NUMERICAL ANALYSIS OF A FIRST-ORDER COMPUTATIONAL ALGORITHM FOR REACTION-DIFFUSION EQUATIONS VIA THE PRIMAL-DUAL HYBRID GRADIENT METHOD

SHU LIU, XINZHE ZUO, STANLEY OSHER, AND WUCHEN LI

Abstract. In [35], a first-order optimization algorithm has been introduced to solve time-implicit schemes of reaction-diffusion equations. In this research, we conduct theoretical studies on this first-order algorithm equipped with a quadratic regularization term. We provide sufficient conditions under which the proposed algorithm and its time-continuous limit converge exponentially fast to a desired time-implicit numerical solution. We show both theoretically and numerically that the convergence rate is independent of the grid size, which makes our method suitable for large scale problems. The efficiency of our algorithm has been verified via a series of numerical examples conducted on various types of reaction-diffusion equations. The choice of optimal hyperparameters as well as comparisons with some classical root-finding algorithms are also discussed in the numerical section.

1. Introduction

Reaction-diffusion equations (RD) are well-known time-dependent partial differential equations (PDEs). They are originally used to model the density evolution of chemical systems with local reaction processes in which substances get transformed, and diffusion processes in which the substances get spread over. Since the same type of equations describe many systems, the RD equation finds its applications in broad scientific areas. This includes the study of phase-field models in which the Allen-Cahn and the Cahn-Hilliard equations [1, 3] are used to depict the development of microstructures of multiple materials; the research on the evolution of species distribution in ecology system [39]; the study of the reaction processes of multiple chemicals [42, 40]; and the modeling & prediction of crimes [45].

Time-implicit schemes are often used when solving RD equations numerically. This is because in simulations, explicit or semi-explicit schemes are often encountered with Courant–Friedrichs–Lewy (CFL) conditions, under which the time step size is restricted to be very small. Conversely, employing time-implicit schemes allows for the use of larger time step sizes, leading to a more efficient computation of the equilibrium state in RD equations. Moreover, computing RD equations with a weak diffusion and a strong reaction term is of great interest to the computational math community. The performance of explicit and semi-implicit schemes could be unstable under these circumstances. However, it has been shown that implicit schemes still work very well on these models [50, 35]. In addition, time-implicit schemes are also known to be energy-stable [50].

In a recent work [35], the primal-dual hybrid gradient (PDHG) algorithm which is an easy-to-implement optimization algorithm, has been used for computing the time-implicit solution of RD equations. The PDHG algorithm (1.6) is a first-order optimization algorithm with tunable hyperparameters. Notably, it does not require computing the inverse of the Jacobian matrix in the time-implicit scheme. It converges robustly regardless of the choice of the initial value, which is a

Key words and phrases. Reaction diffusion equations; Time-implicit schemes; Primal-dual hybrid gradient algorithm; Lyapunov analysis.

S. Liu and X. Zuo are partially funded by AFOSR YIP award No. FA9550-23-1-0087. S. Liu, X. Zuo, and S. Osher are partially funded by AFOSR MURI FA9550-18-502 and ONR N00014-20-1-2787. W. Li's work is supported by AFOSR YIP award No. FA9550-23-1-0087, NSF DMS-2245097, and NSF RTG: 2038080.

key distinction from many classical methods, such as Newton's methods. This property makes the PDHG algorithm easy to implement and computationally efficient for solving nonlinear equations. Another motivating feature of the PDHG method is that it allows for the design of customized preconditioning matrices based on the structure of the specific RD equation, resulting in a notable grid-size-independent convergence rate throughout the algorithm.

Nevertheless, the prototype PDHG algorithm presented in [35] faces theoretical and practical challenges. The time-implicit scheme results in a nonlinear equation, and the PDHG algorithm introduces nonlinear coupling in both the primal and dual variables. In addition, there is a lack of convergence analysis for the proposed PDHG algorithm. Furthermore, the nonlinearity inherent in RD equations poses a challenge in resolving the optimal choice of hyperparameters. In this paper, we provide the convergence study of the PDHG algorithm for computing the time-implicit scheme of RD equations. We also present a series of numerical experiments on the choices of hyperparameters.

Let us consider the general form of the RD equation on a region $\Omega \subset \mathbb{R}^d$ with prescribed boundary (e.g., periodic, Neumann, or Dirichlet) and initial conditions.

$$\frac{\partial u(x,t)}{\partial t} = -\mathcal{G}(a\mathcal{L}u(x,t) + bf(u(x,t))), \quad x \in \Omega, \quad u(\cdot,0) = u_0(\cdot), \tag{1.1}$$

Here we assume $\mathcal{L}, \mathcal{G}$ are self-adjoint, non-negative definite linear operators. $f(\cdot)$ is the reaction term (usually nonlinear). $a \geq 0$ is the diffusion coefficient. And $b \geq 0$ is the reaction coefficient. To compute the numerical solution of (1.1), we adopt the following time-implicit scheme with a time step size $h_t > 0$ and solve for the numerical solution on N_t intervals:

$$\frac{u^{t+1} - u^t}{h_t} = -\mathcal{G}(a\mathcal{L}u^{t+1} + bf(u^{t+1})), \quad 0 \leq t \leq N_t - 1. \tag{1.2}$$

Assume that at each time step, the numerical solution u^t belongs to a certain Hilbert space $\mathcal{X}$ with an inner product $(\cdot,\cdot)$. Let us denote $\boldsymbol{u} = [u^1, \ldots, u^t, \ldots, u^{N_t}]^\top \in \mathcal{X}^{N_t}$. Define the function $\mathcal{F}(\cdot) : \mathcal{X}^{N_t} \to \mathcal{X}^{N_t}$ as

$$\mathcal{F}(\boldsymbol{u}) = [\ldots, u^{t+1} - u^t + h_t\mathcal{G}(a\mathcal{L}u^{t+1} + bf(u^{t+1})), \ldots]^\top_{0 \leq t \leq N_t - 1}. \tag{1.3}$$

Then, solving the time-implicit scheme (1.2) is equivalent to obtaining the root of the problem $\mathcal{F}(\boldsymbol{u}) = 0$.

We now reformulate the time-implicit scheme (1.2) as an inf-sup problem with a tunable parameter ϵ>0 following the treatment in [56]

$$\inf_{\boldsymbol{u} \in \mathcal{X}^{N_t}} \sup_{\boldsymbol{p} \in \mathcal{X}^{N_t}} (\boldsymbol{p}, \mathcal{F}(\boldsymbol{u})) - \frac{\epsilon}{2}\|\boldsymbol{p}\|^2_{\mathcal{X}^{N_t}}. \tag{1.4}$$

Here we write $\boldsymbol{p} = [p_1, \ldots, p_t, \ldots, p_{N_t}] \in \mathcal{X}^{N_t}$. Compared with the saddle point scheme considered in [35], a quadratic regularization term is introduced in (1.4) to enhance the performance of the proposed algorithm both theoretically and numerically. It is not hard to verify that (1.4) is equivalent to the residue-minimizing problem $\inf_{\boldsymbol{u}} \frac{1}{2\epsilon}\|\mathcal{F}(\boldsymbol{u})\|^2_{\mathcal{X}^{N_t}}$, and we further point out that the saddle point of this inf-sup problem (1.4) exists and solves $\mathcal{F}(\boldsymbol{u}) = 0$ whenever the root-finding problem admits a unique solution.

As demonstrated in [35], we deal with the inf-sup saddle problem by applying the primal-dual hybrid gradients (PDHG) algorithm [7, 54]. We further substitute the proximal step of variable $\boldsymbol{u}$ with an explicit update to obtain

$$\begin{aligned} \boldsymbol{p}_{n+1} =& \frac{1}{1+\epsilon\tau_P}\left(\boldsymbol{p}_n + \tau_P\mathcal{F}(\boldsymbol{u}_n)\right), \\ \widetilde{\boldsymbol{p}}_{n+1} =& \boldsymbol{p}_{n+1} + \omega(\boldsymbol{p}_{n+1} - \boldsymbol{p}_n), \\ \boldsymbol{u}_{n+1} =& \boldsymbol{u}_n - \tau_U D\mathcal{F}(\boldsymbol{u}_n)^*\widetilde{\boldsymbol{p}}_{n+1}. \end{aligned} \tag{1.5}$$

Here $\omega > 0$ is the extrapolation coefficient, and $\tau_P, \tau_U > 0$ are PDHG step sizes. $D\mathcal{F}(\boldsymbol{u})$ is a linear operator on $\mathcal{X}^{N_t}$, which denotes the Fréchet derivative of $\mathcal{F}(\cdot)$ at $\boldsymbol{u}$. $D\mathcal{F}(\boldsymbol{u})^*$ is the adjoint operator of $D\mathcal{F}(\boldsymbol{u})$ on $\mathcal{X}^{N_t}$. It is not hard to verify that the equilibrium state of PDHG scheme (1.5) is the desired $(\boldsymbol{u}_*, 0)$ with $\mathcal{F}(\boldsymbol{u}_*) = 0$ whenever $D\mathcal{F}(\boldsymbol{u})^*$ is invertible for arbitrary $\boldsymbol{u} \in \mathcal{X}^{N_t}$.

The PDHG algorithm (1.5) converges slowly when $\mathcal{F}(\cdot)$ possesses a large condition number. To improve the convergence speed, it is necessary to consider preconditioning $\mathcal{F}(\cdot)$. We consider an invertible linear operator $\mathfrak{M} : \mathcal{X}^{N_t} \to \mathcal{X}^{N_t}$, where $\mathfrak{M}$ is extracted from the linear part of $\mathcal{F}(\cdot)$. Then we introduce the preconditioned functional $\widehat{\mathcal{F}}(\boldsymbol{u}) = \mathfrak{M}^{-1}\mathcal{F}(u)$. We apply the PDHG algorithm (1.5) to $\widehat{\mathcal{F}}(u) = 0$ to obtain

$$\begin{aligned} \boldsymbol{p}_{n+1} =& \frac{1}{1+\epsilon\tau_P}(\boldsymbol{p}_n + \tau_P \mathfrak{M}^{-1}\mathcal{F}(\boldsymbol{u}_n)), \\ \widetilde{\boldsymbol{p}}_{n+1} =& \boldsymbol{p}_{n+1} + \omega(\boldsymbol{p}_{n+1} - \boldsymbol{p}_n), \\ \boldsymbol{u}_{n+1} =& \boldsymbol{u}_n - \tau_U D\mathcal{F}(\boldsymbol{u}_n)^*(\mathfrak{M}^{-1})^*\widetilde{\boldsymbol{p}}_{n+1}. \end{aligned} \tag{1.6}$$

The above treatment (1.6) will significantly improve the algorithm's convergence speed while leaving the equilibrium state invariant.

It is worth noting that the original approach proposed in [35] solves the implicit scheme (1.3) while preserving the time causality: the algorithm sequentially computes u^t at each time step, using the previous solution as the initial condition. In contrast, our approach generalizes by accumulating multiple time steps into a single root-finding problem and computing the multi-step solution in a forward manner. More precisely, we solve $\mathcal{F}(\boldsymbol{u}_j) = 0$ for sequential blocks of solutions, where each block is defined as:

$$\boldsymbol{u}_j = \left[u^{j\cdot N_t+1}, \dots, u^{j\cdot N_t+t}, \dots, u^{j\cdot N_t+N_t}\right]^\top \in \mathcal{X}^{N_t}, \quad j = 0, 1, 2, \dots$$

When updating from $\boldsymbol{u}_j$ to $\boldsymbol{u}_{j+1}$, we set $u^0 = u^{j\cdot N_t+N_t}$ as the initial condition in (1.2). Unlike step-by-step update, the new approach preserves time causality among solution blocks $\boldsymbol{u}_j, \boldsymbol{u}_{j+1}$.

In this paper, we analyze the aforementioned preconditioned PDHG algorithm (1.6) to establish sufficient conditions under which the method is guaranteed to converge. We remark that there are two types of convergence analysis, which may cause confusion in this manuscript. One refers to the convergence of the numerical solution to the real solution as the number of grid points increases; the other one refers to the convergence of $(\boldsymbol{u}_n, \boldsymbol{p}_n)$ to the equilibrium state of the PDHG algorithm (1.6) as n increases. In this research, we mainly focus on analyzing the second type of convergence. We now briefly summarize the main contributions:

- (Theoretical aspect) Suppose that the reaction term $f(\cdot)$ is Lipschitz. Assume that the discretization of the differential operators $\mathcal{L}_h, \mathcal{G}_h$ are positive-definite, self-adjoint, and commute. We establish the following theoretical results for our algorithm.
  1. We study the PDHG flow (3.5), which is the time-continuous limit of (2.16) as $\tau_U, \tau_P \to 0, (1+\omega)\tau_P \to \gamma > 0$. We give conditions on h_t, N_t under which we can pick γ, ϵ such that the residual term exponentially decays to 0. The convergence results for general RD equations are discussed in Theorem 5 and Theorem 7; We establish convergence rates that are independent of the grid-size N_x for both Allen-Cahn type and Cahn-Hilliard type equations in Corollary 7.1.
  2. We analyze the convergence speed of the PDHG method (2.16) in Theorem 8. We show that under certain conditions of h_t, N_t, we are able to select suitable hyperparameters $\tau_U, \tau_P, \omega, \epsilon$ that guarantee the exponential convergence of the L_2 error term. We establish convergence rates that are independent of the grid-size N_x for both Allen-Cahn type and Cahn-Hilliard type equations in Corollary 8.1.
- (Numerical aspect) In section 3.2.2 and 3.3 we justify our theoretical results stated above. In section 4.1, we demonstrate the effectiveness of our algorithm on different RD equations,

including the standard Allen-Cahn and Cahn-Hilliard equations, as well as equations with variable mobility terms or higher-order diffusion terms whose linear operator $\mathscr{M}$ (c.f. (2.13)) cannot be directly inverted. In section 4.1.5, we validate that the convergence rate of our method is independent of the grid size N_x. In section 4.3, we investigate the optimal, or at least near-optimal hyperparameters of our algorithm for achieving efficient performance. We demonstrate the efficiency of our method by comparing it with some of the classical methods in section 4.4 and section 4.5.

There exist plenty of references regarding the numerical schemes for RD equations, which include studies on finite difference methods [6, 11, 15, 24, 25, 29, 38, 43, 44, 50, 52], and finite element methods [19, 20, 25, 31, 32, 33, 34, 53]. A series of benchmark problems [12, 27] has also been introduced to verify the effectiveness of the proposed methods. Recently, machine learning or deep learning algorithms such as deep-learning-based backward stochastic differential equations (BSDE) [22, 23], physics-informed neural networks (PINNs) [41, 49, 51], and Gaussian processes [10] have also been applied to deal with various types of nonlinear equations including the RD equations.

The primal-dual hybrid gradients (PDHG) method was first introduced in [7, 54] to deal with constrained optimization problems arising in image processing. This method later finds its applications in various branches such as nonsmooth PDE-constrained optimization [13], Magnetic resonance imaging (MRI) [48], large-scale optimization problems including image denoising and optimal transport [26], computing gradient flows in Wasserstein-like transport metric spaces [4, 5, 19], as well as design fast optimization algorithms [56], etc.

In [16], the authors introduce damping terms to the wave equation to achieve faster stabilization, which resembles the time-continuous limit (the PDHG flow) (3.5) of our proposed algorithm. However, [16] focuses on the linear case while our research deals with nonlinear RD equations. In recent work [9], the authors conduct certain transformations to enhance the convergence of a saddle point algorithm. Although the transformed algorithm shares similarities with our method, the target functionals considered in both researches are distinct. In [8], the authors apply the splitting method to propose an accelerating algorithm for the root-finding problem $\mathcal{A}(x) = 0$, where $\mathcal{A}$ can be decomposed as the sum of the gradient function and the skew-symmetric operator. In contrast, our proposed method deals with a time-dependent root-finding problem, which generally can not be cast into the settings of [8]. We refer our readers to [35] for more detailed discussions on related references.

Our research is inspired by [36] in which the authors apply the PDHG algorithm to compute time-implicit conservation laws. Our former research [35] mainly focuses on the conceptual and experimental aspects of the PDHG method applied to RD equations. In addition, the primal-dual method also finds its application in the computation of Hamilton-Jacobi equations [37]. The aforementioned works [35, 36, 37] do not address the convergence speed of the PDHG algorithm. In this work, we establish the convergence guarantee for the nonlinearly coupled primal-dual system. Moreover, we prove a convergence property of our method, where the convergence rate is independent of the space grid size.

This paper is organized as follows. In section 2, we provide a detailed derivation of our algorithm applied to RD equations. In section 3.1, we establish the existence and uniqueness result regarding the time-implicit scheme of the RD equation. In section 3.2, we focus on the PDHG flow, which is the time-continuous limit of the proposed algorithm. We first establish convergence results for the general root-finding problem and then apply our theory to the time-implicit schemes of RD equations. In section 3.3, we prove exponential convergence of our algorithm. We also investigate necessary conditions that guarantee such convergence. In section 4, we demonstrate the effectiveness of our method on different types of RD equations and make comprehensive comparisons with the IMEX scheme as well as some classical root-finding algorithms.

2. Derivation of the method

In this section, we give a detailed derivation of the PDHG method when applied to the reaction-diffusion (RD) equation (1.1). From now on, we assume that the domain $\Omega = [0, L]^2$ is a square region.

Suppose we solve (1.1) on the time interval $[0, T]$. We divide the time interval into N_t subintervals, and divide the domain Ω into $N_x \times N_x$ grids. Applying time-implicit finite difference scheme yields

$$\frac{\mathrm{u}^{t+1} - \mathrm{u}^t}{h_t} = -\mathcal{G}_h(a\mathcal{L}_h \mathrm{u}^{t+1} + bf(\mathrm{u}^{t+1})), \quad \text{for } t = 0, 1, \ldots, N_t, \text{ with } \mathrm{u}^0 \text{ given.} \tag{2.1}$$

Denote $h_t = \frac{T}{N_t}$, and $h_x = \frac{L}{N_x}$. Write $U^t \in \mathbb{R}^{N_x \times N_x}$ as the numerical solution at the $t-$th time node. We denote $\mathcal{G}_h, \mathcal{L}_h$ as $N_x^2 \times N_x^2$ matrices, which represents the discretization of the operator $\mathcal{L}, \mathcal{G}$ w.r.t. the spatial step size h_x and the boundary condition.

Remark 1 (Allen-Cahn and Cahn-Hilliard equations)**.** *For Allen-Cahn equation* [1]*, we have* $\mathcal{G} = \mathrm{Id}$*,* $\mathcal{L} = -\Delta$*; for Cahn-Hilliard equation* [3]*, we have* $\mathcal{G} = -\Delta$*,* $\mathcal{L} = -\Delta$*. And* $f(\cdot) = W'(\cdot)$ *where* $W(\xi) = \frac{1}{4}(\xi^2 - 1)^2$ *is the double-well potential for both equations. We can impose periodic or homogeneous Neumann boundary conditions for both equations. Furthermore, suppose we apply the central difference scheme to discretize the Laplace operator* Δ*. We obtain* $\Delta_{h_x}^P = I_{N_x} \otimes \mathrm{Lap}_{h_x}^P + \mathrm{Lap}_{h_x}^P \otimes I_{N_x}$ *for periodic boundary condition, and* $\Delta_{h_x}^N = I_{N_x} \otimes \mathrm{Lap}_{h_x}^N + \mathrm{Lap}_{h_x}^N \otimes I_{N_x}$ *for Neumann boundary condition, where* $\otimes$ *is the Kronecker product and we define*

$$\mathrm{Lap}_{h_x}^P = \frac{1}{h_x^2}\begin{bmatrix} -2 & 1 & & & 1 \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ 1 & & & 1 & -2 \end{bmatrix}, \quad \mathrm{Lap}_{h_x}^N = \frac{1}{h_x^2}\begin{bmatrix} -1 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -1 \end{bmatrix}. \tag{2.2}$$

2.1. PDHG method for preconditioned root-finding problem.

In this section, we provide a more detailed derivation for our algorithm.

Let us treat $\mathcal{X} = \mathbb{R}^{N_x^2}$. We denote $U = [\mathrm{u}^{1\top}, \ldots, \mathrm{u}^{N_t\top}]^\top \in \mathbb{R}^{N_t N_x^2}$ as the numerical solution. $\mathcal{L}_h, \mathcal{G}_h$ indicate the discrete approximations of $\mathcal{L}, \mathcal{G}$. We formulate the time-implicit scheme (2.1) as a root-finding problem

$$F(U) = 0, \tag{2.3}$$

with $F : \mathbb{R}^{N_t N_x^2} \to \mathbb{R}^{N_t N_x^2}$ defined as

$$F(U) = \mathscr{D}U + h_t \mathscr{G}_h(a\mathscr{L}_h U + bf(U)) - V. \tag{2.4}$$

Here we denote the time difference matrix $\mathscr{D} = D_{N_t} \otimes I_x$, where I_x is the identity matrix on $\mathbb{R}^{N_x^2}$, and

$$D_N = \begin{bmatrix} 1 & & & & \\ -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{bmatrix} \quad \text{is an } N \times N \text{ matrix.} \tag{2.5}$$

On the other hand, we define

$$\mathscr{G}_h = I_t \otimes \mathcal{G}_h, \quad \mathscr{L}_h = I_t \otimes \mathcal{L}_h, \tag{2.6}$$

with I_t representing the identity matrix on $\mathbb{R}^{N_t}$. The reaction function $f(\cdot)$ acts element-wisely on vector U. The constant vector $V \in \mathbb{R}^{N_t N_x^2}$ depends on both the initial condition and the boundary condition of the equation.

We aim to solve $F(U) = 0$. In [35], an indicator function $\iota(u) = \begin{cases} 0 & \text{if } u = 0; \\ +\infty & \text{if } u \neq 0; \end{cases}$ is introduced to reformulate the root-finding problem as an optimization problem

$$\inf_{U \in \mathbb{R}^{N_t N_x^2}} \iota(F(U)), \tag{2.7}$$

which can be further reduced to an inf-sup saddle problem

$$\inf_{U \in \mathbb{R}^{N_t N_x^2}} \sup_{P \in \mathbb{R}^{N_t N_x^2}} P^\top F(U). \tag{2.8}$$

Inspired by [56], we replace ι in (2.7) by a milder quadratic function $\frac{1}{2\epsilon}\|\cdot\|^2$ to obtain

$$\inf_{U \in \mathbb{R}^{N_t N_x^2}} \frac{1}{2\epsilon}\|F(U)\|^2. \tag{2.9}$$

By introducing the dual variable $P \in \mathbb{R}^{N_t N_x^2}$, one can reformulate (2.9) as an inf-sup problem with a tunable parameter ϵ,

$$\inf_{U \in \mathbb{R}^{N_t N_x^2}} \sup_{P \in \mathbb{R}^{N_t N_x^2}} L(U, P) \triangleq P^\top F(U) - \frac{\epsilon}{2}\|P\|^2. \tag{2.10}$$

We tackle this saddle point problem by leveraging the primal-dual hybrid gradient (PDHG) algorithm and obtain

$$\begin{aligned} P_{n+1} =& \frac{1}{1+\epsilon\tau_P}\left(P_n + \tau_P F(U_n)\right), \\ \widetilde{P}_{n+1} =& P_{n+1} + \omega(P_{n+1} - P_n), \\ U_{n+1} =& U_n - \tau_U DF(U_n)^\top \widetilde{P}_{n+1}. \end{aligned} \tag{2.11}$$

When $DF(U)$ is nonsingular for arbitrary $U \in \mathbb{R}^{N_t N_x^2}$, the equilibrium state of the above discrete dynamic is $(U_*, 0)$ with $F(U_*) = 0$. As discussed in the introduction, a large condition number of $F(\cdot)$ may significantly slow down the convergence speed of (2.11). To mitigate this, we consider suitable preconditioning of $F(\cdot)$. Let us decompose $F(U)$ into its linear part and nonlinear part,

$$\begin{aligned} F(U) =& \mathscr{D}U + h_t \mathscr{G}_h (a\mathscr{L}_h U + bf(U)) - V \\ =& (\mathscr{D} + ah_t\mathscr{G}_h\mathscr{L}_h)U + bh_t\mathscr{G}_h(f(\overline{U}) + J_f(U - \overline{U}) + R(U)) - V. \end{aligned} \tag{2.12}$$

Here we assume $\overline{U}$ is a certain point in $\mathbb{R}^{N_x^2}$ at which we expand $f(U) = f(\overline{U}) + J_f(U - \overline{U}) + R(U)$. We choose matrix J_f as an approximation of the Jacobian matrix $Df(\overline{U}) = \text{diag}(\dots, f'(\overline{U}_{ij}), \dots)$. We denote $R(U) \triangleq f(U) - f(\overline{U}) - J_f(U - \overline{U})$ as the remainder term.

Remark 2. *In practice, we usually choose $J_f = Df(u_{\mathrm{e}}\mathbf{1})$ where $\mathbf{1}$ is the $1-$vector, and u_{e} is one of the stable equilibrium states, i.e. $f(u_{\mathrm{e}}) = 0$. For example, in Allen-Cahn equation, $f(u) = u^3 - u$, then $u_{\mathrm{e}} = \pm 1$, we always have $f'(u_{\mathrm{e}}) = 2$. Thus, we set $J_f = 2I$.*

By writing

$$\mathscr{M} = \mathscr{D} + ah_t\mathscr{G}_h\mathscr{L}_h + bh_t\mathscr{G}_h J_f = \begin{bmatrix} X & & & \\ -I & X & & \\ & \ddots & \ddots & \\ & & -I & X \end{bmatrix} \text{ with } X = I + ah_t\mathcal{G}_h\mathcal{L}_h + bh_t\mathcal{G}_h J_f, \tag{2.13}$$

$$\widetilde{w} = bh_t\mathscr{G}_h(f(\overline{U}) - J_f\overline{U}) - V,$$

we decompose $F(U)$ as $\mathscr{M}U+bh_t\mathscr{G}_hR(U)-\widetilde{w}$. It is beneficial to consider the preconditioned function

$$\widehat{F}(U)=\mathscr{M}^{-1}F(U)=U+\mathscr{M}^{-1}(bh_t\mathscr{G}_hR(U))-\widetilde{w}\overset{\text{denote as}}{=}U+\eta(U). \tag{2.14}$$

We discuss the sufficient condition under which $\mathscr{M}$ is invertible in the following remark.

Remark 3 (Invertibility of $\mathscr{M}$)**.** *Suppose $a,b\geq 0$, $\mathcal{G}_h$, $\mathcal{L}_h$ are self-adjoint, non-negative definite, and commute. Assume $J_f=cI$ with $c\geq 0$. Then $\mathscr{M}$ is invertible for any $h_t>0$. To prove this, it suffices to show that each X is invertible. By similar arguments of the proof in Lemma 13, it is not hard to verify that X is equivalent to $I+ah_t\Lambda_{\mathcal{G}_h}\Lambda_{\mathcal{L}_h}+bch_t\Lambda_{\mathcal{G}_h}$, which is invertible for $h_t>0$. Here $\Lambda_{\mathcal{G}_h},\Lambda_{\mathcal{L}_h}$ are diagonal matrices equivalent to $\mathcal{G}_h,\mathcal{L}_h$.*

The corresponding root-finding problem $\widehat{F}(U)=0$ is equivalent to the original problem (2.3) whenever $\mathscr{M}$ is invertible.

