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 [32], 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 [36]; the study of the reaction processes of multiple chemicals [39, 37]; and the modeling & prediction of crimes [42].

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 [36, 32]. In addition, time-implicit schemes are also known to be energy-stable [46].

In a recent work [32], 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. The method does not require extra effort to compute the inverse of the Jacobian matrix in the time-implicit scheme. This feature is distinct from many classical algorithms, such as...
Newton’s methods for solving nonlinear equations. Nevertheless, the PDHG algorithm presented in [32] faces both 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 first briefly review the treatment in [32]. Consider the following 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 $h_t > 0$.

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

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

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

We first reformulate (1.2) as an inf-sup problem with a tunable parameter $\epsilon > 0$ following the treatment in [32]

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

Here we write $p = [p_1, \ldots, p_t, \ldots, p_{N_t}] \in X^{N_t}$. We point out that the saddle point of inf-sup problem (1.4) solves the time-implicit scheme $\mathcal{F}(u) = 0$ whenever (1.3) admits a unique solution. It is worth mentioning that in the above treatment, we slightly generalize the algorithm proposed in [32] by accumulating multiple time steps in a root-finding problem and introducing a quadratic regularization term in the saddle point problem (1.4).

As demonstrated in [32], we deal with the inf-sup saddle problem by applying the primal-dual hybrid gradients (PDHG) algorithm [7, 50]. We further substitute the proximal step of variable $u$ with an explicit update to obtain

$$
\begin{align*}
\mathcal{P}_{n+1} &= \frac{1}{1 + \epsilon \tau_p} (\mathcal{P}_n + \tau_p \mathcal{F}(u_n)), \\
\mathcal{P}_{n+1} = &\mathcal{P}_{n+1} + \omega(\mathcal{P}_{n+1} - \mathcal{P}_n), \\
\mathcal{U}_{n+1} &= \mathcal{U}_n - \tau_U \mathcal{D}(\mathcal{U}_n) \mathcal{F}(u_n). \tag{1.5}
\end{align*}
$$

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

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 $\mathcal{R} : X^{N_t} \rightarrow X^{N_t}$, where $\mathcal{R}$ is extracted from the linear part of $\mathcal{F}(\cdot)$. Then
we introduce the preconditioned functional $\tilde{F}(u) = M^{-1}F(u)$. We apply the PDHG algorithm (1.5) to $\tilde{F}(u) = 0$ to obtain

$$
\begin{align*}
    p_{n+1} &= \frac{1}{1 + \epsilon \tau P}(p_n + \tau P M^{-1} F(u_n)), \\
    \bar{p}_{n+1} &= p_{n+1} + \omega(p_{n+1} - p_n), \\
    u_{n+1} &= u_n - \tau U D F(u_n)^* (M^{-1})^* \bar{p}_{n+1}.
\end{align*}
$$

(1.6)

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

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 $(u_n, p_n)$ 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 $L_h, G_h$ 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 (2.15) 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 $\gamma, \epsilon$ 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 (2.15) 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 $M$ (c.f. (2.12)) 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 4.2, 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 4.3 and section 4.4.

There exist plenty of references regarding the numerical schemes for RD equations, which include studies on finite difference methods [6, 11, 15, 22, 23, 25, 35, 40, 41, 46, 48], and finite element methods [18, 19, 23, 28, 29, 30, 31, 49]. A series of benchmark problems [12, 25] have 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) [20, 21], physics-informed neural networks (PINNs) [38, 45, 47], 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, 50] 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) \cite{14}, large-scale optimization problems including image denoising and optimal transport \cite{24}, computing gradient flows in Wasserstein-like transport metric spaces \cite{4} \cite{5} \cite{18}, as well as design fast optimization algorithms \cite{52}, etc.

In \cite{16}, the authors introduce damping terms to the wave equation to achieve faster stabilization, which resembles the time-continuous limit (the PDHG flow) \cite{3.3} of our proposed algorithm. However, \cite{16} focuses on the linear case while our research deals with nonlinear RD equations. In recent work \cite{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 \cite{8}, the authors apply the splitting method to propose an accelerating algorithm for the root-finding problem. Although the transformed algorithm shares similarities with our method, the 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 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 \cite{34}. The aforementioned works \cite{32} \cite{33} \cite{34} 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 \cite{1.1}. From now on, we assume that the domain $\Omega = [0, L]^2$ is a square region.

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

$$
\frac{U^{t+1} - U^t}{h_t} = -G_h(aL_hU^{t+1} + bf(U^{t+1})), \quad \text{for } t = 0, 1, \ldots, N_t, \text{ with } U_0 \text{ given.} \quad (2.1)
$$

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 $G_h$, $L_h$ as $N_x^2 \times N_x^2$ matrices, which represents the discretization of the operator $L, G$ 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 \cite{1}, we have $G = \text{Id}$, $L = -\Delta$; for Cahn-Hilliard equation \cite{3}, we have $G = -\Delta$, $L = -\Delta$. 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 $\Delta$. We obtain $\Delta^P_{h_x} = I_{N_x} \otimes \text{Lap}^P_{h_x} + \text{h}x^2$.
We can thus reformulate (2.8) as an inf-sup problem with a tunable parameter

\[ \inf_{q} \sup_{P} P^T F(U). \tag{2.7} \]

Inspired by [52], we replace \( \iota \) by a milder quadratic function \( q_\epsilon(\cdot) = \frac{1}{2\epsilon} \| \cdot \|^2 \) to obtain the minimization problem

\[ \inf_{U \in \mathbb{R}^{N_1 N_2}} q_\epsilon(F(U)). \tag{2.8} \]

Since \( q_\epsilon \) is the Legendre transform of its dual \( q_\epsilon^*(\cdot) = \frac{\epsilon}{2} \| \cdot \|^2 \), i.e., \( q_\epsilon(U) = \sup_{P \in \mathbb{R}^{N_1 N_2}} P^T U - q_\epsilon^*(P) \).

We can thus reformulate (2.8) as an inf-sup problem with a tunable parameter \( \epsilon \),

\[ \inf_{U \in \mathbb{R}^{N_1 N_2}} \sup_{P \in \mathbb{R}^{N_1 N_2}} L(U, P) \triangleq P^T F(U) - \frac{\epsilon}{2} \| P \|^2. \tag{2.9} \]
We tackle this saddle point problem by leveraging the primal-dual hybrid gradient (PDHG) algorithm and obtain

\begin{equation}
\begin{aligned}
P_{n+1} &= \frac{1}{1 + \epsilon \tau} (P_n + \tau P F(U_n)), \\
\bar{P}_{n+1} &= P_{n+1} + \omega (P_{n+1} - P_n), \\
U_{n+1} &= U_n - \tau \nabla F(U_n)^\top \bar{P}_{n+1}.
\end{aligned}
\end{equation}

\tag{2.10}

When \(DF(U)\) is nonsingular for arbitrary \(U \in \mathbb{R}^{N_t N^2_L}\), 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 \(2.10\). To mitigate this, we consider suitable preconditioning of \(F(\cdot)\). Let us decompose \(F(U)\) into its linear part and nonlinear part,

\[F(U) = DU + h_t G_h(a L_h U + b f(U)) - V = (D + ah_t G_h L_h) U + bh_t G_h (f(\bar{U}) + J_f(U - \bar{U}) + R(U)) - V.\]

\tag{2.11}

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

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

By writing

\[M = D + ah_t G_h L_h + bh_t G_h J_f = \begin{bmatrix} X & -I & X \\ -I & \ddots & \ddots \\ & -I & X \end{bmatrix} \quad \text{with} \quad X = I + ah_t G_h L_h + bh_t G_h J_f,\]

\[\bar{V} = bh_t G_h (f(\bar{U}) - J_f \bar{U}) - V,\]

we decompose \(F(U)\) as \(M U + bh_t G_h R(U) - \bar{V}\). It is beneficial to consider the preconditioned function

\[\hat{F}(U) = M^{-1} F(U) = U + M^{-1} (bh_t G_h R(U)) - \bar{V} \quad \text{denote as} \quad U + \eta(U).\]

\tag{2.13}

We discuss the sufficient condition under which \(M\) is invertible in the following remark.

**Remark 3** (Invertibility of \(M\)). Suppose \(a, b \geq 0\), \(G_h, L_h\) are self-adjoint, non-negative definite, and commute. Assume \(J_f = cI\) with \(c \geq 0\). Then \(M\) 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 \(I + ah_t \Lambda^\top G_h + bh \Lambda^\top G_h\), which is invertible for \(h_t > 0\). Here \(\Lambda G_h, \Lambda L_h\) are diagonal matrices equivalent to \(G_h, \Lambda L_h\).

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

We now apply \(2.10\) to the inf-sup saddle problem with respect to \(\hat{F}(\cdot)\)

\begin{equation}
\begin{aligned}
\inf_{U \in \mathbb{R}^{N_t N^2_L}} \sup_{Q \in \mathbb{R}^{N_t N^2_L}} \hat{L}(U, Q) \triangleq Q^\top \hat{F}(U) - \frac{c}{2} \|Q\|_2^2.
\end{aligned}
\end{equation}

\tag{2.14}
And our PDHG method with *implicit* update in $Q$ and *explicit* update in $U$ yields

$$
Q_{k+1} = \frac{1}{1 + \epsilon \tau P}(Q_k + \tau P(\hat{F}(U_k))); \\
\bar{Q}_{k+1} = Q_{k+1} + \omega(\bar{Q}_{k+1} - \bar{Q}_k); \\
U_{k+1} = U_k - \tau U(D\hat{F}(U_k)^\top \bar{Q}_{k+1}).
$$

We then iterate (2.15) so that \{$U_k$\} approaches the desired root $U_\ast$. We terminate the iteration whenever the $\ell^\infty$ norm of the residual term

$$
\text{Res}(U_k) = F(U_k)/h_t = \left[\ldots, \left(\frac{U_{k+1}^t - U_k^t}{h_t} + G_h(a\mathcal{L}_h U_{k+1}^t + b f(U_{k+1}^t))\right)^\top, \ldots\right]_{0 \leq t \leq N_t - 1}
$$

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

2.2. **Complexity of the algorithm.** We apply the Fast Fourier Transform (FFT) [14, 43] to evaluate the multiplication of $\mathcal{L}_h, G_h$ for periodic boundary conditions. We refer interested readers to [32] for more details. Thus, computing $F(U)$ requires $O(N_t N_x^2 \log N_x)$ steps of operations. Furthermore, since $\mathcal{M}$ is block lower triangular, applying back substitution together with FFT to solve the linear system involving $\mathcal{M}$ requires $O(N_t N_x^2 \log N_x)$ steps of operations. Thus, the complexity at each iteration of our algorithm equals $O(N_t N_x^2 \log(N_x))$.

