EXCITONS

Geert Brocks

January 8, 2019

Many physics problems involve solving one or more (partial) differential equations. Hence, how to solve a differential equation numerically, is a vital skill if you use the computer to attack a physics problem. In numerical mathematics you have solved differential equations expressed as initial value problems (IVPs) and/or boundary value problems (BVPs). Here we look at another class of differential equations that are of prime importance in physics: eigenvalue problems (EVPs). EVPs are slightly more difficult to solve than IVPs or BVPs. Besides having to find a function that obeys the differential equation, one also has to find an eigenvalue.

In this project we make use of finite-difference techniques similar to the ones you have had in numerical mathematics. This allows for applying a special numerical technique to solve EVPs, which is called the *shooting method*.[1,2] This method uses a penalty function that is zero at an eigenvalue λ_i, $F(\lambda_i) = 0$, and nonzero otherwise, i.e. $F(\lambda) \neq 0$ if $\lambda \neq \lambda_i$. The problem then becomes one of root searching. With some special knowledge of $F(\lambda)$ this can be done in an efficient way.

We start from a simple central difference scheme to convert an EVP into a finite-difference expression. Depending on the physics problem at hand, there are ways to markedly improve the accuracy and efficiency of this scheme without complicating it too much. Numerov's method is a cheap way to introduce a higher order (and generally better) approximation. It works for a particular type of differential equations only, but this type occurs frequently in physics. Finite differencing is mostly based upon equidistant grids, which are not always most efficient computationally. A logarithmic grid is a non-equidistant grid that is particularly suited for the EVP we are interested in here.

We focus on the radial Schrödinger equation as the prime example of an EVP. The numerical techniques we use to solve this problem are explained in Sec. 1. We apply these techniques to a specific problem: excitons in semiconductors. The experimental and theoretical backgrounds of this problem are given in Secs. 2 and 3, respectively.

[1] Have a guess from which country the term originates. A fantastic method, the best method ever, a method made great again, etcetera.

[2] More general, EVPs are often solved using numerical linear algebra techniques. The latter are not part of the present course. There is no good reason for this, except that one has to make some selection of topics. Linear algebra is more mathematics than physics, and the course is called "computational physics". I admit, a somewhat lame excuse.

Contents

1 Numerical background

Sec. 1.1 defines the differential equation we are going to solve in this project. Secs. 1.2 and 1.3 describe our basic numerical approaches: a finite-difference approximation combined with forward and backward propagations to solve the differential equation, and the shooting method to find the eigenvalue. These basic techniques can be modified to improve their accuracy and their efficiency. The method described in Sec. 1.4 allows one to find the eigenvalue faster. Sec. 1.5 describes a technique that makes the finite-difference approach more accurate, and Sec. 1.6 defines a more optimal finite-difference grid for our problem.

1.1 The radial Schrödinger equation

We consider the radial Schrödinger equation.

$$\left\{-\frac{\hbar^2}{2\mu}\frac{d^2}{dr^2}+\frac{\hbar^2 l(l+1)}{2\mu r^2}+V(r)\right\}\zeta(r)=E\zeta(r);\quad r\in[0,\infty), \tag{1}$$

where μ is the particle's mass, $l=0,1,...$ is its angular momentum quantum number, E is its energy, and $V(r)$ is the potential, see Sec. 3.2. We are interested in bound states, which gives us the conditions

$$E<V(\infty);\quad \lim_{r\to\infty}\zeta(r)=0;\quad \lim_{r\to 0}\zeta(r)=0. \tag{2}$$

We will solve this problem numerically. It resembles a boundary value problem, but it is slightly more complicated. We know that Eq. 1 has solutions obeying the boundary conditions Eq. 2, if the potential is sufficiently well-behaved.[3]However, bound states can be found only at certain discrete energy levels $E=E_n$, which can be labeled by an integer quantum number n.[4] This means that, not only do we have to find the function $\zeta_n(r)$, but also the specific energy E_n that makes this function possible.

Numerical approaches usually start by converting the equation into one based on dimensionless variables. These should be chosen such, that one does not get into numerical troubles right from the start.[5] It makes sense to introduce a parameter

$$r_0=\hbar/\sqrt{2\mu V_0}, \tag{3}$$

where V_0 is a constant of dimension energy that is a sensible unit for the strength of the potential. If the potential has a minimum, $V_0=V_{\min}$, could a sensible choice. For the Coulomb potential, which does not have a minimum, one can choose the Rydberg as a unit of energy, see Sec. 3.3. The real positive number r_0 has dimension m, and it represents the

[3]What exactly that means, we leave to the mathematicians. The potential should have the shape of a well, $V(r)<V(\infty)$, for a sufficiently large range in r, and that well should be sufficiently deep, in order to be able to support bound states.

[4]The energies are "quantized"; after all, it is quantum mechanics. The phenomenon is quite general however. If you enclose electromagnetic waves or acoustic waves in a cavity (a "potential"), then standing waves are only possible at certain discrete frequencies.

[5]SI units are usually not very practical in numerical calculations. For instance, $\hbar=1.05\times10^{-34}$ m^{2}kg/s, $e=1.60\times10^{-19}$ C, $a_0=5.29\times10^{-11}$ m, etc.. Manipulating such small numbers can get you into numerical problems. Multiplying too many of them leads to underflow (a number too small for the computer to represent), and dividing by too many of them gives overflow (a number too large for the comnputer to represent), for instance.

natural unit of length in the problem. If we introduce a dimensionless radius $\rho = r/r_0$, the wave function and the potential can be expressed as $\zeta(\rho)$ and $V(\rho)$,[6] which allows one to rewrite Eq. 1 as

$$\frac{d^2\zeta(\rho)}{d\rho^2} = \{W(\rho) - \lambda\}\,\zeta(\rho) \quad \text{with} \quad W(\rho) = \frac{V(\rho)}{V_0} + \frac{l(l+1)}{\rho^2} \quad \text{and} \quad \lambda = \frac{E}{V_0}, \tag{4}$$

with the constraints

$$\lambda < V(\infty); \quad \lim_{\rho\to\infty} \zeta(\rho) = 0; \quad \lim_{\rho\to 0} \zeta(\rho) = 0. \tag{5}$$

Besides knowing the function values at the boundaries, in some cases we can also derive the asymptotic behavior of the functions near the boundaries. For instance, assuming we have a potential where $\lim_{\rho\to 0} \rho^2 V(\rho) = 0$, then for small ρ Eq. 4 can be approximated by

$$\frac{d^2\zeta(\rho)}{d\rho^2} \approx \frac{l(l+1)}{\rho^2}\zeta(\rho); \quad \rho \text{ small}; \; l > 0. \tag{6}$$

This differential equation is easily solved: $\zeta(\rho) = A\rho^{l+1} + B\rho^{-l}$. In view of Eq. 5, we must set $B = 0$, so

$$\lim_{\rho\to 0} \rho^2 V(\rho) = 0 \implies \zeta(\rho) \approx A\rho^{l+1}; \quad \rho \text{ small}. \tag{7}$$

Although Eq. 6 is valid only for $l > 0$, Eq. 7 is valid for all l, i.e., $l \geq 0$.[7]

If we have a potential where $\lim_{\rho\to\infty} V(\rho) = 0$, then for large ρ Eq. 4 can be approximated by

$$\frac{d^2\zeta(\rho)}{d\rho^2} \approx \lambda\zeta(\rho), \tag{8}$$

which is easily solved: $\zeta(\rho) = C\exp\left[-\rho\sqrt{-\lambda}\right] + D\exp\left[\rho\sqrt{-\lambda}\right]$. Again, because of Eq. 5, we have to set $D = 0$, so

$$\lim_{\rho\to\infty} V(\rho) = 0 \implies \zeta(\rho) \approx C\exp\left[-\rho\sqrt{-\lambda}\right] \quad \rho \text{ large}. \tag{9}$$

We find solutions $\zeta(\rho)$ and λ to Eq. 4, subject to the constraints of Eq. 5. If the potential obeys certain constraints, we have the asymptotic behaviours of Eqs. 7 and 9. If not, it need not be a disaster, but we have to be a bit more alert.

