the integration step size may be done in the same way as it was done for the explicit scheme, i.e. with the equation (ni.13) . Given the step size, Δx_1, we make an estimate of all variables at the end of the step

$$x_{i,k+1,old} = x_{i,k} + f_{i,k} \times \Delta x_1, \qquad i = 1, 2, 3, ...n, n+1. \tag{ni.18}$$

We define

$$F_i \equiv x_{i,k+1} - x_{i,k} - f_{i,k+1} \times \Delta x_1, \qquad i = 2, 3, ...n, n+1, \tag{ni.19}$$

where

$$f_{i,k+1} = f_{i,k+1}\left(x_{j,k+1}\right), \qquad i, j = 1, 2, 3, ...n, n+1, \tag{ni.20}$$

are the derivatives (i.e. right hand sides of differential equations) calculated at the end of the integration step. We do not have to include f_1, in these considerations, as it is constant, $f_1 = 1$, and $F_1 \equiv 0$.

Of course, we would like to have $F_i = 0$, and of course this will not be achieved with the first guess as provided with equations (ni.18) . Therefore, we expand F_i in a Taylor series, and we retain only linear terms. In general f_i depend on all variables x_j, and we have

$$F_i + \sum_{j=2}^{n+1} \frac{\partial F_i}{\partial x_{j,k+1}} \delta x_j = 0, \qquad i = 2, 3, ...n, n+1, \tag{ni.21a}$$

$$\frac{\partial F_i}{\partial x_{j,k+1}} = \delta_{i,j} - \frac{\partial f_{i,k+1}}{\partial x_{j,k+1}} \Delta x_1, \qquad i, j = 2, 3, ...n, n+1, \tag{ni.21b}$$

$$\delta_{i,j} = 1 \qquad \text{for} \quad i = j, \qquad\qquad \delta_{i,j} = 0 \qquad \text{for} \quad i \neq j. \tag{ni.21c}$$

These are n linear equations with n unknowns, so they can be solved using standard matrix inversion techniques, and we may calculate the corrected values of all variables at the end of integration step

$$x_{i,k+1,new} = x_{i,k+1,old} + \delta x_i, \qquad i = 2, 3, ...n, n+1. \tag{ni.22}$$

Using these new values we may calculate new values of F_i, and compare them to zero. If they are too different from zero then we have to repeat the iteration, calculate the new corrections, and so on, until all F_i and all the corrections δx_i are sufficiently close to zero.

The iterative process may converge, or may not converge. There are many ways to modify the technique if the iterations in the implicit integration scheme do not converge. Sometimes we may have to take larger or smaller integration step. Sometimes we may have to reduce **all** the corrections δx_i by the **same** factor. This may be interpreted as follows. We are looking for a solution of n equations $F_i = 0$ in n dimensional space. We have a guess for a solution, the n coordinates $x_{i,k+1,old}$. The vector of corrections δx_i points towards the solution we seek, but if we apply its whole magnitude we may "overshoot". In such case we reduce the **length** of the correction vector, while keeping its **direction**. We accomplish this be dividing all corrections by the same factor. This will slow down the convergence process, but may increase the range over which the iterations converge. Sometimes, instead of using a fixed reducing factor we may require that none of the corrections should exceed some maximum value, $\delta x_{i,max}$, and we reduce all the corrections by the **same** factor selected so that no correction exceeds its allowed limit. This is similar to the method used for selecting the size of the integration step.

5) Simple stellar models - white dwarfs.