2.3. **Relation with G-prox PDHG method.** The G-prox primal-dual hybrid gradients algorithm [24] was recently invented to improve the convergence of optimization and root-finding problems. The algorithm can be formulated as

$$
P_{k+1} = \arg\min_{P \in \mathbb{R}^{N_t N_x^2}} \left\{ \frac{1}{2\tau P} \|P - P_k\|_G^2 - \bar{L}(U_k, P) \right\} = \frac{1}{1 + \epsilon \tau P}(P_k + \tau P G^{-1} F(U_k)); \\
\bar{P}_{k+1} = P_{k+1} + \omega(P_{k+1} - P_k); \\
U_{k+1} = \arg\min_{U \in \mathbb{R}^{N_t N_x^2}} \left\{ \frac{1}{\tau P} \|U - U_k\|_G^2 + \bar{L}(U_k, \bar{P}_{k+1}) \right\}.
$$

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

$$
U_{k+1} = U_k - \tau U(D\hat{F}(U_k))^\top \bar{P}_{k+1}.
$$

Now, we multiply $\mathcal{M}^\top$ on both sides of (2.17) (but with the third line replaced by (2.18)) to obtain

$$
\mathcal{M}^\top P_{k+1} = \frac{1}{1 + \epsilon \tau P}(\mathcal{M}^\top P_k + \tau P \mathcal{M}^{-1} F(U_k)); \\
\mathcal{M}^\top \bar{P}_{k+1} = \mathcal{M}^\top P_{k+1} + \omega(\mathcal{M}^\top P_{k+1} - \mathcal{M}^\top P_k); \\
U_{k+1} = U_k - \tau U(D\hat{F}(U_k))^\top (\mathcal{M}^\top \bar{P}_{k+1}).
$$

By denoting $Q_k = \mathcal{M}^\top P_k$ and noticing that $\hat{F}(U) = \mathcal{M}^{-1} F(U)$, (2.19) reduces exactly to (2.15). 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
\[
\mathcal{F} = \left\{ f \in C^1(\mathbb{R}) \mid f \text{ can be decomposed as } f = V' + \phi, \right. \\
\left. \text{where } V \in C^1(\mathbb{R}) \text{ is convex, and } \phi \in C(\mathbb{R}) \text{ is Lipschitz.} \right\}.
\] (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 \( \mathcal{G}_h \):
\[
\mathcal{G}_h = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} \Lambda & O \\ O & \Lambda \end{bmatrix} \begin{bmatrix} Q_1^\top \\ Q_2^\top \end{bmatrix},
\] (3.2)
where \( \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_r) \) is a diagonal matrix with positive entries \( \lambda_1 \geq \cdots \geq \lambda_r > 0 \), \( r = \text{rank}(\mathcal{G}_h) \).

**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
\[
\lambda_{\min} \left( \frac{\Lambda^{-1}}{h_t} + a Q_1^\top \mathcal{L}_h Q_1 \right) + bK > b \text{ Lip}(\phi),
\] (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) \( \mathcal{G} = \text{Id}, \mathcal{L} = -\Delta, f(x) = x^3 - x \). We set \( \mathcal{G}_h = I_{N^2} \), and \( \mathcal{L}_h = -\Delta_{h_x}^P = \tilde{I}_{N_x} \otimes (\text{Lap}_{h_x}^P) + (\text{Lap}_{h_x}^P) \otimes I_{N_x} \), where \( \text{Lap}_{h_x}^P \) 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 \( h_t < \frac{\alpha^2}{b} \).

Similar results regarding both Allen-Cahn and Cahn-Hilliard equations have also been done in [46]. 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 (2.15) 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 \( \tilde{F}(U) = 0 \) regardless of the exact form of \( \tilde{F}(U) \). Our main results are summarized in Theorem 2 and Corollary 3.

Recall (2.15), we substitute \( \tilde{Q}_{k+1} \) with
\[
\tilde{Q}_{k+1} = Q_k + (1 + \omega)\left( Q_{k+1} - Q_k \right) = Q_k + (1 + \omega)\tau_P \left( -\frac{1}{1 + \epsilon \tau_P} Q_k - \frac{1}{1 + \epsilon \tau_P} \tilde{F}(U_k) \right)
\] (3.3)
\[= \left( 1 - \frac{(1 + \omega)\tau_P \epsilon}{1 + \epsilon \tau_P} \right) Q_k - \left( 1 + \omega \right) \tau_P \tilde{F}(U_k).
\]
Then, the PDHG iteration (2.15) can be formulated as
\[
\frac{Q_{k+1} - Q_k}{\tau p} = -\frac{\epsilon}{1 + \epsilon \tau p} Q_k - \frac{1}{1 + \epsilon \tau p} \hat{F}(U_k); \\
\frac{U_{k+1} - U_k}{\tau U} = -D \hat{F}(U_k)^T \left( \left( 1 - \frac{(1 + \omega) \tau_p \epsilon}{1 + \epsilon \tau p} \right) Q_k - \frac{(1 + \omega) \tau_p}{1 + \epsilon \tau p} \hat{F}(U_k) \right).
\]
(3.4)

Suppose we send the step sizes \( \tau_U, \tau_P \to 0 \), and keep \( \omega \) 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”.
\[
\begin{align*}
\dot{Q} &= -\epsilon Q + \hat{F}(U), \\
\dot{U} &= -D \hat{F}(U)^T ((1 - \gamma \epsilon)Q - \gamma \hat{F}(U)).
\end{align*}
\]
(3.5)

We introduce two notations that will be commonly used in the following discussion,
\[
\sigma = \inf_{U \in \mathbb{R}^{N_tN_M^2}} \{ \sigma_{\min}(D \hat{F}(U)) \} = \inf_{U \in \mathbb{R}^{N_tN_M^2}} \{ \sigma_{\min}(I + bh_t M^{-1} \varphi h DR(U)) \},
\]
(3.6)
\[
\overline{\sigma} = \sup_{U \in \mathbb{R}^{N_tN_M^2}} \{ \sigma_{\max}(D \hat{F}(U)) \} = \sup_{U \in \mathbb{R}^{N_tN_M^2}} \{ \sigma_{\max}(I + bh_t M^{-1} \varphi h DR(U)) \},
\]
(3.7)
where \( \sigma_{\min}(A)(\sigma_{\max}(A)) \) denotes the minimum (maximum) singular value of matrix \( A \). The condition number is defined by
\[
\kappa = \overline{\sigma}/\sigma.
\]
(3.8)

We consider the following Lyapunov function of \((U, Q)\) associated with a parameter \( \mu > 0 \),
\[
\mathcal{I}_\mu(U, Q) = \frac{1}{2} \| \hat{F}(U) \|^2 + \mu \| Q \|^2.
\]
(3.9)

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

**Lemma 2** (Exponential decay of \( \mathcal{I}_\mu(U_t, Q_t) \)). Suppose that \( 0 < \sigma \leq \overline{\sigma} < \infty \). We pick the parameter \( \mu > 0 \) satisfying
\[
\frac{1}{\sigma} - \frac{1}{\overline{\sigma}} < \frac{2}{\sqrt{\mu}}.
\]
(3.10)

Furthermore, we choose \( \gamma, \epsilon > 0 \) satisfying
\[
\max \left\{ \left( 1 - \sqrt{\frac{\mu}{\sigma}} \right)^2, \left( 1 - \sqrt{\frac{\mu}{\overline{\sigma}}} \right)^2 \right\} < \gamma \epsilon < \left( 1 + \sqrt{\frac{\mu}{\overline{\sigma}}} \right)^2.
\]
(3.11)

Under the above choices of \( \mu, \gamma \) and \( \epsilon \), 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
\[
\beta = \min_{z \in [\sigma^2, \overline{\sigma}^2]} \{ \varphi_{\mu, \gamma, \epsilon}(z) \} > 0,
\]
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, \( \beta \) does not take an explicit form. In the following theorem, we relax the bound in Lemma 2 to obtain an explicit convergence rate for \( \| \hat{F}(U_t) \| \).
Theorem 3 (Exponential decay of the residual $\|\hat{F} (U_t)\|$). Assume that $(U_t, Q_t)$ solves (3.5) with an arbitrary initial position $(U_0, Q_0)$. Then, as long as $\sigma$ is bounded away from 0 and $\sigma'$ is finite, one can always pick suitable parameters $\gamma, \epsilon$ such that the residual $\|\hat{F} (U_t)\|$ 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

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

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

3.2.2. Convergence analysis for our specific root-finding problem (2.13). 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.,

$$
a \geq 0, \quad b \geq 0.
$$

2. Assume that

$$
f(\cdot) \text{ is Lipschitz with constant } \operatorname{Lip}(f).
$$

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

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

4. Recall $J_f$ mentioned in (2.11). We assume

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

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

Lemma 4. Suppose (A), (B), (C) hold. When $h_t < \frac{1}{\max \{\operatorname{Lip}(J), \sigma\} \sqrt{\sigma h}}$, we always have $\sigma > 0$ and $\sigma' < \infty$.

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 $\|\hat{F} (U_t)\|$). 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 $h_t < \frac{1}{\max \{\mathcal{G}_h \mathcal{L}_h, \mathcal{L}_h \mathcal{G}_h\} \sqrt{J_f}}$. Suppose $\gamma = \frac{1 - \delta}{\kappa}$, and $\epsilon = (1 - \delta)\kappa$ with $\kappa = \sigma' / \sigma$, and $|\delta| < \frac{1}{\kappa}$. Then $\|\hat{F} (U_t)\|$ converges exponentially fast to 0.

Remark 5. It is worth mentioning that we do not assume condition (3.3) of Theorem 7. Then $\hat{F} (U) = 0$ might not admit a unique solution, but the exponential decay of $\|\hat{F} (U_t)\|$ is still guaranteed.

Although Theorem 5 guarantees the exponential convergence of $\|\hat{F} (U_t)\|$ for arbitrarily large $b$ and $T$ as long as $h_t, \gamma, \epsilon$ 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 $\sigma$ and $\sigma'$ 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 $R(U) = f(U) - f(\hat{U}) - Df(\hat{U})(U - \hat{U})$ mentioned in (2.11). 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 \(\text{Lip}(R)\). Define

\[
\zeta_{a,b,c}(h_t) = \max_{1 \leq k \leq N_x^2} \left\{ \frac{\lambda_k(G_h)}{1 + h_t(a \lambda_k(G_h) \lambda_k(L_h) + bc \lambda_k(G_h))} \right\},
\]

where \(\lambda_k(G_h), \lambda_k(L_h)\) are the eigenvalues of \(G_h, L_h\) which are simultaneously diagonalizable by an orthogonal matrix \(Q\). Recall that in (2.13), we have \(\eta(U) = bh_t, \mathcal{A}^{-1}G_hR(U) - \overline{w}\), then

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

And we also have

\[
\sigma \geq 1 - bT \zeta_{a,b,c}(h_t) \text{Lip}(R), \quad \sigma \leq 1 + bT \zeta_{a,b,c}(h_t) \text{Lip}(R).
\]

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 \(\tilde{F}(U) = 0\), but also establishes exponential convergence for \(\|\tilde{F}(U_t)\|\).

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

\[
bT \text{Lip}(R) \zeta_{a,b,c}(h_t) < 1. \tag{3.12}
\]

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

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

Proof. The unique existence of the root for \(\tilde{F}(\cdot)\) is due to Lemma 15. We now prove the exponential convergence (3.13). According to Lemma 6, by letting \(\theta = bT \text{Lip}(R) \zeta_{a,b,c}(h_t)\), we obtain

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

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

\[
\|\tilde{F}(U_t)\| \leq \exp \left( -\frac{1}{2} (3 - \frac{1}{2\kappa}) \min\{\sigma^2, 1\} t \right) \sqrt{\|\tilde{F}(U_0)\|^2 + \sigma^2\|Q_0\|^2}
\leq \exp \left( -\frac{1}{2} \cdot \frac{5}{2} \frac{(1 - \theta)^3}{8(1 + \theta)} t \right) \sqrt{\|\tilde{F}(U_0)\|^2 + (1 + \theta)\|Q_0\|^2}
= \exp \left( -\frac{5}{32} \frac{(1 - \theta)^3}{1 + \theta} t \right) \sqrt{\|\tilde{F}(U_0)\|^2 + (1 + \theta)\|Q_0\|^2},
\]

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

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

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

\[
\|\widehat{F}(U_t)\| \leq \exp \left( -\frac{5}{32} \cdot \frac{(1 - \theta)^3}{1 + \theta} \right) \sqrt{\|\widehat{F}(U_0)\|^2 + (1 + \theta)\|Q_0\|^2}. \tag{3.15}
\]

**Proof.** Recall that we have \(\theta = bT\text{Lip}(R)\zeta_{a,b,c}(h_t)\). We prove \(\tilde{\theta} \geq \theta\) under both cases.

