Masterarbeit

Existence of Solutions and Asymptotic Behavior of Adversarial Training with General Loss Functions

Lennart Siethoff

Würzburg, 29. Juni 2025

Julius-Maximilians-Universität Würzburg
Lehrstuhl für Mathematik des Maschinellen Lernens
Betreuer: Prof. Dr. Leon Bungert

Abstract

In this thesis, we investigate adversarial training models in the context of classification. We study the total-variation-regularized adversarial training problem, previously examined in [13], and extend it to a broader class of loss functions, establishing the existence of optimal classifiers for both binary and multiclass settings. In addition, we show that the original robust optimization problem can be reformulated in an analytically advantageous way and also admits minimizers. Furthermore, we provide a detailed proof for the Γ-convergence of the weighted non-local total variation, recovering the result in [14] under modified assumptions on the data distribution. Finally, we combine the existence and Γ-convergence findings to deduce an asymptotic result of the regularized adversarial training problem in the binary and multiclass case, proving that minimizers of this problem converge to a Bayes classifier with minimal total variation.

Contents

Chapter 1

Introduction

1.1 Motivation

In recent years research has revealed that various state-of-the-art trained machine learning models are susceptible to adversarially chosen inputs [45, 2, 38], whereby small, often imperceptible changes to the inputs of a training model can lead to incorrect outputs. A particularly striking example of this phenomenon occurs in image classification, where the introduction of certain structured noise, invisible to humans, may cause an otherwise well-performing image classification model to radically mislabel an image [30] (see Figure 1.1). Such *adversarial attacks* pose a major hurdle to deploying machine learning systems in

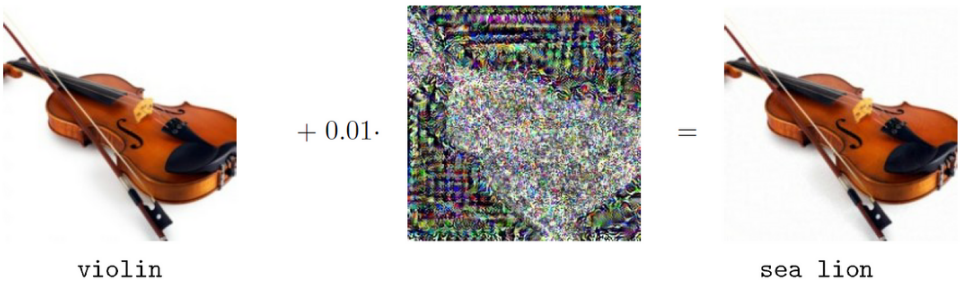

Figure 1.1: Picture taken from [12]. Adversarial attack on an image classification model.

security-critical applications [8], as it was strikingly demonstrated in practice by Tencent Keen Security Lab who were able to influence the behavior of a car's autopilot system through adversarial attacks [24]. The goal of *adversarial training* is to defend against these attacks through training methods that produce robust models, i.e., models which still perform well on adversarially perturbed data. It presents one of the most effective approaches of doing so [4], explaining the growing interest in this field [40].

In this thesis, we will focus on abstract classification problems in the product space $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is a metric space representing an "input space", and the "output space"

$\mathcal{Y}$ is a discrete set of hard labels. In practice, $\mathcal{X}$ is often a subset of $\mathbb{R}^d$, each dimension encoding some attribute of the data. The training data, consisting of correctly classified points, is given by a probability measure $\mu \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$, and in standard classification, one wishes to solve the supervised learning task of minimizing

$$\min_{u \in \mathcal{H}} \mathbb{E}_{(x,y)\sim\mu}[l(u(x), y)] \coloneqq \min_{u \in \mathcal{H}} \int_{\mathcal{X}} \int_{\mathcal{Y}} l(u(x), y) \, \mathrm{d}\mu(x, y), \tag{1.1}$$

where $\mathcal{H}$ is some *hypothesis space* containing *classifiers* $u : \mathcal{X} \to \hat{\mathcal{Y}}$, and $l : \hat{\mathcal{Y}} \times \mathcal{Y} \to [0, \infty]$ is a *loss function*. The loss function quantifies the discrepancy between the true label y of a data point x and the label $u(x)$ predicted by the hypothesis. We will work with hypothesis spaces of *soft classifiers*, meaning the functions $u \in \mathcal{H}$ attain continuous values. In the case of binary classification with the labels $\mathcal{Y} = \{0, 1\}$, we consider functions mapping to $\hat{\mathcal{Y}} = [0, 1]$, where $u(x)$ can be interpreted as a probability that the data point x gets the label 1.

In order to turn (1.1) into an adversarially robust problem, it is modified to account for adversarial attacks, which we define in the following. For $(x, y) \sim \mu$ adversarial attacks are perturbations $\tilde{x}$ of an input x inside the ε-ball $B_\varepsilon(x)$ such that $u(\tilde{x}) \neq y$, i.e., the data point x gets an incorrect label, and the loss is maximal. The parameter $\varepsilon > 0$ is called the *adversarial budget* and controls the strength of the adversarial attacks. In general, such attacks can be computed by solving

$$\sup_{\tilde{x} \in B_\varepsilon(x)} l(u(\tilde{x}, y)),$$

and training with this perturbed data, one obtains a model protected against adversarial attacks inside the ε-ball. This leads to the *(original) robust optimization problem*

$$\min_{u \in \mathcal{H}} \mathbb{E}_{(x,y)\sim\mu} \left[\sup_{\tilde{x} \in B_\varepsilon(x)} l(u(\tilde{x}), y) \right], \tag{1.2}$$

which was first proposed by Mądry et. al. in [41], and is going to be central to this thesis.

