Journal of Scientific Computing (2022) 92:90
https://doi.org/10.1007/s10915-022-01933-5

Hessian Informed Mirror Descent

Li Wang[1] · **Ming Yan**[2]

Received: 11 April 2022 / Accepted: 25 June 2022 / Published online: 26 July 2022

Abstract
Inspired by the recent paper (L. Ying, Journal of Scientific Computing, 84, 1–14 (2020), we explore the relationship between the mirror descent and the variable metric method. When the metric in the mirror decent is induced by a convex function, whose Hessian is close to the Hessian of the objective function, this method enjoys both robustness from the mirror descent and superlinear convergence for Newton type methods. When applied to a linearly constrained minimization problem, we prove the global and local convergence, both in the continuous and discrete settings. As applications, we compute the Wasserstein gradient flows and Cahn-Hillard equation with degenerate mobility. When formulating these problems using a minimizing movement scheme with respect to a variable metric, our mirror descent algorithm offers a fast convergence speed for the underlying optimization problem while maintaining the total mass and bounds of the solution.

Keywords Mirror descent · Variable metric · Wasserstein gradient flow · Degenerate mobility

1 Introduction

We consider the following linearly constrained minimization problem

$$\min_{u\in\Omega: Au=b} f(u)\,, \tag{1.1}$$

where Ω is a closed convex domain in $\mathbb{R}^n$ with nonempty interior, $f:\Omega\to\mathbb{R}$ is a convex differentiable function and $A\in\mathbb{R}^{m\times n}$ with m being a small nonnegative integer. One typical

L.W. is partially supported by NSF grant DMS-1846854. M.Y is partially supported by NSF grant DMS-2012439.

✉ Li Wang
wang8818@umn.edu

Ming Yan
myan@msu.edu

[1] School of Mathematics, University of Minnesota, Twin cities, MN 55455, USA

[2] Department of Computational Mathematics, Science and Engineering and Department of Mathematics, Michigan State University, East Lansing, MI 48824, USA

example is with f taking the form:

$$f(u) = \sum_{i=1}^{n} (g_i(u_i) + u_i V_i) + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} W_{i,j} u_i u_j \,, \tag{1.2}$$

where g_i is a convex function, $W_{i,j} = W(|x_i - x_j|)$ and $V_i = V(x_i)$, with $W(|x|)$ and $V(x)$ assumed to be $\lambda-$convex. The linear constraint $Au = b$ oftentimes encodes the properties such as mass conservation. The convex set Ω can be denoted by, for instance, $\Omega = \{u : u_i \geq 0\}$, which imposes a bound constraint on u. This example arises in aggregation dynamics [10, 26], kinetic description of granular gas [2], the mean field limit of neural networks [24], among many others. In this paper, we assume that the problem (1.1) has a unique solution u^*.

When $A = \mathbf{1}^\top$ (the all one row vector), $b = 1$, and $\Omega = \{u : u_i \geq 0\}$, then the feasible set is the simplex

$$\mathcal{U} = \left\{ u : u_i \geq 0, \qquad \sum_{i=1}^{n} u_i = 1 \right\} . \tag{1.3}$$

In this case, a strongly convex function $\Phi(u)$ is constructed to solve the problem (1.1), e.g., $\Phi(u) = \sum_{i=1}^{n} g_i(u_i)$ for the general case, and $\Phi(u) = \sum_{i=1}^{n} g_i(u_i) + \frac{1}{2} W_{i,i} u_i^2$ if the matrix $[W_{i,j}]$ is positive semidefinite. Ying considered three different types of strongly convex functions g_i in [28]: Kullback-Leibler divergence, reverse Kullback-Leibler divergence, and Hellinger divergence. Then, the mirror descent has the following update formula:

$$\text{mirror descent}: \quad \nabla\Phi(u^{k+1}) = \nabla\Phi(u^k) - \eta_{\mathcal{M}}(\nabla f(u^k) + A^\top c(u^k)) \,, \tag{1.4}$$

where $\eta_{\mathcal{M}}$ is the stepsize and $c(u^k) \in \mathbb{R}^m$ is the unique vector to be determined such that $Au^{k+1} = b$. The nonnegative conditions $\{u_i \geq 0\}_{i=1}^n$ are automatically satisfied because of the log terms in $\{g_i\}_{i=1}^n$, and $c(u^k)$ plays the role of the Lagrangian multiplier for the constraint $Au = b$ in the mirror descent update. More importantly, for the special case when $g(u) = u \log u$, the value $c(u^k)$ can be easily found by a normalization step. For other cases in [28], the value for $c(u^k)$ is efficiently found by iterative algorithms such as Newton and bisection.

In practice, to obtain u^{k+1} from (1.4), let Φ^* be the conjugate function of Φ, which is defined as $\Phi^*(v) = \max_u \; v^\top u - \Phi(u)$. Then we have $u = \nabla\Phi^*(\nabla\Phi(u))$. Therefore, the mirror descent (1.4) has the following equivalent form:

$$\begin{aligned} \text{mirror descent}: \quad u^{k+1} &= \nabla\Phi^*(\nabla\Phi(u^{k+1})) = \nabla\Phi^*(\nabla\Phi(u^k) - \eta_{\mathcal{M}}(\nabla f(u^k) \\ &\quad + A^\top c(u^k))). \end{aligned} \tag{1.5}$$

When Φ is continuous differentiable, then (1.5) reduces to:

$$u^{k+1} = (\nabla\Phi)^{-1}(\nabla\Phi(u^k) - \eta_{\mathcal{M}}(\nabla f(u^k) + A^\top c(u^k))).$$

To put (1.5) in a more general framework, since $u = \nabla\Phi^*(\nabla\Phi(u))$, taking derivative with respect to u, we have $\mathbf{I} = \nabla^2\Phi(u)\nabla^2\Phi^*(\nabla\Phi(u))$. Then the above equation bares the following first order approximation:

$$\begin{aligned} u^{k+1} \approx & \nabla\Phi^*(\nabla\Phi(u^k)) - \eta_{\mathcal{M}}\nabla^2\Phi^*(\nabla\Phi(u^k))(\nabla f(u^k) + A^\top c(u^k)) \\ = & u^k - \eta_{\mathcal{M}}\nabla^2\Phi(u^k)^{-1}(\nabla f(u^k) + A^\top c(u^k)) \,. \end{aligned}$$

It shows that mirror descent is a discretization of

$$\dot{u} = -\nabla^2\Phi(u)^{-1}(\nabla f(u) + A^\top c(u)) \,. \tag{1.6}$$

Since we can multiply Φ by a scalar and change the ordinary differential equation, to simplify the following analysis, we assume that Φ is 1-strongly convex with respect to a given norm $\|\cdot\|_w$, i.e.,

$$\Phi(x) - \Phi(y) - \nabla\Phi(y)^\top(x-y) \geq \frac{1}{2}\|x-y\|_w^2 \,.$$

A more direct discretization of (1.6) is to apply the forward Euler scheme, namely,

$$\text{variable metric}: \quad u^{k+1} = u^k - \eta_{\mathcal{N}}\nabla^2\Phi(u^k)^{-1}(\nabla f(u^k) + A^\top c(u^k)) \,, \tag{1.7}$$

which can be viewed as a first order variant of the mirror descent. Similarly, $\eta_{\mathcal{N}}$ is the stepsize and $c(u^k)$ is a vector to be determined such that $Au^{k+1} = b$. This method is equivalent to

$$\text{variable metric}: \quad u^{k+1} = \arg\min_{u:Au=b} f(u^k) + \langle \nabla f(u^k), u - u^k\rangle + \frac{1}{2\eta_{\mathcal{N}}}\|u-u^k\|^2_{\nabla^2\Phi(u^k)}, \tag{1.8}$$

and it is called variable metric [14] because of the variable metric $\nabla^2\Phi(u^k)$ used in the quadratic term. When $\Phi = f$, (1.8) reduces to the proximal Newton method [22].

In view of the mirror descent method (1.5) and variable metric method (1.7), they both are first order discretizations of the continuous flow (1.6). Despite vast literature on either method individually, there is little discussion on the relation between them. Indeed, for the mirror descent method, emphasize has been put on the treatment of constraints, especially the simplex constraint mentioned previously, which makes the choice of $\Phi(u) = \sum_{i=1}^n u_i \log u_i$ the most popular. On the other hand, in variable metric methods such as Newton type methods, Φ is chosen to incorporate the second order information of the objective function with the goal of improving the local convergence rate. The constraint, however, is often dealt with by a projection step. Inspired by the paper [28], we see that one can merge the advantages of both methods by constructing Φ that has both Hessian information and constraint guarantee. Consequently, by choosing the appropriate Bregman divergence in the mirror descent, we can prove the global convergence of the new method. This proof can easily lend itself to Newton type methods owing to their similarity. In return, following the superlinear convergence for Newton type methods, we can prove the same local convergence for the new method.

Comparing (1.4) and (1.8), it is important to point out the advantage of mirror descent in treating the constraint $x \in \Omega$. In practice, such a constraint can be enforced through the choice of Φ, for instance, $\Phi(u) = \begin{cases} \frac{1}{2}\|u\|_2^2 & \text{for } u \in \Omega \\ +\infty & \text{otherwise} \end{cases}$, or the ones we use in Sect. 4 (i.e., (4.4) and (4.19)). As a result, the iteration u^{k+1} obtained from the mirror descent satisfies the constraint by construction. In contrast, it is not guaranteed that u^{k+1} obtained from (1.7) will satisfy the constraint, and a projection step may be needed. In addition, in order to make sense of (1.7), Φ needs to be twice continuously differentiable, whereas (1.4) or (1.5) allows for a much wider class of Φ, which gives more flexibility to the mirror descent method.

The contributions and organization of this paper are summarized as follows:

- We establish the sublinear convergence of the gradient flow in (1.6) for a general $\Phi(u)$ in Sect. 2 and extend it to linear convergence with an improved rate in the case of strong convexity.
- We prove both the global and local convergence of two discretizations (1.4) and (1.7) in Sect. 3.
- Applications in variable metric gradient flows are presented in Sect. 4 along with numerical experiments.

Finally, the conclusion is drawn in Sect. 5.

2 Convergence of the Gradient Flow (1.6)

In this section, we consider the convergence of (1.6), which guides the convergence analysis of (1.4) and (1.7) in the next section. With the proper choice of distance measure, in particular the Bregman divergence in our case, the global convergence can be established. A similar version of the following theorem can be found in [21].

