Accuracy-control techniques applied to stable transfer-matrix computations

A. Mayer* and J.-P. Vigneron
Laboratoire de Physique du Solide, Facultés Universitaires Notre Dame de la Paix, Rue de Bruxelles 61, B-5000 Namur, Belgium
(Received 17 February 1998; revised manuscript received 9 October 1998)

The transfer-matrix methodology is frequently used to deal with elastic scattering problems that require a solution of the Schrödinger or homogeneous Maxwell equations in the continuous part of their spectra. The numerical stability of the transfer-matrix algorithm can be dramatically improved by a subdivision of the diffusive part of the system into several adjacent layers. However, until now, no accurate recommendation on the number of layers to use was given. This paper presents the transfer-matrix technique and the layer addition algorithm. A model is developed to analyze the accuracy of these techniques and enable a quantitative control. As a result of the model, an expression for the minimum number of layers to consider in order to achieve a given accuracy on the transfer-matrix computation is derived. The theory is illustrated by a simulation of electronic field emission. [S1063-651X(99)02904-9]

PACS number(s): 02.70.−c, 11.80.−m, 61.14.Dc

I. INTRODUCTION

Linear systems of differential equations are frequently encountered in theoretical physics. Such equations indeed appear when dealing with the Schrödinger equation in quantum mechanics or with the Maxwell equations in electromagnetism. A useful property that appears in these situations is the additivity of solutions. When an analytic solution is not obtainable, several numerical techniques exist to deal with these equations in the energy or frequency continuum.

The transfer-matrix methodology [1–6] is one of these techniques. To apply this methodology, the physical system considered should be located between two separate boundaries. Given a set of basic states used for the wave function expansion, the transfer matrices contain, for each state incident on one boundary of the system, the amplitudes of the corresponding transmitted and reflected states.

The method basically depends on the additivity property of solutions and requires the numerical propagation of basic states from one boundary to the other. Some of them can have transmission coefficients several orders of magnitude lower than others. All these numbers are gathered in a matrix that has to be inverted. The method reaches its limit when the condition number of such matrices exceeds the representation possibilities of the machine. The layer addition algorithm, introduced by Pendry [7,8] in dynamic a low-energy electron diffraction (LEED) computation, enables the inversion of matrices associated with a smaller part of the total system and therefore better conditioned.

This paper first presents the transfer-matrix technique and the layer addition algorithm in Secs. II and III. This presentation includes the implementation procedure that gives the most precise results. Since a subdivision of the system into several adjacent layers can improve the accuracy of the result but only qualitative indications on how to split the system exist, a model is developed in Sec. IV to analyze the accuracy of the layer addition algorithm. This model predicts the precision of a transfer-matrix computation and is used in Sec. V to determine the minimum number of subdivisions to consider in order to achieve a given accuracy on the result. In Sec. VI, the theory is applied to the simulation of electronic field emission.

II. TRANSFER-MATRIX METHOD

A. Presentation

Let us consider the scattering in a physical system made of three adjacent regions and let us assume the intermediate region to be the only diffusive part. The scanning tunneling microscope [9] and the Fresnel projection microscope [10] provide examples of such situations. Let us refer to the intermediate region as ''region II'' and the two other regions as ''region I'' and ''region III.'' Let z be a coordinate axis oriented from region I to region III, so that region II corresponds to the interval $0 \leqslant z \leqslant D$.

At this point, we should make the choice of simple basic states to represent the waves in regions I and III. They should preferably be the same in both regions, but this is only a matter of convenience. Let us write these states $\Psi_j^{\mathrm{I},\pm}$ in region I and $\Psi_j^{\mathrm{III},\pm}$ in region III. The sign $\pm$ stands for the direction of propagation relative to the z axis.

Let us write $\mathbf{t}_{0,D}^{++}$ and $\mathbf{t}_{0,D}^{-+}$ for the matrices that contain in each column the amplitudes of the respectively transmitted and reflected basic states corresponding to each basic state with unit amplitude injected from $z=-\infty$. Similarly, $\mathbf{t}_{0,D}^{--}$ and $\mathbf{t}_{0,D}^{+-}$ collect the amplitudes of the transmitted and reflected basic states corresponding to each basic state with unit amplitude coming from $z=+\infty$. The subscripts 0 and D stand for the boundaries which limit the diffusive part of the system.

B. Implementation

Let us now turn to the construction of two transfer matrices $\mathbf{t}_{0,D}^{++}$ and $\mathbf{t}_{0,D}^{-+}$. To obtain them, each outgoing state $\Psi_j^{\mathrm{III},+}$ is considered individually and propagated backwards from $z=D$ to $z=0$, by using the relevant propagation equation. The solution is then written as a combination of incident

*Author to whom correspondence should be addressed. Electronic address: alexandre.mayer@fundp.ac.be

states $\Psi_i^{\mathrm{I},+}$ and reflected states $\Psi_i^{\mathrm{I},-}$. The following set of solutions results from these operations:

$$\overline{\Psi}_j^{+} \overset{z\leqslant 0}{=} \sum_i A_{i,j}\Psi_i^{\mathrm{I},+} + \sum_i B_{i,j}\Psi_i^{\mathrm{I},-} \overset{z\geqslant D}{=} \Psi_j^{\mathrm{III},+}. \tag{1}$$

Since the relevant propagation equation is assumed to be linear, these solutions can be combined in order to derive a set of solutions corresponding to a single incident state $\Psi_j^{\mathrm{I},+}$ in region I:

$$\Psi_j^{+} \overset{z\leqslant 0}{=} \Psi_j^{\mathrm{I},+} + \sum_i (t_{0,D}^{-+})_{i,j}\Psi_i^{\mathrm{I},-} \overset{z\geqslant D}{=} \sum_i (t_{0,D}^{++})_{i,j}\Psi_i^{\mathrm{III},+}. \tag{2}$$

The two transfer matrices $\mathbf{t}_{0,D}^{++}$ and $\mathbf{t}_{0,D}^{-+}$ are related to the **A** and **B** matrices of Eq. (1) through $\mathbf{t}_{0,D}^{++}=\mathbf{A}^{-1}$ and $\mathbf{t}_{0,D}^{-+}=\mathbf{B}\ \mathbf{A}^{-1}$.