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

\[
\theta = bT\text{Lip}(R)\zeta_{a,b,c}(h_t) \leq bT\text{Lip}(R) = \tilde{\theta}.
\]

• (Cahn-Hilliard type) We have

\[
\zeta_{a,b,c}(h_t) = \max_k \left\{ \left( \frac{1}{\lambda_k(\mathcal{G}_h)} + h_t a \lambda_k(\mathcal{L}_h) + h_t bc \right)^{-1} \right\} = \max_k \left\{ \left( \frac{1}{\lambda_k(\mathcal{L}_h)} + h_t a \lambda_k(\mathcal{L}_h) + h_t bc \right)^{-1} \right\} \leq \frac{1}{2\sqrt{ah_t + bch_t}}.
\]

Then,

\[
\theta = bT\text{Lip}(R)\zeta_{a,b,c}(h_t) \leq \frac{bT\text{Lip}(R)}{2\sqrt{ah_t + bch_t}} = \tilde{\theta}.
\]

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

\[\square\]

3.2.3. **Numerical verification.** We apply our algorithm to solve the Allen-Cahn equation (4.1) with \(\epsilon_0 = 0.01\) on a \(64 \times 64\) grid. We use \(\tau_U = \tau_T = 0.5\), \(\omega = 1\), \(\epsilon = 0.1\). At each iteration \(k\), denote \(U_k\) as the numerical solution. We define \(\tau_k = -\log_{10}(\|\widehat{F}(U_{k+1})\|/\|\widehat{F}(U_k)\|)\) to be the convergence rate of the residual term \(\|\widehat{F}(U_k)\|\) at \(k\)th 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 \(\tilde{\theta}\) is small, the convergence rate is \(\frac{5}{32} (1 - 4\tilde{\theta} + \mathcal{O}(\tilde{\theta}^2))\), which is linear w.r.t. \(N_t h_t\) (recall that \(\tilde{\theta} \propto T = N_t h_t\)). 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 \(\tilde{\theta} \propto N_t h_t\) keeps increasing. Furthermore, we verify the dependence of the convergence rate on \(N_t h_t\) via the left plot of Figure 2.

We also apply our algorithm to the Cahn-Hilliard equation (4.2) with \(\epsilon_0 = 0.1\) on a \(64 \times 64\) 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 \(\tilde{\theta} \propto N_t \sqrt{h_t}\) is small. This is reflected in Figure 2. Unlike the case of Allen-Cahn, in which the PDHG algorithm converges as \(\tilde{\theta}\) increases, the iterations for Cahn-Hilliard diverges as \(\tilde{\theta} \propto N_t \sqrt{h_t}\) increases. This is reflected on the right plot of Figure 2.

For a fixed time step size \(h_t\), denote by \(N_{\text{max}}\) the maximum number of time steps that guarantees the convergence of the PDHG algorithm. We plot the relation between \(N_{\text{max}}\) and \(h_t\) on a logarithmic scale in Figure 3. We observe the relation \(N_{\text{max}} = \mathcal{O}(\frac{1}{\sqrt{h_t}})\) 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.
3.3. Lyapunov analysis for the time-discrete case. In this section, we discuss the convergence of the time-discrete PDHG algorithm \((2.15)\). Recall that the equilibrium state of the PDHG dynamic \((2.15)\) 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 \((2.15)\)). 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 \(b\text{Lip}(R)\zeta_{a,b,c}(h_t) < \sqrt{2} - 1\). Suppose we pick \(\theta \geq b\text{Lip}(R)\zeta_{a,b,c}(h_t)\) with \(\theta < \sqrt{2} - 1\).
- (On PDHG algorithm \((2.15)\)) Suppose \((D)\) holds. There exist \(\bar{\gamma} = \omega\tau_P, \varrho = \frac{\tau_P}{\epsilon}, \epsilon > 0\) satisfying

\[
\varrho \bar{\gamma} \epsilon \psi(\theta) - \frac{1}{4} \Omega(\bar{\gamma} \epsilon, \varrho, \theta)^2 > 0.
\]  

\((3.16)\)
Figure 3. $N_{\text{max}} - h_t$ 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 $5 \leq k \leq 11$.

Figure 4. Plots of $\bar{r}$ vs $(N_t, h_t)$. (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 $N_t \sqrt{h_t} = 0.15, 0.05, 0.01, 0.001$.

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

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

and set the extrapolation coefficient $\omega = \frac{\gamma}{\tau_P}$, the PDHG step size for $U$ as $\tau_U = \frac{\tau_P}{\varrho}$. 
Under the above conditions, there exists a unique $U_*$ s.t. $\hat{F}(U_*) = 0$. Furthermore, assume that $\{U_k, Q_k\}$ solves the PDHG algorithm (2.15) with arbitrary initial condition $(U_0, Q_0)$. Write $J_k = J(U_k, Q_k)$. We have

$$J_k \leq \left(\frac{2}{\Phi + \sqrt{\Phi^2 + 4}}\right)^{k+1} \left(J_1 + \frac{\Phi + \sqrt{\Phi^2 + 4}}{2} J_0\right),$$

where

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

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 $b\text{Lip}(R)T < \sqrt{2} - 1$ (resp., $\frac{b\text{Lip}(R)T}{2\sqrt{\alpha h_t + \beta c_h}} < \sqrt{2} - 1$). If we set $\theta = b\text{Lip}(R)T$ (resp., $\theta = \frac{b\text{Lip}(R)T}{2\sqrt{\alpha h_t + \beta c_h}}$), then we have $bT\zeta_{a,b,c}(h_t)\text{Lip}(R) < \theta < \sqrt{2} - 1$.

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

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

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

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

Now the value of $\tau_P$ is determined by (3.17), $\tau_U = \frac{\tau_P}{\rho^2}, \omega = \frac{\omega}{\tau_P}$ can also be determined. In summary, we have the following Corollary.

**Corollary 8.1** ($N_c$-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, $G_h = I, L_h$ is self-adjoint, non-negative definite) Pick $T < \frac{\sqrt{2} - 1}{b\text{Lip}(R)}$, or equivalently, $h_t < \frac{\sqrt{2} - 1}{b\text{Lip}(R) h_t}$. We denote $\theta = b\text{Lip}(R)T < \sqrt{2} - 1$;
- (Cahn-Hilliard type, $G_h = L_h$ is self-adjoint, and non-negative definite)

Pick $T < \frac{(\sqrt{2} - 1)(2\sqrt{\alpha h_t + \beta c_h})}{b\text{Lip}(R)}$, or equivalently, $h_t < \frac{(\sqrt{2} - 1)2a}{b(\text{Lip}(R) - \sqrt{(2 - 1)\epsilon})}$. We denote $\theta = \frac{b\text{Lip}(R)T}{2\sqrt{\alpha h_t + \beta c_h}} = \frac{b\text{Lip}(R)N_t\sqrt{R}}{2\sqrt{\alpha h_t + \beta c_h}} < \sqrt{2} - 1$.

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

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

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
\[ C_0 = \left( J_1 + \Phi + \sqrt{\Phi^2 + 4} J_0 \right), \quad \Phi = \frac{(1 - 2\theta)^2}{8(1 + \theta)^2} \cdot \frac{\left( 1 - \frac{\vartheta^2}{1 - 2\theta} \cdot \frac{1}{u} \right)^2}{\max \{ (1 + \theta)^2, (1 - u)/u \}}. \]

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 \( R(u) = f(u) - cu = u^3 - 3u \). Thus we use \( \sup_{u \in [-1, 1]} |R'(u)| = 3 \) as the value of Lip(\( R \)) in Corollary 8.1 during the calculation.

<table>
<thead>
<tr>
<th>AC(4.1)</th>
<th>h_t</th>
<th>N_t</th>
<th>u</th>
<th>( \tau_P )</th>
<th>( \tau_U )</th>
<th>( \omega )</th>
<th>( \epsilon )</th>
<th>( \bar{\vartheta} )</th>
<th>Actual rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \epsilon_0 = 1.0 )</td>
<td>0.005</td>
<td>20</td>
<td>0.5</td>
<td>0.0498</td>
<td>0.0996</td>
<td>5.0181</td>
<td>1.0</td>
<td>0.3000</td>
<td>0.0112</td>
</tr>
<tr>
<td>( \epsilon_0 = 0.1 )</td>
<td>0.1381</td>
<td>27</td>
<td>( u \in (0.2250, 1) )</td>
<td>0.0374</td>
<td>0.1147</td>
<td>4.3587</td>
<td>1.0</td>
<td>0.2100</td>
<td>0.0141</td>
</tr>
<tr>
<td>( \epsilon_0 = 0.01 )</td>
<td>0.0013</td>
<td>13</td>
<td>( u \in (0.0760, 1) )</td>
<td>0.0936</td>
<td>0.1872</td>
<td>2.6702</td>
<td>1.0</td>
<td>0.1500</td>
<td>0.0307</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CH(4.2)</th>
<th>h_t</th>
<th>( \epsilon_0 )</th>
<th>N_t</th>
<th>u</th>
<th>( \tau_U )</th>
<th>( \omega )</th>
<th>( \epsilon )</th>
<th>( \bar{\vartheta} )</th>
<th>Actual rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \epsilon_0 = 10 )</td>
<td>0.005</td>
<td>10</td>
<td>0.5</td>
<td>0.0842</td>
<td>0.1684</td>
<td>2.9695</td>
<td>1.0</td>
<td>0.1640</td>
<td>0.0260</td>
</tr>
<tr>
<td>( \epsilon_0 = 0.1 )</td>
<td>0.1455</td>
<td>12</td>
<td>( u \in (0.04, 1) )</td>
<td>0.0475</td>
<td>0.0949</td>
<td>5.2662</td>
<td>1.0</td>
<td>0.2874</td>
<td>0.0103</td>
</tr>
<tr>
<td>( \epsilon_0 = 0.01 )</td>
<td>0.0015</td>
<td>9</td>
<td>( u \in (0.0798, 1) )</td>
<td>0.0286</td>
<td>0.0572</td>
<td>8.7392</td>
<td>1.0</td>
<td>0.2741</td>
<td>0.0043</td>
</tr>
</tbody>
</table>

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 \( \{k, \log(\|U_k - U_*\|_2^2)/\|U_k - U_*\|_2^2)\} \) for \( 1 \leq k \leq 400 \) (Allen-Cahn equation (4.1)); and \( 1 \leq k \leq 500 \) (Cahn-Hilliard equation (4.2)).

![Plot](image)

Figure 5. Plot of \( \log(\|U_k - U_*\|_2^2) \) vs \( k \) (\( 1 \leq k \leq 400 \)) when using hyperparameters specified in Table 1 to solve Allen-Cahn equation (4.1) with different \( \epsilon_0 \) on a 128×128 grid.

Remark 6. \( \epsilon_0 \) may also not be the optimal choice of hyperparameters. We provide suggestions on selecting the optimal hyperparameters in section 4.2.
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 \cite{35}, the preconditioned fixed point method \cite{2} and Newton’s method \cite{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 $\| \text{Res}(U_k) \|_\infty < \text{tol}$ with $\text{tol} = 10^{-6}$. Here the residual term $\text{Res}(U_k)$ is defined in (2.16). All numerical examples are imposed with periodic boundary conditions. We adopt the central discretization scheme to discretize the Laplace operator $\Delta$, i.e., we set the discretized Laplace operator as $\text{Lap}_h^P$ defined in (2.2).

Among four equations discussed in this section, equations (4.1), (4.2), and (4.6) have already been considered in \cite{32}, 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

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

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 $\mathcal{M}$, $\mathcal{G}_h = I$, and $\mathcal{L}_h = \Delta_h^P$, 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 \cite{35}. A comparison among the plots of...
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.