As neural networks can be greatly improved by robust models [30], problems of this type are widely studied [19]. The game theoretic properties in the context of Nash-equilibria, stemming from the interpretation of (1.2) as a two-player min-max game, have been investigated in [9, 36]. In [3, 47] the existence of robust classifiers has been shown for different reformulations of (1.2). This thesis focuses on (1.2) rewritten as a regularized optimization problem, building on the approach of Bungert et al. in [13] and [14].

1.2 Outline

In Chapter 2 we give an overview of the relevant mathematical concepts and state all the important theorems used throughout this thesis.

In Chapter 3 we show the existence of robust minimizers to different adversarial training models relating to (1.2). We begin by presenting an important result from [13]. For the *0-1-loss*

$$l(u(x), y) \coloneqq |u(x) - y|,$$

and under mild assumptions on a reference measure ν, the original robust optimization problem (1.2) is equivalent to the analytically advantageous regularized problem

$$\min_{u \in \mathcal{H}} \mathbb{E}_{(x,y)\sim\mu}[|u(x) - y|] + \varepsilon \mathrm{TV}_\varepsilon(u; \mu), \tag{1.3}$$

where the regularizer $\mathrm{TV}_\varepsilon(\,\cdot\,; \mu)$ is the *weighted non-local total variation* defined by

$$\mathrm{TV}_\varepsilon(u; \mu) \coloneqq \frac{1}{\varepsilon} \int_{\mathcal{X}} \nu\text{-}\operatorname*{ess\,sup}_{B_\varepsilon(x)} u - u(x) \,\mathrm{d}\rho_0(x) + \frac{1}{\varepsilon} \int_{\mathcal{X}} u(x) - \nu\text{-}\operatorname*{ess\,inf}_{B_\varepsilon(x)} u \,\mathrm{d}\rho_1(x). \tag{1.4}$$

Building on this, we study a generalization of (1.3), allowing for a larger class of loss functions in the data term, while retaining the regularizing effect of the total variation

$$\min_{u \in \mathcal{H}} \mathbb{E}_{(x,y)\sim\mu} \left[l(u(x), y)\right] + \varepsilon \mathrm{TV}_\varepsilon(u; \mu). \tag{1.5}$$

We work in the context of weak-* topology to show that this minimization problem admits solutions in the binary classification case. The result is easily generalized to the multiclass case, by employing the setup proposed in [28]. Subsequently, we return to the original problem (1.2) and prove that it can be rewritten with an essential supremum instead of the supremum, which allows us to show the existence of minimizers for this problem, as well.

In Chapter 4 we investigate the asymptotic behavior of the non-local total variation, as the adversarial budget ε tends to 0, following the results in [14]. To this end, we use the framework of Γ-convergence to study the associated perimeter functional, defined by

$$\mathrm{Per}_\varepsilon(A; \boldsymbol{\rho}) \coloneqq \mathrm{TV}_\varepsilon(\chi_A; \boldsymbol{\rho})$$

for $A \subset \Omega$, where Ω is an open subset of $\mathbb{R}^d$, and $\boldsymbol{\rho} \coloneqq (\rho_0, \rho_1)$ denotes the dependency on the underlying data distributions for the two classes. We replicate the Γ-convergence results

$$\mathrm{Per}_\varepsilon(\,\cdot\,; \boldsymbol{\rho}) \xrightarrow{\Gamma} \mathrm{Per}(\,\cdot\,; \boldsymbol{\rho}) \quad \text{and} \quad \mathrm{TV}_\varepsilon(\,\cdot\,; \boldsymbol{\rho}) \xrightarrow{\Gamma} \mathrm{TV}(\,\cdot\,; \boldsymbol{\rho})$$

from [14] under modified assumptions for the data distribution $\boldsymbol{\rho}$, where the local versions of the perimeter and the total variation are intrinsically related to the theory of functions of bounded variation. In this setting, the non-local perimeter is closely connected to the Minkwoski content of sets [17], and has been thoroughly studied by Cesaroni et al. in [16, 15].

In Chapter 5 we combine the results from the previous two chapters, leading to insight into

the asymptotic behavior of (1.5) and the convergence of its minimizers, as the adversarial budget tends to 0. In particular, we show that a sequence of minimizers of (1.5) converges, up to a subsequence, as $\varepsilon \to 0$, to a Bayes classifier with minimal total variation, i.e., one recovers minimizers of the standard classification problem (1.1) with some additional regularity. As our last result, we generalize these findings for binary classification to the multiclass case. Further analysis of the asymptotics of adversarial training, especially in the context of l_∞-perturbations, can be found in [46, 48].
Finally, in Chapter 6, we propose further questions and present possible ideas on how one might generalize our existence and asymptotic behavior results.

