Why Error Correction Methods Realize Fast Computations

Hajime Igarashi, Member, IEEE, Yuta Watanabe, Yasuhisa Ito

Graduate School of Information Science and Technology, Hokkaido University, 060-0814, Sapporo, Japan

In this paper, deflation method, A-phi method, explicit and implicit error correction (EC) methods and time-periodic explicit error correction (TP-EEC) method are shown to be based on the same mathematical principle that the slowly converging errors are effectively eliminated through the EC process while the fast converging errors are reduced by iterative linear solvers. The properties of the TP-EEC are discussed in detail and the EC is shown to work as an error filter. The numerical experiments show that the error reduction in the time-periodic eddy current problem is in good agreement with those predicted by the present theory.

Index Terms—Finite Element Method, Error Correction Method, Multigrid Method, Quasi-static Electromagnetic Field

I. INTRODUCTION

The multigrid method decomposes the errors of iterative linear solvers such as Jacobi and Gauss-Seidel methods into fast and slowly converging components by use of the restriction-prolongation matrix [1]. The fast converging errors with oscillating profiles can be eliminated by relatively small number of iterations. The slowly converging errors, on the other hand, are eliminated by the coarse grid correction. The eigenvalue deflation, which has recently been applied to diffusion and magnetostatic problems [2, 3] for acceleration of linear solution, matrix conditioning is improved by replacing small eigenvalues with zeros.

In the explicit and implicit error correction (EEC and IEC) methods [4], unknowns are decomposed into fast and slowly converging components, the latter of which are constructed, for example, for flat finite elements on the basis of physical insight [5]. The similarity between the IEC and the AV (or A-phi) methods has been pointed out in [6], which simultaneously solve the original and correction equations. In this paper, the methods which are based on the correction to eliminate slowly converging errors, as in the multi-grid, deflation methods, EEC and IEC, are called the error correction (EC) method.

It has been shown that the reason why these EC methods work well can be explained through the eigenvalue deflation [7]. However, it would be difficult to give a clear explanation of the time-periodic EEC (TP-EEC) method [8, 9], which accelerates the convergence of the time stepping scheme applied to diffusion-type transient systems with periodical inputs to their steady states, by the eigenvalue deflation principle. For this reason, we need a stronger mathematical framework to explain why the EC methods realize convergence acceleration. This paper presents a unified formulation of the EC methods including TP-EEC to clarify their common mathematical principle. In particular, the effect of TP-EEC is analyzed in detail based on this principle. It will be shown that TP-EEC works as an error filter in IV.

II. FORMULATION OF ERROR CORRECTION

Let us consider a system of linear equations

$$Ax = b,$$  \hspace{1cm} (1)

where A is an $M \times M$ regular matrix. (If A is singular, the following formulation is modified by introducing a generalized inverse matrix.) When matrix conditioning of A is poor, iterative solution of (1) requires long computational time. To accelerate the convergence, the EC method is applied to (1), which will be formulated below.

The linearly independent vectors $w_i$, $i=1, 2, \ldots, k$, $k \ll M$, are introduced, which are “smooth” in a sense that

$$A w_i \approx 0.$$  \hspace{1cm} (2)

When using the conjugate gradient (CG) method as a linear solver, its convergence is slow if the condition number, the ratio of the maximum to minimum eigenvalues of A, is large. Thus the numerical error $e = x_{\text{exact}} - \tilde{x}$ in (1) remained after CG iterations would belong to the sub-space spanned by $w_i$, which correspond to the eigenvectors of the small eigenvalues of A. When using stationary iterative solvers such as the Jacobi and Gauss-Seidel method, A is expressed by $A = M - N$, and iteration $x^n = M^{-1}(Nx^{n-1} + b)$ is carried out. The numerical errors here obey $e^n = M^{-1}\text{Ne}^{n-1}$. Thus if the errors belong to the space spanned by $w_i$, $e^n \approx e^{n-1}$ hold which means slow convergence.