Theorem 1 *(Sublinear convergence) Let $u(t)$ be the solution to (1.6) with $u(0) = u_0$. Then we have*

$$f\left(\frac{1}{T}\int_0^T u(t)\mathrm{d}t\right) - f(u^*) \le \frac{1}{T} D_\Phi(u^*, u_0)\,, \tag{2.1}$$

where D_Φ is the Bregman divergence induced by Φ:

$$D_\Phi(u^*, u_0) = \Phi(u^*) - \Phi(u_0) - \nabla\Phi(u_0)^\top (u^* - u_0)\,.$$

Proof Consider the time derivative of $D_\Phi(u^*, u(t))$, we have

$$\begin{aligned}
\frac{d}{dt} D_\Phi(u^*, u(t)) &= \frac{d}{dt}\left[\Phi(u^*) - \Phi(u(t)) - \nabla\Phi(u(t))^\top (u^* - u(t))\right] \\
&= -\nabla\Phi(u(t))^\top \frac{du(t)}{dt} - \left[\nabla^2\Phi(u)\frac{du(t)}{dt}\right]^\top (u^* - u(t)) \\
&\quad + \nabla\Phi(u(t))^\top \frac{du(t)}{dt} \\
&= (\nabla f(u(t)) + A^\top c(u(t)))^\top (u^* - u(t)) = \nabla f(u(t))^\top (u^* - u(t)) \\
&\le f(u^*) - f(u(t))\,,
\end{aligned}$$

where the third equality uses (1.6) and the inequality comes from the convexity of f. Integrating both sides from 0 to T, we obtain

$$\frac{1}{T}\left[D_\Phi(u^*, u(T)) - D_\Phi(u^*, u_0)\right] \le f(u^*) - \frac{1}{T}\int_0^T f(u(t))\mathrm{d}t\,,$$

which readily implies (2.1) thanks again to the convexity of f and D_Φ being nonnegative. □

If we further assume the strong convexity of f, we can obtain the linear convergence.

Theorem 2 *(Linear convergence) Let $u(t)$ be the solution to (1.6) with $u(0) = u_0$. Define two Bregman divergences induced by Φ and f as $D_\Phi(t) := D_\Phi(u^*, u) = \Phi(u^*) - \Phi(u) -$*

$\langle \nabla \Phi(u), u^* - u \rangle$ *and* $D_f(t) := D_f(u^*, u) = f(u^*) - f(u) - \langle \nabla f(u), u^* - u \rangle$, *respectively. Assume that* $D_f(t) \geq \mu D_\Phi(t)$ *for all* t. *Then we have*

$$D_\Phi(t) \leq D_\Phi(0) \exp^{-\mu t}, \qquad \text{for all } t \geq t_0 .$$

Proof Denote $G(u) = -[\nabla f(u) + A^\top c(u)]$, then (1.6) writes as

$$\frac{d}{dt} \nabla \Phi(u) = G(u), \qquad \text{or} \quad \dot{u} = \nabla^2 \Phi(u)^{-1} G(u). \tag{2.2}$$

The global convergence result shows $G(u^*) = 0$. Then, we have

$$\begin{aligned} \dot{D_\Phi} &= -\langle G(u), u^* - u \rangle = -\langle \nabla f(u^*) + A^\top c(u^*) - \nabla f(u) - A^\top c(u), u^* - u \rangle \\ &= -\langle \nabla f(u^*) - \nabla f(u), u^* - u \rangle \qquad \text{from } (Au = Au^*) \\ &\leq -[f(u^*) - f(u) - \langle \nabla f(u), u^* - u \rangle] \leq -\mu D_\Phi(t). \end{aligned}$$

Therefore, we have $D_\Phi(t) \leq D_\Phi(0) \exp^{-\mu t}$. □

Remark 1 The scalar μ determines the linear convergence rate. If $\Phi(u) = \|u\|^2/2$, then μ is the strongly convex constant with respect to the standard norm, which can be very small in some applications. In such cases, if Φ is chosen according to the Hessian of f, then μ can be much larger than the strongly convex constant of f with respect to the standard norm and results in a much faster convergence.

3 Convergence at the Discrete Level

This section is devoted to the convergence of the discrete schemes (1.4) and (1.7). For global convergence, the proof follows a similar line of reasoning as in the continuous setting but with more involved calculations; whereas the local convergence is obtained via a two stage proof as in other Newton type methods.

3.1 Global Convergence

We first establish the global convergence of (1.4), which is slightly different from that in [6]. We still include it here for completeness.

Theorem 3 *(Global sublinear convergence for mirror descent (1.4)) Assume* Φ *is 1-strongly convex w.r.t. a certain norm* $\|\cdot\|_\omega$, *i.e.,*

$$D_\Phi(x, y) \geq \frac{1}{2} \|x - y\|_\omega^2 . \tag{3.1}$$

Let $\{u^k\}$ *be the solution to (1.4) with the initial* $u^0 = u_0$. *Then we have*

$$f\left(\frac{1}{K} \sum_{k=0}^{K-1} u^k\right) - f(u^*) \leq \frac{1}{\eta_{\mathcal{M}} K} D_\Phi(u^*, u_0) + \frac{1}{K} \sum_{k=0}^{K-1} \frac{\eta_{\mathcal{M}}}{2} \|\nabla f(u^k) + A^\top c(u^k)\|_{\omega,*}^2, \tag{3.2}$$

where $\|\cdot\|_{\omega,*}$ *is the dual norm of* $\|\cdot\|_\omega$.

Proof We mimic the proof of Theorem 1. Consider

$$D_\Phi(u^*, u^{k+1}) - D_\Phi(u^*, u^k) = \Phi(u^k) - \Phi(u^{k+1}) + \nabla\Phi(u^k)^\top(u^* - u^k) - \nabla\Phi(u^{k+1})^\top(u^* - u^{k+1}).$$

Plugging in the relation (1.4) and using $Au^* = Au^k = Au^{k+1} = b$, we have

$$\begin{aligned} &D_\Phi(u^*, u^{k+1}) - D_\Phi(u^*, u^k) \\ &= \Phi(u^k) - \Phi(u^{k+1}) + \nabla\Phi(u^{k+1})^\top(u^{k+1} - u^k) + \eta_{\mathcal{M}}[\nabla f(u^k) + A^\top c(u^k)]^\top(u^* - u^k) \\ &\le -D_\Phi(u^{k+1}, u^k) + \eta_{\mathcal{M}}(\nabla f(u^k) + A^\top c(u^k))^\top(u^k - u^{k+1}) + \eta_{\mathcal{M}}[f(u^*) - f(u^k)]. \end{aligned} \tag{3.3}$$

Using the fact that Φ is 1-strongly convex w.r.t. norm $\|\cdot\|_\omega$, we have

$$\begin{aligned} &- D_\Phi(u^{k+1}, u^k) + \eta_{\mathcal{M}}(\nabla f(u^k) + A^\top c(u^k))^\top(u^k - u^{k+1}) \\ &\le \frac{\eta_{\mathcal{M}}^2}{2}\|\nabla f(u^k) + A^\top c(u^k)\|_{\omega,*}^2, \end{aligned}$$

where we have used the Young's inequality for the term $\eta_{\mathcal{M}}(\nabla f(u^k) + A^\top c(u^k))^\top(u^k - u^{k+1})$. Therefore, we have

$$f(u^k) - f(u^*) \le \frac{1}{\eta_{\mathcal{M}}}\left[D_\Phi(u^*, u^k) - D_\Phi(u^*, u^{k+1})\right] + \frac{\eta_{\mathcal{M}}}{2}\|\nabla f(u^k) + A^\top c(u^k)\|_{\omega,*}^2.$$

Summing from $k = 0$ to $K - 1$ and dividing by K give rise to (3.2). □

The inequality (3.2) is still valid if $A^\top c(u^k)$ is removed, and it reduces to the standard convergence result with bounded gradient [6]. However, we add $A^\top c(u^k)$ here because $\nabla f(u^k) + A^\top c(u^k)$ converges to 0 when u^k converges, while $\nabla f(u^k)$ may not.

Theorem 4 *(Global convergence for variable-metric (1.7)) Assume Φ is 1-strongly convex w.r.t. norm $\|\cdot\|_2$ and $\nabla^2\Phi$ is L-Lipschitz, i.e.,*

$$\|\nabla^2\Phi(x) - \nabla^2\Phi(y)\|_2 \le L\|x - y\|_2.$$

Then the solution $\{u^k\}$ to (1.7) with initial $u^0 = u_0$ satisfies

$$\begin{aligned} f\left(\frac{1}{K}\sum_{k=0}^{K-1} u^k\right) - f(u^*) \le{}& \frac{1}{\eta_{\mathcal{N}} K} D_\Phi(u^*, u_0) + \frac{1}{K}\sum_{k=0}^{K-1}\frac{\eta_{\mathcal{N}}}{2}\|\nabla f(u^k) + A^\top c(u^k)\|_2^2 \\ &+ \frac{L\eta_{\mathcal{N}}}{2}\frac{1}{K}\sum_{k=0}^{K-1}\Big[\|\nabla^2\Phi(u^k)^{-1}(\nabla f(u^k) \\ &+ A^\top c(u^k))\|_2^2\|u^* - u^{k+1}\|_2\Big]. \end{aligned} \tag{3.4}$$

Moreover, if we assume $\Omega\bigcap\{u : Au = b\}$ is bounded, and $\nabla\Phi$ and ∇f are Lipschitz continuous with constant L_1 and L_2 over the bounded set $\Omega \cap \{u : Au = b\}$, respectively, i.e.,

$$\|\nabla^2\Phi(u)\|_2 \le L_1, \qquad \|\nabla^2 f(u)\|_2 \le L_2 \qquad \forall u \in \Omega \cap \{u : Au = b\}.$$

Then

$$\|\nabla f(u^k) + A^\top c(u^k)\|_2 \to 0 \quad as \quad k \to \infty. \tag{3.5}$$

Proof Here we follow the approach in the proof of Theorem 3 but tailor the details according to the update rule (1.7). First we write

$$\begin{aligned} &D_\Phi(u^*, u^{k+1}) - D_\Phi(u^*, u^k) \\ &\quad = \Phi(u^k) - \Phi(u^{k+1}) + \nabla\Phi(u^k)^\top(u^* - u^k) - \nabla\Phi(u^{k+1})^\top(u^* - u^{k+1}) \\ &\quad = -D_\Phi(u^{k+1}, u^k) + [\nabla\Phi(u^k) - \nabla\Phi(u^{k+1})]^\top(u^* - u^{k+1})\,. \end{aligned} \tag{3.6}$$

