Japan Journal of Industrial and Applied Mathematics (2020) 37:803–818
https://doi.org/10.1007/s13160-020-00417-z

ORIGINAL PAPER

A new iterative refinement for ill-conditioned linear systems based on discrete gradient

Kai Liu[1] · Jie Yang[1] · Changying Liu[2]

Received: 26 November 2019 / Revised: 17 March 2020 / Published online: 7 April 2020

Abstract
In this paper, a new iterative refinement for ill-conditioned linear systems is derived based on discrete gradient methods for gradient systems. It is proved that the new method is convergent for any initial values irrespective of the choice of the stepsize h. Some properties of the new iterative refinement are presented. It is shown that the condition number of the coefficient matrix in the linear system to be solved in every step can be improved significantly compared with Wilkinson's iterative refinement. The numerical experiments illustrate that the new method is more effective and efficient than Wilkinson's iterative refinement when dealing with ill-conditioned linear systems.

Keywords Wilkinson's iterative refinement · Ill-conditioned system of linear equations · Discrete gradient method · Gradient system

Mathematics Subject Classification 65L05 · 65L06 · 65F10

1 Introduction

Consider a linear equation of the form

$$Ax = b, \tag{1}$$

The research was supported in part by the Natural Science Foundation of China under Grant 11701271 and by the Natural Science Foundation of the Jiangsu Higher Education Institutions under Grant 16KJB110010.

✉ Jie Yang
jieyiyangwu@163.com

[1] College of Applied Mathematics, Nanjing University of Finance and Economics, Nanjing 210023, People's Republic of China

[2] School of Mathematics and Statistics, Nanjing University of Information Science and Technology, Nanjing 210044, People's Republic of China

where $A \in \mathbb{R}^{n\times n}$ is symmetric positive definite and $x, b \in \mathbb{R}^n$. The condition number of a nonsingular square matrix A with respect to a given matrix norm $\|\cdot\|$ is defined to be

$$\kappa(A) = \|A\| \cdot \|A^{-1}\|. \tag{2}$$

Linear system (1) is called ill-conditioned if the condition number $\kappa(A)$ of the coefficient matrix A is very large, in which case, small perturbations in the observation data can have large effects on the computed solutions. One approach to deal with ill-conditioned linear systems is the scaling strategy, but scaling of the equations and unknowns must proceed on a problem-by-problem basis [1, 2]. Another approach is regularization method in which a parameter is introduced in order to modify the ill-conditioned linear systems, such that it can be well-conditioned. For example, the truncated singular value decomposition [3], maximum entropy regularization [4] , and the Tikhonov regularization method [5] are widely used in this field. The regularization method requires the choice of an adequate regularization parameter [6]. An optimal regularization parameter can balance the perturbation error and the regularization error in the regularized solution.

Besides the above mentioned approaches, an iterative refinement of the solution obtained by a direct solver is also a widely-used method for solving ill-conditioned linear systems, on which we will focus ourselves in this paper. A popular method of the iterative refinement is the well-known Wilkinson's iterative refinement, which reads as follows:

$$\begin{cases} y_k = A^{-1}\left(b - Ax_k\right) \\ x_{k+1} = x_k + y_k \quad k = 0, 1, \ldots . \end{cases} \tag{3}$$

Since A is a symmetric positive definite matrix, the computation of y_k can be done by using Cholesky factorization of $A = LL^T$, where L is lower triangular matrix with positive diagonal entries. Due to rounding errors, the process (3) would not stop at the first step. In general, $x_k + y_k$ is a more accurate approximation to the solution of the linear system (1) than the approximation x_k. Wilkinson's iterative refinement performs iterations on the system whose right-hand side is the residual vector $b - Ax_k$ for successive approximations until satisfactory accuracy results. It is shown in [7] that by using a combination of 32-bit and 64-bit floating point arithmetic, the performance of Wilkinson's iterative refinement can be significantly enhanced while maintaining the 64-bit accuracy of the resulting solution. In [8], three precisions (IEEE half precision, single precision, double precision)are used in Wilkinson's iterative refinement, which makes it possible to solve $Ax = b$ to full single precision. In [9], the authors develop a GMRES (generalized minimal residual)-based iterative refinement method for solving ill-conditioning linear systems.

It is noted that Wilkinson's iterative refinement can be viewed as an explicit Euler method with stepsize $h = 1$ for solving the following continuous dynamic system

$$\begin{cases} \dfrac{dx}{dt} = A^{-1}(b - Ax) \\ x(0) = x_0. \end{cases} \tag{4}$$

The connections between iterative numerical methods in numerical linear algebra and continuous dynamic systems have been studied since 1970s [10, 11]. The dynamic system methods [12], which raised based on their connections, have become fruitfully alternative methods for solving linear systems. Wilkinson's iterative refinement can be viewed as a dynamic system methods. Dynamical systems have some superiors to discrete ones as they can obtain the convergence results more easily under weaker restrictions on the convergence theorems. This may offer a better understanding about the convergence of the dynamic system methods.