As discussed above, the slowly converging errors in the approximated solution $\tilde{x}$ to (1) could be expressed by the linear combination of $w_i$, that is, $e_{\text{slow}} = Wz$, where $W=[w_1, w_2, \ldots, w_k]$. The unknown coefficient vector $z$ can be determined from the condition that the fast converging error $e_{\text{fast}} = e - e_{\text{slow}}$ is A-orthogonal to the space spanned by $w_i$, that is,

$$W^T A (e - Wz) = 0,$$  \hspace{1cm} (3)

which is equivalent to

$$W^T A Wz = W^T (b - A \tilde{x}),$$  \hspace{1cm} (4)

because of $A(x_{\text{exact}} - \tilde{x}) = b - A \tilde{x}$. By solving (4) for $z$, the approximated solution $\tilde{x}$ is corrected by
\[ \tilde{x}_{\text{new}} = \tilde{x} + Wz. \] (5)

To see the effect of the EC, the solution \( z \) to (4) is substituted into (5) to obtain
\[ \tilde{x}_{\text{new}} = \tilde{x} + W(W^tAW)^{-1}W^t(b - A\tilde{x}) \] (6)

By subtracting \( x_{\text{exact}} \) from both sides of (6), we have
\[ e_{\text{new}} = Pe, \] (7)

where
\[ P = I - W(W^tAW)^{-1}W^tA, \] (8)
and \( I \) denotes the identity matrix. In (7), \( P \) is a projection matrix satisfying \( P^2 = P \). The fast and slowly converging errors can be expressed as
\[ e_{\text{fast}} \in \text{Ker}(W^tA) = \{ e | W^tAe = 0 \}, \] (9a)
\[ e_{\text{slow}} \in \text{Range}(W) = \{ e | e = Wy, \ \forall y \in R^k \}. \] (9b)

It follows from (7)-(9) that each error component satisfies \( Pe_{\text{slow}} = 0 \), \( Pe_{\text{fast}} = e_{\text{fast}} \). Hence, the EC eliminates \( e_{\text{slow}} \) while \( e_{\text{fast}} \) remains unchanged, the latter of which is effectively reduced through iterative linear solvers such as CG and Gauss-Seidel methods.

Finally we consider non-linear equations
\[ A(x) = b, \] (10)

instead of (1). When the Newton-Raphson method is applied to (10), one has the linearized equation
\[ \frac{\partial A}{\partial x} e = b - A(\tilde{x}). \] (11)

The A-orthogonality between \( e_{\text{fast}} = e - e_{\text{slow}} \) and \( w_i \) with respect to the Jacobi matrix in (11) leads to
\[ W^t \frac{\partial A}{\partial x} Wz = W^t[b - A(\tilde{x})]. \] (12)

It is, therefore, concluded that \( A \) is just replaced with the Jacobi matrix in (11) to obtain the projection matrix \( P \) in (8) for the non-linear case.

### III. Unified Derivation of Error Correction Methods

It can be found that the following methods can be recognized as variants of the EC method based on (6)-(8) with different choices of \( W \). TABLE I summarizes the choice of \( W \) for each method. The matrix \( W \) can be constructed by finding smooth vectors satisfying (2).

#### A. Multigrid Method

In this method, \( e_{\text{slow}} \) corresponds to spatially smooth errors which cannot be effectively reduced by, e.g., the Gauss-Seidel and CG methods. The error \( e_{\text{slow}} \) is shown to be in the range of the restriction matrix \( R \) which maps vectors in a fine grid to those in a coarse grid. In the multigrid, \( e_{\text{slow}} \) is eliminated by (7) where \( W = R [1] \). 

#### B. Deflation Method

In the deflation method, the decomposition (5) is modified to \( \tilde{x}_{\text{new}} = P\tilde{x} + (1 - P)\tilde{x} \), where \( P \) is defined in (8), and the first and second terms represent fast and slowly converging components, respectively. The latter can be obtained from
\[ (I - P)\tilde{x} = W(W^tAW)^{-1}W^t(b - A\tilde{x}) \] (13)

while the former is obtained by solving
\[ A\tilde{P}x = P^t b, \] (14)

where the commutative property \( AP = P^t A \) is used to derive (14). In the typical deflation method, \( W \) is composed of the eigenvectors corresponding to the small eigenvalues of \( A \). The matrix \( AP \) in (14) is proved to have better conditioning than \( A [2] \). Substitution of (13) into the decomposed form of \( \tilde{x}_{\text{new}} \) results in (7).

#### C. EEC and IEC, TP-EEC

In the EEC method, the restriction matrix \( R \) used in the multigrid method is replaced by a matrix \( W \) whose column vectors \( w_i \) represent slowly converging components. In this method, (4) and
\[ A\tilde{x} + AWz = b, \] (15)

which can be obtained by multiplying \( A \) to (5) from left and setting \( A\tilde{x}_{\text{new}} = b \), are alternatively solved. The IEC method simultaneously solves (4) and (15), that is,
\[ \begin{bmatrix} A & AW \\ W^tA & W^tAW \end{bmatrix} \begin{bmatrix} \tilde{x} \\ z \end{bmatrix} = \begin{bmatrix} b \\ W^tb \end{bmatrix} \] (16)

The augmented matrix in (16) has better conditioning in comparison with \( A \) after pre-conditioning [7]. It can easily be shown that both methods are based on the EC given by (7).

