Improved transfer matrix method without numerical instability

HUIQIONG YIN and RUIBAO TAO(a)

Department of Physics, Fudan University - Shanghai 200433, China

received 19 June 2008; accepted in final form 3 November 2008
published online 16 December 2008

PACS 72.25.Dc – Spin polarized transport in semiconductors
PACS 73.23.-b – Electronic transport in mesoscopic systems
PACS 85.75.Jn – Hybrid Hall devices

Abstract – A new improved transfer matrix method (TMM) is presented. It is shown that the method not only overcomes the numerical instability found in the original TMM, but also greatly improves the scalability of computation. The new improved TMM has no extra cost of computing time as the length of the homogeneous scattering region becomes large. The comparison between the scattering matrix method (SMM) and our new TMM is given. It clearly shows that our new method is much faster than the SMM.

The transfer matrix method (TMM) has been a useful approach for studying the physical properties. There have been many publications on the application of the TMM, such as studies of the Ising model [1–4], quantum spin [5], electronic transport [6,7], and electronic state in quasi-periodic and aperiodic chains [8,9]. The TMM is also widely applied in studying the propagation of electromagnetic waves [10], elastic waves [11] and light waves [12] in multi-layer systems.

The original TMM (OTMM) can be efficiently used for the studies of periodic systems, one boundary problem between two homogeneous media and the scattering problem of a small region sandwiched by two leads. However, the OTMM has a fundamental shortcoming due to its numerical instability when the number of transfer steps is beyond a critical value. This has been a major limitation to the application of the OTMM. Some approaches have been developed to calculate the transport through a longer scattering area such as the scattering matrix method (SMM) [13–15], the extended transfer matrix technique (ETMT) [7] and the Green’s function approach (GFA) [16]. In the SMM, the middle scattering region is cut into some smaller sections, then the scattering matrices of these sections are found by means of the TMM and combined together to get the total scattering matrix (SM) recursively. In the ETMT, Wan et al. used a technique to cancel the increasing modes step by step for each slice to avoid the numerical instability. Although the numerical instability in the SMM and the ETMT does not exist anymore, the CPU time cost grows linearly when the length of the middle scattering region increases and the GFA, too. Based on a continuous Schrödinger equation in an electron wave guide without spin-orbit coupling (SOC) [17] and with SOC [18], the stable solutions of transport were achieved. However, there is an infinite number of evanescent modes that pose tremendous numerical difficulties, and the approximation of the limit number of the evanescent wave must be done. But in some case, the contribution of the evanescent wave is significant, even more in the presence of SOC [18,19]. In this letter, we present a new improved TMM (NITMM) that has overcome such non-physical numerical instability of the OTMM. Our study will focus on the discrete version of the Schrödinger equation with Rashba SOC [20] and will find the exact solution of electron transport with and without SOC. Our new method not only gives numerically stable solution at any length of the sample, but also provides a high performance in computation. That is, no extra computing costs at increasing length of sample. Meanwhile the simplicity of the OTMM is still maintained.

General formulae of TMM. – A 2d strip with the geometry of a bar sandwiched by two semi-infinite leads is studied. The Hamiltonian of 2d electron gas is

\[ \hat{H} = \frac{\hat{p}_x^2 + \hat{p}_y^2}{2m^*} + V(x, y) + H_{so}, \]

where \( V(x, y) \) is the potential including confinement boundary, \( H_{so} \) the remainder of the Hamiltonian that may contain spin-orbit coupling. After discretization of the Schrödinger equation \( \hat{H}\Psi(x, y) = E\Psi(x, y) \), the strip is replaced by a lattice array with \( N \) infinite long

