INDUSTRIAL AND SYSTEMS ENGINEERING

ISE

# Momentum-Based Minimization of the Ginzburg-Landau Functional on Euclidean Spaces and Graphs

OLUWATOSIN AKANDE[1], PATRICK DONDL[2], KANAN GUPTA[3], AKWUM ONWUNTA[1], AND STEPHAN WOJTOWYTSCH[4]

[1]Department of Industrial and Systems Engineering, Lehigh University, Bethlehem, PA, 18015 USA

[2]Abteilung f¨ur Angewandte Mathematik, Albert-Ludwigs-Universit¨at Freiburg, Hermann-Herder-Straße 10, 79104 Freiburg i. Br., Germany

[3]University of Pittsburgh, Department of Mathematics, Thackeray Hall, Pittsburgh, PA 15213, USA

[4]University of Pittsburgh, Department of Mathematics, Thackeray Hall, Pittsburgh, PA 15213, USA

ISE Technical Report 24T-023

LEHIGH UNIVERSITY

# MOMENTUM-BASED MINIMIZATION OF THE GINZBURG-LANDAU FUNCTIONAL ON EUCLIDEAN SPACES AND GRAPHS

OLUWATOSIN AKANDE, PATRICK DONDL, KANAN GUPTA, AKWUM ONWUNTA, AND STEPHAN WOJTOWYTSCH

Abstract. We study the momentum-based minimization of a diffuse perimeter functional on Euclidean spaces and on graphs with applications to semi-supervised classification tasks in machine learning. While the gradient flow in the task at hand is a parabolic partial differential equation, the momentum-method corresponds to a damped hyperbolic PDE, leading to qualitatively and quantitatively different trajectories. Using a convex-concave splitting-based FISTA-type time discretization, we demonstrate empirically that momentum can lead to faster convergence if the time step size is large but not too large. With large time steps, the PDE analysis offers only limited insight into the geometric behavior of solutions and typical hyperbolic phenomena like loss of regularity are not be observed in sample simulations.

## 1. Introduction

From its inception as Dido's problem, separating two regions with as short a boundary as possible is one of the oldest problems in mathematics. Many problems across the sciences are driven by the energetic imperative to minimize a perimeter or weighted perimeter functional, from crystal grain growth to rocks being ground to smooth pebbles in the sea and from liquids with surface tension to soap films and bubbles. The heuristic of minimizing a transition region in a suitable sense has been applied to graphs in the classical setting (min-cut problem) and more recently on weighted graphs in the setting of semi-supervised learning in data science.

In semi-supervised learning tasks, we have a large number of data points but only a small subset of them are labeled. Our goal is to label the remaining points. Two heuristics are common:

(1) *Proximity-based clustering.* We base new labels on the closest labeled point or points. The simplest instance of this heuristic is the $k$-nearest neighbors algorithm, which does not exploit the knowledge of where other unlabeled points are. A more advanced PDE-based version which integrates this information can be based on the eikonal equation [DEK22].
(2) *Perimeter minimization clustering.* We try to assign consistent labels to clusters of data and change labels only in between clusters in regions of low data density (small weighted perimeter)

The second approach leads to similar mathematical structures in data sciences as in the science. PDE-based algorithms are popular in machine learning due to their interpretability: While intuition may not transfer one-to-one to the new setting, such analogy can offer insight into the relative strengths and weaknesses of algorithms and even inspire methods to remedy such pitfalls [CCTS20].

Finding minimizers of perimeter functionals is not easy, neither analytically nor numerically. It has stimulated active research across disciplines from minimal surfaces in differential geometry

*Key words and phrases.* Phase-field model, semi-supervised learning, Graph-PDE, momentum-based optimization.

to mean curvature flow (the $L^2$-gradient flow of the perimeter functional) in partial differential equations and to rigorous computational approximations in numerical analysis.

Direct numerical discretizations of an evolving interface between two regions are possible, but challenging even for surfaces in dimension three [DDE05, DE07]. Mimicking extrinsic perspectives in geometric measure theory [Bra15], extrinsic approaches to mean curvature flows have been proposed. These include the 'thresholding' or Merriman-Bence-Osher (MBO) scheme [MBO92, Eva93] and the Allen-Cahn equation. Both have an $L^2$-gradient-flow structure with respect to a smooth approximation of the perimeter functional [EO15, LO20]. Heuristically, the sharp jump between the two domains is 'smeared out' across a narrow region in a principled way, leading to computationally stable methods which can easily accommodate topological transitions. Due to their stability, diffuse interface models have been popular both in mathematical modeling and computational works. Both the MBO scheme and the Allen-Cahn equation have been studied extensively on graphs for applications in image segmentation and semi-supervised learning [LB17, MBC18, BKS18, BM19, Cri19, MBS20, BKMS20, BvGL21, BvGL$^+$23].

Classical energy-driven approaches (MCF, MBO, Allen-Cahn) have emphasized using a gradient flow to minimize a diffuse perimeter functional, both on Euclidean spaces and graphs. In this work, we consider a momentum-based method or 'accelerated gradient flow' instead.

While gradient flows choose a locally optimal 'descent' direction based on first order information on an objective landscape, momentum methods (also referred to as 'accelerated gradient flows') retain information on past descent directions along their trajectory. Both can be seen as applications of Newton's second law, but while the (inertial) mass vanishes in gradient flow models, it is normalized as 1 in momentum methods. Due to inertia, the velocity does not change instantaneously to adapt to a new gradient. Integrating such global information allows them to converge faster in 'favorable' landscapes where past information is indicative of future geometry.

Rigorous guarantees for momentum methods outperforming gradient descent schemes are available mostly in convex optimization or in landscapes with convex-like properties [GW24]. Qualitative differences can arise between finite-dimensional and infinite-dimensional tasks [SW23]. Still, for non-convex optimization tasks in machine learning such as the training of neural networks, there is a large corpus of empirical evidence that momentum-based methods converge significantly faster than pure gradient descent methods – see e.g. [GW24] and the sources cited therein.

We note that (diffuse) perimeter minimization geometrically differs from most benchmarks in convex (and non-convex) optimization. For instance, compact initial surfaces vanish in finite time under mean curvature flow (the gradient flow of the perimeter functional), while gradient flows typically require infinite time to find minimizers in convex and strongly convex optimization tasks. This article is a curiosity-driven exploration into momentum methods for perimeter minimization: Are they (provably and/or empirically) outperforming gradient flow discretizations?

Our main contributions are as follows.

(1) We introduce the 'accelerated Allen-Cahn' equation and establish its elementary properties (total energy dissipation, conditional convergence to a minimizer).
(2) We show that the accelerated Allen-Cahn equation is hyperbolic and inherits its geometric properties from the wave equation, in particular: finite speed of propagation and a lack of implicit regularization for evolving interfaces, compared to the parabolic mean curvature flow equation.
(3) We demonstrate that it also inherits features from momentum-methods compared to gradient flows (such as 'overshooting' a global minimizer due to inertia).

(4) We conjecture a geometric evolution equation for the hypersurfaces in the singular limit as the width of the diffuse interfaces is taken to zero.
(5) We introduce two time discretizations of the 'accelerated Allen-Cahn equation', both based on a stablizing convex-concave splitting, i.e. a splitting where the gradient of the 'convex part' of the objective function is evaluated implicitly and the gradient of the 'concave part' of the objective function is evaluated explicitly.

    The *Convex-Implicit Nonconvex-Explicit Momentum Algorithm* (CINEMA) is a discretization which provably decreases the sum of kinetic and potential energy in every time step even for large time step sizes, but empirically it does not achieve acceleration over a standard gradient descent scheme with convex-concave splitting.

    The second discretization is an instance of the *Fast iterative shrinkage and thresholding algorithm* (FISTA), also with convex-concave splitting. We do not guarantee it to be energy decreasing, but empirically it achieves substantial acceleration over a gradient flow.
(6) We demonstrate how both algorithms can be implemented with negligible excess computational cost over a mere gradient descent scheme. A key ingredient is the choice of a double-well potential whose convex part is quadratic, leading to a linear problem for the implicit part of the time step.
(7) We compare momentum methods and local-in-time gradient descent methods numerically in Eudlidean spaces and on semi-supervised learning tasks on graphs.

Oversimplifying, we can summarize: The 'accelerated gradient flow' of the diffuse perimeter functional has potentially undesirable geometric properties. However, in a large time step discretization, the algorithm may achieve significantly faster convergence than a convex-concave splitting gradient descent scheme. The choice of a time discretization is crucial.

Note that we are not considering the 'accelerated Allen-Cahn equation' as a physical model, but only as a computational tool to find a (local) minimizer of the diffuse perimeter functional. Momentum-based methods have been previously considered in numerical analysis for solving obstacle problems [Sch18, CY19] with an $L^1$-penalty (which enforces the obstacle constraint exactly). Unlike the Allen-Cahn equation, these tasks fall into the realm of convex optimization where FISTA is provably faster (at least in the sense of minmax rates).

The article is organized as follows. In Section 2, we give more context for the mathematical models underlying this work: The Ginzburg-Landau (diffuse perimeter) functional, momentum-based first order optimization, and partial differential equations on graphs. In Section 3, we analyze the accelerated gradient flow and gradient flow of the Ginzburg-Landau energy on Euclidean spaces and compare their geometric properties. In Section 4, we derive discrete time algorithms for the numerical solution of the evolution equations, which we use for numerical experiments both in Euclidean spaces and on graphs in Section 5. We conclude with a discussion of context and open problems in Section 6.

## 2. Background Material

2.1. **Perimeter minimization and the Ginzburg-Landau functional.** Many physical phenomena are driven by 'perimeter minimization', i.e. by the imperative to minimize the length (or area) separating two different 'phases'/open sets in $\mathbb{R}^2$ (or $\mathbb{R}^3$). Mathematically, this is generally described as minimizing

$$\mathrm{Per}(E) = \int_\Omega \|\nabla 1_E\| \,\mathrm{d}x = \sup\left\{ \int_E \mathrm{div}(\phi)\,\mathrm{d}x : \phi \in C^\infty(\Omega;\mathbb{R}^d),\ \|\phi\|_{L^\infty} \leq 1 \right\}$$

where $1_E$ is the indicator function of the set $E$ and $\Omega$ is a larger containing set. Since $1_E$ is non-smooth, the integral has to be understood in the sense of functions of bounded variation [GW84], which leads to analytic and numerical challenges.

A computationally stable approximation is the Ginzburg-Landau functional (also sometimes referred to as Modica-Mortola energy)

$$\mathrm{Per}_\varepsilon(u) = \int_\Omega \frac{\varepsilon}{2} \|\nabla u\|^2 + \frac{W(u)}{\varepsilon} \,\mathrm{d}x$$

where $W$ is a 'double-well potential': A non-negative function which takes the value zero only if $u \in \{0, 1\}$, for instance $W(u) = u^2(1-u)^2$. If $\varepsilon$ is small, then $u$ takes values very close to 0 or 1 on most of the domain. However, the transition between the potential wells $0, 1$ cannot happen arbitrary quickly due to the presence of the squared gradient. Both contributions to the energy balance when $u$ transitions between being close to 0 and close to 1 on a length scale $\sim \varepsilon$. Indeed, $\mathrm{Per}_\varepsilon$ converges to (a $W$-dependent multiple of) Per as $\varepsilon \to 0^+$ (in the sense of $\Gamma$-convergence). Depending on the boundary conditions for $u$, the limit may also be a perimeter relative to the set $\Omega$ – we refer to [MM77, Mod87, ADA00] for details.

Rather than the characteristic function of the set $E$, which jumps along the boundary $\partial E$, we encounter functions of the form

$$u_\varepsilon(x) = \phi\left(\frac{\mathrm{sdist}_E(x)}{\varepsilon}\right)$$

when studying $\mathrm{Per}_\varepsilon$. Here $\mathrm{sdist}(x) = \mathrm{dist}(x, E^c) - \mathrm{dist}(x, E)$ is the *signed distance function* from $\partial E$, taken to be positive inside of $E$, and $\phi$ is the optimal transition between the potential wells at 0 and 1 in one dimension, i.e. the monotone increasing function which balances the contributions to the energy: $(\phi')^2 = W(\phi)$. Under mild assumptions, there is a unique solution to this ODE such that $\phi(0) = 1/2$ and $\lim_{x\to\infty} \phi(x) = 1$, $\lim_{x\to-\infty} \phi(x) = 0$.

By differentiation, we see that the 'optimal profile' $\phi$ satisfies $\phi'' = W'(\phi)$. The transition between 0 and 1 is 'smeared out' across an area of width $\sim \varepsilon$ with the characteristic shape $\phi$. We recall that sdist, like the regular distance function, has a unit gradient: $\|\nabla \mathrm{sdist}\| \equiv 1$ wherever the distance function is smooth. By standard results, sdist is always $C^2$-smooth close to a $C^2$-boundary and even if $\partial E$ is non-smooth, it is differentiable except on a set of measure zero by Rademacher's theorem.

### 2.2. **The Allen-Cahn Equation.** The Allen-Cahn equation

$$\varepsilon\, \partial_t u = \varepsilon\, \Delta u - \frac{W'(u)}{\varepsilon}$$

is the (time-normalized) $L^2$-gradient flow of the Ginzburg-Landau functional. Different boundary conditions imposed on the Ginzburg-Landau energy – which correspond to different limits as $\varepsilon \to 0^+$ – correspond to different boundary conditions for the PDE.

#### 2.2.1. *Singular limit.*

Like $\mathrm{Per}_\varepsilon$ converges to a perimeter functional, also solutions to the Allen-Cahn equation (i.e. the $L^2$-gradient flow of $\mathrm{Per}_\varepsilon$) converge to solutions of Mean Curvature Flow (i.e. the $L^2$-gradient flow of Per) in a suitable sense. Heuristically, this can be reasoned out as follows:

If $u_{\varepsilon,0} : \Omega \to [0, 1]$ is the initial condition for the Allen-Cahn equation, then $0 < u_\varepsilon(t, x) < 1$ for all $t > 0$ and $x \in \Omega$ by the maximum principle (unless the boundary conditions on $\partial\Omega$ force us to

leave the interval). In particular, we can write

$$u_\varepsilon(t,x) = \phi\left(\frac{r_\varepsilon(t,x)}{\varepsilon}\right), \qquad r_\varepsilon = \varepsilon\,\phi^{-1}(u_\varepsilon).$$

In $r_\varepsilon$, the Allen-Cahn operator can be expressed as

$$(\partial_t - \Delta)u_\varepsilon + \frac{W'(u_\varepsilon)}{\varepsilon^2} \;=\; \frac{\phi''(r_\varepsilon/\varepsilon)}{\varepsilon^2}\big(1 - \|\nabla r_\varepsilon\|^2\big) + \phi'\left(\frac{r_\varepsilon}{\varepsilon}\right)\,\big(\partial_t - \Delta\big)r_\varepsilon.$$

It is in general not possible to simultaneously make both terms zero, i.e. to simultaneously solve $\|\nabla r\|^2 = 1$ and $(\partial_t - \Delta)r = 0$. If $r_\varepsilon \to r$ for some limiting $r$, then we expect that the coefficient of $1/\varepsilon^2$ has to be zero, while there is a slight bit of leeway for the $O(1)$ term. Thus, we expect that $\|\nabla r\| \equiv 1$ everywhere, suggesting that $r$ should be a signed distance function in the spatial coordinates. This does not tell us anything about the time evolution of $r$ and the interface. To find it, we posit that the second PDE $(\partial_t - \Delta)r$ is solved on the interface $r = 0$, i.e. in the place where $\phi'$ is largest.