We now apply (2.11) to the inf-sup saddle problem with respect to $\widehat{F}(\cdot)$

$$\inf_{U\in\mathbb{R}^{N_tN_x^2}}\sup_{Q\in\mathbb{R}^{N_tN_x^2}}\widehat{L}(U,Q)\triangleq Q^\top\widehat{F}(U)-\frac{\epsilon}{2}\|Q\|_2^2. \tag{2.15}$$

And our PDHG method with *implicit* update in Q and *explicit* update in U yields

$$\begin{aligned}Q_{k+1}&=\frac{1}{1+\epsilon\tau_P}(Q_k+\tau_P(\widehat{F}(U_k)));\\ \widetilde{Q}_{k+1}&=Q_{k+1}+\omega(Q_{k+1}-Q_k);\\ U_{k+1}&=U_k-\tau_U(D\widehat{F}(U_k)^\top\widetilde{Q}_{k+1}).\end{aligned} \tag{2.16}$$

We then iterate (2.16) so that $\{U_k\}$ approaches the desired root U_*. We terminate the iteration whenever the ℓ^∞ norm of the residual term

$$\operatorname{Res}(U_k)=F(U_k)/h_t=\left[\dots,\left(\frac{\mathrm{u}_k^{t+1}-\mathrm{u}_k^t}{h_t}+\mathcal{G}_h(a\mathcal{L}_h\mathrm{u}_k^{t+1}+bf(\mathrm{u}_k^{t+1}))\right)^\top,\dots\right]^\top_{0\leq t\leq N_t-1}. \tag{2.17}$$

is less than a certain tolerance *tol*, i.e., $\|\operatorname{Res}(U_k)\|_\infty<tol$.

2.2. **Complexity of the algorithm.** We apply the Fast Fourier Transform (FFT) [14, 46] to evaluate the multiplication of $\mathcal{L}_h,\mathcal{G}_h$ for periodic boundary conditions. Furthermore, the Discrete Cosine Transform (DCT) [47] can be utilized to handle the no-flux or more general Neumann boundary conditions. We refer interested readers to [35] for more details. Thus, computing $F(U)$ requires $\mathcal{O}(N_tN_x^2\log N_x)$ steps of operations. Furthermore, since $\mathscr{M}$ is block lower triangular, applying back substitution together with FFT/DCT to solve the linear system involving $\mathscr{M}$ requires $\mathcal{O}(N_tN_x^2\log N_x)$ steps of operations. Thus, the complexity at each iteration of our algorithm equals $\mathcal{O}(N_tN_x^2\log(N_x))$.

2.3. **Computing with time causality.** As mentioned in Section 1, time causality can be incorporated into the numerical scheme by solving sequential blocks of numerical solutions:

$$U^j=\left[\mathrm{u}^{j\cdot N_t+1\top},\dots,\mathrm{u}^{j\cdot N_t+t\top},\dots,\mathrm{u}^{j\cdot N_t+N_t\top}\right]^\top\in\mathbb{R}^{N_x^2N_t},\quad j=0,1,2,\dots$$

More precisely, to compute U^j over the j-th time interval $[(j-1)\cdot N_t\cdot h_t, j\cdot N_t\cdot h_t]$, the proposed PDHG algorithm is applied to the root-finding problem $F(U^j)=0$, with $\mathrm{u}_0=\mathrm{u}^{(j-1)\cdot N_t+N_t}$. That is, the initial value is set as the final state from the previous block U^{j-1}.

From a practical perspective, increasing N_t leads to higher memory consumption. From a theoretical point of view, as justified in Corollary 7.1 and 8.1, fixing the time step size h_t while selecting a large N_t may result in an ill-conditioned root-finding problem, posing challenges to the convergence of the method. In practice, choosing a moderate N_t (generally not exceeding 5) mitigates these issues and ensures the efficient performance of the algorithm. Further discussions regarding hyperparameter selections are provided in Section 4.3.

A standard choice of the initial values (U_0, Q_0) for the PDHG algorithm (2.16) upon solving $F(U^j) = 0$ is $U_0 = \widetilde{U}^j, Q_0 = 0$, where $\widetilde{U}^j$ denotes the numerical solution precomputed using the IMEX scheme with initial condition $\mathrm{u}^{(j-1)\cdot N_t + N_t}$. A simpler alternative is to set $U_0 = U^{j-1}, Q_0 = 0$. Both choices are efficient in practice as long as N_t is not too large.

2.4. **Relation with G-prox PDHG method.** The G-prox primal-dual hybrid gradients algorithm [26] was recently invented to improve the convergence of optimization and root-finding problems. The algorithm can be formulated as

$$\begin{aligned} P_{k+1} &= \underset{P \in \mathbb{R}^{N_t N_x^2}}{\operatorname{argmin}} \left\{ \frac{1}{2\tau_P} \|P - P_k\|_G^2 - \widehat{L}(U_k, P) \right\} = \frac{1}{1+\epsilon\tau_P}(P_k + \tau_P G^{-1} F(U_k)); \\ \widetilde{P}_{k+1} &= P_{k+1} + \omega(P_{k+1} - P_k); \\ U_{k+1} &= \underset{U \in \mathbb{R}^{N_t N_x^2}}{\operatorname{argmin}} \left\{ \frac{1}{\tau_P} \|U - U_k\|_2^2 + \widehat{L}(U, \widetilde{P}_{k+1}) \right\}. \end{aligned} \tag{2.18}$$

Here we define the G-weighted norm as $\|v\|_G^2 = v^\top G v$, and pick $G = \mathscr{M}\mathscr{M}^\top$. In practice, we substitute the following explicit update of U_k for the implicit update,

$$U_{k+1} = U_k - \tau_U DF(U_k)^\top \widetilde{P}_{k+1}. \tag{2.19}$$

Now, we multiply $\mathscr{M}^\top$ on both sides of (2.18) (but with the third line replaced by (2.19)) to obtain

$$\begin{aligned} \mathscr{M}^\top P_{k+1} &= \frac{1}{1+\epsilon\tau_P}(\mathscr{M}^\top P_k + \tau_P \mathscr{M}^{-1} F(U_k)); \\ \mathscr{M}^\top \widetilde{P}_{k+1} &= \mathscr{M}^\top P_{k+1} + \omega(\mathscr{M}^\top P_{k+1} - \mathscr{M}^\top P_k); \\ U_{k+1} &= U_k - \tau_U D\widehat{F}(U_k)^\top (\mathscr{M}^\top \widetilde{P}_{k+1}). \end{aligned} \tag{2.20}$$

By denoting $Q_k = \mathscr{M}^\top P_k$ and noticing that $\widehat{F}(U) = \mathscr{M}^{-1} F(U)$, (2.20) reduces exactly to (2.16). This verifies the equivalence between the G-prox PDHG algorithm and our proposed method.

3. Numerical analysis of the proposed method

In this section, we study the numerical convergence properties of the proposed PDHG algorithm. In subsection 3.1, we prove the unique solvability of the time-implicit scheme (2.1) of RD equations. In subsection 3.2, we study the convergence of the time-continuous limit of the PDHG algorithm. In subsection 3.3, we prove the convergence of the PDHG algorithm.

3.1. **Unique solvability of the time-implicit scheme.** In this research, we mainly focus on reaction functions f that belong to the functional space $\mathcal{F}$, where

$$\mathcal{F} = \left\{ f \in C^1(\mathbb{R}) \;\middle|\; \begin{array}{l} f \text{ can be decomposed as } f = V' + \phi, \\ \text{where } V \in C^1(\mathbb{R}) \text{ is convex, and } \phi \in C(\mathbb{R}) \text{ is Lipschitz.} \end{array} \right\}. \tag{3.1}$$

The space $\mathcal{F}$ covers a majority of reaction functions that arise in classical RD equations such as the Allen-Cahn and the Cahn-Hilliard equations.

Before we present the result, we assume the spectral decomposition of $\mathcal{G}_h$:

$$\mathcal{G}_h = \left[\begin{array}{c|c} Q_1 & Q_2 \end{array}\right] \left[\begin{array}{cc} \Lambda & \\ & O \end{array}\right] \left[\begin{array}{c} Q_1^\top \\ Q_2^\top \end{array}\right], \tag{3.2}$$

where $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_r)$ is a diagonal matrix with positive entries $\lambda_1 \geq \cdots \geq \lambda_r > 0$, $r = \operatorname{rank}(\mathcal{G}_h)$.

Theorem 1 (Existence and uniqueness of (2.3))**.** *Suppose that $\mathcal{G}_h$, $\mathcal{L}_h$ used in the finite difference scheme* (2.1) *are self-adjoint and positive semidefinite. Assume $\mathcal{G}_h$ has the spectral decomposition as in* (3.2)*. We also assume that $f \in \mathcal{F}$, such that the convex function V satisfies*

$$(V'(x) - V'(y), x - y) \geq K|x - y|^2,$$

for some $K \geq 0$. If the time step size h_t in (2.1) *satisfies*

$$\lambda_{\min}\left(\frac{\Lambda^{-1}}{h_t} + a\, Q_1^\top \mathcal{L}_h Q_1\right) + bK > b\,\mathrm{Lip}(\phi), \tag{3.3}$$

then the root-finding problem (2.3) *admits a unique solution.*

The proof of the theorem is deferred to Appendix A.1.

Remark 4. *The condition* (3.3) *can be simplified for some specific equations.*

- *(Allen-Cahn equation with periodic boundary condition) $\mathcal{G} = \mathrm{Id}, \mathcal{L} = -\Delta$, $f(x) = x^3 - x$. We set $\mathcal{G}_h = I_{N_x^2}$, and $\mathcal{L}_h = -\Delta_{h_x}^P = I_{N_x} \otimes (-\mathrm{Lap}_{h_x}^P) + (-\mathrm{Lap}_{h_x}^P) \otimes I_{N_x}$, where $\mathrm{Lap}_{h_x}^P$ is defined in* (2.2)*. In this case, the condition* (3.3) *yields $h_t < \frac{1}{2b}$.*
- *(Cahn-Hilliard equation with periodic boundary condition) $\mathcal{G} = -\Delta$, $\mathcal{L} = -\Delta$, $f(x) = x^3 - x$. We set $\mathcal{G}_h = \mathcal{L}_h = -\Delta_{h_x}^P$. A sufficient condition for* (3.3) *is $h_t < \frac{a^2}{b^2}$.*

Similar results regarding both Allen-Cahn and Cahn-Hilliard equations have also been done in [50]*. Theorem 1 applies to general RD equation* (1.1)*. We refer interested readers to Appendix A.2 for more detailed discussions.*

3.2. **Lyapunov analysis for the PDHG flow.** We are ready to present the main result of this paper. In subsection 3.2.1, we first prove the convergence of the time-continuous limit of the PDHG algorithm (2.16) for the general root-finding problem. In subsection 3.2.2, we apply the previous theory to the time-implicit scheme of RD equations. In subsection 3.2.3, we provide numerical justifications for the theoretical study. To alleviate the notation, we denote $\|\cdot\|$ as the 2−norm for both vectors and matrices in the following discussion.

3.2.1. *Convergence analysis for the general root-finding problem.* Firstly, we establish the convergence result for a general root-finding problem $\widehat{F}(U) = 0$ regardless of the exact form of $\widehat{F}(U)$. Our main results are summarized in Theorem 2 and Corollary 3.

Recall (2.16), we substitute $\widetilde{Q}_{k+1}$ with

$$\begin{aligned}\widetilde{Q}_{k+1} = Q_k + (1+\omega)(Q_{k+1} - Q_k) &= Q_k + (1+\omega)\tau_P\left(-\frac{\epsilon}{1+\epsilon\tau_P}Q_k + \frac{1}{1+\epsilon\tau_P}\widehat{F}(U_k)\right)\\ &= \left(1 - \frac{(1+\omega)\tau_P\epsilon}{1+\epsilon\tau_P}\right)Q_k + \frac{(1+\omega)\tau_P}{1+\epsilon\tau_P}\widehat{F}(U_k).\end{aligned}$$

Then, the PDHG iteration (2.16) can be formulated as

$$\begin{aligned}\frac{Q_{k+1} - Q_k}{\tau_P} &= -\frac{\epsilon}{1+\epsilon\tau_P}Q_k + \frac{1}{1+\epsilon\tau_P}\widehat{F}(U_k);\\ \frac{U_{k+1} - U_k}{\tau_U} &= -D\widehat{F}(U_k)^\top\left(\left(1 - \frac{(1+\omega)\tau_P\epsilon}{1+\epsilon\tau_P}\right)Q_k + \frac{(1+\omega)\tau_P}{1+\epsilon\tau_P}\widehat{F}(U_k)\right).\end{aligned} \tag{3.4}$$

Suppose we send the step sizes $\tau_U, \tau_P \to 0$, and keep ω increasing such that $(1+\omega)\tau_P \to \gamma > 0$. Then the above time-discrete dynamic will converge to the following time-continuous dynamic of (U_t, Q_t) which we denote as the "PDHG flow".

$$\begin{cases}\dot{Q} = -\,\epsilon Q + \widehat{F}(U),\\ \dot{U} = -\,D\widehat{F}(U)^\top((1-\gamma\epsilon)Q + \gamma\widehat{F}(U)).\end{cases} \tag{3.5}$$

We introduce two notations that will be commonly used in the following discussion,

$$\underline{\sigma} = \inf_{U \in \mathbb{R}^{N_t N_x^2}} \{\sigma_{\min}(D\widehat{F}(U))\} = \inf_{U \in \mathbb{R}^{N_t N_x^2}} \{\sigma_{\min}(I + bh_t \mathscr{M}^{-1} \mathscr{G}_h DR(U))\}, \tag{3.6}$$

$$\overline{\sigma} = \sup_{U \in \mathbb{R}^{N_t N_x^2}} \{\sigma_{\max}(D\widehat{F}(U))\} = \sup_{U \in \mathbb{R}^{N_t N_x^2}} \{\sigma_{\max}(I + bh_t \mathscr{M}^{-1} \mathscr{G}_h DR(U))\}, \tag{3.7}$$

where $\sigma_{\min}(A)(\sigma_{\max}(A))$ denotes the minimum (maximum) singular value of matrix A. The condition number is defined by

$$\kappa = \overline{\sigma}/\underline{\sigma}. \tag{3.8}$$

We consider the following Lyapunov function of (U, Q) associated with a parameter $\mu > 0$,

$$\mathcal{I}_\mu(U, Q) = \frac{1}{2}\|\widehat{F}(U)\|^2 + \frac{\mu}{2}\|Q\|^2. \tag{3.9}$$

The parameter μ enables us to establish the exponential decay of $\mathcal{I}_\mu(U_t, Q_t)$ along the PDHG flow whenever $0 < \underline{\sigma} \leq \overline{\sigma} < \infty$. We have the following Lemma.

Lemma 2 (Exponential decay of $\mathcal{I}_\mu(U_t, Q_t)$)**.** *Suppose that* $0 < \underline{\sigma} \leq \overline{\sigma} < \infty$. *We pick the parameter* $\mu > 0$ *satisfying*

$$\frac{1}{\underline{\sigma}} - \frac{1}{\overline{\sigma}} < \frac{2}{\sqrt{\mu}}. \tag{3.10}$$

Furthermore, we choose $\gamma, \epsilon > 0$ *satisfying*

$$\max\left\{\left(1 - \frac{\sqrt{\mu}}{\overline{\sigma}}\right)^2, \left(1 - \frac{\sqrt{\mu}}{\underline{\sigma}}\right)^2\right\} < \gamma\epsilon < \left(1 + \frac{\sqrt{\mu}}{\overline{\sigma}}\right)^2. \tag{3.11}$$

Under the above choices of μ, γ *and* ϵ, *let* (U_t, Q_t) *be the solution to the PDHG flow* (3.5) *with arbitrary initial condition* (U_0, Q_0). *Then we have,*

$$\mathcal{I}_\mu(U_t, Q_t) \leq \exp\left(-\frac{2\beta\, t}{\max\{1, \mu\}}\right) \mathcal{I}_\mu(U_0, Q_0).$$

Here we denote

$$\beta = \min_{z \in [\underline{\sigma}^2, \overline{\sigma}^2]} \{\varphi_{\mu,\gamma,\epsilon}(z)\} > 0,$$

with $\varphi_{\mu,\gamma,\epsilon}(z) = \frac{1}{2}(\gamma z + \mu\epsilon - \sqrt{(\gamma z - \mu\epsilon)^2 + (\mu - (1-\gamma\epsilon)z)^2})$.

We defer the proof of this Lemma to Appendix B.1. Lemma 2 provides a sharp convergence rate for $\mathcal{I}_\mu(U_t, Q_t)$. However, β does not take an explicit form. In the following theorem, we relax the bound in Lemma 2 to obtain an explicit convergence rate for $\|\widehat{F}(U_t)\|$.

Theorem 3 (Exponential decay of the residual $\|\widehat{F}(U_t)\|$)**.** *Assume that* (U_t, Q_t) *solves* (3.5) *with an arbitrary initial position* (U_0, Q_0). *Then, as long as* $\underline{\sigma}$ *is bounded away from* 0 *and* $\overline{\sigma}$ *is finite, one can always pick suitable parameters* γ, ϵ *such that the residual* $\|\widehat{F}(U_t)\|$ *decays exponentially fast to* 0. *In particular, if we set* $\epsilon = (1-\delta)\kappa$ *and* $\gamma = \frac{1-\delta}{\kappa}$ *with* $|\delta| < \frac{1}{\kappa}$, *then we have*

$$\|\widehat{F}(U_t)\|_2 \leq \exp\left(-(1 - \kappa|\delta|)(3 - \delta)\frac{\min\{\underline{\sigma}^2, 1\}}{8\kappa}\, t\right) \sqrt{\|\widehat{F}(U_0)\|^2 + \underline{\sigma}^2\|Q_0\|^2}.$$

The proof is provided in Appendix B.1. We can further improve the convergence rate by fixing $\gamma\epsilon = 1$ in Theorem 11 of Appendix B.1.

3.2.2. *Convergence analysis for our specific root-finding problem* (2.14). In this section, we discuss the exponential decay of the PDHG flow (3.5) when it is applied to the time-implicit scheme (2.1) of the RD equation (1.1) when the reaction term $f(\cdot)$ is Lipschitz. The main results of this section are Theorem 7 and Corollary 7.1.

Before demonstrating our result, we list several conditions regarding equation (1.1) and its numerical scheme (2.1). These conditions will be used later.

(1) Suppose the coefficients a, b are non-negative, i.e.,

$$a \geq 0, \quad b \geq 0. \tag{A}$$

(2) Assume that

$$f(\cdot) \text{ is Lipschitz with constant } \mathrm{Lip}(f). \tag{B}$$

(3) In the numerical scheme (2.1) of (1.1), suppose

$$\mathcal{L}_h, \mathcal{G}_h \text{ are self-adjoint, non-negative definite, and commute, i.e., } \mathcal{G}_h\mathcal{L}_h = \mathcal{L}_h\mathcal{G}_h. \tag{C}$$

(4) Recall J_f mentioned in (2.12). We assume

$$J_f \text{ is a constant diagonal matrix } cI \text{ with } c \geq 0. \tag{D}$$

Remark 5. *We point out that many reaction-diffusion equations do not possess Lipschitz reaction terms $f(\cdot)$: Double-well polynomial potential in the phase field model, as well as logarithmic Flory-Huggins potential, does not yield Lipschitz reaction functions* [28, 18]. *However, the Lipschitz assumption can still be applied if one can prove an a priori estimation on ℓ^∞ norm of the numerical solution U_k for all PDHG iteration k. This may serve as our future research topic.*

As stated in Theorem 3, we need $\underline{\sigma} > 0$ and $\overline{\sigma} < \infty$ in order to establish the exponential decay of $\|\widehat{F}(U)\|$. Lemma 4 provides a sufficient condition for this to hold.

Lemma 4. *Suppose* (A), (B), (C) *hold. When $h_t < \frac{1}{|b|\lambda_{\max}(\mathcal{G}_h)\mathrm{Lip}(f)}$, we always have $\underline{\sigma} > 0$ and $\overline{\sigma} < \infty$.*

We prove this Lemma in Appendix B.2. Combining Theorem 3 and Lemma 4 leads to the following Theorem 5.

Theorem 5 (First convergence result of $\|\widehat{F}(U_t)\|$)**.** *Consider the RD equation* (1.1) *on* $[0, T]$. *Suppose* (A), (B) *and* (C) *hold. We apply the PDHG flow* (3.5) *to solve the time-implicit scheme* (2.1) *with time step size $h_t < \frac{1}{|b|\|\mathcal{G}_h\|\mathrm{Lip}(f)}$. Suppose $\gamma = \frac{1-\delta}{\kappa}$, and $\epsilon = (1-\delta)\kappa$ with $\kappa = \overline{\sigma}/\underline{\sigma}$, and $|\delta| < \frac{1}{\kappa}$. Then $\|\widehat{F}(U_t)\|$ converges exponentially fast to* 0.

Remark 6. *It is worth mentioning that we do not assume condition* (3.3) *of Theorem 1. Then $\widehat{F}(U) = 0$ might not admit a unique solution, but the exponential decay of $\|\widehat{F}(U_t)\|$ is still guaranteed.*

Although Theorem 5 guarantees the exponential convergence of $\|\widehat{F}(U_t)\|$ for arbitrarily large b and T as long as h_t, γ, ϵ are suitably chosen, both the time step size h_t and the convergence rate may depend on the spatial discretization N_x. To get rid of this dependency, we provide sufficient conditions under which $\underline{\sigma}$ and $\overline{\sigma}$ are bounded away from the constants that are independent of N_x. Thus, we achieve a convergence rate that is independent of N_x. Recall the remainder term $R(U) = f(U) - f(\overline{U}) - Df(\overline{U})(U - \overline{U})$ mentioned in (2.12). We have the following Lemma.

Lemma 6. *Consider the reaction-diffusion type equation* (1.1) *on* $[0, T]$. *Suppose the conditions* (A), (B), (C) *and* (D) *hold. Since* (B) *requires f to be Lipschitz, so does R. And we denote its Lipschitz constant as* $\mathrm{Lip}(R)$. *Define*

$$\zeta_{a,b,c}(h_t) = \max_{1 \leq k \leq N_x^2} \left\{ \frac{\lambda_k(\mathcal{G}_h)}{1 + h_t(a\lambda_k(\mathcal{G}_h)\lambda_k(\mathcal{L}_h) + bc\lambda_k(\mathcal{G}_h))} \right\},$$

where $\lambda_k(\mathcal{G}_h), \lambda_k(\mathcal{L}_h)$ are the eigenvalues of $\mathcal{G}_h$, $\mathcal{L}_h$ which are simultaneously diagonalizable by an orthogonal matrix Q. Recall that in (2.14)*, we have $\eta(U) = bh_t\mathscr{M}^{-1}\mathscr{G}_h R(U) - \widetilde{\mathbf{w}}$, then*

$$\|D\eta(U)\| \leq bT\zeta_{a,b,c}(h_t)\mathrm{Lip}(R)\,.$$

And we also have

$$\underline{\sigma} \geq 1 - bT\zeta_{a,b,c}(h_t)\mathrm{Lip}(R), \quad \overline{\sigma} \leq 1 + bT\zeta_{a,b,c}(h_t)\mathrm{Lip}(R).$$

We prove this lemma in Appendix B.2. A direct corollary of Lemma 6 and Theorem 3 is Theorem 7, which not only guarantees the unique solvability of $\widehat{F}(U) = 0$, but also establishes exponential convergence for $\|\widehat{F}(U_t)\|$.

Theorem 7 (Unique existence of the root & the second convergence result of $\|\widehat{F}(U_t)\|$)**.** *Suppose conditions* (A)*,* (B)*,* (C) *and* (D) *hold. We pick h_t and $T = N_t h_t$ ($N_t \in \mathbb{N}_+$) satisfying*

$$bT\mathrm{Lip}(R)\zeta_{a,b,c}(h_t) < 1. \tag{3.12}$$

Then there exists a unique root of $\widehat{F}$. Furthermore, we denote $\theta = bT\mathrm{Lip}(R)\zeta_{a,b,c}(h_t) < 1$. Suppose we set $\epsilon = \kappa - \frac{1}{2}$ and $\gamma = \frac{1}{\kappa} - \frac{1}{2\kappa^2}$. Then we have

$$\|\widehat{F}(U_t)\| \leq \exp\left(-\frac{5}{32} \cdot \frac{(1-\theta)^3}{1+\theta}\, t\right) \sqrt{\|\widehat{F}(U_0)\|^2 + (1+\theta)\|Q_0\|^2}. \tag{3.13}$$

Proof. The unique existence of the root for $\widehat{F}(\cdot)$ is due to Lemma 15.

We now prove the exponential convergence (3.13). According to Lemma 6, by letting $\theta = bT\mathrm{Lip}(R)\zeta_{a,b,c}(h_t)$, we obtain

$$\underline{\sigma} \geq 1 - \theta, \quad \overline{\sigma} \leq 1 + \theta, \quad \text{and thus} \quad \kappa \leq \frac{1+\theta}{1-\theta}. \tag{3.14}$$

Now recall Theorem 3. To alleviate our discussion, we choose $\delta = \frac{1}{2\kappa}$. After setting the parameters $\epsilon = \kappa - \frac{1}{2}$ and $\gamma = \frac{1}{\kappa} - \frac{1}{2\kappa^2}$, we have

$$\begin{aligned}
\|\widehat{F}(U_t)\| &\leq \exp\left(-\frac{1}{2}(3 - \frac{1}{2\kappa})\frac{\min\{\underline{\sigma}^2, 1\}}{8\kappa}\, t\right) \sqrt{\|\widehat{F}(U_0)\|^2 + \underline{\sigma}^2\|Q_0\|^2} \\
&\leq \exp\left(-\frac{1}{2} \cdot \frac{5}{2} \cdot \frac{(1-\theta)^3}{8(1+\theta)}\, t\right) \sqrt{\|\widehat{F}(U_0)\|^2 + (1+\theta)\|Q_0\|^2} \\
&= \exp\left(-\frac{5}{32} \cdot \frac{(1-\theta)^3}{1+\theta}\, t\right) \sqrt{\|\widehat{F}(U_0)\|^2 + (1+\theta)\|Q_0\|^2},
\end{aligned}$$

where the second inequality is due to (3.14) and the fact that $\kappa \geq 1$. □

We can simplify condition (3.12) for specific types of RD equations. This is summarized in the following Corollary.