1.1.1 Solution strategy

We apply a numerical technique called *shooting*, which amounts to the following. One starts with a reasonable guess of the eigenvalue $\lambda = \lambda^{(0)}$. Then, by a finite difference approximation of the differential equation, Eq. 4, one obtains a reasonable guess for $\zeta^{(0)}(\rho)$ that obeys the boundary conditions, Eq. 5. From $\zeta^{(0)}(\rho)$ one determines a better guess $\lambda^{(1)}$, solve Eq. 4 again to obtain $\zeta^{(1)}(\rho)$, etcetera, until we have convergence.

[6] Mathematicians would complain that I should have used different symbols, but I have too many of them already. The idea is to express the wave and the potential as a function of a dimensionless variable.

[7] It requires a more careful analysis of the asymptotic limit for $l = 0$.

1.2 Solving the differential equation: propagating finite differences

We discretize Eq. 4 on an equidistant grid $\rho_j = jh;\ j = 0, 1, .., N$, where h is the step size, and N sufficiently large to cover the whole wave function, such that $\zeta(Nh) \approx 0$. For the second derivative we start from the central difference approximation

$$\frac{d^2\zeta(\rho)}{d\rho^2} \approx \frac{1}{h}\left[\frac{d\zeta(\rho + h/2)}{d\rho} - \frac{d\zeta(\rho - h/2)}{d\rho}\right] = \frac{1}{h}\left[\left(\frac{\zeta(\rho + h) - \zeta(\rho)}{h}\right) - \left(\frac{\zeta(\rho) - \zeta(\rho - h)}{h}\right)\right]. \tag{10}$$

One then derives the forward propagation algorithm

$$\begin{aligned} &\text{set} \quad \zeta_f(\rho_0); \zeta_f(\rho_1); \\ &\text{do } j = 1, ..., M \quad \zeta_f(\rho_{j+1}) = 2\zeta_f(\rho_j) - \zeta_f(\rho_{j-1}) + h^2\left\{W(\rho_j) - \lambda^{(0)}\right\}\zeta_f(\rho_j); \end{aligned} \tag{11}$$

We set $\zeta_f(\rho_0) = 0$, using the substript f to indicate that we generate this function by forward propagation. To start we also need $\zeta_f(\rho_1)$, where in principle we can choose any value we like. If our potential obeys Eq. 7, we can check whether this value is reasonable (A should not become too weird). The algorithm of Eq. 11 is then used to determine $\zeta_f(\rho_2)$, subsequently $\zeta_f(\rho_3)$, etcetera.

Sometimes one cannot start at or too close to $\rho = 0$, because $W(\rho \downarrow 0) \to \infty$. This is the case for the Coulomb potential, for instance, where $V(\rho) \propto 1/\rho$, or if $l > 0$. One is forced to displace the grid to $\rho_j = a + jh;\ j = 0, 1, .., N$, where a is sufficiently large, such that $W(a)$ does not blow up. Then it helps to have Eq. 7 to set our starting values to $\zeta_f(\rho_0) = \zeta_f(a) = a^{l+1}$ and $\zeta_f(\rho_1) = \zeta_f(a + h) = (a + h)^{l+1}$ (choosing $A = 1$).

In principle, forward propagation by Eq. 11 gives you a solution over the full domain. However, it is not a good idea to propagate all the way up to the end-point $\rho_N = Nh$. We know that our eigenvalue equation, Eq. 4, has well-behaved, bounded solutions $\zeta(\rho)$ only for certain specific values of λ. So, unless by accident we have chosen one of these values as our guess $\lambda^{(0)}$, the $\zeta(\rho)$ we obtain by propagation will be a bit weird. It will contain a part that is exponentially growing, and does not go to zero for $\rho \to \infty$. For exactly the right λ, $\zeta(\rho)$ will be an exponentially decaying function, as in Eq. 9. Numerically, however, this information is easily lost, as it is overwhelmed by an expontially growing part if we don't have exactly the right λ (and we never have exactly the right λ, because the computer operates with numbers at finite precision).

So it is best to propagate Eq. 11 forward only onto a "safe" point $\rho_M = Mh;\ 0 < M < N$, where M is a point somewhere in the middle region where the function $\zeta(\rho)$ is not yet exponential. In practice, this point is often chosen close to the *classical outer turning point*, where $W(\rho_M) - \lambda^{(0)} \approx 0$.[8] Actually, it is best to propagate one point further to ρ_{M+1}, as in Eq. 11, for reasons that will become clear later on.

Eq. 10 also allows for a backward propagation formula

$$\begin{aligned} &\text{set} \quad \zeta_b(\rho_N);\ \zeta_b(\rho_{N-1}); \\ &\text{do } j = N - 1, ..., M \quad \zeta_b(\rho_{j-1}) = 2\zeta_b(\rho_j) - \zeta_b(\rho_{j+1}) + h^2\left\{W(\rho_j) - \lambda^{(0)}\right\}\zeta_b(\rho_j); \end{aligned} \tag{12}$$

[8] Propagating outwards, this is the point where a classical particle would bounce back and reverses its motion. A quantum particle can tunnel into the "forbidden" region. In this tunneling region one finds an exponentially decaying wave function. The classical turning point is easily determined, as $W(\rho_M) - \lambda^{(0)}$ changes sign there.

We set the function values of the two final points $\zeta_b(\rho_N)$ and $\zeta_b(\rho_{N-1})$, using the subscript b to indicate that we generate this function by backward propagation. If Eq. 9 holds, we can use this as a check to see whether our values are reasonable. The backward propagation algorithm can then be used to find $\zeta_b(\rho_{N-2})$, subsequently $\zeta_b(\rho_{N-3})$, etcetera. We stop at the point ρ_{M-1}, so we have a small overlap with the forward propagated solution.

1.3 Finding the eigenvalues: the shooting method

We don't have a full solution yet, only the two halves, $\zeta_f(\rho)$, for $0 \leq \rho \leq \rho_{M+1}$, and, $\zeta_b(\rho)$, for $\rho_{M-1} \leq \rho \leq \rho_N$. We must connect these two halves at the point ρ_M to form one function. A proper wave function should be both *continuous*, as well as *differentiable*. Continuity is easily assured; just define two functions

$$\tilde{\zeta}_f(\rho) = \zeta_f(\rho)/\zeta_f(\rho_M) \quad \text{and} \quad \tilde{\zeta}_b(\rho) = \zeta_b(\rho)/\zeta_b(\rho_M), \tag{13}$$

and we will have $\tilde{\zeta}_f(\rho_M) = \tilde{\zeta}_b(\rho_M) = 1$. Differentiability is not assured, i.e., the two functions will generally **NOT** obey

$$d\tilde{\zeta}_f(\rho_M)/d\rho = d\tilde{\zeta}_b(\rho_M)/d\rho. \tag{14}$$

Only if our guess $\lambda^{(0)}$ happens to be one of the eigenvalues of Eq. 4, we can expect our generated function to be a proper wave function, i.e., continuous and differentiable. One can also turn the argument around: if Eq. 14 happens to be obeyed by our generated function, then it is a proper wave function, therefore $\lambda^{(0)}$ is an eigenvalue of Eq. 4.