It is well known that $\operatorname{div}(\nabla v/\|\nabla v\|)(x)$ is the mean curvature of the level set $\{z : v(z) = v(x)\}$ at $z$ for any smooth function $v$ – see e.g. [ES91, Section 2.1]. If we are correct and $\|\nabla r\| \equiv 1$, then $\Delta r$ is the mean curvature of the interface

$$I(t) = \{x : r_\varepsilon(t,x) = 0\} = \{x : u_\varepsilon(t,x) = 1/2\}.$$

A proof in the context of distance functions is also given in [GT77, Chapter 14.6].

The meaning of $\partial_t r$ is easiest to glean in the setting of moving hyperplanes. Namely, if $H(t) = \{x : x_n = x_n^0 + vt\}$, then the signed distance function is $\operatorname{sdist} = (x_n^0 + vt) - x_n$ and the normal velocity is $v = \partial_t r$ (where $v > 0$ corresponds to the area where $\operatorname{sdist} > 0$ expanding in time).

We thus conjecture that $r_\varepsilon$ is (close to) the signed distance function from an interface $I(t)$ which moves by mean curvature. Far away from the interface, we conjecture that 'nothing happens', i.e. $u_\varepsilon$ remains almost constant in space and time close to the potential wells. Overall, $u_\varepsilon$ would then be close (e.g. in $L^2$) to the characteristic function of a set $E(t)$ whose boundary moves by mean curvature flow.

The heuristic explanation above in fact describes the limiting behavior of the Allen-Cahn equation. Rigorous proofs of this result in various forms are given in [Ilm93, MR11, FLS20].

2.2.2. *Vector-valued extension.* In many applications, there may be more than two phases. If the phases are unordered (i.e. phase 1 can border phase 3 and does not have to pass through phase 2), we can design Ginzburg-Landau type functionals

$$E_\varepsilon(u) = \int_\Omega \frac{\varepsilon}{2}\,\|Du\|_F^2 + \frac{W(u)}{\varepsilon}\,\mathrm{d}x$$

for functions $u : \Omega \to \mathbb{R}^k$ where $k \geq 3$ is the number of classes, $\|Du\|_F$ is the Frobenius norm of the derivative matrix and $W$ is a potential with $k$ wells, usually selected at the unit vectors $e_1, \dots, e_k$, unless prior information suggests that boundaries between different phases should have different 'surface tension'.

An easy way to create such a double-well potential is to select a double-well potential $W_{1D} : \mathbb{R} \to \mathbb{R}$ which vanishes only at $0, 1$ and set

$$W : \mathbb{R}^k \to [0,\infty], \qquad W(u) = \sum_{i=1}^k W_{1D}(u_i) + \lambda \cdot \left(1 - \sum_{i=1}^k u_i\right)^2$$

for some $\lambda \in (0, \infty]$. The first contribution to the potential $W$ vanishes if and only if $W_{1D}(u_i) = 0$ for all $i$, i.e. if and only if $u_i \in \{0, 1\}$ for all $i$. The second term ensures that exactly one of the $u_i$ is 1 and the others are 0.

Also solutions to the vector-valued Allen-Cahn equation approach (multi-phase) mean curvature flow as proved recently in [FM24].

2.2.3. *The hyperbolic Allen-Cahn equation.* Another modification of the classical Allen-Cahn equation is its hyperbolic version

$$\tau\, u_{tt} + \alpha\, u_t = \Delta u + \frac{W'(u)}{\varepsilon^2}.$$

which has been considered more recently by several authors with parameters scaling as $\tau/\alpha = O(\varepsilon)$. Like in the parabolic case, the limiting dynamics are described by mean curvature flow. For details, see e.g. [NGA16, FLM16, Fol17].

The 'accelerated Allen-Cahn equation' considered in this work is given by the same equation, but in a parameter regime where $\tau$ and $\alpha$ are both of order 1. Its behavior is therefore quite different and borrows more from the hyperbolic world of PDEs.

2.3. **Momentum-based first order methods in optimization.** Gradient flows reduce an objective function by adjusting the function inputs towards a locally optimal 'steepest descent' direction. If the objective function has favorable geometric properties – for instance, in convex optimization – it is possible to integrate more global geometric information gained throughout the optimization process to achieve faster convergence. This is the rationale behind both conjugate gradient methods in numerical linear algebra and momentum-based optimizers such as Nesterov's algorithm [Nes83] and FISTA [BT09] in convex optimization.

For instance, if $f$ is a convex objective function which has a minimizer and a Lipschitz-continuous gradient, then gradient descent achieves a decay of $f(x_t^{GD}) - \inf f = O(1/t)$ in $t$ steps while Nesterov's method achieves the much faster decay $f(x_t^{Nest}) - \inf f = O(1/t^2)$ with constants of comparable size hidden in the Landau notation. Non-asymptotic lower bounds demonstrate that Nesterov's method achieves optimal decay, at least up to a constant factor – see e.g. [Nes18] for precise statements.

For information from previous time-steps to remain useful, the objective function $f$ must have favorable geometric properties such as convexity or strong convexity. Recently, multiple works have relaxed the geometric conditions under which faster convergence can be established in several directions – see e.g. [GW24, HADR24] and the references cited therein. While many realistic optimization tasks do not fall into classes where faster convergence for momentum methods can be guaranteed, the use of momentum often leads to a notable improved in practice (for instance in the training of neural networks).

Both momentum-methods and gradient flows are derived from Newton's second law of mechanics

$$m\ddot{x} = -\alpha\, \dot{x} - \nabla f(x)$$

for a particle of mass $m$ under the influence of a linear friction and a potential force $-\nabla f$. Gradient flows formally correspond to $m = 0, \alpha = 1$ while momentum-methods correspond to positive mass $m = 1$. The optimal coefficient of friction $\alpha$ depends on the geometry of $f$: For $\mu$-strongly convex functions, we select (an estimate of) $2\sqrt{\mu}$ while for general convex functions, $\alpha = \alpha(t) = 3/t$ is in fact a function of time.

2.4. **Momentum-based time stepping algorithms in convex optimization.** Notably, the choice of time discretization for the heavy-ball dynamics is crucial. Polyak's [Pol64] heavy-ball method

$$x_{n+1} = x_n - \alpha\,\nabla f(x_n) + \beta(x_n - x_{n-1})$$

is closely related to the explicit (forward) Euler discretization of the heavy ball ODE and generally does not achieve acceleration (or even fails to converge) in situations where Nesterov's scheme remains stable [LRP16, GTD23]. In situations where a convex function can be decomposed into one smooth part and one 'simple' part, the *fast iterative shrinkage and thresholding algorithm* (FISTA) which treats the smooth part explicitly and the simple part implicitly is a stable algorithm with provable acceleration.

Let $H$ be a Hilbert space. Consider the task of minimizing the sum of two functions $F+G : H \to \mathbb{R}$. If $F, G$ are both convex, then Nesterov's method and FISTA are two well-studied methods which are both based on numerical discretizations of the 'heavy ball ODE' $\ddot{x} = -\alpha(t)\,\dot{x} - \nabla(F+G)(x)$. Both can be written in the form

$$x_{n+1} = x_n + \tau v_n - \eta g_n, \qquad v_{n+1} = \rho_n\big(v_n - \tau g_n\big) \tag{1}$$

where $g_n$ is an approximation of the gradient. The general scheme (1) encompasses

(1) Nesterov's method with

$$g_n = \big(\nabla F + \nabla G\big)\big(x_n + \tau v_n\big)$$

and parameters $\eta = \tau^2$ and $\rho_n = \frac{n}{n+3}$ (convex case, corresponding to $\alpha(t) = 3/t$) or $\rho_n = \frac{1-\sqrt{\mu}\,\tau}{1+\sqrt{\mu}\,\tau}$ ($\mu$-strongly convex case, corresponding to $\alpha = 2\sqrt{\mu}$).

The scheme converges (at a rate faster than GD in general in the sense of minimax optimal rates) if $F+G$ is convex and $\tau \le 1/L$ where $L$ is the Lipschitz-constant of $\nabla(F+G)$.

(2) FISTA with

$$g_n = \nabla F(x_{n+1}) + \nabla G\big(x_n + \tau v_n\big)$$

and parameters $\eta = \tau^2$ and $\rho_n = \frac{n}{n+3}$ (convex) or $\rho_n = \frac{1-\sqrt{\mu}\,\tau}{1+\sqrt{\mu}\,\tau}$ ($\mu$-strongly convex).

The scheme converges (at a rate faster than GD) if $F, G$ are convex and $\tau \le 1/L$ where $L$ is the Lipschitz-constant of $\nabla G$ (which may be finite even if $\nabla F$ is discontinuous).

Neither scheme is guaranteed to decrease the 'total energy'

$$(F+G)(x_n) + \frac{\lambda}{2}\,\|v_n\|^2$$

monotonically in non-convex optimization for a suitable $\lambda$ depending on $\tau, \eta, \rho$, and even for very small time steps, we found that FISTA increased $E_\varepsilon$ in numerical experiments in certain time steps when we split $E_\varepsilon$ into its convex part $F$ and concave part $G$ (details below).

*Remark* 2.1 (Restart). It is well-known that momentum may cause us to overshoot minimizers both in continuous and in discrete time: A boulder rolling down a mountainside generally will not come to rest the moment it reaches the bottom of a valley.

A popular remedy are momentum algorithms with adaptive or scheduled restart:

$$(x_{n+1}, v_{n+1}) = \begin{cases} (x_{n+1}, v_{n+1}) & \text{as in Theorem 4.2 if } (F+G)(x_{n+1}) < (F+G)(x_n) \\ (x_n, 0) & \text{else.} \end{cases}$$

The first step in the momentum scheme coincides with a convex-concave gradient descent step since $v_0 = 0$, i.e. $(F+G)(x_1) \le e_1 \le e_0 = (F+G)(x_0)$. Restarting therefore guarantees strict

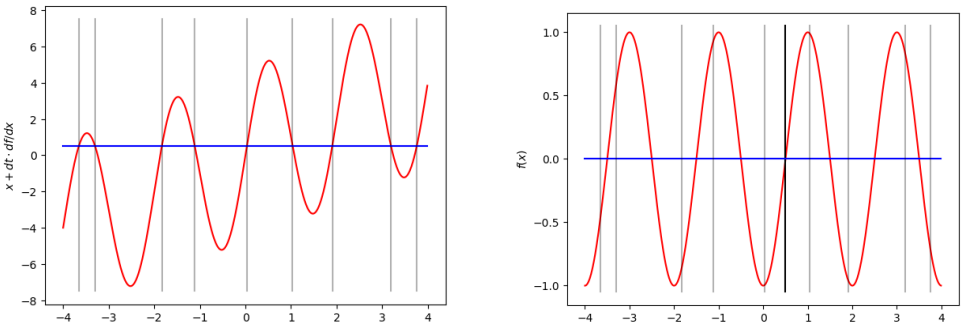


FIGURE 1. **Left:** $x+\tau f'(x)$ for $f(x) = -\cos(\pi x)$ and $\tau = 1.5$ (red line) as well as $z = 0.5$ (blue line). The intersections of the lines are the solutions to the implicit gradient descent equation $x + \tau f'(x) = z$ starting at $z$ and are marked by vertical lines. **Right:** The objective function $f(x)$ (red line) and the energy level $f(z)$ at the initial point (blue line). The admissible implicit Euler solutions and the initial point $z$ are marked by vertical lines. We see that the energy at the possible next steps is above the energy at the starting point for four out of nine options.

energy-monotonicity, unless we restart at $(x_n, 0)$ where $-\nabla G(x_n) \in \partial F(x_n)$, i.e. a stationary point of $F + G$.

2.5. **Convex-concave splitting.** If $F$ is convex and $G$ is concave, their sum can have a very complicated geometry – indeed, any $C^2$-function $f : \mathbb{R}^d \to \mathbb{R}$ with globally bounded Hessian can be written as

$$f(x) = \underbrace{f(x) + \frac{\lambda}{2}\|x\|^2}_{\text{convex}} + \underbrace{\frac{-\lambda}{2}\|x\|^2}_{\text{concave}}.$$

An explicit gradient descent scheme $x_{n+1} = x_n - h\,\nabla f(x_n)$ is only energy-decreasing if $h$ is small enough ($h < 2/L$ is sufficient where $L$ is the Lipschitz-constant of $\nabla f$). The implicit gradient descent scheme $x_{n+1} = x_n - h\,\nabla f(x_{n+1})$ always has an energy decreasing solution $x_{n+1} = \operatorname{argmin}_z \frac{1}{2}\|x - z\|^2 + h\,f(z)$, but if $h$ is too large, there are generally additional solutions with possibly higher energy – see e.g. Figure 1. Finding the energy decreasing solution may not be feasible.

However, if an explicit splitting $f = F+G$ into a convex and part $F$ a concave part $G$ is available, then the mixed explicit/implicit scheme

$$x_{n+1} = x_n - h\big(\nabla F(x_{n+1}) + G(x_n)\big) \tag{2}$$

defines a sequence uniquely and is energy-stable for any step size. Namely, $x_{n+1}$ is the *unique* minimizer of the strongly convex function

$$f^{aux}_{x_n,h}(z) := \frac{1}{2}\|z - x_n\|^2 + h\big(F(z) + F(x) + \langle \nabla F(x), z - x\rangle\big).$$

To see that the scheme is energy-stable, recall that a convex $C^1$-function $F : \mathbb{R}^d \to \mathbb{R}$ satisfies the first order convexity condition

$$F(x) \geq F(z) + \langle \nabla F(z), x - z\rangle \qquad \forall\, x, z \in \mathbb{R}^d.$$

For a concave function $G$ on the other hand, we note that

$$G(z) \leq G(x) + \nabla G(x), z - x\rangle$$

where we reversed the roles of $x$ and $z$ in the notation. In particular, if $x = x_{n+1}$ and $x = x_n$ as in (2), then we may add the two inequalities to obtain

$$\begin{aligned}(F+G)(x_{n+1}) &\leq (F+G)(x_n) + \langle \nabla G(x_n), x_{n+1} - x_n\rangle - \langle \nabla F(x_{n+1}), x_n - x_{n+1}\rangle \\ &= (F+G)(x_n) + \langle \nabla G(x_n) + \nabla F(x_{n+1}),\ x_{n+1} - x_n\rangle \\ &= (F+G)(x_n) - h\left\|\nabla F(x_{n+1}) + \nabla G(x_n)\right\|^2.\end{aligned}$$

In particular, the sequence $(F+G)(x_n)$ is monotone decreasing independently of the step size.

The assumptions on $F, G$ can be relaxed somewhat in terms of regularity, and the scheme may be defined on a Hilbert space rather than $\mathbb{R}^d$. These extensions are indeed necessary for the Allen-Cahn equation. We sill pursue them below, where we study a momentum scheme of the type (1) with convex-concave splitting in the gradients.

For geometric intuition, we provide the following example.

*Example* 2.2. Assume that $F(x) = \frac{1}{2}\, x^T A x$ where $A$ is a symmetric positive semi-definite $d \times d$-matrix, and $g : \mathbb{R}^d \to \mathbb{R}$ is a smooth convex function. Then the convex-concave splitting gradient descent scheme reads

$$x_{n+1} = x_n - h\, A x_{n+1} - h\, \nabla G(x_n),$$

or equivalently

$$\begin{aligned}(I + hA)x_{n+1} &= x_n - h\, \nabla G(x_n) = (I + hA)x_n - h\, (Ax_n + \nabla G)(x_n) \\ &= (I + hA)x_n - h\, \nabla (F+G)(x_n).\end{aligned}$$

This reduces to the preconditioned explicit gradient descent scheme

$$x_{n+1} = x_n - h\big(I + h\, D^2 f(x_n)\big)^{-1} \nabla (F+G)(x_n).$$

In particular in one dimension, the effect of the convex-concave splitting is to automatically select a stable step size for the explicit gradient descent scheme.