The conventional numerical solvers such as Euler method, Runge-Kutta method, linear multistep method are mainly considered for solving the continuous dynamic systems to derive the dynamic system methods. For example, in [13–15], the explicit Euler, the implicit Euler and embedded Runge-Kutta methods are used to derive new type of iterative refinement methods. In the past decades, geometric numerical integration methods for differential equations have emerged as alternatives to traditional numerical methods. Geometric numerical integration methods are such methods that can preserve certain structures such as first integrals, symplectic structure, symmetries, phase-space volume, Lyapunov functions etc. of the continuous systems [16, 17]. Normally, geometric numerical integration methods perform a better behaviour than traditional methods when solving continuous dynamic systems in a long time. The study of geometric numerical integration provides us a variety of state-of-the-art numerical tools for solving the continuous dynamic systems [18].

It is noted that in Wilkinson's iterative refinement, it still suffers from the ill conditioning of the coefficient matrix A when using the Cholesky factorization to solve y_k in every step. In this paper, a new iterative refinement is derived by applying the discrete gradient methods (see e.g., [19–21]), a tool from geometric numerical integration, to the gradient systems associated with the linear system $Ax = b$. For the new iterative refinement, a linear system with a well-conditioned coefficient matrix rather than the ill-conditioned matrix A will be solved in every step. Therefore, it avoids the adverse effects of the ill-conditioning of the matrix A to some extend.

The rest of the paper is organized as follows. In Sect. 2, discrete gradient methods for gradient systems are introduced. In Sect. 3, a new iterative refinement is derived based on discrete gradient methods. The convergence of the new iterative refinement is analysed. Some properties of the new iterative refinement are presented from a pure algebraic point of view in Sect. 4. In Sect. 5, two numerical experiments are carried out to compare between Wilkinson's iterative refinement and the new one proposed in this paper. Some concluding remarks and discussions are summarized in the last section.

2 Discrete gradient methods for gradient systems

Discrete gradient methods for ordinary differential equations were introduced by [22]. A discrete gradient $\overline{\nabla} f$ is defined as follows.

Definition 1 Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable. The function $\overline{\nabla} f : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is called a discrete gradient if it meets

$$\begin{cases} f(p) - f(q) = \overline{\nabla} f(p,q)^T (p-q) \\ \overline{\nabla} f(p,p) = \nabla f(p), \end{cases} \tag{5}$$

for all $p, q \in \mathbb{R}^n, p \neq q$.

There are three well-known discrete gradients, the coordinate increment discrete gradient, the midpoint discrete gradient and the average vector field discrete gradient, which are defined as follows.

- The coordinate increment discrete gradient by [23]:

$$\overline{\nabla}_{CI} f(p,q) = \begin{pmatrix} \frac{f(p^1,q^2,\ldots,q^k) - f(q^1,q^2,\ldots,q^k)}{p^1 - q^1} \\ \frac{f(p^1,p^2,q^3,\ldots,q^k) - f(p^1,q^2,\ldots,q^k)}{p^2 - q^2} \\ \vdots \\ \frac{f(p^1,p^2,\ldots,q^k) - f(p^1,p^2,\ldots,p^{k-1},q^k)}{p^k - q^k} \end{pmatrix}. \tag{6}$$

- The midpoint discrete gradient by [22]:

$$\overline{\nabla}_{MID} f(p,q) = \nabla f\left(\frac{p+q}{2}\right) + (p-q)\frac{f(p) - f(q) - \nabla f(\frac{p+q}{2})^T (p-q)}{\|p-q\|^2}. \tag{7}$$

- The average vector field discrete gradient ([19, 20, 24]):

$$\overline{\nabla}_{AVF} f(p,q) = \int_0^1 \nabla f(\xi p + (1-\xi)q) d\xi. \tag{8}$$

Both the midpoint discrete gradient (7) and the average vector field discrete gradient (8) are second-order approximation to the value of the gradient at the midpoint of the interval $[p, q]$.

For a given differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, we consider gradient systems of the form

$$\begin{cases} \frac{d}{dt} x(t) = -P \nabla f(x(t)) \\ x(0) = x_0 \in \mathbb{R}^n, \end{cases} \tag{9}$$

where $P \in \mathbb{R}^{n\times n}$ is symmetric positive definite matrix. Then the corresponding discrete gradient method with stepsize h for the gradient system (9) can be written as

$$\frac{x_{k+1}-x_k}{h} = -P\overline{\nabla} f(x_{k+1}, x_k). \tag{10}$$

It can be shown that the discrete gradient method (10) preserves the energy dissipation of the gradient system (9). More precisely, we have

$$\begin{aligned}\frac{1}{h}(f(x_{k+1}) - f(x_k)) &= \overline{\nabla} f(x_{k+1}, x_k)^T \frac{x_{k+1}-x_k}{h} \\ &= -\overline{\nabla} f(x_{k+1}, x_k) P \overline{\nabla} f(x_{k+1}, x_k) \le 0,\end{aligned} \tag{11}$$

which corresponds to the decay of the energy function $f(x)$ along the solution:

$$\frac{d}{dt} f(x(t)) = \nabla f(x(t))^T \frac{d}{dt} x(t) = -\nabla f(x(t))^T P \nabla f(x(t)) \le 0.$$