We have now a means of searching for eigenvalues. Define a function of λ, $F(\lambda) = 2h\left[d\tilde{\zeta}_b(\rho_M)/d\rho - d\tilde{\zeta}_f(\rho_M)/d\rho\right]$, or in finite difference

$$F(\lambda) = \left[\tilde{\zeta}_{b,\lambda}(\rho_{M+1}) - \tilde{\zeta}_{b,\lambda}(\rho_{M-1})\right] - \left[\tilde{\zeta}_{f,\lambda}(\rho_{M+1}) - \tilde{\zeta}_{f,\lambda}(\rho_{M-1})\right] \tag{15}$$

where $\tilde{\zeta}_{f,\lambda}$ has been generated by forward propagation, Eq. 11, and $\tilde{\zeta}_{b,\lambda}$ by backward propagation, Eq. 12, with the value λ. The expression for F is obtained by discretization of Eq. 14, using the mid-point expression for the derivative.[9]

Here is the main point: according to Eq. 14, if the function $F(\lambda) = 0$, then λ is an eigenvalue of Eq. 4. In other words, our problem of finding eigenvalues has now become the problem of finding the roots of $F(\lambda)$. We can use any root-searching algorithm that does not require additional information besides $F(\lambda)$ (such as $dF/d\lambda$, for instance). For instance, one can start with the simplest one: *bisection,* which you should know from numerical mathematics.

1.4 Improving the eigenvalue search: shooting with a better aim

Bisection is a very general algorithm, which is guaranteed to give you a solution, but it is also a relatively slow algorithm. It would be nice to have an algorithm that is more tailored to fit our specific problem and is faster.

[9] This is why we need the overlap between the forward and the backward generated functions. You know that the mid-point expression generally gives a better approximation to the derivative than the forward or backward Euler expressions.

For a general value of λ, the function $F(\lambda) \neq 0$, and the function

$$\tilde{\zeta}_\lambda(\rho) = \begin{cases} \tilde{\zeta}_f(\rho) & 0 \leq \rho \leq \rho_M \\ \tilde{\zeta}_b(\rho) & \rho_M \leq \rho \leq \rho_N \end{cases}$$

is not an eigenfunction belonging to the potential $W(\rho)$, cf. Eq. 4. If $F(\lambda) \neq 0$, then $\tilde{\zeta}_\lambda(\rho)$ has a cusp (a discontinuous first derivative) at $\rho = \rho_M$, which makes it a somewhat strange function. Nevertheless there exists a potential that gives a wave function with a cusp: a δ-function potential. A potential of the type $W_\delta(\rho) = \alpha\delta(\rho - \rho_M)$ leads to a cusp in the wave function at $\rho = \rho_M$ of size

$$\left.\frac{d\zeta}{d\rho}\right|_{\rho\downarrow\rho_M} - \left.\frac{d\zeta}{d\rho}\right|_{\rho\uparrow\rho_M} = \alpha\zeta(\rho_M) \tag{16}$$

For a proof of this statement, see Griffiths' quantum mechanics book, section 2.5.

Turning the argument around, the function $\tilde{\zeta}_\lambda(\rho)$ with its cusp at $\rho = \rho_M$ is the solution of Eq. 4 with the potential $W(\rho) + W_\delta(\rho)$, with eigenvalue λ. The parameter α in $W_\delta(\rho)$ can be fixed using Eq. 16. We approximate the left-hand side by $F(\lambda)/(2h)$, Eq. 15, so $\alpha = F(\lambda)/(2h\tilde{\zeta}_\lambda(\rho_M))$. If we have $\tilde{\zeta}_\lambda(\rho_M) = 1$ by construction of Eq. 13, then this factor drops out (which is a sensible thing to do), but to keep it general I let it stand.

According to standard quantum mechanics we can write for the eigenvalue $\lambda = \lambda_W + \lambda_\delta$, with

$$\lambda_W = A\int_0^\infty \tilde{\zeta}_\lambda(\rho)^* \left\{W(\rho) - \frac{d^2}{d\rho^2}\right\} \tilde{\zeta}_\lambda(\rho)d\rho \text{ and } \lambda_\delta = A\int_0^\infty W_\delta(\rho)\left|\tilde{\zeta}_\lambda(\rho)\right|^2 d\rho, \tag{17}$$

$$\text{with } \; A = 1/\int_0^\infty \left|\tilde{\zeta}_\lambda(\rho)\right|^2 d\rho. \tag{18}$$

If the wave function $\tilde{\zeta}_\lambda(\rho)$ is not too crazy, then λ_W should be a decent approximation to the eigenvalue belonging to potential $W(\rho)$, at least a better one than λ. We don't have to calculate the integral explicitly. Just write $\lambda_W = \lambda - \lambda_\delta$, and calculate

$$\lambda_\delta = A\frac{F(\lambda)}{2h}\int_0^\infty \delta(\rho - \rho_M)\left|\tilde{\zeta}_\lambda(\rho)\right|^2 d\rho \approx A\frac{F(\lambda)}{2h\tilde{\zeta}_\lambda(\rho_M)}\left|\tilde{\zeta}_\lambda(\rho_M)\right|^2 = \frac{A}{2h}F(\lambda)\tilde{\zeta}_\lambda(\rho_M).$$

This gives the following algorithm for updating the eigenvalue

$$\text{set } \lambda^{(0)}; \;\; \text{do } n = 1, \ldots \;\; \lambda^{(n)} = \lambda^{(n-1)} - \frac{A}{2h}F\left(\lambda^{(n-1)}\right)\tilde{\zeta}_\lambda(\rho_M) \text{ until converged;} \tag{19}$$

where "until converged" is a criterion of the type $\left|AF\left(\lambda^{(n)}\right)\right| < \epsilon$, with $\epsilon > 0$ a user defined tolerance.

This scheme turns out to be remarkably robust in finding an eigenvalue. It might not be the eigenvalue you want, if your starting guess $\lambda^{(0)}$ is too crazy. In that case you have to play around a bit with your starting guess. In rare cases one might have trouble with convergence. Damping may then help, i.e. multiply F in Eq. 19 with a damping factor β, where $0 < \beta \leq 1$.

1.5 Improving finite differencing: Numerov's method

The finite difference approximation leading up to Eq. 10 is about the simplest one can make, and it would be naive to expect that it is the most accurate one. It would be nice if one could increase the accuracy, without making the algorithm much more complicated. This is possible for problems derived from the Schrödinger equation, thanks to a clever trick discovered by Numerov.[10]

Start from the Taylor expansion

$$f(x \pm h) = f(x) \pm h\frac{df}{dx}(x) + \frac{1}{2!}h^2\frac{d^2f}{dx^2}(x) \pm \frac{1}{3!}h^3\frac{d^3f}{dx^3}(x) + \frac{1}{4!}h^4\frac{d^4f}{dx^4}(x) + ... \tag{20}$$

It is easy to see that approximating the second derivative as in Eq. 10, one makes a discretization error $O(h^4)$. To prove this, just write

$$f(x+h) - 2f(x) + f(x-h) = h^2\frac{d^2f}{dx^2}(x) + \frac{1}{12}h^4\frac{d^4f}{dx^4}(x) + O(h^6). \tag{21}$$

When propagating according to Eqs. 11 or 12, the leading term of what one neglects is $O(h^4)$. In principle, one can increase the order of the error, using the Runge-Kutta method, for instance. This however has the disadvantage that the algorithm becomes more complicated, and one needs evaluations of the function at more points than just the grid points.

There is an alternative that works for differential equations of the type

$$\frac{d^2f}{dx^2}(x) = g(x)f(x). \tag{22}$$

The alternative is called Numerov's method. It is essentially a clever trick to get rid of the h^4 term in Eq. 21. One can write

$$h^2\frac{d^4f}{dx^4}(x) = h^2\frac{d^2}{dx^2}\left[g(x)f(x)\right] = g(x+h)f(x+h) - 2f(x)g(x) + g(x-h)f(x-h) + O(h^4), \tag{23}$$

according to Eqs. 22 and 21. Using Eq. 23 in Eq. 21, and reshuffling a bit, one obtains

$$h^2\frac{d^2f}{dx^2}(x) = f(x+h)q(x+h) - 2f(x)q(x) + f(x-h)q(x-h) + O(h^6) \tag{24}$$
$$\text{with} \quad q(x) = 1 - \frac{h^2}{12}g(x).$$

Using this to modify the propagation algorithms, Eqs. 11 and 12, one has increased the order of the leading error term from $O(h^4)$ to $O(h^6)$, while keeping the same grid and the same number of function evaluations. Is this elegant or not?