This observation also illustrates the limitations of the convex-concave splitting: A large step size may in reality correspond to a fairly small (explicit) time-step. Even with an unconditionally stable scheme, there can be room to accelerate convergence in practice.

2.6. **Partial differential equations on graphs.** A graph $\Gamma = (V, E)$ is a tuple of two sets: The set $V$ of vertices and the set $E$ of edges, where an edge $e = \{v, v'\}$ is a set containing two vertices $v, v' \in V$. We say that $e = \{v, v'\}$ *connects* $v$ and $v'$.

In the following, we will assume that each edge has a weight $w_e \in [0, \infty)$ which is large if its vertices are 'close' (the edge is short) and small if the vertices are 'far apart' (the edge is long). We can think of $w_e$ as a reciprocal length or a reciprocal length squared.

The graphs we consider have a finite set $V = \{v_1, \dots, v_n\}$ of vertices and two vertices can only be connected by at most one edge. If $e$ connects the vertices $v_i, v_j$ in $V = \{v_1, \dots, v_n\}$, we also denote $w_e = w_{ij} = w_{ji}$. A function $u : V \to \mathbb{R}^k$ can also be interpreted as a vector $u_i = u(x_i)$.

In analogy to the Euclidean Dirichlet energy $E_{DR}(u) = \frac{1}{2}\int \|\nabla u\|^2 \,\mathrm{d}x$, one can define the graph Dirichlet energy

$$E_{DR}(u) = \frac{1}{2}\sum_{e \in E} w_e \left|u_i - u_j\right|^2 = \frac{1}{4}\sum_{i,j=1}^{n} w_{ij} \left|u_i - u_j\right|^2$$

where $w_{ij} = 0$ if $v_i, v_j$ are not connected by an edge. The $L^2$-gradient of the Dirichlet energy is the (negative) Laplacian, so we call the gradient

$$(Lu)_k = \partial_{u_k} E_{DR}(u) = \frac{1}{2} \sum_{i,j} w_{ij}(u_i - u_j)(\delta_{ik} - \delta_{jk}) = \sum_j w_{kj} \, (u_k - u_j)$$

of the Dirichlet energy the 'graph Laplacian.' Like the classical (negative) Laplacian, also the graph-Laplacian is symmetric and positive semi-definite and, since

$$\langle Lu, u \rangle = \sum_i u_i \sum_j w_{ij} \, (u_i - u_j) = \frac{1}{2} \left( \sum_{i,j} w_{ij}(u_i - u_j)u_i - \sum_{i,j} w_{ij}(u_i - u_j)u_j \right) = 2 \, E_{DR}(u),$$

it is positive definite on the orthogonal complement of the functions which are constant on the connected components of the graph.[1]

The graph-Laplacian introduced here is unnormalized, but symmetric positive definite. We note that there are two other common notions of a 'graph-Laplacian': The symmetric normalized graph Laplacian ensures that the weights of incoming edges at each node sum up to 1. The unnormalized graph Laplacian is given by the application of the matrix $L$ with entries $L_{ij} = -w_{ij}$ for $i \neq j$ and $L_{ii} = \sum_{j \neq i} w_{ij}$. The normalized graph Laplacian is given by $\tilde{L} = D^{-1/2} L D^{-1/2}$ where $D$ is the diagonal matrix with entries $D_{ii} = L_{ii}$. In simulations below, we use this normalization which falls into the same framework, but with different weights.

The final normalization generalizes a different property of the classical differential operator: Its link to Brownian motion. The 'random walk graph Laplacian' $D^{-1}L$ is the generator of a random walk on $V$. It is usually not symmetric since the probability of jumping from $v_i$ to $v_j$ may be different from the probability of jumping from $v_j$ to $v_i$: The probabilities to jump from $v_i$ to $v_j$ in the next step have to sum up to 1 over $j$ (since there are no other options), but there is no reason that they should sum up to 1 over $i$.

For variational problems – such as minization of the Ginzburg-Landau energy – the unnormalized or symmetric normalized graph Laplacians are suitable, and we focus on them in this article.

2.7. **Semi-supervised learning.** Consider a data space $\mathcal{X}$ and label space $\mathcal{Y}$. In semi-supervised learning applications, we receive a set of data points $S = \{x_i \in \mathcal{X} : i = 1, \dots, N\}$ and a collection of labels $\{y_i \in \mathcal{Y} : i \in I\}$ for a subset of the data. In general, the cardinality of the labeled set is much smaller than that of the dataset. Our task is to find a 'good' function $\ell : S \to \mathcal{Y}$ such that $\ell(x_i) = y_i$ for all $i \in I$.

To decide what a 'good' function $\ell$ is, we require further information. In general, we assume that a notion of similarity or distance between data points is available which is meaningful in the sense that similar data points can generally be expected to have similar labels. We can use this notion to construct a graph with vertex set $V = S$ in which two nodes are connected by an edge if they are 'similar enough'.

Our applications fall into a common framework: We assume that $\mathcal{X} = \mathbb{R}^d$ for some $d$ and $\mathcal{Y} = \{1, \dots, k\}$ for some $k \in \mathbb{N}$. We either connect two vertices $x_i, x_j \in V = S$ by an edge if their Euclidean distance $\|x_i - x_j\|$ is below a cut-off length, or we connect every point to its $K$ nearest neighbors for some $K \in \mathbb{N}$ (where the symmetry of the graph means that some vertices have degree

[1] Two vertices $v, v'$ are in the same connected component if there exists a 'path' $v_{i_1}, \dots, v_{i_m}$ such that $v_{i_1} = v$, $v_{i_m} = v'$ and $e_{i_l i_{l+1}} > 0$.

$\gg K$ if they are the nearest neighbor to that vertex). We either assign the weight 1 to all edges, or we choose edge weights

$$w_{ij} = \exp\left(-\frac{\|x_i - x_j\|^2}{\sigma^2}\right)$$

for the edge connecting $x_i, x_j$ for a suitable $\sigma$. Of course, more advanced constructions are possible and, at times, required.

The well-developed Allen-Cahn framework can be leveraged in semi-supervised learning applications by minimizing the 'energy'

$$E_\varepsilon : (\mathbb{R}^k)^n \to [0,\infty), \qquad E_\varepsilon(u) = \frac{\varepsilon}{4}\sum_{i,j=1}^N w_{ij}\big|u_i - u_j\big|^2 + \frac{1}{\varepsilon}\sum_{i=1}^N W(u_i)$$

subject to the (Dirichlet) 'boundary condition' $u_i = \vec{y}_i$ if $i \in I$. The vector $\vec{y}_i$ is the one-hot encoding of the label $y_i \in \{1,\dots,k\}$ (i.e. the vector which has a one in the $y_i$-th coordinate and zeros in all others). From $u$, we derive the label function $\ell$ by $\ell(x_i) = \arg\max_{1\le m\le k} u_m(x_i)$. As in the Euclidean setting, minimizing $E_\varepsilon$ (approximately) corresponds to minimizing the size of the transition set as measured by the sum over edges between differently labeled points

$$\frac{1}{2N}\sum_{i,j} w_{ij} 1_{\{\ell(x_i)\neq\ell(x_j)\}}.$$

The energy $E$ is non-convex and in general, there exist many (local) minimizers. Using a gradient flow to minimize $E_\varepsilon$ corresponds to solving the graph Allen-Cahn equation [BKS18, BM19, BKMS20, MBS20]. In this article, we investigate the use of momentum methods for the same purpose.

## 3. The Allen-Cahn Equation and Accelerated Allen-Cahn Equation on $\mathbb{R}^d$

3.1. **Basic properties.** For the sake of convenience, we assume that the doublewell potential $W : \mathbb{R} \to [0,\infty)$ is $C^1$-smooth, vanishes only at 0 and 1, and only grows quadratically at $\infty$. This excludes the popular prototype $W(u) = u^2(1-u)^2$, but it includes for instance the potentials used in our simulations. It could be generalized with little additional effort, but more complicated statements. We define

$$F_\varepsilon : L^2(\Omega) \to [0,\infty], \qquad F_\varepsilon(u) = \begin{cases} \int_\Omega \frac{\varepsilon}{2}\|\nabla u\|^2 + \frac{W(u)}{\varepsilon}\,\mathrm{d}x & \text{if } u \in H^1(\Omega) \\ +\infty & \text{else.} \end{cases}$$

The choice of defining $F_\varepsilon$ on the larger space $L^2$ provides a notion of dissipation both for the gradient flow and the momentum method. For an initial condition $u_0 \in H^1(\Omega)$, the accelerated and time-normalized $L^2$-gradient flow of $F_\varepsilon$ is given by the evolution equation

$$\begin{cases} (\partial_{tt} + \alpha\partial_t)u &= \Delta u - \frac{1}{\varepsilon^2}W'(u) & t>0, x\in\Omega \\ u &= u_0 & t=0 \\ \partial_t u &= 0 & t=0 \\ \partial_\nu u &= 0 & t>0,\ x\in\partial\Omega. \end{cases}$$

Naturally, we can require Dirichlet boundary conditions by making $F_\varepsilon$ finite if and only if $u \in g + H^1_0(\Omega)$ for some $g \in H^1(\Omega)$ rather than for all $u \in H^1(\Omega)$. In this case, we would recover a Dirichlet boundary condition also for the evolution equation in place of the homogeneous Neumann boundary condition. The following statement applies to both settings.

**Theorem 3.1** (Total energy decrease). *Let $\alpha, \varepsilon > 0$ and $W \in C^2(\mathbb{R})$ a function such that $W(u) = 0$ if and only if $u \in \{0, 1\}$. Assume that $u \in C^2([0, \infty) \times \Omega)$ solves the PDE*

$$(\partial_{tt} + \alpha\partial_t)u = \Delta u - \frac{1}{\varepsilon^2} W'(u)$$

*in a domain $\Omega$ and that either*

1. *$\partial_t u(t, x) \equiv 0$ for all $t > 0$ and $x \in \partial\Omega$ or*
2. *$\partial_\nu u(t, x) \equiv 0$ for all $t > 0$ and $x \in \partial\Omega$ or*
3. *$\partial\Omega = \emptyset$,*

*i.e. we have Dirichlet boundary conditions which do not depend on time (first case), homogeneous Neumann boundary conditions (second case) or periodic boundary conditions (third case).*

*Then the total energy*

$$E_\varepsilon(u) = F_\varepsilon(u) + \frac{\varepsilon}{2} \|\partial_t u\|^2_{L^2(\Omega)}$$

*is monotone decreasing in time.*

*Proof.* We compute

$$\begin{aligned} E'_\varepsilon(t) &= \frac{2\varepsilon}{2} \langle u_t, u_{tt} \rangle_{L^2(\Omega)} + \int_\Omega \varepsilon \langle \nabla u, \nabla u_t \rangle + \frac{W'(u)}{\varepsilon} u_t \, \mathrm{d}x \\ &= \int_\Omega \varepsilon \, u_{tt} \, u_t + \operatorname{div}(u_t \nabla u) - u_t \, \Delta u + \frac{W'(u)}{\varepsilon} u_t \, \mathrm{d}x \\ &= \int_\Omega \left( \varepsilon \, u_{tt} - \varepsilon \Delta u + \frac{W'(u)}{\varepsilon} \right) u_t \, \mathrm{d}x + \int_{\partial\Omega} u_t \, \partial_\nu u \, \mathrm{d}A = -\alpha\varepsilon \|u_t\|^2_{L^2(\Omega)} \end{aligned}$$

since the boundary integral vanishes if $u_t \equiv 0$ or $\partial_\nu u \equiv 0$ on $\partial\Omega$ (in particular if $\partial\Omega = \emptyset$). The domain integral is evaluated using the PDE. □

We conjecture that the regularity assumptions can be relaxed. Existence, uniqueness and regularity of solutions to the accelerated Allen-Cahn equation are left for future work.

While we do not prove that solutions to the accelerated Allen-Cahn equation have a long term limit, we establish that any limit if it does exist must be a critical point of the energy $F_\varepsilon$.

**Theorem 3.2** (Conditional convergence to a critical point). *Let $u$ be as in Theorem 3.1 Assume that there exist $c, C, R > 0$ such that*

$$c|u| \leq sign(u) \, W'(u) \leq C \, |u| \qquad \forall \ |u| \geq R.$$

1. *There exist a sequence of times $t_n \to \infty$ and $u^* \in H^1(\Omega)$ such that $u(t_n, \cdot) \rightharpoonup u^*$ weakly in $H^1(\Omega)$.*
2. *Assume that $u(t, \cdot) \to u^\infty$ weakly in $H^1(\Omega)$ as $t \to \infty$. Then $-\Delta u^\infty + \frac{W'(u^\infty)}{\varepsilon^2} = 0$.*

The condition on $W$ is purely technical and could easily be removed at the expense of a more complicated statement. For instance, for the classical example $W(u) = u^2(1 - u)^2$, we would have to consider the more complicated space $H^1 \cap L^4$.

*Proof.* **First claim.** The condition on $W'$ ensures that

$$W(u) \geq W(R) + c \int_R^u t \, \mathrm{d}t \geq c \left( \frac{u^2}{2} - \frac{R^2}{2} \right) \qquad \Rightarrow \quad u^2 \leq R^2 + \frac{2}{c} W(u)$$

if $u \geq R$ and similarly for $u \leq -R$. The same bold holds trivially for $|u| \leq R$, so

$$\int_\Omega u^2 \,\mathrm{d}x \leq R^2|\Omega| + \frac{2}{c}\int_\Omega W(u)\,\mathrm{d}x.$$

Since $E_\varepsilon(u)$ remains bounded along the accelerated Allen-Cahn equation, we find that for any sequence $t_n \in (0,\infty)$, the associated sequence $u(t_n,\cdot)$ remains bounded in $H^1$. By [Bre11, Theorem 3.18 and Proposition 9.1] there exists a subsequence $t_{n_k}$ of $t_n$ such that $u(t_{n_k},\cdot)$ converges to a limit weakly in $H^1(\Omega)$.

**Second claim.** Assume for the sake of contradiction that $f := \Delta u^\infty + \frac{W'(u^\infty)}{\varepsilon^2} \neq 0$ in $H^{-1}(\Omega)$. Let $\phi \in H^1_0(\Omega)$ such that $\bar{c} := \langle f, \phi\rangle_{H^{-1},H^1} > 0$. Denote

$$h(t) := \int_\Omega u(t,x)\,\phi(x)\,\mathrm{d}x.$$

Since $u$ converges to a limit weakly in $L^2(\Omega)$ as $t \to \infty$, we find that $h$ converges to a limit. To obtain a contradiction, compute

$$\begin{aligned}\left(\frac{d^2}{dt^2} + \alpha\frac{d}{dt}\right)\int_\Omega u(t,x)\,\phi(x)\,\mathrm{d}x &= \int_\Omega \left(u_{tt}(t,x) + \alpha u_t\right)\,\phi(x)\,\mathrm{d}x \\ &= \langle -\Delta u + W'(u)/\varepsilon^2,\, \phi\rangle_{H^{-1},H^1}.\end{aligned}$$

We note that $-\Delta u(t,\cdot) \rightharpoonup -\Delta u^\infty$ in $H^{-1}(\Omega)$ by definition. For the non-linear term $W'(u(t,\cdot))$, we note that $W'$ grows linearly, i.e. $W'(u(t_n,\cdot)) \to W'(u^\infty)$ strongly in $L^2$ by the compact Sobolev embedding [Dob10, Kapitel 6.7] and hence in $H^{-1}$.