In the same way, each state $\Psi_j^{\mathrm{I},-}$ can be considered individually and propagated backwards from $z=0$ to $z=D$, where the solution is written as a combination of incident states $\Psi_i^{\mathrm{III},-}$ and reflected states $\Psi_i^{\mathrm{III},+}$. Another set of solutions results:

$$\overline{\Psi}_j^{-} \overset{z\leqslant 0}{=} \Psi_j^{\mathrm{I},-} \overset{z\geqslant D}{=} \sum_i A_{i,j}\Psi_i^{\mathrm{III},-} + \sum_i B_{i,j}\Psi_i^{\mathrm{III},+}. \tag{3}$$

Using again the linearity of the propagation equation, these solutions can be combined to establish a set of solutions corresponding to a single incident state $\Psi_j^{\mathrm{III},-}$ in region III:

$$\Psi_j^{-} \overset{z\leqslant 0}{=} \sum_i (t_{0,D}^{--})_{i,j}\Psi_i^{\mathrm{I},-} \overset{z\geqslant D}{=} \Psi_j^{\mathrm{III},-} + \sum_i (t_{0,D}^{+-})_{i,j}\Psi_i^{\mathrm{III},+}. \tag{4}$$

The two transfer matrices $\mathbf{t}_{0,D}^{--}$ and $\mathbf{t}_{0,D}^{+-}$ are given by the relations: $\mathbf{t}_{0,D}^{--}=\mathbf{A}^{-1}$ and $\mathbf{t}_{0,D}^{+-}=\mathbf{B}\ \mathbf{A}^{-1}$.

C. Advantage of a backwards numerical integration

To highlight the advantage of a backwards integration, let us consider the simple case where $\overline{\Psi}_j^{+}$ takes the form of a pair of exponential solutions in region II:

$$\overline{\Psi}_j(z) \overset{0\leqslant z\leqslant D}{=} M_j e^{K_j z} + N_j e^{-K_j z}, \tag{5}$$

where K_j would take the expression $K_j=\sqrt{(2m/\hbar^2)(V-E)+k_j^2}$ with k_j the transverse component of the wave vector associated with the state $\overline{\Psi}_j^{+}$, in the case of tunneling through a potential barrier with height V by particles with energy $E<V$.

The coefficients are obtained from

$$M_j=\frac{1}{2}\left(\overline{\Psi}_j^{+}(z=D)+\frac{1}{K_j}\frac{d\overline{\Psi}_j^{+}(z=D)}{dz}\right)e^{-K_j D}, \tag{6}$$

$$N_j=\frac{1}{2}\left(\overline{\Psi}_j^{+}(z=D)-\frac{1}{K_j}\frac{d\overline{\Psi}_j^{+}(z=D)}{dz}\right)e^{+K_j D}. \tag{7}$$

Since the states in region III are propagative, $d\overline{\Psi}_j^{+}/dz(z=D)=ik_{z,j}\overline{\Psi}_j^{+}(z=D)$, with $k_{z,j}$ the z component of the wave vector associated with the state $\overline{\Psi}_j^{+}$ in region III. It is easy to check that

$$\frac{|M_j e^{K_j D}|}{|N_j e^{-K_j D}|}=1. \tag{8}$$

This relation means that the two parts of the solution contribute equally to the initial value of $\overline{\Psi}_j^{+}$. By integrating numerically from $z=D$ to $z=0$, the exponentially increasing solution $N_j e^{-K_j z}$ will dominate the exponentially decreasing $M_j e^{K_j z}$. This last part of the solution will consequently vanish from the small number of representative digits stored by the computer. This is, however, acceptable since the resulting solution corresponds to the physical one. This appears clearly if one considers the relation

$$\frac{|M_j e^{K_j 0}|}{|N_j e^{-K_j 0}|}=e^{-2K_j D}. \tag{9}$$

Since the coefficients of the matrices **A** and **B** depend linearly on $\overline{\Psi}_j^{+}(z=0)$ and $d\overline{\Psi}_j^{+}/dz(z=0)$, the ratio between the contribution of the two parts of the solution is given by this last result.

From these arguments, it appears that the integration step (and therefore the matrices **A** and **B**) can achieve an excellent accuracy. The same comments apply to the matrices needed to compute $\mathbf{t}_{0,D}^{--}$ and $\mathbf{t}_{0,D}^{+-}$. However, depending on the peculiar values of K_j, these two matrices are likely to be made of numbers with different orders of magnitude. Anticipating the results of Sec. IV, the condition number of such matrices is expected to be of the order of $e^{K_{\max}D}$, with $K_{\max}=\sqrt{(2m/\hbar^2)V}$ in the case of tunneling through a potential barrier with height V. The accuracy is then drastically reduced in the inversion step and can even be completely lost if D were too large. From there come the limits of the transfer-matrix technique. However, the layer addition algorithm allows us to deal with much larger distances.