Corollary 7.1 (N_x-independent convergence rate for specific RD equations)**.** *Suppose the conditions* (A)*,* (B)*,* (C) *and* (D) *hold. We pick $T = N_t h_t$ ($N_t \in \mathbb{N}_+$) such that*

- *(Allen-Cahn type, $\mathcal{G}_h = I$, $\mathcal{L}_h$ is self-adjoint, non-negative definite) $T < \frac{1}{b\mathrm{Lip}(R)}$, or equivalently, pick $h_t < \frac{1}{b\mathrm{Lip}(R)}$ and $N_t \leq \left\lfloor \frac{1}{b\mathrm{Lip}(R)h_t} \right\rfloor$. We denote $\widetilde{\theta} = b\mathrm{Lip}(R)T < 1$.*
- *(Cahn-Hilliard type, $\mathcal{G}_h = \mathcal{L}_h$ are self-adjoint, and non-negative definite) $T < \frac{2\sqrt{ah_t}+bch_t}{b\mathrm{Lip}(R)}$, or equivalently, pick $h_t < \frac{4a}{b^2(\mathrm{Lip}(R)-c)_+^2}$ and $N_t \leq \left\lfloor \frac{2\sqrt{a/h_t}+bc}{b\mathrm{Lip}(R)} \right\rfloor$. We denote $\widetilde{\theta} = \frac{b\mathrm{Lip}(R)T}{2\sqrt{ah_t}+bch_t} < 1$.*

Suppose further that $\epsilon = \kappa - \frac{1}{2}$ *and* $\gamma = \frac{1}{\kappa} - \frac{1}{2\kappa^2}$*, then* $\|\widehat{F}(U_t)\|$ *convergences to* 0 *exponentially fast,*

$$\|\widehat{F}(U_t)\| \le \exp\left(-\frac{5}{32}\cdot\frac{(1-\widetilde{\theta})^3}{1+\widetilde{\theta}}\, t\right)\sqrt{\|\widehat{F}(U_0)\|^2 + (1+\widetilde{\theta})\|Q_0\|^2}. \tag{3.15}$$

Proof. Recall that we have $\theta = bT\mathrm{Lip}(R)\zeta_{a,b,c}(h_t)$. We prove $\widetilde{\theta} \ge \theta$ under both cases.

- (Allen-Cahn type) Note that $\zeta_{a,b,c}(h_t) = \max_k \left\{\frac{1}{1+h_t(a\lambda_k(\mathcal{L}_h)+bc)}\right\} \le 1$. Thus,

$$\theta = bT\mathrm{Lip}(R)\zeta_{a,b,c}(h_t) \le bT\mathrm{Lip}(R) = \widetilde{\theta}.$$

- (Cahn-Hilliard type) We have

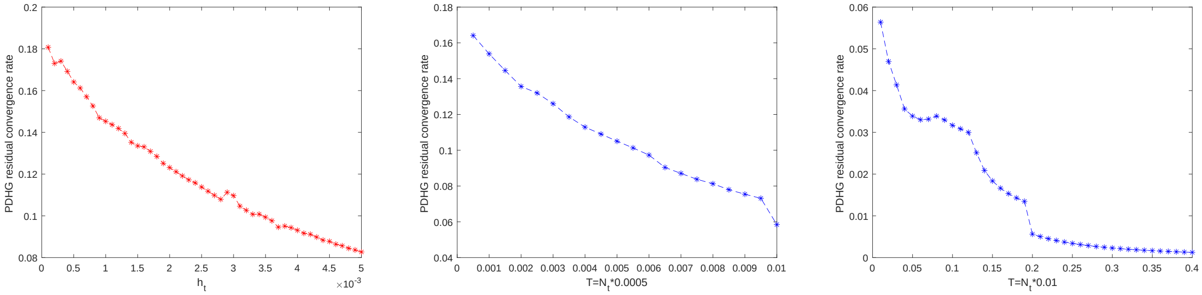
$$\begin{aligned}\zeta_{a,b,c}(h_t) &= \max_k \left\{\left(\frac{1}{\lambda_k(\mathcal{G}_h)} + h_t a\lambda_k(\mathcal{L}_h) + h_t bc\right)^{-1}\right\}\\ &= \max_k \left\{\left(\frac{1}{\lambda_k(\mathcal{L}_h)} + h_t a\lambda_k(\mathcal{L}_h) + h_t bc\right)^{-1}\right\}\\ &\le \frac{1}{2\sqrt{ah_t} + bch_t}.\end{aligned}$$

  Then,

$$\theta = bT\mathrm{Lip}(R)\zeta_{a,b,c}(h_t) \le \frac{bT\mathrm{Lip}(R)}{2\sqrt{ah_t} + bch_t} = \widetilde{\theta}.$$

Since $\widetilde{\theta} < 1$ in both cases, we have $\theta \le \widetilde{\theta} < 1$. Applying Theorem 7 yields (3.13). Note that $\frac{(1-\theta)^3}{1+\theta} \ge \frac{(1-\widetilde{\theta})^3}{1+\widetilde{\theta}}$ for $0 \le \theta \le \widetilde{\theta} < 1$. This implies our result (3.15). □

3.2.3. *Numerical verification.* We apply our algorithm to solve the Allen-Cahn equation (4.1) with $\epsilon_0 = 0.01$ on a 64×64 grid. We use $\tau_U = \tau_P = 0.5$, $\omega = 1$, $\epsilon = 0.1$. At each iteration k, denote U_k as the numerical solution. We define $r_k = -\log_{10}(\|\widehat{F}(U_{k+1})\|/\|\widehat{F}(U_k)\|)$ to be the convergence rate of the residual term $\|\widehat{F}(U_k)\|$ at kth iteration. The residual is expected to converge linearly to 0. We denote by $\bar{r}$ the average convergence rate of the first 500 iterations. By (3.15), when $\widetilde{\theta}$ is small, the convergence rate is $\frac{5}{32}(1 - 4\widetilde{\theta} + \mathcal{O}(\widetilde{\theta}^2))$, which is linear w.r.t. $N_t h_t$ (recall that $\widetilde{\theta} \propto T = N_t h_t$). Such linear relation is verified in the first two figures of Figure 1. In the third figure, we observe fast decay of the average convergence rate $\bar{r}$ as $\widetilde{\theta} \propto N_t h_t$ keeps increasing. Furthermore, we verify the dependence of the convergence rate on $N_t h_t$ via the left plot of Figure 4. We also apply our algorithm



(A) Plot of $\bar{r}$ vs h_t. Fix $N_t = 1$, $h_t = 10^{-4}k$, $1 \le k \le 50$.

(B) Plot of $\bar{r}$ vs N_t. Fix $h_t = 5 \times 10^{-4}$, $1 \le N_t \le 20$.

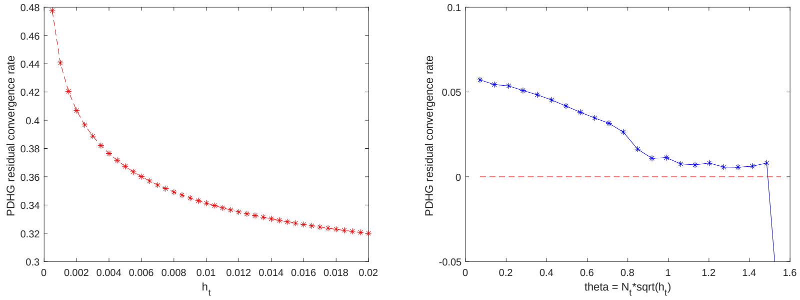
(C) Plot of $\bar{r}$ vs N_t. Fix $h_t = 10^{-2}$, $1 \le N_t \le 40$.

FIGURE 1. Convergence rate of the residual term $\|\widehat{F}(U_k)\|$ w.r.t. h_t, N_t for Allen-Cahn equation.

to the Cahn-Hilliard equation (4.2) with $\epsilon_0 = 0.1$ on a 64×64 grid. We keep the hyperparameters the same as in the case of Allen-Cahn. The average convergence rate $\bar{r}$ is computed by the first 500 iterations of the algorithm. By (3.15), the convergence rate is linear w.r.t. $N_t(\sqrt{h_t} + o(\sqrt{h_t}))$ when $\widetilde{\theta} \propto N_t\sqrt{h_t}$ is small. This is reflected in Figure 2. Unlike the case of Allen-Cahn, in which the PDHG algorithm converges as $\widetilde{\theta}$ increases, the iterations for Cahn-Hilliard diverges as $\widetilde{\theta} \propto N_t\sqrt{h_t}$ increases. This is reflected on the right plot of Figure 2.

For a fixed time step size h_t, denote by $N_{\max}$ the maximum number of time steps that guarantees the convergence of the PDHG algorithm. We plot the relation between $N_{\max}$ and h_t on a logarithmic scale in Figure 3. We observe the relation $N_{\max} = \mathcal{O}(\frac{1}{\sqrt{h_t}})$ when the step size h_t is not too small. The dependence of the convergence rate w.r.t. $N_t\sqrt{h_t}$ is shown in the right plot of Figure 4.

Remark 7. *It is worth mentioning that some of the tested values of h_t in Figure 2, 3 may have exceeded the theoretical bounds for uniqueness (cf. Remark 4) and convergence (cf. 8.1). We point out that these bounds, derived from ensuring convexity and positive definiteness in the numerical analysis, are sufficient but not necessary. The figures primarily aim to illustrate the dependence of convergence rate on $N_t \cdot h_t$ and $N_t \cdot \sqrt{h_t}$, rather than strictly adhering to the bounds.*



(A) Plot of $\bar{r}$ vs h_t. Fix $N_t = 1$, $h_t = 5 \times 10^{-4}k$, $1 \le k \le 40$.

(B) Plot of $\bar{r}$ vs N_t. Fix $h_t = 0.005$, $1 \le N_t \le 22$.

FIGURE 2. Convergence rate of the residual term $\|\widehat{F}(U_k)\|$ w.r.t. h_t, N_t for Cahn-Hilliard equation.

3.3. Lyapunov analysis for the time-discrete case.

In this section, we discuss the convergence of the time-discrete PDHG algorithm (2.16). Recall that the equilibrium state of the PDHG dynamic (2.16) is $(U_*, 0)$ with $\widehat{F}(U_*) = 0$, we consider the following Lyapunov function

$$\mathcal{J}(U, Q) = \frac{1}{2}(\|U - U_*\|^2 + \|Q - 0\|^2) = \frac{1}{2}(\|U - U_*\|^2 + \|Q\|^2).$$

The next theorem provides a sufficient condition on the convergence of $\mathcal{J}$ when $f(\cdot)$ is Lipschitz.

Theorem 8 (Exponential convergence of the PDHG algorithm (2.16))**.** *Consider the following assumptions,*

- *(On PDE* (1.1)*) Assume* (A), (B) *hold.*
- *(On numerical scheme* (2.1) *of PDE) Assume* (C) *holds. Suppose the time step size h_t and $T = N_t h_t$ satisfy $bT\mathrm{Lip}(R)\zeta_{a,b,c}(h_t) < \sqrt{2} - 1$. Suppose we pick $\theta \ge bT\mathrm{Lip}(R)\zeta_{a,b,c}(h_t)$ with $\theta < \sqrt{2} - 1$.*
- *(On PDHG algorithm* (2.16)*) Suppose* (D) *holds. There exist $\widetilde{\gamma} = \omega\tau_P, \varrho = \frac{\tau_P}{\tau_U}, \epsilon > 0$ satisfying*

$$\varrho\widetilde{\gamma}\epsilon\Psi(\theta) - \frac{1}{4}\Omega(\widetilde{\gamma}\epsilon, \varrho, \theta)^2 > 0. \tag{3.16}$$

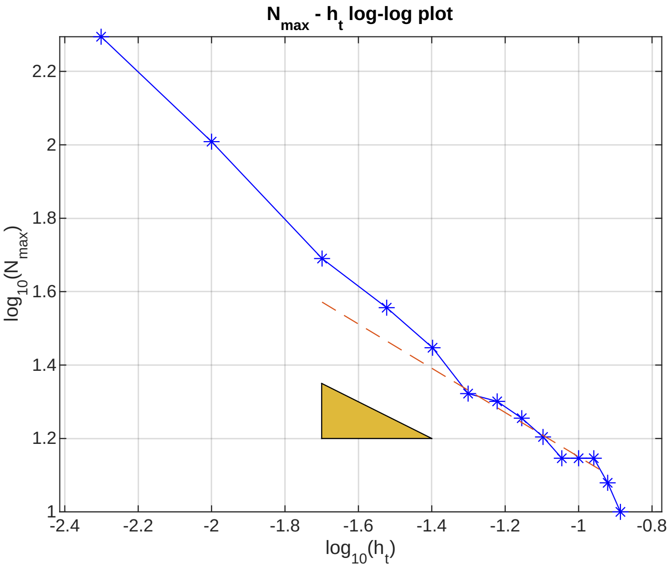
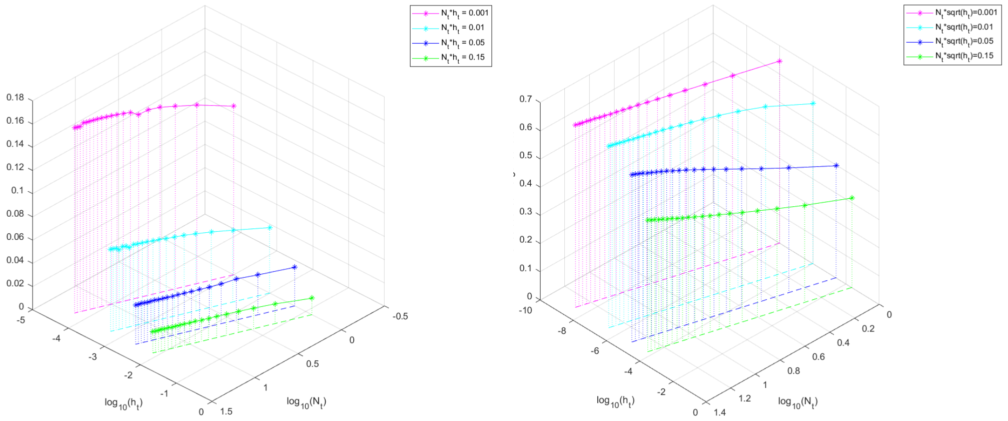


FIGURE 3. $N_{\max} - h_t$ log-log plot for Cahn-Hilliard equation (4.2). We solve the equation on 64 grid with $h_t = 0.01 \cdot k$, $k = 0.5, 1, 2, \ldots, 13$. The yellow triangle has slope equals to $\frac{1}{2}$. The orange dashed line is the linear regression of data points with rather large $h_t = 0.01 \cdot k$ with $5 \le k \le 11$.



(A) We solve Allen-Cahn equation (4.1) Plot of $\bar{r}$ vs $(\log_{10} N_t, \log_{10} h_t)$, with $N_t h_t = 0.15, 0.05, 0.01, 0.001$.

(B) We solve Cahn-Hilliard equation (4.2) Plot of $\bar{r}$ vs $(\log_{10} N_t, \log_{10} h_t)$, with $N_t \sqrt{h_t} = 0.15, 0.05, 0.01, 0.001$.

FIGURE 4. Plots of $\bar{r}$ vs (N_t, h_t).

Here we denote $\Psi(\theta) = 1 - 2\theta - \theta^2$, *and* $\Omega(u, \varrho, \theta) = |1 - u - \varrho| + (|1 - u| + \varrho)\theta$. *We choose PDHG step size for the dual variable as*

$$\tau_P = \frac{\varrho\widetilde{\gamma}\epsilon\Psi(\theta) - \frac{1}{4}\Omega(\widetilde{\gamma}\epsilon, \varrho, \theta)^2}{4(\widetilde{\gamma} + \varrho\epsilon)(1 + \theta)^2 \max\{\widetilde{\gamma}^2(1 + \theta)^2, (1 - \widetilde{\gamma}\epsilon)^2\}}, \tag{3.17}$$

and set the extrapolation coefficient $\omega = \frac{\widetilde{\gamma}}{\tau_P}$, *the PDHG step size for* U *as* $\tau_U = \frac{\tau_P}{\varrho}$.

Under the above conditions, there exists a unique U_* s.t. $\widehat{F}(U_*) = 0$. Furthermore, assume that $\{U_k, Q_k\}$ solves the PDHG algorithm (2.16) *with arbitrary initial condition* (U_0, Q_0)*. Write $\mathcal{J}_k = \mathcal{J}(U_k, Q_k)$. We have*

$$\mathcal{J}_k \leq \left(\frac{2}{\Phi + \sqrt{\Phi^2+4}}\right)^{k+1} \left(\mathcal{J}_1 + \frac{\Phi + \sqrt{\Phi^2+4}}{2}\mathcal{J}_0\right), \tag{3.18}$$

where

$$\Phi = \frac{(\varrho\widetilde{\gamma}\epsilon\Psi(\theta) - \frac{1}{4}\Omega(\widetilde{\gamma}\epsilon, \varrho, \theta)^2)^2}{2(1+\theta)^2 \max\{\widetilde{\gamma}^2(1+\theta)^2, (1-\widetilde{\gamma}\epsilon)^2\}(\widetilde{\gamma}+\varrho\epsilon)^2}.$$

The proof of the theorem is provided in Appendix B.3.

We can simplify the results in Theorem 8 for Allen-Cahn and Cahn-Hilliard type of equations, using similar argument in the proof of Corollary 7.1, for Allen-Cahn (resp., Cahn-Hilliard) type equations. Suppose $b\mathrm{Lip}(R)T < \sqrt{2}-1$ (resp., $\frac{b\mathrm{Lip}(R)T}{2\sqrt{ah_t}+bch_t} < \sqrt{2}-1$). If we set $\theta = b\mathrm{Lip}(R)T$ (resp., $\theta = \frac{b\mathrm{Lip}(R)T}{2\sqrt{ah_t}+bch_t}$), then we have $bT\zeta_{a,b,c}(h_t)\mathrm{Lip}(R) \leq \theta < \sqrt{2}-1$.

Furthermore, we can pick specific values of the hyperparameters $\tau_U, \tau_P, \omega, \epsilon$ to obtain a more concise convergence rate Φ. To do so, we denote $u = \widetilde{\gamma}\epsilon$ and assume that $u < 1$. We set $\varrho = 1 - \widetilde{\gamma}\epsilon = 1 - u$. Then the condition (3.16) leads to $(1-u)u\Psi(\theta) - (1-u)^2\theta^2 > 0$, which yields $\frac{\theta^2}{1-2\theta} < u < 1$. Furthermore, the rate Φ equals

$$\Phi = \frac{(1-u)^2(u(1-2\theta-\theta^2) - (1-u)\theta^2)^2}{2(1+\theta)^2\max\{\widetilde{\gamma}^2(1+\theta)^2, (1-u)^2\}(\widetilde{\gamma} + (1-u)\epsilon)^2}.$$

We further pick $\widetilde{\gamma} = (1-u)\epsilon$. Together with $\widetilde{\gamma}\epsilon = u$, we have $\widetilde{\gamma} = \sqrt{u(1-u)}$, $\epsilon = \sqrt{\frac{u}{1-u}}$. Thus,

$$\Phi = \frac{(1-2\theta)^2}{8(1+\theta)^2} \cdot \frac{\left(1 - \frac{\theta^2}{1-2\theta}\cdot\frac{1}{u}\right)^2}{\max\{(1+\theta)^2, (1-u)/u\}}.$$

Now the value of τ_P is determined by (3.17), $\tau_U = \frac{\tau_P}{\varrho}$, $\omega = \frac{\widetilde{\gamma}}{\tau_P}$ can also be determined. In summary, we have the following Corollary.

Corollary 8.1 (N_x-independent convergence rate for specific RD equations)**.** *Suppose* (A), (B), (C), *and* (D) *hold. Assume h_t, N_t and $T = N_t h_t$ satisfy*

- *(Allen-Cahn type, $\mathcal{G}_h = I$, $\mathcal{L}_h$ is self-adjoint, non-negative definite) Pick $T < \frac{\sqrt{2}-1}{b\mathrm{Lip}(R)}$, or equivalently, $h_t < \frac{\sqrt{2}-1}{b\mathrm{Lip}(R)}$, $N_t \leq \left\lfloor \frac{\sqrt{2}-1}{b\mathrm{Lip}(R)h_t} \right\rfloor$. We denote $\theta = b\mathrm{Lip}(R)T < \sqrt{2}-1$;*
- *(Cahn-Hilliard type, $\mathcal{G}_h = \mathcal{L}_h$ is self-adjoint, and non-negative definite) Pick $T < \frac{(\sqrt{2}-1)(2\sqrt{ah_t}+bch_t)}{b\mathrm{Lip}(R)}$, or equivalently, $h_t < \frac{4(\sqrt{2}-1)^2 a}{b^2(\mathrm{Lip}(R)-(\sqrt{2}-1)c)_+^2}$, $N_t \leq \left\lfloor (\sqrt{2}-1)\frac{2\sqrt{a/h_t}+bc}{b\mathrm{Lip}(R)} \right\rfloor$. We denote $\theta = \frac{b\mathrm{Lip}(R)T}{2\sqrt{ah_t}+bch_t} = \frac{b\mathrm{Lip}(R)N_t\sqrt{h_t}}{2\sqrt{a}+bc\sqrt{h_t}} < \sqrt{2}-1$.*

Then, there is unique U_* with $\widehat{F}(U_*) = 0$. Furthermore, if we choose $u \in (\frac{\theta^2}{1-2\theta}, 1)$ and set

$$\tau_P = \frac{u(1-2\theta)-\theta^2}{8\sqrt{u(1-u)}(1+\theta)^2\max\{u(1+\theta)^2, 1-u\}}, \quad \tau_U = \frac{\tau_P}{1-u}, \quad \omega = \frac{\sqrt{u(1-u)}}{\tau_U}, \quad \epsilon = \sqrt{\frac{u}{1-u}}, \tag{3.19}$$

then U_k converges exponentially fast to U_*, i.e.,

$$\|U_k - U_*\|^2 \leq C_0 \left(\frac{2}{\Phi + \sqrt{\Phi^2+4}}\right)^{k+1}.$$

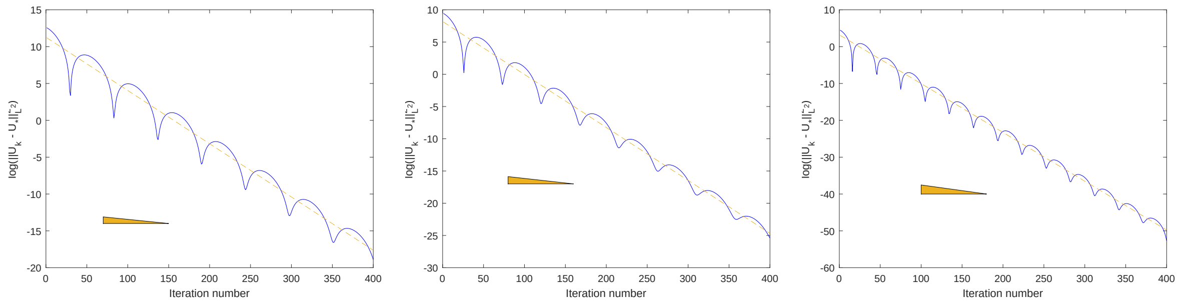
Here

$$C_0 = \left(\mathcal{J}_1 + \frac{\Phi + \sqrt{\Phi^2+4}}{2}\mathcal{J}_0\right), \quad \Phi = \frac{(1-2\theta)^2}{8(1+\theta)^2} \cdot \frac{\left(1 - \frac{\theta^2}{1-2\theta}\cdot\frac{1}{u}\right)^2}{\max\{(1+\theta)^2, (1-u)/u\}}.$$

In the following example, we pick the hyperparameters $h_t, N_t, \tau_U, \tau_P, \omega, \epsilon$ according to Corollary 8.1, and apply it to different types of equations. Our algorithm is guaranteed to converge linearly. The theoretical results presented in Theorem 8 and Corollary 8.1 are not necessarily the sharpest convergence rate. In practice, the actual convergence rate of our PDHG method is generally faster than the theoretical guarantee in Corollary 8.1. This is reflected in the following Table 1. When composing Table 1, recall that $f(u) = u^3 - u$, we set $c = f'(\pm 1) = 2$, and $R(u) = f(u) - cu = u^3 - 3u$. In our numerical result, we observe that $|U_{ij}^t| \le 1$ for any spatial index (i,j) and temporal index t. Thus we use $\sup_{u\in[-1,1]} |R'(u)| = 3$ as the value of $\mathrm{Lip}(R)$ in Corollary 8.1 during the calculation.

		h_t	N_t	u	τ_P	τ_U	ω	ϵ	$\widetilde{\theta}$	Theoretical rate	Actual rate
AC(4.1)	$\epsilon_0 = 1.0$	0.005 < 0.1381	20 ≤ 27	0.5 $u \in (0.2250, 1)$	0.0498	0.0996	5.0181	1.0	0.3000	0.0112	0.0723
	$\epsilon_0 = 0.1$	0.001 < 0.0138	7 ≤ 13	0.5 $u \in (0.0760, 1)$	0.0574	0.1147	4.3587	1.0	0.2100	0.0141	0.0821
	$\epsilon_0 = 0.01$	0.0005 (< 0.0014)	1 (≤ 2)	0.5 $(u \in (0.0321, 1))$	0.0936	0.1872	2.6702	1.0	0.1500	0.0307	0.1325
CH(4.2)	$\epsilon_0 = 10$	0.005 (<1.4553)	10 (≤ 12)	0.5 $(u \in (0.04, 1))$	0.0842	0.1684	2.9695	1.0	0.1640	0.0260	0.0537
	$\epsilon_0 = 1.0$	0.001 (<0.1455)	5 (≤ 9)	0.5 $(u \in (0.0978, 1))$	0.0475	0.0949	5.2662	1.0	0.2874	0.0103	0.0301
	$\epsilon_0 = 0.1$	0.0005 (<0.0015)	1 (≤ 1)	0.5 $(u \in (0.1663, 1))$	0.0286	0.0572	8.7392	1.0	0.2741	0.0043	0.0169

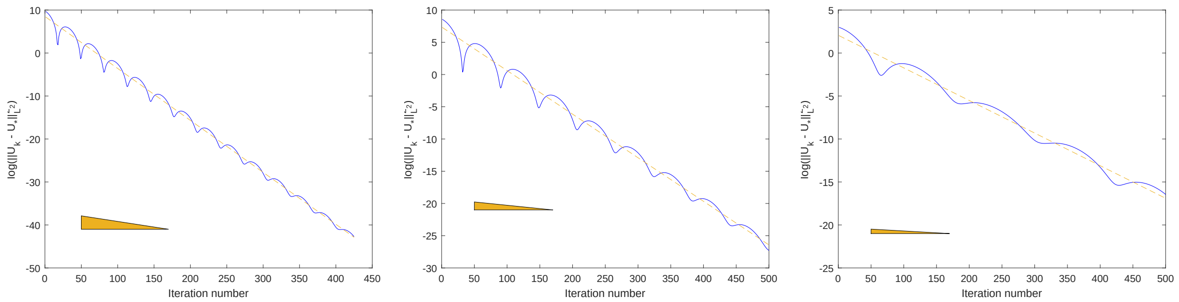
Table 1. Theoretical convergence rate vs actual convergence rate of $\|U_k - U_*\|_2^2$. The constraints in the parentheses in the columns of h_t, N_t, and u are derived from the conditions in Corollary 8.1. The actual rate r is solved from the linear regression model $r \cdot k + b$ given the numerical data $\{k, \log(\|U_{k+1} - U_*\|^2 / \|U_k - U_*\|^2)\}$ for $1 \le k \le 400$ (Allen-Cahn equation (4.1)); and $1 \le k \le 500$ (Cahn-Hilliard equation (4.2)).