1.6 Improving the grid: a logarithmic grid

Using an equidistant grid $\rho_j = jh;\ j = 0, 1, ..., N$ for solving a differential equation is not always optimal. For many potentials in the radial Schrödinger equation, the resulting wave

[10] *B. V. Numerov, Monthly Notices Royal Astronomical Society 84 (1924) pp 592-601.* Note the year; in those days "computers" were actual people sitting in a room, performing computations.

functions have quite a long exponential tail at large ρ. In principle, one only needs a coarse grid to describe such a simple behavior. Using a fine grid is a waste of your time.

However, at smaller ρ the wave functions are more diverse; higher quantum numbers give oscillating functions, for instance. One needs a relatively fine grid to describe such a behavior. In other words, an optimal grid is fine in a region of space where the wave function varies a lot, but coarse in a region where it does not. Algorithms for non-uniform grids are more complicated than the simple finite-difference scheme for equidistant grids we have used so far. However, a certain type of non-uniform grid can be transformed to an equidistant one by a coordinate transformation, allowing us to keep using a simple finite-difference scheme.

Define a function $u(\rho)$, then $du = (du/d\rho)d\rho$, meaning that, if we use $h = du$ as spacing of an equidistant grid in u, then the grid spacing in ρ is given by $d\rho = h(d\rho/du)$. Clearly this grid spacing is non-uniform, if $u(\rho)$ is a non-linear function.

A useful function in our case is

$$u(\rho) = \ln(\rho + \tau_1)\,;\ \text{a grid}\ u_j = a + jh;\ j = 0, 1, ..., N\ \text{then gives}\ \rho_j = \tau_0 d^j - \tau_1;\ j = 1, 2, ..., N, \tag{25}$$

where $\tau_0 = \exp a$, $d = \exp h$ and $0 \leq \tau_1 \leq \tau_0$. Such a grid is called a *(shifted) logarithmic grid*.

The parameters $a$and h define an equidistant grid in u. They determine the spacing between the grid points in ρ. The logarithmic grid does what we want for our case; the spacing between subsequent grid points j and $j + 1$ is small for small ρ_j, and large for large ρ_j. As an extra flexibility, the parameter τ_1 can be used to shift the grid. It can be particularly useful to have a fine control over the point ρ_0 where the grid starts, and the density of grid points there.

As the grid in the new variable u is equidistant, it makes sense to transform the differential equation, Eq. 4 to this new variable. Starting with the second derivative

$$\frac{d^2\zeta}{du^2} = (\rho + \tau_1)^2 \frac{d^2\zeta}{d\rho^2} + (\rho + \tau_1)\frac{d\zeta}{d\rho}, \tag{26}$$

one observes that a first derivative is introduced besides a second derivative. This would hinder the numerical algorithm. For instance, Numerov's method cannot be applied to the differential operator of Eq. 26. So, we would prefer to have an expression without first derivatives. Defining a function $\eta = (\rho + \tau_1)^\alpha \zeta$, we have

$$\begin{aligned}\frac{d^2\eta}{du^2} &= (\rho + \tau_1)^\alpha \frac{d^2\zeta}{du^2} + 2\alpha(\rho + \tau_1)^\alpha \frac{d\zeta}{du} + \alpha^2(\rho + \tau_1)^\alpha \zeta. \\ &= (\rho + \tau_1)^{\alpha+2} \frac{d^2\zeta}{d\rho^2} + (\rho + \tau_1)^{\alpha+1} \frac{d\zeta}{d\rho} + 2\alpha(\rho + \tau_1)^{\alpha+1} \frac{d\zeta}{d\rho} + \alpha^2(\rho + \tau_1)^\alpha \zeta,\end{aligned} \tag{27}$$

where we see that the first derivative drops out if we choose $\alpha = -\frac{1}{2}$. Using this value and the expression of Eq. 4, one obtains

$$\frac{d^2\eta(u)}{du^2} = \left\{\left[X(u) - \lambda + \frac{l(l+1)}{(e^u - \tau_1)^2}\right] e^{2u} + \frac{1}{4}\right\}\eta(u) \quad \text{with} \quad X(u) = \frac{V(\rho(u))}{V_0}. \tag{28}$$

This equation lends itself to a numerical solution in the way explained in the previous sections. It looks more complicated than it is. Given the equidistant grid in u, Eq. 25, one

uses the finite-difference approximation of Eq. 10 and proceeds from there. Eq. 25 gives for each grid point u_j the corresponding point ρ_j, which one can use in calculating the terms on the right-hand side of Eq. 28. Having obtained the function $\eta(u)$, one can always reconstruct the original function $\zeta(\rho)$ by

$$\zeta(\rho) = \eta(u(\rho))\sqrt{\rho + \tau_1} \tag{29}$$

Most of the things we have introduced before stay valid: Numerov's method to solve the differential equation, and finding the eigenvalues by an improved search. There are some details one has to mind. however. For instance, the proper normalization, Eq. 18, is

$$\int_0^\infty |\zeta(\rho)|^2 \, d\rho = \int_0^\infty |\eta(u)|^2 \, e^{2u} du \tag{30}$$

2 Experimental background

2.1 Wannier excitons

A statement that is often found in textbooks on optics or solid state physics is that a semiconductor or insulator only absorbs light with a frequency $\omega \geq \omega_g$, where $E_g = \hbar\omega_g$ is the fundamental band gap of the material. You should not believe all you read in books. The absorption spectrum of a semiconductor shows some remarkable structure for frequencies below the fundamental band gap.[11] One typically observes several sharp peaks for $\omega < \omega_g$. These are called *exciton* peaks, see figures 1 and 2. The absorption spectrum generally becomes continuous for $\omega > \omega_g$, although its shape is different from what one would expect for free electron-hole excitations.

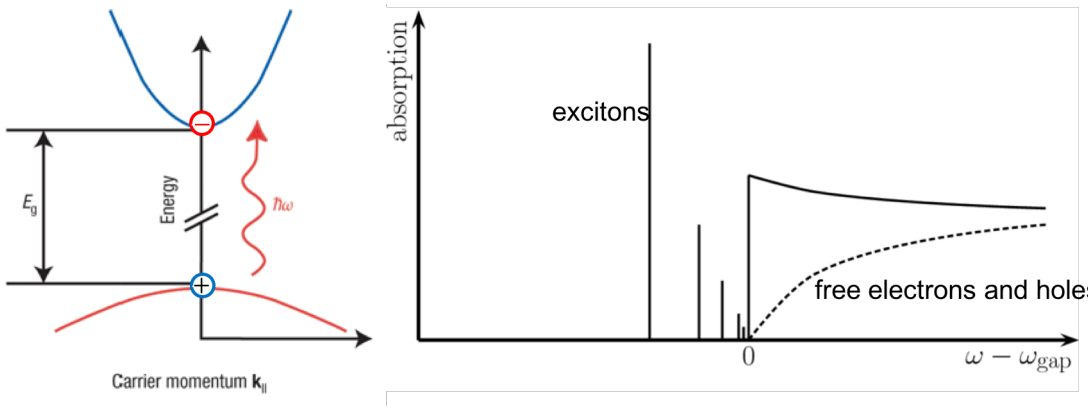

Figure 1: *Left:* schematic band structure of a semiconductor with a direct band gap $E_g = \hbar\omega_g$. Photons of energy $\hbar\omega$ excite electrons across the band gap, leaving electrons in the conduction band, and holes in the valence band. *Right:* schematic picture of the optical absorption of a semiconductor. The dotted line (free electrons and holes) represents the absorption resulting from excitations between the two bands on the left. The solid lines represent the absorption spectrum as it is observed. The isolated peaks at frequencies $\omega < \omega_g$ are called *exciton* peaks. The spacing between these peaks decreases as the frequency increases. For $\omega > \omega_g$ they merge into a continuum. *Figure from: Z.-H. Yang, Y. Li, and C. A. Ullrich, J. Chem Phys. 137, 014513 (2012).*