Chapter 2

Mathematical Preliminaries

2.1 Notation and Definitions

We establish important notation and recall some basic definitions.
Dual space For a normed vector space X, let X^* be the *(topological) dual* of X, i.e., the space of all continuous linear functionals on X.
Sets Throughout this thesis, we denote by

$$\chi_A(x) \coloneqq \begin{cases} 1, & x \in A \\ 0, & \text{else} \end{cases}$$

the *characteristic function* of a set A, and we say a sequence of sets $(A_k)_{k\in\mathbb{N}}$ converge to A in the topology T, if $\chi_{A_k} \to \chi_A$ in the topology T. On a general metric space $(X; d)$ the open ε-ball is defined as $B_\varepsilon(x) \coloneqq \{y \in X : d(x, y) < \varepsilon\}$, while for $X \subset \mathbb{R}^d$ we define $B_\varepsilon(x) \coloneqq \{y \in \mathbb{R}^d : |x-y| < \varepsilon\}$ for the euclidean distance $|\cdot|$ on $\mathbb{R}^d$. For a subset $A \subset \mathbb{R}^d$, the *distance function* to A is defined as $d(x, A) \coloneqq \inf\{|x - y| : y \in A\}$ and the *diameter* of A is defined as $\operatorname{diam}(A) \coloneqq \sup\{|x - y| : x, y \in A\}$.
If A is open and bounded, we say A has C^k *boundary* if for each point $\tilde{x} \in \partial A$ there exists a $r > 0$ and a C^k function $\gamma : \mathbb{R}^{d-1} \to \mathbb{R}$ such that—upon relabeling and reorienting the coordinate axes if necessary—we have

$$A \cap B_r(\tilde{x}) = \{x \in B_r(\tilde{x}) : x_d > \gamma(x_1, \ldots, x_{d-1})\}.$$

Likewise, we say A has *smooth boundary* if ∂A is C^k for $k = 1, 2, \ldots$, and ∂A is analytic if the mapping γ is analytic. Furthermore, we say A has *Lipschitz boundary* if for every $x \in \partial A$ there exists a neighborhood $U(x)$ whose intersection with ∂A is the graph of a Lipschitz continuous function.
Essential supremum/infimum The *essential supremum* and *essential infimum* with

respect to the measure μ are defined by

$$\mu\text{-}\operatorname*{ess\,sup}_{x\in A} u(x) \coloneqq \inf_{\substack{N\subset A,\\ \mu(N)=0}} \sup_{x\in A\setminus N} u(x) \quad \text{and} \quad \mu\text{-}\operatorname*{ess\,inf}_{x\in A} u(x) \coloneqq \sup_{\substack{N\subset A,\\ \mu(N)=0}} \inf_{x\in A\setminus N} u(x).$$

For the d-dimensional Lebesgue measure we don not specify the measure and write

$$\operatorname{ess\,sup} \coloneqq \mathcal{L}^d\text{-}\operatorname{ess\,sup} \quad \text{and} \quad \operatorname{ess\,inf} \coloneqq \mathcal{L}^d\text{-}\operatorname{ess\,inf}.$$

Mollifiers For

$$\phi : \mathbb{R}^d \to \mathbb{R},\ \phi(x) \coloneqq \begin{cases} c\exp\left(\frac{1}{|x|^2-1}\right), & |x|<1 \\ 0, & |x|\geq 1 \end{cases}$$

with the normalization constant c such that $\int_{\mathbb{R}^d}\phi(x)\,\mathrm{d}x = 1$, we denote by

$$\phi_\varepsilon(x) \coloneqq \frac{1}{\varepsilon^n}\phi\left(\frac{x}{\varepsilon}\right)$$

the *standard mollifier*.

Push-forward measure Let $(X_1,\Sigma_1),(X_2,\Sigma_2)$ be two measurable spaces and let $f : X_1 \to X_2$ be measurable. Given a measure μ on X_1 we define the *push-forward measure* on X_2 by the formula

$$f_\sharp\mu \coloneqq \mu(\{x\in X_1 : f(x)\in B\}),$$

where B is an arbitrary set in Σ_2.

Sequences Since most of the results in this thesis are proved up to subsequences, they will often not be relabeled, for visual clarity. This will be pointed out in important steps.

2.2 Preliminaries

2.2.1 Lebesgue Spaces

We introduce Lebesgue spaces using standard notation and basic terminology, and give a concise overview of the relevant results. The reader is directed to [43] for a more thorough exposition.

Let $(X,\mathcal{A},\mu)$ be a measure space. For $1\leq p\leq\infty$ denote by $L^p(X;\mu)$ the classical *Lebesgue space*, consisting of measurable functions $u : X\to\mathbb{R}$ which satisfy $\|u\|_{L^p}<\infty$ for the *L^p-norm*

$$\|u\|_{L^p} \coloneqq \begin{cases} \left(\int_X |u(x)|^p\,\mathrm{d}\mu(x)\right)^{1/p}, & 1\leq p<\infty \\ \operatorname*{ess\,sup}_{x\in X}|u(x)|, & p=\infty. \end{cases} \tag{2.1}$$

As usual, we identify functions that coincide μ-almost everywhere. For $1 \leq p \leq \infty$ the space $(L^p(X;\mu), \|\cdot\|_{L^p})$ is a Banach space and we write $u_k \xrightarrow{L^p(X;\mu)} u$ for a sequence $(u_k)_{k\in\mathbb{N}}$ converging in the (strong) topology induced by the norm. In the case of $X \subset \mathbb{R}^d$ and $\mu = \mathcal{L}^d$, we write $L^p(X;\mu) = L^p(X)$.