(a)E-mail: rbtao@fudan.edu.cn
chains. The lattice constant is assumed to be \( a \). A wider strip needs a larger chain number \( N \) to be represented. The Schrödinger equation can be transformed into the following transfer matrix (TM) equations:

\[
\Phi_{i+1} = T_i \Phi_i, \quad \Phi_{i+m} = \left( \prod_{i=1}^{m} T_i \right) \Phi_i,
\]

where \( \Phi_i = (\phi_{i,N \uparrow}, \phi_{i,N \downarrow}, \phi_{i,1 \uparrow}, \phi_{i,1 \downarrow}, \phi_{i-1,1 \uparrow}, \phi_{i-1,1 \downarrow}, \ldots, \phi_{1,1 \uparrow}, \phi_{1,1 \downarrow})^T \). \( T_i \) is a matrix with dimension \( 4N \times 1 \). The superscript \( t \) means the transpose of the matrix. The element \( \phi_{i,j,\sigma} \) in \( \Phi_i \) is the value of wave function \( \Psi(x, y) \) at \( (i, j) \) with spin index \( \sigma \). The lattice indices are \( \{(i, j) | i \in (-\infty, -L - 1], \text{ right lead in } i \in [L + 1, \infty) \text{, and the middle bar in } i \in [-L + 1, L - 1]. \)

Two interfaces are at \( i = -L \) and \( i = L \). \( T_i \) is a \( 4N \times 4N \) TM between the wave functions \( \Phi_i \) and \( \Phi_{i+1} \). Two leads can be different or the same, and here we assume they are the same. Further, we assume that the leads and the middle bar are homogeneous. Thus, we have four different transfer matrices: \( T_i \) is the TM in two leads, \( T_{so} \) in the bar, \( T_{SL} \) and \( T_{SN} \) at the left and right interface respectively.

Fixing adjustable parameters like the hopping constant \( t_0 \), the on-site potential and the energy of the electron \( E \), we can obtain all elements of four TM matrices \( \{T_i, T_{so}, T_{SL}, T_{SN}\} \). It can be verified that the determinants of \( T_i \) and \( T_{so} \) within homogeneous regions are 1. Then two transform matrices \( U_i \) and \( U_{so} \) can be numerically solved to diagonalize \( T_i \) and \( T_{so} \): \( D_i = U_i T_{so} U_i^{-1} \) and \( D_{so} = U_{so} T_{so} U_{so}^{-1} \). When one knows the matrix \( \Phi_i \) at an arbitrary site \( i \), in principle, the wave function can be found anywhere by TM equations. We denote \( \Phi_{i} = U_i \Phi_i \) in leads and \( \Phi_{\uparrow} = U_{so} \Phi_i \) in the middle bar. The TM equations in diagonal representation can be written as

\[
\Phi_{i+1} = \begin{cases} 
D_i \tilde{\Phi}_i, & i \in (-\infty, -L - 1] \cup [L + 1, \infty), \\
D_{so} \Phi_i, & i \in [-L + 1, L - 1].
\end{cases}
\]

When the strip has no interface and fully homogenous or only one interface and two sides are homogenous, it has been shown that the OTTM works well and has no numerical instability. In this letter, we study the strip with two interfaces. In this case, the OTTM will have a serious overflow problem if the length of the middle bar between two leads is longer than a critical value.