We conclude that for sufficiently large $t$ we have

$$\frac{d}{dt}\left(h'(t) + \alpha h(t)\right) = h''(t) + \alpha h'(t) \geq \bar{c}/2.$$

In particular

$$h'(t) + \alpha h(t) \geq h_* + \frac{\bar{c}}{2}\,(t - t^*)$$

for $t > t^*$ with some $h_*, t_* \in \mathbb{R}$. Since $h$ converges, it also remains bounded. This means that $h'(t) \to +\infty$ as $t \to \infty$, also contradicting the fact that $h$ remains bounded. □

Formally, the accelerated Allen-Cahn equation is a semi-linear version of the telegraph equation

$$\left(\partial_{tt} + \lambda_1\partial_t + \lambda_2\right)u = c^2\,\Delta u$$

with a zeroth-order non-linearity. The telegraph equation is a hyperbolic partial differential equation of second order which arises when decoupling a system of PDEs modelling the electric flow in a transmission line. As an equation from electromagnetism, it is compatible with special relativity: Information cannot propagate faster than the speed of light.

We establish a similar limit on the speed of propagation also for the accelerated Allen-Cahn equation, as it will be crucial for our geometric analysis below. The method is standard, and more general results are known, but harder to access.

**Theorem 3.3** (Finite speed of propagation)**.** *Let $(\bar{t},\bar{x}) \in (0,\infty)\times\Omega$ such that the initial condition satisfies $u_0 \equiv 0$ on $B_{\bar{t}}(\bar{x}) \cap \Omega$ and the initial condition for $\partial_t u$ vanishes on $B_{\bar{t}}(\bar{x})$. Assume that $u$ is as in Theorem 3.1. Then $u(\bar{t},\bar{x}) = 0$.*

*Proof.* Consider the 'past light cone' of $(t_0, x_0)$, i.e. the family of shrinking domains $C(t) = B_{\bar{t}-t}(\bar{x}) \cap \Omega$ and the localized energy

$$e(t) := \int_{C(t)} \frac{1}{2}\, u_t^2 + \frac{1}{2}\, \|\nabla u\|^2 + \frac{W(u)}{\varepsilon^2}\, \mathrm{d}x.$$

We can compute the derivative

$$\begin{aligned} e'(t) &= \int_{C(t)} u_t u_{tt} + \langle \nabla u, \nabla u_t \rangle + \frac{W'(u)}{\varepsilon^2}\, u_t \, \mathrm{d}x - \int_{\Omega \cap \partial B_{\bar{t}-t}(\bar{x})} \frac{u_t^2 + \|\nabla u\|^2}{2} + \frac{W(u)}{\varepsilon^2}\, \mathrm{d}A \\ &= \int_{C(t)} \left( u_{tt} - \Delta u + \frac{W'(u)}{\varepsilon^2} \right) u_t \, \mathrm{d}x - \int_{\Omega} \frac{u_t^2 + \|\nabla u\|^2}{2} + \frac{W(u)}{\varepsilon^2}\, \mathrm{d}A + \int_{\partial C(t)} u_t\, \partial_\nu u \, \mathrm{d}A \end{aligned}$$

since only the portion $\partial B_{\bar{t}-t}(x_0)$ of the boundary is moving. The second boundary integral is obtained by the divergence theorem and therefore goes over the whole boundary $\partial C(t)$, but the boundary conditions make $u_t \partial_\nu u \equiv 0$ on $\partial\Omega \cap \partial C(t)$, so

$$\begin{aligned} e'(t) &= -\alpha\varepsilon \int_{C(t)} |u_t|^2 \, \mathrm{d}x + \int_{\Omega \cap \partial B_{\bar{t}-t}(x_0)} u_t\, \partial_\nu u - \frac{u_t^2 + \|\nabla u\|^2}{2} - \frac{W(u)}{\varepsilon^2}\, \mathrm{d}A \\ &\leq \int_{\partial B_{\bar{t}-t}(x_0)} \frac{1}{2}\big(u_t^2 + (\partial_\nu u)^2\big) - \frac{u_t^2 + \|\nabla u\|^2}{2}\, \mathrm{d}A \leq 0 \end{aligned}$$

by Young's inequality, using that $|\partial_\nu u| \leq \|\nabla u\|$. In particular, since $\|\nabla u\| \equiv u_t \equiv W(u) \equiv 0$ in $C(0)$, we conclude that $e(t) = 0$ for all $t \leq \bar{t}$. □

Naturally, the same would be true if $u \equiv 1$ on $\Omega \cap B_{\bar{t}}(\bar{x})$.

Theorems 3.1 and 3.2 serve as indicators that the accelerated Allen-Cahn equation can alternatively be used to as a tool to minimize the Ginzburg-Landau energy in place of the Allen-Cahn equation. To the best of our knowledge, the existence of a long time limit for solutions to the Allen-Cahn equation (i.e. the uniqueness of the limit for different subsequences $t_n, t_n' \to \infty$) is open in the general case. Even for gradient flows in finite dimensions, a unique limit may not exist – see for instance the summary of counterexamples in the introduction of [DK21].

Theorem 3.3 is a first step towards geometrically analyzing solutions to the accelerated Allen-Cahn equation. We will explore the geometry in greater detail in the following section by deriving the singular limit $\varepsilon \to 0$ of the evolutions.

3.2. **Singular limit.** Even if the initial condition takes values strictly between the potential wells, there is no guarantee that the same is true for the solution to the accelerated Allen-Cahn equation at a positive time. For instance, with homogeneous Neumann (or periodic) boundary conditions, if the inital condition $u^0 \equiv c$ is constant in space, so is the solution $u(t, \cdot) \equiv c(t)$ for all positive times, and the constant $c(t)$ satisfies the one-dimensional heavy ball ODE

$$c'' + \alpha c' = -\frac{W'(c)}{\varepsilon^2}. \tag{3}$$

If $W''(0), W''(1) > 0$, $W$ behaves like a quadratic function at the minimizers and $c$ strongly resembles a harmonic oscillator. Unless $\alpha$ is sufficiently large – roughly $2\sqrt{\min\{W''(0), W''(1)\}}/\varepsilon$ for critical dampening – it is well-known that $c$ 'overshoots' the potential wells – see e.g. [SW23].

We conjecture that a singular limit as $\varepsilon \to 0$ exists, and that it is given by the relation

$$\partial_t v = (1 - v^2)\,(h - \alpha v) \tag{4}$$

between the normal velocity $v$ and the mean curvature $h$. In Figure 2, we compare solutions to the accelerated Allen-Cahn equation to the predicted singular limit for the special case of a circle, where (4) becomes an ordinary differential equation for the radius

$$\ddot{r} = (1 - \dot{r}^2)\left(-\frac{1}{r} - \alpha\,\dot{r}\right)$$

since the PDE (4) is rotationally invariant.[2] The coefficient of friction is chosen as $\alpha = 0.05$. The numerical results suggest a good resemblance between continuum limit and numerical approximation.

Differential operators are implemented using a fast Fourier transform on a spatial spatial grid of $400 \times 400$ points with a time step size of $\tau = 2.5 \cdot 10^{-7}$. Further details of the numerical implementation can be found in Section 4.

Solutions to the ordinary differential equations for the singular limit are found using a predictor/corrector scheme based on the Adams-Bashforth and Adams-Moulton formulas of order five. The initial values for the linear multi-step methods were found by the Runge-Kutta method of order four.

3.3. **A brief comparison of the Allen-Cahn and accelerated Allen-Cahn equations.** There are several notable differences between the Allen-Cahn equation and its 'accelerated' version.

3.3.1. *Interface velocity.* The maximal speed that an interface can have in the accelerated Allen-Cahn equation or its singular limit is 1. This is quite different from the singular limit of the Allen-Cahn equation, where the velocity is routinely unbounded at topological singularities. It does, however, strongly resemble the behavior of the hyperbolic 'leading order' expression $(\partial_{tt} - \Delta)u$ in the accelerated Allen-Cahn equation.

Thus interfaces can move faster – and the perimeter can decrease faster – in the singular limit of the Allen-Cahn equation than in the singular limit of what we optimistically dubbed the 'accelerated Allen-Cahn equation'. However, the analysis in continuous time may not be representative of discrete time simulations: If $f$ is $\mu$-strongly convex, then generically the gradient flow of $f$ decays as $f(x_t) - \inf f \sim \exp(-\mu t)$ while the 'accelerated gradient flow' with $\alpha = 2\sqrt{\mu}$ decays like $f(x_t) - \inf f \sim \exp(-\sqrt{\mu}\,t)$. Which rate of decay is faster depends on whether $\mu < 1$ or $\mu > 1$.

However, if $\nabla f$ is $L$-Lipschitz continuous, then the largest stable step size for the explicit Euler discretization of gradient flow scales as $1/L$, while we can allow larger timesteps $\sim 1/\sqrt{L}$ in Nesterov's discretization of the accelerated gradient flow [SBC14]. In discrete time, we obtain the decay $f(x_k) - \inf f \sim (1 - \sqrt{\mu/L})^k$ for Nesterov's scheme after $k$ steps while gradient descent leads to markedly slower decay $f(x_k) - \inf f \sim (1 - \mu/L)^k$.

A more conclusive answer to which algorithm is 'faster' in practical (e.g. machine learning) applications therefore requires a discrete time analysis. For the highly non-convex problem of perimeter minimization, a fully satisfying answer is beyond the scope of this note, but numerical experiments suggest that momentum may be helpful if a suitable time-discretization is chosen.

3.3.2. *Interface width and shape.* For both equations, we make the ansatz that $u(t, x) = \phi(r(t, x)/\varepsilon)$. For the Allen-Cahn equation, we find based on this that $r$ is the signed distance function from a spatially evolving hypersurface, i.e. $\|\nabla r\| \equiv 1$. For the accelerated Allen-Cahn equation, on the

[2] So is the accelerated Allen-Cahn equation, up to boundary conditions. Our derivation is only valid inside the domain $\Omega$, i.e. for circles away from the boundary.

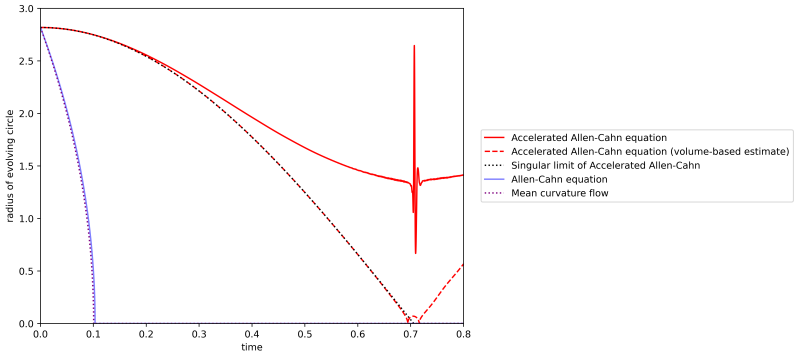


FIGURE 2. We numerically solve the Allen-Cahn equation and the accelerated Allen-Cahn equation for $\varepsilon = 0.01$ on the unit square with periodic boundary conditions. The initial condition is a circle of radius $r = 0.45$ centered at $(0.5, 0.5)$. Both solutions remain circular throughout their evolution.
For the accelerated Allen-Cahn equation, we estimate the circumference of the circle in two ways: By the Ginzburg-Landau energy (solid line) and by the volume-area relation $\mathrm{Per} = 2\sqrt{\pi \left| \int u \, \mathrm{d}x \right|}$ of a circle (dashed line). The second estimate suggests that the solution is indeed close to the singular limit (4), but the Ginzburg-Landau energy dramatically overestimates the true perimeter.
This effect is expected since the interfaces become compressed when the interface moves quickly, leading to suboptimal transitions between the potential wells. It is additionally possible that the spatial resolution is insufficient to accurately capture the effect's precise magnitude in the continuum limit.
Notably, after vanishing in a point, the solution to the accelerated Allen-Cahn equation continues its evolution by expanding the circle again, increasing its perimeter. Briefly, before the circle expands again, the integral of $u$ becomes negative.
For comparison, we present the solution to mean curvature flow/curve shortening flow – the singular limit of the Allen-Cahn equation – and the singular limit (4) for the accelerated Allen-Cahn equation.

other hand, the interface width depends on its velocity $v$: If $v \neq 0$, then $\|\nabla r\| = 1/\sqrt{1 - v^2} > 1$ and the interface is thinner than it would be without acceleration (see Figure 3).

In particular, an evolving 'interface' may become much thinner than we would anticipate from the optimal transition shape. Unless the spatial resolution of a numerical approximation is much finer than the thickness parameter $\varepsilon$ – and much finer than the Allen-Cahn equation would require in the same problem – the accelerated Allen-Cahn equation may not be resolved accurately.

Even for the continuous partial differential equation without numerical approximation, the sharper interfaces may have noticeable numerical impact: Since $u$ does not transition on the optimal length scale between the potential wells, it overestimates the size of the transition area. The faster the interface moves, the larger the gap between the value of the Ginzburg-Landau functional and the true size of the transition area.

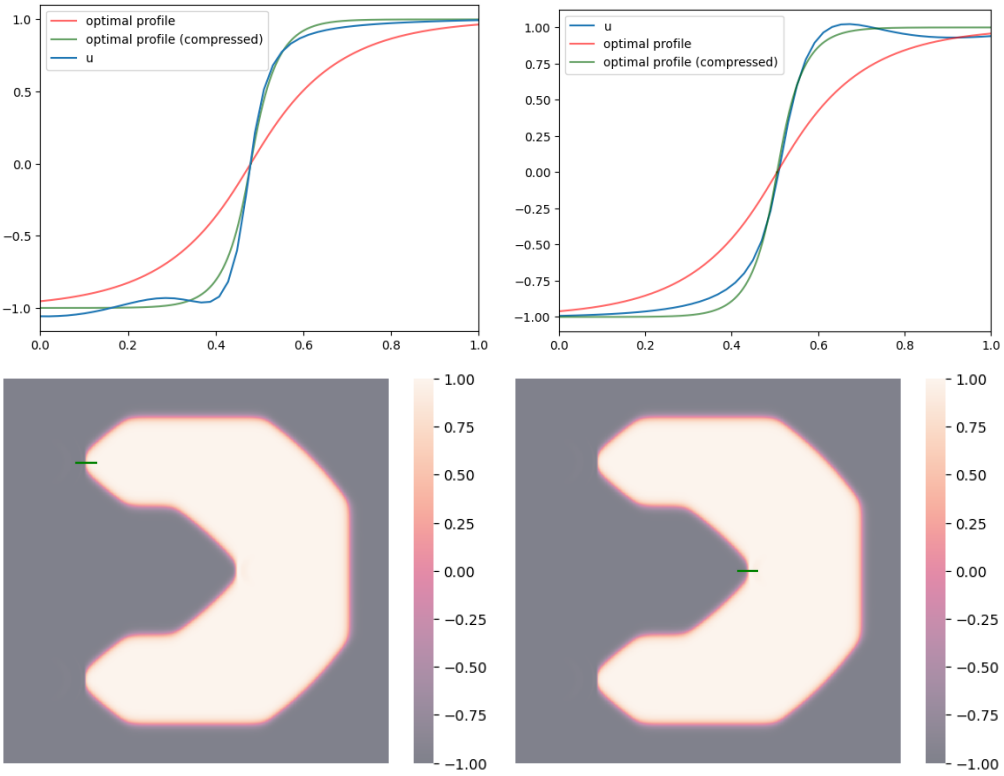


FIGURE 3. **Top line:** An interface moving to the left (left image) and an interface moving right (right image). The optimal profile $\phi$ (red line) is included on the length scale of a stationary interface (red line) and scaled for a good visual fit (green line) to illustrate the compression of fast-moving interface in the Allen-Cahn equation with momentum. **Bottom line:** The green line indicates where the one-dimensional slice which we are viewing in the image above is taken from in a simulation.