Next, compute the difference

$$\begin{aligned} &\nabla\Phi(u^k) - \nabla\Phi(u^{k+1}) \\ &\quad = -\int_0^1 \nabla^2\Phi(u^k + t(u^{k+1} - u^k))(u^{k+1} - u^k)\mathrm{d}t \\ &\quad = \eta_{\mathcal{N}} \int_0^1 \nabla^2\Phi(u^k + t(u^{k+1} - u^k))\nabla^2\Phi(u^k)^{-1}(\nabla f(u^k) + A^\top c(u^k))\mathrm{d}t \\ &\quad = \eta_{\mathcal{N}}(\nabla f(u^k) + A^\top c(u^k)) \\ &\qquad - \eta_{\mathcal{N}} \underbrace{\int_0^1 \left[I - \nabla^2\Phi(u^k + t(u^{k+1} - u^k))\nabla^2\Phi(u^k)^{-1}\right](\nabla f(u^k) + A^\top c(u^k))\mathrm{d}t}_{\mathcal{A}}\,. \end{aligned} \tag{3.7}$$

Plugging (3.7) into (3.6) gives

$$\begin{aligned} &D_\Phi(u^*, u^{k+1}) - D_\Phi(u^*, u^k) \\ &\quad = \eta_{\mathcal{N}}(\nabla f(u^k) + A^\top c(u^k))^\top(u^k - u^{k+1} + u^* - u^k) \\ &\quad - \eta_{\mathcal{N}}\mathcal{A}^\top(u^* - u^{k+1}) - D_\Phi(u^{k+1}, u^k) \\ &\quad \le \eta_{\mathcal{N}}(\nabla f(u^k) + A^\top c(u^k))^\top(u^* - u^{k+1}) + \eta_{\mathcal{N}}[f(u^*) - f(u^k)] \\ &\qquad - \eta_{\mathcal{N}}\mathcal{A}^\top(u^* - u^{k+1}) - D_\Phi(u^{k+1}, u^k) \\ &\quad \le \frac{\eta_{\mathcal{N}}^2}{2}\|\nabla f(u^k) + A^\top c(u^k)\|_2^2 - \eta_{\mathcal{N}}[f(u^*) - f(u^k)] - \eta_{\mathcal{N}}A^\top(u^* - u^{k+1})\,, \end{aligned} \tag{3.8}$$

where we again use the Young's inequality for the first term and 1-strongly convexity of Φ. Because $\nabla^2\Phi$ is L-Lipschitz, we have

$$\begin{aligned} \|\mathcal{A}\|_2 &\le \frac{L}{2}\|u^{k+1} - u^k\|_2\|\nabla^2\Phi(u^k)^{-1}(\nabla f(u^k) + A^\top c(u^k))\|_2 \\ &= \frac{L\eta_{\mathcal{N}}}{2}\|\nabla^2\Phi(u^k)^{-1}(\nabla f(u^k) + A^\top c(u^k))\|_2^2. \end{aligned} \tag{3.9}$$

Therefore, similarly to the previous theorem, we have

$$\begin{aligned} &f(u^k) - f(u^*) \\ &\quad \le \frac{1}{\eta_{\mathcal{N}}}\left[D_\Phi(u^*, u^k) - D_\Phi(u^*, u^{k+1})\right] + \frac{\eta_{\mathcal{N}}}{2}\|\nabla f(u^k) + A^\top c(u^k)\|_2^2 - \mathcal{A}^\top(u^* - u^{k+1}) \\ &\quad \le \frac{1}{\eta_{\mathcal{N}}}\left[D_\Phi(u^*, u^k) - D_\Phi(u^*, u^{k+1})\right] + \frac{\eta_{\mathcal{N}}}{2}\|\nabla f(u^k) + A^\top c(u^k)\|_2^2 \\ &\qquad + \frac{L\eta_{\mathcal{N}}}{2}\|\nabla^2\Phi(u^k)^{-1}(\nabla f(u^k) + A^\top c(u^k))\|_2^2\|u^* - u^{k+1}\|_2\,. \end{aligned}$$

Summing from $k = 0$ to $K - 1$ and dividing it by K, we arrive at (3.4).

To show (3.5), note that, using (1.7),

$$\begin{aligned} f(u^{k+1}) &= f(u^k - \eta_{\mathcal{N}} \nabla^2 \Phi(u^k)^{-1}(\nabla f(u^k) + A^\top c(u^k))) \\ &\le f(u^k) - \eta_{\mathcal{N}} \nabla f(u^k)^\top \nabla^2 \Phi(u^k)^{-1}(\nabla f(u^k) + A^\top c(u^k)) \\ &\quad + \frac{\eta_{\mathcal{N}}^2}{2} L_2 \|\nabla^2 \Phi(u^k)^{-1}(\nabla f(u^k) + A^\top c(u^k))\|_2^2 \\ &= f(u^k) - \eta_{\mathcal{N}} [\nabla f(u^k) + A^\top c(u^k)]^\top \nabla^2 \Phi(u^k)^{-1}(\nabla f(u^k) + A^\top c(u^k)) \\ &\quad + \frac{\eta_{\mathcal{N}}^2}{2} L_2 \|\nabla^2 \Phi(u^k)^{-1}(\nabla f(u^k) + A^\top c(u^k))\|_2^2 , \end{aligned} \tag{3.10}$$

where the last equality uses the fact

$$A \nabla^2 \Phi(u^k)^{-1}(\nabla f(u^k) + A^\top c(u^k)) = \frac{1}{\eta_{\mathcal{N}}} A(u^k - u^{k+1}) = 0 .$$

Then if we choose $\eta_{\mathcal{N}} \le \frac{1}{L_1 L_2}$, (3.10) becomes

$$\begin{aligned} f(u^{k+1}) - f(u^k) &\le -\eta_{\mathcal{N}} \left(\frac{1}{L_1} - \frac{\eta_{\mathcal{N}} L_2}{2} \right) \|\nabla^2 \Phi(u^k)^{-1}(\nabla f(u^k) + A^\top c(u^k))\|_2^2 \\ &\le -\frac{\eta_{\mathcal{N}}}{2L_1} \|\nabla^2 \Phi(u^k)^{-1}(\nabla f(u^k) + A^\top c(u^k))\|_2^2 , \end{aligned}$$

which implies

$$\|\nabla^2 \Phi(u^k)^{-1}(\nabla f(u^k) + A^\top c(u^k))\|_2^2 \le \frac{2L_1}{\eta_{\mathcal{N}}} [f(u^k) - f(u^{k+1})] .$$

Summing over k of the above inequality leads to

$$\sum_k \|\nabla^2 \Phi(u^k)^{-1}(\nabla f(u^k) + A^\top c(u^k))\|_2^2 \le \frac{2L_1}{\eta_{\mathcal{N}}} [f(u^0) - f(u^*)] ,$$

and (3.5) readily follows. □

Comparing (3.8) to (3.3), one sees that the difference lies in the additional term $\mathcal{A}$, which leads to the different results in (3.2) and (3.4). While it is not obvious which one gives faster convergence, we would like to restate one advantage of mirror descent method in treating constraint $u \in Omega$, as mentioned in the introduction. In particular, for cases when Ω has some bound constraints, such as non-negativity, one can directly build it in Φ for (1.4) and the resulting solution is automatically bound preserving. Whereas in (1.7), there is no such a guarantee. See numerical examples in Figs. 4 and 5 for an evidence.

As a side note, we can extend the result in Theorem 4 to general quasi-Newton methods:

$$u^{k+1} = u^k - \eta B_k^{-1}(\nabla f(u^k) + A^\top c(u^k)) , \tag{3.11}$$

where B_k is a symmetric and positive definite matrix that approximates the Hessian as follows:

$$B_{k+1}(u^{k+1} - u^k) = \nabla f(u^{k+1}) - \nabla f(u^k) , \tag{3.12}$$

and $c(u^k)$ is again the to-be-determined vector that warrants $Au^{k+1} = b$.

Theorem 5 *(Global convergence for quasi-Newton (3.11)) Let $\{u^k\}$ be the solution to (3.11)–(3.12) with initial guess u_0. If B_k satisfies*

$$\|B_{k+1} B_k^{-1} - I\|_2 \le \eta L \tag{3.13}$$

for some constant L, which is independent of η and k, then we have

$$f\left(\frac{1}{K}\sum_{k=0}^{K-1}u^k\right)-f(u^*)\le\frac{1}{\eta K}D_f(u^*,u_0)+\frac{\eta}{K}\sum_{k=0}^{K-1}\left[\|\nabla f(u^k)+A^\top c(u^k)\|^2_{B_k^{-1}}+L\|\nabla f(u^k)+A^\top c(u^k)\|_2\|u^*-u^{k+1}\|_2\right]. \tag{3.14}$$

Proof Because there is no function Φ, we use the objective function f to define D_f and control the distance between current iteration and the optimal solution. More precisely, we consider

$$\begin{aligned}D_f(u^*,u^{k+1})-D_f(u^*,u^k)&=[\nabla f(u^k)-\nabla f(u^{k+1})]^\top(u^*-u^{k+1})-D_f(u^{k+1},u^k)\\&\le[B_{k+1}(u^k-u^{k+1})]^\top(u^*-u^{k+1})\\&=[B_k(u^k-u^{k+1})+(B_{k+1}-B_k)(u^k-u^{k+1})]^\top(u^*-u^{k+1})\end{aligned} \tag{3.15}$$

where the first inequality come from (3.12). Using (3.11), we see that $B_k(u^k-u^{k+1})=\eta(\nabla f(u^k)+A^\top c(u^k))$. Thus, (3.15) becomes

$$\begin{aligned}&D_f(u^*,u^{k+1})-D_f(u^*,u^k)\\&\le\eta(\nabla f(u^k)+A^\top c(u^k))^\top(u^*-u^k)+\eta(\nabla f(u^k)+A^\top c(u^k))^\top(u^k-u^{k+1})\\&\quad+[(B_{k+1}-B_k)(u^k-u^{k+1})]^\top(u^*-u^{k+1})\\&\le\eta[f(u^*)-f(u^k)]+\eta^2(\nabla f(u^k)+A^\top c(u^k))^\top B_k^{-1}(\nabla f(u^k)+A^\top c(u^k))\\&\quad+\eta^2L\|\nabla f(u^k)+A^\top c(u^k)\|_2\|u^*-u^{k+1}\|_2,\end{aligned}$$

where we have used the convexity of f and property of B_k in (3.13). Then (3.14) follows from summing the following inequality over k and dividing by K. □

3.2 Local Superlinear Convergence

In this section, we show the local convergence of (1.4). For notation brevity, we omit the subscript in η and simply write (1.4) as

$$\nabla\Phi(u^{k+1})-\nabla\Phi(u^k)=-\eta(\nabla f(u^k)+A^\top c(u^k)). \tag{3.16}$$

First we have the following proposition showing that if the iteration step η is properly chosen, the objective function sufficiently decreases along the flow. This mimics the first stage of Newton's method, where only linear convergence is obtained.