III. LAYER ADDITION ALGORITHM

Since problems arise when systems characterized by large integration distances are treated in a single step, a possibility to cope with them is to split this large distance into several adjacent layers. Their number is chosen so that each one is small enough to enable the computation of its individual transfer matrices. The transfer matrices corresponding to the whole system are obtained by combination of the individual transfer matrices. Pendry [7,8] has developed the appropriate formulas:

$$\mathbf{t}_{z_0,z_i}^{++}=\mathbf{t}_{z_{i-1},z_i}^{++}[\mathbf{I}-\mathbf{t}_{z_0,z_{i-1}}^{+-}\mathbf{t}_{z_{i-1},z_i}^{-+}]^{-1}\mathbf{t}_{z_0,z_{i-1}}^{++}, \tag{10}$$

$$\mathbf{t}_{z_0,z_i}^{-+}=\mathbf{t}_{z_0,z_{i-1}}^{-+}+\mathbf{t}_{z_0,z_{i-1}}^{--}\mathbf{t}_{z_{i-1},z_i}^{-+}[\mathbf{I}-\mathbf{t}_{z_0,z_{i-1}}^{+-}\mathbf{t}_{z_{i-1},z_i}^{-+}]^{-1}\mathbf{t}_{z_0,z_{i-1}}^{++}, \tag{11}$$

$$\mathbf{t}^{--}_{z_0,z_i}=\mathbf{t}^{--}_{z_0,z_{i-1}}[\mathbf{I}-\mathbf{t}^{-+}_{z_{i-1},z_i}\mathbf{t}^{+-}_{z_0,z_{i-1}}]^{-1}\mathbf{t}^{--}_{z_{i-1},z_i}, \tag{12}$$

$$\mathbf{t}^{+-}_{z_0,z_i}=\mathbf{t}^{+-}_{z_{i-1},z_i}+\mathbf{t}^{++}_{z_{i-1},z_i}\mathbf{t}^{+-}_{z_0,z_{i-1}}[\mathbf{I}-\mathbf{t}^{-+}_{z_{i-1},z_i}\mathbf{t}^{+-}_{z_0,z_{i-1}}]^{-1}\mathbf{t}^{--}_{z_{i-1},z_i}, \tag{13}$$

where $z_0=0<\cdots<z_{i-1}<z_i<z_{i+1}<\cdots<z_n=D$. This set of formulas is used to compute iteratively the transfer matrices corresponding to the i first layers, as a combination of the transfer matrices associated with the $i-1$ first layers and the layer i. The transfer matrices of each individual slab are computed with the method given in Sec. II B. The only matrices needed at each step i are the four transfer matrices corresponding to the set of the $i-1$ first layers and those corresponding to the last considered layer i.

IV. ACCURACY OF THE TRANSFER-MATRIX COMPUTATION

A. Mathematical analysis of the accuracy

To represent the accuracy of a result, a distinction is made between the true, but unknown, value of a matrix $\mathbf{A}$ and its known approximate representation $\overline{\mathbf{A}}$. Their components can be related by

$$\overline{A}_{i,j}=(1+\delta_{A;i,j})A_{i,j}, \tag{14}$$

where $\delta_{A;i,j}$ stands for the relative error on the true value of $A_{i,j}$. We define the average relative error on the matrix $\mathbf{A}$ by

$$\epsilon_A=\frac{\sum_{i,j}|\delta_{A;i,j}A_{i,j}|}{\sum_{i,j}|A_{i,j}|}. \tag{15}$$

We thus have a single parameter that takes into account the relative importance of each component. The result is considered meaningless if $\epsilon_A>1$.

The best possible accuracy ϵ_{comp} for a computer-stored result depends on the representation limits of the machine and is related to the largest number x whose representation differs from $x+1$. For a binary representation:

$$\epsilon_{\text{comp}}=2^{-n_{\text{bit}}}, \tag{16}$$

where n_{bit} is the number of bits used to represent the fractional part of reals. It is related to the number of bits n_{expt} used to represent the exponent part and the sign of real numbers coded with r bytes by $n_{\text{bit}}=8*r-n_{\text{expt}}$.

When operating with imperfectly represented matrices, one wishes to know the accuracy of the result. Let us consider the effect on accuracy of three common operations: multiplication, addition, and inversion.

1. Multiplication

Let $\mathbf{A}$ and $\mathbf{B}$ be matrices represented, respectively, with an accuracy ϵ_A and ϵ_B. By application of the definition (15), one finds the following expression for ϵ_{AB}:

$$\epsilon_{AB}=\frac{\sum_{i,j}\left|\sum_k(\delta_{A;i,k}+\delta_{B;k,j})A_{i,k}B_{k,j}\right|}{\sum_{i,j}|(AB)_{i,j}|}, \tag{17}$$

which is evaluated simply by

$$\epsilon_{AB}=\epsilon_A+\epsilon_B. \tag{18}$$

2. Addition

By considering the components of $\overline{\mathbf{A}}$ and $\overline{\mathbf{B}}$ individually, one finds

$$\epsilon_{A+B}=\frac{\sum_{i,j}|[(\delta_{A;i,j}A_{i,j}+\delta_{B;i,j}B_{i,j})/(A_{i,j}+B_{i,j})](A+B)_{i,j}|}{\sum_{i,j}|(A+B)_{i,j}|}, \tag{19}$$

which is evaluated by the (twice) weighted average

$$\epsilon_{A+B}=\frac{\sum_{i,j}[(\epsilon_A|\overline{A}_{i,j}|+\epsilon_B|\overline{B}_{i,j}|)/(|\overline{A}_{i,j}|+|\overline{B}_{i,j}|)]|(\overline{A}+\overline{B})_{i,j}|}{\sum_{i,j}|(\overline{A}+\overline{B})_{i,j}|}. \tag{20}$$