4.1.2. Cahn-Hilliard equation (CH). We consider the Cahn-Hilliard equation
\[
\frac{\partial u}{\partial t} = -a \Delta^2 u + \Delta b W'(u), \quad \text{on } [0, 2\pi]^2 \times [0, T], \quad u(x, 0) = u_0(x).
\] (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 \( \varphi \) is defined as
\[
\varphi(s) = \begin{cases} 
2e^{-\frac{s^2}{\epsilon}} & s < 0; \\
0 & s \geq 0,
\end{cases}
\]
with \( \epsilon = 0.1 \).

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

\[
G_h = \mathcal{L}_h = \Delta^P_{h_x}, \quad \text{and} \quad J_f = 2I.
\]

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|c|c|c|c|}
\hline
\( i \) & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\
\hline
\( x_i \) & \( \pi/2 \) & \( \pi/4 \) & \( \pi/2 \) & \( \pi \) & \( 3\pi/2 \) & \( \pi \) & \( 3\pi/2 \) \\
\( y_i \) & \( \pi/2 \) & \( 3\pi/4 \) & \( 5\pi/4 \) & \( \pi/4 \) & \( \pi/4 \) & \( \pi \) & \( 3\pi/2 \) \\
\( r_i \) & \( \pi/5 \) & \( 2\pi/15 \) & \( \pi/15 \) & \( \pi/10 \) & \( \pi/10 \) & \( \pi/4 \) & \( \pi/4 \) \\
\hline
\end{tabular}
\caption{Centers and radius of the 7 circles.}
\end{table}
4.1.3. A reaction-diffusion equation with variable coefficient (VarCoeff). We consider the following equation with variable coefficient (mobility term) $\sigma(\cdot)$,

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

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

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

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 $\sigma_{pq} = \sigma((p - 1)h_x, (q - 1)h_y)$ for any $p, q \in \mathbb{Q}$.

For the precondition matrix $\mathcal{M}, \mathcal{G}_h = I$, we approximate $\mathcal{L}_h$ by $-\sigma \Delta_h$, whose matrix-vector multiplication and inversion can be efficiently computed via the FFT algorithm. Here $\sigma = \frac{1}{|\Omega|} \int_{\Omega} \sigma(x, y) \, dxdy = 1 + \frac{\epsilon_0}{2}$ denotes the average of $\sigma$ over $\Omega = [0, 2\pi]^2$. We set $J_f = 2I$. We choose $\tau_U = 0.5, \tau_F = 0.95$ 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

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

as the discrete analogy of $E(u)$. The free energy $E_h(U^{t_k})$ 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.

![Figure 8](image1.png)

**Figure 8.** Numerical solution of the time-implicit scheme solved via our PDHG method on a 256 $\times$ 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:

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

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

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

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

$$
\Delta_h (\epsilon_0^2 \Delta_h - W''(\pm 1) + \epsilon_0^2) = \Delta_h (\epsilon_0^2 \Delta_h - 2 + \epsilon_0^2),
$$

\[\text{Figure 10. Comparison between the proposed scheme and the IMEX scheme.}\]
and set $L_h = \epsilon_2^2 \Delta_h$. 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 $\theta$ and $\theta$ 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. **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 $\epsilon$. In the following, we discuss the choice of these hyperparameters.
(1) (Choosing $N_t$) Suppose we want to solve an equation on $[0, T_{\text{total}}]$. We divide the time interval into $M \cdot N_t$ subintervals, i.e.,

$$[0, T_{\text{total}}] = \bigcup_{k=1}^{M} I_k = \bigcup_{k=1}^{M} \left( \bigcup_{j=1}^{N_t} I_{k,j} \right),$$

where each $I_{k,j} = [(k-1)T + (j-1)h_t, (k-1)T + jh_t]$.

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

---

**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 $256 \times 256$ 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 $U_{\text{IMEX}}$ and the benchmark solution $U_*$. Red curve indicates the $L^1$ discrepancy between the time-implicit solution $U_{\text{PDHG}}$ and the benchmark $U_*$. (Right) Plot of $|U_{\text{PDHG}} - U_*|$ (up); and plot of $|U_{\text{IMEX}} - U_*|$ (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 $|U_{\text{PDHG}} - U_*|$ (up); and $|U_{\text{IMEX}} - U_*|$ (down).
The intuition of choosing $\varrho > 1$ is that we want to treat the inner optimization of the functional $\hat{L}(U, Q)$ defined in \eqref{eq:inner} 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 \cite{liu2015high} and theoretically \cite{sahu2007error, liu2015high}. A rather efficient choice of the step sizes $(\tau_U, \tau_P)$ is $(0.5, 0.9)$. This is verified in Table \ref{table:cpu_time} in which we compare the choice $(0.5, 0.9)$ with other combinations of $(\tau_U, \tau_P)$.

Table 3. Comparison of CPU time (s) with different $N_t$s (All problems are solved on $256 \times 256$ grids).

![Iterations needed for convergence vs N_x](image)

**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.

We then apply our proposed method to each subinterval $I_k$ in order to obtain the entire numerical solution on $[0, T_{total}]$. 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 $\|\text{Res}(U_k)\|_\infty < 10^{-6}$. The efficiency of our algorithm under different scenarios is reflected in CPU time demonstrated in Table \ref{table:cpu_time}.
algorithm with tolerance $\epsilon$ (4) (Choosing $\epsilon$) We set $\epsilon$ around 0.1. Recall that $\sup_{Q} \{\bar{L}(U, Q)\} = \frac{\|F(U)\|^2}{2\epsilon}$. Increasing $\epsilon$ will decrease the convexity of the functional $\frac{\|F(U)\|^2}{2\epsilon}$, which will slow down our algorithm. Decreasing $\epsilon$ brings our algorithm closer to our original version of PDHG method [32], in which we discover stronger oscillations towards convergence, which may also affect the efficiency.

4.3. 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 $h_t < h_t$, 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 $h_t$, we reset $h_t = h_t$.

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 [14] 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 [23] 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 $\eta = 10^{-10}$ to solve the linear system at each IMEX step. For (4.3), the IMEX method only works stably for a rather small time step size $h_t \leq 0.5 \cdot 10^{-5}$. As reflected in

![Figure 14](image)

**Figure 14.** (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 $t$.

Table 5, our method works better on long-time computation.

---

1Suppose 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 $\|Ax_k - b\|_\infty \leq \eta$. 

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

<table>
<thead>
<tr>
<th>$\epsilon = 0.1$ for all problems</th>
<th>$\tau_P = 0.9, \tau_P = 0.5$</th>
<th>$\tau_U = 0.65, \tau_P = 0.65$</th>
<th>$\tau_U = 0.5, \tau_P = 0.9$</th>
</tr>
</thead>
<tbody>
<tr>
<td>6th Order $[T = 0.5]$</td>
<td>$N_x = 256, N_t = 50$</td>
<td>62.28</td>
<td>47.92</td>
</tr>
<tr>
<td></td>
<td>$N_x = 128, N_t = 50$</td>
<td>12.31</td>
<td>9.47</td>
</tr>
<tr>
<td>VarCoeff $[T = 0.5]$</td>
<td>$N_x = 256, N_t = 50$</td>
<td>103.23</td>
<td>109.38</td>
</tr>
<tr>
<td></td>
<td>$N_x = 128, N_t = 50$</td>
<td>15.92</td>
<td>13.35</td>
</tr>
</tbody>
</table>
Our method | IMEX
---|---
ht = 0.5 \cdot 10^{-3} & ht = 0.2 \cdot 10^{-3} & ht = 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].

4.4. **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 [35]. 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 15.

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

$$(I + ah_tG_hL_h)U^{t+1} = U^t - bh_tG_hf(U^{t+1}).$$

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

$$U_{k+1} = (I + h_tG_h(aL_h + bcI))^{-1}(U^t - bh_tG_h(f(U_k) - cU_k)), \quad \text{with initial guess } U_0 = U^t.$$  

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 16.

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...
Figure 16. 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 $256 \times 256$ grid. These quantile plots are composed based on 40 independent runs of both algorithms.

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 17. 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.

Figure 17. 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 $256 \times 256$ 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 [32] 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 $N$. 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 three 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?

### References


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

\[ \frac{U - U^0}{h_t} = -G_h(a \mathcal{L}_h U + bf(U)), \]  

(A.1)

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

\[ \frac{\xi}{h_t} + G_h(a \mathcal{L}_h (U^0 + \xi) + bf(U^0 + \xi)) = 0. \]  

(A.2)

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

\[
\min_{\xi \in \text{Ran}(G_h)} \left\{ \frac{\xi^\top G_h^\dagger \xi}{2h_t} + \frac{a}{2} (U^0 + \xi)^\top \mathcal{L}_h (U^0 + \xi) + b W(U^0 + \xi)^\top 1 \right\}.
\]

(A.3)

Here we denote \( W(\cdot) \) as the primitive function of \( f(\cdot) \). Let us define \( \mathcal{V} = \text{Ran}(G_h) \) 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 \( \xi \) 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

\[ \frac{G_h^\dagger \xi}{h_t} + a \Pi \mathcal{V} \mathcal{L}_h (U^0 + \xi) + b \Pi \mathcal{V} f(U + \xi) = 0. \]

Writing the projection \( \Pi \mathcal{V} = G_h^\dagger G_h \), we obtain

\[ G_h \left( \frac{\xi}{h_t} + G_h \mathcal{L}_h (U^0 + \xi) + b \mathcal{G}_h f(U + \xi) \right) = 0. \]

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

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 \( G_h \), and \( x \in \mathbb{R}^r \), (A.3) is equivalent to the following non-constrained optimization problem

\[
\min_{x \in \mathbb{R}^r} \left\{ \frac{x^\top \Lambda^{-1} x}{2h_t} + \frac{a}{2} x^\top Q_1^\top \mathcal{L}_h Q_1 x + a U^0^\top \mathcal{L}_h Q_1 x + b W(U^0 + Q_1 x)^\top 1 \right\}.
\]

(A.4)

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

\[
\nabla \mathcal{J}(x) = \frac{\Lambda^{-1}}{h_t} x + a Q_1^\top \mathcal{L}_h Q_1 x + a Q_1^\top \mathcal{L}_h U^0 + b Q_1^\top (V'(U^0 + Q_1 x) + \phi(U^0 + Q_1 x)).
\]

Then

\[
(x - y, \nabla \mathcal{J}(x) - \nabla \mathcal{J}(y)) = \frac{1}{h_t} (x - y)^\top \Lambda (x - y) + a (x - y)^\top Q_1^\top \mathcal{L}_h Q_1 (x - y) + b (Q_1 (x - y))^\top (V'(U^0 + Q_1 x) - V'(U^0 + Q_1 y)) + b (Q_1 (x - y))^\top (\phi(U^0 + Q_1 x) - \phi(U^0 + Q_1 y))
\]

\[
\geq (x - y)^\top \left( \frac{\Lambda}{h_t} + a Q_1^\top \mathcal{L}_h Q_1 \right) (x - y) + b K\|x - y\|^2 - b \text{Lip}(\phi)\|x - y\|^2
\]

\[
\geq \lambda_{\min} \left( \frac{\Lambda}{h_t} + a Q_1^\top \mathcal{L}_h Q_1 \right) + b K - b \text{Lip}(\phi) \|x - y\|^2.
\]
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, \( \mathcal{G} = \text{Id}, \mathcal{L} = -\Delta \).
  
  Then the condition (3.3) leads to
  \[
  \alpha = \lambda_{\min} \left( \frac{\Lambda^{-1}}{h_t} + aQ_1^T \mathcal{L}_h Q_1 \right) + bK - b \text{Lip}(\phi) > 0.
  \]
  This shows the \( \alpha \)-strongly convexity of \( \widetilde{J} \), which leads to the existence and uniqueness of the minimizer to (A.3), which accomplishes the proof.

