A mean curvature flow arising in adversarial training

Leon Bungert* Tim Laux† Kerrek Stinson ‡

April 23, 2024

Abstract

We connect adversarial training for binary classification to a geometric evolution equation for the decision boundary. Relying on a perspective that recasts adversarial training as a regularization problem, we introduce a modified training scheme that constitutes a minimizing movements scheme for a nonlocal perimeter functional. We prove that the scheme is monotone and consistent as the adversarial budget vanishes and the perimeter localizes, and as a consequence we rigorously show that the scheme approximates a weighted mean curvature flow. This highlights that the efficacy of adversarial training may be due to locally minimizing the length of the decision boundary. In our analysis, we introduce a variety of tools for working with the subdifferential of a supremal-type nonlocal total variation and its regularity properties.

Keywords: mean curvature flow, adversarial training, adversarial machine learning, minimizing movements, monotone and consistent schemes
AMS subject classifications: 28A75, 35D40, 49J45, 53E10, 68T05

1 Introduction

In the last decade, machine learning algorithms and in particular deep learning have experienced an unprecedented success story. Such methods have proven their capabilities, inter alia, for the difficult tasks of image classification and generation. Most recently, the advent of large language models is expected to have a strong impact on various aspects of society.

At the same time, the success of machine learning is accompanied by concerns about the reliability and safety of its methods. Already more than ten years ago it was observed that neural networks for image classification are susceptible to adversarial attacks [35], meaning that imperceptible or seemingly harmless perturbations of images can lead to severe misclassifications. As a consequence, the deployment of such methods in situations that affect the integrity and safety of humans, e.g., for self-driving cars or medical image classification, is risky.

To mitigate these risks, the scientific community has been developing different approaches to robustify machine learning in the presence of potential adversaries. The most prominent of these approaches in the context of classification tasks is *adversarial training* [23, 27], which is

*Institute of Mathematics & Center for Artificial Intelligence and Data Science (CAIDAS), University of Würzburg, Emil-Fischer-Str. 40, 97074 Würzburg, Germany. Email: leon.bungert@uni-wuerzburg.de

†Faculty of Mathematics, University of Regensburg, Universitätsstraße 31, 93053 Regensburg, Germany. Email: tim.laux@ur.de

‡Hausdorff Center for Mathematics, University of Bonn, Endenicher Allee 62, 53115 Bonn, Germany. Email: kerrek.stinson@hcm.uni-bonn.de

a robust optimization problem of the form

$$\inf_{u\in\mathcal{H}} \mathbb{E}_{(x,y)\sim\mu}\left[\sup_{\tilde{x}\in B(x,\varepsilon)} \ell(u(\tilde{x}),y)\right], \tag{1.1}$$

where we use the notation $\mathbb{E}_{z\sim\mu}[f(z)] := \int f(z)\,\mathrm{d}\mu(z)$. The ingredients of adversarial training are readily explained: The probability measure $\mu \in \mathcal{P}(\mathcal{X}\times\mathcal{Y})$ models the distribution of given training pairs in the so-called feature space $\mathcal{X}$ and the label space $\mathcal{Y}$. Here $\mathcal{X}$ is a metric space, and $\mathcal{Y}$ a set, e.g., $\mathcal{Y} = \{0,\dots,K-1\}$ describing $K \in \mathbb{N}$ classes. In realistic situations, one uses an empirical distribution of the form $\mu = \frac{1}{M}\sum_{i=1}^{M}\delta_{(x_i,y_i)}$ where $(x_i,y_i) \in \mathcal{X}\times\mathcal{Y}$ for $i = 1,\dots,M$. The optimization takes place in a so-called hypothesis class $\mathcal{H}$ which is nothing but a class of functions from $\mathcal{X}$ to $\mathcal{Y}$, e.g., linear functions, measurable functions, parametrized neural networks, etc. We let $\ell : \mathcal{Y}\times\mathcal{Y} \to \mathbb{R}$ be a so-called loss function, which is often chosen as a power of a norm or f-divergence. Finally, in essence, the optimal classifier u should satisfy $u(\tilde{x}) \approx y$ for μ-almost every $(x,y) \in \mathcal{X}\times\mathcal{Y}$ and all $\tilde{x} \in B(x,\varepsilon)$. Thereby, (1.1) enforces robustness of the classification in ε-balls around the data points, where $\varepsilon > 0$ is called the *adversarial budget*.

Already in [27] it has been empirically observed that (1.1) indeed allows one to compute neural networks that are significantly more robust than those trained with the standard approach (corresponding to $\varepsilon = 0$ in (1.1)). However, the mathematical understanding of adversarial training and related problems only began growing a few years ago: One line of research connects (1.1) with (multimarginal) optimal transport or distributionally robust optimization problems [21, 32, 33] and uses tools from these disciplines to analyse adversarial training. Existence of solutions to (1.1) in the binary classification case where $\mathcal{Y} = \{0,1\}$, ℓ is the 0-1 loss $\ell(\tilde{y},y) = |\tilde{y}-y|$, and $\mathcal{H}$ is a class of measurable functions was proved in [2, 7]. In [2], the authors consider closed balls $B(x,\varepsilon)$ in (1.1) and work with classifiers which are characteristic functions of universally measurable sets in $\mathbb{R}^N$. In contrast, in [7] open balls are used and the classifiers are characteristic functions of Borel measurable subsets of a generic metric measure space. The authors of [7] also proved that adversarial training for binary classification is equivalent to the following variational regularization problem:

$$\inf_{A\in\mathfrak{B}(\mathcal{X})} \mathbb{E}_{(x,y)\sim\mu}\left[|\mathbf{1}_A(x) - y|\right] + \varepsilon\,\mathrm{Per}_\varepsilon(A). \tag{1.2}$$

Here, Per_ε is a non-local and data-dependent perimeter functional that regularizes the decision boundary ∂A between the two classes. A similar decomposition into a "natural error" and a "boundary error" was studied in [38] and used to derive the TRADES algorithm, which essentially replaces the regularization parameter ε in (1.2) by $\frac{1}{\lambda}$ for $\lambda > 0$. Also for other notions of robustness which are weaker than adversarial robustness, geometric interpretations similar to (1.2) exists, see [7, Section 4] or [6]. We also remark that generalizations of some of the results in [2, 7] to the case of multi-class classification can be found in [19]. Finally, an overview of recent mathematical developments in the field can be found in [20].

The perspective in (1.2) opens the door for the geometric analysis of adversarially robust classifiers. As a first step in this direction, in [7] it was shown that maximal and minimal minimizers of (1.2) possess one-sided regularity properties, and that for $\mathcal{X} = \mathbb{R}^N$ there exists a solution with a boundary that is locally the graph of a $C^{1,1/3}$ function. Subsequently, it was shown in [8] that (even for discontinuous densities with bounded variation) the nonlocal perimeter Per_ε Gamma-converges to a weighted local perimeter and, as a consequence, solutions of (1.2) converge to perimeter-minimal solutions of (1.2) with $\varepsilon = 0$. In [22] it was shown that for sufficiently small $\varepsilon > 0$ adversarially robust classifiers evolve (as parametrized by ε) according

to a geometric flow, when smooth solutions starting from the Bayes classifier exist. Expansions used to derive this result show that, infinitesimally in ε, this flow is a weighted mean curvature flow, which shows that adversarial training is connected to decreasing the length of the decision boundary.

The **main contribution** of the present paper is to make this connection with mean curvature flow rigorous in a general setting and to move beyond the short time regime of [22]. To achieve this, we will introduce a slight modification of the adversarial training problem (1.2). Intuitively, the proposed iterative scheme prepares for attacks by an adversary with *total adversarial budget* $T > 0$ and (instantaneous) adversarial budget $\varepsilon > 0$, allowing the adversary to corrupt the data on scale ε and even to react to modified classifiers at most T/ε times. As we will see in Section 2.1 below, the scheme can be interpreted as a minimizing movements scheme for mean curvature flow, in the spirit of Almgren–Taylor–Wang [1]. To select unique solutions we consider a strongly convex Chambolle-type scheme [11] and prove that it is monotone and consistent with respect to a weighted mean curvature flow, thereby proving convergence of the scheme to smooth flows (Theorem 1).

The **main challenge** and the reason why our results are not just straightforward extensions of existing ones is that the adversarial budget $\varepsilon > 0$ in (1.2) acts both as a time step and as a non-locality parameter for the perimeter Per_ε. Hence, in order to prove consistency with mean curvature flow, we have to perform a careful analysis of the associated total variation functional and its subdifferential, showing that the latter is consistent with the 1-Laplace operator for a suitable class of functions.

We would like to emphasize that adversarial training is not the only method in data science connected to mean curvature flow. In particular, in the field of graph-based learning the so-called Merriman–Bence–Osher (MBO) algorithm has been employed frequently for clustering data sets or solving semi-supervised learning problems, see, e.g., [9, 28, 29, 36]. For rigorous connections of such approaches to mean curvature flow we refer to [24, 25].

Organization of the paper. The rest of the paper is organized as follows. In the next section we precisely introduce the proposed adversarial training scheme and state our main result—convergence of the method to weighted mean curvature flow. In Section 3, we deduce the needed properties for the nonlocal total variation and, in particular, study its subdifferential. Finally, in Section 4 we prove convergence of the adversarial training scheme by verifying that it is monotone and consistent with respect to weighted mean curvature flow.

Notation. For the reader's convenience, we collect notation used throughout the paper here. Typically, $\Omega \subset \mathbb{R}^N$ will be a bounded domain (i.e., a non-empty, open, and connected set). We use $\mathcal{L}^N$ to denote the N-dimensional Lebesgue measure and $\mathfrak{B}(\Omega)$ to denote the Borel measurable subsets of Ω. Furthermore, we use $|\cdot|$ for the Euclidean norm of a vector in $\mathbb{R}^N$ and $\mathbb{1}$ for the $N \times N$ identity matrix. For a set $A \subset \mathbb{R}^N$, we let $\mathbf{1}_A$ be the characteristic function taking the value 1 on A and 0 otherwise. For any set $\Omega \subset \mathbb{R}^N$, we define the inner parallel set of distance $a > 0$ through

$$\Omega_a := \{x \in \Omega \,:\, \mathrm{dist}(x, \mathbb{R}^N \setminus \Omega) > a\}. \tag{1.3}$$

Finally, for $x \in \mathbb{R}^N$ and $\varepsilon > 0$, we denote open balls by $B(x, \varepsilon) := \{y \in \mathbb{R}^N \,:\, |x - y| < \varepsilon\}$.

2 From adversarial training to mean curvature flow

Let $\Omega \subset \mathbb{R}^N$ be a bounded domain, $\mu \in \mathcal{P}(\Omega \times \{0, 1\})$ be a probability measure, and $\ell(\bar{y}, y) = 1_{\bar{y} \neq y}$ be the 0-1 loss function. We are interested in binary classifiers found via adversarial

training (1.1), i.e., minimizers of the problem

$$\inf_{A\in\mathfrak{B}(\Omega)} \mathbb{E}_{(x,y)\sim\mu}\left[\sup_{\tilde{x}\in B(x,\varepsilon)\cap\Omega} \ell(\mathbf{1}_A(\tilde{x}),y)\right]. \tag{2.1}$$

We define the conditional distributions $\varrho_i := \mu(\cdot\times\{i\})$ for $i=0,1$ and $\varrho := \varrho_0+\varrho_1$. We pose the following assumption which we shall use in the whole paper, without further reference.

Assumption 1 (The densities)**.** We assume that ϱ_0 and ϱ_1 (and hence also ϱ) have densities with respect to the N-dimensional Lebesgue measure on Ω which are continuously differentiable functions, i.e., $\varrho_i\in C^1(\Omega)$. For notational convenience we shall identify ϱ and ϱ_i with their densities, meaning $\int f\,\mathrm{d}\varrho_{(i)} = \int f\,\varrho_{(i)}\,\mathrm{d}x$. Furthermore, we assume that $c_\varrho < \varrho < c_\varrho^{-1}$ in Ω for some constant $c_\varrho>0$.

In this situation it follows from the general results in [7] that problem (2.1) is equivalent to

$$\inf_{A\in\mathfrak{B}(\Omega)} \iint_{\Omega\times\{0,1\}} |\mathbf{1}_A(x)-y|\,\mathrm{d}\mu(x,y) + \varepsilon\,\mathrm{Per}_\varepsilon(A), \tag{2.2}$$

where the generalized perimeter functional Per_ε is defined as

$$\mathrm{Per}_\varepsilon(A) := \frac{1}{\varepsilon}\left[\int_\Omega \left(\operatorname*{ess\,sup}_{B(x,\varepsilon)\cap\Omega} \mathbf{1}_A - \mathbf{1}_A(x)\right)\mathrm{d}\varrho_0(x) + \int_\Omega \left(\mathbf{1}_A(x) - \operatorname*{ess\,inf}_{B(x,\varepsilon)\cap\Omega}\mathbf{1}_A\right)\mathrm{d}\varrho_1(x)\right].$$

Note that, in particular, the supremum in (2.1) can be replaced by essential suprema and infima in (2.2). Furthermore, it was proved in [7] that minimizers to both problems (2.1) and (2.2) exist and that the infimal values coincide. Studying the limit of the problem with small adversarial budget, the first and third author showed in [8] that the perimeter functional Γ-converges as $\varepsilon\to 0$ to a weighted but local perimeter. In the current setting with smooth densities this local perimeter is given by $\int_{\partial^* A\cap\Omega}\varrho\,\mathrm{d}\mathcal{H}^{N-1}$, where $\mathcal{H}^{N-1}$ is the Hausdorff (surface) measure and $\partial^* A$ is the measure-theoretic reduced boundary of A. Therefore, for small values of ε the problem (2.2) will effectively minimize the energy

$$\frac{1}{\varepsilon}\iint_{\Omega\times\{0,1\}} |\mathbf{1}_A(x)-y|\,\mathrm{d}\mu(x,y) + \int_{\partial^* A\cap\Omega}\varrho\,\mathrm{d}\mathcal{H}^{N-1},$$

which bears a strong resemblance to the Almgren–Taylor–Wang scheme introduced in [1] for the study of mean curvature flow, with $\varepsilon>0$ acting as the time step size. Consequently, the minimization problem (2.2) should roughly be approximated by a mean curvature flow. As remarked in the introduction, similar conclusions were drawn in [22] on short time horizons.

The natural initial condition for the mean curvature flow is any solution of the adversarial training problem (2.2) with $\varepsilon=0$:

$$\inf_{A\in\mathfrak{B}(\Omega)} \iint_{\Omega\times\{0,1\}} |\mathbf{1}_A(x)-y|\,\mathrm{d}\mu(x,y). \tag{2.3}$$

Solutions are called Bayes classifiers and since we have

$$\iint_{\Omega\times\{0,1\}} |\mathbf{1}_A(x)-y|\,\mathrm{d}\mu(x,y) = \int_\Omega \mathbf{1}_A\,\mathrm{d}\varrho_0 + \int_\Omega 1-\mathbf{1}_A\,\mathrm{d}\varrho_1 = -\int_\Omega \mathbf{1}_A\,\mathrm{d}(\varrho_1-\varrho_0) + \varrho_1(\Omega),$$

problem (2.3) is solved by every set A which is the positive part of a Hahn decomposition of the signed measure $\varrho_1-\varrho_0$. For continuous densities ϱ_0,ϱ_1 any set A which is sandwiched as $\{\varrho_1>\varrho_0\}\subset A\subset\{\varrho_1\geq\varrho_0\}$ is a Bayes classifier.

2.1 The minimizing movements scheme

Now we introduce an iterative adversarial training scheme starting from the Bayes classifier which is a slight modification of (2.2) and has a rigorous connection to mean curvature flow. Precisely, we replace (2.2) by a minimizing movements scheme in the spirit of [1, 26]:

$$\begin{cases} A_0 & \in \operatorname*{arg\,min}_{A\in\mathfrak{B}(\Omega)} \iint_{\Omega\times\{0,1\}} |\mathbf{1}_A(x) - y| \,\mathrm{d}\mu(x,y), \\ A_{k+1} & \in \operatorname*{arg\,min}_{A\in\mathfrak{B}(\Omega)} \int_\Omega |\mathbf{1}_A - \mathbf{1}_{A_k}| \,\frac{\operatorname{dist}(\cdot,\partial A_k)}{\varepsilon} \,\mathrm{d}\varrho + \operatorname{Per}_\varepsilon(A), \qquad k \geq 0, \end{cases} \tag{2.4}$$

where in this special case we "overload" the distance function and define

$$\operatorname{dist}(\cdot,\partial A) := \operatorname{dist}(\cdot, A^1) + \operatorname{dist}(\cdot, \Omega\setminus A^1),$$

with A^1 being the points in A with Lebesgue density 1, as the distance to the boundary of A relative to Ω. The representative set A^1 ensures that the distance function does not change when A is modified by a Lebesgue null-set, and we further note that the function coincides with the distance to $\partial(A^1)\cap\Omega$ when Ω is convex.

We note that this departs from the original adversarial training problem derived in (2.2) by the inclusion of a distance function. At a technical level, this is essential to recover the correct surface velocity for the boundary of the regularized classifier. Furthermore, one can show as in [15, Theorem 5.6] that, if A_0 is a smooth set and $\varepsilon > 0$ is small, the scheme (2.4) *without* the distance function would stagnate, i.e., $A_k = A_0$ for all $k \in \mathbb{N}$. At the level of the application, we motivate this term in the following remark.