Comparing the radius estimates from the Ginzburg-Landau energy of the circle and the estimate from volume, this effect is noticeable in Figure 2 even when the interface is moving relatively slowly, e.g. at times $t \approx 0.4$.

3.3.3. *Ambient/parametrized compatibility.* The limit of solutions to the Allen-Cahn equation – a gradient flow of the Ginzburg-Landau energy – as $\varepsilon \to 0^+$ is (the indicator function of a set whose boundary is moving according to) a mean curvature flow, i.e. a gradient flow of the singular limit of the Ginzburg-Landau energies. More plainly: The limit of the gradient flows is the gradient flow of the limit. Whether such a result is true can in general be subtle [SS04, Ser11, DKW19], and an analogous result is not true for the momentum method. Namely, the singular limit of the evolution equations is

$$\dot{v} = \frac{1-v^2}{1+v^2}(h - \alpha v) \qquad \text{not} \quad \dot{v} = h - \alpha v$$

which we would get by imitating Newton's second law directly on the surface.

3.3.4. *Non-monotonicity of the energy.* As we observe in Figure 2, the Ginzburg-Landau energy $\mathrm{Per}_\varepsilon(u(t))$ is monotone decreasing for the gradient flow (Allen Cahn), but not for the momentum dynamics (accelerated Allen-Cahn). This is unsurprising: Even for a quadratic function in one dimension, the momentum dynamics generally overshoot the minimizer if the coefficient of friction $\alpha$ is too low. What is perhaps more surprising is that a along the momentum dynamics, a circle can shrink, disappear – and then reappear! From our preliminary analysis, it is unclear how the flow should continue past singularities.

## 4. Solving the accelerated Allen-Cahn equation numerically

In this Section, we introduce a computationally stable time discretization of the accelerated Allen-Cahn equation (CINEMA) and compare it to a version of FISTA with convex-concave splitting, using the ideas from Sections 2.4 and 2.5. Our main algorithmic contribution is given in an abstract setting in Section 4.1 and specialized to the accelerated Allen-Cahn equation in Section 4.2. An additional simplification is described in Section 4.3. A direct comparison between the two time-stepping methods is given in Section 4.4.

4.1. **An unconditionally stable momentum algorithm.** We consider the version of (1) for which $g_n = \nabla F(x_{n+1}) + \nabla G(x_n)$, i.e. we treate $F$ implicitly and $G$ explicitly, but unlike FISTA, we evaluate the gradient of $G$ at $x_n$ rather than the 'advanced' point $x_n + \tau v_n$. In terms of implementation, this merely corresponds to exchanging the order of two partial steps. We dub this version the Convex Implicit/Non-convex Explicit Momentum Algorithm (CINEMA).

We prove existence and energy-stability the CINEMA scheme in a slightly more general context which automatically covers the PDE setting.

**Theorem 4.1** (Existence in discrete time)**.** *Let $H$ be a separable Hilbert space. Assume that*

1. *$F : H \to \mathbb{R}$ is a weakly lower semi-continuous and convex function with sub-differential $\partial F$ such that the domain of $\partial F$ is dense in $H$.*
2. *$G : H \to \mathbb{R}$ is concave, continuous, Gateaux-differentiable and the Gateaux-derivative is continuous and linear, i.e. for every $x \in H$ there exists a vector $\nabla G(x) \in H$ such that*

$$\lim_{t\to 0} \frac{G(x+tv) - G(x)}{t} = \langle \nabla G(x), v\rangle \qquad \forall\ v \in H.$$

*Then given $x, v \in H$ and $\tau, \eta > 0$, there exists a unique $z^* \in H$ such that*

$$\frac{x - z^* + \tau v}{\eta} - \nabla G(x) \in \partial F(z^*).$$

The proof is given in the Appendix. Assuming for the moment that $\partial F(z) = \{\nabla F(z)\}$ is a singleton, we re-write

$$\frac{x - z + \tau v}{\eta} - \nabla G(x) \in \partial F(z^*) \qquad \Rightarrow \quad x + \tau v - z - \eta \nabla G(x) - \nabla F(z) = 0$$

which is compatible with the scheme (1) with $g_n = \nabla F(x_{n+1}) + \nabla G(x_n)$.

**Theorem 4.2** (Stability in discrete time)**.** *Let $F$ and $G$ be as in Theorem 4.1. Given $(x_0, v_0)$ and $\tau > 0$, $\eta > \tau^2/2$, and $\rho \in [0, 1]$, there exists a unique sequence $(x_n, v_n)$ which obeys the CINEMA*

*time-stepping scheme*

$$\frac{x_n + \tau v_n - x_{n+1}}{\eta} - \nabla G(x_n) \in \partial F(x_{n+1})$$
$$v_{n+1} = \rho_n\big(v_n - \tau g_n\big)$$

*where* $g_n \in \nabla G(x_n) + \partial F(x_{n+1})$ *is the same element of the sub-differential as in the line above. Additionally, the 'total energy'* $e_n = (F+G)(x_n) + \frac{1}{2\rho^2}\|v_n\|^2$ *satisfies*

$$e_{n+1} \le e_n - \frac{1}{2}(\rho^{-2}-1)\|v_n\|^2 + \left(\frac{\tau^2}{2} - \eta\right)\|g_n\|^2.$$

Also this proof is given in the Appendix.We will see in the next section that the Ginzburg-Landau energy satisfies all of the conditions of Theorem 4.2.

4.2. **The accelerated Allen-Cahn equation.** Let $W$ be a double-well potential such that $W'$ grows at most linearly at infinity. We split the energy

$$F_\varepsilon : L^2(\Omega) \to \mathbb{R}, \qquad F_\varepsilon(u) = \begin{cases} \int_\Omega \frac{\varepsilon}{2}\|\nabla u\|^2 + \frac{W(u)}{\varepsilon}\,\mathrm{d}x & \text{if } u \in H^1(\Omega) \\ +\infty & \text{else} \end{cases}$$

into a convex and a concave part

$$F_\varepsilon^{convex}(u) = \int_\Omega \frac{\varepsilon}{2}\|\nabla u\|^2 + \frac{W_{convex}(u)}{\varepsilon}\,\mathrm{d}x, \qquad F_\varepsilon^{concave}(u) = \int_\Omega \frac{W_{concave}(u)}{\varepsilon}\,\mathrm{d}x$$

where $W = W_{convex} + W_{concave}$ and naturally, $W_{convex}$ is convex and $W_{concave}$ is concave.

**Lemma 4.3.** *Assume that* $W'_{convex}$ *and* $W'_{concave}$ *both grow at most linearly at infinity. Then the functions* $F := F_\varepsilon^{convex}$ *and* $G := F_\varepsilon^{concave}$ *satisfy the conditions of Theorem 4.2.*

The assumption on the derivatives is necessary since the convex-concave decomposition is never unique. Namely, if $f = g + h$ is a decomposition into a convex and a concave function, then also $f = (g+\phi) + (h-\phi)$ where $\phi$ is any convex function.

*Proof.* $F$ is convex by construction with domain $H^1(\Omega)$ and the singleton-valued sub-differential $-\Delta u$ on the domain $H^2(\Omega)$ of the sub-differential (which is dense in $L^2$). To see that $F$ is lower semi-continuous, we note that if $u_n \rightharpoonup u^*$ in $L^2(\Omega)$ and $\liminf_{n\to\infty} F(u_n) < \infty$, then the subsequence which realizes the liminf is indeed bounded in $H^1(\Omega)$ as well. A further subsequence then converges to a limit $u'$ weakly in $H^1(\Omega)$. By the compact embedding of $H^1$ into $L^2$, we find that $u_n \to u'$ strongly in $L^2$, yielding that $u' = u^*$. We conclude that

$$\int_\Omega \|\nabla u^*\|^2\,\mathrm{d}x \le \liminf_{n\to\infty} \int_\Omega \|\nabla u_n\|^2\,\mathrm{d}x$$

by the lower semi-continuity of the norm under weak convergence [Bre11, Proposition 3.5] and that

$$\int_\Omega W_{convex}(u^*)\,\mathrm{d}x \le \liminf_{n\to\infty} \int_\Omega W_{convex}(u_n)\,\mathrm{d}x$$

by convexity. $G$ is concave and continuous by construction. Furthermore

$$\lim_{t\to 0} \frac{G(u+t\phi) - G(u)}{t} = \lim_{t\to 0} \int_\Omega \frac{W_{concave}(u+t\phi) - W_{concave}(u)}{t}\,\mathrm{d}x = \int_\Omega W'_{concave}(u)\phi\,\mathrm{d}x,$$

so $G$ is Gateaux-differentiable and the derivative is a continuous linear functional at every point (and continuous as a map into $L^2$) since $W'_{concave}$ grows at most linearly by assumption. □

The momentum descent step in the scheme of Theorem 4.2 for the $u$-variable is

$$u_{n+1} = u_n + \tau v_n - \eta \left( -\Delta u_{n+1} + \frac{W'_{convex}(u_{n+1})}{\varepsilon^2} + \frac{W'_{concave}(u_n)}{\varepsilon^2} \right)$$

or equivalently

$$(1 - \eta\Delta)\, u_{n+1} + \eta \frac{W'_{convex}(u_{n+1})}{\varepsilon^2} = u_n + \tau v_n - \eta\, \frac{W'_{concave}(u_n)}{\varepsilon^2}.$$

For a particularly simple implementation, we choose a double-well potential such that $W_{convex}(u) = u^2$ (scalar valued case) or $W_{convex}(u) = |u|^2 + \infty \cdot 1_{\{u \notin V\}}$ for the affine subspace $V = \{u : \sum_{i=1}^k u_i = 1\}$ of $\mathbb{R}^k$ (vector-valued case). To compute the next CINEMA time-step, we only need to solve

$$\left(1 + 2\,\frac{\eta}{\varepsilon^2} - \eta\Delta\right) \tilde{u}_{n+1} = u_n + \tau v_n - \frac{\eta}{\varepsilon^2}\, W'_{concave}(u_n)$$

and $u_{n+1} = \tilde{u}_{n+1}$ (scalar-valued case) or $u_{n+1} = \Pi_V \tilde{u}_{n+1}$ where $\Pi_V$ denotes the orthogonal projection onto $V$. A similar expression holds true for FISTA.

The algorithm can easily discretized by spectral methods or finite elements, leading to large but sparse linear systems. The same is true in the graph setting if the graph is sparse.

4.3. **A double-well potential with quadratic convex part.** We are looking for a doublewell potential $W$ such that $W(u) = u^2 + W_{conc}(u)$ where $W_{conc}$ is a differentiable concave function. To simplify the presentation, we construct $W$ with potential wells at $\pm 1$ rather than at $0, 1$. When we want the potential wells at $0, 1$ in simulations, we work with $W((1+u)/2)$, which also has a quadratic convex part $u^2/4$.

The functions we consider are smoother versions of $W(u) = u^2 + 1 - 2|u|$ in order to apply the existence result for our time-stepping scheme.

*Example* 4.4. For $R, \beta, \gamma > 0$, the potential

$$W(u) = W_{\beta,\gamma,R}(u) := u^2 + \beta - \gamma\sqrt{Ru^2 + 1}$$

satisfies $\lim_{u\to\pm\infty} W(u) = \infty$. If $\gamma > 0$, then

$$W_{conc}(u) := \beta - \gamma\sqrt{Ru^2 + 1}$$

is smooth and concave. We see that

$$0 = W'(u) = 2u - \frac{2\gamma R u}{2\sqrt{Ru^2+1}} = \left(2 - \frac{\gamma R}{\sqrt{Ru^2+1}}\right) u = 0$$

if $u = 0$ and at $u = \pm 1$ if $\gamma = \frac{2\sqrt{R+1}}{R}$. The critical point at zero is a local maximum and

$$W(\pm 1) = 1 + \beta - \gamma\sqrt{R+1} = 0$$

if $\beta = \gamma\sqrt{R+1} - 1 = \frac{2(R+1)}{R} = 1 + 2/R$. At the origin, we have

$$W(0) = \beta - \gamma = 1 + \frac{2}{R} - \frac{2\sqrt{R+1}}{R} = \frac{R + 2 - 2\sqrt{R+1}}{R} = \frac{(\sqrt{R+1} - 1)\sqrt{R+1}}{R} > 0$$

for all $R > 1$. This ansatz therefore yields a one parameter family of doublewell potentials with quadratic convex part and potential wells at $\pm 1$, which we denote by $W_R$. As $R \to \infty$, the value approaches 1 from below. See Figure 4 for a visualization. Higher values of $R$ corresponds to a more pronounced wall between the potential wells.

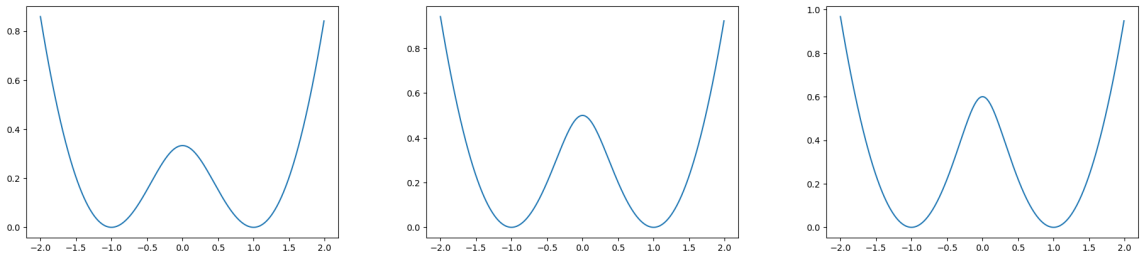

FIGURE 4. The doublewell $W_R$ for $R = 3, 8$ and 15 (moving from left to right).

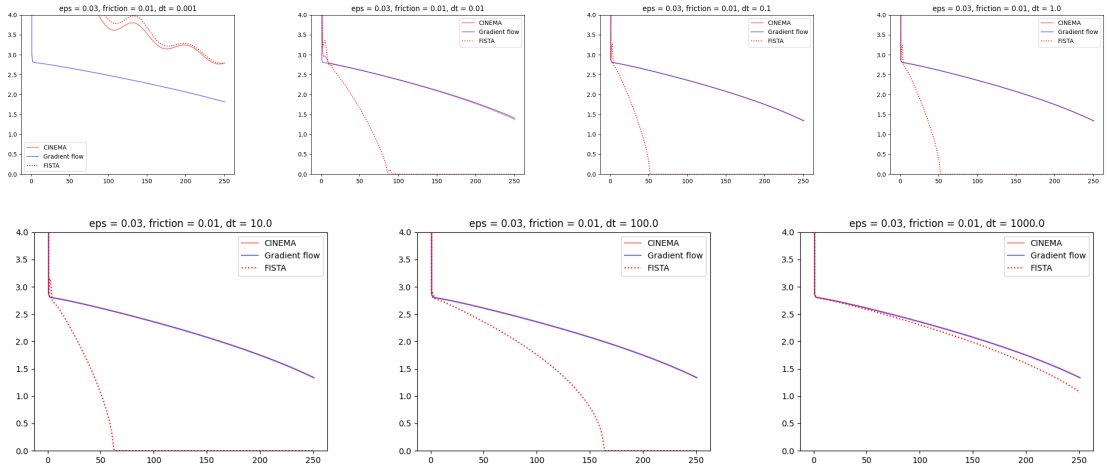


FIGURE 5. We compare FISTA, CINEMA and gradient descent where for all schemes, the convex part of the Ginzburg-Landau energy is treated implicitly and the concave part is treated explicitly. The $x$-axis counts the number of iterations, the $y$-axis the Ginzburg-Landau energy (not total energy).