(A) $\epsilon_0 = 1.0$. (B) $\epsilon_0 = 0.1$. (C) $\epsilon_0 = 0.01$.

Figure 5. Plot of $\log \|U_k - U_*\|^2$ vs k $(1 \le k \le 400)$ when using hyperparameters specified in Table 1 to solve Allen-Cahn equation (4.1) with different ϵ_0 on a 128×128 grid.

Remark 8. (3.19) *may also not be the optimal choice of hyperparameters. We provide suggestions on selecting the optimal hyperparameters in section 4.3.*



(A) $\epsilon = 10$. (B) $\epsilon = 1.0$. (C) $\epsilon = 0.1$.

FIGURE 6. Plot of $\log \|U_k - U_*\|^2$ vs k $(1 \leq k \leq 500)$ when using hyperparameters specified in Table 1 to solve Cahn-Hilliard equation (4.2) with different ϵ_0 on a 128×128 grid.

4. Numerical examples

In this section, we test the proposed algorithm on four types of RD equations, namely the Allen-Cahn equation, the Cahn-Hilliard equation, an RD equation with variable coefficients (mobility term), and a 6th-order reaction-diffusion equation. We verify the independence between the convergence rate of our algorithm and the grid size N_x. We discuss how the hyperparameters of the proposed algorithm are chosen to achieve the optimal (or near-optimal) performance via numerical experiments. We also provide comparisons between the implicit scheme with adaptive step size h_t and the IMEX scheme on long-time range computation. At the end of this section, we make comparisons with three commonly used algorithms for resolving the time-implicit schemes such as the nonlinear SOR [38], the preconditioned fixed point method [2] and Newton's method [11].

For all the numerical examples in this section, if not specified, we always set the hyperparameters $\omega = 1$ and $\epsilon = 0.1$. We terminate the iteration whenever $\|\mathrm{Res}(U_k)\|_\infty < tol$ with $tol = 10^{-6}$. Here the residual term $\mathrm{Res}(U_k)$ is defined in (2.17). All numerical examples are imposed with periodic boundary conditions. We adopt the central discretization scheme to discretize the Laplace operator Δ, i.e., we set the discretized Laplace operator as $\mathrm{Lap}^P_{h_x}$ defined in (2.2).

Among four equations discussed in this section, equations (4.1), (4.2), and (4.6) have already been considered in [35], where more numerical results are demonstrated. In this research, we mainly use them as test equations for validating our theoretical findings and justifying the effectiveness of our method.

All the numerical examples are computed using MATLAB on a laptop with 11th Gen Intel Core i5-1135G7 @ 2.40GHz CPU and 16.0 GB RAM. The corresponding codes are provided at `https://github.com/LSLSliushu/PDHG-method-for-solving-reaction-diffusion-equations/tree/main`.

4.1. **Tested equations.**

Throughout this section, we denote the double potential function $W(u) = \frac{1}{4}(u^2-1)^2$, and thus $W'(u) = u^3 - u$.

4.1.1. *Allen-Cahn equation (AC).* We consider the Allen-Cahn equation

$$\frac{\partial u}{\partial t} = a\Delta u - bW'(u), \quad \text{on } [0,0.5]^2 \times [0,T], \quad u(x,0) = u_0(x). \tag{4.1}$$

We set $a = \epsilon_0, b = \frac{1}{\epsilon_0}$ with $\epsilon_0 = 0.01$. We set the initial condition as $u_0 = 2\chi_{B(x_*,r)} - 1$ where $x_* = (0.25, 0.25), r = 0.2$. For the precondition matrix $\mathscr{M}$, $\mathcal{G}_h = I$, and $\mathcal{L}_h = \Delta^P_{h_x}$, and $J_f = 2I$. We compare our method and the IMEX method in Figure 7. The zero-level set of the solution $u(\cdot,t)$ of this equation is known to be the curvature flow of a circle [38]. A comparison among the plots of the front positions computed by our method, the Nonlinear SOR method. The real solution is presented on the right-hand side of Figure 7.

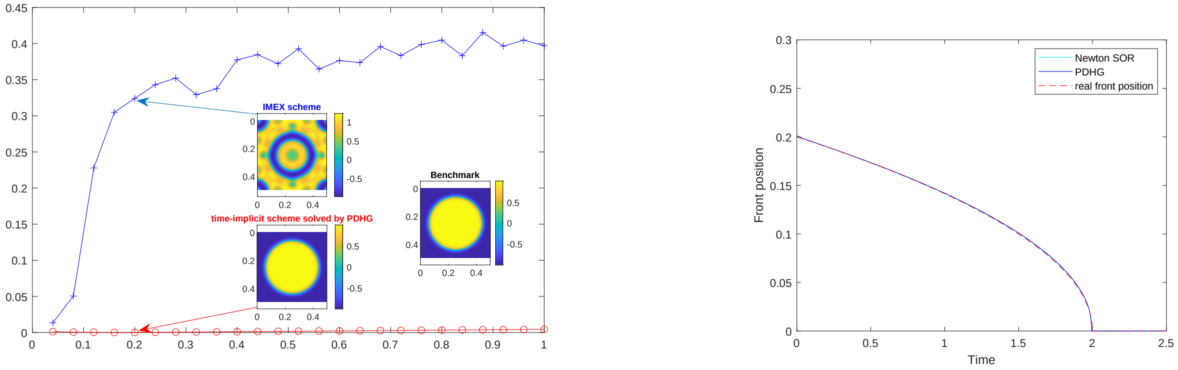


FIGURE 7. We solve equation (4.1) with $\epsilon_0 = 0.01$. We set $\tau_U = 0.55, \tau_P = 0.95$ for our PDHG method. (Left) Comparison between our method (time-implicit scheme solved by the proposed PDHG algorithm) and the IMEX scheme. We discrete the space into 128×128 lattices. We compute both schemes with large time step size $h_t = 0.02$ and compare with the benchmark solution solved from the same IMEX scheme with $h_t = 0.001$. Blue curve indicates the L^1 discrepancy between the IMEX solution on the coarser time grid U_{IMEX} and the benchmark solution $U_\star$. Red curve indicates the L^1 discrepancy between the time-implicit solution U_{PDHG} and $U_\star$. (Right) Comparison between the front position of the numerical solution solved via our PDHG method and the Nonlinear SOR method, as well as the real front position.

4.1.2. *Cahn-Hilliard equation (CH).* We consider the Cahn-Hilliard equation

$$\frac{\partial u}{\partial t} = -a\Delta\Delta u + \Delta b W'(u), \quad \text{on } [0, 2\pi]^2 \times [0, T], \quad u(x, 0) = u_0(x). \tag{4.2}$$

We set $a = \epsilon_0^2$ and $b = 1$. We set the initial condition u_0 as a modified indicator function whose value equals $+1$ if (x, y) falls inside any of the seven circles and -1 otherwise, i.e.,

$$u_0(x, y) = -1 + \sum_{i=1}^{7} \varphi(\sqrt{(x - x_i)^2 + (y - y_i)^2} - r_i),$$

where the mollifier function φ is defined as

$$\varphi(s) = \begin{cases} 2e^{-\frac{\epsilon^2}{s^2}} & s < 0; \\ 0 & s \geq 0 \end{cases}, \quad \text{with } \epsilon = 0.1.$$

The centers and radii of these seven circles are listed in Table 2. For the precondition matrix $\mathscr{M}$,

i	1	2	3	4	5	6	7
x_i	$\pi/2$	$\pi/4$	$\pi/2$	π	$3\pi/2$	π	$3\pi/2$
y_i	$\pi/2$	$3\pi/4$	$5\pi/4$	$\pi/4$	$\pi/4$	π	$3\pi/2$
r_i	$\pi/5$	$2\pi/15$	$\pi/15$	$\pi/10$	$\pi/10$	$\pi/4$	$\pi/4$

TABLE 2. Centers and radius of the 7 circles.

$\mathcal{G}_h = \mathcal{L}_h = \Delta_{h_x}^P$, and $J_f = 2I$.

4.1.3. *A reaction-diffusion equation with variable coefficient (VarCoeff).* We consider the following equation with variable coefficient (mobility term) $\sigma(\cdot)$,

$$\frac{\partial u}{\partial t} = a\nabla\cdot(\sigma(x)\nabla u) - bW'(u), \quad \text{on } [0,2\pi]^2\times[0,T], \quad u(x,0)=u_0(x). \tag{4.3}$$

We choose $a=\epsilon_0, b=\frac{1}{\epsilon_0}$ with $\epsilon_0 = 0.01$. The media $\sigma(x,y) = 1+\frac{\mu}{2}(\sin^2 x+\sin^2 y)$ with $\mu = 5.0$. We set the initial condition $u_0 = \frac{1}{2}(\cos(4x)+\cos(4y))$. We adopt the following time-implicit scheme

$$\frac{U_{ij}^{t+1}-U_{ij}^{t}}{h_t} = \frac{a}{h_x^2}(\sigma_{i+\frac{1}{2},j}(U_{i+1,j}-U_{i,j})-\sigma_{i-\frac{1}{2},j}(U_{i,j}-U_{i-1,j})+\sigma_{i,j+\frac{1}{2}}(U_{i,j+1}-U_{i,j})-\sigma_{i,j-\frac{1}{2}}(U_{i,j}-U_{i,j-1}))-bW'(U_{ij}^{t+1}), \tag{4.4}$$

where $0\leq t\leq N_t-1$, $1\leq i,j\leq N_x$, and $U_{N_x+1,j}=U_{1,j}, U_{0,j}=U_{N_x,j}; U_{i,N_x+1}=U_{i,1}, U_{i,0}=U_{i,N_x}$ for all $1\leq i,j\leq N_x$. And we set $\sigma_{pq}=\sigma((p-1)h_x,(q-1)h_x)$ for any $p,q\in\mathbb{Q}$.

For the precondition matrix $\mathscr{M}$, $\mathcal{G}_h = I$, we approximate $\mathcal{L}_h$ by $-\overline{\sigma}\Delta_{h_x}^P$, whose matrix-vector multiplication and inversion can be efficiently computed via the FFT algorithm. Here $\overline{\sigma} = \frac{1}{|\Omega|}\int_\Omega \sigma(x,y)\,dxdy = 1+\frac{\mu}{2}$ denotes the average of σ over $\Omega=[0,2\pi]^2$. We set $J_f = 2I$. We choose $\tau_U = 0.5, \tau_P = 0.95$ when applying our PDHG method to solve the time-implicit scheme (4.4).

The numerical solutions to (4.3) are provided in Figure 8. A series of residual decay plots throughout our method are demonstrated in Figure 9.

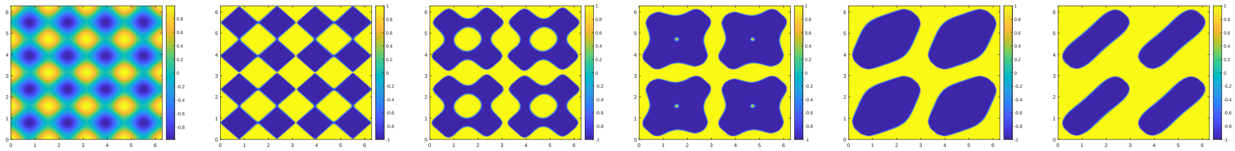
Furthermore, we denote

$$E(u) = \int_\Omega \frac{a}{2}\sigma(x)|\nabla u(x)|^2 + bW(u(x))\,dx,$$

as the free energy functional associated with the reaction-diffusion equation (4.3). Denote

$$E_{h_x}(U) = \sum_{1\leq i,j\leq N_x}\left(\frac{a}{2}(\sigma_{i+\frac{1}{2},j}|U_{i+1,j}-U_{i,j}|^2+\sigma_{i,j+\frac{1}{2}}|U_{i,j+1}-U_{i,j}|^2)+bW(U_{i,j})\right)h_x^2 \tag{4.5}$$

as the discrete analogy of $E(u)$. The free energy $E_{h_x}(U^{t_k})$ versus t_k plot of energy decay is presented in Figure 10. In addition, a comparison between the proposed scheme and the IMEX scheme can be found in Figure 11.



$t=0.0$ $t=0.2$ $t=1.0$ $t=3.6$ $t=10.0$ $t=20.0$

FIGURE 8. Numerical solution of the time-implicit scheme solved via our PDHG method on a 256×256 grid at different time stages $t = 0.0, 0.2, 1.0, 3.6, 10.0, 20.0$.

4.1.4. *A 6th-order Reaction-Diffusion Equation (6th-order).* We consider the following 6th-order Cahn-Hilliard-type equation:

$$\frac{\partial u}{\partial t} = \Delta(\epsilon_0^2\Delta-(W''(u)-\epsilon_0^2)\mathrm{Id})(\epsilon_0^2\Delta u - W'(u)), \quad \text{on } [0,2\pi]^2\times[0,T], \quad u(\cdot,0)=u_0. \tag{4.6}$$

In this example, we choose parameter $\epsilon_0 = 0.18$. We set the initial condition

$$u_0(x,y) = 2e^{\sin x+\sin y-2}+2.2e^{-\sin x-\sin y-2}-1.$$

When we set up the precondition matrix $\mathscr{M}$, we approximate $\mathcal{G}_h$ by

$$\Delta_h(\epsilon_0^2\Delta_h - W''(\pm1)+\epsilon_0^2) = \Delta_h(\epsilon_0^2\Delta_h - 2+\epsilon_0^2),$$

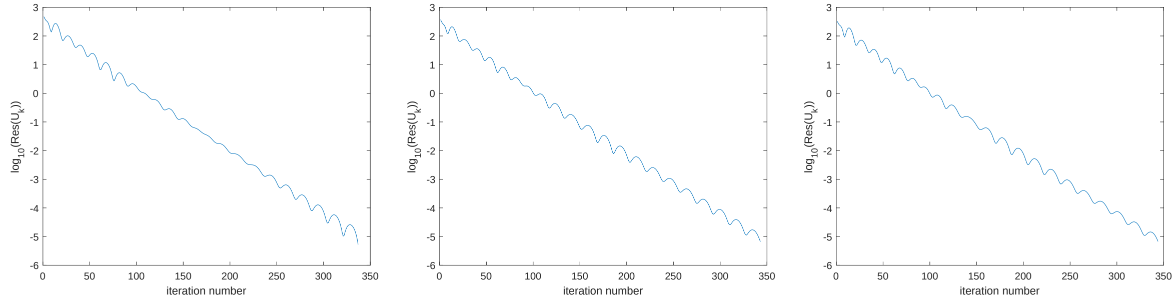


FIGURE 9. The loss plot of $\log_{10}(\text{Res}(U_k))$ vs iteration number k. We solve (4.3) with $h_t = 0.002$. The plots (from left to right) are the loss plots at 30th, 60th, and 90th subinterval.

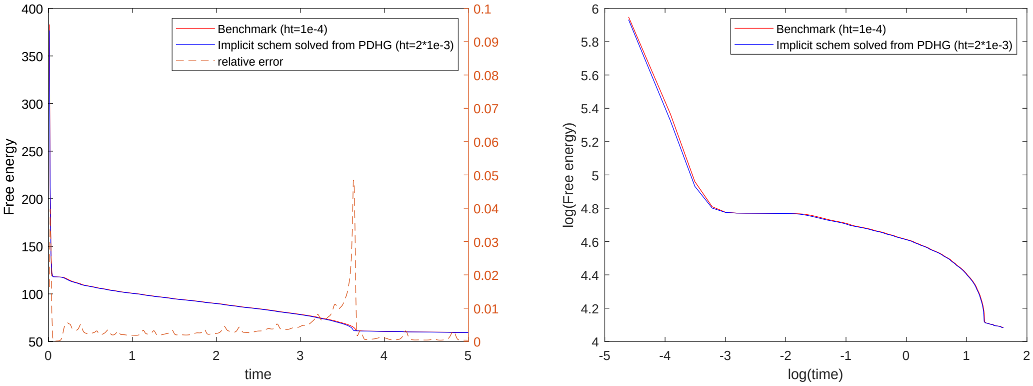


FIGURE 10. We compute the free energy on $[0, 5]$. (Left) Free energy decay (blue) of the time-implicit scheme (solved by PDHG method) with $h_t = 2 \cdot 10^{-3}$, and the reference energy decay (red) solved from IMEX scheme with $h_t = 10^{-4}$. The relative error between them is plotted in orange. (Right) The $\log - \log$ plot of free energy.

and set $\mathcal{L}_h = \epsilon_0^2 \Delta_h$. We pick $J_f = 2I$. We choose $\tau_U = 0.5, \tau_P = 0.95$ for our PDHG method. A comparison between our proposed scheme and the IMEX scheme is provided in Figure 12.

4.1.5. *Grid-size-free algorithm.* As emphasized previously in the introduction, the convergence rate of our algorithm is independent of the grid size N_x. This has also been verified in Corollary 7.1 and Corollary 8.1. (Recall that the quantities $\widetilde{\theta}$ and θ in these corollaries are independent of N_x.) In this subsection, we verify such irrelevance by testing our algorithm on various types of equations with different grid sizes N_x. The numerical results are demonstrated in Figure 13, where the number of iterations required upon convergence directly reflects the convergence rate of our PDHG algorithm.

4.2. **Comparisons with the convex splitting method.** The convex splitting method originally proposed in [15] seeks a specific decomposition of the function $F(\cdot) : \mathbb{R}^n \to \mathbb{R}$, such that the semi-implicit time-discrete scheme applied to the gradient flow

$$\frac{d}{dt}u(t) = -\nabla F(u(t)),$$

is energy stable. To be more specific, suppose F is splitted as $F(u) = F_c(u) - F_e(u)$, where F_c, F_e are convex functions on $\mathbb{R}^n$. Here "c" denotes contraction, and "e" denotes expansion, which indicates

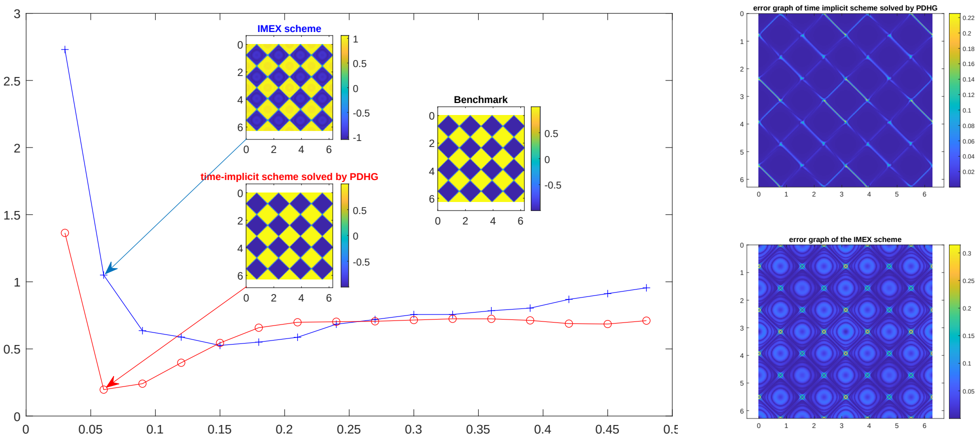


FIGURE 11. (Left) Comparison between our method (time-implicit scheme solved by the proposed PDHG algorithm) and the IMEX scheme. We discretize the space into a 256×256 lattice. We compute both schemes with large time step size $h_t = 0.01$ and compare with the benchmark solution solved from the same IMEX scheme with $h_t = 0.001$. Blue curve indicates the L^1 discrepancy between the IMEX solution on the coarser time grid U_{IMEX} and the benchmark solution $U_\star$. Red curve indicates the L^1 discrepancy between the time-implicit solution U_{PDHG} and the benchmark $U_\star$. (Right) Plot of $|U_{\text{PDHG}} - U_\star|$ (up); and plot of $|U_{\text{IMEX}} - U_\star|$ (down).

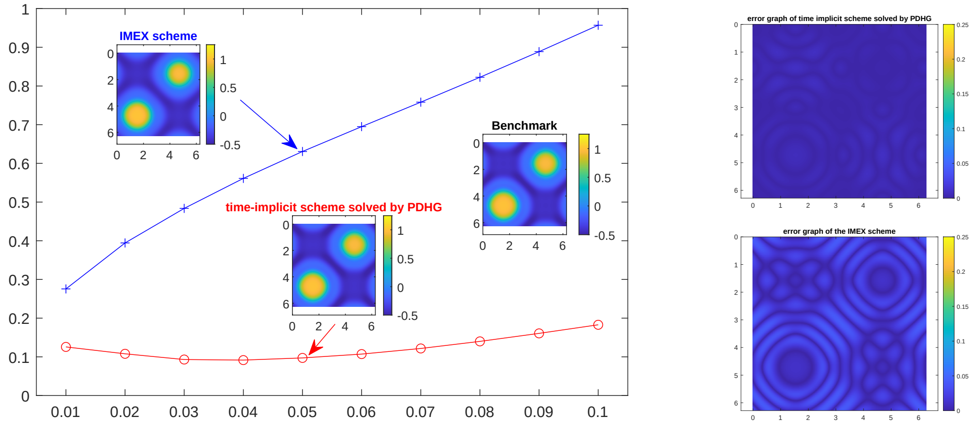


FIGURE 12. Similar to Figure 11. (Left) Comparison between the L^1 discrepancy of our method and the IMEX scheme with $h_t = 0.01$. (Right) Plot of $|U_{\text{PDHG}} - U_\star|$ (up); and $|U_{\text{IMEX}} - U_\star|$ (down).

the effect of the gradient fields ∇F_c and ∇F_e in the gradient flow. Consider the scheme

$$\frac{u^{t+1} - u^t}{2h_t} = -(\nabla F_c(u^{t+1}) - \nabla F_e(u^t)), \quad 0 \le t \le N_t. \tag{4.7}$$

It can be shown that $F(u^{t+1}) \le F(u^t)$ for all $t = 0, 1, 2, \dots,$ i.e., the numerical scheme preserves the decaying of energy.

Many RD equations can be interpreted as gradient flows in certain functional spaces. The convex splitting method has been widely applied to compute these equations. We refer the readers to

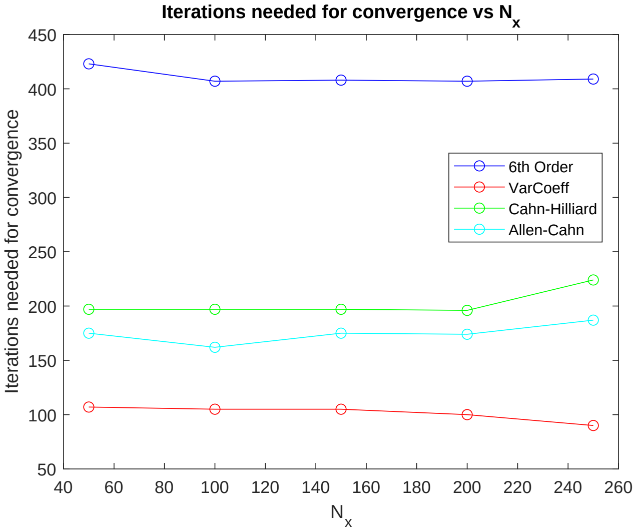


FIGURE 13. Relation between the number of iterations needed for convergence and space discretization N_x. We verify on four different equations with $N_x = 50, 100, 150, 200, 250$. We set $\epsilon_0 = 0.01$ for the Allen-Cahn equation and $\epsilon_0 = 0.1$ for the Cahn-Hilliard equation.

[50] and the references therein for more details. In comparison, we demonstrate that our proposed algorithm, which employs the PDHG method for solving the time-implicit scheme, offers notable advantages over the convex splitting methods. Specifically, it achieves higher accuracy in computing the phase-field models with weak diffusion and strong reaction.

Recall the Allen-Cahn equation (4.1), which can be cast as the L^2 gradient flow of the free energy $\int_\Omega \frac{\epsilon_0}{2}\|\nabla u\|^2\, dx + \frac{1}{\epsilon_0}\int_\Omega W(u)\, dx$. For the numerical solution, we adopt the finite difference scheme and consider the discretized energy function[1] $F(U) = \frac{\epsilon_0}{2}\sum_{i,j} h_x^2 \|\nabla_{h_x} U_{ij}\|^2 + \frac{1}{\epsilon_0}\sum_{i,j} h_x^2 W(U_{ij})$. Following the discussion in [21], we decompose the double-well potential $W(u) = \frac{1}{4}(u^2-1)^2 = W_c(u) - W_e(u)$ in two ways[2]:

$$\text{(A)} \quad W_c(u) = \frac{1}{2}u^2, \quad W_e(u) = -\frac{1}{4}u^4 + \frac{3}{2}u^2 - \frac{1}{4};$$
$$\text{(B)} \quad W_c(u) = \frac{1}{4}u^4 + \frac{1}{4}, \quad W_e(u) = \frac{1}{2}u^2.$$

One then considers $F_c(U) = \frac{\epsilon_0}{2}\sum_{i,j} h_x^2\|\nabla_{h_x}U_{ij}\|^2 + \frac{1}{\epsilon_0}\sum_{i,j} h_x^2 W_c(U_{ij})$ and $F_e(U) = \frac{1}{\epsilon_0}\sum_{i,j} h_x^2 W_e(U_{ij})$. The convex split scheme (4.7) yields

$$(I - \epsilon_0 h_t \Delta_{h_x})U_{ij}^{t+1} + \frac{h_t}{\epsilon_0}W_c'(U_{ij}^{t+1}) = U_{ij}^t + \frac{h_t}{\epsilon_0}W_e'(U_{ij}^t), \quad 0 \le t \le N_t. \tag{4.8}$$

It is worth mentioning that (4.8) reduces to a linear equation if $W_c(\cdot)$ is quadratic. Otherwise, (4.8) is a nonlinear root-finding problem and the proposed PDHG algorithm can be applicable here to resolve for U^{t+1}.

We apply (4.8) using splitting schemes (A) and (B) to (4.1) with $\epsilon_0 = 0.01$ and compare the results with the time-implicit scheme. The numerical results are presented in Figure 14. As shown in the results, a small ϵ_0 in this phase-field model poses a challenge for the convex splitting methods,

[1] Here we denote discrete gradient $\nabla_{h_x} U_{ij} := (\frac{U_{i+1,j}-U_{i-1,j}}{h_x}, \frac{U_{i,j+1}-U_{i,j-1}}{h_x})$.

[2] Although $W_e(\cdot)$ for scheme (A) is not convex on $\mathbb{R}$, it is convex on the finite interval $[-1.7, 1.7]$. This remains a reasonable splitting as long as U_{ij}^t lies in this interval for arbitrary $1 \le i, j \le N_x$, $0 \le t \le N_t$.

as they are unable to accurately capture the movement of the zero-level set of $u(\cdot,t)$. In contrast, the time-implicit scheme maintains computational accuracy. Further comparisons between the time-implicit scheme and the convex splitting method can be found in [50].