Remark 2.1 (The distance function). In the context of training a stable classifier the term $\frac{\operatorname{dist}(\cdot,\partial A_k)}{\varepsilon}$ acts as an adaptive regularization parameter: For points far away from the decision boundary ∂A_k of the previous classifier, the perimeter regularization is unimportant and the first term in (2.4) gets more weight. Close to the boundary, the opposite holds true. If one just performs two iterations of (2.4), the first A_0 equals a Bayes classifier and the second one A_1 a solution to adversarial training, where the class labels are distributed according to the Bayes classifier and weighted according to their distance to the respective other class. Computing this distance function in practice can be done with several different methods, for instance with the fast marching algorithm [34] or the heat flow [16] based on Varadhans formula [37]. In the high-dimensional settings that are characteristic for machine learning problems, such methods are expensive which is why one resorts to so-called fast minimum norm attacks [31] which computes an approximation of the radius of the smallest ball around a data point which contains an adversarial attack. For binary classifiers as in (2.4) this is precisely the distance function to the decision boundary.

As the solutions of (2.4) are not necessarily unique, we consider a selection procedure following Chambolle's approach in [11]. To this end, let us introduce the signed distance function of a set A relative to Ω as

$$\operatorname{sdist}(\cdot, A) := \operatorname{dist}(\cdot, A^1) - \operatorname{dist}(\cdot, \Omega\setminus A^1), \tag{2.5}$$

where as before A^1 denotes the points in A with Lebesgue density 1. Furthermore, we introduce the total variation of a measurable function $u : \Omega \to \mathbb{R}$ which is naturally associated with $\operatorname{Per}_\varepsilon$:

$$\operatorname{TV}_\varepsilon(u) := \frac{1}{\varepsilon}\left[\int_\Omega \left(\operatorname*{ess\,sup}_{B(x,\varepsilon)\cap\Omega} u - u(x)\right) \mathrm{d}\varrho_0(x) + \int_\Omega \left(u(x) - \operatorname*{ess\,inf}_{B(x,\varepsilon)\cap\Omega} u\right) \mathrm{d}\varrho_1(x)\right]. \tag{2.6}$$

By definition it holds $\mathrm{Per}_\varepsilon(A) = \mathrm{TV}_\varepsilon(\mathbf{1}_A)$ and furthermore the perimeter and the total variation are connected through a coarea formula, see [7] and Lemma 3.1 below. The central object of study in this paper is the following modified adversarial training scheme

$$\begin{cases} A_0 & \in \operatorname*{arg\,min}_{A\in\mathfrak{B}(\Omega)} \iint_{\Omega\times\{0,1\}} |\mathbf{1}_A(x) - y| \,\mathrm{d}\mu(x,y), \\ w^* & := \operatorname*{arg\,min}_{w\in L^2(\Omega)} \frac{1}{2\varepsilon}\int_\Omega |w - \mathrm{sdist}(\cdot, A_k^c)|^2 \,\mathrm{d}\varrho + \mathrm{TV}_\varepsilon(w), \qquad k\geq 0, \\ A_{k+1} & := \{w^* > 0\}, \qquad k \geq 0. \end{cases} \tag{2.7}$$

We will prove that (2.7) constitutes a selection mechanism for (2.4); that is the sequence of sets $(A_k)_{k\in\mathbb{N}_0}$ found via (2.7) satisfies (2.4). We note that, in contrast to the work of Chambolle [11], who in our notation considered the scheme $A_{k+1} := \{w^* \leq 0\}$ where $w^* := \arg\min_{w\in L^2(\Omega)} \frac{1}{2\varepsilon}\int_\Omega |w - \mathrm{sdist}(\cdot, A_k)|^2 \,\mathrm{d}x + \mathrm{TV}(w)$ and TV is the standard total variation, we have to flip the sign of the signed distance function and pick the superlevel instead of sublevel set of the resulting minimizer w^* in order for (2.7) to select a solution of (2.4). This is necessary since TV_ε sees orientation, in the sense that $\mathrm{TV}_\varepsilon(-u) \neq \mathrm{TV}_\varepsilon(u)$, in contrast to the standard total variation, for which $\mathrm{TV}(-u) = \mathrm{TV}(u)$.

The objective of this paper is to show that, as the adversarial budget vanishes, meaning $\varepsilon \to 0$, the sequence of sets given by (2.7) converge to a time-parametrized curve $t \mapsto A(t)$ which is a solution of a weighted mean curvature flow equation with the following normal velocity (in the direction $\nu_{A(t)}$)

$$V(t) = -\frac{1}{\varrho}\,\mathrm{div}\left(\varrho\nu_{A(t)}\right) = H_{A(t)} - \nabla\log\varrho\cdot\nu_{A(t)} \quad \text{ on } \partial A. \tag{2.8}$$

Here $\nu_{A(t)}$ is (a smooth unit-length extension of) the inner unit normal to $\partial A(t)$ and $H_{A(t)} := -\,\mathrm{div}\,\nu_{A(t)}$ denotes the mean curvature of $\partial A(t)$. Note our orientation is such that $H_A > 0$ if A is a ball. The convergence to this mean curvature flow is the content of Theorem 1. The mathematical challenges arising in this problem are mostly consequences of the nonlocal TV_ε in (2.7): First, as the TV_ε functional is neither local nor smooth, we will need to carefully study its subdifferential and consistency with the 1-Laplace operator, i.e., the subdifferential of the classical total variation functional. Beyond this, we have not been able to show that minimizers w^* from (2.7) inherit the regularity of their data, e.g., $\mathrm{Lip}(w^*) \leq \mathrm{Lip}(\mathrm{sdist}(\cdot, A^c)) = 1$, an extremely convenient property to have at hand. Circumnavigating this obstacle, we instead prove that minimizers are "almost" Lipschitz by explicitly constructing sub- and supersolutions for conical data. Finally, in (2.7), the parameter ε (appearing in $\frac{1}{2\varepsilon}$ **and** in Per_ε) effectively behaves as the time-step in the discretization of a time interval $(0,T)$ and as a non-locality parameter. Consequently, the non-locality and time-step are of the same magnitude, and we must ensure that this does not prevent localization of the minimizing movements scheme in the limit as $\varepsilon \to 0$.

2.2 Main result

The main consequence of our results is that if the initial Bayes classifier is smooth and compactly contained in Ω, then a time parametrized version of the scheme $(A_k)_{k\in\mathbb{N}_0}$ given in (2.7) converges to a solution of mean curvature flow with initial condition A_0. Precisely, we parametrize the sets $(A_k)_{k\in\mathbb{N}_0}$ in (2.7) with a piecewise-constant curve $t \mapsto A_\varepsilon(t)$ defined by

$$A_\varepsilon(t) := A_k \text{ for } t \in [k\varepsilon, (k+1)\varepsilon). \tag{2.9}$$

With this at hand, we may state our result.

Theorem 1 (Main theorem)**.** *Let $\Omega \subset \mathbb{R}^N$ be a bounded and convex domain. Suppose that in* (2.7) *the Bayes classifier $A_0 \subset\subset \Omega$ has C^2-boundary and that $t \mapsto A(t)$ is a parameterized curve evolving by the weighted mean curvature flow with normal velocity* (2.8) *up to the first singular time T_*, with initial condition A_0.*

Then as $\varepsilon \to 0$, the time parametrized curves $t \mapsto A_\varepsilon(t)$ defined in (2.9)*, coming from the adversarial training scheme* (2.7)*, converge in $L^\infty_{\mathrm{loc}}([0,T_*); L^1(\Omega;\{0,1\}))$ and in the Hausdorff distance to the weighted mean curvature flow parametrized by $t \mapsto A(t)$.*

A couple of remarks on this theorem are in order.

Remark 2.2 (Smooth Bayes classifiers). Note that existence of Bayes classifiers with C^2-boundary is guaranteed, e.g., if the levelset $\{\varrho_0 = \varrho_1\}$ is a C^2-hypersurface in $\mathbb{R}^N$. This follows from the implicit function theorem if ϱ_0, ϱ_1 are C^2-regular in a neighborhood of $\{\varrho_0 = \varrho_1\}$ and if $\nabla\varrho_1 - \nabla\varrho_0 \neq 0$ on $\{\varrho_0 = \varrho_1\}$.

Remark 2.3 (Convexity). Convexity of the domain Ω is exclusively used in Lemma 4.7, a certain comparison principle for (2.7) when A_0 is a ball. In particular, we believe that the assumption could be avoided with some more work.

Remark 2.4 (The first singular time). We also remark that the first singular time in Theorem 1 could, for instance, be due to vanishing bubbles, pinch-off, or intersection with $\partial\Omega$.

Remark 2.5 (Generalized solutions of mean curvature flow). Theorem 1 is a direct consequence of Theorem 2 further down which states *monotonicity* of the scheme (2.7) and *consistency* with smooth sub- and superflows (see Definition 2 below). In [11, Theorem 4] for the Almgren–Taylor–Wang scheme, sub- and superflows are used to define generalized flows that start from more generic initial data so long as the viscosity solution is unique (see also [4, 14, 30] for more general results of that kind). A key element for this to work is that the scheme selects a sequence of open (or closed) sets. However, since we do not have a proof for (Lipschitz) continuity of w^* in (2.7) (see Corollary 4.10 and the discussion preceding it), the iterates $\{w^* > 0\}$ of our scheme are in general neither open nor closed. Alternatively, density estimates can be used to construct open (or closed) selections as in [12], but in our case those are not available because of the non-locality of TV_ε. As a consequence, it is not clear how to use (2.7) to construct viscosity solutions of the weighted mean curvature flow.

Remark 2.6 (Boundary conditions). Herein, we do not address boundary conditions, but we note that—following the numerical experiments in [17]—incorporation of Neumann boundary conditions for the Almgren–Taylor–Wang scheme has only recently been rigorously addressed in [18]. Their techniques appear highly PDE dependent, and it is not clear a similar approach can be used in our nonlocal setting.

3 Properties of the total variation

First, we recall that the total variation admits a coarea formula with respect to the nonlocal perimeter Per_ε.

Lemma 3.1 (Coarea formula [7])**.** *For every $u \in L^1(\Omega)$ it holds that*

$$\mathrm{TV}_\varepsilon(u) = \int_{\mathbb{R}} \mathrm{Per}_\varepsilon(\{u > t\})\,\mathrm{d}t, \tag{3.1}$$

where both sides can take the value $+\infty$.

We remark that the above lemma is stated in [7, Proposition 3.13] using the sets $\{u \geq t\}$. However, as noted in [8, Section 4.1] for sufficiently regular densities (in particular, continuous densities) the statement holds for strict super-level sets, as well.

Next we study some basic properties of the subdifferential of the total variation, regarded as a convex functional on $L^2(\Omega)$ with the standard inner product. We first record the following lemma which will be familiar to readers used to working with 1-homogeneous functionals.

Lemma 3.2. *Let $\mathcal{X}$ be a Banach space with dual pairing $\langle \cdot, \cdot \rangle : \mathcal{X}^* \times \mathcal{X} \to \mathbb{R}$, and let $J : \mathcal{X} \to [0, \infty]$ be a proper functional with $J(cu) = cJ(u)$ for all $u \in \operatorname{dom} J$ and $c \geq 0$. Then the subdifferential of J at $u \in \operatorname{dom} J$, defined as*

$$\partial J(u) := \{p \in \mathcal{X}^* \,:\, J(u) + \langle p, v - u \rangle \leq J(v) \;\; \textit{for all} \;\; v \in \mathcal{X}\}\,, \tag{3.2}$$

has the characterization

$$\partial J(u) = \{p \in \mathcal{X}^* \,:\, \langle p, u \rangle = J(u),\ \langle p, v \rangle \leq J(v) \;\; \textit{for all} \;\; v \in \mathcal{X}\}\,. \tag{3.3}$$

Remark 3.3. Elements $p \in \partial J(u)$ are called subgradients of J at u.

Proof. The inclusion "$\supset$" in (3.3) is trivial. For the converse inclusion, we let $p \in \partial J(u)$ and choose $v = 2u$ in (3.2), yielding $J(u) + \langle p, u \rangle \leq J(2u) = 2J(u)$ and hence $\langle p, u \rangle \leq J(u)$. Choosing $v = 0$ and using $J(0) = 0$ yields the converse inequality $J(u) \leq \langle p, u \rangle$. Hence, it holds $\langle p, u \rangle = J(u)$ which immediately also implies $\langle p, v \rangle \leq J(v)$ for all $v \in \mathcal{X}$, using again (3.2). This concludes the proof of "$\subset$". ☐

It will be important to understand properties of the subdifferential of the total variation TV_ε, regarded as an extended-valued functional on $L^2(\Omega)$. According to (3.2) its subdifferential is given by

$$\partial\, \mathrm{TV}_\varepsilon(u) = \left\{ p \in L^2(\Omega) \,:\, \mathrm{TV}_\varepsilon(u) + \int_\Omega p(v - u)\,\mathrm{d}x \leq \mathrm{TV}_\varepsilon(v) \quad \forall v \in L^2(\Omega) \right\}. \tag{3.4}$$

Using the characterization (3.3) of $\partial\, \mathrm{TV}_\varepsilon(u)$ with $v \equiv \pm 1$, we note that for $p \in \partial\, \mathrm{TV}_\varepsilon(u)$ one has $\int_\Omega p\,\mathrm{d}x = 0$. Characterizing the subdifferential in full generality beyond (3.3) is both not necessary for our purposes and beyond the scope of this paper, for which it suffices to restrict ourselves to suitably nice functions u and a smaller class of test functions than $v \in L^2(\Omega)$. For this we start with a few informal considerations. Since $\mathrm{TV}_\varepsilon(u)$ is positively homogeneous, according to (3.3) it suffices to find p such that $\int_\Omega pv\,\mathrm{d}x \leq \mathrm{TV}_\varepsilon(v)$ for all test functions v with equality for $v = u$ to characterize the subdifferential. If we assumed that u was sufficiently nice such that $\operatorname{ess\,sup}_{B(x,\varepsilon)\cap\Omega} u$ and $\operatorname{ess\,inf}_{B(x,\varepsilon)\cap\Omega}$ were attained at unique points $\Gamma_\varepsilon(x)$ and $\gamma_\varepsilon(x)$, respectively, we could use a change of variables to obtain

$$\int_\Omega v\,\mathrm{d}(\Gamma_\varepsilon)_\sharp \varrho_0 \leq \int_\Omega \operatorname*{ess\,sup}_{B(\cdot,\varepsilon)\cap\Omega} v\,\mathrm{d}\varrho_0 \qquad \text{and} \qquad \int_\Omega v\,\mathrm{d}(\gamma_\varepsilon)_\sharp \varrho_1 \geq \int_\Omega \operatorname*{ess\,inf}_{B(\cdot,\varepsilon)\cap\Omega} v\,\mathrm{d}\varrho_1$$

with equality for $v = u$. Consequently and not being concerned about regularity, the function

$$p := \frac{(\Gamma_\varepsilon)_\sharp \varrho_0 - \varrho_0}{\varepsilon} + \frac{\varrho_1 - (\gamma_\varepsilon)_\sharp \varrho_1}{\varepsilon}$$

would be an element of $\partial\, \mathrm{TV}_\varepsilon(u)$. For this to be rigorous, we would have to make sure that the maps $\Gamma_\varepsilon(x) := \arg\max_{B(x,\varepsilon)\cap\Omega} u$ and $\gamma_\varepsilon(x) := \arg\min_{B(x,\varepsilon)\cap\Omega} u$ are well-defined and the pushforwards $(\Gamma_\varepsilon)_\sharp \varrho_0$ and $(\gamma_\varepsilon)_\sharp \varrho_1$ have densities in $L^2(\Omega)$. Towards this goal, we first of all work with sufficiently regular functions u with non-vanishing gradients, and also with a restricted class of test functions v for which we can prove the subdifferential inequality in (3.4) holds.

Proposition 3.4. *Let $u \in C^2(\overline{\Omega})$ such that $|\nabla u| \geq c$ in $\overline{\Omega}$ for a constant $c > 0$, and let $\Lambda_{\max}$ denote the largest eigenvalue of the Hessian of u over Ω. If $0 < \varepsilon < c/\Lambda_{\max}$ is small enough, then*

- *for every $x \in \Omega_\varepsilon$ the maps*

$$\Gamma_\varepsilon(x) := \operatorname*{arg\,max}_{\overline{B(x,\varepsilon)\cap\Omega}} u \qquad \text{and} \qquad \gamma_\varepsilon(x) := \operatorname*{arg\,min}_{\overline{B(x,\varepsilon)\cap\Omega}} u$$

  are singletons;

- *for every $y \in \Omega_{2\varepsilon}$ the densities of the pushforward measures $(\Gamma_\varepsilon)_\sharp \varrho_0$ and $(\gamma_\varepsilon)_\sharp \varrho_1$ with respect to the Lebesgue measure $\mathcal{L}^N$ are given by*

$$\frac{\mathrm{d}(\Gamma_\varepsilon)_\sharp \varrho_0}{\mathrm{d}\mathcal{L}^N}(y) = \frac{\mathrm{d}\varrho_0}{\mathrm{d}\mathcal{L}^N}\left(y - \varepsilon\frac{\nabla u(y)}{|\nabla u(y)|}\right)\left|\det\left(\nabla\left(y - \varepsilon\frac{\nabla u(y)}{|\nabla u(y)|}\right)\right)\right|, \tag{3.5a}$$

$$\frac{\mathrm{d}(\gamma_\varepsilon)_\sharp \varrho_1(y)}{\mathrm{d}\mathcal{L}^N}(y) = \frac{\mathrm{d}\varrho_1}{\mathrm{d}\mathcal{L}^N}\left(y + \varepsilon\frac{\nabla u(y)}{|\nabla u(y)|}\right)\left|\det\left(\nabla\left(y + \varepsilon\frac{\nabla u(y)}{|\nabla u(y)|}\right)\right)\right|, \tag{3.5b}$$

  where by Assumption 1 it holds $\frac{\mathrm{d}\varrho_i}{\mathrm{d}\mathcal{L}^N} = \varrho_i$ for $i \in \{0,1\}$.