Some stellar models, for example polytropes and white dwarfs, are described with two ordinary differential equations that may be written in dimensionless form. The equations have one free parameter: ploytropic index n, and dimensionless Fermi energy at the center, respectively. If we choose the value of any of those parameters then the conditions at the center can be treated as the initial conditions for the integrations to be carried out all the way from the center to the surface, the surface defined to be at a radius where density vanishes. As an example we may take the case of white dwarfs, which have a structure described with equations

$$\frac{dx_2}{dx_1} = x_1^2 \left(x_3 - 1\right)^{1.5}, \tag{ni.23a}$$

$$\frac{dx_3}{dx_1} = -\frac{x_2}{x_1^2} x_3^{1/2}. \tag{ni.23b}$$

Dimensionless variables x_1, x_2, and x_3, are defined with:

$$r \equiv \alpha_r x_1, \qquad M_r \equiv \alpha_m x_2, \qquad 1 + x^2 \equiv x_3. \tag{ni.24}$$

where M_r is mass within shell with radius r, and $x = p_F/mc$ is dimensionless Fermi momentum. The two scaling parameters are

$$\alpha_r = \left(\frac{B}{8\pi G}\right)^{1/2} \frac{1}{A\mu_e} = 5.455 \times 10^8 \mu_e^{-1} \left(\,\mathrm{cm}\,\right) = 0.00784\ R_\odot \mu_e^{-1}, \tag{ni.25a}$$

$$\alpha_m = \frac{1}{\left(2\pi\right)^{1/2}} \left(\frac{B}{G}\right)^{1.5} \frac{1}{\left(2A\mu_e\right)^2} = 2.00 \times 10^{33} \mu_e^{-2} \left(\,\mathrm{g}\,\right) = 1.005\ M_\odot \mu_e^{-2}. \tag{ni.25b}$$

The boundary conditions are

$$x_3 = x_{3,c}, \qquad x_2 = 0, \qquad \text{at} \;\; x_1 = 0, \qquad \text{(inner boundary conditions)}, \tag{ni.26a}$$

$$x_3 = 1, \qquad x_2 = x_{2,s}, \qquad \text{at} \;\; x_1 = x_{1,s}, \qquad \text{(outer boundary condition)}, \tag{ni.26b}$$

where $x_{3,c}$ is the central value of x_3, and $x_{2,s}$ and $x_{1,s}$ are the surface values of dimensionless mass and radius, respectively. The central density, the total mass and radius of a white dwarf are given as

$$\rho_c = 0.981 \times 10^6 \mu_e \left(x_{3,c} - 1\right)^{1.5} \ \left(\,\mathrm{g\ cm^{-3}}\right), \qquad M = \alpha_m x_{2,s}, \qquad R = \alpha_r x_{1,s}, \tag{ni.27}$$

where $\mu_e = 2/(1 + X)$ is the mean number of nucleons per electron, and X is hydrogen abundance by mass fraction.

Let us choose $x_{3,c}$ as a free parameter, and let us use the inner boundary conditions (ni.26a) as the initial conditions. There is one technical problem at the center: $x_1 = 0$ and $x_2 = 0$ simultaneously, and the right hand side of the equation (ni.23b) cannot be calculated. This is a typical problem: we cannot begin numerical integrations right at the center. We have to expand the solution of the differential equations in a power series, and calculate analytically values of all variables near the center, at a small but finite radius. The first two terms in the expansion are:

$$x_2 = \frac{1}{3} x_c^3 x_1^3 - \frac{1}{20} x_{3,c}^{1/2} x_c^4 x_1^5, \tag{ni.28a}$$

$$x_3 = x_{3,c} - \frac{1}{6} x_{3,c}^{1/2} x_c^3 x_1^2, \tag{ni.28b}$$

where

$$x_c^2 \equiv x_{3,c} - 1, \qquad \rho_c = A\mu_e x_c^3 = 0.981 \times 10^6\ \mu_e x_c^3\ \left(\,\mathrm{g\ cm^{-3}}\right), \tag{ni.29}$$

and ρ_c is the central density. You should verify this solution by inserting equations (ni.28) into (ni.23) .