We denote \( D_{so} \) and \( D_i \) as \( D^{(c)} \), \( c = so, l \). The diagonal elements \( \lambda^{(c)}_i = D^{(c)}_{i} \) can be classified into two types: \( |\lambda^{(c)}_i| = 1 \) and \( |\lambda^{(c)}_i| \neq 1 \), which relate to propagating and evanescent modes, respectively. For mode \( |\lambda^{(c)}_i| = 1 \), the diagonal element can be rewritten as \( e^{+i k_i a} \), where \( k_i > 0 \) is a real number, and \( a \) is the lattice constant. \( e^{i k_i a} \) is called as “right-going” wave, and \( e^{-i k_i a} \) is the “left-going” one. For mode \( |\lambda^{(c)}_i| \neq 1 \), the diagonal element can be rewritten as \( e^{\pm i (\eta_i + \phi_i)} \), where \( \eta_i \) is a positive real number and \( \phi_i \) the phase. The \( e^{\pm i (\eta_i + \phi_i)} \) is a “right-growing” or, say, “left-decaying” mode, and \( e^{\pm i (\eta_i - \phi_i)} \) the “right-decaying” or “left-growing” one. Due to the definition \( T_i = |D^{(c)}_i| = 1 \) in the homogenous region, any modes \( e^{\pm i k_i a} \) or \( e^{\pm i (\eta_i + \phi_i)} \) must appear in pairs. Then we can always arrange the diagonal elements into the order: \( \text{diag}\{D^{(c)}_i\} = \{e^{ik_1a}, e^{ik_2a}, e^{i(-\eta_1+\phi_1)}, \ldots, e^{i(-\eta_n+\phi_n)}\} \). The first \( p + q \) states in \( \text{diag}\{D\} \) correspond to \( p \) right-going and \( q \) right-decaying modes, and the second \( p + q \) states to \( p \) left-going and \( q \) left-decaying modes. At eigen energy \( E \), totally we have \( 2p \) propagating modes, and \( 2q \) evanescent modes. \( 2p + 2q = 4N, N \) is the number of chains. For each energy \( E \), there are \( 4N \) modes distributed in \( 2N \) channels corresponding to \( 2N \) different \( |\lambda_i| \), where \( |\lambda_{1 \ldots p}| = 1 \) and \( |\lambda_{p+1 \ldots 2N}| > 1 \). If \( q \) equals zero, all states are extended.

When \( p = 0 \), no propagating wave can exist in the strip. Changing energy \( E \) results in the change of the \( p \) and \( q \).

We assume that a right-going electron wave, \( e^{ik_1a} \), is injected from the first channel of the left lead. In general there should be some reflection waves in all channels in the left lead, due to the scattering of interfaces. We denote the reflection waves in \( 2N \) channels as \( \{r_1 e^{ik_1a}, r_{p+1} e^{ik_1a}, \ldots, r_{p+q} e^{ik_1a}\} \). We can further set the wave function at position \( i = -L \) to be \( \Phi_{-L} = (1, 0, \ldots, 0, r_1 e^{i \phi_1}, \ldots, r_{p+q} e^{i \phi_p})^T \). The phases of reflection waves have been absorbed in coefficients \( \{r_i\} \). We have \( \Phi_{-L} = D^L \Phi_{-L-1} \). \( \Phi_{-L-1} = T_{SL} \Phi_{L-1} \). \( \Phi_{L+1} = (D_{so})^{2L-1} \tilde{T}_{SL} \tilde{\Phi}_{L-1} = \tilde{S} \tilde{\Phi}_{L-1} \). When \( i = -L \) can be expressed by \( \tilde{\Phi}_{L+1} = (t_1, t_1, t_2, t_4 N, 0, \ldots, 0)^T \), where \( \{t_i\} \) are the transmission coefficients. Thus, we obtain \( 2N \) equations:

\[
(\tilde{\Phi}_{L+1})_0 = 0, \quad (\tilde{\Phi}_{L+1})_1 = 2N + 1, \ldots, 4N.
\]
problem comes from the term \((D_{so})^{2L-1}\) that may have some increasing modes with diagonal elements like \(|\lambda_{p+i}|^{2L-1}\). If the value of \(|\lambda_{p+i}|^{2L-1} \gg 1\), many elements \(W_{lm} = \sum_{i=1}^{4N}\tilde{T}_{S_R}(2N+i,\lambda_{ai}^{2L-1}(\tilde{T}_{S_L}),2N+m)\) are of the order of \(|\lambda_{p+i}|^{2L-1}\). Hence the calculation of \(W^{-1}\) will meet the numerical overflow of the order of \((\lambda^{2L-1})^n\), 1 \(\ll n \ll 2N\). Therefore, the OTTM can only accurately solve the solution for the bar system with smaller length \(L\). For example, we have calculated a Rashba SOC system with \(N = 200\) by means of the OTTM. The longest length of the bar is around 15, and numerical instability occurs for \(2L > 15\). Larger \(N\) results in smaller \(2L\).