- *the function $p \in L^2(\Omega_{2\varepsilon})$, defined by*

$$p := \frac{\mathrm{d}}{\mathrm{d}\mathcal{L}^N}\left[\frac{(\Gamma_\varepsilon)_\sharp \varrho_0 - \varrho_0}{\varepsilon} + \frac{\varrho_1 - (\gamma_\varepsilon)_\sharp \varrho_1}{\varepsilon}\right], \tag{3.6}$$

  satisfies the inequality

$$\mathrm{TV}_\varepsilon(u) + \int_\Omega p\varphi \,\mathrm{d}x \leq \mathrm{TV}_\varepsilon(u+\varphi) \tag{3.7}$$

  for all $\varphi \in L^2(\Omega)$ with $\varphi = 0$ almost everywhere in $\Omega \setminus \Omega_{2\varepsilon}$.

Proof. We will derive the first two statements only for Γ_ε; the ones for γ_ε follow from replacing u by $-u$.

Step 1 (Optimality condition). First, we note that for any $x \in \Omega_\varepsilon$ there exists a point $y^* \in \operatorname{arg\,max}_{\overline{B(x,\varepsilon)\cap\Omega}} u$ since u is continuous. Second, by the Karush–Kuhn–Tucker optimality conditions (or direct verification) we get that y^* satisfies

$$\nabla u(y^*) - \lambda^*(y^* - x) = 0, \qquad |y^* - x| \leq \varepsilon, \tag{3.8}$$

for a Lagrange multiplier $\lambda^* \geq 0$ which is such that $\lambda^*(|y^* - x|^2 - \varepsilon^2) = 0$. Since by assumption $|\nabla u| \geq c > 0$ on Ω, the maximum has to be taken on the boundary of $B(x,\varepsilon)$, i.e., $|y^* - x| = \varepsilon$. Therefore, we obtain from (3.8) that the Lagrange multiplier is given by $\lambda^* = \frac{|\nabla u(y^*)|}{|y^* - x|} = \frac{|\nabla u(y^*)|}{\varepsilon}$.

Step 2 (Unique maximum). Next we prove that the maximizer y^* is uniquely determined. For this, we define the Lagrangian

$$L(y,\lambda) := -u(y) + \frac{\lambda}{2}\left(|y - x|^2 - \varepsilon^2\right) \qquad \text{for } \lambda \in [0,\infty)$$

and observe that it holds

$$\nabla_y^2 L(y^*,\lambda^*) = -\nabla^2 u(y^*) + \lambda^* \mathbb{1} = -\nabla^2 u(y^*) + \frac{|\nabla u(y^*)|}{\varepsilon}\mathbb{1} \succeq \left(-\Lambda_{\max} + \frac{c}{\varepsilon}\right)\mathbb{1} \succ 0$$

by our assumption on ε. We let $M_\varepsilon := \frac{c}{\varepsilon} - \Lambda_{\max}$ and, supposing that $\tilde{y} \in \partial B(x,\varepsilon)$ is another maximizer, we get using Taylor expansions and applying (3.8) that

$$\begin{aligned}
-u(\tilde{y}) &= L(\tilde{y}, \lambda^*) \\
&= L(y^*, \lambda^*) + \nabla_y L(y^*, \lambda^*)(\tilde{y} - y^*) + \frac{1}{2}(\tilde{y} - y^*)^T \nabla_y^2 L(y^*, \lambda^*)(\tilde{y} - y^*) + o\left(|\tilde{y} - y^*|^2\right) \\
&\geq -u(y^*) + \frac{M_\varepsilon}{2} |\tilde{y} - y^*|^2 - \omega(|\tilde{y} - y^*|) |\tilde{y} - y^*|^2
\end{aligned}$$

where ω is the modulus of continuity of $\nabla_y^2 L(y, \lambda^*)$ in y, which is the same as the modulus of continuity of $\nabla^2 u$ (and thereby independent of ε). Using that $u(\tilde{y}) = u(y^*)$, we find

$$\frac{M_\varepsilon}{2} |\tilde{y} - y^*|^2 \leq \omega(|\tilde{y} - y^*|) |\tilde{y} - y^*|^2 . \tag{3.9}$$

We note that M_ε is positive for ε small enough and even $\lim_{\varepsilon \to 0} M_\varepsilon = \infty$. Therefore, for $\varepsilon > 0$ small enough, (3.9) becomes a contradiction unless $\tilde{y} = y^*$. Hence, we have shown that $\Gamma_\varepsilon(x) = \{y^*\}$ is a singleton.

Step 3 (Computation of the pushforward). Vice versa, solving(3.8) for x, we see that Γ_ε is one-to-one with inverse

$$\Gamma_\varepsilon^{-1}(y^*) = y^* - \varepsilon \frac{\nabla u(y^*)}{|\nabla u(y^*)|}, \tag{3.10}$$

and in particular, this is a well-defined injective C^1 function on $\Gamma(\Omega_\varepsilon)$. Therefore we can use the definition of the pushforward and the area formula to show for any continuous function $\varphi \in C(\overline{\Omega})$ that

$$\begin{aligned}
\int_{\Gamma_\varepsilon(\Omega_\varepsilon)} \varphi \, \mathrm{d}(\Gamma_\varepsilon)_\sharp \varrho_0 &= \int_{\Omega_\varepsilon} \varphi \circ \Gamma_\varepsilon \, \mathrm{d}\varrho_0 \\
&= \int_{\Gamma_\varepsilon(\Omega_\varepsilon)} \varphi(y) \left|\det(\nabla \Gamma_\varepsilon^{-1}(y))\right| \varrho_0(\Gamma_\varepsilon^{-1}(y)) \, \mathrm{d}y \\
&= \int_{\Gamma_\varepsilon(\Omega_\varepsilon)} \varphi(y) \left|\det\left(\nabla\left(y - \varepsilon \frac{\nabla u(y)}{|\nabla u(y)|}\right)\right)\right| \varrho_0\left(y - \varepsilon \frac{\nabla u(y)}{|\nabla u(y)|}\right) \mathrm{d}y.
\end{aligned}$$

Restricting ourselves to arbitrary continuous functions φ with $\operatorname{supp}\varphi \subset \Omega_{2\varepsilon} \subset \Gamma_\varepsilon(\Omega_\varepsilon)$, we obtain the claimed identity (3.6) for the pushforwards. Note that $\Omega_{2\varepsilon} \subset \Gamma_\varepsilon(\Omega_\varepsilon)$ follows from (3.10).

Step 4 (Local subgradient). To obtain the last claim (3.7), let $\mathcal{N}$ be the set of points, which are not Lebesgue points for $u + \varphi$ and hence $\mathcal{L}^N(\mathcal{N}) = 0$. In particular, also the sets $\mathcal{N}_1 := \Gamma_\varepsilon(\Omega_\varepsilon) \cap \mathcal{N}$ and $\mathcal{N}_2 := \gamma_\varepsilon(\Omega_\varepsilon) \cap \mathcal{N}$ have zero Lebesgue measure. Since Γ_ε^{-1} and γ_ε^{-1} are diffeomorphisms from $\Gamma_\varepsilon(\Omega_\varepsilon)$ and $\gamma_\varepsilon(\Omega_\varepsilon)$, respectively, to Ω_ε, we have

$$\begin{aligned}
\mathcal{L}^N(\{x \in \Omega_\varepsilon \,:\, \Gamma_\varepsilon(x) \in \mathcal{N} \text{ or } \gamma_\varepsilon(x) \in \mathcal{N}\}) &= \mathcal{L}^N(\{x \in \Omega_\varepsilon \,:\, \Gamma_\varepsilon(x) \in \mathcal{N}_1 \text{ or } \gamma_\varepsilon(x) \in \mathcal{N}_2\}) \\
&\subset \mathcal{L}^N(\Gamma_\varepsilon^{-1}(\mathcal{N}_1) \cup \gamma_\varepsilon^{-1}(\mathcal{N}_2)) = 0,
\end{aligned}$$

hence, for almost every $x \in \Omega_\varepsilon$ the points $\Gamma_\varepsilon(x)$ and $\gamma_\varepsilon(x)$ are Lebesgue points of $u + \varphi$. Necessarily, it follows that for such $x \in \Omega_\varepsilon$ it holds

$$(u + \varphi) \circ \Gamma_\varepsilon(x) \leq \operatorname*{ess\,sup}_{B(x,\varepsilon)} (u + \varphi) \quad \text{and} \quad \operatorname*{ess\,inf}_{B(x,\varepsilon)} (u + \varphi) \leq (u + \varphi) \circ \gamma_\varepsilon(x).$$

Hence, we obtain for any $\varphi \in L^2(\Omega)$ with $\varphi = 0$ almost everywhere on $\Omega \setminus \Omega_{2\varepsilon}$ that

$$\begin{aligned}
\varepsilon \int_\Omega p\varphi \,\mathrm{d}x &= \int_{\Gamma_\varepsilon(\Omega_\varepsilon)} \varphi \,\mathrm{d}(\Gamma_\varepsilon)_\sharp \varrho_0 - \int_\Omega \varphi \,\mathrm{d}\varrho_0 + \int_\Omega \varphi \,\mathrm{d}\varrho_1 - \int_{\gamma_\varepsilon(\Omega_\varepsilon)} \varphi \,\mathrm{d}(\gamma_\varepsilon)_\sharp \varrho_1 \\
&= \int_{\Omega_\varepsilon} \varphi \circ \Gamma_\varepsilon \,\mathrm{d}\varrho_0 - \int_\Omega \varphi \,\mathrm{d}\varrho_0 + \int_\Omega \varphi \,\mathrm{d}\varrho_1 - \int_{\Omega_\varepsilon} \varphi \circ \gamma_\varepsilon \,\mathrm{d}\varrho_1 \\
&= \int_{\Omega_\varepsilon} \Big((u+\varphi) \circ \Gamma_\varepsilon - u \circ \Gamma_\varepsilon \Big) \,\mathrm{d}\varrho_0 + \int_{\Omega_\varepsilon} \Big(u \circ \gamma_\varepsilon - (u+\varphi) \circ \gamma_\varepsilon \Big) \,\mathrm{d}\varrho_1 \\
&\quad - \int_\Omega \varphi \,\mathrm{d}\varrho_0 + \int_\Omega \varphi \,\mathrm{d}\varrho_1 \\
&\leq \int_\Omega \left(\operatorname*{ess\,sup}_{B(\cdot,\varepsilon)\cap\Omega} (u+\varphi) - \operatorname*{ess\,sup}_{B(\cdot,\varepsilon)\cap\Omega} u \right) \,\mathrm{d}\varrho_0 - \int_\Omega \left(\operatorname*{ess\,inf}_{B(\cdot,\varepsilon)\cap\Omega} u - \operatorname*{ess\,inf}_{B(\cdot,\varepsilon)\cap\Omega} (u+\varphi) \right) \,\mathrm{d}\varrho_1 \\
&\quad - \int_\Omega \varphi \,\mathrm{d}\varrho_0 + \int_\Omega \varphi \,\mathrm{d}\varrho_1 \\
&\quad - \int_{\Omega\setminus\Omega_\varepsilon} \left(\operatorname*{ess\,sup}_{B(\cdot,\varepsilon)\cap\Omega} (u+\varphi) - \operatorname*{ess\,sup}_{B(\cdot,\varepsilon)\cap\Omega} u \right) \,\mathrm{d}\varrho_0 + \int_{\Omega\setminus\Omega_\varepsilon} \left(\operatorname*{ess\,inf}_{B(\cdot,\varepsilon)\cap\Omega} u - \operatorname*{ess\,inf}_{B(\cdot,\varepsilon)\cap\Omega} (u+\varphi) \right) \,\mathrm{d}\varrho_1.
\end{aligned}$$

The last two integrals vanish because $\varphi = 0$ in $\Omega \setminus \Omega_{2\varepsilon}$. Hence

$$\begin{aligned}
\varepsilon \int_\Omega p\varphi \,\mathrm{d}x &= \int_\Omega \left(\operatorname*{ess\,sup}_{B(\cdot,\varepsilon)\cap\Omega} (u+\varphi) - (u+\varphi) \right) \,\mathrm{d}\varrho_0 + \int_\Omega \left(u+\varphi - \operatorname*{ess\,inf}_{B(\cdot,\varepsilon)\cap\Omega} (u+\varphi) \right) \,\mathrm{d}\varrho_1 \\
&\quad - \int_\Omega \left(\operatorname*{ess\,sup}_{B(\cdot,\varepsilon)\cap\Omega} u - u \right) \,\mathrm{d}\varrho_0 - \int_\Omega \left(u - \operatorname*{ess\,inf}_{B(\cdot,\varepsilon)\cap\Omega} u \right) \,\mathrm{d}\varrho_1 \\
&= \varepsilon \,\mathrm{TV}_\varepsilon(u+\varphi) - \varepsilon \,\mathrm{TV}_\varepsilon(u).
\end{aligned}$$

This shows (3.7) and concludes the proof. □

Next we prove that the subgradient identified in the previous lemma is consistent with a weighted 1-Laplacian operator which, neglecting boundary conditions, is the subgradient of a local weighted total variation.

Proposition 3.5. *Under the conditions of Proposition 3.4 it holds that*

$$\frac{\mathrm{d}}{\mathrm{d}\mathcal{L}^N} \left[\frac{(\Gamma_\varepsilon)_\sharp \varrho_0 - \varrho_0}{\varepsilon} + \frac{\varrho_1 - (\gamma_\varepsilon)_\sharp \varrho_1}{\varepsilon} \right] = -\operatorname{div}\left(\varrho \frac{\nabla u}{|\nabla u|} \right) + o_{\varepsilon\to 0}(1) \quad \textit{uniformly in } \Omega_{2\varepsilon}. \tag{3.11}$$

Proof. We start by investigating the density of $(\Gamma_\varepsilon)_\sharp \varrho_0$ given in Proposition 3.4. Let us fix $y \in \Omega_{2\varepsilon}$. Using a Taylor expansion of $\varrho_0 \in C^1(\overline{\Omega})$ and utilizing $|\nabla u(y)| \geq c$, we have uniformly in $\Omega_{2\varepsilon}$ that

$$\varrho_0 \left(y - \varepsilon \frac{\nabla u(y)}{|\nabla u(y)|} \right) = \varrho_0(y) - \varepsilon \nabla \varrho_0(y) \cdot \frac{\nabla u(y)}{|\nabla u(y)|} + o(\varepsilon).$$

Furthermore, using a Taylor expansion of the determinant and utilizing $u \in C^2(\overline{\Omega})$ with $|\nabla u| \geq c$, we get

$$\det\left(\nabla \left(y - \varepsilon \frac{\nabla u(y)}{|\nabla u(y)|} \right) \right) = 1 - \varepsilon \operatorname{div}\left(\frac{\nabla u(y)}{|\nabla u(y)|} \right) + O(\varepsilon^2).$$

In particular, we see that the determinant is non-negative if $\varepsilon > 0$ is sufficiently small. Multiplying the two expressions and using the product rule yields

$$\begin{aligned}
&\varrho_0\left(y - \varepsilon\frac{\nabla u(y)}{|\nabla u(y)|}\right)\left|\det\left(\nabla\left(y - \varepsilon\frac{\nabla u(y)}{|\nabla u(y)|}\right)\right)\right| \\
&= \left(\varrho_0(y) - \varepsilon\nabla\varrho_0(y)\cdot\frac{\nabla u(y)}{|\nabla u(y)|} + o(\varepsilon)\right)\left(1 - \varepsilon\operatorname{div}\left(\frac{\nabla u(y)}{|\nabla u(y)|}\right) + O(\varepsilon^2)\right) \\
&= \varrho_0(y) - \varepsilon\nabla\varrho_0(y)\cdot\frac{\nabla u(y)}{|\nabla u(y)|} - \varepsilon\varrho_0(y)\operatorname{div}\left(\frac{\nabla u(y)}{|\nabla u(y)|}\right) + o(\varepsilon) \\
&= \varrho_0(y) - \varepsilon\operatorname{div}\left(\varrho_0(y)\frac{\nabla u(y)}{|\nabla u(y)|}\right) + o(\varepsilon).
\end{aligned}$$

Using Proposition 3.4 we therefore obtain

$$\frac{\mathrm{d}}{\mathrm{d}\mathcal{L}^N}\left[\frac{(\Gamma_\varepsilon)_\sharp\varrho_0 - \varrho_0}{\varepsilon}\right] = -\operatorname{div}\left(\varrho_0(y)\frac{\nabla u(y)}{|\nabla u(y)|}\right) + o_{\varepsilon\to 0}(1).$$

Repeating the same arguments for $(\gamma_\varepsilon)_\sharp\varrho_1$ and using $\varrho = \varrho_0 + \varrho_1$ leads to the final conclusion. □

Next, we prove that the total variation is a lower semicontinuous functional with respect to the weak L^2 topology. In [7], it was already proved that it is weak-$*$ lower semicontinuous in L^∞, but this is insufficient for our purposes, because we will use additivity of the subdifferential (see the proof of Proposition 4.2).