For the expansion to be accurate we require the second term to be much smaller than the first term in both equations (ni.28) . This may be used as a criterion for the choice of x_1 in the expansion, i.e. how big should be the radius of the innermost sphere covered with analytical expansion. When a choice of x_1 is made we may use equations (ni.28) to calculate x_2 and x_3. Now we may begin numerical integrations from that point. We shall continue the integrations until x_3 becomes less than 1 at the end of some integration step. We may interpolate between the beginning and the end of that step to find the location at which $x_3 = 1$, i.e. we can find the location of the white dwarf surface. Equations (ni.27) may now be used to calculate stellar mass and radius, as well as central density, in physical units.

6) Initial conditions, boundary conditions.

Suppose now that instead of choosing the central value of x_3, which is equivalent to choosing central density ρ_c, we have chosen the total stellar mass M, or the total stellar radius R as a free parameter, fixed its value, and tried to solve the structure equations (ni.23) . This time we would not know what should be the value of $x_{3,c}$ or ρ_c, and we would have to leave $x_{3,c}$ as an adjustable parameter in the inner boundary condition, while another condition, e.g. fixed M, would be given at the surface, i.e. as the outer boundary condition. This time it would be very much more difficult to find the solution, as it would have to satisfy two **boundary** conditions at the two opposite ends of integration interval. A simple, one way integration which was possible when we had the **initial** conditions given at the center would not be sufficient. A number of trial integrations would be necessary.

In many cases we are not interested in one stellar model with some specific value of mass or central density, but rather in a series of models covering a large range of these parameters. In such a case it is not important which parameter, total mass M, or central density ρ_c is chosen as a parameter that labels the models. We should choose the one that is more convenient. In our case it is clearly more convenient to choose central density of a white dwarf. In this case we are solving the **initial value problem**, and the whole structure, and the total mass of every model come out from a single numerical integration. If instead the total mass is chosen as the parameter that selects models of white dwarfs then for each model we have to solve the **boundary value problem**, which requires many iterative integrations for every model.

7) Zero age main sequence stars (ZAMS) - a fittting method.

Zero age main sequence stars are defined as stars that have already ignited hydrogen, hydrogen burning produces as much energy as is radiated from the surface, but there is no reduction of the original hydrogen content. These stars are in thermal and hydrostatic equilibria, they are chemically homogeneous, and have the original chemical composition of the interstellar matter from which they have just formed. Of course, these requirements cannot be rigorously met, all at the same time, but it is convenient to use them to define a simplified but useful type of stellar models.

Our problem is posed as follows. We define the chemical composition. In the simplest case we choose just two parameters: X and Z. X is hydrogen abundance, and Z is heavy element abundance by mass fraction. Helium abundance is $Y = 1 - X - Z$. We want to find a solution of the four stellar structure equations for a given total stellar mass M. The four equations may be written as

$$\frac{dT}{dM_r} = \frac{T}{P}\nabla_T \frac{dP}{dM_r}, \tag{ni.30a}$$

$$\frac{d\rho}{dM_r} = \frac{\rho}{P}\nabla_\rho \frac{dP}{dM_r}, \tag{ni.30b}$$

$$\frac{dr}{dM_r} = \frac{1}{4\pi r^2 \rho}, \tag{ni.30c}$$

$$\frac{dL_r}{dM_r} = \epsilon, \tag{ni.30d}$$

where

$$\frac{dP}{dM_r} = -\frac{GM_r}{4\pi r^4}. \tag{ni.31a}$$

$$\nabla_T = min\left[\nabla_{rad}, \nabla_{ad}\right], \tag{ni.31b}$$

$$\nabla_{rad} \equiv \frac{\kappa L_r}{16\pi cGM_r}\frac{3P}{aT^4}, \tag{ni.31c}$$

$$\nabla_{ad} \equiv \left(\frac{\partial \ln T}{\partial \ln P}\right)_S \tag{ni.31d}$$

$$\nabla_\rho \equiv \frac{d\ln\rho}{d\ln P} = \left[1 - \left(\frac{\partial \ln P}{\partial \ln T}\right)_\rho \nabla_T\right] / \left(\frac{\partial \ln P}{\partial \ln \rho}\right)_T. \tag{ni.31e}$$

and ϵ, κ, P, ∇_{ad}, are all assumed to be known functions of temperature, density, and chemical composition.