3 A new iterative refinement for the solution of (1)

In this section, we will derive a new iterative refinement based on discrete gradient methods. To this end, we consider the energy function

$$f(x) = \frac{1}{2} x^T A x - x^T b, \tag{12}$$

where $A \in \mathbb{R}^{n\times n}$ and $b \in \mathbb{R}^n$.

For the energy function (12), both the midpoint discrete gradient (7) and the average vector field discrete gradient (8) give the same discrete gradient $\overline{\nabla} f(x, y) = A\frac{x+y}{2} - b$. Then, the corresponding discrete gradient method reads

$$\frac{x_{k+1}-x_k}{h} = -P\left(A\frac{x_{k+1}+x_k}{2} - b\right), \quad k = 0, 1, 2, \dots, \tag{13}$$

which is equivalent to the following iterative refinement

$$\left(I + \frac{h}{2}PA\right)(x_{k+1} - x_k) = hP(b - Ax_k), \quad k = 0, 1, 2, \dots, \tag{14}$$

or equivalently

$$\begin{cases} \left(\frac{1}{h}P^{-1} + \frac{1}{2}A\right) y_k = b - Ax_k \\ x_{k+1} = x_k + y_k, \quad k = 0, 1, 2 \dots . \end{cases} \tag{15}$$

To show that the iterative refinement (14) gives sequences $\{x_s\}_{s=0}^{+\infty}$ that converge to the unique solution of the linear system (1), we recall some definitions.

Definition 2 A function $f : \mathbb{R}^n \to \mathbb{R}$ is called

- convex, if and only if, for all $x, y \in \mathbb{R}^n$ and $\lambda \in [0, 1]$,
$$f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda) f(y);$$
- strictly convex, if and only if, for all $x, y \in \mathbb{R}^n$, $x \neq y$ and $\lambda \in (0, 1)$,
$$f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda) f(y);$$
- coercive, if and only if, $f(x_k) \to \infty$ for $\|x_k\| \to \infty$.

Lemma 1 *The energy function* $f(x) = \frac{1}{2}x^T Ax - x^T b$ *is strictly convex and coercive.*

Proof By some manipulation, it can be verified that

$$f(\lambda x + (1-\lambda)y) = \lambda f(x) + (1-\lambda) f(y) + \frac{1}{2}\lambda(\lambda - 1)(x-y)^T A(x-y).$$

Since $\lambda \in (0, 1)$ and A is positive definite, we have

$$f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda) f(y) \tag{16}$$

for $x \neq y$. Therefore, the energy function $f(x)$ is strictly convex.

The coerciveness of the function $f(x)$ can be proved as follows:

$$f(x) = \frac{1}{2}x^T Ax - x^T b \ge \frac{1}{2}\lambda^A_{min}\|x\|_2^2 - \|x\|_2 \cdot \|b\|_2 \to \infty, \quad as \quad \|x\|_2 \to \infty. \tag{17}$$

Here, $\lambda^A_{min} > 0$ is the smallest eigenvalue of the matrix A and $\|\cdot\|_2$ is the Euclidean norm of $\mathbb{R}^n$. □

Now we are in a position to present the main theorem concerning the convergence of the iterative refinement (14).

Theorem 1 *The sequence* $\{x_n\}_{n=0}^{\infty}$ *generated by the iterative refinement* (14) *converges to the unique solution* $x^* = A^{-1}b$ *of the linear system* (1) *for any initial value* x_0.

Proof First of all, one should note that the unique solution $x^* = A^{-1}b$ of the linear system (1) is the unique minimizer to the unconditioned optimization problem

$$\min_{x \in \mathbb{R}^n} f(x).$$

As a matter of fact,

$$f(x) = f(x^*) + \frac{1}{2}(x - x^*)^T A(x - x^*) \ge f(x^*), \quad \forall x \in \mathbb{R}^n.$$

Therefore, x^* is a minimizer of $f(x)$.

Assume that the function f has two different minimizers called x^* and $\overline{x}^*$, that is $x^* \neq \overline{x}^*$ and $f(x^*) = f(\overline{x}^*) \leq f(x)$ for all $x \in \mathbb{R}^n$. Since f is strictly convex, picking $\lambda \in (0, 1)$ and we obtain

$$f(\lambda x^* + (1-\lambda)\overline{x}^*) < \lambda f(x^*) + (1-\lambda) f(\overline{x}^*) = \lambda f(x^*) + (1-\lambda) f(x^*) = f(x^*).$$

Since $\lambda x^* + (1-\lambda)\overline{x}^* \in \mathbb{R}^n$, this is a contradiction to x^*(or $\overline{x}^*$, respectively) being a minimizer. Hence x^* is the unique minimizer of the unique minimizer to the unconditioned optimization problem $\min_{x \in \mathbb{R}^n} f(x)$.