What is going on can be understood qualitatively using the following picture. Absorbing a photon excites an electron from the valence band of the semiconductor to the conduction band, leaving a hole in the former, and an electron in the latter. One might assume that these are then free to move through the crystal (free electrons and holes), but this is not completely true. Electrons and holes are negatively and positively charged particles, respectively, which means they attract one another. The attraction between an electron and a hole can result in a bound electron-hole pair, whose energy is obviously lower than that of a free electron and hole. Such bound pairs are called *excitons*. Excitons can be created directly by optical absorption. $\hbar\omega_g$ is the photon energy with which a free electron and hole are created, see figure 1. Likewise, $\hbar\omega_e$ is the photon energy with which an exciton is created. The binding energy between the electron and hole in the exciton is then $E_x = \hbar\omega_g - \hbar\omega_e > 0$. This is also

[11] Or the spectrum of an insulator. I use "semiconductor" to indicate any material with a band gap.

called the *exciton binding energy*. As in quantum mechanics binding energies are quantized, excitons give sharp discrete peaks in a spectrum.

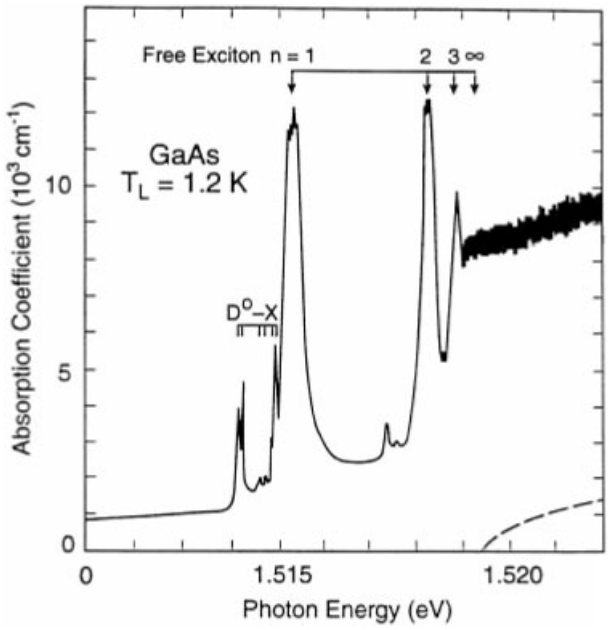

Figure 2: Absorption spectrum of GaAs taken at low temperature ($T = 1.2$K). The peaks marked n are exciton peaks. The peaks marked D^0-X are due to electrons bound to donor impurities in the material, which will not be discussed here. *Figure from: C. Weisbuch and H. Benisty, phys. stat. sol. (b) 424, 2345 (2005).*

GaAs has been one of the first semiconductors widely applied in opto-electronics (LEDs, solid state lasers). Looking at its absorption spectrum, figure 2, you observe that the exciton binding energy is small, $E_b \lesssim 5$ meV. This energy is much smaller than k_BT at room temperature (26 meV), so in GaAs excitons tend to be broken up into separated electrons and holes at room temperature, and dissappear from the absorption spectrum. In that sense the textbooks are right. However, whereas the exciton binding energies in GaAs are tiny, this is not the case in all materials. For instance, the exciton spectrum of the semiconductor Cu_2O has been the subject of intensive research. Many exciton lines can be observed and some of the exciton wave functions have a giant extension, see Fig. 3. Our job is to calculate the exciton binding energies. Predicting exciton binding energies in general requires complicated qunatum mechanical calculations, but in a certain limit it becomes much easier. Lucky for us, this limit is highly relevant for our case.

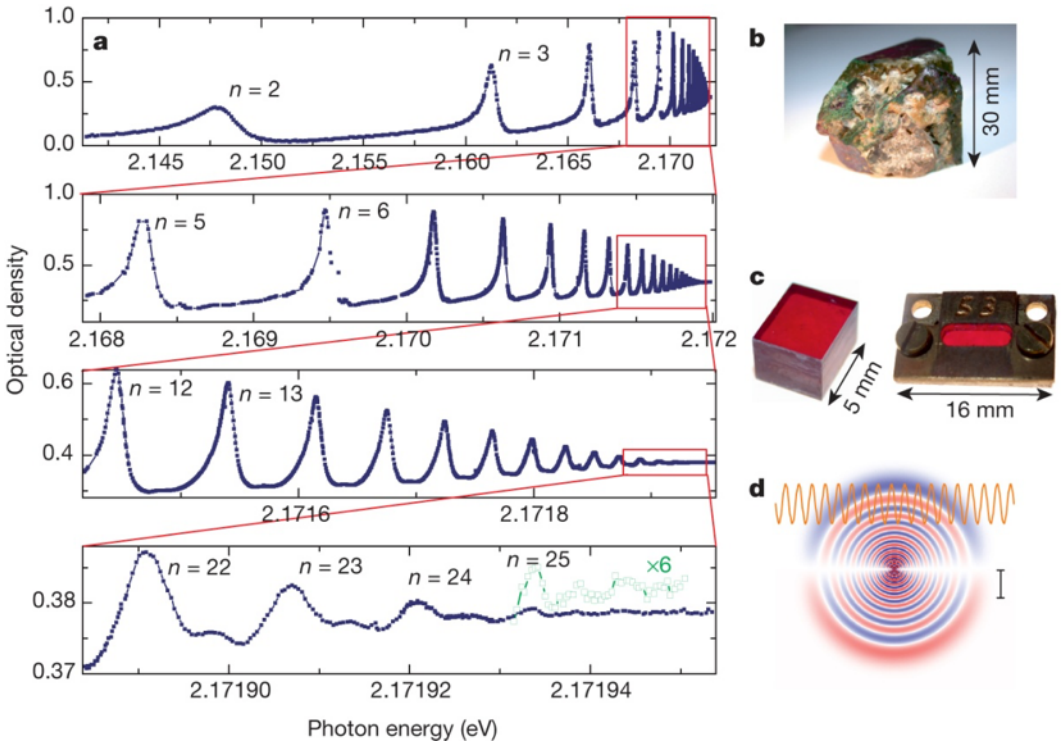

Figure 3: High-resolution absorption spectra of yellow P ($l = 1$) excitons in Cu_2O. **a** Spectra are measured with a single-frequency laser on a natural sample of thickness 34 mm at 1.2 K. Peaks correspond to exciton states with different principal quantum number n. The panels below show close-ups of the areas marked by rectangles in each panel above. **b** Photograph of the natural Cu_2O crystal from which samples of different size and crystal orientation were cut. **c** A large crystal and a thin crystal mounted strain-free in a brass holder. **d** Wavefunction of the P exciton with $n = 25$. To visualize the giant extension, the corresponding light wavelength is shown as the period of the sine function. The bar corresponds to the extension of 10^3 lattice constants.*Figure from: T. Kazimierczuk, D. Frölich, S. Scheel, H. Stolz and M. Bayer, Nature 514, 343-347 (2014).*

If the average distance between an electron and a hole in a bound state is much larger that the spacing between nearest-neighbor atoms in the crystal lattice, one can assume that the details of that lattice do not play a great role. The electron can be represented as a (quasi-)electron with an effective mass m_-, and the hole by a (quasi-)hole with an effective mass m_+. These effective masses are given by $m_-^{-1} = \frac{1}{\hbar^2}\frac{d^2E_C}{dk^2}$ and $m_+^{-1} = -\frac{1}{\hbar^2}\frac{d^2E_V}{dk^2}$, with E_C, E_V the conduction and valence band energies, respectively, assuming a direct band gap and isotropic, parabolic bands, see your solid state physics textbook and figure 1. This is the model we are going to use. It was formulated by Wannier, and the resulting excitons are called *Wannier excitons.*[12]

We first have to derive an equation that the Wannier exciton obeys. I will do that in the next section. Solving that equation is where computational physics comes in.