The four equations (ni.30) have to be supplemented with the boundary conditions. These may be written as:

$$\rho = 10^{-12}\left(\mathrm{g\ cm^{-3}}\right), \qquad T = \left(\frac{L}{8\pi R^2\sigma}\right)^{1/4}(\mathrm{K}), \qquad \text{at}\ \ M_r = M, \tag{ni.32a}$$

and

$$r = 0, \qquad L_r = 0, \qquad \text{at}\ \ M_r = 0. \tag{ni.32b}$$

These are the outer boundary conditions (ni.32a) and the inner boundary conditions (ni.32b) . We have two adjustable parameters at the surface: stellar radius and luminosity, R and L, and two adjustable parameters at the center: central temperature and density, T_c and ρ_c.

This is a truly boundary value problem, which cannot be reduced by any trick to any initial value problem. It has to be solved by integrating the four stellar structure equations from the surface inwards, to some fitting point at $M_r = M_f$, and integrating the same equations from the center to M_f, and finally trying to match the two sets of integrations at the fitting point. At the center the stellar structure equations suffer from the same problem as the equations for the white dwarf case: the right hand side of the equations (ni.30a) , (ni.30b) and (ni.31a) is of the 0/0 type at the center. Therefore, we have to make an analytical expansion there in order to start numerical integrations. The general approach to the fitting procedure is as follows.

Let us consider a stellar model in a hydrostatic and thermal equilibria, with the total mass M, and the profile of chemical composition $X(M_r)$, specified. The four differential equations describing stellar structure may be written in a form

$$\frac{dx_i}{dx_o} = y_i, \qquad i = 1, 2, 3, 4, \tag{ni.33}$$

where x_o is the independent space-like variable, usually M_r, and x_i are the four dependent variables, like T, ρ, r, and L_r. The boundary conditions have two adjustable parameters at the center, z_1 and z_2, and two parameters at the surface, z_3 and z_4. We may have for example: $z_1 = \rho_c$, $z_2 = T_c$, $z_3 = R$, $z_4 = L$.

The equations of stellar structure are integrated from the surface inwards, down to the fitting point at $M_r = M_f$, and from the center outwards to the same fitting point. The results of the envelope integrations at the fitting point may be written as

$$x_{i,e} = x_{i,e}\left(z_3, z_4\right), \qquad i = 1, 2, 3, 4, \tag{ni.34}$$

and the results of the core integrations as

$$x_{i,c} = x_{i,c}\left(z_1, z_2\right), \qquad i = 1, 2, 3, 4. \tag{ni.35}$$

At the fitting point the differences between the core and envelope integrations are calculated

$$\Delta x_i \equiv x_{i,c} - x_{i,e}, \qquad i = 1, 2, 3, 4. \tag{ni.36}$$

The model is found when $\Delta x_i = 0$. In general, this is not so when wrong values of the boundary parameters are used. The iterative process of finding the correct values is based on the linearized equation

$$\Delta x_i + \sum_{j=1}^{4} c_{ij} \delta z_j = 0, \qquad i = 1, 2, 3, 4, \tag{ni.37a}$$

where

$$c_{ij} \equiv \frac{\partial \Delta x_i}{\partial z_j} \qquad i, j = 1, 2, 3, 4, \tag{ni.37b}$$

These equations are solved to find the corrections to the boundary parameters, δz_j, and the new values of the boundary parameters, $z_j + \delta z_j$. In order to solve the equations (ni.37a) the determinant of the matrix $|c_{ij}|$ has to be calculated. We are basically using a Newton - Raphson technique to find the solution of the four non-linear equations (ni.36) . When the iterations converge, i.e. when Δx_i and δz_j are sufficiently small, the zero age main sequence model has been found.