It can be seen from (11) that

$$f(x_{n+1}) - f(x_n) = -h\overline{\nabla} f(x_{n+1}, x_n)^T P \overline{\nabla} f(x_{n+1}, x_n) \leq 0. \tag{18}$$

Therefore,

$$f(x_0) \geq f(x_1) \geq \cdots \geq f(x_k) \geq f(x_{k+1}) \geq \cdots \geq f(x^*), \tag{19}$$

hence the limit $\lim_{t \to \infty} f(x_k)$ exists. Meanwhile,

$$\frac{1}{h}(f(x_k) - f(x_{k+1})) = \overline{\nabla} f(x_{k+1}, x_k)^T P \overline{\nabla} f(x_{k+1}, x_k) \geq \lambda_{min}^P \|\overline{\nabla} f(x_{k+1}, x_k)\|_2^2 \tag{20}$$

with $\lambda_{min}^P > 0$ the minimum eigenvalue of the matrix P. By summing equations (20) from $k = 0$ to $k = m-1$, $m > k$, we obtain

$$\sum_{k=0}^{m-1} \|\overline{\nabla} f(x_{k+1}, x_k)\|_2^2 \leq \frac{1}{h\lambda_{min}^P}(f(x_0) - f(x_m)) \tag{21}$$

and thus

$$\sum_{k=0}^{\infty} \|\overline{\nabla} f(x_{k+1}, x_k)\|_2^2 < \infty. \tag{22}$$

From equation (10), we have

$$\overline{\nabla} f(x_{k+1}, x_k) = P^{-1} \frac{x_k - x_{k+1}}{h}.$$

Substituting it into (11) yields

$$h(f(x_k) - f(x_{k+1})) = (x_{k+1} - x_k)^T P^{-1} (x_{k+1} - x_k) \geq (\lambda_{min}^P)^{-1} \|x_k - x_{k+1}\|_2. \tag{23}$$

Summing equations (23) from $k = 0$ to $k = m-1$, $m > k$ gives

$$\sum_{k=0}^{m-1} \|x_k - x_{k+1}\|_2^2 \leq h\lambda_{min}^P (f(x_0) - f(x_m)) \tag{24}$$

and thus

$$\sum_{k=0}^{\infty} \|x_{k+1} - x_k\|_2^2 < \infty. \tag{25}$$

Therefore, by (22) and (25), we have

$$\lim_{k\to\infty}(x_{k+1} - x_k) = \lim_{k\to\infty} \overline{\nabla} f(x_{k+1}, x_k) = 0. \tag{26}$$

Since

$$\nabla f(x_k) = \overline{\nabla} f(x_{k+1}, x_k) + \frac{1}{2}A(x_k - x_{k+1}), \tag{27}$$

then

$$\lim_{k\to\infty} \nabla f(x_k) = \lim_{k\to\infty} \overline{\nabla} f(x_{k+1}, x_k) + \lim_{k\to\infty} \frac{1}{2}A(x_k - x_{k+1}) = 0. \tag{28}$$

Moreover, since the set defined by $V_f(x_0) = \{x \in \mathbb{R}^n | f(x) \leq f(x_0)\}$ is bounded, due to the coerciveness of $f(x)$, the sequence $\{x_k\}_{k=0}^{\infty}$ has at least one accumulation point x^{**} by the Bolzano–Weierstrass theorem. For a subsequence $\{x_{k_l}\}_{l=0}^{\infty}$ with $\lim_{l\to\infty} x_{k_l} = x^{**}$, we have

$$0 = \lim_{l\to\infty} \nabla f(x_{k_l}) = \nabla f(x^{**}) = Ax^{**} - b$$

due to the continuity of ∇f. Hence, $x^{**} = A^{-1}b$ is the minimizer to $f(x)$. According to the uniqueness of the minimizer, all accumulation point of the sequence $\{x_k\}_{k=0}^{\infty}$, which are minimizers, must be identical. Therefore the sequence converges to the unique minimizer $x^* = A^{-1}b$. □

Remark 1 Since the energy function (12) is quadratic, the midpoint discrete gradient method or the average vector field discrete gradient method coincides with the traditional midpoint method, which is a very typical implicit Runge-Kutta method. However, it can be observed from the above discussion that the discrete gradient method preserves the decay of the energy function, whose unique minimizer is exactly the solution to the linear system $Ax = b$. The analysis of the convergence of the iteration (15) based on discrete gradient method clearly provides more insight into the connection between the iterative method and the continuous dynamic systems than that based on typical Runge-Kutta methods.