Lemma 3.6. *Let $(u_k)_{k\in\mathbb{N}} \subset L^2(\Omega)$ be a sequence which converges weakly to $u \in L^2(\Omega)$. Then it holds that*

$$\mathrm{TV}_\varepsilon(u) \le \liminf_{k\to\infty} \mathrm{TV}_\varepsilon(u_k).$$

Proof. It suffices to show that

$$\mathrm{TV}^0_\varepsilon(u) := \frac{1}{\varepsilon}\int_\Omega \left(\operatorname*{ess\,sup}_{B(x,\varepsilon)\cap\Omega} u - u(x)\right)\mathrm{d}\varrho_0(x)$$

is weakly lower semi-continuous. As $\mathrm{TV}^0_\varepsilon$ is convex, it suffices to prove lower semi-continuity with respect to strongly converging sequences. In particular, we assume $u_k \to u$ in $L^2(\Omega)$, and may further suppose that $u_k(x) \to u(x)$ for almost every $x \in \Omega$.

Let $\delta > 0$. The strong convergence of u_k to u in particular implies for all $y \in \Omega$ and $0 < r < \delta$:

$$\frac{1}{\mathcal{L}^N(B(y,r)\cap\Omega)}\int_{B(y,r)\cap\Omega} u\,\mathrm{d}z = \lim_{k\to\infty}\frac{1}{\mathcal{L}^N(B(y,r)\cap\Omega)}\int_{B(y,r)\cap\Omega} u_k\,\mathrm{d}z.$$

Taking Lebesgue points of u, for almost every $y \in \Omega$ we get

$$\begin{aligned}
u(y) &= \lim_{r\to 0}\frac{1}{\mathcal{L}^N(B(y,r)\cap\Omega)}\int_{B(y,r)\cap\Omega} u\,\mathrm{d}z = \lim_{r\to 0}\lim_{k\to\infty}\frac{1}{\mathcal{L}^N(B(y,r)\cap\Omega)}\int_{B(y,r)\cap\Omega} u_k\,\mathrm{d}z \\
&\le \liminf_{k\to\infty}\operatorname*{ess\,sup}_{B(y,\delta)\cap\Omega} u_k.
\end{aligned}$$

As $\delta > 0$ is arbitrary, we conclude that for almost every $x \in \Omega$

$$\operatorname*{ess\,sup}_{B(x,\varepsilon)\cap\Omega} u \leq \liminf_{k\to\infty} \operatorname*{ess\,sup}_{B(x,\varepsilon)\cap\Omega} u_k.$$

As we have the pointwise convergence of u_k, we improve this to

$$\operatorname*{ess\,sup}_{B(x,\varepsilon)\cap\Omega} u - u(x) \leq \liminf_{k\to\infty} \left(\operatorname*{ess\,sup}_{B(x,\varepsilon)\cap\Omega} u_k - u_k(x) \right),$$

from which Fatou's lemma concludes the claimed lower semi-continuity. □

Next we establish an important submodularity property of the total variation TV_ε. In fact it follows from submodularity of the perimeter Per_ε (proved in [7, Proposition 3.3]) and the coarea formula (see [13, Proposition 3.2]) but for self-containedness we elaborate on the proof.

Lemma 3.7. *It holds for all $u, v \in L^2(\Omega)$ that*

$$\mathrm{TV}_\varepsilon(u \vee v) + \mathrm{TV}_\varepsilon(u \wedge v) \leq \mathrm{TV}_\varepsilon(u) + \mathrm{TV}_\varepsilon(v).$$

Proof. The statement follows from the directly obtained inequalities

$$\operatorname*{ess\,sup}_{B(\cdot,\varepsilon)\cap\Omega} (u \vee v) \leq \Big(\operatorname*{ess\,sup}_{B(\cdot,\varepsilon)\cap\Omega} u \Big) \vee \Big(\operatorname*{ess\,sup}_{B(\cdot,\varepsilon)\cap\Omega} v \Big),$$
$$\operatorname*{ess\,sup}_{B(\cdot,\varepsilon)\cap\Omega} (u \wedge v) \leq \Big(\operatorname*{ess\,sup}_{B(\cdot,\varepsilon)\cap\Omega} u \Big) \wedge \Big(\operatorname*{ess\,sup}_{B(\cdot,\varepsilon)\cap\Omega} v \Big),$$

the reverse analogues for the ess inf, and the elementary identity $a \vee b + a \wedge b = a + b$ for numbers $a, b \in \mathbb{R}$. □

4 Convergence of the adversarial training scheme

For a set $A \in \mathfrak{B}(\Omega)$, let us define the *one-step operator* $S_\varepsilon(A)$ of the adversarial training scheme (2.7) via

$$S_\varepsilon(A) := \{w_\varepsilon^* > 0\} \quad \text{where} \quad w_\varepsilon^* := \operatorname*{arg\,min}_{u\in L^2(\Omega)} \frac{1}{2\varepsilon} \int_\Omega |u - \mathrm{sdist}(\cdot, A^c)|^2 \, \mathrm{d}\varrho + \mathrm{TV}_\varepsilon(u). \tag{4.1}$$

For convenience, we assume that w_ε^* is a Lebesgue representative with $\{w_\varepsilon^* > 0\} = (\{w_\varepsilon^* > 0\})^1$, where we recall the notation introduced in (2.5). In particular, with this representative convention in place, one can verify by hand that

$$\mathrm{sdist}(\cdot, \{w_\varepsilon^* > 0\}^c) = \mathrm{dist}(\cdot, \{w_\varepsilon^* > 0\}^c) - \mathrm{dist}(\cdot, \{w_\varepsilon^* > 0\}),$$

circumventing the need for a well-chosen representative in sdist (see (2.5)) for another application of S_ε.

In Proposition 4.2 below, we will first prove that the operator S_ε does in fact select a solution of (2.4); in other words, the convex minimization problem arising in (2.7) is consistent with the scheme (2.4). Second, to prove Theorem 1, we will connect the operator (4.1) to the limiting equation by showing that the operator is monotone and consistent with respect to weighted mean curvature flow in the following sense.

Definition 1 (Monotonicity)**.** The operator S_ε defined in (4.1) is *monotone* if $A' \subset A \subset \Omega$ implies $S_\varepsilon(A') \subset S_\varepsilon(A)$.

While monotonicity is a property of the operator by itself, consistency, on the other hand, directly connects the scheme to mean curvature flow.

For consistency, we rely on the notions of sub- and superflows typically used to construct barrier solutions for mean curvature flow, see for example [3, Chapter 9]. If $[t_0, t_1] \ni t \mapsto A(t) \subset\subset \Omega$ is a smooth curve of smooth sets which evolve with normal speed $V(t) = -\frac{1}{\varrho} \operatorname{div}\left(\varrho \nu_{A(t)}\right)$, where $\nu_{A(t)}$ is the inner normal vector to the boundary of $A(t)$, i.e., as in (2.8), then the signed distance function $d(x,t) := \operatorname{sdist}(x, A^c(t))$ satisfies

$$\partial_t d(x,t) = \frac{1}{\varrho(x)} \operatorname{div}\left(\varrho(x) \nabla d(x,t)\right) \tag{4.2}$$

for any (x,t) with $d(x,t) = 0$. This is because $\nu_{A(t)} = \nabla d(x,t)$. The PDE (4.2) forms the basis of our sub- and superflow definitions for weighted mean curvature flow, adapted from [14, Definition 2.1]. Informally, a superflow is a smooth evolution of sets that moves strictly faster than mean curvature flow, while a subflow moves slower. We emphasize that our meaning is the same as in other works, but the inequalities are reversed as the gradient of the signed distance function $\operatorname{sdist}(x, A^c)$ points into the evolving set (more consistent with BV-solution concepts).

Definition 2 (Sub- and superflows)**.** Let $A(t) \subset\subset \Omega$, $t \in [t_0, t_1]$. We say that $A(t)$ is a *subflow* of (4.2) if

- there exists a relatively open set $B \subset \Omega \times [t_0, t_1]$ with $\bigcup_{t_0 \le t \le t_1} \partial A(t) \times \{t\} \subset B$;
- the function $d(x,t) := \operatorname{sdist}(x, A^c(t))$ is continuously differentiable in time and twice continuously differentiable in space in B, which we abbreviate as $d \in C^{2,1}_{x,t}(B)$;
- there exists $\delta > 0$ such that

$$\partial_t d(x,t) \ge \frac{1}{\varrho(x)} \operatorname{div}\left(\varrho(x) \nabla d(x,t)\right) + \delta \tag{4.3}$$

  for any $(x,t) \in B$.

We say that $A(t)$ is a *superflow* whenever $\delta < 0$ and the reverse inequality holds in (4.3).

Definition 3 (Consistency)**.** The operator S_ε defined in (4.1) is *consistent* if

- for every subflow $[t_0, t_1] \ni t \mapsto A(t)$ in the sense of Definition 2 there exists $\varepsilon_0 > 0$ such that $S_\varepsilon(A(t)) \subset A(t + \varepsilon)$ for all $0 < \varepsilon < \varepsilon_0$ and all $t \in [t_0, t_1 - \varepsilon]$;
- for every superflow the same holds with the converse inclusion.

The interpretation of this definition is that for $\varepsilon > 0$ sufficiently small the scheme $S_\varepsilon(A)$ defined in (4.1) moves faster than a subflow and slower than a superflow starting at A. With these definitions in hand, we may state the principle result of this paper.

Theorem 2 (Monotonicity and consistency)**.** *If $\Omega \subset \mathbb{R}^N$ is a bounded and convex domain the operator S_ε is monotone and consistent with the weighted mean curvature flow* (2.8) *in the sense of Definitions 1 and 3, respectively.*

Theorem 2 will follow directly follow from Propositions 4.4 and 4.11 below. We briefly defer the proofs of the aforementioned selection principle and Theorem 2, as we can now directly conclude Theorem 1.

Proof of Theorem 1. Let A_0 and $t \mapsto A(t)$ be as in the hypothesis of the theorem and recall $t \mapsto A_\varepsilon(t)$ defined in (2.9). Then let $t \mapsto A_{\mathrm{sub}}(t)$ be any subflow with initial condition $A_0 \subset A_{\mathrm{sub}}(0)$ and parameter $\delta > 0$. First, we will show that for any time t before the singular time of A_{sub}, we have that

$$A_\varepsilon(t) \subset A_{\mathrm{sub}}(\varepsilon \lfloor t/\varepsilon \rfloor). \tag{4.4}$$

for all ε sufficiently small. Similarly, the converse set containment holds for a superflow A_{sup} leading to

$$A_{\mathrm{sup}}(\varepsilon \lfloor t/\varepsilon \rfloor) \subset A_\varepsilon(t) \subset A_{\mathrm{sub}}(\varepsilon \lfloor t/\varepsilon \rfloor) \quad \text{for all } \varepsilon > 0 \text{ sufficiently small.} \tag{4.5}$$

It turns out (4.4) is a simple consequence of Theorem 2. Let $(A_k)_{k \in \mathbb{N}_0}$ be the sets in the definition of A_ε in (2.9) coming from iteratively applying the scheme. By monotonicity and consistency we have

$$A_1 = S_\varepsilon(A_0) \subset S_\varepsilon(A_{\mathrm{sub}}(0)) \subset A_{\mathrm{sub}}(\varepsilon).$$

Applying S_ε to both sides once again, using monotonicity and consistency, we have

$$A_2 = S_\varepsilon(A_1) \subset S_\varepsilon(A_{\mathrm{sub}}(\varepsilon)) \subset A_{\mathrm{sub}}(2\varepsilon).$$

Iterating and recalling the definition of A_ε in (2.9), we conclude (4.4) and hence also (4.5).

Consequently, using (4.5) and letting T be the earliest singular time of A_{sub} and A_{sup}, we estimate for any $s < T$ that

$$\limsup_{\varepsilon \to 0} \sup_{t \in [0,s]} \mathcal{L}^N(A_\varepsilon(t) \triangle A(t)) \leq \sup_{t \in [0,s]} \left(\mathcal{L}^N(A_{\mathrm{sub}}(t) \triangle A(t)) + \mathcal{L}^N(A_{\mathrm{sup}}(t) \triangle A(t)) \right), \tag{4.6}$$

where we have used continuity of the sub- and superflows to replace $\varepsilon \lfloor t/\varepsilon \rfloor$ with t. Similarly, one can get an estimate for the Hausdorff distance $d_H(A, B) := \sup_{x \in A} \operatorname{dist}(x, B) \vee \sup_{x \in B} \operatorname{dist}(x, A)$. Using (4.5) again we have

$$\begin{aligned} \limsup_{\varepsilon \to 0} \sup_{t \in [0,s]} d_H(A_\varepsilon(t), A(t)) &\leq \sup_{t \in [0,s]} \sup_{y \in A_{\mathrm{sub}}(t)} d(x, A(t)) \vee \sup_{x \in A(t)} \operatorname{dist}(x, A_{\mathrm{sup}}(t)) \\ &\leq \sup_{t \in [0,s]} d_H(A_{\mathrm{sub}}(t), A(t)) + d_H(A_{\mathrm{sup}}(t), A(t)). \end{aligned} \tag{4.7}$$

It remains to argue that the right hand side of (4.6) and (4.7) can be made arbitrarily small by approximating $t \mapsto A(t)$ by sub- and superflows. Briefly, let $d(x, t) = \operatorname{sdist}(x, A^c_{\mathrm{sub}}(t))$. Note that if $\partial_t d = \frac{1}{\varrho} \operatorname{div}(\varrho \nabla d) + \delta$ on $\partial A_{\mathrm{sub}}(t)$, then as a curvature flow this may be written as

$$V_{\mathrm{sub}}(t) = -\frac{1}{\varrho} \operatorname{div}\left(\varrho \nu_{A_{\mathrm{sub}}(t)}\right) - \delta = H_{A_{\mathrm{sub}}(t)} - \nabla \log \varrho \cdot \nu_{A_{\mathrm{sub}}(t)} - \delta \quad \text{on } \partial A_{\mathrm{sub}}(t), \tag{4.8}$$

following the convention of (2.8). As (4.8) is a perturbation of (2.8), one can show that if (2.8) has a strong solutions up to time T_*, then for any $T < T^*$, there is $\delta_T > 0$ sufficiently small such that the flow (4.8) has a strong solution up to time T for all $|\delta| < \delta_T$. With this in mind, we see that the right-hand side of (4.6) and (4.7) can be made arbitrarily small by choosing sub- and superflows satisfying the inequalities of Definition 2 with equality on the interface and then sending $\delta \to 0$ (in the definition of sub-/superflow one must replace δ by $\delta/2$ to get the neighborhood B); at the same time this will allow one to take $T \to T_*$, concluding the theorem. ∎

4.1 Well-posedness

We first show that the optimization problem in (4.1) has a unique solution, up to equality on Lebesgue null-sets. This follows from the following more general statement.

Proposition 4.1. *For any $f \in L^2(\Omega)$ there exists a unique element $w \in L^2(\Omega)$ such that*

$$\frac{1}{2\varepsilon}\int_\Omega |w-f|^2 \,\mathrm{d}\varrho + \mathrm{TV}_\varepsilon(w) = \inf_{u\in L^2(\Omega)} \frac{1}{2\varepsilon}\int_\Omega |u-f|^2 \,\mathrm{d}\varrho + \mathrm{TV}_\varepsilon(u).$$

Furthermore, if $f \in L^\infty(\Omega)$ then $w \in L^\infty(\Omega)$ with $\|w\|_{L^\infty(\Omega)} \le \|f\|_{L^\infty(\Omega)}$.

Proof. We define the functional $E : L^2(\Omega) \to [0,\infty]$ via

$$E(u) := \frac{1}{2\varepsilon}\int_\Omega |u-f|^2 \,\mathrm{d}\varrho + \mathrm{TV}_\varepsilon(u).$$

Thanks to Lemma 3.2 the functional $u \mapsto \mathrm{TV}_\varepsilon(u)$ is convex and hence E is strictly convex. This implies uniqueness. Existence of the minimizer w is a consequence of lower semi-continuity of the functional (Lemma 3.6) and the direct method. For the claimed L^∞-bound we note that we can replace w by the truncation $\hat{w} := (-C) \vee w \wedge C$ with $C := \|f\|_{L^\infty(\Omega)}$ which satisfies $E(\hat{w}) \le E(w)$ (as may be directly checked using Lemma 3.7) and therefore by uniqueness it holds that $\hat{w} = w$, and the bounds for w follow. □

4.2 Selection property

Next we show that S_ε selects a solution of (2.4). Our proof (lightly) deviates from that of [11, Proposition 2.2] for the standard TV functional due to the asymmetry of TV_ε with respect to super- and sublevel sets. In particular, we have $\mathrm{Per}_\varepsilon(A) \neq \mathrm{Per}_\varepsilon(A^c)$.

Proposition 4.2 (Selection principle)**.** *Letting S_ε be defined as in* (4.1)*, it holds that*

$$S_\varepsilon(A) \in \operatorname*{arg\,min}_{E\in\mathfrak{B}(\Omega)} \int_\Omega |\mathbf{1}_E - \mathbf{1}_A| \,\frac{\mathrm{dist}(\cdot,\partial A)}{\varepsilon}\,\mathrm{d}\varrho + \mathrm{Per}_\varepsilon(A).$$

Proof. We may of course assume $A \neq \emptyset$ and $A \neq \Omega$. We use the abbreviation $d(x) := \mathrm{sdist}(x, A^c)$ and recall that $|d(x)| = \mathrm{dist}(x,\partial A)$ following the notation introduced after (2.4). As the L^2-fidelity term and the TV_ε functional in the minimization problem (4.1) are both convex and lower semi-continuous, we have that

$$0 \in \frac{w_\varepsilon^* - d}{\varepsilon}\varrho + \partial\,\mathrm{TV}_\varepsilon(w_\varepsilon^*).$$

We define $p := -\frac{w_\varepsilon^*-d}{\varepsilon}\varrho \in \partial\,\mathrm{TV}_\varepsilon(w_\varepsilon^*)$ and $E_s := \{w_\varepsilon^* > s\}$.