**New improved TM method.** In fact, the physics here must be finite. The solution for \(\{r_i\}\) should be stable. The numerical overflow is an artifact of the OTTM. Even though \(|\lambda|^{2L-1} \gg 1\) for evanescent modes when \(2L - 1 \gg 1\), the wave function after \(2L - 1\) steps of the transfer by \((D_{so})^{2L-1}\) should still be finite. Thus, the values of \((\tilde{T}_{S_R}\tilde{\Phi}_{-L})2N+i,1,2,\ldots,q\) corresponding to the increasing modes of \(|\lambda_{p+i}| > 1\) must be very small to assure \((D_{so})^{2L-1}\tilde{T}_{S_R}\tilde{\Phi}_{-L}\) to be finite. That is the physical requirement. Thus, we introduce \(q\) new auxiliary parameters \(\{\zeta_i\}\) and assume that
\[
(\tilde{T}_{S_R}\tilde{\Phi}_{-L})2N+i = \zeta_ie^{-(2L-1)(\eta_{a+i\delta_i}), i = 1, \ldots, q. (6)
\]

The \(q\) parameters \(\{\zeta_i\}\) have to be determined together with the \(2N\) reflection coefficients \(\{r_i\}\). The elements in the matrix \(\tilde{\Phi}_{L+1}\) are the linear combination of the \(2N\) coefficients \(\{r_i\}\) and the \(q\) auxiliary parameters \(\{\zeta_i\}\). We have
\[
(\tilde{\Phi}_{L+1})_a = D_a + \sum_{i=1}^{2N} C_{ai}^a r_i + \sum_{j=1}^{q} B_{aj}^a \zeta_j,
\]
where
\[
\begin{cases}
D_a = \sum_{\beta=1}^{2N+p} (\tilde{T}_{S_R})_{a\beta}(D_{so})^{2L-1}\tilde{T}_{S_L})_{\beta 1}, \\
C_{ai} = \sum_{\beta=1}^{2N+p} (\tilde{T}_{S_R})_{a\beta}(D_{so})^{2L-1}\tilde{T}_{S_L})_{\beta 2N+i}, \\
B_{aj} = (\tilde{T}_{S_R})_{a2N+p+i}, \quad a = 1, 2, \ldots, 4N.
\end{cases}
(8)
\]

No left-going wave in the right lead is requested, i.e. \((\tilde{\Phi}_{L+1})_a = 0, \quad a = 2N + 1, \ldots, 4N\), which yield the following 2N equations:
\[
D_a + \sum_{i=1}^{2N} C_{ai}^a r_i + \sum_{j=1}^{q} B_{aj}^a \zeta_j = 0.
(9)
\]

With 2N equations (9) and \(q\) auxiliary equations (6), totally we have \(2N + q\) equations that can uniquely determine \(2N + q\) unknown parameters \(\{r_i\}\) and \(\{\zeta_i\}\). All coefficients of \(\{\zeta_i\}\) and \(\{r_i\}\) are finite here so that the inverse of the coefficient matrix can be calculated without overflow and a unique solution of \(\{\zeta_i\}\) and \(\{r_i\}\) is obtained. Now the numerical overflow problem does not exist any more. The method is stable for any length. After all \(\{r_i, \zeta_i\}\) are found, the transmission coefficients \(t_{\alpha}\) can be obtained by equations
\[
t_{\alpha} = (\tilde{\Phi}_{L+1})_{\alpha}, \quad \alpha = 1, 2, \ldots, 2N.
(10)
\]