4 Some properties of the new iterative refinement

The previous section analyzed the convergence of the iterative refinement based the connection between the discrete and continuous systems. In this section, we will present some properties of the new iterative refinement from the pure algebraic point of view. The following lemma is crucial for the proof of the properties.

Lemma 2 *Let $A, P \in \mathbb{R}^{n\times n}$ be two symmetric positive definite matrices. Then the eigenvalues of PA satisfy $\lambda(PA) > 0$.*

Proof Since matrices A, P are symmetric positive definite, then there exists two invertible matrices M and N such that

$$P = M^T M, A = N^T N.$$

Therefore, we have

$$NPAN^{-1} = (MN^T)^T MN^T, \tag{29}$$

which implies that the matrix PA is similar to the matrix $(MN^T)^T MN^T$. Since M and N are invertible, thus $(MN^T)^T MN^T$ is symmetric positive definite. Therefore, $\lambda(PA) > 0$ holds. □

The iterative refinement (15) can be regarded as a stationary iterative method with the iterative matrix

$$J = \left(I + \frac{h}{2}PA\right)^{-1}\left(I - \frac{h}{2}PA\right). \tag{30}$$

The following theorem presents a theoretical result concerning the spectral radius of the iteration matrix J.

Theorem 2 *Assume that $\mu_i\,(i = 1, 2, \ldots, n)$ are eigenvalues of $n \times n$ matrix PA and*

$$0 < \mu_1 \leqslant \mu_2 \leqslant \cdots \leqslant \mu_n,$$

then the spectral radius of the iterative method (15) is

$$\rho(J) = \max\left(\left|\frac{1 - \frac{h}{2}\mu_1}{1 + \frac{h}{2}\mu_1}\right|, \left|\frac{1 - \frac{h}{2}\mu_n}{1 + \frac{h}{2}\mu_n}\right|\right) < 1. \tag{31}$$

Therefore, the asymptotic rate of convergence of (15) is

$$R(J) = -\ln \rho(J) = \min\left(-\ln\left|\frac{1 - \frac{h}{2}\mu_1}{1 + \frac{h}{2}\mu_1}\right|, -\ln\left|\frac{1 - \frac{h}{2}\mu_n}{1 + \frac{h}{2}\mu_n}\right|\right).$$

Proof Since $\mu_i (i = 1, 2, \ldots, n)$ are the eigenvalues of PA, we obtain that

$$\frac{1 - \frac{h}{2}\mu_i}{1 + \frac{h}{2}\mu_i}, \quad i = 1, 2, \ldots, n \tag{32}$$

are the eigenvalues of $\left(I+\frac{h}{2}PA\right)^{-1}\left(I-\frac{h}{2}PA\right)$, hence

$$\rho(J)=\rho\left[\left(I+\frac{h}{2}PA\right)^{-1}\left(I-\frac{h}{2}PA\right)\right]=\max_{1\leqslant i\leqslant n}\left|\frac{1-\frac{h}{2}\mu_i}{1+\frac{h}{2}\mu_i}\right|. \tag{33}$$

Because of the strictly monotone decreasing of the function $\frac{1-x}{1+x}$ on $[0,\infty)$, the maximum and minimum of $\frac{1-\frac{h}{2}\mu_i}{1+\frac{h}{2}\mu_i}, i=1,\ldots,n$ are $\frac{1-\frac{h}{2}\mu_1}{1+\frac{h}{2}\mu_1}$ and $\frac{1-\frac{h}{2}\mu_n}{1+\frac{h}{2}\mu_n}$, respectively. Thus, we have

$$\rho(J)=\max\left(\left|\frac{1-\frac{h}{2}\mu_1}{1+\frac{h}{2}\mu_1}\right|,\left|\frac{1-\frac{h}{2}\mu_n}{1+\frac{h}{2}\mu_n}\right|\right)<1.$$

And the asymptotic convergence rate of (15) is

$$R(J)=-\ln\rho(J)=\min\left(-\ln\left|\frac{1-\frac{h}{2}\mu_1}{1+\frac{h}{2}\mu_1}\right|,-\ln\left|\frac{1-\frac{h}{2}\mu_n}{1+\frac{h}{2}\mu_n}\right|\right).$$

□

Theorem 2 also shows the convergence of the iterative refinement (15) since $\rho(J)<1$.

In Wilkinson's iterative refinement, one must solve a linear system with coefficient matrix A for every step. Thus, it still suffers from the ill-conditioning of the matrix A for every step. In contrast, for the new iterative refinement, the coefficient matrix is $I+\frac{h}{2}PA$. The simplest choice of P would be the identity matrix I. In this case, the condition number of $I+\frac{h}{2}A$ can be shown to be smaller than that of A.

Theorem 3 *If the eigenvalues of the matrix A satisfy $0<\lambda_1\leqslant\lambda_2\leqslant\cdots\leqslant\lambda_n$, then we have*

$$\kappa_2\left(I+\frac{h}{2}A\right)\leq\kappa_2(A) \tag{34}$$

with $\kappa_2(A)=\|A\|_2\cdot\|A^{-1}\|_2$ called the spectral condition numbers of A.