Proposition 1 *Assume Φ is 1-strongly convex with respect to the standard norm, i.e., $\nabla^2\Phi\succeq I$, and ∇f is $L-$Lipschitz. If η is chosen by $\eta\le\frac{2}{L}(1-\alpha)$ for $\alpha\in(0,0.5)$, then we have the following sufficient descent condition*

$$f(u^{k+1})\le f(u^k)+\alpha\nabla f(u^k)^\top(u^{k+1}-u^k). \tag{3.17}$$

Proof Define

$$\Delta u=u^{k+1}-u^k. \tag{3.18}$$

then we have

$$\begin{aligned} f(u^{k+1}) - f(u^k) &= \int_0^1 \nabla f(u^k + t\Delta u)^\top \Delta u \mathrm{d}t \\ &= \nabla f(u^k)^\top \Delta u + \int_0^1 [\nabla f(u^k + t\Delta u) - \nabla f(u^k)]^\top \Delta u \mathrm{d}t \\ &\leq \alpha \nabla f(u^k)^\top \Delta u + (1-\alpha)\nabla f(u^k)^\top \Delta u + \frac{L}{2}\|\Delta u\|^2 . \end{aligned} \tag{3.19}$$

From (3.16), one sees that

$$\begin{aligned} \nabla f(u^k)^\top \Delta u &= (\nabla f(u^k) + A^\top c(u^k))^\top \\ \Delta u &= -\frac{1}{\eta}(\nabla \Phi(u^{k+1}) - \nabla \Phi(u^k))^\top \Delta u \leq -\frac{1}{\eta}\|\Delta u\|^2 . \end{aligned} \tag{3.20}$$

Plugging it into (3.19), we have

$$f(u^{k+1}) - f(u^k) \leq \alpha \nabla f(u^k)^\top \Delta u + \left(\frac{L}{2} - \frac{1-\alpha}{\eta}\right)\|\Delta u\|^2 .$$

Then choosing $\eta \leq \frac{2}{L}(1-\alpha)$ makes $\frac{L}{2} - \frac{1-\alpha}{\eta} \leq 0$ and therefore (3.17) holds. □

Remark 2 1) (3.20) indicates that $f(u^{k+1})$ is indeed decreasing over iteration.
2) α can be in the range of (0, 1) for the above proposition to hold. However, for the later use in Lemma 1, we still restrict $\alpha \in (0, 0.5)$.

Next, we show that after sufficiently large number of iterations, the second stage of Newton type methods is reached and results in superlinear convergence.

Lemma 1 *Let Δu be defined in (3.18). If*

$$G_k := \int_0^1 \nabla^2 \Phi(u^k + s\Delta u)\mathrm{d}s \tag{3.21}$$

satisfies the Dennis-Moré condition [15]:

$$\frac{\|(G_k - \nabla^2 f(u^*))(u^{k+1} - u^k)\|}{\|u^{k+1} - u^k\|} \to 0, \quad \text{as } k \to \infty , \tag{3.22}$$

and $\nabla^2 f$ is L_2-Lipschitz around u^*, then $\eta = 1$ in (3.16) satisfies the sufficient descent condition (3.17) for sufficiently large k.

Proof We have

$$\begin{aligned} & f(u^{k+1}) - f(u^k) \\ &= \nabla f(u^k)^\top \Delta u + \Delta u^\top \int_0^1 (1-t)\nabla^2 f(u^k + t\Delta u)\mathrm{d}t \Delta u \\ &= \frac{1}{2}\nabla f(u^k)^\top \Delta u + \frac{1}{2}\Delta u^\top [\nabla \Phi(u^k) - \nabla \Phi(u^k + \Delta u)] \\ &\quad + \Delta u^\top \int_0^1 (1-t)\nabla^2 f(u^k + t\Delta u)\mathrm{d}t \Delta u \\ &= \frac{1}{2}\nabla f(u^k)^\top \Delta u - \frac{1}{2}\Delta u^\top \left[G_k - \nabla^2 f(u^*)\right]\Delta u + \Delta u^\top \int_0^1 (1-t) \\ &\quad \times \left[\nabla^2 f(u^k + t\Delta u) - \nabla^2 f(u^*)\right]\mathrm{d}t \Delta u . \end{aligned}$$

From (3.22), one has $\| \left[G_k - \nabla^2 f(u^*)\right] \Delta u\| \leq o(\|\Delta u\|)$. Therefore,

$$f(u^{k+1}) - f(u^k) \leq \frac{1}{2}\nabla f(u^k)^\top \Delta u + o(\|\Delta u\|^2) + \frac{L_2}{2}\|\Delta u\|^2 \|u^{k+1} - u^*\| .$$

For k sufficiently large, by global convergence Δu is sufficiently small, then the last two terms on the above inequality can be controlled by the first term on the right hand side, so the descent condition (3.17) will be satisfied. □

Once η is chosen to be 1, the superlinear convergence of (3.16) can be obtained.

Theorem 6 *(Local superlinear convergence of (3.16)) If G_k defined in (3.21) satisfies the Dennis-Moré condition (3.22), and assume that around x^*, f is β-strongly convex and $\nabla^2 f$ is L-Lipschitz. Then the mirror descent (3.16) converges superlinearly, i.e., $\|u^{k+1} - u^*\| \leq o(\|u^k - u^*\|)$.*

Proof From Lemma 1, the unit step length is allowed after sufficiently many iterations, and therefore we have

$$\nabla \Phi(u^{k+1}) - \nabla \Phi(u^k) = -(\nabla f(u^k) + A^\top c(u^k)) ,$$

which can be rewritten as

$$G_k(u^{k+1} - u^k) = -(\nabla f(u^k) + A^\top c(u^k)) ,$$

where G_k is defined in (3.21). Then we have

$$\begin{aligned}
&(G_k - \nabla^2 f(u^*))(u^{k+1} - u^k) \\
&\quad = -(\nabla f(u^k) + A^\top c(u^k)) - \nabla^2 f(u^*)(u^{k+1} - u^k) \\
&\quad = \nabla f(u^{k+1}) - \nabla f(u^k) - \nabla^2 f(u^*)(u^{k+1} - u^k) - (\nabla f(u^{k+1}) + A^\top c(u^k)). \qquad (3.23)
\end{aligned}$$

From the Lipschitz continuity of $\nabla^2 f$ around u^*, we have $\|\nabla f(u^{k+1}) - \nabla f(u^k) - \nabla^2 f(u^*)(u^{k+1} - u^k)\|/\|u^{k+1} - u^k\| \to 0$ as $k \to \infty$. Then using the Dennis-Moré condition (3.22), (3.23) implies

$$\lim_{k\to\infty} \frac{\|\nabla f(u^{k+1}) + A^\top c(u^k)\|}{\|u^{k+1} - u^k\|} = 0 ,$$

which readily leads to

$$\lim_{k\to\infty} \frac{\|\nabla f(u^{k+1}) + A^\top c(u^k) - \nabla f(u^*) - A^\top c(u^*)\|}{\|u^{k+1} - u^k\|} = 0 .$$

The above equation also implies

$$\lim_{k\to\infty} \frac{\langle \nabla f(u^{k+1}) + A^\top c(u^k) - \nabla f(u^*) - A^\top c(u^*), (u^{k+1} - u^*)/\|u^{k+1} - u^*\| \rangle}{\|u^{k+1} - u^k\|} = 0 .$$

Then using the fact that $Au^{k+1} = Au^*$, it reduces to

$$\lim_{k\to\infty} \frac{\langle \nabla f(u^{k+1}) - \nabla f(u^*), u^{k+1} - u^* \rangle}{\|u^{k+1} - u^k\| \|u^{k+1} - u^*\|} = 0 .$$

Since $\langle \nabla f(u^{k+1}) - \nabla f(u^*), u^{k+1} - u^* \rangle \geq \frac{\beta}{2}\|u^{k+1} - u^*\|^2$ for sufficiently large k and $\|u^{k+1} - u^k\| \leq \|u^{k+1} - u^*\| + \|u^k - u^*\|$, we have

$$\lim_{k\to\infty} \frac{\beta}{2} \frac{\|u^{k+1} - u^*\|}{\|u^{k+1} - u^*\| + \|u^k - u^*\|} = 0 ,$$

which implies $\lim_{k\to\infty} \frac{\|u^{k+1}-u^*\|}{\|u^k-u^*\|} = 0$. □

We also mention that in general (3.22) is not satisfied, so instead of having the superlinear convergence, we will have a linear convergence but with an increased rate as compared to the standard gradient descent. More specifically, we have the following theorem.

Theorem 7 *Let Φ be 1-strongly convex with respect to $\|\cdot\|_w$, and the sequence $\{u^k\}$ be obtained from (3.16). Also, assume that $D_f(u^*, u^k) \geq \mu D_\Phi(u^*, u^k)$. Then we have*

$$D_\Phi(u^*, u^{k+1}) \leq (1-\eta\mu) D_\Phi(u^*, u^k),$$

if $0 < \eta \leq 2(f(u^k) - f(u^*))/\|\nabla f(u^k)\|_{w,*}^2$.