Theorem 2.1 (Dual of L^p/Riesz representation theorem)**.** *For $1 < p < \infty$ and q such that $1/p + 1/q = 1$ it holds that*

$$(L^p(X;\mu))^* \cong L^q(X;\mu).$$

If $(X, \mathcal{A}, \mu)$ is a σ-finite measure space, the isomorphism above also holds for $p = 1$ and $q = \infty$. In both cases, for $\phi \in (L^p(X;\mu))^*$ and $v_\phi \in L^q(X;\mu)$ the isomorphism is characterized by

$$\langle \phi, u \rangle = \int_X v_\phi u \,\mathrm{d}\mu \quad \forall u \in L^p(X;\mu).$$

In the special case of $p = 2$, the space $L^2(X;\mu)$ is self-dual and becomes a Hilbert space with the inner product

$$\langle u, v \rangle := \int_X uv \,\mathrm{d}\mu \quad \forall u, v \in L^2(X;\mu).$$

The following fundamental convergence result for L^p-spaces will be especially useful in this thesis for proving compactness properties.

Theorem 2.2. *Let $(u_k)_{k\in\mathbb{N}}$ be a sequence in $L^p(X;\mu)$ with $1 \leq p \leq \infty$ and let $u \in L^p$ be such that $u_k \xrightarrow{L^p(X;\mu)} u$. Then, there exists a subsequence $(u_{k_l})_{l\in\mathbb{N}}$ such that $u_{k_l}(x) \to u(x)$ for almost every $x \in X$.*

Proof. See, e.g., [11, Theorem 4.9]. □

Furthermore, we give a fundamental separability result for Lebesgue spaces, relevant for the characterization of compactness.

Theorem 2.3. *Let $(X, \mathcal{A}, \mu)$ be a separable measure space. Then $L^p(X;\mu)$ is separable for any $1 \leq p < \infty$.*

Proof. See, e.g., [11, Theorem 4.13]. □

We state *Fatou's lemma* on general measure spaces, which will prove essential in showing lower semicontinuity properties.

Theorem 2.4 (Fatou's lemma)**.** *Let $(X, \mathcal{A}, \mu)$ be a measure space and $(u_k)_{k\in\mathbb{N}}$ a sequence of non-negative measurable functions on X. Then*

$$\int_X \liminf_{k\to\infty} u_k \,\mathrm{d}\mu \leq \liminf_{k\to\infty} \int_X u_k \,\mathrm{d}\mu.$$

As an easy corollary, one has the *reverse Fatou lemma* for the limes superior.

Corollary 2.5 (Reverse Fatou lemma)**.** *Let $(X, \mathcal{A}, \mu)$ be a measure space and $(u_k)_{k\in\mathbb{N}}$ a sequence of measurable functions on X. If there exists a non-negative integrable function v on X such that $u_k \leq v$ for all $k \in \mathbb{N}$, then*

$$\limsup_{k\to\infty} \int_X u_k \,\mathrm{d}\mu \leq \int_X \limsup_{k\to\infty} u_k \,\mathrm{d}\mu.$$

Proof. Since $v - u_k$ is non-negative, we apply Fatou's lemma to see

$$\begin{aligned}\int_X v \,\mathrm{d}\mu - \limsup_{k\to\infty} \int_X u_k \,\mathrm{d}\mu &= \liminf_{k\to\infty} \left(\int_X v \,\mathrm{d}\mu - \int_X u_k \,\mathrm{d}\mu \right) = \liminf_{k\to\infty} \int_X v - u_k \,\mathrm{d}\mu \\ &\geq \int_X \liminf_{k\to\infty} v - u_k \,\mathrm{d}\mu = \int_X v \,\mathrm{d}\mu - \int_X \limsup_{k\to\infty} u_k \,\mathrm{d}\mu.\end{aligned}$$

Canceling $\int_X v \,\mathrm{d}\mu$ and multiplying by -1 proves the corollary. □

Since, in Section 4.1 we wish to reformulate the non-local perimeter in terms of sets, we introduce the notion of *Lebesgue density points* and state the main result concerning the density of sets.

Definition 2.6 (Lebesgue density points)**.** For $t \in [0, 1]$ the points where a measurable set $A \subset \mathbb{R}^d$ has (Lebesgue) density t are defined as

$$A^t := \left\{ x \in \mathbb{R}^d : \lim_{r \searrow 0} \frac{\mathcal{L}^d(A \cap B_r(x))}{\mathcal{L}^d(B_r(x))} = t \right\}.$$

Theorem 2.7 (Lebesgue density theorem)**.** *Let $A \subset \mathbb{R}^d$ be a measurable set. Then the limit*

$$\lim_{r \searrow 0} \frac{\mathcal{L}^d(A \cap B_r(x))}{\mathcal{L}^d(B_r(x))}$$

exists and equals 1 for $\mathcal{L}^d$-almost every $x \in A$ and equals 0 for $\mathcal{L}^d$-almost every $x \in \mathbb{R}^d \setminus A$.