A multi-well potential for the vector-valued case can be derived from $W$ as described in Section 2.

4.4. **CINEMA vs FISTA.** Let us compare CINEMA to the well-established FISTA scheme. In Figure 6, we compare three different versions of the time-stepping scheme (1) for minimizing the double-well potential $W_R : \mathbb{R} \to \mathbb{R}$ described in Section 4.3 (which corresponds to minimizing the Ginzburg-Landau energy among constant functions).

In all three versions, $F$ is the convex part and $G$ is the concave part of $W$, but the algorithms differ in where the gradients are evaluated: For Nesterov's algorithm, we evaluate both gradients at $x_n + \tau v_n$. For FISTA, we evaluate $\nabla G$ at $x_n + \tau v_n$ and $\nabla F$ (implicitly) at $x_{n+1}$. For CINEMA, we evaluate $\nabla G$ at $x_n$ and $\nabla F$ at $x_{n+1}$. In all algorithms, $\eta = \tau^2$ and $\rho = 1/(1 + \alpha\tau)$ with $\alpha = 0.01$.

Unsurprisingly, the explicit Nesterov scheme quickly seizes to be stable. More surprisingly, FISTA fails to reduce the 'total energy' $e_n = W(u_n) + \frac{1}{2\rho^2}|v_n|^2$ over a range of time step sizes, even as the energy is low enough to ensure that we are indeed in a region where $W$ is convex. Other monotone

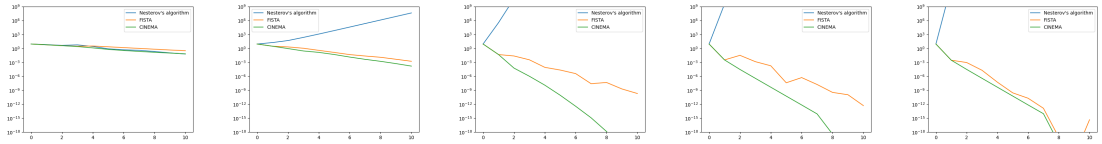

FIGURE 6. We plot the total energy $e_n = W(u_n) + \frac{1}{2\rho^2}|v_n|^2$ for Nesterov's algorithm (blue line), FISTA (orange line) and the algorithm proposed above (green line) for time step sizes $\tau \in \{0.5, 1, 10, 100, 1000\}$ (from left to right).

quantities exist. CINEMA is decreasing the energy fastest among the algorithms and reliably over several orders of magnitude in the time-step size.

The situation is quite different in the PDE-experiments we present in Figure 5. Here we compare the FISTA and CINEMA discretizations of the accelerated Allen-Cahn equation starting from a circle in the same experimental setting as outlined in Section 5.1.1. We see that for small time-steps ($\tau = 10^{-3}$), FISTA and CINEMA behave very similarly, and for very large time step sizes ($\tau = 10^3$), both FISTA and CINEMA essentially recover the behavior of a pure gradient descent scheme with convex-concave splitting.

Over the vast range of intermediate time step sizes (roughly $10^{-2} \leq \tau \leq 10^2$), we find that FISTA leads to much faster decrease of the Ginzburg-Landau energy while CINEMA never accelerates substantially over the gradient descent scheme with convex-concave splitting.

To interpret this observation, we compare the two time-stepping schemes. CINEMA corresponds to solving

$$u_{n+1} = u_n + \tau v_n + \eta \left( \Delta u_{n+1} - \frac{2u_{n+1} + W'_{R,conc}(u_n)}{\varepsilon^2} \right)$$

or equivalently

$$(-\eta\,\Delta + 1 + 2\eta)\, u_{n+1} = u_n + \tau v_n - \frac{\eta}{\varepsilon^2}\, W'_{R,concave}(u_n). \tag{5}$$

Formally, all that changes in FISTA is the place where the derivative of $W_{R,concave}$ is evaluated:

$$(-\eta\,\Delta + 1 + 2\eta)\, u_{n+1} = u_n + \tau v_n - \frac{\eta}{\varepsilon^2}\, W'_{R,concave}(u_n + \tau v_n).$$

Both schemes have comparable computational complexity. However, we conjecture that their different performance can be explained as follows: The FISTA scheme factors into a momentum step and a gradient descent step

$$u_{n+1/2} = u_n + \tau v_n, \qquad (1 + \eta/\varepsilon^2 - \eta\Delta)\, u_{n+1} = u_{n+1/2} - \frac{\eta}{\varepsilon^2}\, W_{concave'}(u_{n+1/2}).$$

The momentum step is not energy-driven and may lead to an intermediate 'loss of regularity' which is regained by the gradient descent step. The velocity – which concentrates on the narrow interface – is not regularized. By comparison, the two steps happen simultaneously and cannot be separated for CINEMA. This leads to unconditional stability, but also this comes at the cost of oversmoothing the velocity with the Laplacian by coupling the momentum step and the gradient descent step. See also Figure 10 for a visualization of the FISTA scheme with large time step size.

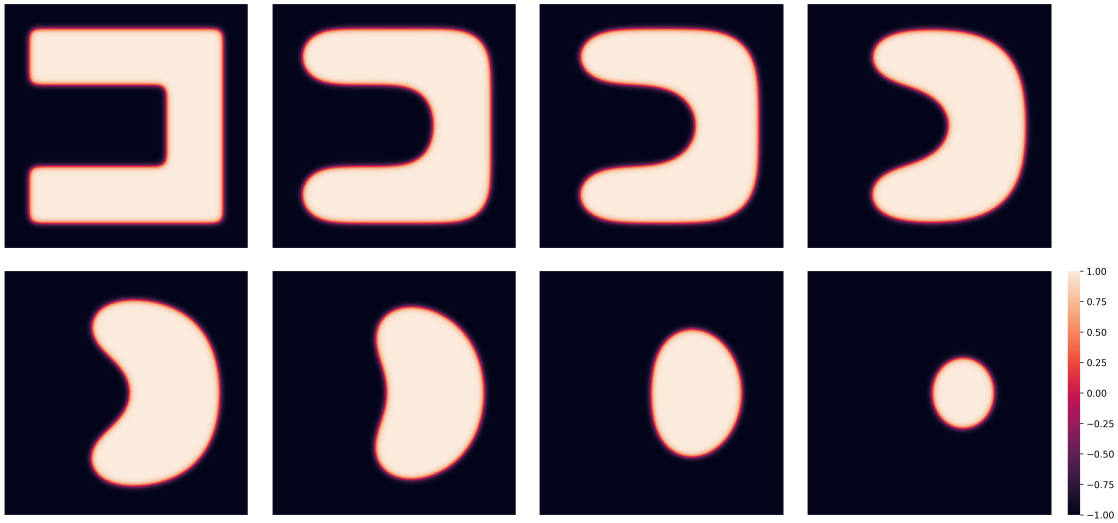


FIGURE 7. Evolution of a Jordan curve under the Allen-Cahn approximation to curve shortening flow at times $t \in \{0, 0.005, 0.01, 0.02, 0.03, 0.04, 0.06, 0.08\}$ (left to right, top to bottom).

## 5. Numerical experiments

5.1. **Simulations in Euclidean Spaces.** We present numerical solutions to the accelerated Allen-Cahn equation in two dimensions to develop geometric intuition.

5.1.1. *Simulations in* $\mathbb{R}^2$. We compare the gradient flow and accelerated flow for the Ginzburg-Landau functional on the unit square with periodic boundary conditions. The gradient flow is implemented by a convex-concave splitting and the accelerated gradient flow is implemented as CINEMA on a $n \times n$-grid with $n = 400$ with $\varepsilon = 0.01$ and the coefficient of friction $\alpha = 3$ for the accelerated version. The time step size is $\tau = 10^{-5}$. The equation (5) is solved in the Fourier domain, transforming the right hand side by an FFT and the solution by the inverse FFT, and analogously for the regular Allen-Cahn equation.

The initial condition is (a relaxation of) the characteristic function of a C-shaped set. For increased numerical stability, we take ten Allen-Cahn time steps with step size $\tilde{\tau} = 10^{-4}$ from the discontinuous initialization.

In Figure 7, we illustrate the numerical behavior of the Allen-Cahn approximation to mean curvature flow: corners are instantaneously smoothed out and the curve convexifies before disappearing in a 'round point' (i.e. the shape becomes approximately circular at the time of disappearance). This behavior is representative of MCF for curves in dimension two (also known as curve shortening flow) where it is known that no singularities occur if the inital curve is embedded.

In Figure 8, we illustrate the accelerated Allen-Cahn approximation to the geometric motion governed by (4). There are notable geometric differences to the regular Allen-Cahn equation.

At the ninety degree corners in the initial condition, the curvature is infinite (or essentially infinite after the Allen-Cahn relaxation), so the corner begins to instantaneously move inwards with velocity close to 1. On the straight interfaces, on the other hand, the curvature is zero, so

the interfaces remain initially unchanged due to the finite speed of propagation of information. In effect, this creates new corners with sharp angles of forty-five degrees. There is no parabolic 'smoothing' effect in the evolution.

In fast-moving segments of the boundary, the transition between the potential wells is markedly faster than in stationary segments. This increased sharpness is particularly visible in the vertical segments of the boundary in the second row (also compare Figure 3).

The 'vertical' momentum in the accelerated Allen-Cahn equation translates into 'horizontal' momentum for the interface: After the white area disappears in the bottom row, it reappears again.

We compare the decrease in the Ginzburg-Landau energy along the two different evolutions in Figure 9. Notably, the Ginzburg-Landau energy is not monotone decreasing along the accelerated Allen-Cahn equation, but the total energy (the sum of Ginzburg-Landau energy and kinetic energy) is. As in Figure 2, there is a singularity in the Ginzburg-Landau energy curve at the time that the shape 'disappears' briefly, but it is less pronounced since $\alpha$ is much larger in this simulation. Unsurprisingly, the energy decrease is much faster along the Allen-Cahn equation in 'physical' time.

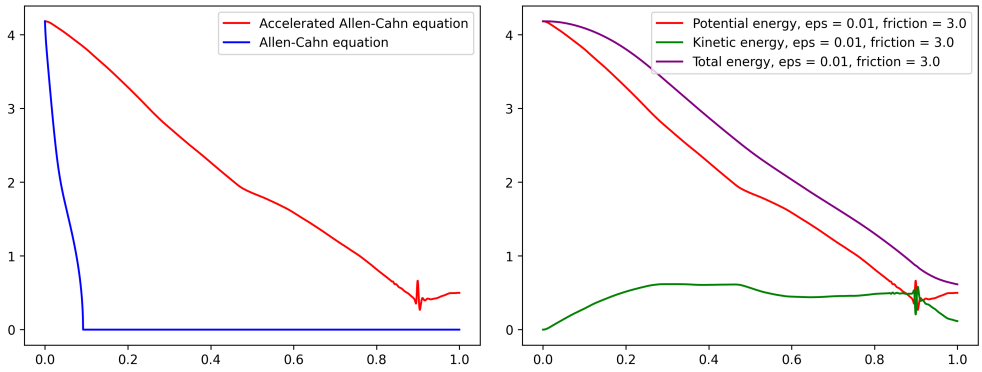


FIGURE 9. **Left:** The decrease of boundary length in the Allen-Cahn equation and 'accelerated Allen-Cahn equation' as measured by the Ginzburg-Landau functional. **Right:** The evolution of 'potential energy' (the Ginzburg-Landau functional), kinetic energy $\frac{\varepsilon}{2}\|u_t\|_{L^2}^2$ and the total energy (their sum). While the potential energy is not monotone decreasing, the total energy decreases, as is expected for a momentum-based optimizer.

In Figure 10, we consider the evolution of the same initial condition with the FISTA discretization and the much larger time step size $\tau = 1$ and with $\eta = \tau^2 = 1$ and $\rho = 1/(1+\alpha\tau)$ for $\alpha = 0.1$. In this regime, the momentum method geometrically resembles mean curvature more than the accelerated mean curvature flow and does not develop non-smooth interfaces. The curve shrinks significantly faster than under the convex-concave splitting discretization of the Allen-Cahn equation.

5.2. **Simulations on graphs.** In this section, we report simulations on graphs using the synthetic 'blobs' dataset [Den12] with a densely connected graph. The weights are assigned as $w_{ij} = \exp(-\|x_i - x_j\|^2/\sigma^2)$ in both cases, where $\sigma$ is suitably chosen and $\|\cdot\|$ denotes the Euclidean norm. The graph Laplacian in the experiment is the symmetric normalized graph Laplacian.

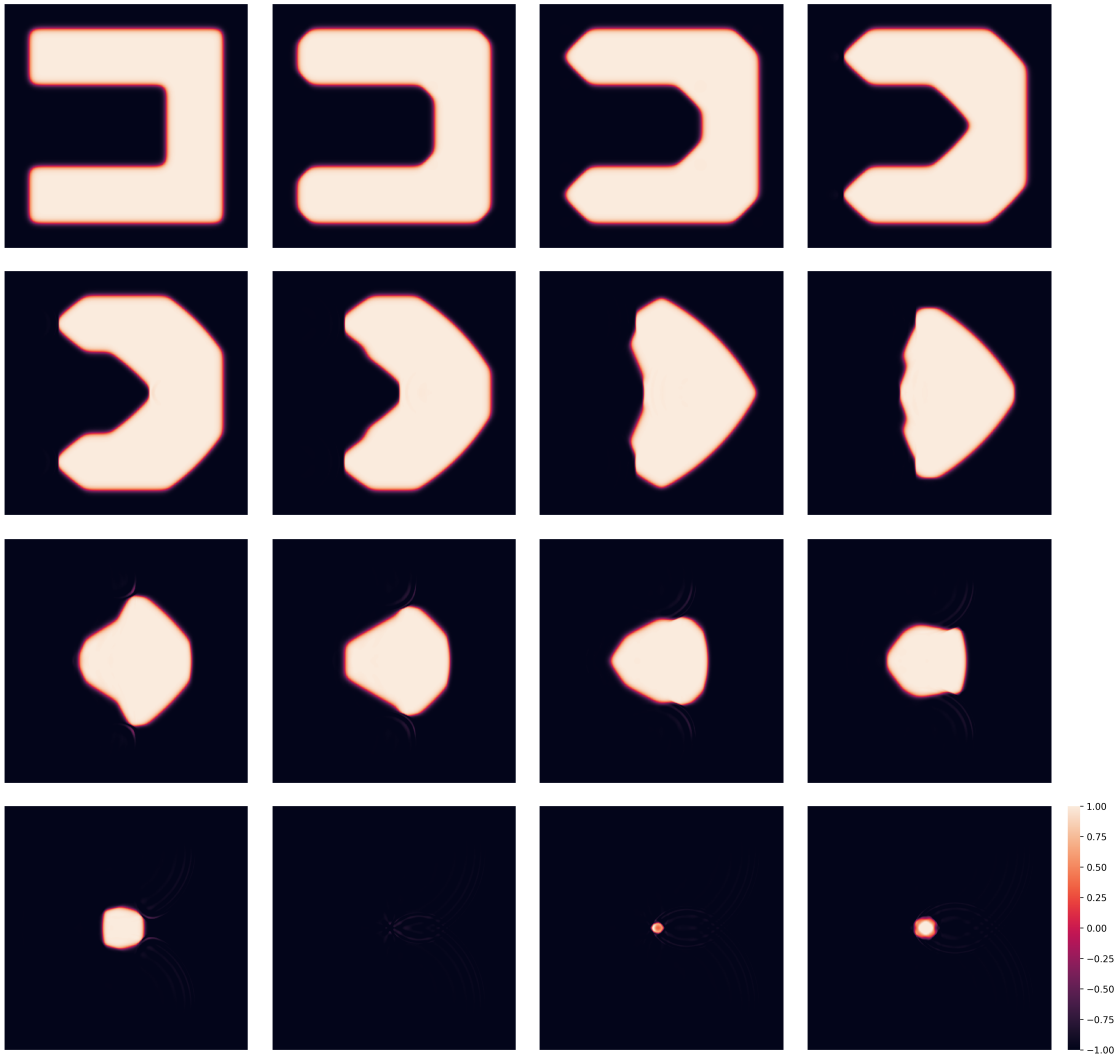


FIGURE 8. Evolution of a Jordan curve under the accelerated Allen-Cahn approximation to 'accelerated curve shortening flow' at times $t \in \{0, 0.05, 0.1, 0.15\}$ (top row, left to right), $t \in \{0.225, 0.3, 0.4, 0.45\}$ (second row) $t \in \{0.55, 0.6, 0.65, 0.7\}$ (third row) and $t \in \{0.80.9, 0.925, 0.95\}$ (bottom row).