Proof The spectral condition numbers of A and $I+\frac{h}{2}A$ are

$$\kappa_2(A) = \frac{\lambda_n}{\lambda_1}$$

and

$$\kappa_2\Big(I + \frac{h}{2}A\Big) = \frac{1 + \frac{h}{2}\lambda_n}{1 + \frac{h}{2}\lambda_1},$$

respectively. Therefore,

$$\kappa_2\Big(I + \frac{h}{2}A\Big) - \kappa_2(A) = \frac{2\lambda_1 - 2\lambda_n}{\lambda_1(2 + h\lambda_1)},$$

Since $\lambda_1 \leq \lambda_n$, we get

$$\kappa_2\Big(I + \frac{h}{2}A\Big) \leq \kappa_2(A).$$

□

Remark 2 Since we have

$$\kappa_2\Big(I + \frac{h}{2}A\Big) - \kappa_2(A) = \frac{2\lambda_1 - 2\lambda_n}{\lambda_1(2 + h\lambda_1)} = \frac{1 - \frac{\lambda_n}{\lambda_1}}{1 + \frac{h}{2}\lambda_1} \approx -\frac{\kappa_2(A)}{1 + \frac{h}{2}\lambda_1},$$

the condition number of the matrix could be reduced significantly if the stepsize h is chosen such that $\frac{h}{2}\lambda_1$ is very small. In practice, the ill-conditioning of many matrices is caused by the extremely small magnitude of λ_1, in which case, the improvement of the condition number is obvious.

Remark 3 There are many choices of the matrix P. In this paper, we also consider the case $P = D^{-1}$, where $D = diag(a_{11}, a_{22}, \ldots, a_{nn})$. Since $a_{ii} > 0, i = 1, \ldots, n$ because of the symmetric positive definiteness of the matrix A, P is symmetric positive definite. In this case, the iterative refinement can be formulated as

$$\begin{cases} \Big(\frac{1}{h}D + \frac{1}{2}A\Big)y_k = (b - Ax_k), \\ x_{k+1} = x_k + y_k. \quad k = 0, 1, \ldots. \end{cases} \tag{35}$$

5 Numerical comparison

In this section, some numerical examples are performed to show the effectiveness of the proposed iterative refinement. All experiments are executed by using the MATLAB programming package with double precision by default. In all the numerical examples, Wilkinson's iterative refinement (3) and the proposed iterative refinement (15) with $P = I$ and $P = D^{-1}$ are applied to solve the linear system. The Cholesky factorisation is used to solve the linear system involved in every step for all three iterative refinements. The stepsize h in the new iterative refinements is chosen as $h = 2$.

Problem 1 Consider the notorious ill-conditioned linear system of algebraic equations with coefficient matrices of $n \times n$ Hilbert matrices $H_n = (h_{ij})_{n\times n}$

$$H_n x = b, \tag{36}$$

where

$$h_{ij} = \frac{1}{i+j-1}, \quad i,j = 1,2,\ldots,n.$$

Two cases of the right-hand side term b are considered :

$$b_i = \sum_{k=1}^{n} h_{ik}, \quad i = 1,2,\ldots,n$$

and

$$b_i = \sum_{k=1}^{n} k \times h_{ik}, \quad i = 1,2,\ldots,n,$$

where the exact solutions of the system (36) are given by $x^* = [1,1,\ldots,1]^T$ and $x^* = [1,2,\ldots,n]^T$, respectively.

Letting $n = 20, 50, 70$ and 100, the spectral condition numbers for Wilkinson's iterative refinement and the iterative refinement proposed in this paper are shown in Table 1, and the corresponding relative error estimates of computed solution $\|(x_{1000} - x^*)./x^*\|_\infty$ are listed in Tables 2 and 3, where the notation ./ is interpreted componentwisely.

Table 1 Spectral condition numbers $k(H_n)$ with $h = 2$

n	20	50	70	100
$\kappa(H_n)$	2.60×10^{18}	4.14×10^{18}	1.24×10^{19}	6.98×10^{19}
$\kappa\left(\frac{1}{h}I + \frac{1}{2}H_n\right)$	2.91	3.08	3.13	3.18
$\kappa\left(\frac{1}{h}D + \frac{1}{2}H_n\right)$	9.71×10	2.58×10^2	3.67×10^2	5.31×10^2

Table 2 The relative error estimate of computed solution $\|(x_{1000} - x^*)./x^*\|_\infty$ for Problem 1 with 1000 steps (exact solution $x^* = [1, 1, \ldots, 1]^T$)

n	20	50	70	100
Wilkinson's method	NaN	NaN	NaN	NaN
New method with $P = I$	2.04×10^{-2}	2.29×10^{-2}	2.29×10^{-2}	2.32×10^{-2}
New method with $P = D^{-1}$	5.36×10^{-3}	5.27×10^{-3}	5.49×10^{-3}	5.34×10^{-3}