- (Cahn-Hilliard equation with periodic boundary condition) In this case, \( \mathcal{G}_h = I_{N_x^2}, \mathcal{L}_h = -\Delta_{h_x}^P = I_{N_x} \otimes (-\text{Lap}_{h_x}^P) + (-\text{Lap}_{h_x}^P) \otimes I_{N_x} \), where \( \text{Lap}_{h_x}^P \) is defined in (2.2). Then
  \[
  \lambda_k^P = \frac{4}{h_x^2} \sin^2 \left( \frac{\pi k}{N_x} \right), \quad \text{with } 1 \leq k \leq N_x,
  \]
  are the eigenvalues of \( -\text{Lap}_{h_x}^P \). And the eigenvalues of \( \frac{\Lambda^{-1}}{h_t} + aQ_1^T \mathcal{L}_h Q_1 = \frac{I}{h_t} + a\mathcal{L}_h \) are \( \lambda_{k,l} = \frac{1}{h_t} + a(\lambda_k^P + \lambda_l^P) \), with \( 1 \leq k, l \leq N_x \). Thus, \( \lambda_{\min}(\frac{\Lambda^{-1}}{h_t} + aQ_1^T \mathcal{L}_h Q_1) = \frac{1}{h_t} \).
  Furthermore, we can decompose \( f(x) = V'(x) + \phi(x) \), where
  \[
  V(x) = \begin{cases} \frac{1}{4}(x^2 - 1)^2, & |x| > 1; \\ 0, & |x| \leq 1. \end{cases}
  \]
  \[
  \phi(x) = \begin{cases} 0, & |x| > 1; \\ x^3 - x, & |x| \leq 1. \end{cases}
  \]

  Then one can verify that \( K = 0 \) and \( \text{Lip}(\phi) = 2 \). In this case, condition (3.3) implies
  \[
  h_t < \frac{1}{\text{Lip}(\phi)b} = \frac{1}{2b}.
  \]

- (Cahn-Hilliard equation with periodic boundary condition) In this case, \( \mathcal{G} = -\Delta, \mathcal{L} = -\Delta \).
  
  Then one can verify that \( K = 0 \) and \( \text{Lip}(\phi) = 2 \). In this case, condition (3.3) implies
  \[
  h_t < \frac{1}{\text{Lip}(\phi)b} = \frac{1}{2b}.
  \]

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

Lemma 9. Suppose \( \overline{\lambda} \geq \lambda > 0 \). Assume \( \mu > 0 \) satisfies \( \frac{1}{\sqrt{\lambda}} - \frac{1}{\sqrt{\lambda}} < \frac{2}{\sqrt{\mu}} \). Define
  \[
  A = \max \left\{ \left(1 - \frac{\sqrt{\mu}}{\sqrt{\lambda}}\right)^2, \left(1 - \frac{\sqrt{\mu}}{\sqrt{\lambda}}\right)^2 \right\}, \quad \text{and } B = \left(1 + \frac{\sqrt{\mu}}{\sqrt{\lambda}}\right)^2,
  \]

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

\[ B_\lambda = \begin{bmatrix} -\frac{1}{2}(\mu - (1 - \gamma \epsilon)\lambda) & -\frac{1}{2}(\mu - (1 - \mu)\lambda) \\ \mu \epsilon & 0 \end{bmatrix} \]  

is always positive definite.

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

\[ \max \left\{ \left| 1 - \frac{\sqrt{\mu}}{\sqrt{\lambda}} \right|, \left| 1 - \frac{\sqrt{\mu}}{\sqrt{\lambda}} \right| \right\} < 1 + \sqrt{\frac{\mu}{\lambda}}. \]

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 $B_\lambda$ 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

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

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

Now, for fixed $\lambda \in [\underline{\lambda}, \bar{\lambda}]$, 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

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

and

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

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

\[ \square \]

**Lemma 10** (Positive definiteness of $H_\mu$). Consider the matrix $H_\mu$,

\[ H_\mu = \begin{bmatrix} \gamma \Sigma & \frac{1}{2}(\mu I - (1 - \gamma \epsilon)\Sigma) \\ -\frac{1}{2}(\mu I - (1 - \gamma \epsilon)\Sigma) & \mu \epsilon \end{bmatrix}, \]

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

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

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

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

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

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

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

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

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

Thus $\lambda_2(B_\lambda) = \varphi_{\mu,\gamma,\epsilon}(\lambda)$. Since $B_\lambda$ 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 $B_\lambda > \beta I$ for any $\lambda \in [\underline{\lambda}, \overline{\lambda}]$.

To estimate $H_\mu$ 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 $\Sigma$. Since $H_\mu$ is symmetric, $H_\mu$ 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 $B_{\lambda_j}$ defined as in (B.1). Since each $\lambda_j \in \lambda(\Sigma) \subset [\underline{\lambda}, \overline{\lambda}]$, the above argument applies to every $B_{\lambda_j}$, i.e., $B_{\lambda_j} > \beta I$ for any $1 \leq j \leq N$. This leads to $H_\mu > \beta I$. \qed

We are ready to prove Lemma 2.

Proof of Lemma 2. We denote

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

and compute

$$\frac{d}{dt} \mathcal{I}_\mu(U_t, Q_t) = \hat{F}(U)^T D\hat{F}(U) \hat{U} + \mu Q^T Q
\begin{align*}
= & \hat{F}(U)^T D\hat{F}(U) \hat{F}(U) \hat{F}(U)^T (Q + \gamma Q) + \mu Q^T (-\epsilon Q + \hat{F}(U)) \\
= & \hat{F}(U)^T \Sigma(Q + \gamma(-\epsilon Q + \hat{F}(U))) - \mu\|Q\|^2 + \mu Q^T \hat{F}(U) \\
= & \hat{F}(U)^T (\mu I - (1 - \gamma)\Sigma)Q - \gamma \hat{F}(U)^T \Sigma \hat{F}(U) - \mu\|Q\|^2 \\
= & [\hat{F}(U)^T, Q^T] \left[ -\frac{1}{2}(\mu I - (1 - \gamma)\Sigma) \begin{array}{c}
\frac{1}{2}(\mu I - (1 - \gamma)\Sigma) \\
\mu e
\end{array} \right] \\
= & [\hat{F}(U)^T, Q^T] \begin{bmatrix}
\hat{F}(U) \\
Q
\end{bmatrix}
\begin{bmatrix}
\mu I - (1 - \gamma)\Sigma \\
\mu e
\end{bmatrix}. \quad (B.4)
\end{align*}$$

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

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

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

$$\lambda(\Sigma) \subset [\sigma^2, \sigma^2].$$
Now we apply Lemma 10 with $\lambda = \sigma^2$, $\bar{\lambda} = \bar{\sigma}^2$. We prove that $H_\mu \succ \beta I$ for any $U_t \in \mathbb{R}^N$. As a result, we obtain the following inequality:

$$\frac{d}{dt} I_\mu(U_t, Q_t) = -[\bar{F}(U)^	op, Q^	op] H_\mu [\bar{F}(U)^	op, Q^	op] \preceq -\beta(\|\bar{F}(U_t)\|^2 + \|Q_t\|^2).$$

Furthermore, one has

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

which yields

$$\|\bar{F}(U)\|^2 + \|Q\|^2 \geq \frac{2}{\max\{1, \mu\}} I_\mu(U, Q).$$

This finally leads to

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

And the Grönwall’s inequality gives

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

We now prove Theorem 3.

**Proof of Theorem 3.** Let us pick the hyperparameter $\mu = \sigma^2$, one can verify that $\mu$ 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 $[\sigma^2, \bar{\sigma}^2]$. For $z \in [\sigma^2, \bar{\sigma}^2]$, we further calculate

$$\varphi_{\mu, \gamma, \epsilon}(z) = \frac{1}{2}(\gamma z + \mu \epsilon - \sqrt{(\gamma z + \mu \epsilon)^2 - (4\gamma_\epsilon \mu z - (\mu - (1 - \gamma_\epsilon)z)^2)})$$

$$= \frac{1}{4\gamma_\epsilon \mu z - (\mu - (1 - \gamma_\epsilon)z)^2}$$

$$\frac{2\gamma z + \mu \epsilon + \sqrt{(\gamma z + \mu \epsilon)^2 - (4\gamma_\epsilon \mu z - (\mu - (1 - \gamma_\epsilon)z)^2)}}{\gamma z + \mu \epsilon}$$

$$\geq \frac{4\gamma_\epsilon \mu z - (\mu - (1 - \gamma_\epsilon)z)^2}{4(\gamma z + \mu \epsilon)}$$

$$= \frac{-(1 - \gamma_\epsilon)^2 z^2 + 2\mu(1 + \gamma_\epsilon)z - \mu^2}{4(\gamma z + \mu \epsilon)}$$

$$= \frac{-(1 + \gamma_\epsilon)(z - \mu)^2 + 4\gamma_\epsilon z^2}{4(\gamma z + \mu \epsilon)}$$

$$= \frac{(2\sqrt{\gamma_\epsilon}z - (1 + \gamma_\epsilon)z + \mu)(2\sqrt{\gamma_\epsilon}z + (1 + \gamma_\epsilon)z - \mu)}{4(\gamma z + \mu \epsilon)}$$

$$\geq \frac{1 - \sqrt{\gamma_\epsilon}z, z \leq \sigma^2}{4(\gamma \sigma^2 + \mu \epsilon)}\left((\sqrt{\mu} - |1 - \sqrt{\gamma_\epsilon}|\sqrt{z})(\sqrt{\mu} + |1 - \sqrt{\gamma_\epsilon}|\sqrt{z})(1 + \sqrt{\gamma_\epsilon})|z - \mu\right)$$

$$\geq \frac{4(\gamma z + \mu \epsilon)}{4(\gamma \sigma^2 + \mu \epsilon)}.$$ (B.5)

Since we have set

$$\gamma = \frac{1 - \delta}{\kappa}, \epsilon = (1 - \delta)\kappa, \mu = \sigma^2.$$
Substituting them into (3.5) yields
\[
\varphi_{\mu, \gamma, \epsilon}(z) \geq \frac{(a - \beta\|z\|)(\beta\|z\| + a)(2 - \beta)^2 z - a^2)}{8(1 - \beta) a^2} \\
= \frac{1}{8(1 - \beta)} \left(1 - |\beta\| z^2 \right) \left(|\beta\| z^2 + \frac{a^2}{\beta}\right) \left((2 - \beta)^2 z - a^2\right) \\
\geq \frac{1}{8(1 - \beta)} \left(1 - \kappa |\beta|\right) \left(|\beta| + 1\right) \left(1 - \beta\right)(3 - \beta)a^2 \\
\geq \frac{1}{8\kappa} \left(1 - \kappa |\beta|\right)(3 - \beta)a^2.
\]

If we denote \( \beta = \min_{z \in \mathbb{R}^n} \{\varphi_{\mu, \gamma, \epsilon}(z)\} \), then we have
\[
\frac{\beta}{\max\{1, \mu\}} \geq \frac{(1 - \kappa |\beta|)(3 - \beta)}{8\kappa} \frac{\sigma^2}{\max\{1, \sigma^2\}} = \frac{1}{8} (1 - \kappa |\beta|)(3 - \beta) \frac{\min\{\sigma^2, 1\}}{\kappa}.
\]
Thus, the result of Theorem 2 yields
\[
\mathcal{I}_\mu(U_t, Q_t) \leq \exp\left(-\frac{1}{4} (1 - \kappa |\beta|)(3 - \beta) \frac{\min\{\sigma^2, 1\}}{\kappa} t\right) \mathcal{I}_\mu(U_0, Q_0).
\]
Taking square root on both sides of the above inequality and using the fact that
\[
\|\hat{F}(U_t)\| \leq \sqrt{\mathcal{I}_\mu(U_t, Q_t)},
\]
we obtain
\[
\|\hat{F}(U_t)\| \leq \exp\left(-\frac{1}{8} (1 - \kappa |\beta|)(3 - \beta) \frac{\min\{\sigma^2, 1\}}{\kappa} t\right) \sqrt{\mathcal{I}_\mu(U_0, Q_0)}.
\]
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.,
\[
\mathcal{I}_\mu(U_t, Q_t) \leq \exp(-2\lambda t) \mathcal{I}_\mu(U_0, Q_0),
\]
where
\[
\lambda = \min\{\epsilon - \frac{1}{2}(1 - \gamma \epsilon)\sigma^2_1/\mu - 1, \epsilon - \frac{1}{2}(1 - \gamma \epsilon)\sigma^2_n/\mu - 1\},
\]
\[
\gamma = \frac{\frac{1}{2}(1 - \gamma \epsilon)\sigma^2_1 - \epsilon - \frac{1}{2}(1 - \gamma \epsilon)\sigma^2_n - \mu}{\frac{1}{2}(1 - \gamma \epsilon)\sigma^2_1 - \mu}.
\]
In particular, when \(\gamma \epsilon = 1, \mu = 0\), and
\[
\gamma = \frac{1}{2} \frac{\sigma^2_2 - \sigma^2_n}{\sigma^2_1 + \sigma^2_n} + \sqrt{\frac{1}{4} \left(\frac{\sigma^2_2 - \sigma^2_n}{\sigma^2_1 + \sigma^2_n}\right)^2 + 4\sigma^2_n},
\]
we have \(\lambda = 2\sigma^2_n \sigma^2_1 + \sigma^2_n^2 - \frac{1}{2} \sigma^2_n \left(\frac{\sigma^2_2 + \sigma^2_n}{\sigma^2_1 + \sigma^2_n}\right)^3 + O(\sigma^6_n)\).