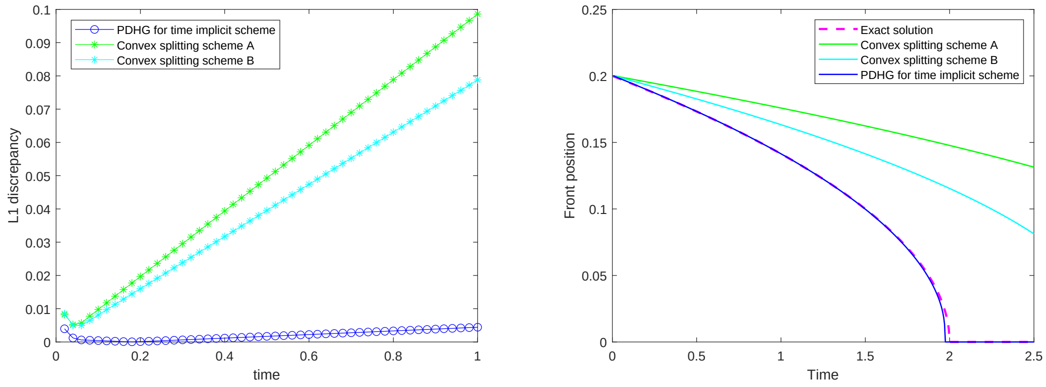


FIGURE 14. Comparison between the time implicit scheme and the convex splitting scheme: (Left) We discrete the space into 128 × 128 lattices. Similar to Figure 7, we compute both schemes with rather large time step size $h_t = 0.02$ and plot the L^1 discrepancy curve versus time (with the benchmark solution solved from the IMEX scheme with $h_t = 0.001$). (Right) We discretize the space into 256×256 lattices, pick $h_t = 0.005$, and plot the front position versus time for different numerical schemes.

4.3. **Hyperparameter selection.** Given the spatial and the temporal step sizes h_x, h_t of the implicit scheme, there are 5 hyperparameters to be determined for our algorithm: $N_t, \tau_U, \tau_P, \omega$, and ϵ. In the following, we discuss the choice of these hyperparameters.

(1) (Choosing N_t) As mentioned previously in section 2.3, one can distribute the computational task into multiple blocks and apply PDHG algorithm to evaluate each block of solutions sequentially. Suppose we aim to solve an equation on $[0, T_{\text{total}}]$. We may divide the time interval into $M \cdot N_t$ subintervals, i.e.,

$$[0,T_{\text{total}}] = \bigcup_{k=1}^{M} I_k = \bigcup_{k=1}^{M} \left(\bigcup_{j=1}^{N_t} I_{k,j} \right), \quad \text{where each } I_{k,j} = [(k-1)T + (j-1)h_t, (k-1)T + jh_t].$$

$$\text{with } T = T_{\text{total}}/M, h_t = T/N_t.$$

We then apply our proposed method to each subinterval I_k in order to obtain the entire numerical solution on $[0, T_{\text{total}}]$. We test our algorithm with different combinations of $M \cdot N_t$ on various types of equations. Unless specified otherwise, we choose $\omega = 1, \epsilon = 0.1$. We set the stopping criteria as $\|\mathrm{Res}(U_k)\|_\infty < 10^{-6}$. The efficiency of our algorithm under different scenarios is reflected in CPU time demonstrated in Table 3. Among the series of experiments, we observe that it is usually efficient to pick $N_t \leq 5$.

(2) (Choosing τ_U, τ_P) Theoretically, choosing τ_U, τ_P as suggested in Corollary 8.1 will guarantee the convergence of our method. In practice, we can pick a larger τ_U, τ_P to achieve faster convergence. Generally speaking, the optimal step size τ_P is around 0.9, and the optimal ratio $\varrho = \frac{\tau_P}{\tau_U}$ should be slightly less than 2. The intuition of choosing $\varrho > 1$ is that we want to treat the inner optimization of the functional $\widehat{L}(U,Q)$ defined in (2.15) w.r.t. the dual variable Q more thoroughly. In fact, it is common in bi-level optimization to choose a larger, more aggressive step size for the inner-level optimization problem both practically [17] and theoretically [30, 55]. A rather efficient choice of the

Equation Name $[\tau_U, \tau_P, T_{\text{total}}]$	$M \times N_t$								
	1×100	2×50	4×25	10×10	20×5	25×4	$33 \times 3 + 1$	50×2	100×1
AC($\epsilon_0 = 0.01$) $[0.5, 0.5, 1.0]$	–	–	1198.41	219.52	137.71	138.65	**88.53**	106.41	92.72
AC($\epsilon_0 = 0.1$) $[0.5, 0.5, 1.0]$	–	–	90.28	57.73	34.37	50.43	41.37	26.62	**24.20**
AC($\epsilon_0 = 1$) $[0.5, 0.5, 1.0]$	64.28	38.11	23.42	24.24	13.05	13.29	12.51	10.89	**10.72**
CH $[0.5, 0.5, 1.0]$	775.15	208.93	170.77	252.99	148.96	183.34	101.41	**77.35**	86.37
6th Order $[0.8, 0.8, 0.1]$	–	–	374.82	389.90	285.12	384.52	199.11	**188.58**	208.30
Varcoeff $[0.95, 0.5, 1.0]$	–	–	305.73	206.72	204.34	153.88	144.67	142.22	**61.46**

Table 3. Comparison of CPU time (s) with different N_ts (All problems are solved on 256×256 grids).

step sizes (τ_U, τ_P) is $(0.5, 0.9)$. This is verified in Table 4, in which we compare the choice $(0.5, 0.9)$ with other combinations of (τ_U, τ_P).

$\epsilon = 0.1$ for all problems		$\tau_U = 0.9, \tau_P = 0.5$	$\tau_U = 0.65, \tau_P = 0.65$	$\tau_U = 0.5, \tau_P = 0.9$
6th Order $[T = 0.5]$	$N_x = 256, N_t = 50$	62.28	47.92	**30.53**
	$N_x = 128, N_t = 50$	12.31	9.47	**8.54**
VarCoeff $[T = 0.5]$	$N_x = 256, N_t = 50$	103.23	109.38	**82.38**
	$N_x = 128, N_t = 50$	15.92	13.35	**9.54**

Table 4. Comparison on speeds among different ratios $\varrho = \frac{\tau_P}{\tau_U}$ for different equations.

(3) (Choosing ω) We pick $\omega = 1$ in our experiments. If one increases or decreases ω, one should modify τ_P correspondingly so that $\widetilde{\gamma} = \omega\tau_P$ remains unchanged. Once $\widetilde{\gamma} \approx 0.9$ is fixed, we generally achieve the optimal (or near-optimal) performance of our algorithm.

(4) (Choosing ϵ) We set ϵ around 0.1. Recall that $\sup_Q \{\widehat{L}(U, Q)\} = \frac{\|\widehat{F}(U)\|^2}{2\epsilon}$. Increasing ϵ will decrease the convexity of the functional $\frac{\|\widehat{F}(U)\|^2}{2\epsilon}$, which will slow down our algorithm. Decreasing ϵ brings our algorithm closer to our original version of PDHG method [35], in which we discover stronger oscillations towards convergence, which may also affect the efficiency.

4.4. **Long-time computation via adaptive time step size.** It is an important topic how one can efficiently compute the RD equation for large time T to study its behavior near the equilibrium state. Since we can pick large time step size h_t under the implicit scheme, our proposed method offers an opportunity for faster computations to approximate the equilibrium state of RD equations.

To be more precise, we adopt adaptive time step size h_t during the update of time-implicit scheme (2.1). Suppose we set up an upper bound $\bar{h}_t > 0$ for time step size h_t. As $h_t < \bar{h}_t$, we increase h_t by 10% if the proposed PDHG algorithm converges in less than $\bar{n}$ steps. Otherwise, we decrease h_t by 50%. If h_t exceeds $\bar{h}_t$, we reset $h_t = \bar{h}_t$.

We implement this strategy of adaptive time step size on equation (4.3) with $T = 20$. As we pick $\epsilon_0 = 0.01$, (4.3) possesses weak mobility-diffusion and strong reaction. We solve the equation with $N_x = 256$, and set the initial time step size $h_t = 0.01$, we set $\bar{h}_t = 0.08$. As shown in Figure 15, our method works efficiently in this example, with an average $h_t \approx 0.04$. We also compute the same equation by using the classical IMEX method [25] in which we treat the linear part as implicit and the nonlinear part as explicit. We apply the preconditioned conjugate gradient (PCG) algorithm with tolerance[3] $\eta = 10^{-10}$ to solve the linear system at each IMEX step. For (4.3), the

[3] Suppose we apply PCG algorithm to solve the linear equation $Ax = b$ with A positive definite. Denote x_k as the solution obtained at the k-th iteration of the PCG algorithm, then we terminate the PCG iteration if $\|Ax_k - b\|_\infty \leq \eta$.

Our method	IMEX		
	$h_t = 0.5 \cdot 10^{-3}$	$h_t = 0.2 \cdot 10^{-3}$	$h_t = 10^{-4}$
1481.76 s	1814.40 s	4158.18 s	6216.81 s

Table 5. Comparison of CPU time (s) between our treatment and the classical IMEX method on computing the equation (4.3) on $[0, 20]$.

IMEX method only works stably for a rather small time step size $h_t \leq 0.5 \cdot 10^{-3}$. As reflected in Table 5, our method works better on long-time computation.

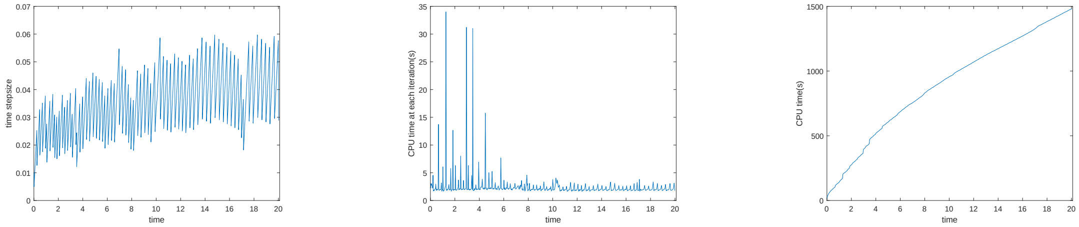


Figure 15. (Left) Plot of time step size h_t versus physical time t; (Middle) Plot of PDHG iterations versus physical time t; (Right) Plot of accumulated CPU time (s) versus physical time t.

4.5. **Comparison on computational efficiency.** In this section, we compare the computational efficiency (in CPU time) of the proposed method with some classical algorithms used for solving time-implicit schemes of the reaction-diffusion equations.

(1) (**Nonlinear SOR**) The Nonlinear SOR (NL SOR) method is the nonlinear version of the successive over-relaxation (SOR) algorithm. It is used to solve the implicit scheme of the Allen-Cahn equation (4.1) in [38]. We set the tolerance of the Newton's method used in NL SOR as 10^{-10}. We set $\tau_U = 0.55, \tau_P = 0.95$ for our PDHG method. We compare NL SOR with our algorithm in Figure 16.

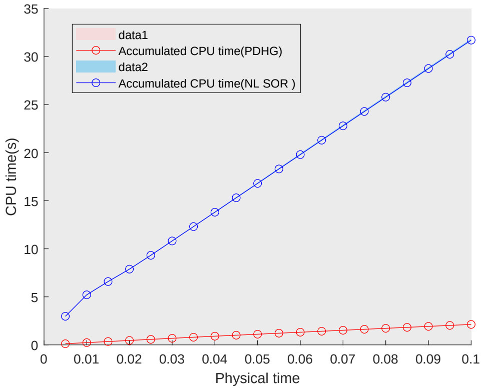


Figure 16. Accumulated CPU time comparison between our method (red) and Nonlinear SOR method (blue) applied to Allen-Cahn equation (4.1) with $\epsilon_0 = 0.1$ and $h_t = 0.005$. We solve the equation on a 128×128 grid. The quantile plots are composed based on 40 independent runs of both algorithms.

(2) (**Fixed point method**) The fixed point method is also a frequently used algorithm to solve the time-implicit scheme of the RD equation. We reformulate the time-implicit scheme (2.1) as

$$(I + ah_t\mathcal{G}_h\mathcal{L}_h)U^{t+1} = U^t - bh_t\mathcal{G}_h f(U^{t+1}).$$

For fixed U^t, we establish the following fixed point iteration for solving U^{t+1},

$$U_{k+1} = (I + h_t\mathcal{G}_h(a\mathcal{L}_h + bcI))^{-1}(U^t - bh_t\mathcal{G}_h(f(U_k) - cU_k)), \quad \text{with initial guess } U_0 = U^t.$$

Here c is a tunable constant that can be chosen as the value of $f'(\cdot)$ at equilibrium state. When $f(u) = W(u) = \frac{1}{4}(1-u^2)^2$, we set $c = f'(\pm 1) = 2$. The linear system is solved by the PCG algorithm with tolerance $\eta = 10^{-10}$. We set $\tau_U = 0.5, \tau_P = 0.95$ for our PDHG method. We apply both algorithms to (4.3) with $\epsilon_0 = 0.1$. We compare the fixed point method with our algorithm in Figure 17.

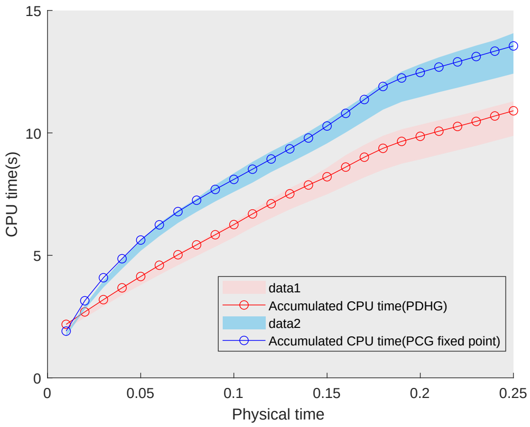


FIGURE 17. Accumulated CPU time comparison between our method (red) and PCG-fixed point iteration (blue). We solve (4.3) with $\epsilon_0 = 0.1$ and $h_t = 0.01$ on a 256×256 grid. These quantile plots are composed based on 40 independent runs of both algorithms.

(3) (**Newton's method**) Newton's method with the PCG algorithm as its linear solver serves as a popular tool for solving implicit schemes of RD equations with a higher order of spatial differentiation. Here we consider Newton's method introduced in section 3 of [11]. In [11], Newton's method is applied to the spectral discretization of the solution while here we apply Newton's method to the finite difference scheme. We set $\tau_U = 0.5, \tau_P = 0.95$ for our PDHG method. We apply both methods to (4.6). According to our experiments, we observe that when the time step size $h_t \leq 0.005$, Newton's method works more efficiently than the PDHG algorithm. When $h_t > 0.005$, the PDHG method is faster. Such observation is reflected in Figure 18. Table 6 demonstrates that the PDHG method is more efficient than Newton's method when the latter is applied to multi-interval computation with smaller time step sizes.

Method	PDHG	PCG Newton's method			
$h_t \times n$	0.01×50	0.005×100	0.001×500	0.0005×1000	0.00025×2000
CPU time(s)	263.90	422.28	299.02	470.71	773.01

TABLE 6. Time costs of applying the PDHG method and Newton's method to (4.6) on 256×256 grid.

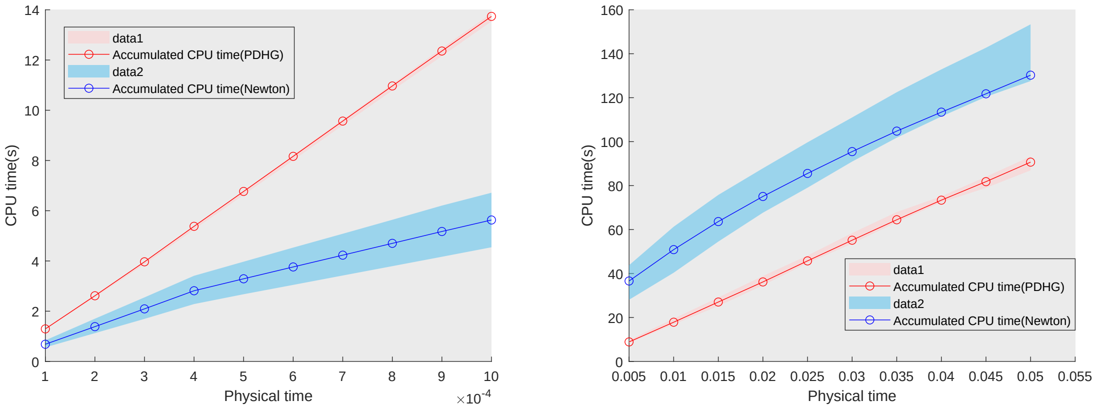


FIGURE 18. Accumulated CPU time comparison between our method (red) and Newton's method (blue). Solving equation (4.6) with $h_t = 0.001$ (Left) and $h_t = 0.005$ (Right). We solve the equation on a 256×256 grid. These quantile plots are composed based on 40 independent runs of both algorithms.

5. Conclusion

In this research, we reformulate the PDHG algorithm proposed in [35] by introducing a quadratic regularization term to solve implicit schemes of RD equations. Theoretically, we establish unique existence results for the time-implicit schemes of general RD equations. We further prove the exponential convergence for both the PDHG flow and the proposed discrete-time PDHG algorithm. In addition, we show that the convergence rates are independent of the grid size N_x. Our theoretical results are also supported by numerous numerical experiments. We test the proposed PDHG method via four different types of reaction-diffusion equations. Based on these numerical examples, we verify the optimal (or near-optimal) way to set the hyperparameters of our algorithm. We also verify the efficiency of our method by comparing it with several classical root-finding algorithms, such as the nonlinear SOR method, the fixed point method, and Newton's method.

We end the discussion by mentioning important future directions.

- The convergence rate achieved in this research is not the sharpest rate. Can we establish a sharp convergence rate in terms of the algorithm's hyperparameters?
- Currently, all of the proposed preconditioners are time-independent. How can we design a more sophisticated time-dependent preconditioner to assist the convergence of the generalized PDHG algorithm?
- As we accumulate multiple time intervals together to formulate a saddle-point scheme for the root-finding problem, we cancel the causalities among different time nodes. Will this causality-free optimization strategy render the possibility of parallel computing for the proposed PDHG time-implicit solvers?
- While the proposed algorithm performs efficiently on reaction-diffusion equations, we aim to extend our approach to simulate general equations in physical modeling, including Fokker-Planck equations and their generalizations in complex systems.
- Extend the proposed primal-dual approach to nonlinear, high-dimensional equations by integrating deep learning algorithms.

References

[1] Samuel M Allen and John W Cahn. A microscopic theory for antiphase boundary motion and its application to antiphase domain coarsening. *Acta metallurgica*, 27(6):1085–1095, 1979.

[2] Zhong-Zhi Bai, Yu-Mei Huang, and Michael K Ng. On preconditioned iterative methods for burgers equations. *SIAM Journal on Scientific Computing*, 29(1):415–439, 2007.

[3] John W Cahn. On spinodal decomposition. *Acta metallurgica*, 9(9):795–801, 1961.

[4] José A. Carrillo, Katy Craig, Li Wang, and Chaozhen Wei. Primal Dual Methods for Wasserstein Gradient Flows. *Foundations of Computational Mathematics*, 22(2):389–443, 2022.

[5] Jose A. Carrillo, Li Wang, and Chaozhen Wei. Structure preserving primal dual methods for gradient flows with nonlinear mobility transport distances, 2023.

[6] Hector D Ceniceros and Carlos J García-Cervera. A new approach for the numerical solution of diffusion equations with variable and degenerate mobility. *Journal of Computational Physics*, 246:1–10, 2013.

[7] Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of mathematical imaging and vision*, 40:120–145, 2011.

[8] Long Chen and Jingrong Wei. Accelerated gradient and skew-symmetric splitting methods for a class of monotone operator equations. *arXiv preprint arXiv:2303.09009*, 2023.

[9] Long Chen and Jingrong Wei. Transformed primal-dual methods for nonlinear saddle point systems. *Journal of Numerical Mathematics*, 2023.

[10] Yifan Chen, Bamdad Hosseini, Houman Owhadi, and Andrew M Stuart. Solving and learning nonlinear pdes with gaussian processes. *Journal of Computational Physics*, 447:110668, 2021.

[11] Andrew Christlieb, Jaylan Jones, Keith Promislow, Brian Wetton, and Mark Willoughby. High accuracy solutions to energy gradient flows from material science models. *Journal of computational physics*, 257:193–215, 2014.

[12] Jon M Church, Zhenlin Guo, Peter K Jimack, Anotida Madzvamuse, Keith Promislow, Brian Wetton, Steven M Wise, and Fengwei Yang. High accuracy benchmark problems for Allen-Cahn and Cahn-Hilliard dynamics. *Communications in computational physics*, 26(4), 2019.

[13] Christian Clason and Tuomo Valkonen. Primal-dual extragradient methods for nonlinear nonsmooth PDE-constrained optimization. *SIAM Journal on Optimization*, 27(3):1314–1339, 2017.

[14] James W Cooley and John W Tukey. An algorithm for the machine calculation of complex fourier series. *Mathematics of computation*, 19(90):297–301, 1965.

[15] David J Eyre. An unconditionally stable one-step scheme for gradient systems. *Unpublished article*, 6, 1998.

[16] Fariba Fahroo and Kazufumi Ito. Optimum damping design for an abstract wave equation. *Kybernetika*, 32(6):557–574, 1996.

[17] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

[18] Paul J Flory. Thermodynamics of high polymer solutions. *The Journal of chemical physics*, 10(1):51–61, 1942.

[19] Guosheng Fu, Stanley Osher, and Wuchen Li. High order spatial discretization for variational time implicit schemes: Wasserstein gradient flows and reaction-diffusion systems. *arXiv preprint arXiv:2303.08950*, 2023.

[20] Guosheng Fu, Stanley Osher, Will Pazner, and Wuchen Li. Generalized optimal transport and mean field control problems for reaction-diffusion systems with high-order finite element computation. *arXiv preprint arXiv:2306.06287*, 2023.

[21] Shuting Gu and Xiang Zhou. Convex splitting method for the calculation of transition states of energy functional. *Journal of Computational Physics*, 353:417–434, 2018.

[22] Jiequn Han, Arnulf Jentzen, and Weinan E. Solving high-dimensional partial differential equations using deep learning. *Proceedings of the National Academy of Sciences*, 115(34):8505–8510, 2018.

[23] Jiequn Han, Arnulf Jentzen, et al. Deep learning-based numerical methods for high-dimensional parabolic partial differential equations and backward stochastic differential equations. *Communications in mathematics and statistics*, 5(4):349–380, 2017.

[24] Dianming Hou, Lili Ju, and Zhonghua Qiao. A linear second-order maximum bound principle-preserving bdf scheme for the allen-cahn equation with a general mobility. *Mathematics of Computation*, 2023.

[25] Willem H Hundsdorfer, Jan G Verwer, and WH Hundsdorfer. *Numerical solution of time-dependent advection-diffusion-reaction equations*, volume 33. Springer, 2003.

[26] Matt Jacobs, Flavien Léger, Wuchen Li, and Stanley Osher. Solving large-scale optimization problems with a convergence rate independent of grid size. *SIAM Journal on Numerical Analysis*, 57(3):1100–1123, 2019.

[27] Andrea M Jokisaari, PW Voorhees, Jonathan E Guyer, James Warren, and OG Heinonen. Benchmark problems for numerical implementations of phase field models. *Computational Materials Science*, 126:139–151, 2017.

[28] JS Langer. Models of pattern formation in first-order phase transitions. In *Directions in condensed matter physics: Memorial volume in honor of shang-keng ma*, pages 165–186. World Scientific, 1986.

[29] Yibao Li, Hyun Geun Lee, Darae Jeong, and Junseok Kim. An unconditionally stable hybrid numerical method for solving the Allen-Cahn equation. *Computers & Mathematics with Applications*, 60(6):1591–1606, 2010.

[30] Tianyi Lin, Chi Jin, and Michael Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, pages 6083–6093. PMLR, 2020.

[31] Chun Liu, Cheng Wang, and Yiwei Wang. A structure-preserving, operator splitting scheme for reaction-diffusion equations with detailed balance. *Journal of Computational Physics*, 436:110253, 2021.

[32] Chun Liu, Cheng Wang, and Yiwei Wang. A second-order accurate, operator splitting scheme for reaction-diffusion systems in an energetic variational formulation. *SIAM Journal on Scientific Computing*, 44(4):A2276–A2301, 2022.

[33] Chun Liu, Cheng Wang, Yiwei Wang, and Steven M Wise. Convergence analysis of the variational operator splitting scheme for a reaction-diffusion system with detailed balance. *SIAM Journal on Numerical Analysis*, 60(2):781–803, 2022.

[34] Chun Liu and Yiwei Wang. On lagrangian schemes for porous medium type generalized diffusion equations: A discrete energetic variational approach. *Journal of Computational Physics*, 417:109566, 2020.

[35] Shu Liu, Siting Liu, Stanley Osher, and Wuchen Li. A first-order computational algorithm for reaction-diffusion type equations via primal-dual hybrid gradient method. *Journal of Computational Physics*, page 112753, 2024.

[36] Siting Liu, Stanley Osher, Wuchen Li, and Chi-Wang Shu. A primal-dual approach for solving conservation laws with implicit in time approximations. *Journal of Computational Physics*, 472, 2023.

[37] Tingwei Meng, Wenbo Hao, Siting Liu, Stanley J Osher, and Wuchen Li. Primal-dual hybrid gradient algorithms for computing time-implicit hamilton-jacobi equations. *arXiv preprint arXiv:2310.01605*, 2023.

[38] Barry Merriman, James K Bence, and Stanley J Osher. Motion of multiple junctions: A level set approach. *Journal of computational physics*, 112(2):334–363, 1994.

[39] James D Murray. *Mathematical biology II: Spatial models and biomedical applications*, volume 3. Springer New York, 2001.

[40] John E Pearson. Complex patterns in a simple system. *Science*, 261(5118):189–192, 1993.

[41] Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378:686–707, 2019.

[42] J Schnakenberg. Simple chemical reaction systems with limit cycle behaviour. *Journal of theoretical biology*, 81(3):389–400, 1979.

[43] Jie Shen, Jie Xu, and Jiang Yang. A new class of efficient and robust energy stable schemes for gradient flows. *SIAM Review*, 61(3):474–506, 2019.

[44] Jie Shen and Xiaofeng Yang. Numerical approximations of Allen-Cahn and Cahn-Hilliard Equations. *Discrete Contin. Dyn. Syst*, 28(4):1669–1691, 2010.

[45] Martin B. Short, P. Jeffrey Brantingham, Andrea L. Bertozzi, and George E. Tita. Dissipation and displacement of hotspots in reaction-diffusion models of crime. *Proceedings of the National Academy of Sciences*, 107(9):3961–3965, 2010.

[46] Gilbert Strang. A proposal for Toeplitz matrix calculations. *Studies in Applied Mathematics*, 74(2):171–176, 1986.