Proof. See, e.g., [35, Corollary 2.14]. □

In particular this theorem implies that $\chi_{A^1}(x) = \chi_A(x)$ for $\mathcal{L}^d$-almost every $x \in \mathbb{R}^d$. We give a simple example, illustrating the notion of density points.

Example 2.8.

- Let $A := \{(x, y) \in \mathbb{R}^2 : 0 < x, y < 1\}$ be a open square with edges parallel to the coordinate axes. Then one can easy see that every $x \in \partial A$ is a point of density $1/2$, except for the vertices, where the density is $1/4$. Every point in A has density 1 and every point in $\mathbb{R}^d \setminus \overline{A}$ has density 0. Hence, it holds that $A^1 = A$.
- For sets with 'rougher' boundary, for example, $B := A \cup \{(x, 0) : 0 < x < 2\}$, taking the points of density one, we again have $B^1 = A$, meaning that the distance to B^1 ignores subsets of B with measure 0. This notion will be formalized in Lemma 4.2.

2.2.2 Weak and Weak-* Topology

In the existence proofs in Chapter 3 we will not work with the restrictive strong topology on L^p-spaces. Instead we will exploit properties of the *weak* and *weak-* topology*, which we define here through its convergent sequences. For a detailed introduction to weak topologies we refer to [11].

Definition 2.9 (Weak and weak-* convergence)**.** Let X be a Banach space of $\mathbb{R}$ with dual X^*. We say a sequence $(x_k)_{k\in\mathbb{N}}$ in X *converges weakly* to $x \in X$ and write $x_k \rightharpoonup x$ if

$$x^*(x_k) \to x^*(x), \quad \forall x^* \in X^*.$$

Furthermore, we call a sequence $(y_k)_{k\in\mathbb{N}}$ in X^* *weak-* convergent* to $y \in X^*$ and write $y_k \rightharpoonup^* y$ if

$$y_k(x) \to y(x) \quad \forall x \in X.$$

Remark 2.10 (Weak and weak-* convergence in L^p-spaces)**.** With the Riesz representation theorem one can easily characterize weak and weak-* convergence in L^p-spaces. Let $(X, \mathcal{A}, \mu)$ be a measure space and let $(u_k)_{k\in\mathbb{N}} \subset L^p(X;\mu)$. Since $L^p(X;\mu)$ is reflexive for $1 < p < \infty$, the characterization for weak and weak-* convergence coincide and one has $u_k \rightharpoonup u$ and $u_k \rightharpoonup^* u$ if

$$\lim_{k\to\infty} \int_X u_k \phi \,\mathrm{d}\mu = \int_X u\phi \,\mathrm{d}\mu \quad \forall \phi \in L^q(X;\mu),$$

for q such that $1/p + 1/q = 1$. Furthermore, if μ is a σ-finite measure we have $u_k \rightharpoonup^* u$ for $(u_k)_{k\in\mathbb{N}} \subset L^\infty(X;\mu)$ if

$$\lim_{k\to\infty} \int_X u_k \phi \,\mathrm{d}\mu = \int_X u\phi \,\mathrm{d}\mu \quad \forall \phi \in L^1(X;\mu).$$

The following fundamental compactness result in the weak-* topology is our primary motivation for working with this topology.

Theorem 2.11 (Banach–Alaoglu)**.** *Let X be a Banach space of $\mathbb{R}$ with dual X^*. Then the closed unit ball of X^**

$$B_{X^*} := \{x \in X^* : \|x\|_{X^*} \leq 1\}$$

is weak- compact.*

Proof. See, e.g., [11, Theorem 3.16]. □

By a scaling argument it follows directly from Banach–Alaoglu's theorem that any bounded subset of L^∞-spaces is weak-* precompact, which will be essential to prove the

existence of solutions to optimization problems with the direct method. Since this strategy also requires lower semicontinuity of the functionals we wish to optimize, we define the weak versions and state a helpful characterization of weak lower semicontinuity in Hilbert spaces.

Definition 2.12 (Weak and weak-* lower semicontinuity)**.** Let X be a Banach space of $\mathbb{R}$. We say a function $f : X \to [-\infty, \infty]$ is *weakly lower semicontinuous* (*weak-* lower semicontinuous*) in x if, for every sequence $(x_k)_{k\in\mathbb{N}}$ in X with $x_k \rightharpoonup x$ ($x_k \rightharpoonup^* x$) it holds that

$$f(x) \leq \liminf_{k\to\infty} f(x_k).$$

For convex functions on Hilbert spaces weak and strong lower semicontinuity coincide. This holds, since the epigraph of a convex lower semicontinuous function is closed and convex, which implies that it is weakly closed. For a detailed proof see [6, Theorem 9.1].

Theorem 2.13. *Let H be a Hilbert space and let $f : H \to]-\infty, \infty]$ be a convex function. Then weak lower semicontinuity of f in $x \in H$ is equivalent to lower semicontinuity in $x \in H$ with respect to the strong topology on H.*

2.2.3 Γ-convergence

We present the basic definitions and important results in the theory of Γ-convergence. This notion of convergence for functionals was first introduced by Ennio De Giorgi in a series of papers [29, 22, 23], and has since been established as primary tool used in asymptotic analysis of variational problems [10]. In Chapter 4, we will study the Γ-convergence of the non-local total variation and perimeter, leading to insight into the asymptotic behavior of the related adversarial training problems.