Step 1 (Subdifferential of the super-level sets). We **claim** that for almost every $|s| \le \mathrm{diam}(\Omega) =: C$,

$$p \in \partial\,\mathrm{TV}_\varepsilon(\mathbf{1}_{E_s}) =: \partial\,\mathrm{Per}_\varepsilon(E_s).$$

We first note that by the ε-coarea formula in Lemma 3.1 and that $\|w_\varepsilon^*\|_{L^\infty} \le C$ by Proposition 4.1, we have

$$\mathrm{TV}_\varepsilon(w_\varepsilon^*) = \int_{-C}^{C} \mathrm{Per}_\varepsilon(E_s)\,\mathrm{d}s.$$

Similarly, by the layer cake formula and Fubini's theorem

$$\int_\Omega p w_\varepsilon^* \,\mathrm{d}x = \int_\Omega p(x) \left(\int_{-C}^{C} \mathbf{1}_{E_s}(x) \,\mathrm{d}s - C \right) \mathrm{d}x = \int_{-C}^{C} \int_\Omega p(x) \mathbf{1}_{E_s}(x) \,\mathrm{d}x \,\mathrm{d}s,$$

where in the last equality we recall that $\int_\Omega p \,\mathrm{d}x = 0$ as noted immediately after (3.4). By Lemma 3.2, we have $\int_\Omega p w_\varepsilon^* \,\mathrm{d}x = \mathrm{TV}_\varepsilon(w_\varepsilon^*)$, so that we may combine the above displays to find that

$$\int_{-C}^{C} \mathrm{Per}_\varepsilon(E_s) \,\mathrm{d}s = \int_{-C}^{C} \int_\Omega p(x) \mathbf{1}_{E_s}(x) \,\mathrm{d}x \,\mathrm{d}s. \tag{4.9}$$

Once again by the characterization of the subdifferential in Lemma 3.2, we have $\int_\Omega p \mathbf{1}_{E_s} \,\mathrm{d}x \leq \mathrm{Per}_\varepsilon(E_s)$, so that (4.9) implies

$$\mathrm{Per}_\varepsilon(E_s) = \int_\Omega p \mathbf{1}_{E_s} \,\mathrm{d}x \tag{4.10}$$

for almost every s with $|s| \leq C$. Applying the characterization (3.3) of the subdifferential gives the claim.

Step 2 (Subdifferential for $\{w_\varepsilon^* > 0\}$). The claim of Step 1 can be improved to every $s \in [-C, C)$: Fixing such an s and taking a sequence $s_k \downarrow s$ such that (4.10) holds for each s_k, we have that $\mathbf{1}_{E_{s_k}} \to \mathbf{1}_{E_s}$ pointwise and thereby in L^1. We pass to the limit in (4.10) as $k \to \infty$ using lower semi-continuity of Per_ε, the L^1 convergence, and that $p \in \partial \,\mathrm{TV}_\varepsilon(w_\varepsilon^*)$ (for the last inequality below) to find

$$\mathrm{Per}_\varepsilon(E_s) \leq \int_\Omega p \mathbf{1}_{E_s} \,\mathrm{d}x \leq \mathrm{Per}_\varepsilon(E_s).$$

Thus $p \in \partial \,\mathrm{Per}_\varepsilon(E_s)$ for any $s \in [-C, C)$, and in particular $p \in \partial \,\mathrm{Per}_\varepsilon(E_0) = \partial \,\mathrm{Per}_\varepsilon(\{w_\varepsilon^* > 0\})$.

Step 3 (Conclusion). We apply the definition of subdifferential at E_0 for any set E to find that

$$\mathrm{Per}_\varepsilon(E_0) + \int_\Omega p \left(\mathbf{1}_E - \mathbf{1}_{E_0} \right) \mathrm{d}x \leq \mathrm{Per}_\varepsilon(E). \tag{4.11}$$

Noting by definition of $p = \frac{d - w_\varepsilon^*}{\varepsilon} \varrho$ that

$$\int_\Omega p \left(\mathbf{1}_E - \mathbf{1}_{E_0} \right) \mathrm{d}x = \frac{1}{\varepsilon} \int_\Omega d \left(\mathbf{1}_E - \mathbf{1}_{E_0} \right) \mathrm{d}\varrho - \frac{1}{\varepsilon} \int_\Omega w_\varepsilon^* (\mathbf{1}_E - \mathbf{1}_{E_0}) \,\mathrm{d}\varrho \geq \frac{1}{\varepsilon} \int_\Omega d \left(\mathbf{1}_E - \mathbf{1}_{E_0} \right) \mathrm{d}\varrho,$$

where we used that $w_\varepsilon^* (\mathbf{1}_E - \mathbf{1}_{E_0}) \leq 0$ almost everywhere, we can rearrange (4.11) as

$$\mathrm{Per}_\varepsilon(E_0) + \frac{1}{\varepsilon} \int_\Omega (-d) \mathbf{1}_{E_0} \,\mathrm{d}\varrho \leq \mathrm{Per}_\varepsilon(E) + \frac{1}{\varepsilon} \int_\Omega (-d) \mathbf{1}_E \,\mathrm{d}\varrho.$$

However, one can verify that $\int_\Omega |\mathbf{1}_E - \mathbf{1}_A| \frac{\mathrm{dist}(\cdot, \partial A)}{\varepsilon} \,\mathrm{d}\varrho = \frac{1}{\varepsilon} \int_\Omega (-d) \mathbf{1}_E \,\mathrm{d}\varrho + \frac{1}{\varepsilon} \int_\Omega d \mathbf{1}_A \,\mathrm{d}\varrho$, so that the previous display is equivalently written

$$\mathrm{Per}_\varepsilon(E_0) + \int_\Omega |\mathbf{1}_{E_0} - \mathbf{1}_A| \frac{\mathrm{dist}(\cdot, \partial A)}{\varepsilon} \,\mathrm{d}\varrho \leq \mathrm{Per}_\varepsilon(E) + \int_\Omega |\mathbf{1}_E - \mathbf{1}_A| \frac{\mathrm{dist}(\cdot, \partial A)}{\varepsilon} \,\mathrm{d}\varrho,$$

concluding the proposition. □

4.3 Monotonicity

For proving monotonicity we start with a simple comparison principle for solutions of the optimization problem in (4.1).

Proposition 4.3 (Comparison principle I)**.** *For $d, d' \in L^\infty(\Omega)$ with $d' \leq d$ almost everywhere in Ω assume that w, w' satisfy*

$$w^{(\prime)} = \operatorname*{arg\,min}_{u \in L^2(\Omega)} \frac{1}{2\varepsilon} \int_\Omega \left| u - d^{(\prime)} \right|^2 \mathrm{d}\varrho + \mathrm{TV}_\varepsilon(u).$$

Then it holds $w' \leq w$ almost everywhere in Ω.

Proof. Using optimality of w and w' we have

$$\begin{aligned} &\frac{1}{2\varepsilon} \int_\Omega \left(|w - d|^2 + |w' - d'|^2 \right) \mathrm{d}\varrho + \mathrm{TV}_\varepsilon(w) + \mathrm{TV}_\varepsilon(w') \\ &\leq \frac{1}{2\varepsilon} \int_\Omega \left(|w \vee w' - d|^2 + |w \wedge w' - d'|^2 \right) \mathrm{d}\varrho + \mathrm{TV}_\varepsilon(w \vee w') + \mathrm{TV}_\varepsilon(w \wedge w'). \end{aligned} \tag{4.12}$$

Using Lemma 3.7 to cancel the total variations we obtain from the above that

$$\int_\Omega \left(|w - d|^2 + |w' - d'|^2 \right) \mathrm{d}\varrho \leq \int_\Omega \left(|w \vee w' - d|^2 + |w \wedge w' - d'|^2 \right) \mathrm{d}\varrho.$$

Expanding squares, canceling terms, and reordering this inequality, we reduce to

$$\begin{aligned} 0 &\leq \int_\Omega \left((w - (w \vee w')) \, d + (w' - (w \wedge w')) \, d' \right) \mathrm{d}\varrho \\ &= \int_{\Omega \cap \{w' > w\}} (w' - w) \, (d' - d) \, \mathrm{d}\varrho. \end{aligned}$$

Using that $\varrho > c_\varrho$ we infer that $\Omega \cap \{w' > w\} \cap \{d' < d\}$ has zero Lebesgue measure. As in [11, Lemma 2.1] one can argue that in fact $w' \leq w$ holds almost everywhere. □

With this comparison principle at hand, the proof of monotonicity for Theorem 2 is straightforward.

Proposition 4.4 (Monotonicity)**.** *The operator S_ε defined in* (4.1) *is monotone in the sense of Definition 1.*

Proof. Let w' and w denote the solutions of the problem in (4.1) for A' and A, respectively. Since $A' \subset A$ we have $\mathrm{sdist}(\cdot, (A')^c) \leq \mathrm{sdist}(\cdot, A^c)$, and by Proposition 4.3, it follows that $w' \leq w$ outside of a Lebesgue null-set $\mathcal{N}$. This immediately implies

$$S_\varepsilon(A') = \{w' > 0\} = (\{w' > 0\} \setminus \mathcal{N})^1 \subset (\{w > 0\} \setminus \mathcal{N})^1 = \{w > 0\} = S_\varepsilon(A),$$

where we have used that $\{w^{(\prime)} > 0\} = (\{w^{(\prime)} > 0\})^1$ by choice of representative in (4.1) and $(\{w^{(\prime)} > 0\} \setminus \mathcal{N})^1 = (\{w^{(\prime)} > 0\})^1$ holds for any Lebesgue null-set $\mathcal{N}$. □

4.4 Consistency

This subsection is devoted to the proof of consistency for Theorem 2. Consistency connects the numerical scheme directly to mean curvature flow, and as such, will require a delicate analysis. Our approach is motivated by Chambolle and Novaga's in [14] for an anisotropic but local TV functional.

We briefly summarize the strategy of the proof. To show that the evolving set of a subflow stays outside the adversarial scheme, we show that the subflow for mean curvature flow can be modified to construct a subsolution for a static TV_ε problem. To compare the modified subsolution to w_ε^* found using (4.1), we will apply the variational comparison principle proven below in Proposition 4.6 below on a tubular neighborhood of the interface. For this to work, we must know that the modified subsolution is greater than w_ε^* on the boundary of the tubular neighborhood. This information comes from Lemma 4.7 below. In fact, this lemma can be used to show that up to an error $O(\sqrt{\varepsilon})$, the minimizer w_ε^* is Lipschitz continuous (see Corollary 4.10), and related estimates will allow us to recover the boundary conditions.

We first note that global minimizers give rise to local minimizers, so long as the boundary conditions are frozen on an ε-neighborhood.

Lemma 4.5 (Restricted minimizer)**.** *Let $d \in L^\infty(\Omega)$, and let $w \in L^\infty(\Omega)$ solve*

$$w = \arg\min \left\{ \frac{1}{2\varepsilon} \int_\Omega |u-d|^2 \, \mathrm{d}\varrho + \mathrm{TV}_\varepsilon(u) \,:\, u \in L^2(\Omega) \right\}.$$

Let $\Omega' \subset \Omega$ be an open subset and recall the notation in (1.3)*. Then it also holds that*

$$w = \arg\min \left\{ \frac{1}{2\varepsilon} \int_{\Omega'} |u-d|^2 \, \mathrm{d}\varrho + \mathrm{TV}_\varepsilon(u;\Omega') \,:\, u \in L^2(\Omega'),\ u = w \text{ in } \Omega' \setminus \Omega'_\varepsilon \right\},$$

where $\mathrm{TV}_\varepsilon(u;\Omega')$ is the total variation as defined in (2.6) *with Ω replaced by Ω'.*

Proof. Let $u \in L^2(\Omega')$ be a function such that $u = w$ on $\Omega' \setminus \Omega'_\varepsilon$. We extend u to a function in $L^2(\Omega)$ by setting $u := w$ on $\Omega \setminus \Omega'$. Hence, using also the minimization property of w it holds

$$\begin{aligned}
&\frac{1}{2\varepsilon} \int_{\Omega'} |w-d|^2 \, \mathrm{d}\varrho + \mathrm{TV}_\varepsilon(w;\Omega') - \left(\frac{1}{2\varepsilon} \int_{\Omega'} |u-d|^2 \, \mathrm{d}\varrho + \mathrm{TV}_\varepsilon(u;\Omega') \right) \\
&\quad = \frac{1}{2\varepsilon} \int_{\Omega} |w-d|^2 \, \mathrm{d}\varrho + \mathrm{TV}_\varepsilon(w;\Omega') - \left(\frac{1}{2\varepsilon} \int_{\Omega} |u-d|^2 \, \mathrm{d}\varrho + \mathrm{TV}_\varepsilon(u;\Omega') \right) \\
&\quad \leq \mathrm{TV}_\varepsilon(u) - \mathrm{TV}_\varepsilon(u;\Omega') + \mathrm{TV}_\varepsilon(w;\Omega') - \mathrm{TV}_\varepsilon(w) \\
&\quad = \frac{1}{\varepsilon} \int_{\Omega\setminus\Omega'} \left(\operatorname*{ess\,sup}_{B(x,\varepsilon)\cap\Omega} u - u(x) \right) \mathrm{d}\varrho_0(x) + \frac{1}{\varepsilon} \int_{\Omega\setminus\Omega'} \left(u(x) - \operatorname*{ess\,inf}_{B(x,\varepsilon)\cap\Omega} u \right) \mathrm{d}\varrho_1(x) \\
&\qquad - \frac{1}{\varepsilon} \int_{\Omega\setminus\Omega'} \left(\operatorname*{ess\,sup}_{B(x,\varepsilon)\cap\Omega} w - w(x) \right) \mathrm{d}\varrho_0(x) - \frac{1}{\varepsilon} \int_{\Omega\setminus\Omega'} \left(w(x) - \operatorname*{ess\,inf}_{B(x,\varepsilon)\cap\Omega} w \right) \mathrm{d}\varrho_1(x)
\end{aligned}$$

Utilizing that $u = w$ on $\Omega \setminus \Omega'_\varepsilon$ one easily sees that all terms in the right hand side of this inequality cancel which renders it equal to zero. Therefore, since u was arbitrary, w is a minimizer as claimed. ∎

Next, we establish a comparison principle for solutions of the minimization problem in (4.1). Because the subdifferential of TV_ε from Lemma 3.2 is not a differential operator, we use variational instead of PDE techniques to prove this comparison principle.

Proposition 4.6 (Comparison principle II)**.** *Let $d \in L^\infty(\Omega)$, let $w \in L^\infty(\Omega)$ solve*

$$w = \arg\min \left\{ \frac{1}{2\varepsilon} \int_\Omega |u-d|^2 \, \mathrm{d}\varrho + \mathrm{TV}_\varepsilon(u) \, : \, u \in L^2(\Omega) \right\},$$

and assume that $v \in L^\infty(\Omega')$ satisfies

$$\frac{1}{2\varepsilon} \int_{\Omega'} |v-d|^2 \, \mathrm{d}\varrho + \mathrm{TV}_\varepsilon(v;\Omega') \leq \frac{1}{2\varepsilon} \int_{\Omega'} |v \vee w - d|^2 \, \mathrm{d}\varrho + \mathrm{TV}_\varepsilon(v \vee w;\Omega')$$

for an open subset $\Omega' \subset \Omega$. If $v \geq w$ almost everywhere on $\Omega' \setminus \Omega'_\varepsilon$ or if $\Omega' = \Omega$, then $v \geq w$ holds almost everywhere in Ω'.

Proof. Let us first abbreviate the energy $E(u) := \frac{1}{2\varepsilon} \int_{\Omega'} |u-d|^2 \, \mathrm{d}\varrho + \mathrm{TV}_\varepsilon(u;\Omega')$ for $u \in L^2(\Omega')$. Lemma 4.5 implies that w is a minimizer of E with fixed data w on $\Omega' \setminus \Omega'_\varepsilon$. By the assumption that $v \geq w$ on $\Omega' \setminus \Omega'_\varepsilon$, we get that $v \wedge w = w$ on $\Omega' \setminus \Omega'_\varepsilon$ and hence $v \wedge w$ is a feasible competitor for w on Ω'. In the case $\Omega' = \Omega$ it is trivial that $v \wedge w$ is a competitor for w on Ω.

Assuming that $v \wedge w \neq w$ on a set of positive measure in Ω' and using the strict convexity of E, we get

$$E(w) < E(v \wedge w).$$

Furthermore, by assumption we get that

$$E(v) \leq E(v \vee w).$$

Summing these two inequalities and using Lemma 3.7 to cancel the total variations we get

$$\begin{aligned} \int_{\Omega'} |v-d|^2 + |w-d|^2 \, \mathrm{d}\varrho &< \int_{\Omega'} |v \vee w - d|^2 + |v \wedge w - d|^2 \, \mathrm{d}\varrho \\ &= \int_{\Omega'} |v-d|^2 + |w-d|^2 \, \mathrm{d}\varrho \end{aligned}$$

which is a contradiction. Hence, we have $v \wedge w = w$ almost everywhere on Ω', proving the claim. ∎

In the next lemma we derive a subsolution of the optimization problem in (4.1) where the data is a cone which corresponds to controlling the action of the scheme (2.7) on a ball. Note that, as opposed to the case of constant densities and a local total variation, we cannot compute the explicit solution. However, a subsolution suffices for our purposes. For consistency with the language in Definition 2, a subsolution is minimal with respect to competitors that are greater than it.