Proof From the definition of Bregman divergence, we have

$$\begin{aligned}
&D_\Phi(u^*, u^{k+1}) - D_\Phi(u^*, u^k) \\
&\quad = [\Phi(u^*) - \Phi(u^{k+1}) - \langle \nabla\Phi(u^{k+1}), u^* - u^{k+1}\rangle] \\
&\qquad - [\Phi(u^*) - \Phi(u^k) - \langle \nabla\Phi(u^k), u^* - u^k\rangle] \\
&\quad = -[\Phi(u^{k+1}) - \Phi(u^k) - \langle \nabla\Phi(u^k), u^{k+1} - u^k\rangle] + \eta\langle \nabla f(u^k), u^* - u^{k+1}\rangle \\
&\quad \leq -\|u^k - u^{k+1}\|_w^2/2 + \eta\langle \nabla f(u^k), u^k - u^{k+1}\rangle - \eta(f(u^k) - f(u^*)) - \eta D_f(u^*, u^k) \\
&\quad \leq \eta^2\|\nabla f(u^k)\|_{w,*}^2/2 - \eta(f(u^k) - f(u^*)) - \eta\mu D_\Phi(u^*, u^k)
\end{aligned}$$

Therefore, if we choose $\eta \leq 2(f(u^k) - f(u^*))/\|\nabla f(u^k)\|_{w,*}^2$, then we have

$$D_\Phi(u^*, u^{k+1}) \leq (1-\eta\mu) D_\Phi(u^*, u^k).$$

The theorem is proved. □

This theorem is consistent with Theorem 2. When Φ is properly chosen, it will mitigate the ill-conditioning inherited from f in the sense that μ is increased, and therefore leads to a much improved rate of convergence.

4 Applications and Numerical Experiments

Apart from the examples mentioned in [28], we consider two additional applications of the mirror descent (1.4) in evolutionary PDEs: the Wasserstein gradient flow and Cahn-Hillard equation with degenerate mobility. In particular, viewing the Wasserstein gradient flow as a weighted H^{-1} gradient flow and using a minimizing movement scheme, we obtain an ill-conditioned optimization problem. The same problem is encountered in the Cahn-Hillard equation when the mobility is degenerate. In both cases, our mirror descent can provide preconditioning mechanisms while preserving the bounds of the solution (e.g., positivity) and mass conservation.

4.1 Wasserstein Gradient Flow

Let's consider the following Wasserstein gradient flow

$$\partial_t \rho(t,x) = -\nabla_{\mathcal{W}_2}\mathcal{E}(\rho(t,x)) := \nabla\cdot\left(\rho(t,x)\nabla\frac{\delta\mathcal{E}}{\delta\rho}(\rho(t,x))\right), \tag{4.1}$$

where $\mathcal{W}_2$ is the quadratic Wasserstein metric and δ denotes the first variation. Here $\rho(t,x)$ with $x \in \Omega \subset \mathbb{R}^n$ is the particle density function, and energy $\mathcal{E}(\rho(t,x))$ takes the form

$$\mathcal{E}(\rho(t,x)) = \int_\Omega [U(\rho(t,x)) + V(x)\rho(t,x)]\,\mathrm{d}x + \frac{1}{2}\int_{\Omega\times\Omega} W(x-y)\rho(t,x)\rho(t,y)\mathrm{d}x\mathrm{d}y\,.$$

The no-flux boundary condition $\rho(t,x)\nabla\frac{\delta\mathcal{E}}{\delta\rho}\cdot\hat{n}\big|_{\partial\Omega} = 0$ is imposed to ensure the mass conservation. This equation has diverse applications in physics and biology, such as granular materials [12], chemotaxis [20], animal swarming [4, 11], and many others.

Numerically solving (4.1) has been quite challenging to satisfy three desired properties: non-negativity, mass conservation, and energy dissipation. Besides the Eulerian and Lagrangian methods that have been developed in the literature, we particularly mention the variational approach following the seminal JKO scheme by Jordan, Kinderlehrer, and Otto [19]. Given a time step $\tau > 0$, the JKO scheme recursively defines a sequence $\rho_n(x)$ via a minimizing movement approach. This approach has revolutionized PDE analysis, whereas its impact in numerics has only be revealed recently with the aid of modern optimization algorithms [7, 9, 13, 18, 23, 25].

In this paper, we consider a similar but slightly different approach. In particular, we obtain the solution sequence $\{\rho_n\}$, an approximation to the exact solution $\rho_n(x) \approx \rho(n\tau, x)$ as follows:

$$\rho_0 = \rho_{\text{in}}, \quad \rho_{n+1} = \arg\min_{\rho\in\mathbb{K}}\left\{\frac{1}{2\tau}\|\rho-\rho_n\|^2_{\Delta^{-1}_{\rho_n}} + \mathcal{E}(\rho)\right\} := \arg\min_{\rho\in\mathbb{K}} f(\rho)\,, \tag{4.2}$$

where $\|u\|^2_{\Delta^{-1}_{\rho_n}} = \int_\Omega u(x)\Delta^{-1}_{\rho_n}u(x)\mathrm{d}x$, and $\mathbb{K} = \left\{\rho : \rho\in\mathcal{P}(\Omega),\ \int_\Omega |x|^2\rho\,\mathrm{d}x < +\infty\right\}$, where $\mathcal{P}$ is the set of probability measure. Here Δ_{ρ_n} is the negative weighted Laplacian $\Delta_{\rho_n} = -\nabla\cdot(\rho_n\nabla)$, and $\Delta^{-1}_{\rho_n}$ is its pseudo-inverse. It has been shown that the weighted H^{-1} norm is a first order approximation to the Wasserstein distance [27], and therefore will not violate the first order accuracy of the JKO scheme [3]. In view of (4.2), one sees that the three desired properties mentioned above are all satisfied. Indeed, the positivity and mass conservation are obtained by requiring the minimizer in $\mathbb{K}$, the energy dissipation $\mathcal{E}(\rho_{n+1}) \le \mathcal{E}(\rho_n)$ is also immediate since ρ_{n+1} is the minimizer.

4.1.1 Mirror Descent Algorithm

To solve the optimization problem (4.2), a direct projected gradient descent takes the following form

$$\rho^{k+1} = \text{proj}_{\mathbb{K}}\left\{\rho^k - \eta\left[\frac{1}{\tau}\Delta^{-1}_{\rho_n}(\rho^k-\rho_n) + \frac{\delta\mathcal{E}}{\delta\rho}(\rho^k)\right]\right\}. \tag{4.3}$$

where η is the iteration stepsize and the superscript k, which shall not be confused with the subscript n, denotes the iteration index. Since the value of ρ_n can be arbitrarily close to zero, Δ_{ρ_n} is very stiff, and therefore the gradient descent (4.3) will take extremely long time to converge. To this end, we propose the following mirror descent algorithm.

Choosing Φ in (1.4) to be

$$\Phi(\rho) = \frac{1}{2\tau}\|\rho-\rho_n\|^2_{\Delta^{-1}_{\rho_n}} + \varepsilon\int\rho\log\rho\mathrm{d}x\,, \tag{4.4}$$

then the mirror descent reads

$$\frac{\delta\Phi}{\delta\rho}(\rho^{k+1}) - \frac{\delta\Phi}{\delta\rho}(\rho^k) = -\eta\left[\frac{1}{\tau}\Delta^{-1}_{\rho_n}(\rho^k-\rho_n) + \frac{\delta\mathcal{E}}{\delta\rho}(\rho^k)\right],$$

which simplifies to

$$\rho^{k+1} + \varepsilon\tau\Delta_{\rho_n}\log\rho^{k+1} = \rho^k + \varepsilon\tau\Delta_{\rho_n}\log\rho^k - \eta\left[\rho^k - \rho_n + \tau\Delta_{\rho_n}\frac{\delta\mathcal{E}}{\delta\rho}(\rho^k)\right]. \quad (4.5)$$

The reason for choosing Φ in the form of (4.4) is two-fold. On one hand, the stiffness introduced by $\Delta_{\rho_n}^{-1}$ constitutes the major difficulty in the optimization task in (4.2), and therefore including this term in Φ will significantly mitigate the stiffness. On the other hand, thanks to the additional entropy term in (4.4), the positivity of ρ is preserved in (4.5). Moreover, since Δ_{ρ_n} preserves mass, i.e., $\int \Delta_{\rho_n} u(x)\mathrm{d}x = 0$, mass conservation is also guaranteed in (4.5), that is, $\int \rho^{k+1}(x)\mathrm{d}x = \int \rho^k(x)\mathrm{d}x = \int \rho_n(x)\mathrm{d}x$.

In practice, we will further discretize (4.5) in space. Let us consider one dimension for instance. Denote $[x_L, x_R]$ as the computational domain and Δx the size of the spatial grid. Choose $x_j = x_L + (j - \frac{1}{2})\Delta x$, and denote

$$\rho_j \approx \rho(x_j), \quad \rho_{n,j} \approx \rho(t_n, x_j), \quad 1 \le j \le N_x,\ n \in \mathbb{N}_+,$$

where $t_n = n\tau$, and $N_x\Delta x = x_R - x_L$. First we discretize Δ_{ρ_n}, and denote its discrete counterpart as D_{ρ_n}. Then we propose

$$(\mathsf{D}_{\rho_n}u)_j = \begin{cases} -\frac{1}{\Delta x^2}\left[\frac{\rho_{n,j}+\rho_{n,j+1}}{2}u_{j+1} - \frac{\rho_{n,j+1}+\rho_{n,j}}{2}u_j\right], & j = 1 \\ -\frac{1}{\Delta x^2}\left[\frac{\rho_{n,j}+\rho_{n,j+1}}{2}u_{j+1} - \frac{\rho_{n,j+1}+2\rho_{n,j}+\rho_{n,j-1}}{2}u_j\right. & \\ \left. +\frac{\rho_{n,j}+\rho_{n,j-1}}{2}u_{j-1}\right], & 2 \le j \le N_x - 1 \\ -\frac{1}{\Delta x^2}\left[-\frac{\rho_{n,j}+\rho_{n,j-1}}{2}u_j + \frac{\rho_{n,j}+\rho_{n,j-1}}{2}u_{j-1}\right], & j = N_x \end{cases}. \quad (4.6)$$

Note specifically that for $j = 1$ and $j = N_x$, our discretization takes into account the boundary condition in (4.1). Indeed, if either $\rho\big|_{\partial\Omega} = 0$ or $u\cdot\hat{n}\big|_{\partial\Omega} = 0$, then at discrete level on the left boundary, we have $\rho_{n,0} + \rho_{n,1} = 0$ or $u_1 - u_0 = 0$ correspondingly, and in either the second line of (4.6) reduces to the first line. Same arguments applies to the right boundary. As a result, D_{ρ_n} preserves the mass, i.e., $1^\top\mathsf{D}_{\rho_n} = 0$.