[12] *G. H. Wannier, Phys. Rev. 52, 191 (1937).* They are called *Wannier-Mott* excitons in some of the literature. I don't know who was first, or whether the English are trying to be chauvinistic by adding an Englishman. Can't be, an Englishman is never chauvinistic (well, hardly ever).

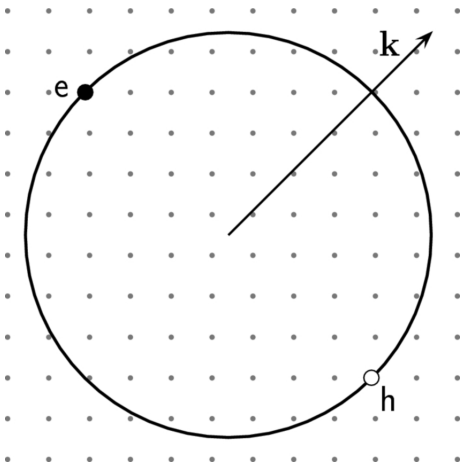

Figure 4: Classical picture of a Wannier exciton. The electron and hole are orbiting around one another at an average distance that is much larger than the lattice spacing. The electron-hole pair as a whole can have a center-of-mass motion with momentum $\mathbf{P} = \hbar\mathbf{k}$, see the next section. *Figure from: Z.-H. Yang, Y. Li, and C. A. Ullrich J. Chem Phys. 137, 014513 (2012).*

3 Theoretical background

3.1 Schrödinger equation for Wannier excitons

The states of an electron and a hole are described by a two-particle Schrödinger equation

$$\left\{-\frac{\hbar^2}{2m_-}\nabla_-^2 - \frac{\hbar^2}{2m_+}\nabla_+^2 + V(\mathbf{r}_-, \mathbf{r}_+)\right\}\Psi(\mathbf{r}_-, \mathbf{r}_+) = E\Psi(\mathbf{r}_-, \mathbf{r}_+), \tag{31}$$

where $\mathbf{r}_-$, m_- are the position and mass of the electron, $\mathbf{r}_+$, m_+ are the position and mass of the hole, $\nabla_-^2 = \frac{\partial^2}{\partial x_-^2} + \frac{\partial^2}{\partial y_-^2} + \frac{\partial^2}{\partial z_-^2}$ and $\nabla_+^2 = \frac{\partial^2}{\partial x_+^2} + \frac{\partial^2}{\partial y_+^2} + \frac{\partial^2}{\partial z_+^2}$. The potential $V(\mathbf{r}_-, \mathbf{r}_+)$ describes the interaction between the electron and the hole. Equation 31 represents a six-dimensional partial differential equation. Solving this requires a big computer, so we do what physicists usually do when they encounter a problem that is too big: cut it down, i.e., simplify and approximate, until we arrive at an equation we can solve.

Let us assume that the potential only depends on the relative position, $\mathbf{r} = \mathbf{r}_- - \mathbf{r}_+$, of the electron and hole, and not on the individual positions of electron and hole separately, in other words $V(\mathbf{r}_-, \mathbf{r}_+) = V(\mathbf{r})$. We define the set of coordinates

$$\mathbf{r} = \mathbf{r}_- - \mathbf{r}_+; \quad \mathbf{R} = \frac{m_-\mathbf{r}_- + m_+\mathbf{r}_+}{m_- + m_+}, \tag{32}$$

where $\mathbf{R}$ can be recognized as the position of the center of mass of the electron-hole pair. We now make a coordinate transformation from the coordinates $\mathbf{r}_-, \mathbf{r}_+$ to $\mathbf{r}, \mathbf{R}$. You can do it yourself; just apply the chain rule for derivatives in Eq. 31, using the inverse relations of Eq. 32

$$\mathbf{r}_- = \mathbf{R} + \frac{m_+}{m_- + m_+}\mathbf{r}; \quad \mathbf{r}_+ = \mathbf{R} - \frac{m_-}{m_- + m_+}\mathbf{r}. \tag{33}$$

One obtains the Schrödinger equation

$$\left\{-\frac{\hbar^2}{2M}\nabla_R^2 - \frac{\hbar^2}{2\mu}\nabla_r^2 + V(\mathbf{r})\right\}\Psi(\mathbf{R}, \mathbf{r}) = E\Psi(\mathbf{R}, \mathbf{r}), \tag{34}$$

with $M = m_- + m_+$ the total mass of the electron-hole pair, and

$$\mu = \frac{m_- m_+}{m_- + m_+} \quad \text{or} \quad \frac{1}{\mu} = \frac{1}{m_-} + \frac{1}{m_+}, \tag{35}$$

a parameter that is called the *reduced mass*. Note that the name is well-chosen, as $\mu < m_-, m_+$.

This might not look like much of a simplification, as Eq. 34 is still a six-dimensional partial differential equation. However the equation now lends itself to the separation-of-variables technique.[13] We try a solution of the form $\Psi(\mathbf{R}, \mathbf{r}) = \phi(\mathbf{R})\psi(\mathbf{r})$ and solve the two equations

$$\begin{cases} -\frac{\hbar^2}{2M}\nabla_R^2 \phi(\mathbf{R}) & = E_P \phi(\mathbf{R}) \\ \left\{ -\frac{\hbar^2}{2\mu}\nabla_r^2 + V(\mathbf{r}) \right\} \psi(\mathbf{r}) & = E_x \psi(\mathbf{r}) \end{cases} \tag{36}$$

separately. Mathematics tells us that, if we solve these two equations, then the product $\phi(\mathbf{R})\psi(\mathbf{r})$ is a solution of Eq. 34, with eigenvalue $E = E_P + E_x$. The first of these two equations is trivial to solve. It does not contain a potential, so it is the Schrödinger equation of a free particle with solutions

$$\phi(\mathbf{R}) = A \exp\left[\frac{i}{\hbar M} \mathbf{P} \cdot \mathbf{R} \right]; \quad E_P = \frac{P^2}{2M}, \tag{37}$$

where $\mathbf{P}$ is the total momentum of the electron-hole pair. It is easy to prove that $\hat{\mathbf{P}} = \hat{\mathbf{p}}_- + \hat{\mathbf{p}}_+$, with $\hat{\mathbf{P}} = -\frac{\hbar}{i}\nabla_R$, $\hat{\mathbf{p}}_- = -\frac{\hbar}{i}\nabla_-$, and $\hat{\mathbf{p}}_+ = -\frac{\hbar}{i}\nabla_+$ the momentum operators for the total momentum, and the momenta of the electron and the hole, respectively. Apparently the electron-hole pair moves through space like a free particle with total mass M and total momentum $\mathbf{P}$.

Optical excitations create excitons in the zero-momentum state, $\mathbf{P} = 0$. This is because, like any other physical process, optical excitation has to conserve the total momentum. Before excitation the total momentum of the system is zero, i.e., for each electron with (crystal) momentum $\hbar\mathbf{k}$, there is also an electron with momentum $-\hbar\mathbf{k}$, so after excitation, when we have the exciton, the total momentum also must be zero.[14]

3.2 The radial Schrödinger equation

We are interested in the relative motion of the electron and the hole, as described by the second equation in Eq. 36. This three-dimensional partial differential equation is still not easy to solve, so we have to make a further approximation. We assume that the potential only depends on the relative distance, $r = |\mathbf{r}|$, between electron and hole, i.e., $V(\mathbf{r}) = V(r)$. The equation then becomes

$$\left\{ -\frac{\hbar^2}{2\mu}\nabla^2 + V(r) \right\} \psi(\mathbf{r}) = E_x \psi(\mathbf{r}), \tag{38}$$

[13] Physicists are what the Americans call *one-trick ponies*. Separation of variables is the one trick they know to solve partial differential equations.