Lemma 4.7 (Subsolution for cone data)**.** *Let $\Omega \subset \mathbb{R}^N$ a bounded and convex domain, $x_0 \in \Omega$, and define $d(x) := |x - x_0|$ for $x \in \mathbb{R}^N$. Let furthermore*

$$w := \operatorname*{arg\,min}_{u \in L^2(\Omega)} \frac{1}{2\varepsilon} \int_\Omega |u-d|^2 \, \mathrm{d}\varrho + \mathrm{TV}_\varepsilon(u).$$

There exist constants $C_1, C_2 \geq 1$ with $2C_2 \geq C_1 > C_2$ and depending only on $\mathrm{Lip}(\varrho_0)$, $\mathrm{Lip}(\varrho_1)$, c_ϱ, $\mathrm{diam}(\Omega)$, and the dimension N, such that for $\varepsilon > 0$ sufficiently small it holds for almost every $x \in \Omega$ that

$$w(x) \leq \overline{w}(x) := \begin{cases} C_1\sqrt{\varepsilon} & \text{if } |x-x_0| \leq C_2\sqrt{\varepsilon}, \\ |x-x_0| + \dfrac{C_2(C_1-C_2)\varepsilon}{|x-x_0|} & \text{else.} \end{cases}$$

Remark 4.8. For a negative cone, i.e., $d(x) = -|x - x_0|$, it is not immediately obvious that $-\overline{w}$ will be a supersolution, in the sense that $w \geq -\overline{w}$. The reason for this is that $\mathrm{TV}_\varepsilon(-u) \neq \mathrm{TV}_\varepsilon(u)$. However, since all constants in the definition of $\overline{w}$ only depend on the Lipschitz constants of ϱ_0 and ϱ_1, the density lower bound c_ϱ, and the dimension N, one can just exchange the roles of the densities in the definition of TV_ε and reduce to the subsolution case of Lemma 4.7.

Proof. For a lighter notation we assume without loss of generality that $x_0 = 0 \in \Omega$. Throughout the proof, we let $C_1 \geq 1$ and $C_2 \geq 1$ be the constants from the lemma statement, deferring their specific choice to the last step of the proof. The strategy is to construct $p \in L^2(\Omega)$ that satisfies

$$(\overline{w} - d)\varrho + \varepsilon p \geq 0 \qquad \text{in } \Omega, \tag{4.13}$$

$$\int_\Omega p\varphi \,\mathrm{d}x \leq \mathrm{TV}_\varepsilon(\overline{w} + \varphi) - \mathrm{TV}_\varepsilon(\overline{w}) \qquad \text{for all } \varphi \in L^2(\Omega),\ \varphi \geq 0. \tag{4.14}$$

These two properties immediately imply that for the energy $E(u) := \frac{1}{2\varepsilon}\int_\Omega |u - d|^2 \,\mathrm{d}\varrho + \mathrm{TV}_\varepsilon(u)$ for $u \in L^2(\Omega)$ it holds $E(\overline{w}) \leq E(\varphi + \overline{w})$ for all non-negative test functions $\varphi \geq 0$. Then Proposition 4.6 with the choice $\varphi := w \vee \overline{w} - \overline{w} \geq 0$ implies the claim that $w \leq \overline{w}$. The required function p is reminiscent of a subgradient of TV_ε at $\overline{w}$ with the difference being that it only satisfies the subdifferential inequality (4.14) for non-negative test functions.

Step 1 (Construction of a "subgradient"). Since $\overline{w}$ is a radial function, constructing p is not difficult. In the spirit of Proposition 3.4, we will define argmin and argmax operators corresponding to the capped cone $\overline{w}$ on $\mathbb{R}^N$. We first note that $\overline{w}$ is radially non-decreasing. Indeed, the derivative of the function $f(r) := r + \frac{C_2(C_1 - C_2)\varepsilon}{r}$ satisfies for $r \geq C_2\sqrt{\varepsilon}$:

$$f'(r) = 1 - \frac{C_2(C_1 - C_2)\varepsilon}{r^2} \geq 1 - \frac{C_2(C_1 - C_2)}{C_2^2} = 1 - \frac{C_1 - C_2}{C_2} = \frac{2C_2 - C_1}{C_2} \geq 0.$$

To construct p we begin by defining the argmax map

$$\Gamma_\varepsilon(x) := \sigma_\varepsilon(|x|)\frac{x}{|x|} \tag{4.15}$$

where the piecewise linear and increasing function σ_ε is defined as

$$\sigma_\varepsilon(t) := t + \varepsilon \min\left\{\frac{t}{C_2\sqrt{\varepsilon} - \varepsilon}, 1\right\} = \begin{cases} \dfrac{t}{1 - \sqrt{\varepsilon}/C_2} & \text{if } 0 \leq t < C_2\sqrt{\varepsilon} - \varepsilon, \\ t + \varepsilon & \text{if } t \geq C_2\sqrt{\varepsilon} - \varepsilon, \end{cases}$$

which in particular satisfies $\sigma_\varepsilon(0) = 0$ so that Γ_ε is well defined at $x = 0$. Note that we can assume $\varepsilon < 1$ so that $C_2\sqrt{\varepsilon} - \varepsilon > 0$. It is important to note that here Γ_ε is a "global" argmax map of $\overline{w}$ that does not see the geometry of the domain Ω, i.e., $\Gamma_\varepsilon(x) \in \arg\max_{B(x,\varepsilon)} \overline{w}$. Note that Γ_ε is invertible and by construction an ε-perturbation of the identity. More precisely, the function σ_ε is invertible on $[0, \infty) \to [0, \infty)$ with inverse $\tau_\varepsilon := \sigma_\varepsilon^{-1}$. Hence, the inverse of Γ_ε is given by

$$\gamma_\varepsilon(x) := \Gamma_\varepsilon^{-1}(x) = \tau_\varepsilon(|x|)\frac{x}{|x|}$$

and thanks to the convexity of Ω it holds that $\gamma_\varepsilon(\Omega) \subset \Omega$. It is immediate from the piecewise definition of σ_ε that

$$\tau_\varepsilon(s) := s - \varepsilon \min\left\{\frac{s}{C_2\sqrt{\varepsilon}}, 1\right\} = \begin{cases} (1 - \sqrt{\varepsilon}/C_2)s & \text{if } 0 \leq s < C_2\sqrt{\varepsilon}, \\ s - \varepsilon & \text{if } s \geq C_2\sqrt{\varepsilon}. \end{cases} \tag{4.16}$$

Using this it is straightforward to see that γ_ε is an argmin map for $\overline{w}$ on $\mathbb{R}^N$.

We can now define the functions

$$\begin{aligned}
p_0(x) &:= \frac{1}{\varepsilon}\bigg(\varrho_0(\Gamma_\varepsilon^{-1}(x))\left|\det\nabla\Gamma_\varepsilon^{-1}(x)\right| - \varrho_0(x)\bigg),\\
p_1(x) &:= \frac{1}{\varepsilon}\bigg(\varrho_1(x) - \varrho_1(\gamma_\varepsilon^{-1}(x))\left|\det\nabla\gamma_\varepsilon^{-1}(x)\right|\bigg)\mathbf{1}_{\gamma_\varepsilon(\Omega)},\\
p(x) &:= p_0(x) + p_1(x).
\end{aligned}$$

Note that the reason why we have to introduce the characteristic function $\mathbf{1}_{\gamma_\varepsilon(\Omega)}$ in the definition of p_1 is that the argmax map $\Gamma_\varepsilon = \gamma_\varepsilon^{-1}$ exits Ω, i.e., $\gamma_\varepsilon^{-1}(\Omega) \not\subset \Omega$.

Step 2 (Validity of the "subdifferential" inequality). Next we prove (4.14) by estimating the L^2-inner product of p_i and φ for $i = 0, 1$, where slightly different arguments are required. We start with $i = 0$. Using a change of variables and $\Gamma_\varepsilon(x) \in \arg\max_{B(x,\varepsilon)\cap\Omega}\overline{w}$ for $x \in \Gamma_\varepsilon^{-1}(\Omega)$ we find

$$\begin{aligned}
\varepsilon\int_\Omega p_0\varphi\,\mathrm{d}x &= \int_\Omega \left(\varrho_0(\Gamma_\varepsilon^{-1})\left|\det\nabla\Gamma_\varepsilon^{-1}\right| - \varrho_0\right)\varphi\,\mathrm{d}x\\
&= \int_{\Gamma_\varepsilon^{-1}(\Omega)} \varphi\circ\Gamma_\varepsilon\,\mathrm{d}\varrho_0 - \int_\Omega \varphi\,\mathrm{d}\varrho_0\\
&= \int_{\Gamma_\varepsilon^{-1}(\Omega)} (\overline{w}+\varphi)\circ\Gamma_\varepsilon\,\mathrm{d}\varrho_0 - \int_\Omega \overline{w}+\varphi\,\mathrm{d}\varrho_0 - \left[\int_{\Gamma_\varepsilon^{-1}(\Omega)} \overline{w}\circ\Gamma_\varepsilon\,\mathrm{d}\varrho_0 - \int_\Omega \overline{w}\,\mathrm{d}\varrho_0\right]\\
&\le \int_\Omega \left(\operatorname*{ess\,sup}_{B(\cdot,\varepsilon)\cap\Omega}(\overline{w}+\varphi) - (\overline{w}+\varphi)\right)\mathrm{d}\varrho_0 - \int_\Omega\left(\operatorname*{ess\,sup}_{B(\cdot,\varepsilon)\cap\Omega}\overline{w} - \overline{w}\right)\mathrm{d}\varrho_0\\
&\quad + \int_{\Omega\setminus\Gamma_\varepsilon^{-1}(\Omega)} -\operatorname*{ess\,sup}_{B(\cdot,\varepsilon)\cap\Omega}(\overline{w}+\varphi) + \operatorname*{ess\,sup}_{B(\cdot,\varepsilon)\cap\Omega}\overline{w}\,\mathrm{d}\varrho_0.
\end{aligned}$$

The last integral on the right-hand side is non-positive because $\varphi \ge 0$. In the last inequality above we used $\overline{w}\circ\Gamma_\varepsilon(x) = \operatorname{ess\,sup}_{B(x,\varepsilon)\cap\Omega}\overline{w}$ for $x \in \Gamma_\varepsilon^{-1}(\Omega)$, and that $(\varphi+\overline{w})\circ\Gamma_\varepsilon \le \operatorname{ess\,sup}_{B(\cdot,\varepsilon)\cap\Omega}(\varphi+\overline{w})$ almost everywhere in $\Gamma_\varepsilon^{-1}(\Omega)$. The reasoning for the latter inequality to hold is analogous to the one in the proof of Proposition 3.4, using that Γ_ε is a Lipschitz isomorphism and hence preserves Lebesgue null-sets.

The case $i = 1$ is treated similarly, although not entirely symmetrically. Using the fact that $\varphi \ge 0$ and that by convexity $\gamma_\varepsilon(\Omega) \subset \Omega$ we get

$$\begin{aligned}
\varepsilon\int_\Omega p_1\varphi\,\mathrm{d}x &= \int_{\gamma_\varepsilon(\Omega)} \left(\varrho_1 - \varrho_1(\gamma_\varepsilon^{-1})\left|\det\nabla\gamma_\varepsilon^{-1}\right|\right)\varphi\,\mathrm{d}x\\
&= \int_{\gamma_\varepsilon(\Omega)} \varphi\,\mathrm{d}\varrho_1 - \int_{\gamma_\varepsilon^{-1}(\gamma_\varepsilon(\Omega))} \varphi\circ\gamma_\varepsilon\,\mathrm{d}\varrho_1\\
&\le \int_\Omega \varphi\,\mathrm{d}\varrho_1 - \int_\Omega \varphi\circ\gamma_\varepsilon\,\mathrm{d}\varrho_1\\
&= \int_\Omega (\overline{w}+\varphi)\,\mathrm{d}\varrho_1 - \int_\Omega (\overline{w}+\varphi)\circ\gamma_\varepsilon\,\mathrm{d}\varrho_1 - \left[\int_\Omega \overline{w}\,\mathrm{d}\varrho_1 - \int_\Omega \overline{w}\circ\gamma_\varepsilon\,\mathrm{d}\varrho_1\right]\\
&\le \int_\Omega \left((\overline{w}+\varphi) - \operatorname*{ess\,inf}_{B(\cdot,\varepsilon)\cap\Omega}(\overline{w}+\varphi)\right)\mathrm{d}\varrho_1 - \int_\Omega\left(\overline{w} - \operatorname*{ess\,inf}_{B(\cdot,\varepsilon)\cap\Omega}\overline{w}\right)\mathrm{d}\varrho_1.
\end{aligned}$$

Here again, we need to argue as before that γ_ε preserves Lebesgue null-sets (as a diffeomorphism) for the validity of the last inequality. Adding the two inequalities we have just established and dividing by $\varepsilon > 0$ proves (4.14).

Step 3 (Optimality conditions for supersolution). It remains to prove (4.13), i.e., the inequality $(\overline{w}-d)\varrho+\varepsilon p\geq 0$. The basic idea is that $\overline{w}-d\approx\sqrt{\varepsilon}$ by the choice of $\overline{w}$, so that the inequality will follow if we can show $\varepsilon p\geq-\sqrt{\varepsilon}$ (up to a constant multiple). Importantly, within this step, we let $C_\varrho>0$ be a constant (possibly changing from line to line) depending only on $\mathrm{Lip}(\varrho_i)$, c_ϱ, $\mathrm{diam}(\Omega)$, and the dimension N.

For this we first compute $\det\nabla\Gamma_\varepsilon^{-1}(x)$ which appears in the definition of p_0. The Jacobian of Γ_ε^{-1} is given by

$$\begin{aligned}\nabla\Gamma_\varepsilon^{-1}(x)&=\tau_\varepsilon'(|x|)\frac{x}{|x|}\otimes\frac{x}{|x|}+\frac{\tau_\varepsilon(|x|)}{|x|}\left(\mathbb{1}-\frac{x}{|x|}\otimes\frac{x}{|x|}\right)\\&=\frac{\tau_\varepsilon(|x|)}{|x|}\left[\mathbb{1}+\left(\tau_\varepsilon'(|x|)\frac{|x|}{\tau_\varepsilon(|x|)}-1\right)\frac{x}{|x|}\otimes\frac{x}{|x|}\right].\end{aligned}$$

The derivative of τ_ε defined in (4.16) is given by $\tau_\varepsilon'(t)=1-\frac{\sqrt{\varepsilon}}{C_2}\mathbf{1}_{\{t\leq C_2\sqrt{\varepsilon}\}}$, so that

$$\left(\tau_\varepsilon'(|x|)\frac{|x|}{\tau_\varepsilon(|x|)}-1\right)=\begin{cases}0 & \text{if } |x|\leq C_2\sqrt{\varepsilon},\\ \frac{\varepsilon}{|x|-\varepsilon} & \text{if } |x|\geq C_2\sqrt{\varepsilon}.\end{cases}$$

Using this, we make a case distinction based on $|x|$: For $|x|<C_2\sqrt{\varepsilon}$, using also the elementary inequality $(1+x)^N\geq 1+Nx$ for $x\geq-1$, we get

$$\det(\nabla\Gamma_\varepsilon^{-1}(x))=\left(\frac{\tau_\varepsilon(|x|)}{|x|}\right)^N\det(\mathbb{1})=\left(1-\frac{\sqrt{\varepsilon}}{C_2}\right)^N\geq 1-\frac{\sqrt{\varepsilon}N}{C_2}\tag{4.17}$$

since $\varepsilon\leq 1\leq C_2^2$. For $|x|\geq C_2\sqrt{\varepsilon}$ we can use a Taylor expansion of the determinant to get

$$\begin{aligned}\det(\nabla\Gamma_\varepsilon^{-1}(x))&=\left(1-\frac{\varepsilon}{|x|}\right)^N\det\left(\mathbb{1}+\frac{\varepsilon}{|x|-\varepsilon}\frac{x}{|x|}\otimes\frac{x}{|x|}\right)\\&\geq\left(1-\frac{\varepsilon N}{|x|}\right)\left(1+\frac{\varepsilon}{|x|-\varepsilon}+O\left(\left(\frac{\varepsilon}{|x|-\varepsilon}\right)^2\right)\right)\geq\left(1-\frac{\varepsilon N}{|x|}\right),\end{aligned}\tag{4.18}$$

where we note that for $\varepsilon>0$ sufficiently small the term $\frac{\varepsilon}{|x|-\varepsilon}$ dominates the quadratic one and, consequently, both can be dropped.

Using Lipschitz continuity of ϱ_0 and the explicit formula for τ_ε in (4.16) we also have

$$\begin{aligned}\varrho_0(\Gamma_\varepsilon^{-1}(x))=\varrho_0\left(\tau_\varepsilon(|x|)\frac{x}{|x|}\right)&\geq\varrho_0(x)-\mathrm{Lip}(\varrho_0)\left|x-\left(|x|-\varepsilon\min\left\{\frac{|x|}{C_2\sqrt{\varepsilon}},1\right\}\right)\frac{x}{|x|}\right|\\&=\varrho_0(x)-\varepsilon C_\varrho\min\left\{\frac{|x|}{C_2\sqrt{\varepsilon}},1\right\}.\end{aligned}\tag{4.19}$$

Combining (4.17), (4.18) and (4.19) we obtain the following lower bound for p_0:

$$\begin{aligned}\varepsilon p_0(x)&\geq\left(\varrho_0(x)-\varepsilon C_\varrho\min\left\{\frac{|x|}{C_2\sqrt{\varepsilon}},1\right\}\right)\left(1-\frac{\varepsilon N}{\max\{|x|,C_2\sqrt{\varepsilon}\}}\right)-\varrho_0(x)\\&\geq-\sqrt{\varepsilon}C_\varrho\left(\min\left\{\frac{|x|}{C_2},\sqrt{\varepsilon}\right\}+\frac{\sqrt{\varepsilon}}{\max\{|x|,C_2\sqrt{\varepsilon}\}}\right).\end{aligned}$$

From this we obtain two lower bounds—a generic one and an improved estimate away from the cone tip:

$$\varepsilon p_0(x)\geq-\sqrt{\varepsilon}C_\varrho\left(\sqrt{\varepsilon}+\frac{1}{C_2}\right)\geq-\sqrt{\varepsilon}C_\varrho\qquad\text{for all } x\in\Omega,\tag{4.20}$$

$$\varepsilon p_0(x) \geq -\sqrt{\varepsilon} C_\varrho \left(\sqrt{\varepsilon} + \frac{\sqrt{\varepsilon}}{|x|} \right) \geq -\frac{\varepsilon}{|x|} C_\varrho \qquad \text{if } |x| \geq C_2 \sqrt{\varepsilon}, \tag{4.21}$$

where in the last inequality we have absorbed $\operatorname{diam}(\Omega)$ into C_ϱ.