As an example, we use our NITMM to study the electron transport through a Rashba bar sandwiched by two semi-infinite metal leads. The Hamiltonian in the bar region is
\[
\hat{H} = \frac{\hat{p}_x^2 + \hat{p}_y^2}{2m^*} + \frac{\alpha}{\hbar} (\hat{\sigma}_x \hat{p}_x - \hat{\sigma}_y \hat{p}_y) + V_{conf}(x, y),
(11)
\]
where \(\hat{\sigma}_x, \hat{\sigma}_y, \hat{\sigma}_z\) are Pauli matrices, \(\alpha\) the strength of the Rashba SOC, and \(V_{conf}(x, y)\) the transverse confining potential. Here an open-boundary condition in the \(y\)-direction is applied. The TM in the leads and the Rashba bar can be written as the following super-matrix:
\[
\begin{pmatrix}
A & B & \cdots & 0 \\
B^* & A & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & A \\
0 & 0 & \cdots & B^*
\end{pmatrix}
(12)
\]
where \(A\) and \(B\) are two \(4 \times 4\) sub-matrices, \(A^*(B^*)\) is the complex conjugate matrix of \(A(B)\).

\[
A = \begin{pmatrix}
a & b & c & f \\
1 & 0 & 0 & 0 \\
-e & -f & a & b \\
0 & 0 & 1 & 0
\end{pmatrix}, \quad B = \begin{pmatrix}
g & 0 & h & 0 \\
0 & 0 & 0 & 0 \\
-h^* & 0 & g^* & 0 \\
0 & 0 & 0 & 0
\end{pmatrix}
(13)
\]
where \(\{a, b, c, e, f, g, h\}\) are the functions of the hopping constant \(t_0\), SOC strength \((t_{so}\) in the Rashba region, 0 in the lead region) and the eigen energy \(E\). The expressions of \(\{a, b, c, e, f, g, h\}\) are \(a = -c(E - w_{xy}i)/t_0, b = c(t_{so}^2/t_0^2 - 1), e = -ct_{so}(E - w_{xy}i)/t_0^2, f = -2ct_{so}t_0, g = c(iu_{so}/t_0)\) for some \(c = (1 + t_{so}^2)/t_0\) to zero, \(h = -c(1 - i)t_{so}/t_0, c = (1 + t_{so}^2)/t_0\). Finally the wave functions at any site of the strip can be obtained. We compare the result of our new improved method with that of the OTTM for \(2L \leq 15\) for the \(N = 200\) system within which the calculation of the OTTM is accurate and has no numerical instability. We obtain exactly the same results for the transmission and reflection coefficients. However, our new method can accurately calculate the case of \(2L\) as large as we want. The main cost of computer time is from the diagonalization of the 800 \times 800 TM. It is clearly shown that our new
Fig. 1: (Colour on-line) a), b), c) are the spin polarization ($x = 10$) for the conductor with width $200a$ in the $y$-direction and length $20a$ in the $x$-direction, while d), e), f) ($x = 300$) for the conductor with width $200a$ in the $y$-direction and length $600a$ in the $x$-direction, with the unit $\hbar/2$. The continuous line includes the evanescent waves (EW). The dotted line shows only the extending states.

Fig. 2: (Colour on-line) The upper orange line stands for transmission rate, while the lower black line is the reflection rate.

Fig. 3: (Colour on-line) The time cost of the SMM (orange upper line) increases as the first order of lengths of the scattering region, whereas that of NITMM (black lower line) is the zeroth order of lengths of the scattering region. The maximum length of a single block within which the SMM is applicable is less than 400a, since there is a matrix inversion operation in the SMM formulae (see footnote 1). Here we choose the maximum length as 300a to ensure the precision of calculation.