It easy to check that if $\epsilon_A=\epsilon_B=\epsilon$, then we have also $\epsilon_{A+B}=\epsilon$. A useful property is

$$\min(\epsilon_A,\epsilon_B)\leqslant\epsilon_{A+B}\leqslant\max(\epsilon_A,\epsilon_B). \tag{21}$$

3. Inversion

When solving the equation $\mathbf{A}\ \mathbf{x}=\mathbf{b}$, the relative errors of all elements of the equation are related [11] by

$$\frac{|\Delta\mathbf{x}|}{|\mathbf{x}|}\leqslant\frac{\text{cond}(A)}{1-\text{cond}(A)\|\delta\mathbf{A}\|/\|\mathbf{A}\|}\left(\frac{|\delta\mathbf{b}|}{|\mathbf{b}|}+\frac{\|\delta\mathbf{A}\|}{\|\mathbf{A}\|}\right), \tag{22}$$

with $\text{cond}(A)$ the condition number of the matrix $\mathbf{A}$, majored by

$$\text{cond}(A)\leqslant\frac{\max|\lambda_i|}{\min|\lambda_i|}, \tag{23}$$

with $\max|\lambda_i|$ and $\min|\lambda_i|$ the maximum and minimum absolute value of the eigenvalues of $\mathbf{A}$.

In the computation of the inverse of $\mathbf{A}$, $\mathbf{b}$ corresponds to a column of the identity matrix and $\mathbf{x}$ to the corresponding column of $\mathbf{A^{-1}}$. Considering $\|\delta\mathbf{A}\|/\|\mathbf{A}\|=\epsilon_A$ and $|\Delta\mathbf{x}|/|\mathbf{x}|=\epsilon_{A^{-1}}$, one finds

$$\epsilon_{A^{-1}}=\epsilon_A\text{cond}(A) \tag{24}$$

by assuming $\epsilon_{A^{-1}}=\epsilon_A\text{cond}(A)$ negligible compared to 1. This condition is fulfilled until the accuracy is completely lost.

B. Accuracy of a transfer-matrix computation with no layer subdivision

Remembering the considerations of Sec. II C, the accuracy of the matrix $\mathbf{A}$ is expected to be very good. Let us assume it to be $\epsilon_A=\epsilon_{\text{comp}}$. However, for a potential barrier with height V and length D the condition number cond(A) is given by the ratio between the maximum possible value of $A_{i,j}$ (i.e., $e^{K_{\max}D}$) and the minimum possible value (i.e., 1 if propagative solutions exist in region II); so one finds

$$\epsilon_{1\text{slab}}=e^{K_{\max}D}\epsilon_{\text{comp}}, \tag{25}$$

with $K_{\max}=\sqrt{(2m/\hbar^2)V}$.

C. Accuracy of a transfer-matrix computation with a layer subdivision

Let us assume the distance D to be split into n layers with the same length $D_{\text{slab}}=D/n$. The four transfer matrices will be computed with the same accuracy:

$$\epsilon_{\text{slab}}=e^{K_{\max}D/n}\epsilon_{\text{comp}}. \tag{26}$$

Let us consider the construction of the matrices $\mathbf{t}^{-+}$. They are updated iteratively by using Eq. (11). In situations of tunneling by particles with an energy E smaller than the potential barrier height, all incident states are essentially reflected. In typical field emission applications, the top of the potential barrier is located at the beginning of the interval $[0,D]$. The first term of Eq. (11) stands for states that are reflected by the first layer encountered. The second term stands for states that are reflected by the last layer considered. Since this term implies two tunneling processes across the first layer (which grows in the iterative construction), it is negligible compared to the first term and the accuracy of the transfer matrices $\mathbf{t}^{-+}$ has the accuracy of the matrix corresponding to the first layer. We can thus write the accuracy of the matrix $\mathbf{t}^{-+}$ corresponding to the i first layers:

$$\epsilon_{t_i^{-+}}=\epsilon_{\text{slab}}. \tag{27}$$

Since this result is not used in the next part of the development, we do not have to care about the generality of the assumptions made to derive it.

Let us now consider the construction of the matrices $\mathbf{t}^{+-}$. They are updated iteratively by using Eq. (13). In general, the situation differs from the previous case since the first layer encountered is not reflective for all incoming states. The first term of Eq. (13) stands for states that are reflected by the layer i. The corresponding accuracy is ϵ_{slab}. The second term stands for states that are reflected by the set of $i-1$ first layers and are transmitted twice through the layer i. Since propagative solutions exist in the layer i, all components of the second term are not negligible compared to those of the first term. These propagative solutions are reflected by the set of $i-1$ first layers and there is essentially no multiple scattering at the interface with the layer i. The accuracy associated with the second term of Eq. (13) in the step i is then given by $2\epsilon_{\text{slab}}+\epsilon_{t_{i-1}^{+-}}$. By taking the largest value, we have a recursive equation for $\epsilon_{t_i^{+-}}$, whose solution is