We continue with proving a similar bound for p_1. For this we remember that $\gamma_\varepsilon^{-1}(x) = \Gamma_\varepsilon(x) = \sigma_\varepsilon(|x|)\frac{x}{|x|}$. Analogous to before, we get that the Jacobian is

$$\nabla \gamma_\varepsilon^{-1}(x) = \frac{\sigma_\varepsilon(|x|)}{|x|} \left[\mathbb{1} + \left(\sigma_\varepsilon'(|x|) \frac{|x|}{\sigma_\varepsilon(|x|)} - 1 \right) \frac{x}{|x|} \otimes \frac{x}{|x|} \right],$$

and we find $\sigma_\varepsilon'(t) = 1 + \frac{\sqrt{\varepsilon}}{C_2 - \sqrt{\varepsilon}} \mathbf{1}_{\{t \leq C_2\sqrt{\varepsilon} - \varepsilon\}}$ and compute

$$\left(\sigma_\varepsilon'(|x|) \frac{|x|}{\sigma_\varepsilon(|x|)} - 1 \right) = \begin{cases} 0 & \text{if } |x| \leq C_2\sqrt{\varepsilon} - \varepsilon, \\ -\dfrac{\varepsilon}{|x| + \varepsilon} & \text{if } |x| \geq C_2\sqrt{\varepsilon} - \varepsilon. \end{cases}$$

Making case distinctions, as before, and also using that $(1+x)^N \leq 1 + 2Nx$ for sufficiently small x, it holds for $|x| \leq C_2\sqrt{\varepsilon} - \varepsilon$ that

$$\det(\nabla \gamma_\varepsilon^{-1}(x)) = \left(1 + \frac{\varepsilon}{C_2\sqrt{\varepsilon} - \varepsilon} \right)^N \leq 1 + \frac{2\sqrt{\varepsilon} N}{C_2 - \sqrt{\varepsilon}} \tag{4.22}$$

whenever $\varepsilon > 0$ is sufficiently small (depending on C_2). Similarly, for $|x| \geq C_2\sqrt{\varepsilon} - \varepsilon$, we have

$$\begin{aligned} \det(\nabla \gamma_\varepsilon^{-1}(x)) &\leq \left(1 + \frac{2\varepsilon N}{|x|} \right) \left(1 - \frac{\varepsilon}{|x| + \varepsilon} + O\left(\left(\frac{\varepsilon}{|x| + \varepsilon} \right)^2 \right) \right) \\ &\leq \left(1 + \frac{2\varepsilon N}{|x|} \right) \end{aligned} \tag{4.23}$$

for $\varepsilon > 0$ sufficiently small. Using Lipschitz continuity of ϱ_1 we have

$$\begin{aligned} \varrho_1(\gamma_\varepsilon^{-1}(x)) = \varrho_1 \left(\sigma_\varepsilon(|x|) \frac{x}{|x|} \right) &\leq \varrho_1(x) + \operatorname{Lip}(\varrho_1) \left| x - \left(|x| + \varepsilon \min\left\{ \frac{|x|}{C_2\sqrt{\varepsilon} - \varepsilon}, 1 \right\} \right) \frac{x}{|x|} \right| \\ &= \varrho_1(x) + \varepsilon C_\varrho \min\left\{ \frac{|x|}{C_2\sqrt{\varepsilon} - \varepsilon}, 1 \right\}. \end{aligned} \tag{4.24}$$

Combining (4.22), (4.23) and (4.24) we obtain the following lower bound on $p_1(x)$ for $x \in \gamma_\varepsilon(\Omega)$:

$$\begin{aligned} \varepsilon p_1(x) &\geq \varrho_1(x) - \left(\varrho_1(x) + \varepsilon C_\varrho \min\left\{ \frac{|x|}{C_2\sqrt{\varepsilon} - \varepsilon}, 1 \right\} \right) \left(1 + \frac{2\varepsilon N}{\max\{|x|, C_2\sqrt{\varepsilon} - \varepsilon\}} \right) \\ &= -\sqrt{\varepsilon} C_\varrho \left(\min\left\{ \frac{|x|}{C_2 - \sqrt{\varepsilon}}, \sqrt{\varepsilon} \right\} + \frac{\sqrt{\varepsilon}}{\max\{|x|, C_2\sqrt{\varepsilon} - \varepsilon\}} \right) \\ &\quad - \sqrt{\varepsilon} C_\varrho \min\left\{ \frac{|x|}{C_2 - \sqrt{\varepsilon}}, \sqrt{\varepsilon} \right\} \frac{\varepsilon}{\max\{|x|, C_2\sqrt{\varepsilon} - \varepsilon\}} \\ &\geq -\sqrt{\varepsilon} C_\varrho \left(\min\left\{ \frac{|x|}{C_2 - \sqrt{\varepsilon}}, \sqrt{\varepsilon} \right\} + \frac{\sqrt{\varepsilon}}{\max\{|x|, C_2\sqrt{\varepsilon} - \varepsilon\}} \right) \end{aligned}$$

if we restrict $\sqrt{\varepsilon} \leq \frac{1}{2}$ which, in particular, means $\varepsilon \leq C_2\sqrt{\varepsilon} - \varepsilon$. Again we deduce two lower bounds, using also that for $x \in \Omega \setminus \gamma_\varepsilon(\Omega)$ we even have $p_1(x) = 0$ by definition of p_1,

$$\varepsilon p_1(x) \geq -\sqrt{\varepsilon} C_\varrho \qquad x \in \Omega, \tag{4.25}$$

$$\varepsilon p_1(x) \geq -\frac{\varepsilon}{|x|} C_\varrho \qquad\qquad |x| \geq C_2\sqrt{\varepsilon} - \varepsilon. \tag{4.26}$$

Adding the bounds (4.20), (4.21), (4.25) and (4.26) we obtain the cumulative lower bound

$$\varepsilon p(x) \geq -\sqrt{\varepsilon} C_\varrho \qquad \text{for all } x \in \Omega, \tag{4.27}$$

$$\varepsilon p(x) \geq -\frac{\varepsilon}{|x|} C_\varrho \qquad \text{if } |x| \geq C_2\sqrt{\varepsilon}. \tag{4.28}$$

Finally, we can now turn to proving (4.13), making a case distinction based on $|x|$. If $0 \leq |x| \leq C_2\sqrt{\varepsilon}$ we can use the definition of $\overline{w}$ and the lower bound (4.27) to get

$$\begin{aligned}(\overline{w}(x) - d(x))\varrho(x) + \varepsilon p(x) &\geq (C_1\sqrt{\varepsilon} - |x|)\varrho(x) + \varepsilon p(x) \\ &\geq \sqrt{\varepsilon}\,(C_1 - C_2 - C_\varrho)\,\varrho(x) \geq 0\end{aligned}$$

if we choose the gap between C_1 and C_2 sufficiently large (depending only on C_ϱ). In the case $|x| \geq C_2\sqrt{\varepsilon}$ we can use the sharper lower bound (4.28) to obtain

$$\begin{aligned}(\overline{w}(x) - d(x))\varrho(x) + \varepsilon p(x) &\geq \frac{C_2(C_1 - C_2)\varepsilon}{|x|}\varrho(x) - \frac{\varepsilon}{|x|} C_\varrho \varrho(x) \\ &= \frac{\varepsilon\varrho(x)}{|x|}\,(C_2(C_1 - C_2) - C_\varrho) \geq 0\end{aligned}$$

if we choose the gap between C_1 and C_2 sufficiently large (again depending only on C_ϱ). Hence, we have proved (4.13) which concludes the proof. □

The first corollary of Lemma 4.7 (in fact of Remark 4.8) is that it allows us to control the evolution of a ball under the scheme (2.7).

Corollary 4.9 (Supersolution for balls)**.** *Under the conditions of Lemma 4.7 there exists a constant $C > 0$, depending only on* $\mathrm{Lip}(\varrho_0)$, $\mathrm{Lip}(\varrho_1)$, c_ϱ, $\mathrm{diam}(\Omega)$, *and the dimension N, such that for any $x_0 \in \Omega$, $0 < R < \mathrm{dist}(x_0, \partial\Omega)$, and $\varepsilon > 0$ sufficiently small it holds that*

$$S_\varepsilon(B(x_0, R)) \supset B(x_0, R - C\sqrt{\varepsilon}).$$

Proof. We apply Remark 4.8 to $d(x) := -|x - x_0|$ and note that $\mathrm{sdist}(x, B(x_0, R)^c) = d(x) + R$ to infer that w_ε^* in the definition of $S_\varepsilon(B(x_0, R))$ satisfies for almost all $x \in \Omega$ that

$$w_\varepsilon^*(x) \geq R - C\sqrt{\varepsilon} - |x - x_0|\,.$$

Here, $C > 0$ is a constant depending on C_1, C_2 in the definition of $\overline{w}$ in Lemma 4.7. Hence, we obtain $S_\varepsilon(B(x_0, R)) = \{w_\varepsilon^* > 0\} \supset B(x_0, R - C\sqrt{\varepsilon})$. □

We highlight the almost-regularity of L^2-minimizers with Lipschitz data as an application of the cone Lemma 4.7. Our proof of consistency basically relies on the same argument, allowing us to avoid Lipschitz regularity. We note that in the periodic setting with constant densities, one can use a simple comparison argument to show that minimizers of the functional in (4.29) (just below) are Lipschitz regular; but the moment one destroys translational invariance of the problem, such regularity becomes much more challenging. See for instance [18, Theorem 3.1] for the related TV problem with boundary conditions. Similar almost-Lipschitz regularity results for solutions of nonlocal problems can be found, for instance, in [5, 10].

Corollary 4.10 (Almost-Lipschitz regularity.)**.** *Let $\Omega \subset \mathbb{R}^N$ be a bounded, open, and convex set. Suppose that $w \in L^\infty(\Omega)$ satisfies*

$$w := \arg\min \left\{ \frac{1}{2\varepsilon} \int_\Omega |u - f|^2 \, \mathrm{d}\varrho + \mathrm{TV}_\varepsilon(u) \,:\, u \in L^2(\Omega) \right\}, \tag{4.29}$$

for a 1-Lipschitz function $f \in C(\Omega)$. Then for almost all $x, x_0 \in \Omega$ it holds that

$$|w(x) - w(x_0)| \le |x - x_0| + C\sqrt{\varepsilon}, \tag{4.30}$$

where $C > 0$ depends only on ϱ, $\mathrm{diam}(\Omega)$, and the dimension N.

Proof. Fix $x_0 \in \Omega$. Let w_{shift} be the minimizer

$$w_{\mathrm{shift}} := \arg\min \left\{ \frac{1}{2\varepsilon} \int_\Omega |u(x) - |x - x_0||^2 \, \mathrm{d}\varrho(x) + \mathrm{TV}_\varepsilon(u) \,:\, u \in L^2(\Omega) \right\},$$

and $\overline{w}$ be the function defined in Lemma 4.7. As f is 1-Lipschitz, we have $f(\cdot) \le |\cdot - x_0| + f(x_0)$, and so by Proposition 4.3 applied to w and $w' = w_{\mathrm{shift}} + f(x_0)$ and Lemma 4.7 applied to w_{shift}, we have $w \le w_{\mathrm{shift}} + f(x_0) \le \overline{w} + f(x_0)$. Applying the same reasoning with supersolutions (using Remark 4.8) and noting that $\overline{w}(\cdot) \le |\cdot - x_0| + C\sqrt{\varepsilon}$ gives that

$$|w(x) - f(x_0)| \le |x - x_0| + C\sqrt{\varepsilon}$$

for any $x \in \Omega$. Using this inequality twice (once with $x = x_0$), applying the triangle inequality, and increasing C directly gives (4.30). ☐

Finally, we conclude the proof of Theorem 2 by showing that the operator is consistent.

Proposition 4.11 (Consistency)**.** *The operator S_ε defined in* (4.1) *is consistent in the sense of Definition 3.*

Proof. The proof follows the strategy of [14, Proposition 4.1] with non-trivial modifications since the scheme (4.1) involves the nonlocal total variation instead of the local one. We show that Definition 3 is satisfied for subflows, with superflows being analogous.

Step 1 (Construction of an variational subsolution). We let $[t_0, t_1] \ni t \mapsto A(t)$ be a subflow in the sense of Definition 2 contained in the neighborhood B. Recall that we define $d(x,t) := \mathrm{sdist}(x, A^c(t))$.

For fixed $t \in [t_0, t_1]$ and $r > 0$ we define

$$\Omega' := \{|d(\cdot, t)| < r\}$$

to be the tube of width r around the boundary $\partial A(t)$. Here we choose r sufficiently small such that $\overline{\Omega'} \subset B \cap (\mathbb{R}^N \times \{t\})$, so that $d(x, t+\tau)$ is in $C^{2,1}_{x,\tau}(\Omega' \times [-\varepsilon, \varepsilon])$ for all small ε; note the choice of r can be made independent of t depending only on the smooth subflow. Let $\psi : \mathbb{R} \to \mathbb{R}$ be smooth with $\psi(s) \ge s$, $\psi(s) = s$ for s in a neighborhood of 0, and $\psi'(s) \ge c > 0$, and define $v_\varepsilon(x) := \psi(d(x, t+\varepsilon))$. By Definition 2, it holds that

$$\begin{aligned} \frac{v_\varepsilon(x) - d(x,t)}{\varepsilon} &\ge \frac{d(x, t+\varepsilon) - d(x,t)}{\varepsilon} \\ &= \fint_0^\varepsilon \frac{\mathrm{d}}{\mathrm{d}\tau} d(x, t+\tau) \, \mathrm{d}\tau \\ &= \fint_0^\varepsilon \partial_t d(x, t+\tau) \, \mathrm{d}\tau \end{aligned}$$

$$\geq \fint_0^\varepsilon \frac{1}{\varrho(x)} \operatorname{div}\left(\varrho(x)\nabla d(x,t+\tau)\right) \,\mathrm{d}\tau + \delta.$$

Letting ω denote a modulus of continuity of $\frac{1}{\varrho(x)} \operatorname{div}\left(\varrho(x)\nabla d(x,\tau)\right)$ in τ (uniformly in x) on B, we have

$$\frac{v_\varepsilon(x) - d(x,t)}{\varepsilon} \geq \frac{1}{\varrho(x)} \operatorname{div}\left(\varrho(x)\nabla d(x,t+\varepsilon)\right) + \delta - \omega(\varepsilon). \tag{4.31}$$

Note furthermore that $\nabla v_\varepsilon(x) = \psi'(d(x,t+\varepsilon))\nabla d(x,t+\varepsilon)$. On one hand this implies

$$|\nabla v_\varepsilon| \geq c \quad \text{ in } \Omega', \tag{4.32}$$

which will be useful later. On the other hand, we see that

$$\frac{\nabla v_\varepsilon(x)}{|\nabla v_\varepsilon(x)|} = \frac{\nabla d(x,t+\varepsilon)}{|\nabla d(x,t+\varepsilon)|} = \nabla d(x,t+\varepsilon).$$

Using this and reordering (4.31) we get

$$\frac{v_\varepsilon(x) - d(x,t)}{\varepsilon}\varrho(x) - \operatorname{div}\left(\varrho(x)\frac{\nabla v_\varepsilon(x)}{|\nabla v_\varepsilon(x)|}\right) - \varrho(x)\left(\delta - \omega(\varepsilon)\right) \geq 0. \tag{4.33}$$

Let $\varphi \in L^\infty(\Omega')$ be a non-negative test function with $\operatorname{supp}\varphi \subset \Omega'_{2\varepsilon}$. Let us also define the energy

$$E_\varepsilon(u;\Omega') := \frac{1}{2\varepsilon}\int_{\Omega'} |u - d(\cdot,t)|^2 \,\mathrm{d}\varrho + \mathrm{TV}_\varepsilon(u;\Omega'),$$

where $\mathrm{TV}_\varepsilon(u;\Omega')$ denotes the total variation (2.6) with Ω replaced by Ω'. We let

$$V_\varepsilon(y) := \frac{\mathrm{d}}{\mathrm{d}\mathcal{L}^N}\left[\frac{(\Gamma_\varepsilon)_\sharp\varrho_0 - \varrho_0}{\varepsilon} + \frac{\varrho_1 - (\gamma_\varepsilon)_\sharp\varrho_1}{\varepsilon}\right]$$

be the density of the pushforward, where the right-hand side is defined as in Proposition 3.4 with u replaced by v_ε, which is admissible due to (4.32). Multiplying (4.33) by φ, using its non-negativity, and integrating over Ω' yields

$$\begin{aligned}
E_\varepsilon(v_\varepsilon;\Omega') &\leq E_\varepsilon(v_\varepsilon;\Omega') + \int_{\Omega'}\left(\frac{v_\varepsilon(x) - d(x,t)}{\varepsilon}\varrho(x) - \operatorname{div}\left(\varrho(x)\frac{\nabla v_\varepsilon(x)}{|\nabla v_\varepsilon(s)|}\right)\right)\varphi(x)\,\mathrm{d}x \\
&\quad - \left(\delta - \omega(\varepsilon)\right)\int_{\Omega'}\varphi(x)\,\mathrm{d}\varrho(x) \\
&= E_\varepsilon(v_\varepsilon;\Omega') + \int_{\Omega'}\left(\frac{v_\varepsilon(x) - d(x,t)}{\varepsilon}\varrho(x) + V_\varepsilon(x)\right)\varphi(x)\,\mathrm{d}x \\
&\quad - \int_{\Omega'}\left(V_\varepsilon(x) + \operatorname{div}\left(\varrho(x)\frac{\nabla v_\varepsilon(x)}{|\nabla v_\varepsilon(s)|}\right)\right)\varphi(x)\,\mathrm{d}x \\
&\quad - \left(\delta - \omega(\varepsilon)\right)\int_{\Omega'}\varphi(x)\,\mathrm{d}\varrho(x) \\
&\leq E_\varepsilon(v_\varepsilon + \varphi;\Omega') + \left(\omega(\varepsilon) + o_{\varepsilon\to 0}(1) - \delta\right)\int_{\Omega'}\varphi(x)\,\mathrm{d}\varrho(x),
\end{aligned}$$

where in the last step, we completed a square, used Propositions 3.4 and 3.5 on Ω' together with the gradient bound (4.32) for v_ε and the fact that $\operatorname{supp}\varphi \subset \Omega'_{2\varepsilon}$. Since $\delta > 0$ we can choose $\varepsilon > 0$ sufficiently small such that the second term is non-positive which implies

$$E_\varepsilon(v_\varepsilon; \Omega') \leq E_\varepsilon(v_\varepsilon + \varphi; \Omega') \tag{4.34}$$

f or $\varepsilon > 0$ sufficiently small.