Table 3 The relative error estimate of computed solution $\|(x_{1000} - x^*)./x^*\|_\infty$ for Problem 1 with 1000 steps (exact solution $x^* = [1, 2, \ldots, n]^T$)

n	20	50	70	100
Wilkinson's method	NaN	NaN	NaN	NaN
New method with $P = I$	1.43×10^{-1}	4.04×10^{-1}	7.35×10^{-1}	1.58
New method with $P = D^{-1}$	6.45×10^{-2}	1.22×10^{-1}	1.57×10^{-1}	1.93×10^{-1}

Problem 2 Consider the ill-conditioned linear system of algebraic equations with $n \times n$ matrices $A_n = (a_{ij})_{n \times n}$

$$A_n x = b, \tag{37}$$

where

$$a_{ij} = \begin{cases} 1 & i \neq j \\ 1 + \varepsilon^2 & i = j \end{cases} \tag{38}$$

and $\varepsilon = 3 \times 10^{-7}$. Set

$$b_i = \sum_{k=1}^{n} a_{ik}, \quad i = 1, 2, \ldots, n,$$

the exact solution of the system(37) is given by

$$x^* = [1, 1, \ldots, 1]^T.$$

Letting $n = 20, 50, 70$ and 100, the spectral condition numbers for Wilkinson's iteration and new method proposed in this paper are shown in Table 3, and the corresponding error estimates of computed solution are listed in Table 4. We must point out that both the new iterative refinements and Wilkinson's iterative refinement fail to give satisfactory computed solutions with significant figures for the system (37) with the exact solution $x^* = [1, 2, \ldots, n]^T$.

Table 4 The relative error estimate of computed solution $\|(x_{1000} - x^*)./x^*\|_\infty$ for Problem 2 with 1000 steps

n	20	50	70	100
Wilkinson's method	5.14×10^{-2}	9.63×10^{-2}	6.56×10^{-2}	5.97×10^{-2}
New method with $P = I$	6.06×10^{-12}	8.58×10^{-12}	1.91×10^{-12}	2.06×10^{-9}
New method with $P = D^{-1}$	6.09×10^{-12}	9.36×10^{-12}	1.32×10^{-12}	2.02×10^{-9}

From Tables 1 and 5, we can see that the condition number of the coefficient matrix in new iterative refinement is much less than that in Wilkinson's iterative refinement. Therefore, comparing with the ill-conditioned coefficient matrix in Wilkinson's iterative refinement, the coefficient matrix of the linear system required to be solved in every step for the new iterative refinement is well conditioned.

It can be observed from Tables 2 and 3 that Wilkinson's iterative refinement gives the computed solution without any significant figures for the linear system with Hilbert matrix within 1000 steps. On the other hand, the new iterative refinements produce much satisfactory results. Moreover, the new iterative refinements give a much better approximation of the solution than Wilkinson's iterative refinement in Problem 2. Generally speaking, the new iterative refinement proposed in the paper performs better that Wilkinson's iterative refinement.

Remark 4 Under the same settings, we also solve the two problems by iterative refinements with two precisions. More precisely, the double-precision arithmetic is used for the iterative refinements while quadruple-precision arithmetic is used for computing the residual vector.

For Problem 1, the accuracy of the solution computed by the new iterative refinements is improved compared with the case of one precision. For example, the error of the computed solution for the Hilbert linear system with exact solution $x^* = [1, 2, \ldots, n]^T$ decreases from 10^{-1} to 10^{-3}. However, Wilkinson's iterative refinement, even though with two precisions, fails again. The reason is that generally speaking, when it comes to Wilkinson's iterative refinement, the condition number of the coefficient matrix must not be too large and normally it is assumed that $\kappa(A) < 1/\varepsilon$, where ε is the machine precision. However, in Problem 1,the condition number of the coefficient matrix is greater that $1/\varepsilon$ for double precision with $\varepsilon = 2.22 \times 10^{-16}$.

Table 5 Spectral condition numbers $k(A_n)$ with $h = 2$

n	20	50	70	100
$\kappa(A_n)$	2.25×10^{14}	2.04×10^{15}	4.25×10^{16}	1.33×10^{16}
$\kappa\left(\frac{1}{h}I + \frac{1}{2}A_n\right)$	2.10×10	5.10×10	7.10×10	1.01×10^2
$\kappa\left(\frac{1}{h}D + \frac{1}{2}A_n\right)$	2.10×10	5.10×10	7.10×10	1.01×10^2

For Problem 2, the condition number of the coefficient matrices do not exceed $1/\varepsilon$. Both the new iterative refinements and Wilkinson's iterative refinement with two precisions produce computed solutions with accuracy up to 10^{-16}.