[14] I am cheating a bit here. In principle, the photon that is absorbed, has a momentum, and this momentum is given to the exciton. However, on the scale of an electronic Brillouin zone, the momentum $\hbar\mathbf{K}$ of a photon is tiny, and it is safe to neglect it.

where I have dropped the subscript r on the ∇; from now on this is the only ∇ we will use. This looks like an equation we know: the Schrödinger equation of a particle in a central potential. We know how to solve this.[15] Switch to spherical coordinates

$$\nabla^2 = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial}{\partial r}\right) + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right) + \frac{1}{r^2\sin^2\theta}\frac{\partial^2}{\partial\varphi^2}. \tag{39}$$

Smart 19th century physicists and mathematicians have solved the following eigenvalue equation

$$\left\{\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right) + \frac{1}{\sin^2\theta}\frac{\partial^2}{\partial\varphi^2}\right\} Y_m^l(\theta,\varphi) = -l(l+1)Y_m^l(\theta,\varphi), \tag{40}$$

and they have shown that this equation only has physically well-behaved solutions if $l = 0, 1, 2, ..,$ and $m = -l, -l+1, ..., l$. The solutions $Y_m^l(\theta,\varphi)$ are called *spherical harmonics*. In atomic physics the functions for $l = 0, 1, 2, 3$ are known as s, p, d, f functions, respectively.

We use this knowledge to try a separation-of-variables solution $\psi(r,\theta,\varphi) = \chi(r)Y_m^l(\theta,\varphi)$ to Eq. 38. Using Eqs. 39 and 40 we then arrive at an equation for $\chi(r)$

$$\left\{-\frac{\hbar^2}{2\mu}\left[\frac{d^2}{dr^2} + \frac{2}{r}\frac{d}{dr} - \frac{l(l+1)}{r^2}\right] + V(r)\right\}\chi(r) = E_x\chi(r), \tag{41}$$

where the partial derivative has become a normal derivative, as there is only one coordinate left. The combination of first- and second derivatives in Eq. 41 is a bit of a nuisance.[16]

There is a standard trick to get rid of the first derivative. Define a function $\zeta(r) = r^\alpha\chi(r)$. Then

$$\frac{d^2\zeta}{dr^2} = r^\alpha\frac{d^2\chi}{dr^2} + 2\alpha r^{\alpha-1}\frac{d\chi}{dr} + \alpha(\alpha-1)r^{\alpha-2}\chi. \tag{42}$$

Choosing $\alpha = 1$, one can then derive from Eq. 41

$$\left\{-\frac{\hbar^2}{2\mu}\frac{d^2}{dr^2} + \frac{\hbar^2 l(l+1)}{2\mu r^2} + V(r)\right\}\zeta(r) = E_x\zeta(r), \tag{43}$$

which looks like a one-dimensional Schrödinger equation with an effective potential

$$V_{\text{eff}}(r) = \frac{\hbar^2 l(l+1)}{2\mu r^2} + V(r). \tag{44}$$

Of course the domain is restricted to $r \in [0,\infty)$. Note that $\lim_{r\to 0}\zeta(r) = 0$ if $\chi(r)$ is a well-behaved, i.e., bounded, function, which it should be as we would like to to be able to construct probability distributions from these wave functions.

The simplicity of Eq. 43 is slightly deceptive. It is impossible to solve it analytically for a general potential $V(r)$. Only for special potentials such as $V(r) \propto r^{-1}$ (as in the Hydrogen atom), or $V(r) \propto r^2$ (as in the harmonic oscillator) we can solve the problem using pen and paper only. Before you start moaning about quantum mechanics, the situation in classical mechanics is no different. Also Newton's equations can only be solved for special potentials, often the same ones as in quantum mechanics. For instance, the solution for $V(r) \propto r^{-1}$ gives you Keppler's laws for planetary motion. As Newton has already found that solution, you may wonder how much progress we have actually made in three centuries. Not much,

[15] Physicists are also hoarders: they never throw a solution away. You never know how it might be useful later on.

[16] For instance, Numerov's method cannot be applied to this.

but a little; the computer has been invented. That allows you to solve Eq. 43 numerically for any potential $V(r)$.[17]

3.3 The interaction potential: Coulomb and Haken potential

In a first effort to specify the potential, we assume is that the electron and hole in a Wannier exciton attract one another via a screened Coulomb interaction

$$V(r) = -\frac{e^2}{4\pi\epsilon}\frac{1}{r}, \tag{45}$$

where ϵ is the dielectric permittivity of the material. This makes Eq. 43 very similar to the Schrödinger equation of a hydrogen atom, for which we know the analytical solution. The eigenvalues are given by the expression[18]

$$E_{x,n} = -\frac{R_x}{n^2}; \quad n = l, l+1, \ldots \text{ with } R_x = \frac{\mu e^4}{32\pi^2\epsilon^2\hbar^2}, \tag{46}$$

a constant, called the *Rydberg constant*. In the expression for the hydrogen atom, μ is simply the electron mass m in vacuum, and ϵ is ϵ_0, the permittivity of vacuum. The Rydberg constant for the hydrogen atom is $R_H \approx 13.606$ eV. The Rydberg constant for the exciton can be expressed as

$$R_x = R_H \left(\mu\epsilon_0^2\right) / \left(m\epsilon^2\right). \tag{47}$$

The electron and hole effective masses in GaAs are $m_- \approx 0.067m$, respectively $m_+ \approx 0.45m$, which gives a reduced mass of $\mu = 0.058m$, see Eq. 35. The static dielectric permittivity of GaAs is $\epsilon \approx 13.1\epsilon_0$. From Eq. 47 one then obtains $R_x \approx 4.6$ meV. The energy spectrum given by Eq. 46 then agrees reasonable well with what is observed in experiment, see Fig. 2.

From the solutions of the Schrödinger equation of the hydrogen atom, we also know that the extend of the wave function is measured in units of the Bohr radius a. In particular, the most probable value of the radius in the ground state is

$$r_{1S} = (4\pi\epsilon\hbar^2)/(\mu e^2) = a. \tag{48}$$

For the hydrogen atom we have a Bohr radius $a_0 \approx 0.0529177$ nm. For our excition in GaAs we have

$$a_x = a_0(m\epsilon)/(\mu\epsilon_0), \tag{49}$$

or $a_x \approx 12.0$ nm, which is gigantic. Such a large Bohr radius justifies our use of the Wannier model, figure 4 and Eqs. 31 and 45.[19]

The small binding energy and the large Bohr radius of the exciton is the result of two factors. Firstly, the effective mass μ is considerably smaller than the electron mass in vacuum, which makes the (positive) kinetic energy (the second derivative term in Eq. 43) relatively large compared to the (negative) potential energy. Secondly, the dielectric permittivity of

[17] A mathematician would object, as for sure he can come up with a potential for which it is impossible to find a solution. Physicists call such potentials "unreasonable". Reasonable people avoid those.

[18] See your favorite quantum mechanics book.

[19] Critics may object that this is a Baron von Münchhausen reasoning: you use theory A to demonstrate that theory A is justfified. The critics are right; all we have shown is that the model is consistent with itself. Believe it or not, it could be worse. But, then again, physics is not mathematics, and we always have experiment to tell us whether a theory is nonsense or not.