[47] Gilbert Strang. The discrete cosine transform. *SIAM review*, 41(1):135–147, 1999.

[48] Tuomo Valkonen. A primal-dual hybrid gradient method for nonlinear operators with applications to MRI. *Inverse Problems*, 30(5):055012, 2014.

[49] Colby L Wight and Jia Zhao. Solving allen-cahn and cahn-hilliard equations using the adaptive physics informed neural networks. *arXiv preprint arXiv:2007.04542*, 2020.

[50] Jinchao Xu, Yukun Li, Shuonan Wu, and Arthur Bousquet. On the stability and accuracy of partially and fully implicit schemes for phase field modeling. *Computer Methods in Applied Mechanics and Engineering*, 345:826–853, 2019.

[51] Jingjing Xu, Jia Zhao, and Yanxiang Zhao. Numerical approximations of the allen-cahn-ohta-kawasaki (acok) equation with modified physics informed neural networks (pinns). *arXiv preprint arXiv:2207.04582*, 2022.

[52] Xiaofeng Yang. Linear, first and second-order, unconditionally energy stable numerical schemes for the phase field model of homopolymer blends. *Journal of Computational Physics*, 327:294–316, 2016.

[53] Jianfeng Zhu, Yong-Tao Zhang, Stuart A Newman, and Mark Alber. Application of discontinuous Galerkin methods for reaction-diffusion systems in developmental biology. *Journal of Scientific Computing*, 40:391–418, 2009.

[54] Mingqiang Zhu and Tony Chan. An efficient primal-dual hybrid gradient algorithm for total variation image restoration. *UCLA CAM Report*, 34:8–34, 2008.

[55] Xinzhe Zuo, Zixiang Chen, Huaxiu Yao, Yuan Cao, and Quanquan Gu. Understanding train-validation split in meta-learning with neural networks. In *The Eleventh International Conference on Learning Representations*, 2022.

[56] Xinzhe Zuo, Stanley Osher, and Wuchen Li. Primal-dual damping algorithms for optimization. *arXiv preprint arXiv:2304.14574*, 2023.

Appendix A. Proofs of section 3.1

A.1. **Proof of Theorem 1.**

Proof of Theorem 1. To prove this result, we only need to prove that the following single-step scheme

$$\frac{U-U^0}{h_t} = -\mathcal{G}_h(a\mathcal{L}_h U + bf(U)), \tag{A.1}$$

admits a unique solution U for arbitrary U^0. By writing $\xi = U - U^0$, we reformulate (A.1) as

$$\frac{\xi}{h_t} + \mathcal{G}_h(a\mathcal{L}_h(U^0+\xi) + bf(U^0+\xi)) = 0. \tag{A.2}$$

We first show that ξ solves (A.2) iff ξ is the critical point of the following variational problem

$$\min_{\xi\in\mathrm{Ran}(\mathcal{G}_h)} \left\{ \frac{\xi^\top \mathcal{G}_h^\dagger \xi}{2h_t} + \frac{a}{2}(U^0+\xi)^\top \mathcal{L}_h(U^0+\xi) + bW(U^0+\xi)^\top \mathbf{1} \right\}. \tag{A.3}$$

Here we denote $W(\cdot)$ as the primitive function of $f(\cdot)$. Let us define $\mathcal{V} = \mathrm{Ran}(\mathcal{G}_h)$ and $\mathcal{J}(\xi)$ as the function in (A.3) for simplicity. Define $\Pi_{\mathcal{V}}$ as the orthogonal projection from $\mathbb{R}^{N_x\times N_x}$ onto the subspace $\mathcal{V}$.

We know that ξ is a critical point of $\mathcal{J}$ on space $\mathcal{V}$ iff

$$\Pi_{\mathcal{V}}\nabla\mathcal{J}(\xi) = 0.$$

By direct calculation, this is equivalent to

$$\frac{\mathcal{G}_h^\dagger \xi}{h_t} + a\,\Pi_{\mathcal{V}}\mathcal{L}_h(U_0+\xi) + b\,\Pi_{\mathcal{V}} f(U+\xi) = 0.$$

Writing the projection $\Pi_{\mathcal{V}} = \mathcal{G}_h^\dagger\mathcal{G}_h$, we obtain

$$\mathcal{G}_h^\dagger\left(\frac{\xi}{h_t} + a\,\mathcal{G}_h\mathcal{L}_h(U_0+\xi) + b\,\mathcal{G}_h f(U+\xi)\right) = 0.$$

Since the vector inside the above bracket belongs to $\mathcal{V}$, the above is equivalent to (A.2).

We now prove the existence and uniqueness of the minimizer to the variational problem (A.3) under condition (3.3), which implies the theorem.

By a change of variable $\xi = Q_1 x$, where Q_1 is defined as in the spectral decomposition (3.2) of $\mathcal{G}_h$, and $x\in\mathbb{R}^r$, (A.3) is equivalent to the following non-constrained optimization problem

$$\min_{x\in\mathbb{R}^r}\left\{ \frac{x^\top\Lambda^{-1}x}{2h_t} + \frac{a}{2}x^\top Q_1^\top\mathcal{L}_h Q_1 x + a\,U^{0\top}\mathcal{L}_h Q_1 x + b\,W(U^0+Q_1x)^\top\mathbf{1}\right\}. \tag{A.4}$$

Denote $\widetilde{\mathcal{J}}(x)$ as the function in the above problem. Computing $\nabla\widetilde{\mathcal{J}}$ yields

$$\nabla\widetilde{\mathcal{J}}(x) = \frac{\Lambda^{-1}}{h_t}x + a\,Q_1^\top\mathcal{L}_hQ_1x + aQ_1^\top\mathcal{L}_hU^0 + b\,Q_1^\top(V'(U^0+Q_1x) + \phi(U^0+Q_1x)).$$

Then

$$\begin{aligned}
&(x-y,\ \nabla\widetilde{\mathcal{J}}(x) - \nabla\widetilde{\mathcal{J}}(y))\\
=&\frac{1}{h_t}(x-y)^\top\Lambda(x-y) + a(x-y)^\top Q_1^\top\mathcal{L}_hQ_1(x-y) + b(Q_1(x-y))^\top(V'(U^0+Q_1x) - V'(U^0+Q_1y))\\
&+ b(Q_1(x-y))^\top(\phi(U^0+Q_1x) - \phi(U^0+Q_1y))\\
\geq&\ (x-y)^\top\left(\frac{\Lambda}{h_t} + aQ_1^\top\mathcal{L}_hQ_1\right)(x-y) + bK\|x-y\|^2 - b\mathrm{Lip}(\phi)\|x-y\|^2\\
\geq&\ \left(\lambda_{\min}\left(\frac{\Lambda^{-1}}{h_t} + aQ_1^\top\mathcal{L}_hQ_1\right) + bK - b\mathrm{Lip}(\phi)\right)\|x-y\|^2.
\end{aligned}$$

Then the condition (3.3) leads to

$$\alpha = \lambda_{\min}\left(\frac{\Lambda^{-1}}{h_t} + aQ_1^\top \mathcal{L}_h Q_1\right) + bK - b\mathrm{Lip}(\phi) > 0.$$

This shows the α-strongly convexity of $\widetilde{\mathcal{J}}$, which leads to the existence and uniqueness of the minimizer to (A.3), which accomplishes the proof. □

A.2. **Simplified conditions for specific reaction-diffusion equations.** The condition (3.3) can be simplified for specific types of equations. We discuss two examples.

- (Allen-Cahn equation with periodic boundary condition) In this case, $\mathcal{G} = \mathrm{Id}, \mathcal{L} = -\Delta$. $f(x) = x^3 - x$. We set $\mathcal{G}_h = I_{N_x^2}$, and $\mathcal{L}_h = -\Delta_{h_x}^P = I_{N_x} \otimes (-\mathrm{Lap}_{h_x}^P) + (-\mathrm{Lap}_{h_x}^P) \otimes I_{N_x}$, where $\mathrm{Lap}_{h_x}^P$ is defined in (2.2). Then

  $$\lambda_k^P = \frac{4}{h_x^2}\sin^2\left(\frac{\pi k}{N_x}\right), \quad \text{with } 1 \le k \le N_x,$$

  are the eigenvalues of $-\mathrm{Lap}_{h_x}^P$. And the eigenvalues of $\frac{\Lambda^{-1}}{h_t} + a\ Q_1^\top \mathcal{L}_h Q_1 = \frac{I}{h_t} + a\mathcal{L}_h$ are $\lambda_{k,l} = \frac{1}{h_t} + a(\lambda_k^P + \lambda_l^P)$, with $1 \le k, l \le N_x$. Thus, $\lambda_{\min}(\frac{\Lambda^{-1}}{h_t} + a\ Q_1^\top \mathcal{L}_h Q_1) = \frac{1}{h_t}$.

  Furthermore, we can decompose $f(x) = V'(x) + \phi(x)$, where

  $$V(x) = \begin{cases} \frac{1}{4}(x^2-1)^2, & |x| > 1; \\ 0, & |x| \le 1. \end{cases} \qquad \phi(x) = \begin{cases} 0, & |x| > 1; \\ x^3 - x, & |x| \le 1. \end{cases}$$

  Then one can verify that $K = 0$ and $\mathrm{Lip}(\phi) = 2$. In this case, condition (3.3) implies

  $$h_t < \frac{1}{\mathrm{Lip}(\phi) b} = \frac{1}{2b}.$$

- (Cahn-Hilliard equation with periodic boundary condition) In this case, $\mathcal{G} = -\Delta$, $\mathcal{L} = -\Delta$. $f(x) = x^3 - x$. We set $\mathcal{G}_h = \mathcal{L}_h = I \otimes (-\mathrm{Lap}_{h_x}^P) + (-\mathrm{Lap}_{h_x}^P) \otimes I$. We have

  $$\lambda_{\min}\left(\frac{\Lambda^{-1}}{h_t} + a\ \Lambda\right) = \min_{1 \le k,l \le N_x - 1}\left\{\frac{1}{(\lambda_k^P + \lambda_l^P)h_t} + a(\lambda_k^P + \lambda_l^P)\right\} \ge 2\sqrt{\frac{a}{h_t}}.$$

  Thus, a sufficient condition for (3.3) is

  $$h_t < \frac{4a^2}{b^2 \mathrm{Lip}(\phi)^2} = \frac{a^2}{b^2}.$$

It is worth mentioning that the conditions on h_t for both Allen-Cahn and Cahn-Hilliard equations are independent of the spatial step size h, which makes it possible for our scheme to overcome the CFL condition required in the time-explicit scheme.

Appendix B. Proofs of section 3.2

B.1. **Proofs of section 3.2.1.** To prove Lemma 2, we need the following Lemma 9 and Lemma 10.

Lemma 9. *Suppose $\overline{\lambda} \ge \underline{\lambda} > 0$. Assume $\mu > 0$ satisfies $\frac{1}{\sqrt{\underline{\lambda}}} - \frac{1}{\sqrt{\overline{\lambda}}} < \frac{2}{\sqrt{\mu}}$. Define*

$$A = \max\left\{\left(1 - \frac{\sqrt{\mu}}{\sqrt{\overline{\lambda}}}\right)^2, \left(1 - \frac{\sqrt{\mu}}{\sqrt{\underline{\lambda}}}\right)^2\right\}, \quad \text{and } B = \left(1 + \frac{\sqrt{\mu}}{\sqrt{\overline{\lambda}}}\right)^2,$$

then we always have $A < B$. Then for any $\lambda \in [\underline{\lambda},\ \overline{\lambda}]$, and $\gamma, \epsilon > 0$ with $A < \gamma\epsilon < B$, the matrix B_λ

$$B_\lambda = \begin{bmatrix} \gamma\lambda & -\frac{1}{2}(\mu - (1-\gamma\epsilon)\lambda) \\ -\frac{1}{2}(\mu - (1-\gamma\epsilon)\lambda) & \mu\epsilon \end{bmatrix} \tag{B.1}$$

is always positive definite.

Proof of Lemma 9. First, we have $\left|1 - \frac{\sqrt{\mu}}{\sqrt{\overline{\lambda}}}\right| < 1 + \frac{\sqrt{\mu}}{\sqrt{\overline{\lambda}}}, 1 - \frac{\sqrt{\mu}}{\sqrt{\underline{\lambda}}} < 1 + \frac{\sqrt{\mu}}{\sqrt{\underline{\lambda}}}$; and the condition $\frac{1}{\sqrt{\underline{\lambda}}} - \frac{1}{\sqrt{\overline{\lambda}}} < \frac{2}{\sqrt{\mu}}$ yields $-(1 - \frac{\sqrt{\mu}}{\sqrt{\underline{\lambda}}}) < 1 + \frac{\sqrt{\mu}}{\sqrt{\overline{\lambda}}}$. This yields

$$\max\left\{\left|1 - \frac{\sqrt{\mu}}{\sqrt{\overline{\lambda}}}\right|, \left|1 - \frac{\sqrt{\mu}}{\sqrt{\underline{\lambda}}}\right|\right\} < 1 + \frac{\sqrt{\mu}}{\sqrt{\overline{\lambda}}}.$$

Taking squares on both sides of the above inequality gives $A < B$.

On the other hand, since $\gamma\lambda > 0$, and $\mu\epsilon > 0$, we know B_λ is positive definite if and only if $\det(B_\lambda) > 0$. In order to alleviate our notations, let us denote the quadratic polynomial $q_{\mu,\lambda}(\cdot)$ as

$$q_{\mu,\lambda}(x) = \lambda^2 x^2 - 2\lambda(\mu + \lambda)x + (\mu - \lambda)^2.$$

Then we know $\det(B_\lambda) = -\frac{1}{4}q_{\mu,\lambda}(\gamma\epsilon)$.

Now, for fixed $\lambda \in [\underline{\lambda}, \overline{\lambda}]$, the two roots of $q_{\mu,\lambda}(x)$ are $\left(1 \pm \frac{\sqrt{\mu}}{\sqrt{\lambda}}\right)^2$. Thus $q_{\mu,\lambda}(x) < 0$ if

$$x \in I_\lambda \triangleq \left(\left(1 - \frac{\sqrt{\mu}}{\sqrt{\lambda}}\right)^2, \left(1 + \frac{\sqrt{\mu}}{\sqrt{\lambda}}\right)^2\right).$$

On the other hand, we have

$$\sup_{\lambda \in [\underline{\lambda}, \overline{\lambda}]} \left\{\left(1 - \frac{\sqrt{\mu}}{\sqrt{\lambda}}\right)^2\right\} = \max\left\{\left(1 - \frac{\sqrt{\mu}}{\sqrt{\overline{\lambda}}}\right)^2, \left(1 - \frac{\sqrt{\mu}}{\sqrt{\underline{\lambda}}}\right)^2\right\} = A,$$

and

$$\inf_{\lambda \in [\underline{\lambda}, \overline{\lambda}]} \left\{\left(1 + \frac{\sqrt{\mu}}{\sqrt{\lambda}}\right)\right\} = \left(1 + \frac{\sqrt{\mu}}{\sqrt{\overline{\lambda}}}\right)^2 = B.$$

As a result, $\bigcap_{\lambda \in [\underline{\lambda}, \overline{\lambda}]} I_\lambda = (A, B)$. Thus, we have shown that for any $\lambda \in [\underline{\lambda}, \overline{\lambda}]$, and $A < \gamma\epsilon < B$, $q_{\mu,\lambda}(\gamma\epsilon) < 0$. This directly leads to the assertion of the lemma. □

Lemma 10 (Positive definiteness of $\boldsymbol{H}_\mu$)**.** *Consider the matrix $\boldsymbol{H}_\mu$,*

$$\boldsymbol{H}_\mu = \begin{bmatrix} \gamma\Sigma & -\frac{1}{2}(\mu I - (1 - \gamma\epsilon)\Sigma) \\ -\frac{1}{2}(\mu I - (1 - \gamma\epsilon)\Sigma) & \mu\epsilon I \end{bmatrix},$$

with Σ symmetric and positive definite. Suppose $0 < \underline{\lambda} \leq \overline{\lambda}$ are two positive numbers such that the spectrum $\lambda(\Sigma) \subset [\underline{\lambda}, \overline{\lambda}]$. We further assume that $\frac{1}{\sqrt{\underline{\lambda}}} - \frac{1}{\sqrt{\overline{\lambda}}} < \frac{2}{\sqrt{\mu}}$. We adopt the notation A, B in Lemma 9, i.e.,

$$A = \max\left\{\left(1 - \frac{\sqrt{\mu}}{\sqrt{\overline{\lambda}}}\right)^2, \left(1 - \frac{\sqrt{\mu}}{\sqrt{\underline{\lambda}}}\right)^2\right\}, \quad \text{and } B = \left(1 + \frac{\sqrt{\mu}}{\sqrt{\overline{\lambda}}}\right)^2.$$

By Lemma 9, we have $A < B$. We also assume that $\gamma, \epsilon > 0$ satisfy $A < \gamma\epsilon < B$.

Define the function $\varphi_{\mu,\gamma,\epsilon}(\cdot)$ as

$$\varphi_{\mu,\gamma,\epsilon}(z) = \frac{1}{2}(\gamma z + \mu\epsilon - \sqrt{(\gamma z - \mu\epsilon)^2 + (\mu - (1 - \gamma\epsilon)z)^2}). \tag{B.2}$$

We denote $\beta = \min_{\lambda \in [\underline{\lambda},\ \overline{\lambda}]} \{\varphi_{\mu,\gamma,\epsilon}(\lambda)\}$, then $\beta > 0$. And we have $\boldsymbol{H}_\mu \succeq \beta I$.

Proof of Lemma 10. For any $\lambda \in [\underline{\lambda}, \overline{\lambda}]$, consider the matrix B_λ as defined in (B.1), i.e.,

$$B_\lambda = \begin{bmatrix} \gamma\lambda & -\frac{1}{2}(\mu - (1 - \gamma\epsilon)\lambda) \\ -\frac{1}{2}(\mu - (1 - \gamma\epsilon)\lambda) & \mu\epsilon \end{bmatrix}.$$

By Lemma 9, we know B_λ is positive definite. By a directly calculation, the eigenvalues of B_λ are given by (we assume $\lambda_1(B_\lambda) \geq \lambda_2(B_\lambda)$),

$$\lambda_{1,2}(B_\lambda) = \frac{\gamma\lambda + \mu\epsilon \pm \sqrt{(\gamma\lambda - \mu\epsilon)^2 + (\mu - (1-\gamma\epsilon)\lambda)^2}}{2}. \tag{B.3}$$

Thus $\lambda_2(B_\lambda) = \varphi_{\mu,\gamma,\epsilon}(\lambda)$. Since B_λ is positive definite, $\lambda_2(B_\lambda) = \varphi_{\mu,\gamma,\epsilon}(\lambda) > 0$.

As a result, $\varphi_{\mu,\gamma,\epsilon}(\lambda) > 0$ for $\lambda \in [\underline{\lambda}, \overline{\lambda}]$. Since $\varphi_{\mu,\gamma,\epsilon}(\cdot)$ is continuous on the compact set $[\underline{\lambda}, \overline{\lambda}]$, we know the infimum value $\beta > 0$. At the same time, it not hard to verify that $B_\lambda \succ \beta I$ for any $\lambda \in [\underline{\lambda}, \overline{\lambda}]$.

To estimate $\boldsymbol{H}_\mu$ from below, let us denote $\lambda(\Sigma) = \{\lambda_1, \lambda_2, \ldots, \lambda_N\}$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N > 0$ as the eigenvalues of matrix Σ. Since $\boldsymbol{H}_\mu$ is symmetric, $\boldsymbol{H}_\mu$ is similar to the following block diagonal matrix via an orthogonal transform

$$\begin{bmatrix} B_{\lambda_1} & & & \\ & B_{\lambda_2} & & \\ & & \ddots & \\ & & & B_{\lambda_N} \end{bmatrix},$$

with each B_{λ_j} defined as in (B.1). Since each $\lambda_j \in \lambda(\Sigma) \subset [\underline{\lambda}, \overline{\lambda}]$, the above argument applies to every B_{λ_j}, i.e., $B_{\lambda_j} \succ \beta I$ for any $1 \leq j \leq N$. This leads to $\boldsymbol{H}_\mu \succ \beta I$. □

We are ready to prove Lemma 2.

Proof of Lemma 2. We denote

$$\Sigma = D\widehat{F}(U_t) D\widehat{F}(U_t)^\top = (I + D\eta(U_t))(I + D\eta(U_t)^\top),$$

and compute

$$\begin{aligned} \frac{d}{dt}\mathcal{I}_\mu(U_t, Q_t) &= \widehat{F}(U)^\top D\widehat{F}(U)\dot{U} + \mu\, Q^\top \dot{Q} \\ &= -\widehat{F}(U)^\top D\widehat{F}(U) D\widehat{F}(U)^\top (Q + \gamma\dot{Q}) + \mu\, Q^\top(-\epsilon Q + \widehat{F}(U)) \\ &= -\widehat{F}(U)^\top \Sigma (Q + \gamma(-\epsilon Q + \widehat{F}(U))) - \mu\epsilon\|Q\|^2 + \mu Q^\top \widehat{F}(U) \\ &= \widehat{F}(U)^\top(\mu I - (1-\gamma\epsilon)\Sigma)Q - \gamma \widehat{F}(U)^\top \Sigma \widehat{F}(U) - \mu\epsilon\|Q\|^2 \\ &= -[\widehat{F}(U)^\top, Q^\top] \underbrace{\begin{bmatrix} \gamma\Sigma & -\frac{1}{2}(\mu I - (1-\gamma\epsilon)\Sigma) \\ -\frac{1}{2}(\mu I - (1-\gamma\epsilon)\Sigma) & \mu\epsilon I \end{bmatrix}}_{\text{denote as } \boldsymbol{H}_\mu} \begin{bmatrix} \widehat{F}(U) \\ Q \end{bmatrix} \\ &= -[\widehat{F}(U)^\top, Q^\top]\ \boldsymbol{H}_\mu\ [\widehat{F}(U)^\top, Q^\top]^\top. \end{aligned} \tag{B.4}$$

We denote $\sigma_1(U_t) \geq \cdots \geq \sigma_N(U_t)$ as the singular values of the Jacobian matrix $D\widehat{F}(U_t)$. It is not hard to verify that the spectrum of Σ

$$\lambda(\Sigma) = \{\sigma_1^2(U_t), \ldots, \sigma_N^2(U_t)\}.$$

According to definition (3.6) and (3.7), we have

$$\lambda(\Sigma) \subset [\underline{\sigma}^2,\ \overline{\sigma}^2].$$

Now we apply Lemma 10 with $\underline{\lambda} = \underline{\sigma}^2$, $\overline{\lambda} = \overline{\sigma}^2$. We prove that $\boldsymbol{H}_\mu \succ \beta I$ for any $U_t \in \mathbb{R}^N$. As a result, we obtain the following inequality:

$$\frac{d}{dt}\mathcal{I}_\mu(U_t, Q_t) = -[\widehat{F}(U)^\top, Q^\top]\ \boldsymbol{H}_\mu\ [\widehat{F}(U)^\top, Q^\top]^\top \leq -\beta(\|\widehat{F}(U_t)\|^2 + \|Q_t\|^2).$$

Furthermore, one has

$$\max\{1, \mu\}(\|\widehat{F}(U)\|^2 + \|Q\|^2) \geq \|\widehat{F}(U)\|^2 + \mu\|Q\|^2,$$

which yields

$$\|\widehat{F}(U)\|^2 + \|Q\|^2 \geq \frac{2}{\max\{1,\mu\}}\,\mathcal{I}_\mu(U,Q).$$

This finally leads to

$$\frac{d}{dt}\mathcal{I}_\mu(U_t,Q_t) \leq -\frac{2\,\beta}{\max\{1,\mu\}}\,\mathcal{I}_\mu(U_t,Q_t).$$

And the Grönwall's inequality gives

$$\mathcal{I}_\mu(U_t,Q_t) \leq \exp\left(-\frac{2\,\beta}{\max\{1,\mu\}}t\right)\mathcal{I}_\mu(U_0,Q_0).$$

□

We now prove Theorem 3.