In TP-EEC method, the EEC method is extended to time-periodic eddy current problems which have long time constant compared to the excitation period. The EEC will be discussed in the next section.

#### D. AV method for eddy current analysis

Let us consider quasi-static fields governed by
\[ \text{rot}(v \text{rot}\mathbf{A}) + \sigma \left( \frac{\partial \mathbf{A}}{\partial t} + \nabla \nabla V \right) = \mathbf{J}, \] (17)

where \( \mathbf{J} \) denotes the current source, and magnetic permeability \( \mu = 1/\nu \) and current conductivity \( \sigma \) are assumed to be constant for simplicity.

The vector potential \( \mathbf{A} \) in (17) is expressed by means of the edge-based basis functions \( N_e \), namely, \( \mathbf{A} = \sum_e a_e N_e \). Because the space spanned by \( N_e \) includes the discrete gradient field, the scalar potential \( V \) can be ignored in the FE equation. This is the A method, which solves
\[ \left( C + K_0^0 \right) a^n = C a^{n-1} + f^n, \] (18)

where \( K_{ee}^0 = (v \text{rot}N_e, v \text{rot}N_e'), \quad C = M/\Delta t - (1 - \theta)K_0^0, \quad M_{ee'} = (\sigma N_e, N_{e'}), \quad f^n = \theta b^n + (1 - \theta)b^{n-1}, \quad b_e = (N_e, J), \quad 0 \leq 0 \leq 1 \). Note here that \( K_0^0 \) is a singular matrix. The serious problem in the A method is that \( C + K_0^0 \) becomes ill-conditioning when \( \Delta t \) becomes large [10]. To overcome this difficulty, the IEC method is applied to (18) by decomposing \( a \) into \( a + G V \), where \( G \) is the discrete counterpart of the grad operator, whose column vectors span \( \text{Ker}(K_0^0) \). Note that
G has the property in (2) when $\Delta t$ is large. The IEC equation, (16), is now given by

$$(C + K)u^n = Cu^{n-1} + f^n$$

(19)

where $u \in R^m$ consists of the unknowns $u'$ and $V$ assigned at edges $e$ and nodes $n$, and

$$K = \begin{bmatrix} K_0 & 0 \\ 0 & 0 \end{bmatrix}.$$  

$$C = \frac{1}{\Delta t} \begin{bmatrix} M & S \\ S^t & N \end{bmatrix} - (1 - \theta)K.$$  

(20a)

The entities of the matrices in (20) are defined as $S_{en} = (\sigma V_n, \nabla N_n), \ N_{nm} = (\nabla V_n, \nabla N_n).$ The method solving (19), called AV method, gives good convergence in comparison with the A method at each time step. However, a number of time steps are required to obtain steady-state solutions when the dominant time constant of the system governed by (19) is much longer than $\Delta t$. To accelerate convergence, the TP-EEC method is applied to (19), which will be formulated in IV.

**TABLE I HERE**

### IV. TIME PERIODIC ERROR CORRECTION

#### A. Formulation

It is assumed here that $J$ is periodic so that $u^{n+N\Delta t} = u^n$ holds in the steady state. In TP-EEC, equation (19) is expressed over one period in the form

$$\begin{bmatrix} C + K & 0 & \cdots & 0 \\ -C & C + K & \cdots & 0 \\ 0 & -C & C + K & \cdots \\ \vdots & \vdots & \ddots & \ddots \\ -C & -C & \cdots & C + K \end{bmatrix} \begin{bmatrix} u^1 \\ u^2 \\ \vdots \\ u^{N} \end{bmatrix} = \begin{bmatrix} f^1 \\ f^2 \\ \vdots \\ f^{N} \end{bmatrix}.$$ 

(21)

which corresponds to (1), where $M=Nm$. Note that (19) is solved through time stepping in TP-EEC while (21) is considered to derive the correction equation. Assuming that (21) is solved using the Gauss-Seidel method, where the residuals are assumed to be zero for $n>1$. On the other hand, because

$$f^1 - (C + K)u^1 + Cu^0 = 0,$$ 

(22)

would hold, the residual at $n=1$ is expressed as

$$r^1 = f^1 - (C + K)u^1 + Cu^N = -C(u^0 - u^N).$$ 

(23)

Moreover, the slowly converging error is assumed to be temporary constant as

$$W = [1, I, \ldots, I]^t \in R^{M \times m},$$ 

(24)

which would satisfy (2) when $\Delta t$ is sufficiently smaller than the dominant time constant of the system. The correction equation corresponding to (4) is given by

$$K \zeta = -C(u^0 - u^N)/N.$$ 

(25)

By solving (25) for $\zeta$, the error correction (5) is carried out at the end of each period. Note that it can be shown that (25) is equivalent to $K\zeta = \Sigma_a(f_a - Ku_a)/N$. Hence, (25) is solvable if $f_a \in \text{Range}[K]$.