Step 2 (Conclusion, assuming ordered boundary values). Supposing that

$$v_\varepsilon \geq w_\varepsilon^* \quad \text{ on } \Omega' \setminus \Omega'_{2\varepsilon}, \tag{4.35}$$

the non-negative test function $\varphi_\varepsilon := v_\varepsilon \vee w_\varepsilon^* - v_\varepsilon$ (where w_ε^* solves the scheme (4.1)) can be inserted into (4.34). Consequently, we have that $E_\varepsilon(v_\varepsilon) \leq E_\varepsilon(v_\varepsilon \vee w_\varepsilon^*)$, and Proposition 4.6 implies that $v_\varepsilon \geq w_\varepsilon^*$ on Ω' (one must technically deal with null-sets, but this may be done as in the proof of Proposition 4.4), and hence, we find

$$S_\varepsilon(A(t)) \cap \Omega' = \{w_\varepsilon^* > 0\} \cap \Omega' \subset \{v_\varepsilon > 0\} \cap \Omega' = \{d(\cdot, t+\varepsilon) > 0\} \cap \Omega' = A(t+\varepsilon) \cap \Omega'.$$

Similarly, we will see in the next step,

$$\{w_\varepsilon^* > 0\} \setminus \Omega' \subset \{d(\cdot, t) > 0\} \setminus \Omega'. \tag{4.36}$$

Further, outside of the set Ω', for sufficiently small ε (depending only on the smooth subflow), we have $A(t) \setminus \Omega' = A(t+\varepsilon) \setminus \Omega'$. Putting these last two pieces together, we recover

$$S_\varepsilon(A(t)) \setminus \Omega' = \{w_\varepsilon^* > 0\} \setminus \Omega' \subset \{d(\cdot, t) > 0\} \setminus \Omega' = A(t+\varepsilon) \setminus \Omega'.$$

Uniting the subset relations for $S_\varepsilon(A(t))$ concludes the proof. We now turn to the proof of (4.35) and (4.36).

Step 3 (Ordered boundary values). To prove (4.35), we have to pick a suitable function ψ in the definition of $v_\varepsilon = \psi(d(\cdot, t+\varepsilon))$. So far we have only used that $\psi(s) \geq s$, $\psi(s) = s$ in a neighborhood of 0, and that $\psi' \geq c > 0$.

First, we argue that one can find ψ such that $v_\varepsilon \geq w_\varepsilon^*$ in the part of the 2ε-neighborhood of the boundary of Ω' that lies inside of $A(t)$: For all $\varepsilon > 0$ small enough and $x \in \partial\Omega'$ with $d(x,t) = r$, it holds that $d(x, t+\varepsilon) \geq \frac{7r}{8}$ since d is uniformly continuous in time. Since d is also uniformly continuous in space, we get $d(x, t+\varepsilon) \geq \frac{3r}{4}$ for all x in $(\Omega' \setminus \Omega'_{2\varepsilon}) \cap A(t)$ if $\varepsilon > 0$ is sufficiently small. On the other hand, by Proposition 4.1 it holds that $w_\varepsilon^* \leq \|d(\cdot,t)\|_{L^\infty(\Omega)} \leq \operatorname{diam}\Omega$. So if we choose ψ such that $\psi(\frac{3r}{4}) \geq \operatorname{diam}\Omega$, we get

$$v_\varepsilon(x) = \psi(d(x, t+\varepsilon)) \geq \psi\left(\frac{3r}{4}\right) \geq \operatorname{diam}\Omega \geq w_\varepsilon^*(x)$$

for all $x \in (\Omega' \setminus \Omega'_{2\varepsilon}) \cap A(t)$.

Next, we argue that also in the 2ε-neighborhood of the exterior part of the boundary of Ω' one can find an appropriate ψ such that $v_\varepsilon \geq w_\varepsilon^*$: The argument for this is more involved than for the inner part since in principle w_ε^* could be arbitrarily close to zero outside of $A(t)$ whereas the signed distance function $d(\cdot, t+\varepsilon)$ in the definition of v_ε might be substantially negative. If this happened, to obtain $v_\varepsilon \geq w_\varepsilon^*$ we could be forced to take $\psi(-3r/4) = 0$ breaking the constraint $\psi' \geq c$.

Instead, fix a point $x \in \partial\Omega'$ such that $d(x,t) = -r$. Since the distance function is 1-Lipschitz

$$d(\cdot, t) \leq |\cdot - x| - r,$$

and therefore Proposition 4.3 ensures that $w_\varepsilon^* \leq w_\varepsilon$ in Ω where

$$w_\varepsilon := \operatorname*{arg\,min}_{u \in L^2(\Omega)} \frac{1}{2\varepsilon} \int_\Omega |u - (|\cdot - x| - r)|^2 \, \mathrm{d}\varrho + \mathrm{TV}_\varepsilon(u).$$

However, by Lemma 4.7, $w_\varepsilon \leq \overline{w} - r$, and it follows that $w_\varepsilon^* \leq -r + C_1\sqrt{\varepsilon}$ for all $|x' - x| \leq C_2\sqrt{\varepsilon}$. Noting that this reasoning can be uniformly applied at all points $x \in \partial\Omega'$ with $d(x,t) = -r$, we see that for sufficiently small $\varepsilon > 0$,

$$w_\varepsilon^* \leq -\frac{3}{4}r \quad \text{on} \quad (\Omega' \setminus \Omega'_{2\varepsilon}) \setminus A(t). \tag{4.37}$$

Restricting ψ to satisfy $\psi(t) \geq -\frac{3r}{4}$ for any $t \geq -\operatorname{diam}\Omega$ we therefore get

$$v_\varepsilon(x) = \psi(d(x, t+\varepsilon)) \geq -\frac{3r}{4} \geq w_\varepsilon^*(x)$$

for all $x \in (\Omega' \setminus \Omega'_{2\varepsilon}) \setminus A(t)$. Hence, we have shown (4.35). The same reasoning used to obtain (4.37), but now at a point for which $d(x,t) \leq -r$, shows

$$\{w_\varepsilon^* > 0\} \setminus (\Omega' \cup A(t)) \subset \{d(\cdot, t) > 0\} \setminus (\Omega' \cup A(t)) = \emptyset$$

which directly gives (4.36), completing the proof.

Note we have proven that there exists an $\varepsilon_0 > 0$ sufficiently small such that consistency in Definition 3 is satisfied at a given time t for all $\varepsilon < \varepsilon_0$, but actually, our estimate for ε_0 is uniform in $t \in [t_0, t_1]$. □

Acknowledgments

The authors would like to thank Antonin Chambolle for fruitful discussions around the construction subsolutions for cone data which happened during the Oberwolfach workshop 2349 "Variational Methods for Evolution". Parts of this work were done when LB was affiliated with the Technical University of Berlin, supported by Germany's Excellence Strategy – The Berlin Mathematics Research Center MATH+ (EXC-2046/1, project ID: 390685689). The authors were affiliated with the Hausdorff Center for Mathematics during parts of this project and the funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC-2047/1 – 390685813 is greatly appreciated. KS was also supported by the DFG project 211504053 - SFB 1060.

References

[1] F. Almgren, J. E. Taylor, and L. Wang, "Curvature-driven flows: A variational approach," *SIAM Journal on Control and Optimization*, vol. 31, no. 2, pp. 387–438, 1993 (cit. on pp. 3–5).

[2] P. Awasthi, N. Frank, and M. Mohri, "On the existence of the adversarial bayes classifier," in *Advances in Neural Information Processing Systems 34, NeurIPS 2021*, M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021, pp. 2978–2990 (cit. on p. 2).

[3] G. Bellettini, *Lecture Notes on Mean Curvature Flow, Barriers and Singular Perturbations*. Scuola Normale Superiore, 2013 (cit. on p. 14).

[4] G. Bellettini and M. Novaga, "Minimal barriers for geometric evolutions," *Journal of Differential Equations*, vol. 139, no. 1, pp. 76–103, 1997 (cit. on p. 7).

[5] L. Bungert, J. Calder, and T. Roith, "Uniform convergence rates for lipschitz learning on graphs," *IMA Journal of Numerical Analysis*, vol. 43, no. 4, pp. 2445–2495, 2023 (cit. on p. 25).

[6] L. Bungert, N. García Trillos, M. Jacobs, D. McKenzie, Đ. Nikolić, and Q. Wang, *It begins with a boundary: A geometric view on probabilistically robust learning*, 2023. arXiv: 2305.18779 [cs.LG] (cit. on p. 2).

[7] L. Bungert, N. García Trillos, and R. Murray, "The geometry of adversarial training in binary classification," *Information and Inference: A Journal of the IMA*, vol. 12, no. 2, pp. 921–968, Jun. 2023, ISSN: 2049-8772 (cit. on pp. 2, 4, 6–8, 12, 13).

[8] L. Bungert and K. Stinson, "Gamma-convergence of a nonlocal perimeter arising in adversarial machine learning," *Calculus of Variations and Partial Differential Equations*, 2024, forthcoming (cit. on pp. 2, 4, 8).

[9] J. Calder, B. Cook, M. Thorpe, and D. Slepcev, "Poisson learning: Graph based semi-supervised learning at very low label rates," in *International Conference on Machine Learning*, PMLR, 2020, pp. 1306–1316 (cit. on p. 3).

[10] J. Calder, N. García Trillos, and M. Lewicka, "Lipschitz regularity of graph laplacians on random data clouds," *SIAM Journal on Mathematical Analysis*, vol. 54, no. 1, pp. 1169–1222, 2022 (cit. on p. 25).

[11] A. Chambolle, "An algorithm for mean curvature motion," *Interfaces and Free Boundaries*, vol. 6, no. 2, pp. 195–218, 2004 (cit. on pp. 3, 5–7, 16, 18).

[12] A. Chambolle, D. De Gennaro, and M. Morini, "Minimizing movements for anisotropic and inhomogeneous mean curvature flows," *Advances in Calculus of Variations*, no. 0, 2023 (cit. on p. 7).

[13] A. Chambolle, A. Giacomini, and L. Lussardi, "Continuous limits of discrete perimeters," *ESAIM: Mathematical Modelling and Numerical Analysis*, vol. 44, no. 2, pp. 207–230, 2010 (cit. on p. 13).

[14] A. Chambolle and M. Novaga, "Approximation of the anisotropic mean curvature flow," *Mathematical Models and Methods in Applied Sciences*, vol. 17, no. 06, pp. 833–844, 2007 (cit. on pp. 7, 14, 19, 26).

[15] T. F. Chan and S. Esedoglu, "Aspects of total variation regularized L^1 function approximation," *SIAM Journal on Applied Mathematics*, vol. 65, no. 5, pp. 1817–1837, 2005 (cit. on p. 5).

[16] K. Crane, C. Weischedel, and M. Wardetzky, "Geodesics in heat: A new approach to computing distance based on heat flow," *ACM Transactions on Graphics (TOG)*, vol. 32, no. 5, pp. 1–11, 2013 (cit. on p. 5).

[17] T. Eto and Y. Giga, "On a minimizing movement scheme for mean curvature flow with prescribed contact angle in a curved domain and its computation," *Annali di Matematica Pura ed Applicata*, pp. 1–27, Nov. 2023 (cit. on p. 7).

[18] T. Eto and Y. Giga, *A convergence result for a minimizing movement scheme for mean curvature flow with prescribed contact angle in a curved domain*, 2024. arXiv: 2402.16180 [math.AP] (cit. on pp. 7, 25).

[19] N. Garcá Trillos, M. Jacobs, and J. Kim, *On the existence of solutions to adversarial training in multiclass classification*, 2023. arXiv: 2305.00075 [cs.LG] (cit. on p. 2).

[20] N. García Trillos and M. Jacobs, "An analytical and geometric perspective on adversarial robustness," *Notices of the American Mathematical Society*, vol. 70, no. 08, 2023 (cit. on p. 2).

[21] N. García Trillos, M. Jacobs, and J. Kim, "The multimarginal optimal transport formulation of adversarial multiclass classification," *Journal of Machine Learning Research*, vol. 24, no. 45, pp. 1–56, 2023 (cit. on p. 2).

[22] N. García Trillos and R. Murray, "Adversarial classification: Necessary conditions and geometric flows," *Journal of Machine Learning Research*, vol. 23, no. 187, pp. 1–38, 2022 (cit. on pp. 2–4).

[23] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *3rd International Conference on Learning Representations, ICLR 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015 (cit. on p. 1).

[24] T. Laux and J. Lelmi, *Large data limit of the mbo scheme for data clustering: Γ-convergence of the thresholding energies*, 2022. arXiv: 2112.06737 [math.AP] (cit. on p. 3).

[25] T. Laux and J. Lelmi, "Large data limit of the mbo scheme for data clustering: Convergence of the dynamics," *Journal of Machine Learning Research*, vol. 24, no. 344, pp. 1–49, 2023 (cit. on p. 3).

[26] S. Luckhaus and T. Sturzenhecker, "Implicit time discretization for the mean curvature flow equation," *Calculus of Variations and Partial Differential Equations*, vol. 3, no. 2, pp. 253–271, 1995 (cit. on p. 5).

[27] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *6th International Conference on Learning Representations, ICLR 2018, Conference Track Proceedings*, 2018 (cit. on pp. 1, 2).

[28] E. Merkurjev, A. L. Bertozzi, and F. Chung, "A semi-supervised heat kernel pagerank mbo algorithm for data classification," *Communications in Mathematical Sciences*, vol. 16, no. 5, pp. 1241–1265, 2018 (cit. on p. 3).

[29] E. Merkurjev, T. Kostic, and A. L. Bertozzi, "An mbo scheme on graphs for classification and image processing," *SIAM Journal on Imaging Sciences*, vol. 6, no. 4, pp. 1903–1930, 2013 (cit. on p. 3).

[30] M. Novaga and A. Chambolle, "Implicit time discretization of the mean curvature flow with a discontinuous forcing term," *Interfaces and Free Boundaries*, vol. 10, no. 3, pp. 283–300, 2008 (cit. on p. 7).

[31] M. Pintor, F. Roli, W. Brendel, and B. Biggio, "Fast minimum-norm adversarial attacks through adaptive norm constraints," in *Advances in Neural Information Processing Systems 34: NeurIPS 2021*, M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021, pp. 20 052–20 062 (cit. on p. 5).

[32] M. S. Pydi and V. Jog, "Adversarial risk via optimal transport and optimal couplings," in *International Conference on Machine Learning*, PMLR, 2020, pp. 7814–7823 (cit. on p. 2).

[33] M. S. Pydi and V. Jog, "The many faces of adversarial risk," *Advances in Neural Information Processing Systems*, vol. 34, pp. 10 000–10 012, 2021 (cit. on p. 2).

[34] J. A. Sethian, "A fast marching level set method for monotonically advancing fronts.," *proceedings of the National Academy of Sciences*, vol. 93, no. 4, pp. 1591–1595, 1996 (cit. on p. 5).

[35] C. Szegedy *et al.*, "Intriguing properties of neural networks," in *2nd International Conference on Learning Representations, ICLR 2014, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2014 (cit. on p. 1).

[36] Y. Van Gennip, N. Guillen, B. Osting, and A. L. Bertozzi, "Mean curvature, threshold dynamics, and phase field theory on finite graphs," *Milan Journal of Mathematics*, vol. 82, pp. 3–65, 2014 (cit. on p. 3).

[37] S. R. S. Varadhan, "On the behavior of the fundamental solution of the heat equation with variable coefficients," *Communications on Pure and Applied Mathematics*, vol. 20, no. 2, pp. 431–455, 1967 (cit. on p. 5).

[38] H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. E. Ghaoui, and M. I. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*, K. Chaudhuri and R. Salakhutdinov, Eds., ser. Proceedings of Machine Learning Research, vol. 97, PMLR, 2019, pp. 7472–7482 (cit. on p. 2).