$$\epsilon_{t_i^{+-}}=\epsilon_{\text{slab}}(2i-1). \tag{28}$$

Equations (10) and (12) relevant to the matrices $\mathbf{t}^{++}$ and $\mathbf{t}^{--}$ imply the transmission across the set of $i-1$ first layers and the layer i. The factors $[\mathbf{I}-\mathbf{t}^{+-}_{z_0,z_{i-1}}\mathbf{t}^{-+}_{z_{i-1},z_i}]^{-1}$ and $[\mathbf{I}-\mathbf{t}^{-+}_{z_{i-1},z_i}\mathbf{t}^{+-}_{z_0,z_{i-1}}]^{-1}$ stand for multiple scattering at the interface between the two layers considered. Since the matrix to invert is made of numbers with the same order of magnitude, its accuracy is given by $\epsilon_{t_i^{+-}}+\epsilon_{\text{slab}}$. After the inversion, the accuracy of these factors is multiplied by the condition number cond(P), where P stands for $[\mathbf{I}-\mathbf{t}^{+-}_{z_0,z_{i-1}}\mathbf{t}^{-+}_{z_{i-1},z_i}]$ or $[\mathbf{I}-\mathbf{t}^{-+}_{z_{i-1},z_i}\mathbf{t}^{+-}_{z_0,z_{i-1}}]$. This number is expected to be small due to the fact that the components of the matrices to invert are all of the same order of magnitude. Typical values of cond(P) $=5$ are encountered in applications. We thus have the following recurrent relation for the accuracy of the two transfer matrices associated with reflection at step i:

$$\epsilon_i=\epsilon_{\text{slab}}+\epsilon_{i-1}+\text{cond}(P)(\epsilon_{t_{i-1}^{+-}}+\epsilon_{\text{slab}}). \tag{29}$$

By using the expression (28) for $\epsilon_{t_{i-1}^{+-}}$ and (26) for ϵ_{slab}, one finds the accuracy of a transfer-matrix computation, when performed by a subdivision into n layers:

$$\epsilon_n=2^{-n_{\text{bit}}}e^{\sqrt{(2m/\hbar^2)V}D/n}\times\{\text{cond}(P)n^2+[1+\text{cond}(P)]n-2\,\text{cond}(P)\}. \tag{30}$$

V. PRACTICAL CONSIDERATIONS

For practical purposes, the behavior of Eq. (30) is dominated by the factor $e^{\sqrt{(2m/\hbar^2)V}D/n}$, which decreases with increasing n. It is, however, to be noted that for extremely large values of n, the factor $\{\text{cond}(P)n^2+[1+\text{cond}(P)]n-2\,\text{cond}(P)\}$ can make the relative error $\epsilon_n>1$. This occurs when $2^{n_{\text{bit}}}\simeq\text{cond}(P)n^2$. In double precision ($n_{\text{bit}}=53$), n has to take values around 10^7. The hypothesis behind the model presented in Sec. IV C should not be valid for so many layers and the relative error should grow more rapidly. This is expected since the contribution of evanescent states is not negligible compared to propagative states for too thin layers and $\epsilon_{t^{+-}}$ should grow more rapidly.

Since the distance D appears only in the factor $e^{\sqrt{(2m/\hbar^2)V}D/n}$, it is possible to deal with large distances just by increasing the number of layers, as long as n does not take extreme values.

A useful piece of information is the minimum number of layers to consider in order to obtain a relative error $\epsilon_n<1$. If we consider only the factor $e^{\sqrt{(2m/\hbar^2)V}D/n}$, it is given by

$$n_{\min}=\frac{\sqrt{(2m/\hbar^2)V}D}{n_{\text{bit}}\ln(2)}, \tag{31}$$

for which $e^{\sqrt{(2m/\hbar^2)V}D/n_{\min}}=2^{n_{\text{bit}}}$. This peculiar value of $n_{\min}$ corresponds to the minimum number of layers to consider in order to obtain significant transfer matrices in each layer. As a result of the presence of the factor $\{\text{cond}(P)n^2+[1$

$+\mathrm{cond}(P)]n-2\,\mathrm{cond}(P)\}$, a larger number of layers has to be considered. By considering the relation $e^{\sqrt{(2m/\hbar^2)}VD/n_{\min}}=2^{n_{\mathrm{bit}}}$, it can be seen that $n=2n_{\min}$ gives results with approximatively one-half of the represented digits significant and that this number increases by 50% by each further increase of $n_{\min}$. We recommend to use $n=4n_{\min}$.

Another way to improve the accuracy consists in considering only, at each step of the computation, the incident states that have the highest transmission probability. This restriction aims at reducing the local value of $K_{\max}$ to some fixed value $\sqrt{(2m/\hbar^2)\Delta E}$, where ΔE can take the value of a few electronvolts. The transfer matrices corresponding to a smaller number of basic states need less storage space and the time needed to compute them is also reduced. The neglected basic states are considered to be completely reflected by the layer and the corresponding coefficients in the transfer matrices associated with transmission (reflection) are set to the value 0 (1).

VI. APPLICATION TO THE SIMULATION OF FIELD EMISSION

A. Preliminaries

To illustrate this theory, let us consider the electronic field emission from a small tip and the diffraction of the extracted beam by a carbon fiber facing the emitter. The extraction field results from the application of a potential bias V established between the metallic support of the tip and a conducting grid located at a distance D. This grid supports the carbon fiber.

Region I (i.e., the metallic support of the tip) is a Sommerfeld metal, delimited by the plane $z=0$ and characterized by empirical values of W (work function) and E_F (Fermi energy). The potential energy in region III (i.e., the region beyond the conducting grid $z>D$) is set conventionally to the constant value 0. The potential energy value in region I is then $V_{\mathrm{met}}=eV-W-E_F$. With these assumptions, region II is the only diffusive part of the problem and, the Schrödinger equation being linear, the transfer-matrix methodology can be applied.

B. Wave function expansion

Let us assume the axial direction z to be an $\overline{n}$-fold symmetry axis and let us use polar coordinates in the plane normal to the symmetry axis (i.e., ϕ for the azimuthal angle and ρ for the radial distance to the axis). The wave function is then expanded along basic functions ψ that contain the ϕ and ρ dependences. The set of these functions is forced to be enumerable, by specifying that the scattering electron remain localized inside a cylinder with radius R [5].