#### B. Effect of TP-EEC method

To make discussion on the effect of the TP-EEC method clear, we begin with a toy problem discussed in [8]

$$u + \tau \frac{du}{dt} = b(t),$$ 

(26)

where $b(t)$ is a periodic function. From (26), one can obtain the finite difference equation of the form

$$(C + 1)u^n - Cu^{n-1} = f^n,$$ 

(27)

where $C = 0 - 1 + \tau/\Delta t$. A system of equations for one period can be constructed from (27) in the same manner as (21). The slowly convergent error is assumed be $W = [1, 1, \ldots, 1]^t$. The projection matrix $P$ in (8) for this case is given by

$$P = 1 - \frac{1}{N}[1, 1, \ldots, 1]^t[I, I, \ldots, I].$$ 

(28)

It is clear from (28) that (7) results in subtraction of the mean value over $e_{\text{mean}}$ one period from $E = [e^1, e^2, \ldots, e^N]^t$, namely, $e_n = e - e_{\text{mean}}$. In other words, the EC works as an error filter. The above discussion is rationalized as follows: since the error in (27) obeys $e^0 = e^1/(1 + e)$, $E$ can be written in the form

$$E = \left[ \begin{array}{c} 1 \\ \vdots \\ -N + 1 \\ e \\ \vdots \\ 2 \\ N expensive \end{array} \right] + O(e^2),$$ 

(29)

where $e = 1$ is assumed. The first term in (29) is eliminated by (7) while the second term, which belongs to $\text{Ker}(W^tA)$, remains unchanged.

Now we consider the effect of the TP-EEC method applied to (19) with a periodic current. The projection matrix $P$ can be constructed from (8), (21) and (24) as

$$P = 1 - \frac{1}{N}[1, I, \ldots, I]^t[I, I, \ldots, I].$$ 

(30)

The error $e$ in (19) obeys $e^n = (1 + C^{-1}K)^{-1}e^{n-1}$. Assuming the spectral radius $\rho(C^{-1}K)$ is smaller than one, the error $E = [e^1, e^2, \ldots, e^N]^t$ over one period can be expanded in the form of the Neumann series as

$$E = \left[ \begin{array}{c} 1 \\ \vdots \\ -N + 1 \\ C^{-1}K \\ \vdots \\ \cdots \end{array} \right] + O(e^2),$$ 

(31)

where $e = 1$ is assumed. Hence, the effect of the TP-EEC method applied to (19) works essentially in the same way as in that applied to the toy problem mentioned above.

It is clear from the above discussions that the constant and linearly changing components in the error, which appear in the first and second terms in (31) can be simultaneously
filtered out by introducing $W$ of the form

$$W = \left[ \begin{array}{c} I_1, I_2, \ldots, I_N \end{array} \right]^T \in \mathbb{R}^{N \times 2m},$$

(32)

where $\alpha_k = 1 + 2(k-1)/(N-1), k = 1, 2, \ldots, N$. The effect of (32) has been experimentally verified in [11]. In summary, in this method, slow error components are a priori assumed to be temporary constant, linear etc. and their linear combinations with the coefficients determined by solving the correction equation (4) are subtracted from the original error.

V. NUMERICAL RESULTS

Figure 1 shows a model for verification of the discussion mentioned in the previous section, where a metallic plate, $\sigma = 5 \times 10^6$ S/m, is placed on the top of a magnetic core, $\mu_r = 2 \times 10^6$. The driving frequency is set to 200 Hz and $\Delta t = 1000$. The mean error $e_{\text{mean}}$ is computed for each period and is subtracted from the error $e^t$ at $N$-th time step. From the discussion in IV, one expects that $e^t - e_{\text{mean}}$ is approximately equal to $e^t$ in the next period. From TABLE II, it can be seen that they are in good agreement.

VI. CONCLUSION

In this paper, it has been shown that the multi-grid, deflation, EEC, IEC and AV methods are based on the decomposition of unknowns into slowly and fast converging components and error correction with the projection matrix $P$. The mathematical properties of the TP-EEC have been clarified. The experimental results are consistent with the theoretical predictions.

REFERENCES