Denote

$$\boldsymbol{\rho} = (\rho_1, \rho_2, \cdots, \rho_{N_x})^\top, \quad \boldsymbol{\rho}_n = (\rho_{n,1}, \rho_{n,2}, \cdots, \rho_{n,N_x})^\top,$$

we can rewrite (4.5) in the discrete form

$$\boldsymbol{\rho}^{k+1} + \varepsilon\tau\mathsf{D}_{\rho_n}\log\boldsymbol{\rho}^{k+1} = \boldsymbol{\rho}^k + \varepsilon\tau\mathsf{D}_{\rho_n}\log\boldsymbol{\rho}^k - \eta\left[\boldsymbol{\rho}^k - \boldsymbol{\rho}_n + \tau\mathsf{D}_{\rho_n}\frac{\delta\mathcal{E}}{\delta\rho}(\boldsymbol{\rho}^k)\right]. \quad (4.7)$$

Then the remaining task is to solve the nonlinear equations for $\boldsymbol{\rho}^{k+1}$, for which we use the Newton's method. Let $\boldsymbol{y} = \log\boldsymbol{\rho}$ and define $\boldsymbol{h}$ to be $\boldsymbol{h}(\boldsymbol{\rho}) = e^{\boldsymbol{y}} + \varepsilon\tau\mathsf{D}_{\rho_n}\boldsymbol{y} - \boldsymbol{b}$, where $\boldsymbol{b} = \boldsymbol{\rho}^k + \varepsilon\tau\mathsf{D}_{\rho_n}\log\boldsymbol{\rho}^k - \eta[\boldsymbol{\rho}^k - \boldsymbol{\rho}_n + \tau\mathsf{D}_{\rho_n}\delta\mathcal{E}(\boldsymbol{\rho}^k)]$. Then the Newton's method takes the form

$$\boldsymbol{y}^{(l+1)} = \boldsymbol{y}^{(l)} - \mathsf{A}_l^{-1}\boldsymbol{h}(\boldsymbol{y}^{(l)}), \quad (4.8)$$

where $\mathsf{A}_l = \mathrm{diag}[e^{\boldsymbol{y}^{(l)}}] + \varepsilon\tau\mathsf{D}_{\rho_n}$. Note that since the components of $\boldsymbol{y}$ can vary drastically, A will be ill-conditioned, and therefore the computation $\mathsf{A}_l^{-1}\boldsymbol{h}(\boldsymbol{y}^{(l)})$ in (4.8) may be susceptible to errors. To fix this issue, we propose the following preconditioner $\mathsf{P}_l = \mathrm{diag}[e^{-\boldsymbol{y}^{(l)}}]$, and rewrite (4.8) into

$$\boldsymbol{y}^{(l+1)} = \boldsymbol{y}^{(l)} - (\mathsf{P}_l\mathsf{A}_l)^{-1}[\mathsf{P}_l\boldsymbol{h}(\boldsymbol{y}^{(l)})]. \quad (4.9)$$

Note that since P_l is a diagonal matrix, the preconditioner is cheap to apply.

In summary, we have the following algorithms.

Algorithm 1: Mirror descent algorithm for (4.2)

Input $\boldsymbol{\rho}_n$,D_{ρ_n}, $\text{Iter}_{\max}$, Δx, τ
Output $\boldsymbol{\rho}_{n+1}$
$\boldsymbol{\rho}^0 = \boldsymbol{\rho}_n$, $\boldsymbol{y}^0 = \log \boldsymbol{\rho}^0$
$k = 0$
while $k \leq Iter_{\max}$ *and stopping criteria is not achieved* **do**
  $\boldsymbol{y} = \boldsymbol{y}^k$,
  while *error* $> 1e-6$ **do**
    $\mathsf{A} = \text{diag}[e^{\boldsymbol{y}}] + \varepsilon\tau\mathsf{D}_{\rho_n}$, $\mathsf{P} = \text{diag}[e^{-\boldsymbol{y}}]$
    $\tilde{\boldsymbol{y}} = \boldsymbol{y} - (\mathsf{PA})^{-1}[\mathsf{P}\boldsymbol{h}(\boldsymbol{y})]$
    error $= \|\tilde{\boldsymbol{y}} - \boldsymbol{y}\|/\|\boldsymbol{y}\|$
    $\boldsymbol{y} = \tilde{\boldsymbol{y}}$
  end
  $k = k+1$, $\boldsymbol{y}^{k+1} = \tilde{\boldsymbol{y}}$
end
$\boldsymbol{\rho}_{n+1} = e^{\boldsymbol{y}^{k+1}}$

Algorithm 2: Variational scheme for (4.1)

Input $\boldsymbol{\rho}_0 = \boldsymbol{\rho}_{\text{in}}$
Output $\boldsymbol{\rho}_n$ for $1 \leq n \leq N_t$
for $n <= N_t$ **do**
  Apply Algorithm 1 to $\boldsymbol{\rho}_n$ to get $\boldsymbol{\rho}_{n+1}$
end

4.2 Numerical Examples

We consider two examples of (4.1) and demonstrate the efficiency of mirror descent. In both examples, we stop Algorithm 1 when the relative error is less than a preset tolerance, i.e.,

$$\frac{\|\boldsymbol{\rho}^{k+1} - \boldsymbol{\rho}^k\|}{\|\boldsymbol{\rho}^k\|} \leq \text{Tol}, \qquad where \quad \boldsymbol{\rho}^k = \exp^{\mathbf{v}^k}. \tag{4.10}$$

4.2.1 Porous Medium Equation

We first consider the porous medium equation

$$\partial_t \rho = \Delta \rho^m, \quad m > 1, \tag{4.11}$$

which can be seen as the Wasserstein gradient flow of $\mathcal{E}(\rho) = \int \frac{1}{m-1}\rho^m \mathrm{d}x$. A well-known family of exact solutions is given by the Barenblatt profiles, which are densities of the form

$$\rho(x,t) = (t+t_0)^{-\frac{1}{m+1}}\left(C - \alpha\frac{m-1}{2m(m+1)}x^2(t+t_0)^{-\frac{2}{m+1}}\right)_+^{\frac{1}{m-1}}, \quad \text{for } C, t_0 > 0. \tag{4.12}$$

In our tests, we choose $m = 2$, $t_0 = 10^{-3}$, and $C = 0.8$. The results using our mirror descent algorithm are gathered in Fig. 1. On the left, the numerical solutions are compared to the

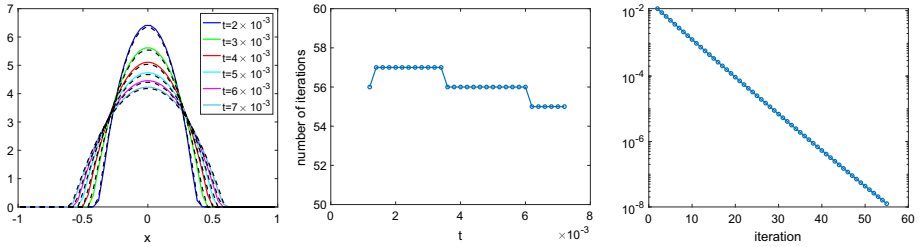

Fig. 1 Porous medium equation with $m = 2$ and computational domain $[-1, 1]$. Left: evolution of ρ compared with exact solution. Middle: number of iterations needed in Algorithm 1 within each outer time step. The tolerance in (4.10) is chosen to be Tol=10^{-8}. Right: iteration error of Algorithm 1 in the first step of outer variational scheme. Numerical parameters are $\Delta x = 0.04$, $\tau = 2 \times 10^{-4}$, $\varepsilon = 0.005$ in (4.4) and $\eta = 0.2$ in Algorithm 1

analytical formulas, and good agreement is demonstrated. In the middle, we have shown the number of iterations needed in Algorithm 1 within each outer time step, and one sees that around the same iterations are needed for a given tolerance Tol= 10^{-8}. On the right, we plot the relative error (4.10) versus the iteration in the first outer time step. The decay of relative error behaves quite similar at later times.

It is interesting to mention that we have also implemented the variable metric method (1.7), or equivalently (1.8). In particular, we choose $\Phi(\rho) = \frac{1}{2\tau}\|\rho - \rho_n\|_{\Delta_{\rho_n}^{-1}}$ instead of (4.4), then (1.7) becomes

$$\rho^{k+1} = \rho^k - \eta \Delta_{\rho_n}(\nabla f(\rho^k) + A^\top c(\rho^k)), \tag{4.13}$$

where $f(\rho)$ is defined in (4.2), $\mathcal{E}(\rho) = \int \frac{1}{m-1}\rho^m \mathrm{d}x$, and A is an all one row vector that encodes the mass conservation of ρ. As explained before, Δ_{ρ_n} preserves the mass exactly, therefore $c(\rho^k) \equiv 0$ here and the algorithm reduces to

$$\rho^{k+1} = \rho^k - \eta(\rho^k - \rho_n) - \eta \Delta_{\rho_n} \frac{\delta \mathcal{E}}{\delta \rho}(\rho^k).$$

Compare it to (4.5), we see that the major difference is that here there is no mechanism to automatically guarantee the positivity. However, with a proper choice of iteration step η, the positivity may still be preserved. In this specific example, we apply (4.13) with the same parameters as in the mirror descent algorithm, and we have obtained exactly the same behavior of the solution as displayed in Fig. 1, so the plots are omitted.