We begin by giving the sequential definition of Γ-convergence, well-suited for our analysis, and refer to [10] for a comprehensive introduction to this topic.

Definition 2.14 (Γ-convergence)**.** Let X be a metric space and let $(J_k)_{k\in\mathbb{N}}$ be a sequence of functionals $J_k : X \to [0, \infty]$. Then we say J_k Γ-*converges* to the Γ-limit J and write $J_k \xrightarrow{\Gamma} J$ if the following two conditions hold:

(i) **(liminf-inequality)** For every convergent sequence $(x_k)_{k\in\mathbb{N}} \subset X$ with limit $x \in X$ we have

$$J(x) \leq \liminf_{k\to\infty} J_k(x_k). \tag{2.2}$$

(ii) **(limsup-inequality)** For every $x \in X$ there exists a *recovery sequence* $(x_k)_{k\in\mathbb{N}} \subset X$ converging to x such that

$$\limsup_{k\to\infty} J_k(x_k) \leq J(x). \tag{2.3}$$

The Γ-convergence of a sequence $J_k \xrightarrow{\Gamma} J$ is particularly useful in the context of minimizing problems, as it gives insight into the optimization problem in the limit

$$\inf_X J. \tag{2.4}$$

To see this we consider a *minimizing sequence* $(x_k)_{k\in\mathbb{N}}$, i.e., $x_k \in \arg\min J_k$ for all $k \in \mathbb{N}$. If we assume such a sequence to be convergent to some $x \in X$ then it is easy to see that x is a solution to (2.4) and the minimal values of J_k converge to $\min_X J$:
For an arbitrary $y \in X$ with the recovery sequence $(y_k)_{k\in\mathbb{N}}$ we have

$$J(x) \overset{(2.2)}{\leq} \liminf_{k\to\infty} J_k(x_k) \leq \limsup_{k\to\infty} J_k(y_k) \overset{(2.3)}{\leq} J(y) \tag{2.5}$$

implying that x is a minimizer of J and since there exists a recovery sequence $(\overline{x}_k)_{k\in\mathbb{N}}$ for x, we combine

$$J(x) \geq \limsup_{k\to\infty} J_k(\overline{x}_k) \geq \limsup_{k\to\infty} \min_X J_k \tag{2.6}$$

with the first inequality in (2.5) to see that the minimal values indeed converge.

However, the assumption that a minimizing sequence converges is often overly restrictive and therefore, we introduce the weaker condition of equi-coerciveness under which the fundamental theorem of Γ-convergence (Theorem 2.16) holds.

Definition 2.15 (Equi-coerciveness)**.** We say a sequence $J_k : X \to [0,\infty]$ is *equi-coercive* if for all $t \in [0,\infty)$ there exists a compact set K_t such that $\{x \in X : J_k(x) < t\} \subset K_t$ for all $k \in \mathbb{N}$.

Let $(J_k)_{k\in\mathbb{N}}$ be an equi-coercive sequence with Γ-limit J and a minimizing sequence $(x_k)_{k\in\mathbb{N}}$. If $\lim_{k\to\infty} \inf_X J_k < \infty$ we have $(x_k)_{k\in\mathbb{N}} \subset \{J_k < t\}$ for some $t \in [0,\infty)$ and by equi-coerciveness there exists a subsequence $(x_{k_j})_{j\in\mathbb{N}} \to x \in X$. Then, by the same argument as before

$$x \in \arg\min J \quad \text{and} \quad \min_X J = \lim_{k\to\infty} \min_X J_k,$$

proving the following theorem:

Theorem 2.16 (Fundamental theorem of Γ-convergence)**.** *Let X be a metric space and let $(J_k)_{k\in\mathbb{N}}$ be an equi-coercive sequence of functionals on X with the Γ-limit J. Then every cluster point of a minimizing sequence $(x_k)_{k\in\mathbb{N}}$ is a minimum point for J and the minimal values of J_k converge to $\min_X J$.*

It is possible to weaken the assumptions under which one obtains existence of minimizers of the Γ-limit further [10]. For example, for the veracity of the previous theorem, the infimal values of J_k do not need to be attained and the statement still holds

if $\lim_{k\to\infty} J_k(x_k) = \lim_{k\to\infty} \inf_X J_k$, but since we will show existence of minimizers in Chapter 3, the previous theorem will suffice for our applications.

A useful fact for proving the limsup-inequality for a functional J_k is the density argument [10, Remark 1.29], which states that it suffices to show that the limsup-inequality holds on a dense subset of X if the Γ-limit satisfies a certain continuity property. Once again, we show a slightly weaker result; in particular, that the limsup-inequality follows if it holds for a constant recovery sequence for all x in the dense subset.

Remark 2.17 (Density argument)**.** Let X be a metric space, let $(J_k)_{k\in\mathbb{N}}$ be a sequence of functionals $J_k : X \to [0,\infty]$ and let D be a dense subset of X. Suppose we have

(i) $\limsup_{k\to\infty} J_k(x) \leq J(x)$ for all $x \in D$ and

(ii) $\lim_{k\to\infty} J(x_k) = J(x)$ for all $x \in X$ and $(x_k)_{k\in\mathbb{N}} \subset D$ such that $x_k \to x$.