The basic states $\Psi^{\mathrm{I},\pm}$ and $\Psi^{\mathrm{III},\pm}$ introduced in Sec. II B to describe the wave function in regions I and III take then the specific form

$$\Psi^{\mathrm{I},\pm}_{(m,j)}=e^{\pm i\sqrt{2m(E-V_{\mathrm{met}})/\hbar^2-k_{m,j}^2}z}\psi_{(m,j)}(\rho,\phi), \tag{32}$$

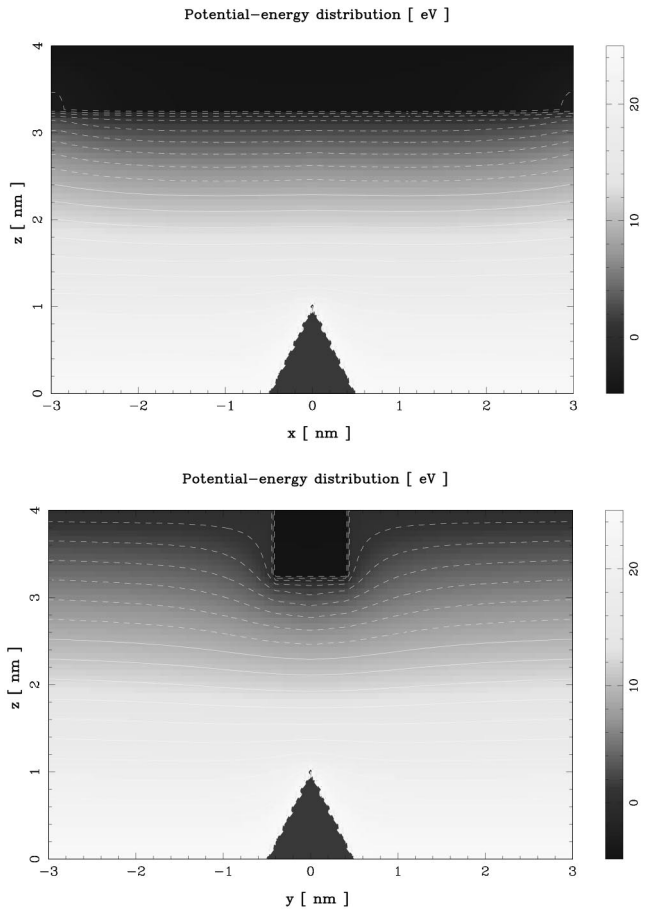

FIG. 1. Potential-energy distribution (in eV) in the x-z plane (top part) and y-z plane (bottom part). A 25 V bias is applied over the 4 nm separation between the metal surface and the conducting grid. This grid supports a carbon fiber oriented along the x axis.

$$\Psi^{\mathrm{III},\pm}_{(m,j)}=e^{\pm i\sqrt{2mE/\hbar^2-k_{m,j}^2}z}\psi_{(m,j)}(\rho,\phi), \tag{33}$$

with

$$\psi_{(m,j)}(\rho,\phi)=\left(\frac{J_m(k_{m,j}\rho)e^{im\phi}}{\sqrt{2\pi\int_0^R\rho[J_m(k_{m,j}\rho)]^2d\rho}}\right), \tag{34}$$

where all functions involved in these expressions have a pair of subscripts (m,j). The radial wave vectors $k_{m,j}$ are solutions of $J'_m(k_{m,j}R)=0$.

C. Propagation equations

To propagate the wave functions $\overline{\Psi}^{+}_{(m,j)}$ and $\overline{\Psi}^{-}_{(m,j)}$ through region II, we use the following expression:

$$\overline{\Psi}^{\pm}_{(m,j)} = \sum_{m,j} \Phi_{(m,j)}(z)\psi_{(m,j)}(\rho,\phi), \tag{35}$$

where the z dependence is contained in the coefficients $\Phi_{(m,j)}(z)$ of the expansion.

When this expression is introduced in the stationary Schrödinger equation, the wave function expansion coefficients $\Phi_{(m,j)}(z)$ verify the exact set of coupled equations

$$\frac{d^2\Phi_{(m,j)}(z)}{dz^2} + \left[\frac{2m}{\hbar^2}E - k^2_{m,j} - \frac{2m}{\hbar^2}V_0(z)\right]\Phi_{(m,j)}(z) = \sum_q \sum_{j'} M^{q,j'}_{m,j}(z)\Phi_{(m-q\overline{n},j')}(z), \tag{36}$$

where E is the electron energy and the coupling coefficients $M^{q,j'}_{m,j}(z)$ are defined by the expression

$$M^{q,j'}_{m,j}(z) = \frac{2m}{\hbar^2} \frac{\int_0^R \rho \overline{V}_q(\rho,z) J_m(k_{m,j}\rho) J_{m-q\overline{n}}(k_{m-q\overline{n},j'}\rho)\,d\rho}{\sqrt{\int_0^R \rho[J_m(k_{m,j}\rho)]^2 d\rho}\sqrt{\int_0^R \rho[J_{m-q\overline{n}}(k_{m-q\overline{n},j'}\rho)]^2 d\rho}}. \tag{37}$$

In these expressions, $V_0(z)$ and $\overline{V}_q(\rho,z)$ are the coefficients used in the n-fold symmetric potential energy:

$$V(\rho,\phi,z) = V_0(z) + \sum_{q=-\infty}^{+\infty} \overline{V}_q(\rho,z)e^{iq\overline{n}\phi}, \tag{38}$$

where the choice of $V_0(z)$ is arbitrary but should correspond to the main part of the potential for better efficiency.