5.2.1. *A toy example.* For the synthetic blobs dataset, we generate 2000 data points in $\mathbb{R}^2$, belonging to five distinct classes using a function from scikit-learn [PVG+11]. We consider two situations: One, in which the clusters ('blobs') are spatially separated and where correct classification based on Euclidean distances is easy (the standard deviation for the draw of the clusters is 1.5) [PVG+11].

In the second experiment, we use a cluster standard deviation of 2.5, resulting in a dataset with some overlap between clusters.

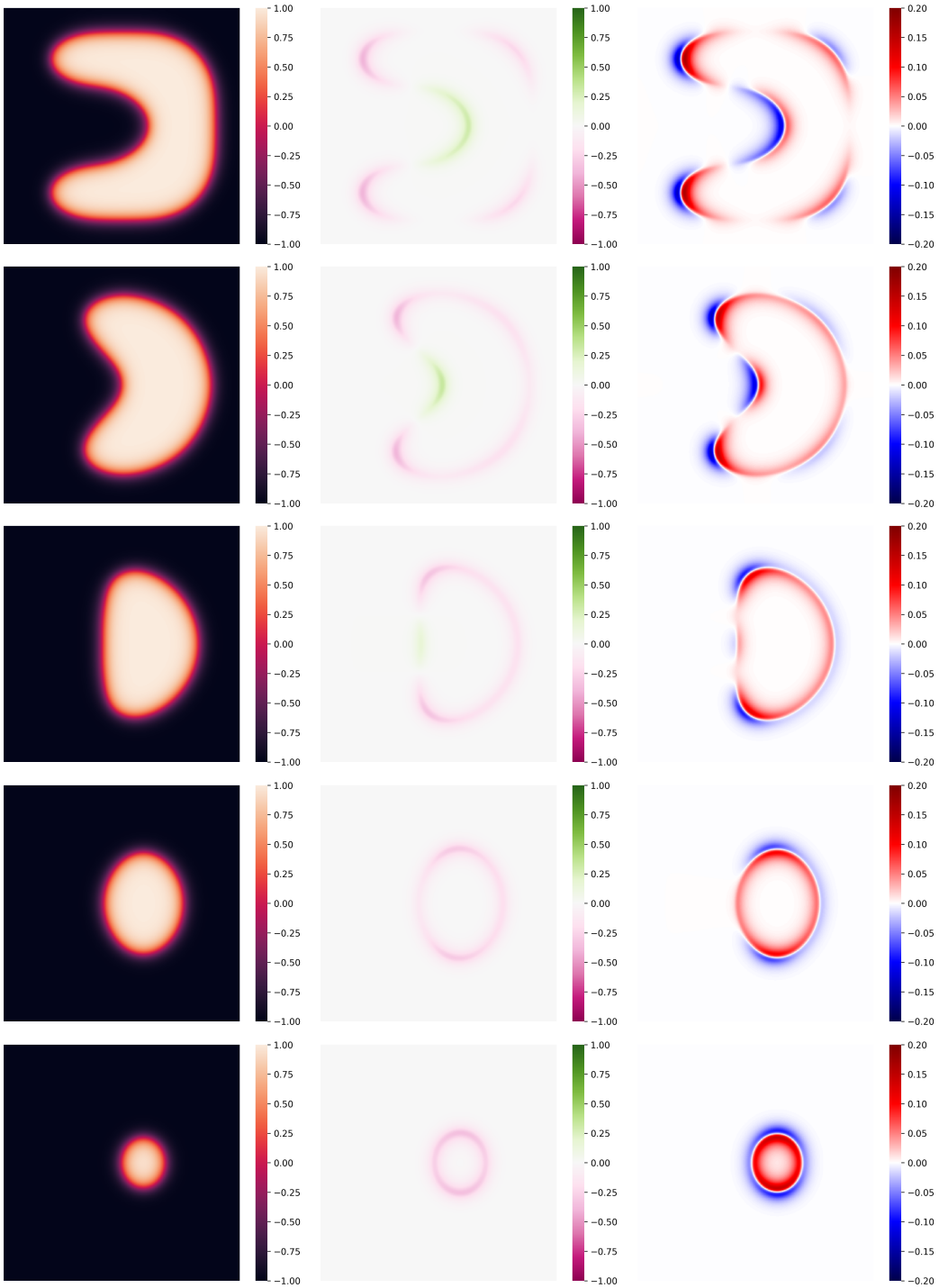

FIGURE 10. The evolution of $u$ under the FISTA-approximation to the accelerated Allen Cahn equation for $\alpha = 0.1$ with convex-concave splitting and large time steps ($\tau = 1$) after 10, 20, 30, 40 and 50 time steps (top to bottom). We give $u$ in the middle column, the velocity variable $v = u_t$ in the middle column and the energy gradient $\Delta u_{n+1} - \frac{2u_{n+1}+W'_{concave}(u_n)}{\varepsilon^2}$ in the right column.
The gradient variable accumulates in the velocity into a signed quantity driving the interface, but in the gradient descent step, it is an unsigned quantity which undoes the sharpening interface caused by the momentum step. As expected, the interface velocity remains below 1. Compared to the PDE, the large time-step scheme does not lead to singular interfaces with corners.

After generating the dataset, we randomly select 1% of the data points as points with known labels. At all other points, the label is predicted based on a phase parameter $u : D \to \mathbb{R}^k$, where $D$ is the data set and $k = 5$ is the number of classes. The labels at the known data points (encoded as one-hot vectors) are enforced as a boundary condition. The double-well potential is above.

We selected parameters $\sigma = 0.1$ for the matrix generating the Gaussian weights, $\varepsilon = 0.1$ for the parameter of the phase-field energy and a time-step size $\tau = 1.0$ for the convex-concave split FISTA discretization (accelerated Allen-Cahn equation) and the convex-concave split gradient descent algorithm (Allen-Cahn equation). The optimization algorithm ran for $T = 1,000$ iterations. As our main focus is the comparison between optimizers, the parameters associated with graph construction were not optimized over. As initial condition, we selected $u(x_i) = 1/k$ for all points with unknown labels.

In Figures 11 and 12, we visualize the blobs dataset with true labels and the solutions to the Allen-Cahn equation and accelerated Allen-Cahn equation.

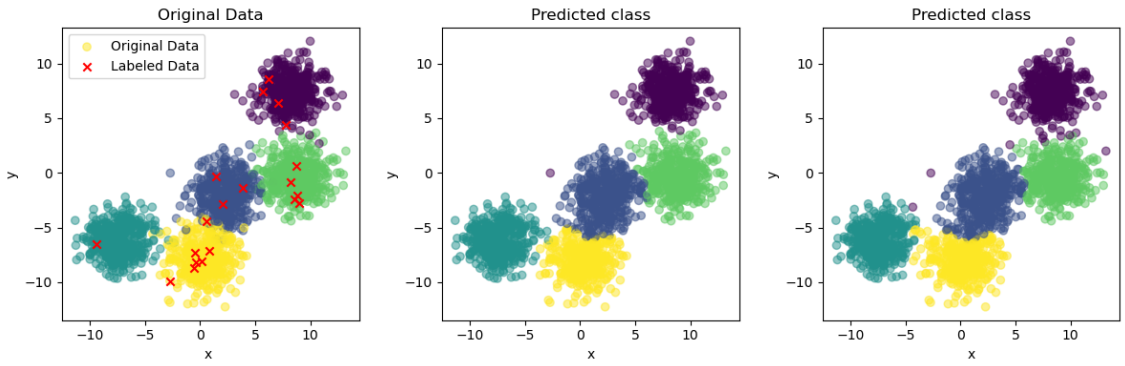


FIGURE 11. Performance comparison using the blobs datasets. **Left:** True labels. The initially labeled data points for the boundary condition are marked by red crosses. **Center:** The final state of the accelerated Allen-Cahn method. **Right:** The final state of the Allen-Cahn method.
In the absence of geometric evidence, points are predicted as class zero (purple) by the tie-breaking mechanism. Evidently, momentum enables the method to overcome the distance separating outlying datapoints from the bulk and more outliers are labeled correctly by the momentum method in the light green and turquoise clusters.

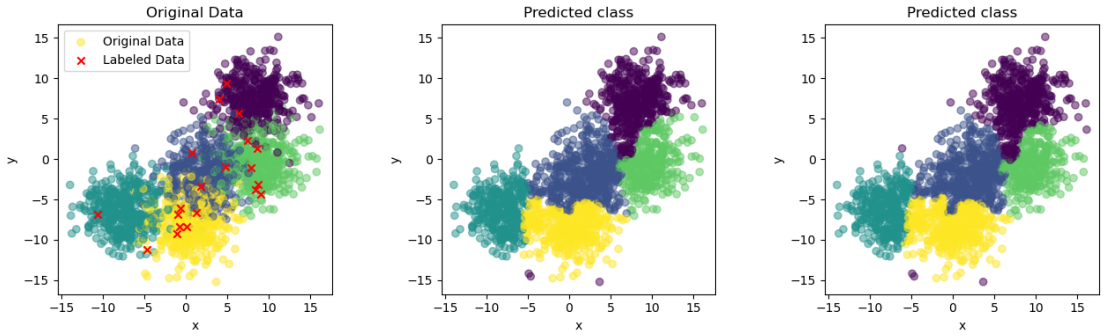


FIGURE 12. Performance comparison using the blobs datasets. **Left:** The original blobs data with an overlap. **Center:** The solution of the accelerated Allen-Cahn method. **Right:** The solution of the Allen-Cahn method.

In Figure 13, we visualize the classification accuracies by comparing the correct labels to the labels predicted by the Allen-Cahn method compared to the accelerated Allen-Cahn method on the for well-separated data as in 11. In Figure 14, we plot how the Ginzburg-Landau energy decreases for the two methods. Notably, FISTA with convex-concave splitting converges much faster than gradient descent with convex-concave splitting in both accuracy and energy.

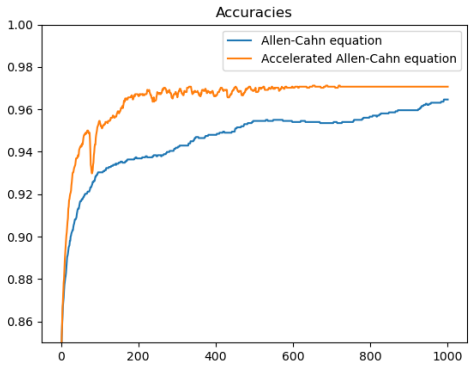


FIGURE 13. Accuracies of the Allen-Cahn and the accelerated Allen-Cahn method

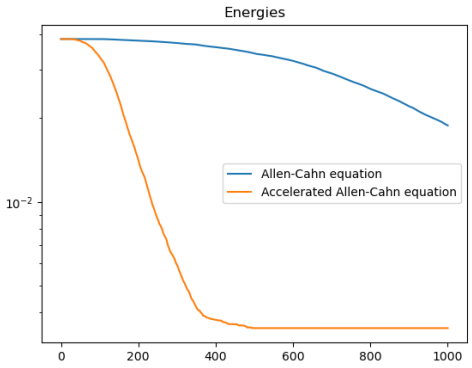


FIGURE 14. Energies of the Allen-Cahn and the accelerated Allen-Cahn method

In Figures 15 and 16, we describe the classification accuracies and Ginzburg-Landau energy on the data with overlap as in Figure 12 using the Allen-Cahn and accelerated Allen-Cahn methods also.

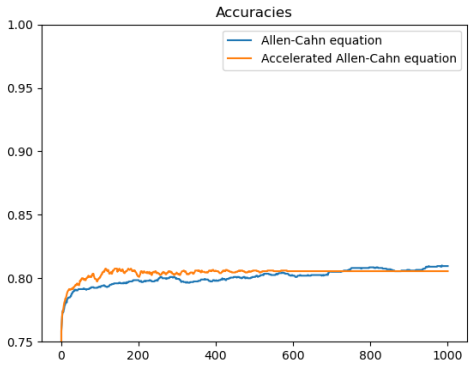


FIGURE 15. Accuracies of the gradient flow and the momentum method

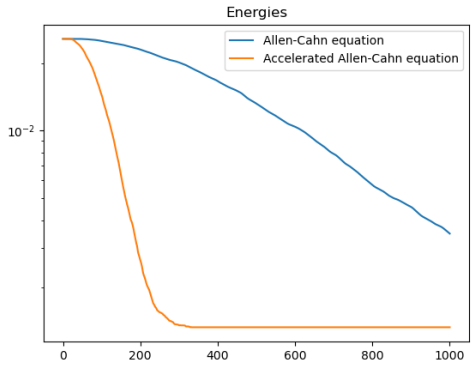


FIGURE 16. Energies of the Allen-Cahn and the Accelerated Allen-Cahn method

## 6. Conclusion and open questions

We have studied the 'accelerated Allen-Cahn equation' (momentum-based) in a setting where it is not guaranteed to outperform the regular Allen-Cahn equation (gradient flow). Many of our findings are reminiscent of observations in neural network training:

(1) While we observe convergence to the unique global minimizer (the zero function) in experiments with compactly supported initial condition, the trajectory which the flows take there are very different (in the parlance of machine learning, the 'implicit bias' of both algorithms is different in terms of which regions of space are explored).
(2) A large time step size appears to nudge also the momentum method towards preferring smoother interfaces, suggesting a potentially better 'implicit bias'. A similar phenomenon is observed in neural network training, where the implicit bias may be best at the 'edge of stability', i.e. for the largest admissible step size.