Then the limsup-inequality already holds for all $x \in X$.

To see this, consider for $x \in X$ a sequence $(x_k)_{k\in\mathbb{N}} \subset D$ with $x_k \to x$. Then we have

$$\begin{aligned} J(x) &\overset{\text{(ii)}}{=} \lim_{k\to\infty} J(x_k) \\ &= \limsup_{k\to\infty} J(x_k) \\ &\overset{\text{(i)}}{=} \limsup_{k\to\infty} \limsup_{l\to\infty} J_l(x_k) \\ &\geq \limsup_{l\to\infty} J_l(x_k) - \delta \quad \forall k \geq K(\delta) \text{ for some } K(\delta) \in \mathbb{N}, \delta > 0. \end{aligned}$$

Therefore, the last inequality still holds when we send $k \to \infty$, and the different indices can be synchronized:

$$J(x) \geq \limsup_{l\to\infty} J_l(x_l) - \delta$$

Sending $\delta \to 0$ yields the desired result.

2.2.4 Functions of Bounded Variation

In this section we give a short introduction to the theory of functions of bounded variation and the space BV—referring to [1, 25] for a thorough treatment of these topics—-which will prove integral to the investigation of Γ-convergence of the non-local total variation and perimeter to their local counterparts. Throughout this section, Ω denotes an open subset of $\mathbb{R}^d$.

We begin by motivating the notion of a weak or distributional derivative: For continuously differentiable functions $f, g : \overline{\Omega} \to \mathbb{R}$ for a regular domain $\Omega \subset \mathbb{R}^d$ one has the

integration by parts formula in higher dimensions (see, e.g., [42, Theorem 37.2]) which, for a test function g satisfying $g(x) = 0$ for all $x \in \partial\Omega$, reduces to

$$\int_\Omega g \cdot \nabla f \, \mathrm{d}x = -\int_\Omega \nabla g \cdot f \, \mathrm{d}x.$$

In view of this equality one defines:

Definition 2.18. Let $u \in L^1(\Omega)$. We say the finite signed Radon measure Du_i is the *weak* or *distributional derivative* of u with respect to x_i if

$$\int_\Omega u \frac{\partial \phi}{\partial x_i} \, \mathrm{d}x = -\int_\Omega \phi \, \mathrm{d}D_i u \quad \forall \phi \in C_c^\infty(\Omega). \tag{2.7}$$

Definition 2.19 (The space BV)**.** We define the *space of functions of bounded variation* $BV(\Omega)$ as the space of all functions $u \in L^1(\Omega)$ for which the first-order partial distributional derivatives exist as finite signed Radon measures for all $i = 1, \dots, d$.

For $u \in BV(\Omega)$ we set $Du \coloneqq (Du_1, \dots, Du_d)$, and define the *total variation measure* of Du as

$$|Du|(A) \coloneqq \sup \left\{ \sum_{k=1}^{\infty} |Du(A_k)|, \{A_k\} \subset \mathcal{B}(\Omega) \text{ is a partition of } A \right\}.$$

We further provide the definition of the total variation of an L^1-function, which is intrinsically tied to the asymptotic behavior of the regularized optimization problem (1.5).

Definition 2.20 (Total variation)**.** For $u \in L^1(\Omega)$ the *total variation* of u in Ω is defined by

$$\mathrm{TV}(u; \Omega) \coloneqq \sup \left\{ \int_\Omega u \, \mathrm{div}\phi \, \mathrm{d}x : \phi \in C_c^1(\Omega; \mathbb{R}^d), |\phi| \leq 1 \right\}, \tag{2.8}$$

where $C_c^1(\Omega; \mathbb{R}^d)$ is the space of continuously differentiable functions $\phi : \Omega \to \mathbb{R}^d$ with compact support.

This definition is motivated by the fact that for $u \in BV(\Omega)$ one has $|Du|(\Omega) < \infty$ and $\mathrm{TV}(u; \Omega) = |Du|(\Omega)$. Therefore, the space of functions of bounded variation can also be characterized as

$$BV(\Omega) = \{u \in L^1(\Omega) : \mathrm{TV}(u; \Omega) < \infty\}.$$

Furthermore, this characterization motivates us to define the total variation on any Borel set $\Omega' \subset \Omega$ as

$$\mathrm{TV}(u; \Omega') \coloneqq |Du|(\Omega').$$

Remark 2.21 (Properties of the total variation)**.**

- The total variation is a lower semicontinuous functional on $L^1(\Omega)$, i.e., for a sequence $(u_k)_{k \in \mathbb{N}} \subset L^1(\Omega)$ with $u_k \xrightarrow{L^1(\Omega)} u$ it holds

$$\mathrm{TV}(u; \Omega) \leq \liminf_{k \to \infty} \mathrm{TV}(u_k; \Omega).$$

- For continuously differentiable functions u it holds

$$\mathrm{TV}(u;\Omega) = \int_\Omega \mathrm{d}|Du|(x) = \int_\Omega |\nabla u(x)|\,\mathrm{d}x.$$