**Proof of Theorem 11.** We would like to find \(\lambda\) such that
\[
\frac{d}{dt} \mathcal{I} + 2\lambda \mathcal{I} \leq 0.
\]
Then by Gronwall's inequality, we obtain exponential convergence. We have
\[
\frac{d}{dt} \mathcal{I} + 2\lambda \mathcal{I} = [\hat{F}(U)^\top, Q^\top] \left[ \begin{array}{c} \lambda \mathcal{I} - \gamma \Sigma \\
\frac{1}{2}(\mu \mathcal{I} - (1 - \gamma \epsilon) \Sigma) \\
\frac{1}{2}(\mu \mathcal{I} - (1 - \gamma \epsilon) \Sigma)
\end{array} \right] \left[ \begin{array}{c} \hat{F}(U) \\
Q \\
\end{array} \right].
\]
Using Lemma A.1 from [52], 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 \( \sigma^2 = \sigma_1^2 \geq \sigma_2^2 \geq \cdots \geq \sigma_n^2 = \sigma^2 \). 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
\[
\lambda \leq \min_{i=1,2} \min_{\sigma_i \leq \sigma \leq \sigma_1} g_i(\sigma).
\]
Since \( g_i(\sigma) \)'s are piece-wise linear and have only one kink, it is easy to check that
\[
\min_{\sigma_i \leq \sigma \leq \sigma_1} g_i(\sigma) = \min \{ g_i(\sigma_1), g_i(\sigma_n) \}.
\]
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 \( \lambda \) large, we would like to equate \( g_1(\sigma_n) \) and \( g_2(\sigma_n) \). This yields
\[
\epsilon = \frac{\gamma \sigma_n^2 + \frac{1}{2} \sigma_2^2 - \sigma_n^2 - \frac{1}{4} (\sigma_1^2 - \sigma_n^2)}{1 - \frac{1}{4} \gamma (\sigma_1^2 - \sigma_n^2)}.
\]
In the special case of \( \gamma \epsilon = 1 \), we obtain
\[
1 = \gamma \epsilon = \frac{\gamma \sigma_n^2 + \frac{1}{2} \sigma_2^2 - \sigma_n^2 - \frac{1}{4} \gamma (\sigma_1^2 - \sigma_n^2)}{1 - \frac{1}{4} \gamma (\sigma_1^2 - \sigma_n^2)}.
\]
We can solve for \( \gamma \) and we get (keeping the positive root)
\[
\gamma = \frac{-\frac{1}{2} \sigma_2^2 - \sigma_n^2 + \sqrt{\frac{1}{4} \left( \frac{\sigma_2^2 - \sigma_n^2}{\sigma_1^2 + \sigma_n^2} \right)^2 + 4 \sigma_n^2}}{2 \sigma_n^2}.
\]
Consequently, the convergence rate is
\[
\lambda = \gamma \sigma_n^2 = -\frac{1}{4} \frac{\sigma_2^2 - \sigma_n^2}{\sigma_1^2 + \sigma_n^2} + \frac{1}{2} \sqrt{\frac{1}{4} \left( \frac{\sigma_2^2 - \sigma_n^2}{\sigma_1^2 + \sigma_n^2} \right)^2 + 4 \sigma_n^2} = 2 \sigma_n^2 \frac{{\sigma_2^2 + \sigma_n^2}}{\sigma_1^2 - \sigma_n^2} - \frac{1}{2} \sigma_n^4 \left( \frac{\sigma_2^2 + \sigma_n^2}{\sigma_1^2 - \sigma_n^2} \right)^3 + \mathcal{O}(\sigma_n^6).
\]
\[
\Box
\]
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 an $n \times m$ matrix with $\sigma_{\min}(A_k) \geq \alpha > 0$ and $\sigma_{\max}(A_k) \leq \overline{\alpha}$, i.e., $\|A_k v\| \geq \alpha \|v\|$, $\|A_k v\| \leq \overline{\alpha} \|v\|$ for any $v \in \mathbb{R}^m$. Then $\|A^{-1}\| \leq \sum_{k=1}^{n} \alpha^{-k}$, and $\|A\| \leq \overline{\alpha} + 1$, i.e., $\sigma_{\min}(A) \geq \frac{1}{\sum_{k=1}^{n} \alpha^{-k}}$, and $\sigma_{\max}(A) \leq \overline{\alpha} + 1$.

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

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

Thus we can write $A^{-1}$ as

$$A^{-1} = \begin{bmatrix}
A_{11} & \cdots & O \\
A_{22} & \cdots & 0 \\
\vdots & \ddots & \vdots \\
O & \cdots & A_{nn}
\end{bmatrix} + \begin{bmatrix}
O & \cdots & 0 \\
0 & \cdots & A_{n,n-1} \\
\vdots & \ddots & \vdots \\
0 & \cdots & 0
\end{bmatrix} + \begin{bmatrix}
0 & \cdots & 0 \\
0 & \cdots & O \\
\vdots & \ddots & \vdots \\
O & \cdots & O
\end{bmatrix}$$

de note as $J_1 + J_2 + \cdots + J_n$.

Here, each $J_k$ ($1 \leq k \leq n$) is an $nm \times nm$ block-(sub)diagonal matrix whose $k$-th subdiagonal is $\text{diag}(A_{k,1}, A_{k+1,2}, \ldots, A_{n,k+1})$.

And each $A_{ij}$ is defined as

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

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

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

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

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

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

On the other hand, we write $A$ as

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

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

$$J = \begin{bmatrix}
0 & \cdots & 0 \\
1 & \cdots & 0 \\
\vdots & \ddots & \vdots \\
0 & \cdots & 1
\end{bmatrix}, \quad \text{(B.10)}$$

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

$$\|A\| \leq \|\text{diag}(A_1, \ldots, A_n)\| + \|J \otimes I\| \leq \overline{\alpha} + 1.$$ 

\qed
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 $\Lambda_G, \Lambda_L$ are the diagonal matrices equivalent to $G, L$. Furthermore, $\sigma_{\min}(I + GL) = \sigma_{\min}(I + LG) \geq 1 + \lambda_{\min}(G)\lambda_{\min}(L) \geq 1$.

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

We now prove Lemma 4.

Proof of Lemma 4 We first recall

$$\sigma = \inf_{U \in \mathbb{R}^{N^2}} \{ \sigma_{\min}(DF(U)) \} = \inf_{U \in \mathbb{R}^{N^2}} \{ \sigma_{\min}(\mathcal{M}^{-1}DF(U)) \},$$

$$\bar{\sigma} = \sup_{U \in \mathbb{R}^{N^2}} \{ \sigma_{\max}(DF(U)) \} = \sup_{U \in \mathbb{R}^{N^2}} \{ \sigma_{\max}(\mathcal{M}^{-1}DF(U)) \},$$

where we denote $F(U) = \mathcal{D} U + h_t \mathcal{G}_h(a \mathcal{L}_h U + b f(U))$.

We have

$$\sigma_{\min}(\mathcal{M}^{-1}DF(U)) = \frac{1}{\sigma_{\max}(DF(U))} \geq \frac{1}{\|DF(U)^{-1}\| \|\mathcal{M}\|^2} = \frac{\sigma_{\min}(DF(U))}{\|\mathcal{M}\|^2}. \tag{B.11}$$

And

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

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

$$DF(U) = \begin{bmatrix} X_1 \\ -I & X_2 \\ -I & X_3 \\ \vdots & \vdots \\ -I & X_{N^2} \end{bmatrix},$$

where each $X_i = I + a h_t \mathcal{G}_h \mathcal{L}_h + b h_t \mathcal{G}_h \text{diag}(f'(U^i))$. (Here we denote $U = (U^1, \ldots, U^{N^2})^T$.)

Then for each $X_i$, we have

$$\sigma_{\min}(X_i) \geq \sigma_{\min}(I + a h_t \mathcal{G}_h \mathcal{L}_h) - \sigma_{\max}(b h_t \mathcal{G}_h \text{diag}(f'(U^i)))$$

$$\geq \sigma_{\min}(I + a h_t \mathcal{G}_h \mathcal{L}_h) - h_t |b| \lambda_{\min}(\mathcal{G}_h) \lambda_{\min}(\mathcal{L}_h).$$

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

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

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

On the other hand, one can also verify that

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

by denoting $\bar{\sigma} = \|I + a h_t \mathcal{G}_h \mathcal{L}_h\| + h_t |b| \|\mathcal{G}_h\| \text{Lip}(f)$, we know $\bar{\sigma}$ is also independent of $U$.

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

$$\sigma_{\min}(\mathcal{F}(U)) \geq \frac{1}{\left(\sum_{k=1}^{N^2} \alpha^{-k}\right) \|\mathcal{M}\|}, \quad \sigma_{\max}(\mathcal{F}(U)) \leq (1 + \bar{\sigma}) \|\mathcal{M}^{-1}\|. \tag{B.13}$$

Since $\alpha, \bar{\sigma}, \|\mathcal{M}\|$ and $\|\mathcal{M}^{-1}\|$ 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 \( \mathcal{G}_h \) be defined as in (2.6), and \( \mathcal{M} \) be defined as in (2.12). Then

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

**Proof of Lemma 14.** Recall that we have

\[
\mathcal{M} = \begin{bmatrix}
X & -I & X \\
-1 & X & -I \\
& \ddots & \ddots & \ddots \\
X^{-(N_t-1)} & X^{-(N_t-2)} & \cdots & X^{-1}
\end{bmatrix}, \quad X = I + ah_t \mathcal{G}_h \mathcal{L}_h + bh_t \mathcal{G}_h J_f.
\]

By Lemma 13, we have \( X = Q(I + ah_t \Lambda \mathcal{G}_h \Lambda \mathcal{L}_h + bh_t \Lambda \mathcal{G}_h) Q^\top \), 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

\[
\| X^{-1} \| \leq \frac{1}{1 + h_t(\lambda_{\min}(a \mathcal{G}_h \mathcal{L}_h + b \mathcal{G}_h))} \leq 1.
\]  

(B.13)