It is to be noted that the coupling between components with different m subscripts occurs only when the corresponding m subscripts are separated by a multiple of the symmetry axis order $\overline{n}$. There are therefore $\overline{n}$ independent groups of coupled components that can be treated independently in the transfer-matrix implementation. For details on how to use Eqs. (36) and (37) and for the computation of the current density associated to all incident basic states in region I, see Refs. [12,13].

D. Characterization of the physical system

Let us consider an electric bias of 25 V and a metal-grid distance of 4 nm. The bulk of the metal is characterized by a Fermi energy of 19.1 eV and a work function of 4.5 eV (values for tungsten). The carbon fiber is assumed to have a dielectric constant of 16.5 (value for diamond [14]) and a work function of 4.82 eV (value for carbon materials [14]). It is infinite along the x axis and has a 0.8 nm section along the y and z directions. The potential distribution in region II is computed by overrelaxation (see Ref. [12]) and represented in Fig. 1.

E. Accuracy considerations

It is possible to predict the accuracy of the transfer-matrix computation as a function of the number of layers used to split the distance D by using expression (30). A better estimation is obtained by using the recurrent relations $\epsilon_{t_i^{+-}} = 2\epsilon_{\text{slab}} + \epsilon_{t_{i-1}^{+-}}$ and $\epsilon_i = \epsilon_{\text{slab}} + \epsilon_{i-1} + \text{cond}(P)(\epsilon_{t_{i-1}^{+-}} + \epsilon_{\text{slab}})$ and considering the local potential barrier height in each slab

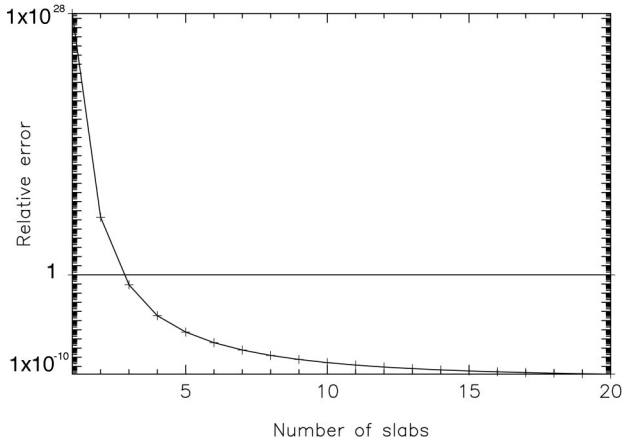

FIG. 2. Accuracy of the transfer matrices for a 25 eV potential barrier with a length of 4 nm, as a function of the number of layers used to split this distance. Values of $\text{cond}(P) = 10$, $n_{\text{bit}} = 53$ (i.e., double precision) are assumed. The horizontal line corresponds to a 100% relative error.

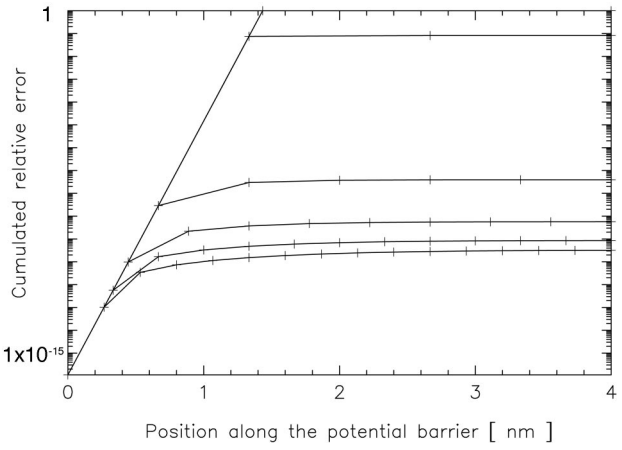

FIG. 3. Cumulated relative error of the transfer matrices along the length of a 25 eV potential barrier, for 1, 3, 6, 9, 12, and 15 layers (from top to bottom). Values of $\text{cond}(P) = 10$, $n_{\text{bit}} = 53$ (i.e., double precision) are assumed. The upper line corresponds to a 100% relative error.