Comparison between our new method and previous methods (SMM and ETMT). – The SMM has been mostly applied for studying the wave transport and has no problems of numerical instability [13]. In the SMM, in order to avoid the numerical instability, one has to divide the middle bar into many small sections. Then one should find the SM for each section via the TMM. In the sample with $N = 100$, the largest length of such pieces should be less than 400 lattices such that no numerical instability occurs. Here we use the length of 300a to ensure the precision of calculation. Then a recursion approach is applied to combine every sub SM to reach the final total SM. Figure 3 shows the detailed formulae of recursive SMM can be found on p. 223, formula (56), ref. [13].
that the CPU cost of the SMM linearly increases with growing length of the scattering region, but our new TMM is independent of the length of the middle bar as shown in the same figure. The numerical solutions obtained by the SMM and our new TMM are exactly the same. In addition, when one studies the transmission of a polarized incident wave going through the middle bar, considering only a single-channel injection of the electron wave is not sufficient for the SMM. All channels injection from left and right must be calculated individually to obtain the SM. Thus, the computing time to obtain the SM of the first block near the interface is almost 4N times of the NITMM. Furthermore, the computing time grows linearly for adding every block by a recursive algorithm in the SMM. In the ETMT [7], the magnetoconductance of a quantum wire with several antidots was studied. The numerical technique used there could avoid numerical instability occurred in the original transfer matrix method, and was applied to a variety of 3d systems involving complicated atomic and many-body potentials. However, due to its iterative calculations, the computing time of the ETMT also increases linearly with the increase in the length of the quantum wire even for the homogeneous region. Our method shows the superiority in treating the long homogeneous transfer region since it does not cost extra computing time as the length of the homogeneous scattering region becomes large. The computing time is in the zeroth order of the homogenous region length L, $O(L^0)$. For the model we calculated here, our method is much faster than the SMM for large-length system, and we believe that it is also faster than the ETMT. Our method can also be applied to the studies on the transport under a uniform magnetic field. Extensions of our NITMM to other problems will be our future work.

This work is supported by the National Natural Science Foundation of China (Nos. 10674027 and 10547001) and 973 project of China under grant No. 2006CB921300.

REFERENCES

[1] Schultz T. D., Mattis D. C. and Lieb E. H., Rev. Mod. Phys., 36 (1964) 856.
[2] Morgenstern I. and Binder K., Phys. Rev. Lett., 43 (1979) 1615.
[3] Morgenstern I. and Binder K., Phys. Rev. B, 22 (1980) 288.
[4] Cheung H.-F. and McMillan W. L., J. Phys. C: Solid State Phys., 16 (1983) 7027.
[5] Suzuki M., Phys. Rev. B, 31 (1985) 2957.
[6] Sautet P. and Joachim C., Phys. Rev. B, 38 (1988) 12238.
[7] Wan C. C., Jesus T. D. and Guo H., Phys. Rev. B, 57 (1998) 11907.
[8] Tao R., J. Phys. A: Math. Gen., 27 (1994) 5069.
[9] Huang X. and Gong C., Phys. Rev. B, 58 (1998) 739.
[10] Chilwell J. and Hodgkinson I., J. Opt. Soc. Am. A, 1 (1984) 742.
[11] Wang L. and Rokhlin S. I., Ultrasonics, 39 (2001) 413.
[12] Li Z.-Y. and Lin L.-L., Phys. Rev. E, 67 (2003) 046607.
[13] Pendry J. B., J. Mod. Opt., 41 (1994) 209.
[14] Pendry J. B. and MacKinnon A., Phys. Rev. Lett., 69 (1992) 2772.
[15] Bell P. M., Pendry J. B., Moreno L. M. and Ward A. J., Comput. Phys. Commun., 85 (1995) 306.
[16] Thouless D. J. and Kirkpatrick S., J. Phys. C, 14 (1981) 235.
[17] Wu H. and Sprung D. W. L., Appl. Phys. A, 58 (1994) 581.
[18] Yao J. and Yang Z., Phys. Rev. B, 73 (2006) 033314.
[19] Li Z. and Yang Z., Phys. Rev. B, 76 (2007) 033307.
[20] Rashba E. I., Sov. Phys. Solid State, 2 (1960) 1109.