Proof of Theorem 3. Let us pick the hyperparameter $\mu = \underline{\sigma}^2$, one can verify that μ satisfies (3.10). Furthermore, $\sqrt{\gamma\epsilon} = 1-\delta$. Since $|\delta| < \frac{1}{\kappa}$, $1-\frac{1}{\kappa} < \sqrt{\gamma\epsilon} < 1+\frac{1}{\kappa}$. This verifies that $\sqrt{\gamma\epsilon}$ satisfies (3.11). Now Theorem 2 guarantees that $\varphi_{\mu,\gamma,\epsilon} > 0$ on $[\underline{\sigma}^2, \overline{\sigma}^2]$. For $z \in [\underline{\sigma}^2, \overline{\sigma}^2]$, we further calculate

$$\begin{aligned}
\varphi_{\mu,\gamma,\epsilon}(z) =& \frac{1}{2}(\gamma z + \mu\epsilon - \sqrt{(\gamma z + \mu\epsilon)^2 - (4\gamma\epsilon\mu z - (\mu - (1-\gamma\epsilon)z)^2)})\\
=& \frac{1}{2}\frac{4\gamma\epsilon\mu z - (\mu-(1-\gamma\epsilon)z)^2}{\gamma z + \mu\epsilon + \sqrt{(\gamma z + \mu\epsilon)^2 - (4\gamma\epsilon\mu z - (\mu - (1-\gamma\epsilon)z)^2)}}\\
\geq& \frac{4\gamma\epsilon\mu z - (\mu-(1-\gamma\epsilon)z)^2}{4(\gamma z + \mu\epsilon)}\\
=& \frac{-(1-\gamma\epsilon)^2 z^2 + 2\mu(1+\gamma\epsilon)z - \mu^2}{4(\gamma z + \mu\epsilon)}\\
=& \frac{-((1+\gamma\epsilon)z - \mu)^2 + 4\gamma\epsilon z^2}{4(\gamma z + \mu\epsilon)}\\
=& \frac{(2\sqrt{\gamma\epsilon}z - (1+\gamma\epsilon)z + \mu)(2\sqrt{\gamma\epsilon}z + (1+\gamma\epsilon)z - \mu)}{4(\gamma z + \mu\epsilon)}\\
=& \frac{(\sqrt{\mu} - |1-\sqrt{\gamma\epsilon}|\sqrt{z})(\sqrt{\mu} + |1-\sqrt{\gamma\epsilon}|\sqrt{z})((1+\sqrt{\gamma\epsilon})^2 z - \mu)}{4(\gamma z + \mu\epsilon)}\\
\overset{1-\sqrt{\gamma\epsilon}=\delta,\ z\leq\overline{\sigma}^2}{\geq}& \frac{(\sqrt{\mu} - |\delta|\sqrt{z})(\sqrt{\mu} + |\delta|\sqrt{z})((2-\delta)^2 z - \mu)}{4(\gamma\overline{\sigma}^2 + \mu\epsilon)}.
\end{aligned} \tag{B.5}$$

Since we have set

$$\gamma = \frac{1-\delta}{\kappa},\ \epsilon = (1-\delta)\kappa,\ \mu = \underline{\sigma}^2.$$

Substituting them into (B.5) yields

$$\begin{aligned}
\varphi_{\mu,\gamma,\epsilon}(z) &\geq \frac{(\underline{\sigma} - |\delta|\sqrt{z})(|\delta|\sqrt{z} + \underline{\sigma})((2-\delta)^2 z - \underline{\sigma}^2)}{8(1-\delta)\ \underline{\sigma}\ \overline{\sigma}}\\
&= \frac{1}{8(1-\delta)}\left(1 - |\delta|\frac{\sqrt{z}}{\underline{\sigma}}\right)\left(|\delta|\frac{\sqrt{z}}{\overline{\sigma}} + \frac{\underline{\sigma}}{\overline{\sigma}}\right)((2-\delta)^2 z - \underline{\sigma}^2)\\
&\geq \frac{1}{8(1-\delta)}(1-\kappa|\delta|)\left(\frac{|\delta|+1}{\kappa}\right)(1-\delta)(3-\delta)\underline{\sigma}^2\\
&\geq \frac{1}{8\kappa}(1-\kappa|\delta|)(3-\delta)\underline{\sigma}^2.
\end{aligned}$$

If we denote $\beta = \min_{z\in[\underline{\sigma}^2,\overline{\sigma}^2]}\{\varphi_{\mu,\gamma,\epsilon}(z)\}$, then we have

$$\frac{\beta}{\max\{1,\mu\}} \geq \frac{(1-\kappa|\delta|)(3-\delta)}{8\kappa}\frac{\underline{\sigma}^2}{\max\{1,\underline{\sigma}^2\}} = \frac{1}{8}(1-\kappa|\delta|)(3-\delta)\frac{\min\{\underline{\sigma}^2,1\}}{\kappa}.$$

Thus, the result of Theorem 2 yields

$$\mathcal{I}_\mu(U_t,Q_t) \leq \exp\left(-\frac{1}{4}(1-\kappa|\delta|)(3-\delta)\frac{\min\{\underline{\sigma}^2,1\}}{\kappa}\,t\right)\mathcal{I}_\mu(U_0,Q_0).$$

Taking square root on both sides of the above inequality and using the fact that

$$\|\widehat{F}(U_t)\| \leq \sqrt{\mathcal{I}_\mu(U_t,Q_t)},$$

we obtain

$$\|\widehat{F}(U_t)\| \leq \exp\left(-\frac{1}{8}(1-\kappa|\delta|)(3-\delta)\frac{\min\{\underline{\sigma}^2,1\}}{\kappa}\,t\right)\sqrt{\mathcal{I}_\mu(U_0,Q_0)}.$$

This implies our theorem. □

Theorem 11 (Exponential decay of $\mathcal{I}_\mu(U_t,Q_t)$)**.** *Assume that (U_t,Q_t) solves* (3.5) *with arbitrary initial position (U_0,Q_0). Then we have the exponential decay of the Lyapunov function $\mathcal{I}_\mu(U_t,Q_t)$, i.e.,*

$$\mathcal{I}_\mu(U_t,Q_t) \leq \exp\left(-2\lambda t\right)\ \mathcal{I}_\mu(U_0,Q_0)\,,$$

where

$$\lambda = \min\{\epsilon - \frac{1}{2}|(1-\gamma\epsilon)\sigma_1^2/\mu - 1|, \epsilon - \frac{1}{2}|(1-\gamma\epsilon)\sigma_n^2/\mu - 1|,$$
$$\gamma\sigma_1^2 - \frac{1}{2}|(1-\gamma\epsilon)\sigma_1^2 - \mu|\}, \gamma\sigma_n^2 - \frac{1}{2}|(1-\gamma\epsilon)\sigma_n^2 - \mu|\}\,.$$

In particular, when $\gamma\epsilon = 1$, $\mu = 0$, and

$$\gamma = \frac{-\frac{1}{2}\frac{\sigma_1^2-\sigma_n^2}{\sigma_1^2+\sigma_n^2} + \sqrt{\frac{1}{4}\left(\frac{\sigma_1^2-\sigma_n^2}{\sigma_1^2+\sigma_n^2}\right)^2 + 4\sigma_n^2}}{2\sigma_n^2},$$

we have $\lambda = 2\sigma_n^2\frac{\sigma_1^2+\sigma_n^2}{\sigma_1^2-\sigma_n^2} - \frac{1}{2}\sigma_n^4\left(\frac{\sigma_1^2+\sigma_n^2}{\sigma_1^2-\sigma_n^2}\right)^3 + \mathcal{O}(\sigma_n^6)$.

Proof of Theorem 11. We would like to find λ such that

$$\frac{d\mathcal{I}}{dt} + 2\lambda\mathcal{I} \leq 0\,.$$

Then by Gronwall's inequality, we obtain exponential convergence. We have

$$\frac{d\mathcal{I}}{dt} + 2\lambda\mathcal{I} = [\widehat{F}(U)^\top, Q^\top]\left[\begin{array}{cc} \lambda I - \gamma\Sigma & \frac{1}{2}(\mu I - (1-\gamma\epsilon)\Sigma) \\ \frac{1}{2}(\mu I - (1-\gamma\epsilon)\Sigma) & \lambda\mu I - \mu\epsilon I \end{array}\right]\left[\begin{array}{c} \widehat{F}(U) \\ Q \end{array}\right]\,.$$

Using Lemma A.1 from [56], it suffices to have

$$\lambda - \gamma\sigma_i^2 + \frac{1}{2}|(1-\gamma\epsilon)\sigma_i^2 - \mu| \leq 0\,, \tag{B.6a}$$

$$\lambda\mu - \mu\epsilon + \frac{1}{2}|(1-\gamma\epsilon)\sigma_i^2 - \mu| \leq 0\,, \tag{B.6b}$$

for all $\overline{\sigma}^2 = \sigma_1^2 \geq \sigma_2^2 \geq \cdots \geq \sigma_n^2 = \underline{\sigma}^2$. Let us define $g_1(\sigma) = \epsilon - \frac{1}{2}|(1-\gamma\epsilon)\sigma^2/\mu - 1|$, and $g_2(\sigma) = \gamma\sigma^2 - \frac{1}{2}|(1-\gamma\epsilon)\sigma^2 - \mu|$. Then (B.6) implies that

$$\lambda \leq \min_{i=1,2}\min_{\sigma_n\leq\sigma\leq\sigma_1} g_i(\sigma)\,.$$

Since $g_i(\sigma)$'s are piece-wise linear and have only one kink, it is easy to check that

$$\min_{\sigma_n \leq \sigma \leq \sigma_1} g_i(\sigma) = \min\{g_i(\sigma_1), g_i(\sigma_n)\}\,.$$

This proves the first part of our lemma. When taking $\mu = \frac{1}{2}(1-\gamma\epsilon)(\sigma_1^2+\sigma_n^2)$, one can show by a straightforward calculation that $g_1(\sigma_n) = g_1(\sigma_1)$. This also implies that $g_2(\sigma_1) \geq g_2(\sigma_n)$. Therefore, to make λ large, we would like to equate $g_1(\sigma_n)$ and $g_2(\sigma_n)$. This yields

$$\begin{aligned}\epsilon - \frac{1}{2}\frac{\sigma_1^2-\sigma_n^2}{\sigma_1^2+\sigma_n^2} &= \gamma\sigma_n^2 - \frac{1}{4}(1-\gamma\epsilon)(\sigma_1^2-\sigma_n^2)\\ \epsilon &= \frac{\gamma\sigma_n^2 + \frac{1}{2}\frac{\sigma_1^2-\sigma_n^2}{\sigma_1^2+\sigma_n^2} - \frac{1}{4}(\sigma_1^2-\sigma_n^2)}{1-\frac{1}{4}\gamma(\sigma_1^2-\sigma_n^2)}\,. \qquad \text{(B.7)}\end{aligned}$$

In the special case of $\gamma\epsilon = 1$, we obtain

$$1 = \gamma\epsilon = \frac{\gamma^2\sigma_n^2 + \frac{1}{2}\gamma\frac{\sigma_1^2-\sigma_n^2}{\sigma_1^2+\sigma_n^2} - \frac{1}{4}\gamma(\sigma_1^2-\sigma_n^2)}{1-\frac{1}{4}\gamma(\sigma_1^2-\sigma_n^2)}\,. \qquad \text{(B.8)}$$

We can solve for γ and we get (keeping the positive root)

$$\gamma = \frac{-\frac{1}{2}\frac{\sigma_1^2-\sigma_n^2}{\sigma_1^2+\sigma_n^2} + \sqrt{\frac{1}{4}\left(\frac{\sigma_1^2-\sigma_n^2}{\sigma_1^2+\sigma_n^2}\right)^2 + 4\sigma_n^2}}{2\sigma_n^2}\,.$$

Consequently, the convergence rate is

$$\begin{aligned}\lambda = \gamma\sigma_n^2 &= -\frac{1}{4}\frac{\sigma_1^2-\sigma_n^2}{\sigma_1^2+\sigma_n^2} + \frac{1}{2}\sqrt{\frac{1}{4}\left(\frac{\sigma_1^2-\sigma_n^2}{\sigma_1^2+\sigma_n^2}\right)^2 + 4\sigma_n^2}\\ &= 2\sigma_n^2\frac{\sigma_1^2+\sigma_n^2}{\sigma_1^2-\sigma_n^2} - \frac{1}{2}\sigma_n^4\left(\frac{\sigma_1^2+\sigma_n^2}{\sigma_1^2-\sigma_n^2}\right)^3 + \mathcal{O}(\sigma_n^6)\,. \qquad \text{(B.9)}\end{aligned}$$

□

B.2. **Proofs of section 3.2.2.** To prove Lemma 4, we need the following Lemma 12 and Lemma 13.

Lemma 12. *Suppose A is an $nm \times nm$ matrix defined as*

$$A = \begin{bmatrix} A_1 & & & & \\ -I & A_2 & & & \\ & -I & A_3 & & \\ & & \ddots & \ddots & \\ & & & -I & A_n \end{bmatrix},$$

where each A_k is anm $\times$ m matrix with $\sigma_{\min}(A_k) \geq \underline{\alpha} > 0$ and $\sigma_{\max}(A_k) \leq \overline{\alpha}$, i.e., $\|A_k v\| \geq \underline{\alpha}\|v\|$, $\|A_k v\| \leq \overline{\alpha}\|v\|$ for any $v \in \mathbb{R}^m$. Then $\|A^{-1}\| \leq \sum_{k=1}^n \underline{\alpha}^{-k}$, and $\|A\| \leq \overline{\alpha}+1$, i.e., $\sigma_{\min}(A) \geq \frac{1}{\sum_{k=1}^n \underline{\alpha}^{-k}}$, and $\sigma_{\max}(A) \leq \overline{\alpha}+1$.

Proof of Lemma 12. By a direct calculation, we have

$$A^{-1} = \begin{bmatrix} A_1^{-1} & & & & \\ (A_1A_2)^{-1} & A_2^{-1} & & & \\ (A_1A_2A_3)^{-1} & (A_2A_3)^{-1} & A_3^{-1} & & \\ \vdots & \vdots & \vdots & \ddots & \\ (A_1A_2\ldots A_n)^{-1} & (A_2\ldots A_n)^{-1} & (A_3\ldots A_n)^{-1} & \ldots & A_n^{-1} \end{bmatrix}.$$

Thus we can write A^{-1} as

$$A^{-1} = \begin{bmatrix} A_{11} & & & O \\ & A_{22} & & \\ & & \ddots & \\ O & & & A_{nn} \end{bmatrix} + \begin{bmatrix} O & & & \\ A_{21} & \ddots & & \\ & \ddots & \ddots & \\ O & & A_{n,n-1} & O \end{bmatrix} + \cdots + \begin{bmatrix} O & & & \\ \vdots & \ddots & & \\ O & & \ddots & \\ A_{n1} & O & \dots & O \end{bmatrix}$$

$$\overset{\text{denote as}}{=} J_1 + J_2 + \cdots + J_n.$$

Here, each J_k $(1 \le k \le n)$ is an $nm \times nm$ block-(sub)diagonal matrix whose k-th subdiagonal is

$$\mathrm{diag}(A_{k,1}, A_{k+1,2}, \dots, A_{n,n-k+1}).$$

And each A_{ij} is defined as

$$A_{ij} = (A_j A_{j+1} \dots A_i)^{-1}, \quad \text{if } i \ge j.$$

Then one can bound $\|A^{-1}\|$ as

$$\|A^{-1}\| \le \sum_{k=1}^{n} \|J_k\|.$$

To bound each $\|J_k\|$ from above, consider any $v = [v_1^\top, v_2^\top, \dots, v_n^\top]^\top \in \mathbb{R}^{nm}$ with each $v_j \in \mathbb{R}^m$, we have

$$\|J_k v\|^2 = \sum_{j=k}^{n} \|A_{j,j-k+1} v_j\|^2 = \sum_{j=k}^{n} \|(A_{j-k+1} \dots A_j)^{-1} v_j\|^2 \le \underline{\alpha}^{-2k} \sum_{j=k}^{n} \|v_j\|^2 \le \underline{\alpha}^{-2k} \|v\|^2.$$

This yields $\|J_k v\| \le \underline{\alpha}^{-k}\|v\|$ which further gives $\|J_k\| \le \underline{\alpha}^{-k}$. Thus, we have proved $\|A^{-1}\| \le \sum_{k=1}^{n} \underline{\alpha}^{-k}$, which directly leads to the result $\sigma_{min}(A) \ge \frac{1}{\sum_{k=1}^{n} \underline{\alpha}^{-k}}$.

On the other hand, we write A as

$$A = \mathrm{diag}(A_1, \dots, A_n) - J \otimes I,$$

where J is an $n \times n$ matrix defined as

$$J = \begin{bmatrix} 0 & & & \\ 1 & \ddots & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{bmatrix}, \tag{B.10}$$

and I is an $n \times n$ identity matrix. Then we have

$$\|A\| \le \|\mathrm{diag}(A_1, \dots, A_n)\| + \|J \otimes I\| \le \overline{\alpha} + 1.$$

□

Lemma 13. *Suppose G, L are self-adjoint, nonnegative definite matrices. Assume $GL = LG$. Then $I + GL$ (or $I + LG$) is orthogonally equivalent to $I + \Lambda_G \Lambda_L$, where Λ_G, Λ_L are the diagonal matrices equivalent to G, L. Furthermore, $\sigma_{\min}(I + GL) = \sigma_{\min}(I + LG) \ge 1 + \lambda_{\min}(G)\lambda_{\min}(L) \ge 1$.*

Proof of Lemma 13. Since G, L commutes, they can be diagonalized simultaneously, i.e., there exists an orthogonal matrix Q, s.t. $G = Q\Lambda_G Q^\top$, and $L = Q\Lambda_L Q^\top$, where $\Lambda_G, \Lambda_L \succeq O$ are diagonal matrices. Then $I + GL = I + LG = Q(I + \Lambda_G \Lambda_L)Q^\top$. And thus $\sigma_{\min}(I + GL) = \sigma_{\min}(I + \Lambda_G \Lambda_L) \ge 1 + \lambda_{\min}(G)\lambda_{\min}(L) \ge 1$. □

We now prove Lemma 4.

Proof of Lemma 4. We first recall

$$\begin{aligned}\underline{\sigma} &= \inf_{U\in\mathbb{R}^{N_x^2}}\{\sigma_{\min}(D\widehat{F}(U))\} = \inf_{U\in\mathbb{R}^{N_x^2}}\{\sigma_{\min}(\mathscr{M}^{-1}DF(U))\},\\ \overline{\sigma} &= \sup_{U\in\mathbb{R}^{N_x^2}}\{\sigma_{\max}(D\widehat{F}(U))\} = \sup_{U\in\mathbb{R}^{N_x^2}}\{\sigma_{\max}(\mathscr{M}^{-1}DF(U))\},\end{aligned}$$

where we denote $F(U) = \mathscr{D}U + h_t\mathscr{G}_h(a\mathscr{L}_hU + bf(U))$.

We have

$$\sigma_{\min}(\mathscr{M}^{-1}DF(U)) = \frac{1}{\sigma_{\max}(DF(U)^{-1}\mathscr{M})} \geq \frac{1}{\|DF(U)^{-1}\|\|\mathscr{M}\|_2} = \frac{\sigma_{\min}(DF(U))}{\|\mathscr{M}\|}. \tag{B.11}$$

And

$$\sigma_{\max}(\mathscr{M}^{-1}DF(U)) \leq \sigma_{\max}(DF(U))\|\mathscr{M}^{-1}\|. \tag{B.12}$$

Now we estimate the singular values of $DF(U)$, since

$$DF(U) = \begin{bmatrix} X_1 & & & & \\ -I & X_2 & & & \\ & -I & X_3 & & \\ & & \ddots & \ddots & \\ & & & -I & X_{N_t}\end{bmatrix},$$

where each $X_i = I + ah_t\mathcal{G}_h\mathcal{L}_h + bh_t\mathcal{G}_h\mathrm{diag}(f'(U^i))$. (Here we denote $U = (U^{1\top},\dots,U^{N_t\top})^\top$.)

Then for each X_i, we have

$$\begin{aligned}\sigma_{\min}(X_i) &\geq \sigma_{\min}(I + ah_t\mathcal{G}_h\mathcal{L}_h) - \sigma_{\max}(bh_t\mathcal{G}_h\mathrm{diag}(f'(U^i)))\\ &\geq \sigma_{\min}(I + ah_t\mathcal{G}_h\mathcal{L}_h) - h_t|b|\|\mathcal{G}_h\|\|\mathrm{diag}(f'(U^i))\|.\end{aligned}$$

By Lemma 13, the first term above is no less than $1 + ah_t\lambda_{\min}(\mathcal{G}_h)\lambda_{\min}(\mathcal{L}_h) \geq 1$. It is not hard to verify that $\|\mathcal{G}_h\| = \lambda_{\max}(\mathcal{G}_h)$, $\|\mathrm{diag}(f'(U^i))\| \leq \mathrm{Lip}(f)$. This leads to

$$\sigma_{\min}(X_i) \geq 1 - h_t|b|\lambda_{\max}(\mathcal{G}_h)\mathrm{Lip}(f).$$

We denote $\underline{\alpha} = 1 - h_t|b|\lambda_{\max}(\mathcal{G}_h)\mathrm{Lip}(f)$. Then $\underline{\alpha} > 0$, and is independent of U.

On the other hand, one can also verify that

$$\sigma_{\max}(X_i) = \|X_i\| \leq \|I + ah_t\mathcal{G}_h\mathcal{L}_h\| + h_t|b|\|\mathcal{G}_h\|\mathrm{Lip}(f),$$

by denoting $\overline{\alpha} = \|I + ah_t\mathcal{G}_h\mathcal{L}_h\| + h_t|b|\|\mathcal{G}_h\|\mathrm{Lip}(f)$, we know $\overline{\alpha}$ is also independent of U.

We now apply Lemma 12 to $DF(U)$ with $\sigma_{\min}(X_i) \geq \underline{\alpha}$ and $\sigma_{\max}(X_i) \leq \overline{\alpha}$. Together with (B.11) and (B.12), we have

$$\sigma_{\min}(D\widehat{F}(U)) \geq \frac{1}{\left(\sum_{k=1}^{N_t}\underline{\alpha}^{-k}\right)\|\mathscr{M}\|}, \quad \sigma_{\max}(D\widehat{F}(U)) \leq (1+\overline{\alpha})\|\mathscr{M}^{-1}\|.$$

Since $\underline{\alpha}$, $\overline{\alpha}$, $\|\mathscr{M}\|$ and $\|\mathscr{M}^{-1}\|$ are all independent of U, we are done. □

To prove Lemma 6, we need the following Lemma 14.

Lemma 14. *Suppose we keep all the assumptions from Lemma 6. Let $\mathscr{G}_h$ be defined as in* (2.6), *and $\mathscr{M}$ be defined as in* (2.13). *Then*

$$\|\mathscr{M}^{-1}\mathscr{G}_h\| \leq N_t\left(\max_{1\leq k\leq N_x^2}\left\{\frac{\lambda_k(\mathcal{G}_h)}{1 + h_t(a\lambda_k(\mathcal{G}_h)\lambda_k(\mathcal{L}_h) + bc\lambda_k(\mathcal{G}_h))}\right\}\right).$$

Proof of Lemma 14. Recall that we have

$$\mathscr{M} = \begin{bmatrix} X & & & & \\ -I & X & & & \\ & -I & X & & \\ & & \ddots & \ddots & \\ & & & -I & X \end{bmatrix}, \quad X = I + ah_t\mathcal{G}_h\mathcal{L}_h + bh_t\mathcal{G}_hJ_f.$$

By Lemma 13, we have $X = Q(I + ah_t\Lambda_{\mathcal{G}_h}\Lambda_{\mathcal{L}_h} + bch_t\Lambda_{\mathcal{G}_h})Q^\top$, where we have also used that $\mathcal{G}_h, \mathcal{L}_h$ commute, and $J_f = cI$. Here we write $\Lambda_{\mathcal{G}_h}, \Lambda_{\mathcal{L}_h}$ as the diagonal matrices which are orthogonally similar to $\mathcal{G}_h, \mathcal{L}_h$ w.r.t. orthogonal matrix Q. It is not hard to verify that

$$\|X^{-1}\| \leq \frac{1}{1 + h_t(\lambda_{\min}(a\mathcal{G}_h\mathcal{L}_h + bc\mathcal{G}_h))} \leq 1. \tag{B.13}$$

Now one can compute

$$\begin{aligned}\mathscr{M}^{-1}\mathscr{G}_h &= \begin{bmatrix} X^{-1} & & & & \\ X^{-2} & X^{-1} & & & \\ X^{-3} & X^{-2} & X^{-1} & & \\ \vdots & \vdots & \vdots & \ddots & \\ X^{-N_t} & X^{-(N_t-1)} & X^{-(N_t-2)} & \dots & X^{-1} \end{bmatrix} \begin{bmatrix} \mathcal{G}_h & & & & \\ & \mathcal{G}_h & & & \\ & & \mathcal{G}_h & & \\ & & & \ddots & \\ & & & & \mathcal{G}_h \end{bmatrix} \\ &= \underbrace{\begin{bmatrix} I & & & & \\ X^{-1} & I & & & \\ X^{-2} & X^{-1} & I & & \\ \vdots & \vdots & \vdots & \ddots & \\ X^{-(N_t-1)} & X^{-(N_t-2)} & X^{-(N_t-3)} & \dots & I \end{bmatrix}}_{\mathscr{N}} \underbrace{\begin{bmatrix} X^{-1}\mathcal{G}_h & & & & \\ & X^{-1}\mathcal{G}_h & & & \\ & & X^{-1}\mathcal{G}_h & & \\ & & & \ddots & \\ & & & & X^{-1}\mathcal{G}_h \end{bmatrix}}_{\widetilde{\mathscr{G}_h}} \\ &\overset{\text{denote as}}{=} \mathscr{N}\widetilde{\mathscr{G}_h}.\end{aligned}$$

Similar to the treatment in Lemma 4, we estimate $\|\mathscr{N}\|$ by decomposing $\mathscr{N}$ as

$$\mathscr{N} = I \otimes X^0 + J \otimes X^{-1} + J^2 \otimes X^{-2} + \cdots + J^{N_t-1} \otimes X^{-(N_t-1)},$$

where we recall that J is defined as in (B.10); And X^0 is treated as the identity matrix.

Then we estimate $\|\mathscr{N}\|$ as

$$\|\mathscr{N}\| \leq \left(\sum_{k=0}^{N_t-1} \|J^k \otimes (X^{-1})^k\|\right).$$

Since $\|A \otimes B\| = \|A\| \cdot \|B\|$ for any dimensions of square matrices A, B, using (B.13) and $\|J\| \leq 1$ yields

$$\|\mathscr{N}\| \leq \sum_{k=0}^{N_t-1} \|(X^{-1})^k\| \leq \sum_{k=0}^{N_t-1} \|X^{-1}\|^k \leq N_t. \tag{B.14}$$

On the other hand, we have

$$\widetilde{\mathscr{G}_h} = X^{-1}\mathcal{G}_h = Q((I + ah_t\Lambda_{\mathcal{G}_h}\Lambda_{\mathcal{L}_h} + bch_t\Lambda_{\mathcal{G}_h})^{-1}\Lambda_{\mathcal{G}_h})Q^\top.$$

If we denote $\{\lambda_k(\mathcal{G}_h)\}$, $\{\lambda_k(\mathcal{L}_h)\}$ $(1 \leq k \leq N_x^2)$ as the corresponding eigenvalues of $\mathcal{G}_h, \mathcal{L}_h$ w.r.t. Q, we know

$$\|\widetilde{\mathscr{G}_h}\| = \max_{1 \leq k \leq N_x^2} \left\{ \frac{\lambda_k(\mathcal{G}_h)}{1 + h_t(a\lambda_k(\mathcal{G}_h)\lambda_k(\mathcal{L}_h) + bc\lambda_k(\mathcal{G}_h))} \right\}. \tag{B.15}$$

Now combining (B.14) and (B.15) and using $\|\mathscr{M}^{-1}\mathscr{G}_h\| \leq \|\mathscr{N}\|\|\widetilde{\mathscr{G}_h}\|$, we finish the proof. □

We now prove Lemma 6.