4.2.2 Aggregation Equation

Next we consider a nonlocal aggregation equation of the form

$$\partial_t \rho = \nabla \cdot (\rho \nabla W * \rho), \quad W(x) = \frac{|x|^2}{2} - \ln(|x|), \tag{4.14}$$

where the interaction kernel W is repulsive at short length scales and attractive at longer distances. This equation admits a unique equilibrium profile

$$\rho_\infty(x) = \frac{1}{\pi}\sqrt{(2 - x^2)_+}. \tag{4.15}$$

In practice, to avoid evaluation of $W(x)$ at $x = 0$, we set $W(0)$ to equal the average value of W on the cell of width $2h$ centered at 0, i.e., $W(0) = \frac{1}{2h}\int_{-h}^{h} W(x)\mathrm{d}x$, where we compute

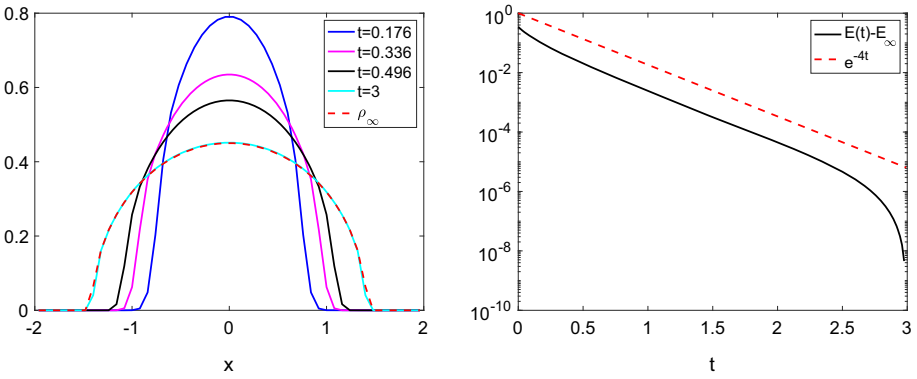

Fig. 2 Aggregation equation with computational domain $[-2, 2]$ and initial data . Left: evolution of ρ, ρ_∞ is given by the analytical formula (4.15). Right: exponential decay of energy. Numerical parameters are $\Delta x = 0.08$, $\tau = 0.016$, $\varepsilon = 0.1$ in (4.4) and $\eta = 0.8$ in Algorithm 1

this value analytically. (See also [9] for a similar treatment.) In Fig. 2, we compute (4.14) with initial data

$$\rho(x, 0) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} + 10^{-8}.$$

The left picture displays the evolution of ρ. At $t = 3$, the solution has reached the steady state, which matches the analytical formula represented by the dashed curve. The right plot shows the exponential decay of the energy, where the red dashed line indicates the decay rate. We also explore the convergence of our algorithm in Fig. 3. As seen in the upper left picture, the number of iterations needed in reaching the tolerance has shown some heterogeneity with respect to the outer time. More specifically, at a few times, such as $t = 0.528$, a significantly larger number of iterations is needed. More detailed plots on how the relative error (4.10) evolves are displayed in the upper right and lower left figures, in which a few representative plots of the error are given. At $t = 0.528$, which corresponding to the first peak, we also plot the solution ρ at this time and the previous time (i.e., $t = 0.512$), with a zoom-in plot near the left propagating front of the solution. It is shown that, at the location $x = -1.24$, there is a sharp transition in the solution. That is, ρ_{10} goes from 5.5996×10^{-7} to 2.1726×10^{-4}, which results in around 389 increase in magnitude. Similar increase are observed at time corresponding to the rest two peaks. So we believe that the deterioration in the convergence is due to such a rapid transition in the solution.

In comparison, we also considered the variable metric algorithm (1.7). With the choice of $\Phi(\rho) = \frac{1}{2\tau}\|\rho - \rho_n\|_{\Delta_{\rho_n}^{-1}}$, and the algorithm takes the same form as (4.13), but with $\mathcal{E}(\rho) = \frac{1}{2}\int \rho(W * \rho)\mathrm{d}x$. The evolution of ρ is given in Fig. 4. Here we choose a smaller iteration step $\eta = 0.01$, but the positivity of the solution can still not be preserved, and oscillation round zero values of ρ is generated and amplified along time (compare ρ at $t = 1.616$ with $t = 3$).

4.3 Cahn-Hillard Equation with Degenerate Mobility

Cahn-Hillard equation has first been introduced to study phase separation in binary alloys, and later extended to many other fields such as image inpainting [8] and math biology [17].

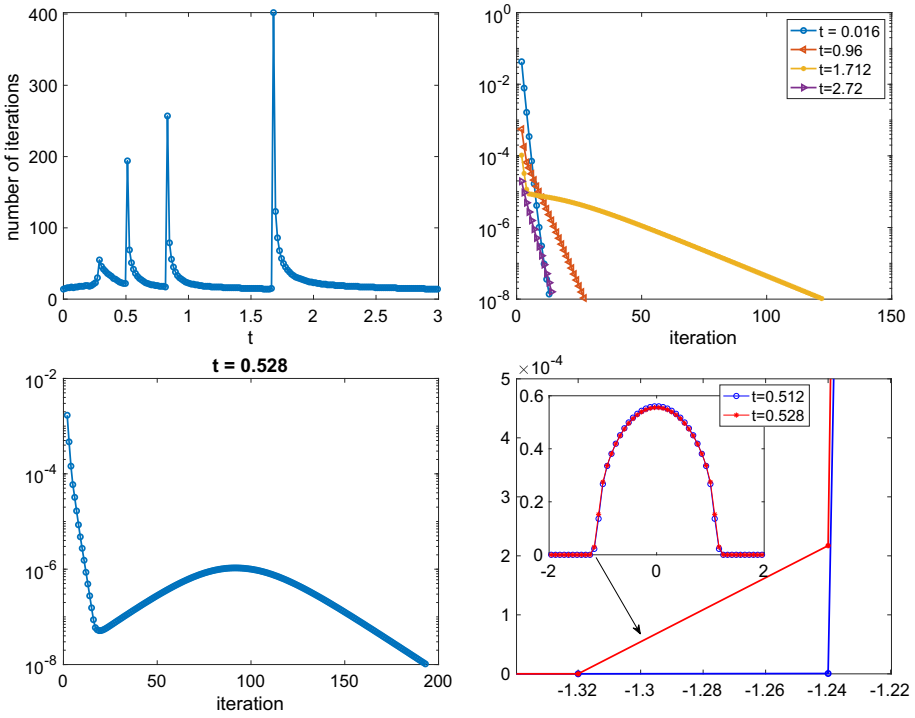

Fig. 3 Convergence in the aggregation equation. Upper left: number of iterations needed within each outer time step. Upper right: several typical plot of the convergence (e.g., relative error versus iteration) at different outer times. Lower left: a hard to converge scenario at $t = 0.528$. Lower right: plot of the corresponding solution at two consecutive times $t = 0.512$ and $t = 0.528$

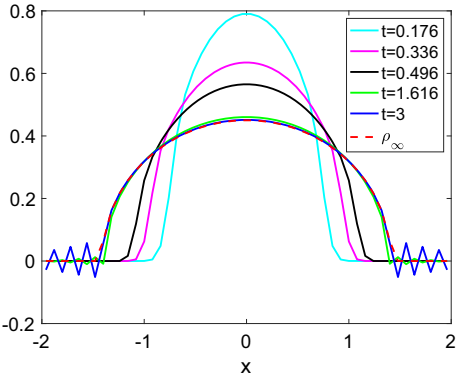

Fig. 4 Computation of aggregation equation using variable metric algorithm (4.13) with $\eta = 0.01$, $\Delta x = 0.08$, $\tau = 0.016$

To put it on the same foot as (4.1), we write it in the gradient flow form:

$$\partial_t u = \nabla \cdot \left(M(u) \nabla \frac{\delta \mathcal{E}}{\delta u} \right), \tag{4.16}$$

where u represents the difference in the local concentration of two components in the alloy, and $M(u) = 1 - u^2 \geq 0$ is a diffusional mobility. $\mathcal{E}$ is the energy functional

$$\mathcal{E}(u) = \int_\Omega (\frac{\alpha^2}{2}|\nabla u|^2 + \Psi(u))\mathrm{d}x\,, \tag{4.17}$$

where the first term penalizes large gradients and models the capillary effects. The second term is the homogeneous free energy. A typical form is the Ginzburg-Landau potential $\Psi(u) = \frac{1}{4}(1-u^2)^2$ or logarithmic potential $\Psi(u) = \frac{\theta}{2}\left[(1+u)\log\left(\frac{1+u}{2}\right) + (1-u)\log\left(\frac{1-u}{2}\right)\right] + \frac{\theta_c}{2}(1-u^2)$ for $u \in (-1, 1)$, where $\theta < \theta_c$ are two positive constants.

As before, we solve (4.16) using the minimizing movement scheme. More precisely, we obtain u_{n+1} by solving

$$u_0 = u_{\text{in}}, \quad u_{n+1} = \arg\min_{u\in\mathbb{U}} \left\{ \frac{1}{2\tau}\|u-u_n\|^2_{\Delta^{-1}_{M(u_n)}} + \mathcal{E}(u) \right\}, \tag{4.18}$$

where $\mathbb{U} = \{\int_\Omega u(x)\mathrm{d}x = \int_\Omega u_{\text{in}}(x)\mathrm{d}x, -1 \le u \le 1\}$, and $\|f\|^2_{\Delta^{-1}_{M(u_n)}} = \int_{\mathbb{R}^d} f(x)\Delta^{-1}_{M(u_n)} f(x)\mathrm{d}x$. Here $\Delta_{M(u_n)}$ is the negative weighted Laplacian $\Delta_{M(u_n)} = -\nabla\cdot(M(u_n)\nabla)$, and $\Delta^{-1}_{M(u_n)}$ is the pseudo-inverse of $\Delta_{M(u_n)}$.

4.3.1 Mirror Descent Algorithm

It has been proven that u will stay within the interval $(-1, 1)$ due either to the singularity in the free energy or degeneracy of the mobility [1, 16]. In order to maintain such a bound, we choose Φ in (1.4) to be:

$$\Phi(u) = \frac{1}{2\tau}\|u-u_n\|^2_{\Delta^{-1}_{M(u_n)}} + \varepsilon_1\int (u+1)\log(u+1)\mathrm{d}x + \varepsilon_2\int (1-u)\log(1-u)\mathrm{d}x\,, \tag{4.19}$$

then the mirror descent becomes

$$\frac{\delta\Phi}{\delta u}(u^{k+1}) - \frac{\delta\Phi}{\delta u}(u^k) = -\eta\left[\frac{1}{\tau}\Delta^{-1}_{M(u_n)}(u^k - u_n) + \frac{\delta\mathcal{E}}{\delta u}(u^k)\right],$$

which simplifies to

$$\begin{aligned} &u^{k+1} + \tau\Delta_{M(u_n)}[\varepsilon_1\log(1+u^{k+1}) + \varepsilon_2\log(1-u^{k+1})] \\ &\quad = u^k + \tau\Delta_{M(u_n)}[\varepsilon_1\log(1+u^k) - \varepsilon_2\log(1-u^k)] - \eta\left[u^k - u_n + \tau\Delta_{M(u_n)}\frac{\delta\mathcal{E}}{\delta u}(u^k)\right]. \end{aligned} \tag{4.20}$$

The discretization of $\Delta_{M(u_n)}$ is the same as in (4.6) except that one replace ρ_n by $M(u_n)$. In solving (4.20) for u^{k+1}, Newton's method will be used and a similar preconditioner as in (4.9) wil be employed. We omit the details as they are very similar to Sect. 4.1.1.