6 Conclusions

Discrete gradient method, a well-known tool from geometric numerical integration, can preserve the energy dissipation for gradient systems. When applied to the systems associated with the linear system $Ax = b$, a new iterative refinement is obtained, the convergence of which can be shown based on the property of dissipation preservation. We have illustrated theoretically and numerically that the new iterative refinement is an improvement of the traditional Wilkinson's iterative refinement.

It should be noted that the choice of stepsize h is very subtle in the new iterative refinement. In this work, a stepsize with a moderate magnitude is chosen, i.e., $h = 2$. Generally speaking, the optimal stepsize is not attainable in practice. However, the new method still provides us a better alternative of the traditional Wilkinson's iterative refinement.

Acknowledgements The authors sincerely thank the editors and referees for their kind and valuable comments of revision which improved the presentation of the paper.

References

1. Golub, G.H., Van Loan, C.F.: Matrix Computations. The John Hapkins University Press, Baltimore (1996)
2. Volokh, K.Y., Vilnay, O.: Pin-pointing solution of ill-conditioned square systems of linear equations. Appl. Math. Lett. **13**, 119–124 (2000)
3. Varah, J.M.: On the numerical solution of ill-conditioned linear system with applications to ill-posed problems. SIAM J. Numer. Anal. **10**(2), 257–267 (1973)
4. Hansen, P.: Regularization tools: a MATLAB package for analysis and solution of discrete ill-posed problems. Numer. Algorithms **6**, 1–35 (1994)
5. Tikhonov, A.N., Arsenin, V.Y.: Solutions of Ill-posed Problems. Winston and Sons, Washignton D.C (1977)
6. Liu, C.-S., Atluri, S.N.: An iterative method using an optimal desent vector, for solving an ill-conditioned system $Bx = b$, better and faster than the conjugate gradient method. Comput. Model. Eng. Sci. **80**(4), 275–298 (2011)
7. Buttari, A., Dongarra, J., Langou, J., Langou, J.L., Luszczek, P.: Mixed precision iterative refinement techniques for the solution of dense linear systems. Int. J. High Perform. C **21**(4), 457–466 (2007)
8. Carson, E., Higham, N.J.: Accelerating the solution of linear systems by iterative refinement in three precisions. SIAM J. Sci. Comput. **40**(2), A817–A847 (2018)
9. Carson, E., Higham, N.J.: A new analysis of iterative refinement and its application to accurate solution of ill-conditioned sparse linear systems. SIAM J. Sci. Comput. **39**(6), A2834–A2856 (2017)
10. Chu, M.T.: On the continuous realization of iterative processes. SIAM Rev. **30**, 375–387 (1988)
11. Chu, M.T.: Linear algebra algorithms as dynamical systems. Acta Numer. **19**, 1–86 (2008)
12. Ramm, A.G.: Dynamical systems method for solving operator equations. Commun. Nonlinear Sci. Numer. Simul. **9**, 383–402 (2004)

13. Wu, X., Shao, R., Zhu, Y.: New iterative improvement of a solution for an ill-conditioned system of linear equations based on a linear dynamic system. Comput. Math. Appl. **44**, 1109–1116 (2002)
14. Wu, X., Wang, Z.: A new iterative refinement with roundoff error analysis. Numer. Linear Algebr. **18**, 275–282 (2011)
15. Wu, X., Fang, Y.: Wilkinson's iterative refinement of solution with automatic step-size control for linear system of equations. Appl. Math. Comput. **193**, 506–513 (2007)
16. Hairer, E., Lubich, C., Wanner, G.: Geometric numerical integration: structure-preserving algorithms for ordinary differential equations. Springer Science and Business Media **31**(2) (2006)
17. McLachlan, R.I., Quispel, G.R.W.: Splitting methods. Acta Numer. **11**, 341–434 (2002)
18. Miyatake, Y., Sogabe, T., Zhang, S.: On the equivalence between SOR-type methods for linear systems and the discrete gradient methods for gradient systems. J. Comput. Appl. Math. **342**, 58–69 (2018)
19. McLachlan, R.I., Quispel, G.R.W., Robidoux, N.: Geometric integration using discrete gradient. R. Soc. Lond. Philos. Trans. Ser. A Math. Phys. Eng. Sci. **357**, 1021–1045 (1999)
20. Quispel, G.R.W., McLaren, D.I.: A new class of energy-preserving numerical integration methods. J. Phys. A **41**, 045206 (2008)
21. Quispel, G.R.W., Turner, G.S.: Discrete gradient methods for solving ODEs numerically while preserving a first integral. J. Phys. A **29**, L341–L349 (1996)
22. Gonzalez, O.: Time integration and discrete Hamilton systems. J. Nonlinear Sci. **6**, 449–467 (1996)
23. Itoh, T., Abe, K.: Hamiltonian-conserving discrete canonical equations based on variational difference quotients. J. Comput. Phys. **76**, 85–102 (1988)
24. Harten, A., Lax, P.D., van Leer, B.: On upstream differencing and Godunov-type schemes for hyperbolic conservation laws. SIAM Rev. **25**, 35–61 (1983)