Now one can compute

\[
\mathcal{M}^{-1} \mathcal{G}_h = \begin{bmatrix}
X^{-1} & \cdots & X^{-1} \\
X^{-2} & \cdots & X^{-2} \\
\vdots & \ddots & \ddots \\
X^{-(N_t-1)} & \cdots & X^{-(N_t-1)}
\end{bmatrix} \begin{bmatrix}
\mathcal{G}_h & \mathcal{G}_h & \cdots & \mathcal{G}_h \\
\mathcal{G}_h & \mathcal{G}_h & \cdots & \mathcal{G}_h \\
\vdots & \ddots & \ddots & \vdots \\
\mathcal{G}_h & \cdots & \mathcal{G}_h & \mathcal{G}_h
\end{bmatrix} = \mathcal{N} \tilde{\mathcal{G}}_h.
\]

Similar to the treatment in Lemma 4, we estimate \( \| \mathcal{M} \| \) by decomposing \( \mathcal{N} \) as

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

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

Then we estimate \( \| \mathcal{M} \| \) as

\[
\| \mathcal{M} \| \leq \left( \sum_{k=0}^{N_t-1} \| J^k \otimes (X^{-1})^k \| \right).
\]

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

\[
\| \mathcal{M} \| \leq \sum_{k=0}^{N_t-1} \| (X^{-1})^k \| \leq \sum_{k=0}^{N_t-1} \| X^{-1} \|^k \leq N_t.
\]  

(B.14)
On the other hand, we have
\[ \tilde{G}_h = X^{-1}G_h = Q((I + a_{\ell}A_{\ell}h\Lambda_{h} + b_{\ell}c{\ell}A_{\ell}h)^{-1}A_{\ell}h)Q^T. \]
If we denote \( \{\lambda_k(G_h)\}, \{\lambda_k(L_h)\} \) \((1 \leq k \leq N^2_2)\) as the corresponding eigenvalues of \( G_h, L_h \) w.r.t. \( Q \), we know
\[
\|\tilde{G}_h\| = \max_{1 \leq k \leq N^2_2} \left\{ \frac{\lambda_k(G_h)}{1 + h_t(a\lambda_k(G_h)\lambda_k(L_h) + bc\lambda_k(G_h))} \right\}. \tag{B.15}
\]
Now combining (B.14) and (B.15) and using \( \|A^{-1}g_h\| \leq \|A\|\|\tilde{g}_h\| \), we finish the proof. \( \square \)

We now prove Lemma 6.

**Proof of Lemma 6.** By Lemma 14 and the fact that \( \|DR(\cdot)\| \leq \text{Lip}(R) \), we have
\[
\|D\eta(U)\| = \|bh_t\cdot A^{-1}g_h DR(U)\| \leq bh_t \cdot \|A^{-1}g_h\| \cdot \text{Lip}(R) \leq bT\zeta_{a,b,c}(h_t)\text{Lip}(R).
\]
Recall that
\[ D\tilde{F}(U) = I + D\eta(U). \]

Now for any \( v \in \mathbb{R}^{N^2_2} \), we have
\[
\|D\tilde{F}(U)v\| = \|v + D\eta(U)v\| \geq \|v\| - \|D\eta(U)||v|| \geq (1 - bT\zeta_{a,b,c}(h_t)||v||. \tag{B.16}
\]
Since the right-hand side of (B.16) is independent of \( U \), this will lead to a lower bound on \( \sigma \), i.e.\[
\sigma \geq 1 - bT\zeta_{a,b,c}(h_t)\text{Lip}(R).
\]
By a similar argument, we have
\[
\|D\tilde{F}(U)v\| \leq \|v\| + \|D\eta(U)||v|| \leq (1 + bT\zeta_{a,b,c}(h_t)\text{Lip}(R)||v||.
\]
This will finally lead to \[ \tilde{\sigma} \leq 1 + bT\zeta_{a,b,c}(h_t)\text{Lip}(R). \] \( \square \)

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

**Proof of Lemma 15.** (3.12) leads to
\[
\max_{1 \leq k \leq N^2_2} \left\{ \frac{\lambda_k(G_h)}{1 + h_t(a\lambda_k(G_h)\lambda_k(L_h) + bc\lambda_k(G_h))} \right\} < \frac{1}{bT\text{Lip}(R)},
\]
which is equivalent to
\[
\min_{1 \leq k \leq N^2_2, \lambda_k(G_h) > 0} \left\{ \frac{1}{\lambda_k(G_h)}h_t(a\lambda_k(L_h) + bc) \right\} > bT\text{Lip}(R).
\]
Since \( T \geq h_t \), the right-hand side of the above inequality is larger than or equal to \( bh_t\text{Lip}(R) \). Thus the above inequality yields
\[
\min_{1 \leq k \leq N^2_2, \lambda_k(G_h) > 0} \left\{ \frac{1}{\lambda_k(G_h)}h_t + a\lambda_k(L_h) + bc \right\} > b\text{Lip}(R). \tag{B.17}
\]
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 \( \lambda_k(Q_1^T \varphi_1 \varphi_1^T Q_1) = \lambda_k(L_h) \). As a result, (B.17) is equivalent to (3.3) in Theorem 1 which leads to the unique existence of the root-finding problem \( \tilde{F}(U) = 0 \). \( \square \)
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

\[
ku\Psi(\theta) - \frac{1}{4} \Omega(u, k, \theta)^2 \geq ku\Psi(\theta) - \frac{1}{2}(1 + \theta)^2((1 - u)^2 + k^2))
\]

\[
= ku(1 + \theta)^2 \left( \frac{\Psi(\theta)}{(1 + \theta)^2} - \frac{(1 - u)^2 + k^2}{2ku} \right)
\]

\[
\geq ku(1 + \theta)^2 \left( \frac{\Psi(\theta)}{(1 + \theta)^2} - \frac{\sqrt{k^2 + 1} - 1}{k} \right).
\]

Denote \( c = \frac{\Psi(\theta)}{(1 + \theta)^2} \). For any \( \theta \in [0, \sqrt{2} - 1] \), \( c \in (0, 1] \). As shown in Figure 18, it is not hard to verify that \( \frac{\sqrt{k^2 + 1} - 1}{k} > 0 \) is guaranteed to have a positive solution \( k > 0 \). This proves the lemma. □

\[
\frac{\sqrt{k^2 + 1} - 1}{k}
\]

Figure 18. Graph of \( \frac{\sqrt{k^2 + 1} - 1}{k} \).

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

\[
v^\top(F(y) - F(x)) = v^\top DF(x + t_v(y - x))(y - x).
\]

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

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

\[
a_{k+2} - a_k \leq -\Phi a_{k+1}, \quad k \geq 0
\]

with \( \Phi > 0 \). Then

\[
a_k \leq \left( \frac{2}{\Phi + \sqrt{\Phi^2 + 4}} \right)^{k+1} \left( a_1 + \frac{\Phi + \sqrt{\Phi^2 + 4}}{2} a_0 \right)
\]

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

Taking the index in (B.19) as \( k - 1 \) and \( k \), one obtains
\[
\begin{align*}
r_+^{k-1} \left( a_1 + \frac{1}{r_+} a_0 \right) &\geq a_k + \frac{1}{r_+} a_{k-1} > a_k; \\
r_+^k \left( a_1 + \frac{1}{r_+} a_0 \right) &\geq a_{k+1} + \frac{1}{r_+} a_k > \frac{1}{r_+} a_k.
\end{align*}
\]

This yields
\[
\begin{align*}
a_k &\leq r_+^{k-1} \left( a_1 + \frac{1}{r_+} a_0 \right), \quad \text{and} \quad a_k \leq r_+^{k+1} \left( a_1 + \frac{1}{r_+} a_0 \right), \quad k \geq 1.
\end{align*}
\]

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
\[
bT \zeta_{a,b,c}(h_t) \text{Lip}(R) < \sqrt{2} - 1 < 1,
\]

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

Now we suppose \( \{U_k, Q_k\} \) solves (2.15). We write \( J_k = J(U_k, Q_k) \) for convenience. Then we want to bound \( J_{k+1} - J_k \) from above. We calculate
\[
J_{k+1} - J_k = (U_{k+1} - U_k) \cdot \left( \frac{1}{2} (U_{k+1} + U_k) - U_\ast \right) + (Q_{k+1} - Q_k) \cdot \left( \frac{Q_{k+1} + Q_k}{2} \right)
\leq (U_{k+1} - U_k) \cdot \left( \frac{1}{2} (U_{k+1} + U_k) - U_\ast \right) + (Q_{k+1} - Q_k) \cdot Q_{k+1}
= (U_{k+1} - U_k) \cdot (U_k - U_\ast) + \frac{1}{2} \|U_{k+1} - U_k\|^2 + (Q_{k+1} - Q_k) \cdot Q_{k+1}.
\]

The inequality is due to the convexity of the quadratic function \( \|Q\|^2 \). From (2.15), we know
\[
U_{k+1} - U_k = -\tau U D \bar{F}(U_k)^\top (Q_{k+1} + \omega \tau P(\bar{F}(U_k) - \epsilon Q_{k+1}));
= -\tau U D \bar{F}(U_k)^\top ((1 - \gamma \epsilon)Q_{k+1} + \gamma \bar{F}(U_k));
Q_{k+1} - Q_k = \tau P(\bar{F}(U_k) - \epsilon Q_{k+1}).
\]
Let us define $\bar{\gamma} = \omega \tau_P$ and $\varrho = \frac{\tau_P \gamma}{\nu_U}$. Using $F(U_\ast) = 0$, we obtain
\[
\mathcal{J}_{k+1} - \mathcal{J}_k = -\tau_U (U_k - U_\ast)^\top D\hat{F}(U_k)^\top ((1 - \bar{\gamma}\epsilon)Q_{k+1} + \bar{\gamma}(\hat{F}(U_k) + \tau_P Q_{k+1}^\top (\hat{F}(U_k) - \epsilon Q_{k+1})) + \frac{1}{2} ||U_{k+1} - U_k||^2
\]
\[
= -\tau_U \left( \bar{\gamma}(U_k - U_\ast)^\top D\hat{F}(U_k)^\top (\hat{F}(U_k) + (1 - \bar{\gamma}\epsilon)(U_k - U_\ast)^\top D\hat{F}(U_k)^\top Q_{k+1}
\right.
\]
\[
\left. - \varrho \hat{F}(U_k)^\top Q_{k+1} + \varrho \epsilon ||Q_{k+1}||^2 \right) + \frac{\tau_P^2}{2} ||D\hat{F}(U_k)^\top ((1 - \bar{\gamma}\epsilon)Q_{k+1} + \bar{\gamma}\hat{F}(U_k))||^2
\]
\[
= -\tau_U \left( \bar{\gamma}(U_k - U_\ast)^\top D\hat{F}(U_k)^\top (\hat{F}(U_k) - \hat{F}(U_\ast)) + (1 - \bar{\gamma}\epsilon)(U_k - U_\ast)^\top D\hat{F}(U_k)^\top Q_{k+1}
\right.
\]
\[
\left. - \varrho (\hat{F}(U_k) - \hat{F}(U_\ast))^\top Q_{k+1} + \varrho \epsilon ||Q_{k+1}||^2 \right) + \frac{\tau_P^2}{2} ||D\hat{F}(U_k)^\top ((1 - \bar{\gamma}\epsilon)Q_{k+1} + \bar{\gamma}\hat{F}(U_k))||^2.
\]
By Lemma 17, term (A) and term (C) are given by
\[
(A) = \bar{\gamma}(U_k - U_\ast)^\top D\hat{F}(U_k)^\top D\hat{F}(U_{k,\nu_1})(U_k - U_\ast),
\]
\[
(C) = -\varrho(U_k - U_\ast)^\top D\hat{F}(U_{k,\nu_2})^\top Q_{k+1},
\]
where $U_{k,\nu_j} = U_\ast + \nu_j(U_k - U_\ast)$ with $\nu_1, \nu_2 \in (0, 1), j = 1, 2$.