4.3.2 An Example

Here we consider a one dimensional example in [5]. Choose $\Psi(u) = \frac{1}{2}(1-u^2)$, $\alpha = 0.1$ in (4.17) and let initial condition be

$$u_{\text{in}}(x) = \begin{cases} \cos\left(\frac{x-\frac{1}{2}}{\alpha}\right) - 1\,, & \text{if } |x-\frac{1}{2}| \le \frac{\pi\alpha}{2} \\ -1\,, & \text{other} \end{cases}.$$

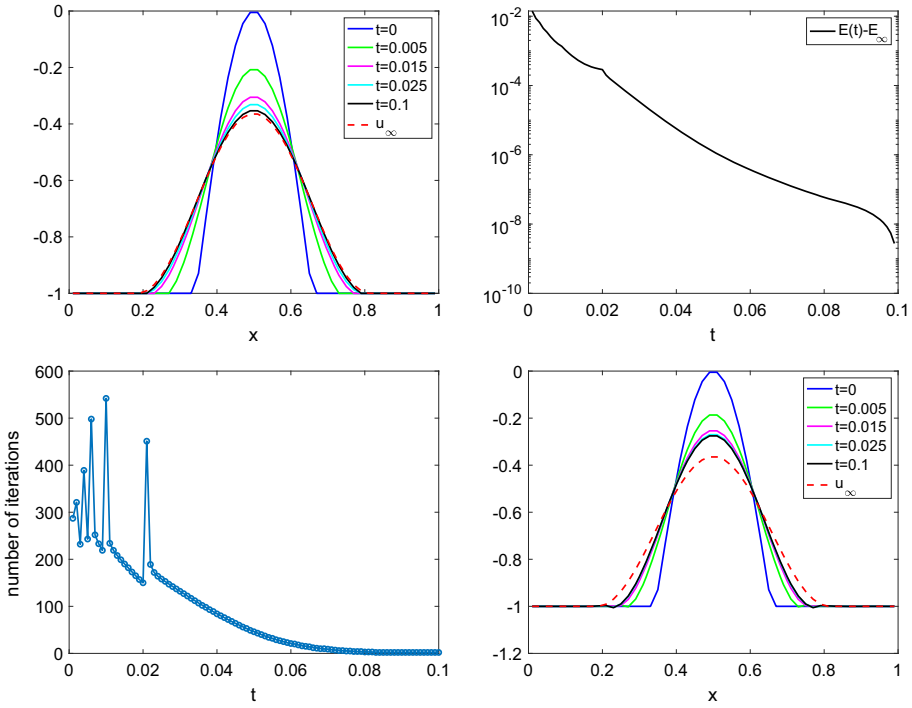

Fig. 5 Cahn-Hillard equation with degenerate mobility. Upper left: evolution of u using mirror descent algorithm. The red dashed curve is given by (4.21). Upper right: exponential decay of the relative energy. Lower left: number of iterations needed within each outer time step. Lower right: evolution of u using variable metric algorithm. Numerical parameters are $\Delta x = 0.02$, $\tau = 10^{-3}$, $\varepsilon = 0.5$, $\eta = 0.02$ (mirror descent), and $\eta = 0.00025$ (variable metric)

Then the steady state takes the form

$$u_\infty(x) = \begin{cases} \frac{1}{\pi}\left[1 + \cos\left(\frac{\pi - \frac{1}{2}}{\alpha}\right)\right] - 1 & \text{if } |x - \frac{1}{2}| \le \pi\alpha \\ -1 & \text{other} \end{cases} . \tag{4.21}$$

The results are collected in Fig. 5. The upper figures show the evolution of the density ρ and decay of the energy. The lower left figure displays the number of iterations within each outer time steps, and the spikes again correspond to the rapid transition of the solution near -1. All three figures are obtained via the mirror descent algorithm. On the other hand, we implemented the variable metric algorithm with $\Phi(u) = \frac{1}{2\tau}\frac{1}{2\tau}\|u - u_n\|^2_{\Delta^{-1}_{M(u_n)}}$ and the profile of u is given in the lower right plot of Fig. 5. Here with a much smaller choice of iteration step, i.e., $\eta = 0.00025$ as compared to $\eta = 0.02$ in mirror descent, the lower bound of u is still violated, and results in a wrong steady state.

5 Conclusion

In this paper, we consider a mirror descent algorithm, where the metric is induced by a convex function, whose Hessian is an approximation of the Hessian of the objective function. The

advantage of this algorithm is two-fold. On one hand, the mirror descent framework gives a natural way to incorporate the bound constraint of the solution. On the other hand, the Hessian information used in building the metric leads to improved rate of convergence. To put such an advantage on a rigorous footing, we first formulate a gradient flow of the algorithm, in which the constraints are incorporated as a to-be-determined vector. From this formulation, we can draw connection between the mirror descent and more general variable metric algorithms. Then the improved rate of convergence is proved following the two stage approach in Newton type methods. In return, the proof we obtained for the mirror descent can lend itself to quasi-Newton methods to show the global convergence. We also apply the algorithm to two cases, the Wasserstein gradient flow and Cahn-Hillard equation with degenerate mobility, and demonstrate its effectiveness.

Funding The authors have not disclosed any funding.

Data Availability Enquiries about data availability should be directed to the authors.

Declarations

Competing interests The authors have not disclosed any competing interests.

References

1. Abels, H., Wilke, M.: Convergence to equilibrium for the Cahn-Hilliard equation with a logarithmic free energy. Nonlinear Analysis: Theory, Methods & Applications **67**, 3176–3193 (2007)
2. Agueh, M.: Local existence of weak solutions to kinetic models of granular media. Arch. Ration. Mech. Anal. **221**, 917–959 (2016)
3. Ambrosio, L., Gigli, N., Savaré, G.: Gradient Flows: in Metric Spaces and in the Space of Probability Measures. Springer Science & Business Media (2008)
4. Barbaro, A.B.T., Cañizo, J.A., Carrillo, J.A., Degond, P.: Phase transitions in a kinetic flocking model of Cucker-Smale type. Multiscale Model. Simul. **14**, 1063–1088 (2016)
5. Barrett, J.W., Blowey, J.F., Garcke, H.: Finite element approximation of the Cahn-Hilliard equation with degenerate mobility. SIAM J. Numer. Anal. **37**, 286–318 (1999)
6. Beck, A., Teboulle, M.: Mirror descent and nonlinear projected subgradient methods for convex optimization. Oper. Res. Lett. **31**, 167–175 (2003)
7. Benamou, J.-D., Carlier, G., Laborde, M.: An augmented Lagrangian approach to Wasserstein gradient flows and applications. ESAIM: PROCEEDINGS AND SURVEYS **54**, 1–17 (2016)
8. Bertozzi, A.L., Esedoglu, S., Gillette, A.: Inpainting of binary images using the Cahn-Hilliard equation. IEEE Trans. Image Process. **16**, 285–291 (2006)
9. Carrillo, J.A., Craig, K., Wang, L., Wei, C.: Primal dual methods for Wasserstein gradient flows. Found. Comut. Math. **22**, 289–443 (2022)
10. Carrillo, J.A., Craig, K., Yao, Y.: Aggregation-diffusion Equations: Dynamics, Asymptotics, and Singular limits, in Active Particles, Volume 2, Springer, p 65–108 (2019)
11. Carrillo, J.A., Fornasier, M., Toscani, G., Vecil, F.: Particle, kinetic, and hydrodynamic models of swarming, Modeling and Simulation in Science, Engineering and Technology, p 297–336 (2010)
12. Carrillo, J.A., McCann, R., Villani, C.: Kinetic equilibration rates for granular media and related equations: entropy dissipation and mass transportation estimates. Revista Matematica Iberoamericana **19**, 971–1018 (2003)
13. Chizat, L., Peyré, G., Schmitzer, B., Vialard, F.-X.: Scaling algorithms for unbalanced optimal transport problems. Math. Comput. **87**, 2563–2609 (2018)
14. Chouzenoux, E., Pesquet, J.-C., Repetti, A.: Variable metric forward-backward algorithm for minimizing the sum of a differentiable function and a convex function. J. Optim. Theory Appl. **162**, 107–132 (2014)
15. Dennis, J.E., Moré, J.J.: A characterization of superlinear convergence and its application to quasi-Newton methods. Math. Comput. **28**, 549–560 (1974)
16. Elliott, C.M., Garcke, H.: On the Cahn-Hilliard equation with degenerate mobility. SIAM J. Math. Anal. **27**, 404–423 (1996)

17. Garcke, H., Lam, K.F., Nürnberg, R., Sitka, E.: A multiphase Cahn-Hilliard-Darcy model for tumour growth with necrosis. Math. Models Methods Appl. Sci. **28**, 525–577 (2018)
18. Jacobs, M., Lee, W., Léger, F.: The back-and-forth method for Wasserstein gradient flows, arXiv preprint arXiv:2011.08151, (2020)
19. Jordan, R., Kinderlehrer, D., Otto, F.: The variational formulation of the Fokker-Plank equation. SIAM J. Math. Anal. **29**, 1–17 (1998)
20. Keller, E., Segel, L.: Traveling bands of chemotactic bacteria: a theoretical analysis. J. Theoret. Biol. **30**, 6420–6437 (1971)
21. Krichene, W., Bayen, A., Bartlett, P.L.: Accelerated mirror descent in continuous and discrete time, Advances in neural information processing systems, 28 (2015)
22. Lee, J.D., Sun, Y., Saunders, M.A.: Proximal Newton-type methods for minimizing composite functions. SIAM J. Optim. **24**, 1420–1443 (2014)
23. Li, W., Lu, J., Wang, L.: Fisher information regularization schemes for Wasserstein gradient flows. J. Comput. Phys. **416**, 109449 (2020)
24. Mei, S., Montanari, A., Nguyen, P.-M.: A mean field view of the landscape of two-layer neural networks. Proc. Natl. Acad. Sci. **115**, E7665–E7671 (2018)
25. Peyré, G.: Entropic approximation of Wasserstein gradient flows. SIAM J. Imag. Sci. **8**, 2323–2351 (2015)
26. Topaz, C., Bertozzi, A., Lewis, M.: A nonlocal continuum model for biological aggregation. Bull. Math. Bio. **68**, 1601–1623 (2006)
27. Villani, C.: Topics in Optimal Transportation, American Mathematical Soc. **58**, (2003)
28. Ying, L.: Mirror descent algorithms for minimizing interacting free energy. J. Sci. Comput. **84**, 1–14 (2020)