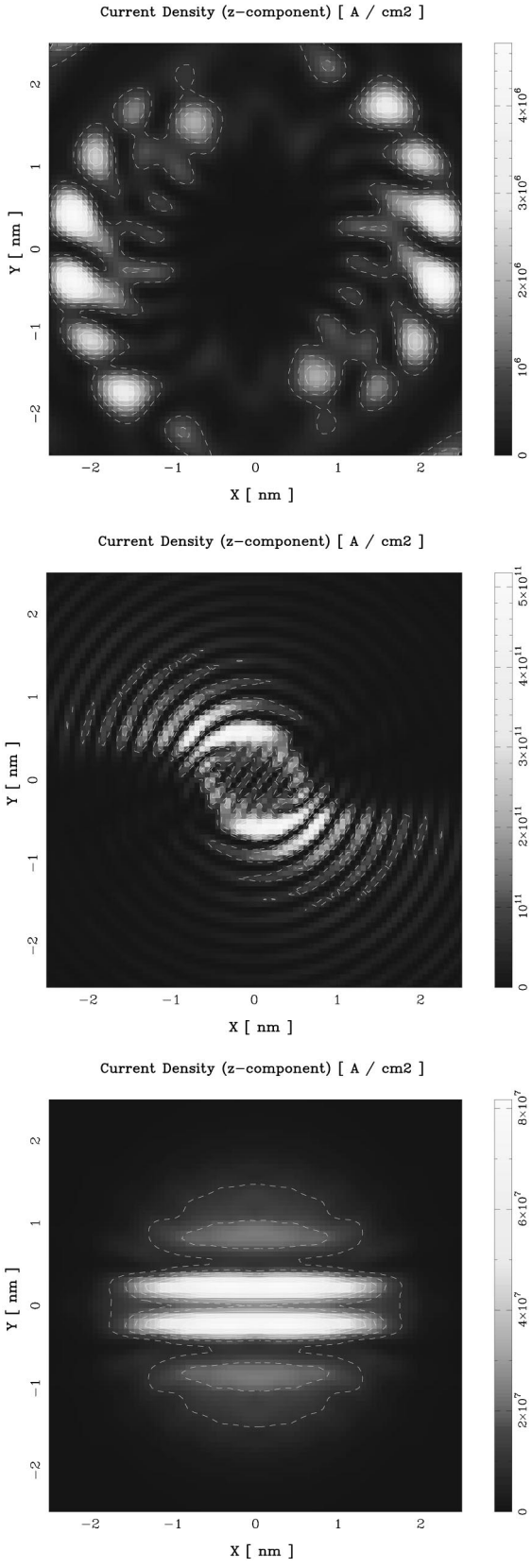

FIG. 4. Current density (z component) computed on the extraction grid. A 25 V bias is applied over the 4 nm separation between this grid and a metal surface. The computation is performed respectively with no layer subdivision (top part), a subdivision into two layers (middle part), and a subdivision into three layers (bottom part).

to define the corresponding value of ϵ_{slab}. The result of this estimation is illustrated in Fig. 2.

By using relation (31), the minimum number of layers to use is found to be 3, in agreement with the results presented in Fig. 2. Expressions (30) and (31) are derived by a model that considers the maximum potential-energy value encountered over the whole distance D to estimate ϵ_{slab} in each layer. This is the reason for ϵ_n and $n_{\min}$ being in general overestimated. Since larger values of n are associated with a better accuracy, expression (31) remains a useful result. It can be seen in Fig. 2 that the gain in significant digits is reduced by a factor of 2 at each additional increase of $n_{\min}$ layers. This appears more clearly in Fig. 3, where the relative error is represented as a function of the position in the length D for 1, 3, 6, 9, 12, and 15 layers.

The result for one slab illustrates the limits of the transfer-matrix methodology, when it is applied without the layer addition algorithm: for a 25 eV potential barrier, all significant digits are lost after 1.4 nm. The figure shows clearly the improvement in accuracy due to a layer subdivision and confirms $4n_{\min}$ to be a good recommendation.

F. Results

The result of the simulations are presented in Fig. 4. The figure shows the current density corresponding to all states incident in region I and evaluated in the plane $z=D$. The different parts of the figure correspond to a computation without layer subdivision, with a subdivision into two layers, and a subdivision into three layers. No change is visible when the number of layers is further increased (we tried up to 400 layers). In agreement with conclusions drawn from Fig. 2, a minimum of three layers is needed to obtain a significant result.

VII. CONCLUSION

The transfer-matrix methodology was presented. It appears limited by numerical instabilities that were related to the physical characteristics of the system considered (potential barrier height and length). The layer addition algorithm comes as a solution to these problems.

To evaluate these methods and determine the minimum number of layers to use in order to obtain significant results, a formalism was developed that accounts for the relative error on a computer-stored result. Simple rules were derived to update the accuracy evaluation in the case of matrix multiplication, addition, and inversion. These rules can be used to evaluate the accuracy of a transfer-matrix computation and give predictions of this accuracy by physical considerations.

The theory was illustrated by a field emission simulation. The model supplies useful information on the dependence of accuracy on the number of layers. The predicted minimum number to use is in agreement with the results of the simulations.

ACKNOWLEDGMENTS

A.M. was supported by the Belgian National Fund for Scientific Research (FNRS). J.-P.V. acknowledges the national program on the Interuniversity Research Project (PAI). The authors acknowledge the use of the Namur Scientific Computing Facility, a common project between the FNRS, IBM-Belgium, and the FUNDP.

[1] P. St. J. Russel, T. A. Birks, and F. D. Lloyds-Lucas, in *Confined Electrons and Photons*, edited by E. Burstein and C. Weisbuch (Plenum, New York, 1995).
[2] W. D. Sheng and J. B. Xia, J. Phys.: Condens. Matter **8**, 3635 (1994).
[3] J. P. Vigneron, I. Derycke, T. Laloyaux, P. Lambin, and A. A. Lucas, Scanning Microsc. Suppl. **7**, 261 (1993).
[4] A. J. Ward and J. B. Pendry, J. Mod. Opt. **44**, 1703 (1997).
[5] T. Laloyaux, A. A. Lucas, J. P. Vigneron, P. Lambin, and H. Morawitz, J. Microsc. **152**, 53 (1988).
[6] T. Laloyaux, I. Derycke, J. P. Vigneron, and A. A. Lucas, Phys. Rev. B **47**, 7508 (1993).
[7] J. B. Pendry, J. Mod. Opt. **41**, 209 (1994).
[8] J. B. Pendry and A. MacKinnon, Phys. Rev. Lett. **69**, 2272 (1992).
[9] G. Binning and H. Rohrer, Helv. Phys. Acta **55**, 726 (1982).
[10] V. T. Binh, V. Semet, and N. Garcia, Ultramicroscopy **58**, 307 (1995).
[11] M. Schatzman, *Analyse Numerique. Cours et exercices pour la licence* (InterEditions, Paris, 1991).
[12] A. Mayer and J. P. Vigneron, Phys. Rev. B **56**, 12 599 (1997).
[13] A. Mayer and J. P. Vigneron, J. Phys.: Condens. Matter **10**, 869 (1998).
[14] W. E. Forsythe, *Smithsonian Physical Tables*, 9th ed. (Smithsonian Institution, Washington, D.C., 1954), pp. 427 and 635.