Recall that $D\hat{F}(U) = I + D\eta(U)$. To simplify the notation, we write
\[
\bar{\sigma}_\eta = \sup_{U \in \mathbb{R}^n} \{||D\eta(U)||\}.
\]
By Lemma 6, we have $\bar{\sigma}_\eta \leq bT\zeta_{a,b,c}(h_k)\text{Lip}(R)$. We now estimate term (A) as
\[
(A) = \bar{\gamma}(U_k - U_\ast)^\top D\hat{F}(U_k)^\top D\hat{F}(U_{k,\nu_1})(U_k - U_\ast)
\]
\[
= (U_k - U_\ast)^\top (I + D\eta(U_\ast))(I + D\eta(U_{k,\nu_1}))(U_k - U_\ast)
\]
\[
= ||U_k - U_\ast||^2 + (U_k - U_\ast)^\top D\eta(U_\ast)^\top (U_k - U_\ast) + (U_k - U_\ast)^\top D\eta(U_{k,\nu_1})(U_k - U_\ast)
\]
\[
+ (U_k - U_\ast)^\top D\eta(U_{k,\nu_1})^\top D\eta(U_{k,\nu_1})(U_k - U_\ast)
\]
\[
\geq (1 - 2\bar{\sigma}_\eta - \bar{\sigma}_\eta^2)||U_k - U_\ast||^2.
\]
We can further estimate the terms (B), (C), and (E) as
\[
(B) = (1 - \bar{\gamma}\epsilon)(U_k - U_\ast)^\top (I + D\eta(U_\ast))Q_{k+1};
\]
\[
(C) = -\varrho(U_k - U_\ast)^\top D\hat{F}(U_{k,\nu_2})^\top Q_{k+1} = -\varrho(U_k - U_\ast)^\top (I + D\eta(U_{k,\nu_2}))^\top Q_{k+1}.
\]
Thus
\[
(B) + (C) = (1 - \bar{\gamma}\epsilon)(U_k - U_\ast)^\top (I + D\eta(U_\ast))Q_{k+1} - \varrho(U_k - U_\ast)^\top (I + D\eta(U_{k,\nu_2}))^\top Q_{k+1}
\]
\[
= (1 - \bar{\gamma}\epsilon - \varrho)(U_k - U_\ast)^\top Q_{k+1} + (U_k - U_\ast)^\top ((1 - \bar{\gamma}\epsilon)D\eta(U_\ast) - \varrho D\eta(U_{k,\nu_2}))^\top Q_{k+1}
\]
\[
\geq - ||1 - \bar{\gamma}\epsilon - \varrho||U_k - U_\ast||Q_{k+1}|| - (1 - \bar{\gamma}\epsilon)\bar{\sigma}_\eta||U_k - U_\ast||||Q_{k+1}||
\]
\[
= - (1 - \bar{\gamma}\epsilon - \varrho + (1 - \bar{\gamma}\epsilon)\bar{\sigma}_\eta)||U_k - U_\ast||||Q_{k+1}||.
\]
And
\[
(E) \leq \bar{\sigma}_\eta^2(||1 - \bar{\gamma}\epsilon||||Q_{k+1}|| + \bar{\gamma}||\hat{F}(U_k)||)^2
\]
\[
\leq \bar{\sigma}_\eta^2(||1 - \bar{\gamma}\epsilon||||Q_{k+1}|| + \bar{\gamma}\bar{\sigma}_\eta||U_k - U_\ast||)^2
\]
\[
\leq 2\bar{\sigma}_\eta^2((1 - \bar{\gamma}\epsilon)^2||Q_{k+1}||^2 + \bar{\gamma}^2\bar{\sigma}_\eta^2||U_k - U_\ast||^2).
\]
The second inequality on (E) is due to
\[
\|\hat{F}(U_k)\| = \|\hat{F}(U_k) - \hat{F}(U_\ast)\| = \left\| \int_0^1 \left( \frac{d}{ds} \hat{F}(U_\ast + s(U_k - U_\ast)) \right) \, ds \right\|
\]
\[
\leq \int_0^1 \|U_k - U_\ast\| \, ds = \sigma \|U_k - U_\ast\|.
\]
Combining the estimations on term (A)-(E), we obtain
\[
J_{k+1} - J_k = -\tau_U \left( \gamma(1 - 2\sigma \eta - \sigma^2 \eta^2)\|U_k - U_\ast\|^2 - (|1 - \gamma \epsilon - \varrho| + (|1 - \gamma \epsilon| + \varrho)\sigma \eta)\|U_k - U_\ast\| \right)_{Q_k+1}
\]
\[
+ \varrho \epsilon\|Q_k+1\|^2 - \tau_U (\sigma^2 (1 - \gamma \epsilon)^2\|Q_k+1\|^2 + \gamma^2 \sigma^2 \|U_k - U_\ast\|^2)
\]
\[
= -\tau_U \left[ \|U_k - U_\ast\|, \|Q_k+1\| \right] (\Gamma - \tau_U \Theta) \left[ \|U_k - U_\ast\| \right].
\]
Here
\[
\Gamma = \left[ \begin{array}{c}
\gamma(1 - 2\sigma \eta - \sigma^2 \eta^2) \\
-\frac{1}{2}(|1 - \gamma \epsilon - \varrho| + (|1 - \gamma \epsilon| + \varrho)\sigma \eta)
\end{array} \right],
\]
\[
\Theta = \left[ \begin{array}{c}
\gamma^2 \sigma^4 \\
\sigma^2 (1 - \gamma \epsilon)^2
\end{array} \right].
\]
Recall that we assume \(bT \zeta_{a,b,c}(h) \text{Lip}(R) \leq \theta\), this leads to \(\sigma \leq \theta\). By Lemma 6, we also have \(\sigma \leq 1 + \theta\). Thus, \(\gamma(1 - 2\sigma \eta - \sigma^2 \eta^2) > \gamma(1 - 2\theta - \theta^2) > 0\) as \(\theta \in [0, \sqrt{2} - 1]\). Hence,
\[
det(\Gamma) = \varrho \gamma \epsilon(1 - 2\sigma \eta - \sigma^2 \eta^2) - \frac{1}{4}(|1 - \gamma \epsilon - \varrho| + (|1 - \gamma \epsilon| + \varrho)\sigma \eta)^2
\]
\[
\geq \varrho \gamma \epsilon(1 - 2\theta - \theta^2) - \frac{1}{4}(|1 - \gamma \epsilon - \varrho| + (|1 - \gamma \epsilon| + \varrho)\theta)^2.
\]
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 \(\gamma, \omega, \epsilon\), such that (3.16) holds. The condition (3.16) leads to \(\det(\Gamma) > 0\), which guarantees the positive definiteness of \(\Gamma\).

Furthermore, we have \(\Gamma \succeq \lambda_2(\Gamma) I\), where \(\lambda_2(\Gamma)\) represents the smallest eigenvalue of \(\Gamma\) and \(I\) is an identity matrix. One can bound \(\lambda_2(\Gamma)\) from below as
\[
\lambda_2(\Gamma) = \frac{\gamma(1 - 2\sigma \eta - \sigma^2 \eta^2) + \varrho \epsilon - \sqrt{\gamma(1 - 2\sigma \eta - \sigma^2 \eta^2) + \varrho \epsilon} - 2\text{det}(\Gamma)}{2}
\]
\[
\geq \frac{4(\varrho \gamma \epsilon(1 - 2\theta - \theta^2) - \frac{1}{4}(|1 - \gamma \epsilon - \varrho| + (|1 - \gamma \epsilon| + \varrho)\theta)^2)}{2(\gamma(1 - 2\sigma \eta - \sigma^2 \eta^2) + \varrho \epsilon + \sqrt{\gamma(1 - 2\sigma \eta - \sigma^2 \eta^2) + \varrho \epsilon} - 2\text{det}(\Gamma))}
\]
\[
\geq \frac{\varrho \gamma \epsilon(1 - 2\theta - \theta^2) - \frac{1}{4}(|1 - \gamma \epsilon - \varrho| + (|1 - \gamma \epsilon| + \varrho)\theta)^2}{\gamma(1 - 2\sigma \eta - \sigma^2 \eta^2) + \varrho \epsilon}
\]
\[
\geq \frac{\varrho \gamma \epsilon \Psi(\theta) - \frac{1}{4} \Omega(\gamma \epsilon, \varrho, \theta)^2}{\gamma + \varrho \epsilon}.
\]
On the other hand, we have
\[
\Theta \prec \sigma^2 \max\{\gamma^2 \sigma^2, |1 - \gamma \epsilon|^2\} I \prec (1 + \theta)^2 \max\{\gamma^2 (1 + \theta)^2, |1 - \gamma \epsilon|^2\} I.
\]
Thus we have
\[
\Gamma - \tau \Theta > \left( \frac{\tilde{\gamma}e\Psi(\theta) - \frac{1}{4}\Omega(\tilde{\gamma}e, \theta)^2}{\tilde{\gamma} + qe} - \tau(1 + \theta)^2 \max\{\tilde{\gamma}^2(1 + \theta)^2, |1 - \tilde{\gamma}e|^2\} \right) I.
\]

Plug this estimation to (B.20), we obtain
\[
J_{k+1} - J_k \leq -\tau C(\theta, \tilde{\gamma}, \epsilon, \varrho, \tau)(\|U_k - U^*\|^2 + \|Q_{k+1}\|^2).
\]
Since we set the PDHG step size as
\[
0 < \tau < \bar{\tau}(\theta, e^\gamma, \epsilon, \varrho, \tau) = \frac{\tilde{\gamma}e\Psi(\theta) - \frac{1}{4}\Omega(\tilde{\gamma}e, \theta)^2}{2(\tilde{\gamma} + qe)(1 + \theta)^2 \max\{\tilde{\gamma}^2(1 + \theta)^2, (1 - \tilde{\gamma}e)^2\}},
\]
this guarantees \(C(\theta, \tilde{\gamma}, \epsilon, \varrho, \tau) > 0\).

Furthermore, as a function of \(\tau\), \(\tau C(\theta, \tilde{\gamma}, \epsilon, \varrho, \tau)\) reaches its maximum value at \(\tau = \frac{1}{2}\bar{\tau}(\theta, \tilde{\gamma}, \epsilon, \varrho, \tau)\). We then set (here \(\bar{\tau}\) denotes \(\bar{\tau}(\theta, \tilde{\gamma}, \epsilon, \varrho)\))
\[
\Phi = \frac{1}{2}\bar{\tau}C(\theta, \tilde{\gamma}, \epsilon, \varrho, \tau) = \frac{\tilde{\gamma}e\Psi(\theta) - \frac{1}{4}\Omega(\tilde{\gamma}e, \theta)^2}{2(1 + \theta)^2 \max\{\tilde{\gamma}^2(1 + \theta)^2, (1 - \tilde{\gamma}e)^2\}(\tilde{\gamma} + qe)^2}.
\]
Thus we have
\[
J_{k+1} - J_k \leq -\Phi \cdot \frac{1}{2}(\|U_k - U^*\|^2 + \|Q_{k+1}\|^2).
\]
Now we prove the exponential decay of \(J_k\). To do so, we sum up the above inequality at time index \(k\) and \(k + 1\) to obtain,
\[
J_{k+2} - J_k \leq -\Phi \cdot \frac{1}{2}(\|U_{k+1} - U^*\|^2 + \|Q_{k+2}\|^2 + \|U_k - U^*\|^2 + \|Q_{k+1}\|^2), \quad k \geq 0.
\]
It is not hard to see that the right-hand side of the above inequality is no larger than \(-\Phi J_{k+1}\). Hence,
\[
J_{k+2} - J_k \leq -\Phi J_{k+1}. \quad (B.21)
\]

Now, by Lemma 18, we obtain
\[
J_k \leq \left( \frac{2}{\Phi + \sqrt{\Phi^2 + 4}} \right)^{k+1} \left( J_1 + \frac{\Phi + \sqrt{\Phi^2 + 4}}{2} J_0 \right), \quad \text{for } k \geq 1.
\]
This concludes our proof.