GaAs is much larger than that of vacuum, which weakens the potential, Eq. 45. The physics behind the latter is that GaAs is a polarizable medium. The material between the electron and the hole is polarized by the fields of these particles and the polarization screens these fields.

Although Eq. 46 gives a decent approximation to observed exciton spectra of many semiconductors, it is not perfect. If one examines the spectra more closely, there are many cases where the positions of the observed exciton lines deviate somewhat from the positions predicted by Eq. 46. Even if we assume that Eq. 45 represents the correct form for the electron-hole interaction, we now that the permittivity $\epsilon(\omega)$ is a function of the frequency, and not a simple function either. A good example is (liquid) water, where the static dielectric permittivity $\epsilon(0) \approx 80\epsilon_0$, whereas at frequencies corresponding to visible light $\epsilon(\omega) \approx 1.8\epsilon_0$. So, what frequency should we take?

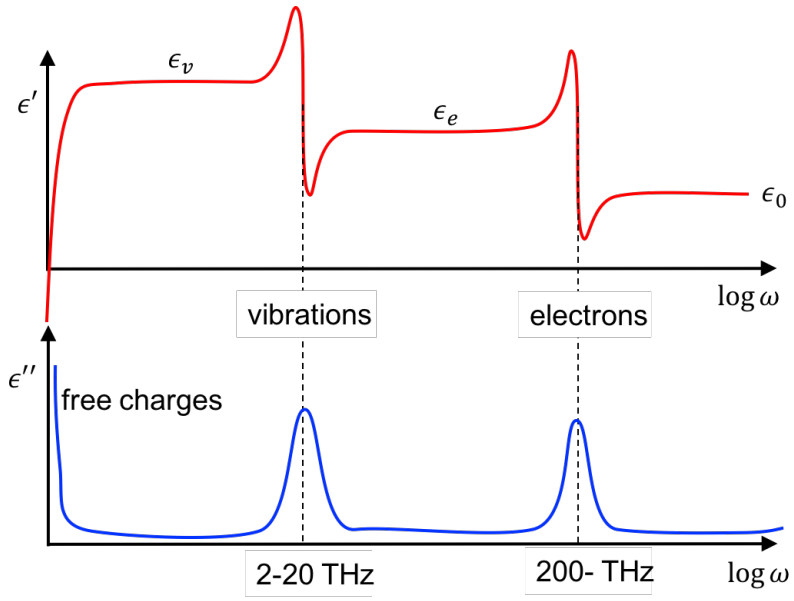

Figure 5: Schematic permittivity $\epsilon = \epsilon' + i\epsilon''$ of a solid. The real part $\epsilon'(\omega)$ consists of resonances separated by plateaus. Each resonance corresponds to the response of an oscillator. The oscillators are marked by "vibrations" (the lattice vibrations in a solid, i.e., the phonons, with typical resonance frequencies $\omega \sim$ 2-20 THz, or $\hbar\omega \sim 10^{-2}$-10^{-1} eV), and "electronic" (the oscillations associated with electronic excitations, with resonance frequencies $\omega > 10^{15}$ Hz, or $\hbar\omega > 1$ eV, assuming a corresponding band gap). The imaginary part $\epsilon''(\omega)$ shows peaks at the resonance frequencies, and is small between the resonances.

Figure 5 shows schematically the permittivity as a function of frequency. It varies wildly around resonances, but between resonances the behaviour is much simpler: the real part $\epsilon'(\omega)$ is approximately constant, and the imaginary part $\epsilon''(\omega)$ is approximately zero. As long as we stay away from the resonance frequencies we can use a real permittivity, $\epsilon(\omega) \approx \epsilon'(\omega)$. The level of the plateaus in $\epsilon'(\omega)$ is determined by which oscillator degrees of freedom participate. At very low frequency all degrees of freedom participate.

If the external driving field has a frequency above a few GHz, it is too fast for the free charge carriers to follow, and they are effectively frozen out. Only the vibrational and the electronic oscillators then participate, leading to an approximately constant $\epsilon'(\omega) \approx \epsilon_v$.

Frequencies above 2-20 THz are too high for the lattice vibrations to follow, and at high frequencies only the electronic oscillators participate, which gives $\epsilon'(\omega) \approx \epsilon_e$. The fewer oscillators participate in the response, the smaller ϵ, so $\epsilon_e < \epsilon_v$. Finally, at very high frequencies (X rays, for instance), even the electronic degrees of freedom cannot follow. Nothing oscillates anymore, and $\epsilon'(\omega) \approx \epsilon_0$.

The characteristic frequency of an exciton can be extracted from Eq. 46 by $\omega_{x,n} = |E_{x,n}|/\hbar$. For GaAs $\omega_{x,1} \approx 1.1$ THz, which is below the resonance frequencies of most of the vibrational oscillators. All other states have even lower frequencies, $\omega_{x,n>1} < \omega_{x,1}$, so one may expect that both the vibrational, as well as the electronic oscillators participate at these frequencies. In other words, one can use $\epsilon \approx \epsilon_v$ in Eq. 45, which is what we did above.[20]

We are in trouble for semiconductors where the Rydberg constant is much larger. For Cu_2O, for instance, the lowest exciton level has an energy $E_{x,1} \approx -0.15$ eV, which gives $\omega_{x,1} \approx 36$ THz, which is above the resonance frequencies of the vibrational oscillators. It seems logical therefore to use $\epsilon \approx \epsilon_e$ in Eq. 45. However, for the exciton level with $n = 5$, Eq. 46 gives $E_{x,5} \approx -6$ meV, or $\omega_{x,5} \approx 1.5$ THz. This is below the vibrational resonance frequencies, making $\epsilon \approx \epsilon_v$ a more logical choice. For intermediate exciton levels, $1 < n < 5$, we need some interpolation between ϵ_e and ϵ_v.

Most interpolations in the literature use a range separation. Strongly bound excitons like the $n = 1$ state have a relatively small radius, whereas weakly bound excitons like the $n = 5$ state have a large radius. One can write the potential as

$$V(r) = -\frac{e^2}{4\pi}\frac{1}{r}\left[f(r)\frac{1}{\epsilon_e} + \frac{1}{\epsilon_v}\left(1 - f(r)\right)\right], \tag{50}$$

where $f(r)$ is a switching function $0 \le f(r) \le 1$, with $f(0) = 1$ and $f(\infty) = 0$. A popular one has been derived by Haken[21]

$$f_H(r) = \frac{1}{2}\left[\exp\left(-\frac{r}{r_-}\right) + \exp\left(-\frac{r}{r_+}\right)\right] \quad \text{with} \quad r_\pm = \frac{\hbar}{\sqrt{2m_\pm E_v}}, \tag{51}$$

where $m_\pm$ are the effective masses of the hole and the electron, and E_v is a typical vibration energy.

There is actually some theory behind this. Haken considered the effect of the coupling between the charge carriers (free electron and holes) and phonons (lattice vibrations) in a simplified model. This leads to the formation of so-called *polarons,* i.e., charge carriers "dressed" with phonons. A polaron put in more classical terms (and therefore, of limited correctness) is a charge carrier that deforms the crystal lattice in its neighborhood, and drags this deformation along when it moves. The polaron has a typical size $r_\pm$ (in classical terms, the extend of the lattice deformation). Dielectric response is the response of charged particles in the material to an electric field. For $r > r_\pm$ these charged particles are polarons, in which all degrees of freedom participate (electrons/holes and phonons). On the inside a polaron, i.e., for $r < r_\pm$, it looks like a free charge carrier (the lattice is static, always the same deformation). So for $r < r_\pm$ the charged particles are electrons/holes, and only those degrees of freedom participate to the response (the phonons are "frozen out").

[20] Again, a von Münchhausen trick; one uses a theory to make the same theory plausible.

[21] *H. Haken, Z. Phys. 146, 527 (1956); Fortschr. Phys. 6, 271 (1958).*