Proof of Lemma 6. By Lemma 14 and the fact that $\|DR(\cdot)\| \leq \mathrm{Lip}(R)$, we have

$$\|D\eta(U)\| = \|bh_t\mathscr{M}^{-1}\mathscr{G}_h DR(U)\| \leq bh_t \cdot \|\mathscr{M}^{-1}\mathscr{G}_h\| \cdot \mathrm{Lip}(R) \leq bT\zeta_{a,b,c}(h_t)\mathrm{Lip}(R).$$

Recall that

$$D\widehat{F}(U) = I + D\eta(U).$$

Now for any $v \in \mathbb{R}^{N_x^2}$, we have

$$\|D\widehat{F}(U)v\| = \|v + D\eta(U)v\| \geq \|v\| - \|D\eta(U)\|\|v\|. \geq (1 - bT\mathrm{Lip}(R)\zeta_{a,b,c}(h_t))\|v\|. \tag{B.16}$$

Since the right-hand side of (B.16) is independent of U, this will lead to a lower bound on $\underline{\sigma}$, i.e.

$$\underline{\sigma} \geq 1 - bT\zeta_{a,b,c}(h_t)\mathrm{Lip}(R).$$

By a similar argument, we have

$$\|D\widehat{F}(U)v\| \leq \|v\| + \|D\eta(U)\|\|v\| \leq (1 + bT\zeta_{a,b,c}(h_t)\mathrm{Lip}(R))\|v\|.$$

This will finally lead to

$$\overline{\sigma} \leq 1 + bT\zeta_{a,b,c}(h_t)\mathrm{Lip}(R).$$

□

Lemma 15 (Sufficient condition on the unique solvability of $\widehat{F}(U) = 0$)**.** *Suppose conditions* (A)*,* (B)*,* (C) *and* (D) *hold. We pick h_t and $T = N_t h_t$ ($N_t \in \mathbb{N}_+$) satisfying $bT\mathrm{Lip}(R)\zeta_{a,b,c}(h_t) < 1$. Then there exists a unique root of $\widehat{F}$.*

Proof of Lemma 15. (3.12) leads to

$$\max_{1\leq k\leq N_x^2}\left\{\frac{\lambda_k(\mathcal{G}_h)}{1 + h_t(a\lambda_k(\mathcal{G}_h)\lambda_k(\mathcal{L}_h) + bc\lambda_k(\mathcal{G}_h))}\right\} < \frac{1}{bT\mathrm{Lip}(R)},$$

which is equivalent to

$$\min_{1\leq k\leq N_x^2, \lambda_k(\mathcal{G}_h)>0}\left\{\frac{1}{\lambda_k(\mathcal{G}_h)} + h_t(a\lambda_k(\mathcal{L}_h) + bc)\right\} > bT\mathrm{Lip}(R).$$

Since $T \geq h_t$, the right-hand side of the above inequality is larger than or equal to $bh_t\mathrm{Lip}(R)$. Thus the above inequality yields

$$\min_{1\leq k\leq N_x^2, \lambda_k(\mathcal{G}_h)>0}\left\{\frac{1}{\lambda_k(\mathcal{G}_h)h_t} + a\lambda_k(\mathcal{L}_h) + bc\right\} > b\mathrm{Lip}(R). \tag{B.17}$$

Recall the decomposition of $f(u) = cu + (f(u) - cu) = cu + R(u)$. By (D), $c \geq 0$. We can then set $K = c, \phi = R$ in Theorem 1. Furthermore, (C) implies $\lambda_k(Q_1^\top \mathcal{L}_h Q_1) = \lambda_k(\mathcal{L}_h)$. As a result, (B.17) is equivalent to (3.3) in Theorem 1, which leads to the unique existence of the root-finding problem $\widehat{F}(U) = 0$. □

B.3. **Proofs of section 3.3.** Before we prove Theorem 8, we need Lemma 16, 17 and 18.

Lemma 16. *Suppose $\theta \in [0, \sqrt{2} - 1)$, there exist $u, k > 0$, s.t.*

$$ku\Psi(\theta) - \frac{1}{4}\Omega(u, k, \theta)^2 > 0,$$

where $\Psi(\theta) = 1 - 2\theta - \theta^2,\ \Omega(u, k, \theta) = |1 - u - k| + \theta(|1 - u| + k)$.

Proof of Lemma 16. We note that $\Omega(u,k,\theta)^2 \leq ((1+\theta)(|1-u|+k))^2 \leq 2(1+\theta)^2((1-u)^2+k^2)$. Then for any $u, k > 0$, we have

$$\begin{aligned} ku\Psi(\theta) - \frac{1}{4}\Omega(u,k,\theta)^2 &\geq ku\Psi(\theta) - \frac{1}{2}(1+\theta)^2((1-u)^2+k^2)) \\ &= ku(1+\theta)^2\left(\frac{\Psi(\theta)}{(1+\theta)^2} - \frac{((1-u)^2+k^2)}{2ku}\right) \\ &\geq ku(1+\theta)^2\left(\frac{\Psi(\theta)}{(1+\theta)^2} - \frac{\sqrt{k^2+1}-1}{k}\right). \end{aligned}$$

Denote $c = \frac{\Psi(\theta)}{(1+\theta)^2}$. For any $\theta \in [0, \sqrt{2}-1)$, $c \in (0,1]$. As shown in Figure 19, it is not hard to verify that $\frac{\sqrt{k^2+1}-1}{k}$ increases monotonically from 0 to 1 on $\mathbb{R}_+$. Thus, $\frac{\Psi(\theta)}{(1+\theta)^2} - \frac{\sqrt{k^2+1}-1}{k} > 0$ is guaranteed to have a positive solution $k > 0$. This proves the lemma. □

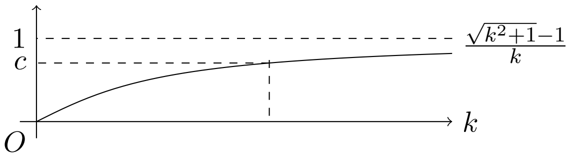


FIGURE 19. Graph of $\frac{\sqrt{k^2+1}-1}{k}$.

Lemma 17. *Suppose $F : \mathbb{R}^d \to \mathbb{R}^d$ is differentiable on $\mathbb{R}^d$. Let $\boldsymbol{v} \in \mathbb{R}^d$. Then, for any $x, y \in \mathbb{R}^d$, there exists $t_{\boldsymbol{v}} \in (0,1)$ such that*

$$\boldsymbol{v}^\top(F(y)-F(x)) = \boldsymbol{v}^\top DF(x+t_{\boldsymbol{v}}(y-x))(y-x).$$

Proof of Lemma 17. Define $h(t) = \boldsymbol{v}^\top(F(x+t(y-x))-F(x))$. Since $h(\cdot)$ is differentiable on $(0,1)$, by mean value theorem, there exists $t_{\boldsymbol{v}} \in (0,1)$ such that $h(1)-h(0) = h'(t_{\boldsymbol{v}})$. □

Lemma 18. *Suppose a positive sequence $\{a_k\}_{k\geq 0}$ satisfies the following recurrence inequality*

$$a_{k+2} - a_k \leq -\Phi\, a_{k+1}, \quad k \geq 0 \tag{B.18}$$

with $\Phi > 0$. Then

$$a_k \leq \left(\frac{2}{\Phi+\sqrt{\Phi^2+4}}\right)^{k+1}\left(a_1 + \frac{\Phi+\sqrt{\Phi^2+4}}{2}a_0\right) \quad \textit{for } k \geq 1.$$

Proof of Lemma 18. We consider the characteristic polynomial $r^2 + \Phi r - 1 = 0$. It has two roots $r_+ = \frac{-\Phi+\sqrt{\Phi^2+4}}{2} > 0$ and $r_- = \frac{-\Phi-\sqrt{\Phi^2+4}}{2} < 0$. Then $\Phi = \frac{1-r_+^2}{r_+} = \frac{1}{r_+} - r_+$. Plugging this back to (B.18) yields

$$a_{k+2} + \left(\frac{1}{r_+} - r_+\right)a_{k+1} - a_k \leq 0, \quad k \geq 0,$$

which further leads to

$$a_{k+2} + \frac{1}{r_+}a_{k+1} \leq r_+\left(a_{k+1} + \frac{1}{r_+}a_k\right) \quad k \geq 0.$$

Thus, we obtain

$$a_{k+1} + \frac{1}{r_+}a_k \leq r_+^k\left(a_1 + \frac{1}{r_+}a_0\right), \quad \text{for any } k \geq 0. \tag{B.19}$$

Taking the index in (B.19) as $k-1$ and k, one obtains

$$\begin{aligned} r_+^{k-1}\left(a_1+\frac{1}{r_+}a_0\right) &\geq a_k+\frac{1}{r_+}a_{k-1} > a_k; \\ r_+^{k}\left(a_1+\frac{1}{r_+}a_0\right) &\geq a_{k+1}+\frac{1}{r_+}a_{k} > \frac{1}{r_+}a_k. \end{aligned}$$

This yields

$$a_k \leq r_+^{k-1}\left(a_1+\frac{1}{r_+}a_0\right), \quad \text{and} \quad a_k \leq r_+^{k+1}\left(a_1+\frac{1}{r_+}a_0\right), \quad k\geq 1.$$

Since $r_+<1$, we finally obtain

$$a_k \leq \left(\frac{2}{\Phi+\sqrt{\Phi^2+4}}\right)^{k+1}\left(a_1+\frac{\Phi+\sqrt{\Phi^2+4}}{2}a_0\right), \quad k\geq 1.$$

□

We now prove Theorem 8.

Proof of Theorem 8. According to Lemma 15, under conditions (A), (B), (C), (D), and

$$bT\zeta_{a,b,c}(h_t)\mathrm{Lip}(R) < \sqrt{2}-1 < 1,$$

it is straightforward to check the unique existence of the root-finding problem $\widehat{F}(U)=0$.

Now we suppose $\{U_k, Q_k\}$ solves (2.16). We write $\mathcal{J}_k=\mathcal{J}(U_k,Q_k)$ for convenience. Then we want to bound $\mathcal{J}_{k+1}-\mathcal{J}_k$ from above. We calculate

$$\begin{aligned} \mathcal{J}_{k+1}-\mathcal{J}_k =& (U_{k+1}-U_k)\cdot\left(\frac{1}{2}(U_{k+1}+U_k)-U_*\right)+(Q_{k+1}-Q_k)\cdot\left(\frac{Q_{k+1}+Q_k}{2}\right) \\ \leq& (U_{k+1}-U_k)\cdot\left(\frac{1}{2}(U_{k+1}+U_k)-U_*\right)+(Q_{k+1}-Q_k)\cdot Q_{k+1} \\ =& (U_{k+1}-U_k)\cdot(U_k-U_*)+\frac{1}{2}\|U_{k+1}-U_k\|^2+(Q_{k+1}-Q_k)\cdot Q_{k+1}. \end{aligned}$$

The inequality is due to the convexity of the quadratic function $\|Q\|^2$. From (2.16), we know

$$\begin{aligned} U_{k+1}-U_k &= -\tau_U D\widehat{F}(U_k)^\top(Q_{k+1}+\omega\tau_P(\widehat{F}(U_k)-\epsilon Q_{k+1})); \\ &= -\tau_U D\widehat{F}(U_k)^\top((1-\widetilde{\gamma}\epsilon)Q_{k+1}+\widetilde{\gamma}\widehat{F}(U_k)); \\ Q_{k+1}-Q_k &= \tau_P(\widehat{F}(U_k)-\epsilon Q_{k+1}). \end{aligned}$$

Let us define $\widetilde{\gamma}=\omega\tau_P$ and $\varrho=\frac{\tau_P}{\tau_U}$. Using $F(U_*)=0$, we obtain

$$\begin{aligned} &\mathcal{J}_{k+1}-\mathcal{J}_k \\ =& -\tau_U(U_k-U_*)^\top D\widehat{F}(U_k)^\top((1-\widetilde{\gamma}\epsilon)Q_{k+1}+\widetilde{\gamma}\widehat{F}(U_k))+\tau_P Q_{k+1}^\top(\widehat{F}(U_k)-\epsilon Q_{k+1})+\frac{1}{2}\|U_{k+1}-U_k\|^2 \\ =& -\tau_U\Big(\widetilde{\gamma}(U_k-U_*)^\top D\widehat{F}(U_k)^\top\widehat{F}(U_k)+(1-\widetilde{\gamma}\epsilon)(U_k-U_*)^\top D\widehat{F}(U_k)^\top Q_{k+1} \\ &\qquad -\varrho\widehat{F}(U_k)^\top Q_{k+1}+\varrho\epsilon\|Q_{k+1}\|^2\Big)+\frac{\tau_U^2}{2}\|D\widehat{F}(U_k)^\top((1-\widetilde{\gamma}\epsilon)Q_{k+1}+\widetilde{\gamma}\widehat{F}(U_k))\|^2 \\ =& -\tau_U\Big(\underbrace{\widetilde{\gamma}(U_k-U_*)^\top D\widehat{F}(U_k)^\top(\widehat{F}(U_k)-\widehat{F}(U_*))}_{(A)}+\underbrace{(1-\widetilde{\gamma}\epsilon)(U_k-U_*)^\top D\widehat{F}(U_k)^\top Q_{k+1}}_{(B)} \\ & \underbrace{-\varrho(\widehat{F}(U_k)-\widehat{F}(U_*))^\top Q_{k+1}}_{(C)}+\underbrace{\varrho\epsilon\|Q_{k+1}\|^2}_{(D)}\Big)+\frac{\tau_U^2}{2}\underbrace{\|D\widehat{F}(U_k)^\top((1-\widetilde{\gamma}\epsilon)Q_{k+1}+\widetilde{\gamma}\widehat{F}(U_k))\|^2}_{(E)}. \end{aligned}$$

By Lemma 17, term (A) and term (C) are given by

$$\begin{aligned}(A) &= \widetilde{\gamma}(U_k - U_*)^\top D\widehat{F}(U_k)^\top D\widehat{F}(U_{k,\nu_1})(U_k - U_*)\,,\\ (C) &= -(U_k - U_*)^\top D\widehat{F}(U_{k,\nu_2})^\top Q_{k+1}\,,\end{aligned}$$

where $U_{k,\nu_j} = U_* + \nu_j(U_k - U_*)$ with $\nu_1, \nu_2 \in (0,1)$, $j = 1, 2$.

Recall that $D\widehat{F}(U) = I + D\eta(U)$. To simplify the notation, we write

$$\overline{\sigma}_\eta = \sup_{U\in\mathbb{R}^n} \{\|D\eta(U)\|\}\,.$$

By Lemma 6, we have $\overline{\sigma}_\eta \le bT\zeta_{a,b,c}(h_t)\mathrm{Lip}(R)$. We now estimate term (A) as

$$\begin{aligned}(A) =&\widetilde{\gamma}(U_k - U_*)D\widehat{F}(U_k)^\top D\widehat{F}(U_{k,\nu_1})(U_k - U_*)\\ =&(U_k - U_*)^\top (I + D\eta(U_k)^\top)(I + D\eta(U_{k,\nu_1}))(U_k - U_*)\\ =&\|U_k - U_*\|^2 + (U_k - U_*)^\top D\eta(U_k)^\top (U_k - U_*) + (U_k - U_*)^\top D\eta(U_{k,\nu_1})(U_k - U_*)\\ &+ (U_k - U_*)^\top D\eta(U_k)^\top D\eta(U_{k,\nu_1})(U_k - U_*)\\ \ge&(1 - 2\overline{\sigma}_\eta - \overline{\sigma}_\eta^2)\|U_k - U_*\|^2.\end{aligned}$$

We can further estimate the terms (B), (C), and (E) as

$$\begin{aligned}(B) &= (1 - \widetilde{\gamma}\epsilon)(U_k - U_*)^\top (I + D\eta(U_k))Q_{k+1};\\ (C) &= -\varrho(U_k - U_*)^\top D\widehat{F}(U_{k,\nu_2})^\top Q_{k+1} = -\varrho(U_k - U_*)^\top (I + D\eta(U_{k,\nu_2}))^\top Q_{k+1}.\end{aligned}$$

Thus

$$\begin{aligned}(B) + (C) =&(1 - \widetilde{\gamma}\epsilon)(U_k - U_*)^\top (I + D\eta(U_k))Q_{k+1} - \varrho(U_k - U_*)^\top (I + D\eta(U_{k,\nu_2}))^\top Q_{k+1}\\ =&(1 - \widetilde{\gamma}\epsilon - \varrho)(U_k - U_*)^\top Q_{k+1} + (U_k - U_*)^\top ((1 - \widetilde{\gamma}\epsilon)D\eta(U_k) - \varrho D\eta(U_{k,\nu_2}))^\top Q_{k+1}\\ \ge&-|1 - \widetilde{\gamma}\epsilon - \varrho|\|U_k - U_*\|\|Q_{k+1}\| - (|1 - \widetilde{\gamma}\epsilon| + \varrho)\bar{\sigma}_\eta\|U_k - U_*\|\|Q_{k+1}\|\\ =&-(|1 - \widetilde{\gamma}\epsilon - \varrho| + (|1 - \widetilde{\gamma}\epsilon| + \varrho)\bar{\sigma}_\eta)\|U_k - U_*\|\|Q_{k+1}\|\,.\end{aligned}$$

And

$$\begin{aligned}(E) &\le \overline{\sigma}^2(|1 - \widetilde{\gamma}\epsilon| \cdot \|Q_{k+1}\| + \widetilde{\gamma}\|\widehat{F}(U_k)\|)^2\\ &\le \overline{\sigma}^2(|1 - \widetilde{\gamma}\epsilon| \cdot \|Q_{k+1}\| + \widetilde{\gamma}\overline{\sigma}\|U_k - U_*\|)^2\\ &\le 2\overline{\sigma}^2((1 - \widetilde{\gamma}\epsilon)^2\|Q_{k+1}\|^2 + \widetilde{\gamma}^2\overline{\sigma}^2\|U_k - U_*\|^2).\end{aligned}$$

The second inequality on (E) is due to

$$\begin{aligned}\|\widehat{F}(U_k)\| =&\|\widehat{F}(U_k) - \widehat{F}(U_*)\| = \left\| \int_0^1 \left(\frac{d}{ds}\widehat{F}(U_* + s(U_k - U_*)) \right)\, ds \right\|\\ =&\left\| \int_0^1 D\widehat{F}(U_* + s(U_k - U_*))(U_k - U_*)\, ds \right\|\\ \le&\int_0^1 \overline{\sigma}\|U_k - U_*\|\, ds = \overline{\sigma}\|U_k - U_*\|.\end{aligned}$$

Combining the estimations on term (A)-(E), we obtain

$$
\begin{aligned}
&\mathcal{J}_{k+1}-\mathcal{J}_k\\
&=-\tau_U\Big(\widetilde{\gamma}(1-2\overline{\sigma}_\eta-\overline{\sigma}_\eta^2)\|U_k-U_*\|^2-(|1-\widetilde{\gamma}\epsilon-\varrho|+(|1-\widetilde{\gamma}\epsilon|+\varrho)\bar{\sigma}_\eta)\|U_k-U_*\|\|Q_{k+1}\|\\
&\qquad\qquad+\varrho\epsilon\|Q_{k+1}\|^2-\tau_U(\overline{\sigma}^2(1-\widetilde{\gamma}\epsilon)^2\|Q_{k+1}\|^2+\widetilde{\gamma}^2\overline{\sigma}^2\|U_k-U_*\|^2)\Big)\\
&=-\tau_U[\|U_k-U_*\|,\|Q_{k+1}\|]\,(\Gamma-\tau_U\Theta)\begin{bmatrix}\|U_k-U_*\|\\ \|Q_{k+1}\|\end{bmatrix}.
\end{aligned}
\tag{B.20}
$$

Here

$$
\Gamma=\begin{bmatrix}\widetilde{\gamma}(1-2\overline{\sigma}_\eta-\overline{\sigma}_\eta^2) & -\frac{1}{2}(|1-\widetilde{\gamma}\epsilon-\varrho|+(|1-\widetilde{\gamma}\epsilon|+\varrho)\bar{\sigma}_\eta)\\ -\frac{1}{2}(|1-\widetilde{\gamma}\epsilon-\varrho|+(|1-\widetilde{\gamma}\epsilon|+\varrho)\bar{\sigma}_\eta) & \varrho\epsilon\end{bmatrix},
$$

$$
\Theta=\begin{bmatrix}\widetilde{\gamma}^2\overline{\sigma}^4 & \\ & \overline{\sigma}^2(1-\widetilde{\gamma}\epsilon)^2\end{bmatrix}.
$$

Recall that we assume $bT\zeta_{a,b,c}(h_t)\mathrm{Lip}(R)\leq\theta$, this leads to $\overline{\sigma}_\eta\leq\theta$. By Lemma 6, we also have $\overline{\sigma}\leq 1+\theta$. Thus, $\widetilde{\gamma}(1-2\overline{\sigma}_\eta-\overline{\sigma}_\eta^2)>\widetilde{\gamma}(1-2\theta-\theta^2)>0$ as $\theta\in[0,\sqrt{2}-1)$. Hence,

$$
\begin{aligned}
\det(\Gamma)=&\varrho\widetilde{\gamma}\epsilon(1-2\overline{\sigma}_\eta-\overline{\sigma}_\eta^2)-\frac{1}{4}(|1-\widetilde{\gamma}\epsilon-\varrho|+(|1-\widetilde{\gamma}\epsilon|+\varrho)\bar{\sigma}_\eta)^2\\
\geq&\varrho\widetilde{\gamma}\epsilon(1-2\theta-\theta^2)-\frac{1}{4}(|1-\widetilde{\gamma}\epsilon-\varrho|+(|1-\widetilde{\gamma}\epsilon|+\varrho)\theta)^2.
\end{aligned}
$$

We denote $\Psi(\theta)=1-2\theta-\theta^2$ and $\Omega(u,\varrho,\theta)=|1-u-\varrho|+(|1-u|+\varrho)\theta$. Lemma 16 guarantees that there exist $\widetilde{\gamma},\omega,\epsilon$, such that (3.16) holds. The condition (3.16) leads to $\det(\Gamma)>0$, which guarantees the positive definiteness of Γ.

Furthermore, we have $\Gamma\succeq\lambda_2(\Gamma)I$, where $\lambda_2(\Gamma)$ represents the smallest eigenvalue of Γ and I is an identity matrix. One can bound $\lambda_2(\Gamma)$ from below as

$$
\begin{aligned}
\lambda_2(\Gamma)=&\frac{\widetilde{\gamma}(1-2\overline{\sigma}_\eta-\overline{\sigma}_\eta^2)+\varrho\epsilon-\sqrt{(\widetilde{\gamma}(1-2\overline{\sigma}_\eta-\overline{\sigma}_\eta^2)+\varrho\epsilon)^2-4\det(\Gamma)}}{2}\\
\geq&\frac{4(\varrho\widetilde{\gamma}\epsilon(1-2\theta-\theta^2)-\frac{1}{4}(|1-\widetilde{\gamma}\epsilon-\varrho|+(|1-\widetilde{\gamma}\epsilon|+\varrho)\theta)^2)}{2(\widetilde{\gamma}(1-2\overline{\sigma}_\eta-\overline{\sigma}_\eta^2)+\varrho\epsilon+\sqrt{(\widetilde{\gamma}(1-2\overline{\sigma}_\eta-\overline{\sigma}_\eta^2)+\varrho\epsilon)^2-4\det(\Gamma)})}\\
\geq&\frac{\varrho\widetilde{\gamma}\epsilon(1-2\theta-\theta^2)-\frac{1}{4}(|1-\widetilde{\gamma}\epsilon-\varrho|+(|1-\widetilde{\gamma}\epsilon|+\varrho)\theta)^2}{\widetilde{\gamma}(1-2\overline{\sigma}_\eta-\overline{\sigma}_\eta^2)+\varrho\epsilon}\\
\geq&\frac{\varrho\widetilde{\gamma}\epsilon\Psi(\theta)-\frac{1}{4}\Omega(\widetilde{\gamma}\epsilon,\varrho,\theta)^2}{\widetilde{\gamma}+\varrho\epsilon}.
\end{aligned}
$$

On the other hand, we have

$$
\Theta\prec\overline{\sigma}^2\max\{\widetilde{\gamma}^2\overline{\sigma}^2,|1-\widetilde{\gamma}\epsilon|^2\}I\prec(1+\theta)^2\max\{\widetilde{\gamma}^2(1+\theta)^2,|1-\widetilde{\gamma}\epsilon|^2\}I.
$$

Thus we have

$$
\Gamma-\tau\Theta\succ\underbrace{\left(\frac{\varrho\widetilde{\gamma}\epsilon\Psi(\theta)-\frac{1}{4}\Omega(\widetilde{\gamma}\epsilon,\varrho,\theta)^2}{\widetilde{\gamma}+\varrho\epsilon}-\tau(1+\theta)^2\max\{\widetilde{\gamma}^2(1+\theta)^2,|1-\widetilde{\gamma}\epsilon|^2\}\right)}_{\text{denote as }C(\theta,\widetilde{\gamma},\epsilon,\varrho,\tau)}I.
$$

Plug this estimation to (B.20), we obtain

$$
\mathcal{J}_{k+1}-\mathcal{J}_k\leq-\tau C(\theta,\widetilde{\gamma},\epsilon,\varrho,\tau)(\|U_k-U_*\|^2+\|Q_{k+1}\|^2).
$$

Since we set the PDHG step size as

$$0 < \tau < \bar{\tau}(\theta, \widetilde{\gamma}, \epsilon, \varrho, \tau) \triangleq \frac{\varrho\widetilde{\gamma}\epsilon\Psi(\theta) - \frac{1}{4}\Omega(\widetilde{\gamma}\epsilon, \varrho, \theta)^2}{2(\widetilde{\gamma} + \varrho\epsilon)(1+\theta)^2 \max\{\widetilde{\gamma}^2(1+\theta)^2, (1-\widetilde{\gamma}\epsilon)^2\}},$$

this guarantees $C(\theta, \widetilde{\gamma}, \epsilon, \varrho, \tau) > 0$.

Furthermore, as a function of τ, $\tau C(\theta, \widetilde{\gamma}, \epsilon, \varrho, \tau)$ reaches its maximum value at $\tau = \frac{1}{2}\bar{\tau}(\theta, \widetilde{\gamma}, \epsilon, \varrho, \tau)$. We then set (here $\bar{\tau}$ denotes $\bar{\tau}(\theta, \widetilde{\gamma}, \epsilon, \varrho)$)

$$\Phi = \frac{1}{2}\bar{\tau}C(\theta, \widetilde{\gamma}, \epsilon, \varrho, \frac{1}{2}\bar{\tau}) = \frac{(\varrho\widetilde{\gamma}\epsilon\Psi(\theta) - \frac{1}{4}\Omega(\widetilde{\gamma}\epsilon, \varrho, \theta)^2)^2}{2(1+\theta)^2 \max\{\widetilde{\gamma}^2(1+\theta)^2, (1-\widetilde{\gamma}\epsilon)^2\}(\widetilde{\gamma} + \varrho\epsilon)^2}.$$

Thus we have

$$\mathcal{J}_{k+1} - \mathcal{J}_k \leq -\Phi \cdot \frac{1}{2}(\|U_k - U_*\|^2 + \|Q_{k+1}\|^2).$$

Now we prove the exponential decay of $\mathcal{J}_k$. To do so, we sum up the above inequality at time index k and $k+1$ to obtain,

$$\mathcal{J}_{k+2} - \mathcal{J}_k \leq -\Phi \cdot \frac{1}{2}(\|U_{k+1} - U_*\|^2 + \|Q_{k+2}\|^2 + \|U_k - U_*\|^2 + \|Q_{k+1}\|^2), \quad k \geq 0.$$

It is not hard to see that the right-hand side of the above inequality is no larger than $-\Phi\mathcal{J}_{k+1}$. Hence,

$$\mathcal{J}_{k+2} - \mathcal{J}_k \leq -\Phi\mathcal{J}_{k+1}. \tag{B.21}$$

Now, by Lemma 18, we obtain

$$\mathcal{J}_k \leq \left(\frac{2}{\Phi + \sqrt{\Phi^2 + 4}}\right)^{k+1} \left(\mathcal{J}_1 + \frac{\Phi + \sqrt{\Phi^2 + 4}}{2}\mathcal{J}_0\right), \quad \text{for } k \geq 1.$$

This concludes our proof. □

Email address: shuliu@math.ucla.edu

Department of Mathematics, University of California, Los Angeles

Email address: zxz@math.ucla.edu

Department of Mathematics, University of California, Los Angeles

Email address: sjo@math.ucla.edu

Department of Mathematics, University of California, Los Angeles

Email address: wuchen@mailbox.sc.edu

Department of Mathematics, University of South Carolina