Neither analogy is perfect. Stochastic effects play a crucial role in neural network training, and arbitrarily large time steps are admissible with a convex-concave splitting. Additionally, the implicit bias studied in the training of neural networks typically concerns the long time limit of the trajectory of an optimization algorithm, while we are considering the full trajectory (and the long time limits coincide, at least in the experiments with compactly supported initial conditions in $\mathbb{R}^d$). Still, we maintain that

(1) momentum-based optimization is a promising avenue also for (non-convex) Ginzburg-Landau functionals,
(2) important questions remain about the appropriate time discretization (in particular, the stability of FISTA), and
(3) there are interesting parallels to momentum-based optimization in other high-dimensional settings.

In particular, the question remains: Can momentum-based optimization in a PDE setting give insight into momentum-based optimization in high-dimensional settings with less clear geometric interpretation?

## Acknowledgements

KG and SW gratefully acknowledge support through NSF grant 2307273 'Neural Networks for Stationary and Evolutionary Variational Problems'.

## Appendix A. Proofs for CINEMA

*Proof of Theorem 4.1.* For $x, v \in H$, we consider the function

$$M : H \to \mathbb{R}, \qquad M(z) = \frac{1}{2}\|z - x\|^2 + \langle \eta \nabla G(x) - \tau v, z\rangle + \eta F(z).$$

We observe that for for fixed $x$ and $v$, $M$ is bounded below since

$$\begin{aligned} M(z) &= \frac{1}{2}\|z\|^2 + \frac{1}{2}\|x\|^2 + \langle \eta \nabla G(x) - \tau v - x, z\rangle + \eta F(z) \\ &\geq \frac{1}{2}\|z\|^2 - \|\tau v + x\|\,\|z\| + \|\eta \nabla G(x)\|\|z\| + \eta\big(F(\xi) - \|\partial F(\xi)\|\,\|z - \xi\|\big) \end{aligned}$$

for some $\xi$ in the domain of $\partial F$. The expression on the right is bounded from below even if $F$ is not. Suppose that $z_n$ is a minimizing sequence of $M$, i.e. $\lim_n M(z_n) = \inf_{z\in H} M(z)$. Then the inequality above implies that $z_n$ is a bounded sequence. By the Banach-Alaoglu theorem and the fact that $H$ is a a separable Hilbert space, $z_n$ has a weakly convergent subsequence. For notational convenience, we can assume that $z_n$ itself is weakly convergent and $z^*$ is its weak limit. Then $\lim_n \langle \eta \nabla G(x) - \tau v, z_n\rangle = \langle \eta \nabla G(x) - \tau v, z^*\rangle$. Further, since $F$ and the norm $\|\cdot\|_H$ are both convex and lower semi-continuous, they are also weakly lower semi-continuous, i.e. $F(z^*) \leq \liminf_n F(z_n)$ and $\|z^* - x\|^2 \leq \liminf_n \|z_n - x\|^2$. Hence, we have shown that

$$M(z^*) \leq \liminf_n M(z_n) = \inf_z M(z) \leq M(z^*).$$

We can conclude that $z^*$ is the minimizer of $M$ and thus $0 \in \partial M(z^*)$, which is equivalent to

$$\frac{x - z^* + \tau v}{\eta} - \nabla G(x) \in \partial F(z^*). \qquad \square$$

*Proof of Theorem 4.2.* Existence of the scheme follows from Theorem 4.1. For convenience, we denote $s_{n+1} \in \partial F(x_{n+1})$ and $g_n = s_{n+1} + \nabla G(x_n)$. Using the first order convexity condition for $F$, we find that

$$F(x_{n+1}) - F(x_n) \leq \langle s_{n+1}, x_{n+1} - x_n\rangle = \langle s_{n+1}, \tau v_n - \eta g_n\rangle.$$

On the other hand, due to the concavity of $G$, we have

$$G(x_{n+1}) - G(x_n) \leq \langle \nabla G(x_n), x_{n+1} - x_n\rangle = \langle \nabla G(x_n), \tau v_n - \eta g_n\rangle.$$

Adding the last two inequalities,

$$F(x_{n+1}) + G(x_{n+1}) \leq F(x_n) + G(x_n) - \eta\|g_n\|^2 + \langle g_n, \tau v_n\rangle.$$

From this inequality and the definition of $v_{n+1}$ it follows that

$$\begin{aligned} e_{n+1} &= F(x_{n+1}) + G(x_{n+1}) + \frac{1}{2\rho^2}\|v_{n+1}\|^2 \\ &\leq F(x_n) + G(x_n) - \eta\|g_n\|^2 + \langle g_n, \tau v_n\rangle \\ &\quad + \frac{1}{2}\|v_n\|^2 + \frac{\tau^2}{2}\|g_n\|^2 - \langle \tau v_n, s_{n+1} + \nabla G(x_n)\rangle \\ &= F(x_n) + G(x_n) + \frac{1}{2}\|v_n\|^2 + \left(\frac{\tau^2}{2} - \eta\right)\|g_n\|^2 \\ &= e_n - \frac{1}{2}(\rho^{-2} - 1)\|v_n\|^2 + \left(\frac{\tau^2}{2} - \eta\right)\|g_n\|^2. \end{aligned}$$

□

## References


[ADA00] Luigi Ambrosio, Norman Dancer, and Giovanni Alberti. Variational models for phase transitions, an approach via Γ-convergence. *Calculus of variations and partial differential equations: topics on geometrical evolution problems and degree theory*, pages 95–114, 2000.

[BKMS20] Jessica Bosch, Steffen Klamt, Pedro Mercado, and Martin Stoll. Graph-based PDEs: Laplacians, eigeninformation, and semi-supervised learning. In *XXI Householder Symposium on Numerical Linear Algebra*, page 84, 2020.

[BKS18] Jessica Bosch, Steffen Klamt, and Martin Stoll. Generalizing diffuse interface methods on graphs: nonsmooth potentials and hypergraphs. *SIAM Journal on Applied Mathematics*, 78(3):1350–1377, 2018.

[BM19] Andrea L Bertozzi and Ekaterina Merkurjev. Graph-based optimization approaches for machine learning, uncertainty quantification and networks. In *Handbook of Numerical Analysis*, volume 20, pages 503–531. Elsevier, 2019.

[Bra15] Kenneth A Brakke. *The motion of a surface by its mean curvature.(MN-20)*, volume 20. Princeton University Press, 2015.

[Bre11] H Brezis. Functional analysis, sobolev spaces and partial differential equations, 2011.

[BT09] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[BvGL21] Jeremy Budd, Yves van Gennip, and Jonas Latz. Classification and image processing with a semi-discrete scheme for fidelity forced allen–cahn on graphs. *GAMM-Mitteilungen*, 44(1):e202100004, 2021.

[BvGL+23] Jeremy M Budd, Yves van Gennip, Jonas Latz, Simone Parisotto, and Carola-Bibiane Schönlieb. Joint reconstruction-segmentation on graphs. *SIAM Journal on Imaging Sciences*, 16(2):911–947, 2023.

[CCTS20] Jeff Calder, Brendan Cook, Matthew Thorpe, and Dejan Slepcev. Poisson learning: Graph based semi-supervised learning at very low label rates. In *International Conference on Machine Learning*, pages 1306–1316. PMLR, 2020.

[Cri19] R Cristoferi. Clustering of big data: Consistency of a nonlocal ginzburg-landau type model. *Rendiconti del Seminario Matematico*, 77(2):7–31, 2019.

[CY19] Jeff Calder and Anthony Yezzi. Pde acceleration: a convergence rate analysis and applications to obstacle problems. *Research in the Mathematical Sciences*, 6(4):35, 2019.

[DDE05] Klaus Deckelnick, Gerhard Dziuk, and Charles M Elliott. Computation of geometric partial differential equations and mean curvature flow. *Acta numerica*, 14:139–232, 2005.

[DE07] Gerhard Dziuk and Charles M Elliott. Finite elements on evolving surfaces. *IMA journal of numerical analysis*, 27(2):262–292, 2007.

[DEK22] Oliver RA Dunbar, Charles M Elliott, and Lisa Maria Kreusser. Models for information propagation on graphs. *arXiv preprint arXiv:2201.07577*, 2022.

[Den12] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[DK21] Steffen Dereich and Sebastian Kassing. Convergence of stochastic gradient descent schemes for lojasiewicz-landscapes. *arXiv preprint arXiv:2102.09385*, 2021.

[DKW19] Patrick W Dondl, Matthias W Kurzke, and Stephan Wojtowytsch. The effect of forest dislocations on the evolution of a phase-field model for plastic slip. *Archive for Rational Mechanics and Analysis*, 232(1):65–119, 2019.

[Dob10] Manfred Dobrowolski. *Angewandte Funktionalanalysis: Funktionalanalysis, Sobolev-Räume und elliptische Differentialgleichungen*. Springer-Verlag, 2010.

[EO15] Selim Esedoglu and Felix Otto. Threshold dynamics for networks with arbitrary surface tensions. *Communications on pure and applied mathematics*, 68(5):808–864, 2015.

[ES91] Lawrence C Evans and Joel Spruck. Motion of level sets by mean curvature. i. In *Fundamental Contributions to the Continuum Theory of Evolving Phase Interfaces in Solids: A Collection of Reprints of 14 Seminal Papers*, pages 328–374. Springer, 1991.

[Eva93] Lawrence C Evans. Convergence of an algorithm for mean curvature motion. *Indiana University mathematics journal*, pages 533–557, 1993.

[FLM16] Raffaele Folino, Corrado Lattanzio, and Corrado Mascia. Metastable dynamics for hyperbolic variations of the allen-cahn equation. *arXiv preprint arXiv:1607.06796*, 2016.

[FLS20] Julian Fischer, Tim Laux, and Theresa M Simon. Convergence rates of the Allen–Cahn equation to mean curvature flow: A short proof based on relative entropies. *SIAM Journal on Mathematical Analysis*, 52(6):6222–6233, 2020.

[FM24] Julian Fischer and Alice Marveggio. Quantitative convergence of the vectorial Allen–Cahn equation towards multiphase mean curvature flow. *Annales de l'Institut Henri Poincaré C*, pages 1–62, 2024.

[Fol17] Raffaele Folino. Slow motion for a hyperbolic variation of allen–cahn equation in one space dimension. *Journal of Hyperbolic Differential Equations*, 14(01):1–26, 2017.

[GT77] David Gilbarg and Neil S Trudinger. *Elliptic partial differential equations of second order*, volume 224. Springer, 1977.

[GTD23] Baptiste Goujaud, Adrien Taylor, and Aymeric Dieuleveut. Provable non-accelerations of the heavy-ball method. *arXiv preprint arXiv:2307.11291*, 2023.

[GW84] Enrico Giusti and Graham H Williams. *Minimal surfaces and functions of bounded variation*, volume 80. Springer, 1984.

[GW24] Kanan Gupta and Stephan Wojtowytsch. Nesterov acceleration in benignly non-convex landscapes. *arXiv preprint arXiv:2410.08395*, 2024.

[HADR24] J Hermant, J-F Aujol, C Dossal, and A Rondepierre. Study of the behaviour of nesterov accelerated gradient in a non convex setting: the strongly quasar convex case. *arXiv preprint arXiv:2405.19809*, 2024.

[Ilm93] Tom Ilmanen. Convergence of the Allen-Cahn equation to Brakke's motion by mean curvature. *Journal of Differential Geometry*, 38(2):417–461, 1993.

[LB17] Xiyang Luo and Andrea L Bertozzi. Convergence of the graph allen–cahn scheme. *Journal of Statistical Physics*, 167:934–958, 2017.

[LO20] Tim Laux and Felix Otto. The thresholding scheme for mean curvature flow and de Giorgi's ideas for minimizing movements. In *The role of metrics in the theory of partial differential equations*, volume 85, pages 63–94. Mathematical Society of Japan, 2020.

[LRP16] Laurent Lessard, Benjamin Recht, and Andrew Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016.

[MBC18] Ekaterina Merkurjev, Andrea L Bertozzi, and Fan Chung. A semi-supervised heat kernel pagerank MBO algorithm for data classification. *Communications in mathematical sciences*, 16(5):1241–1265, 2018.

[MBO92] Barry Merriman, James Kenyard Bence, and Stanley Osher. *Diffusion generated motion by mean curvature*. Department of Mathematics, University of California, Los Angeles, 1992.

[MBS20] Pedro Mercado, Jessica Bosch, and Martin Stoll. Node classification for signed social networks using diffuse interface methods. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part I*, pages 524–540. Springer, 2020.

[MM77] Luciano Modica and Stefano Mortola. Un esempio di Γ-convergenza. *Boll. Un. Mat. Ital. B*, 14:285–299, 1977.

[Mod87] Luciano Modica. The gradient theory of phase transitions and the minimal interface criterion. *Archive for Rational Mechanics and Analysis*, 98:123–142, 1987.

[MR11] Luca Mugnai and Matthias Röger. Convergence of perturbed Allen-Cahn equations to forced mean curvature flow. *Indiana University Mathematics Journal*, pages 41–75, 2011.

[Nes83] Yurii Evgen'evich Nesterov. A method of solving a convex programming problem with convergence rate $o(1/k^2)$. *Doklady Akademii Nauk*, 269(3):543–547, 1983.

[Nes18] Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

[NGA16] IG Nizovtseva, PK Galenko, and DV Alexandrov. The hyperbolic allen–cahn equation: exact solutions. *Journal of Physics A: Mathematical and Theoretical*, 49(43):435201, 2016.

[Pol64] Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17, 1964.

[PVG+11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay. *sklearn.datasets.make_blobs*. scikit-learn, 2011. Accessed: 2024-10-07.

[SBC14] Weijie Su, Stephen Boyd, and Emmanuel Candès. A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. *Advances in neural information processing systems*, 27, 2014.

[Sch18] Hayden Schaeffer. A penalty method for some nonlinear variational obstacle problems. *Communications in Mathematical Sciences*, 16(7):1757–1777, 2018.

[Ser11] Sylvia Serfaty. Gamma-convergence of gradient flows on hilbert and metric spaces and applications. *Discrete Contin. Dyn. Syst*, 31(4):1427–1451, 2011.

[SS04] Etienne Sandier and Sylvia Serfaty. Gamma-convergence of gradient flows with applications to ginzburg-landau. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 57(12):1627–1672, 2004.

[SW23] Jonathan W Siegel and Stephan Wojtowytsch. A qualitative difference between gradient flows of convex functions in finite-and infinite-dimensional Hilbert spaces. *arXiv preprint arXiv:2310.17610*, 2023.

Oluwatosin Akande, Industrial and Systems Engineering, Lehigh University, 200 West Packer Avenue, Bethlehem, PA 18015, USA

*Email address*: oaa323@lehigh.edu

Patrick Dondl, Abteilung für Angewandte Mathematik, Albert-Ludwigs-Universität Freiburg, Hermann-Herder-Strasse 10, 79104 Freiburg i. Br., Germany

*Email address*: patrick.dondl@mathematik.uni-freiburg.de

Kanan Gupta, University of Pittsburgh, Department of Mathematics, Thackeray Hall, Pittsburgh, PA 15213, USA

*Email address*: kanan.g@pitt.edu

Akwum Onwunta, Industrial and Systems Engineering, Lehigh University, 200 West Packer Avenue, Bethlehem, PA 18015, USA

*Email address*: ako221@lehigh.edu

Stephan Wojtowytsch, University of Pittsburgh, Department of Mathematics, Thackeray Hall, Pittsburgh, PA 15213, USA

*Email address*: s.woj@pitt.edu