Remark 2.22. To motivate our choice of working in the space $BV(\Omega)$, we point out a key difference between the Sobolev space $W^{1,1}(\Omega)$ (see, e.g., [33] for a definition) and $BV(\Omega)$, relevant to our results:
The space $W^{1,1}(\Omega)$ only contains functions whose first order weak derivatives are L^1-functions, leading to the fact that the limits of sequences which are uniformly bounded in the $W^{1,1}$-norm, do not necessarily lie in $W^{1,1}(\Omega)$. Consider the following example, relating to the results in Chapter 4: Let $(u_k)_{k\in\mathbb{N}} \in W^{1,1}([-1,1])$,

$$u_k(x) \coloneqq \begin{cases} 0, & x \le 0 \\ kx, & 0 < x < \frac{1}{k} \\ 1, & \frac{1}{k} \le x \end{cases}$$

be a sequence of functions converging to the *Heaviside step function*

$$H(x) \coloneqq \begin{cases} 0, & x \le 0 \\ 1, & x > 0, \end{cases}$$

depicted in Figure 2.1.

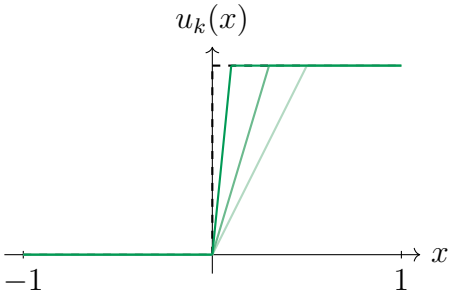

Figure 2.1: The function u_k for $k = 2, 5, 10$ and the Heaviside step function (dashed).

The limit is not contained in $W^{1,1}(\Omega)$, but since its distributional derivative is given by the delta distribution with mass at 0, we have $H \in BV(\Omega)$.

In general, in $BV(\Omega)$ one has the compactness result [1, Theorem 3.23], while the Sobolev space has no similar property. We state this result for an open set Ω with Lipschitz boundary.

Theorem 2.23 (Compactness in BV)**.** *Let $\Omega \subset \mathbb{R}^d$ be an open set with Lipschitz boundary. Then every sequence $(u_k)_{k\in\mathbb{N}} \subset BV(\Omega)$ satisfying*

$$\limsup_{k\to\infty} \left(\int_\Omega |u_k| \, \mathrm{d}x + \mathrm{TV}(u_k; \Omega) \right) < \infty$$

admits a subsequence $(u_{k_l})_{l\in\mathbb{N}}$ converging in $L^1(\Omega)$ to $u \in BV(\Omega)$ and $(Du_{k_l})_{l\in\mathbb{N}}$ weakly- converges to Du.*

Proof. See, e.g., [1, Theorem 3.23]. □

Closely linked to the study of BV-spaces, is the theory of sets of finite perimeter.

Definition 2.24 (Sets of finite perimeter)**.** For an $\mathcal{L}^d$-measurable subset $A \subset \mathbb{R}^d$, we define its *perimeter* in Ω as

$$\mathrm{Per}(A; \Omega) \coloneqq \mathrm{TV}(\chi_A; \Omega)$$

and say A is a *set of finite perimeter* (in Ω) if

$$\chi_A \in BV(\Omega).$$

The following theorem states that the total variation of any L^1-function can be expressed through the perimeters of its level sets.

Theorem 2.25 (Coarea formula in BV)**.** *For any open set $\Omega \subset \mathbb{R}^d$ and $u \in L^1(\Omega)$ one has*

$$\mathrm{TV}(u; \Omega) = \int_{-\infty}^{\infty} \mathrm{Per}(\{u(x) > t\}; \Omega) \, \mathrm{d}t. \tag{2.9}$$

In particular, if $u \in BV(\Omega)$ the set $\{u > t\}$ has finite perimeter in Ω for $\mathcal{L}^1$-almost every $t \in \mathbb{R}$ and

$$|Du|(B) = \int_{-\infty}^{\infty} |D\chi_{\{u>t\}}|(B) \, \mathrm{d}t \tag{2.10}$$

for any Borel set $B \subset \Omega$.

Proof. See, e.g., [1, Theorem 3.40]. □

Furthermore, by [37, Remark 4.3] the coarea formula also holds for the *weighted total variation* and the *weighted perimeter* given by

$$\mathrm{TV}(u; \rho, \Omega) \coloneqq \int_\Omega \rho(x) \, \mathrm{d}|Du|(x), \quad \mathrm{Per}(u; \rho, \Omega) \coloneqq \mathrm{TV}(\chi_A; \rho, \Omega) \tag{2.11}$$

for a measurable function $\rho : \Omega \to \mathbb{R}$.

Next we provide the definitions of the $(d-1)$-dimensional Hausdorff measure and the reduced boundary, as we wish to relate Definition 2.24 to the classical notion of perimeter, given by the $(d-1)$-dimensional Hausdorff measure of the boundary.

Definition 2.26 (Hausdorff measure)**.** Let $A \subset \mathbb{R}^d$ and $0 \leq s < \infty$. We define the *s-dimensional Hausdorff measure* of the set A as

$$\mathcal{H}^s(A) \coloneqq \liminf_{\delta \to 0} \left\{ \sum_{i=0}^{\infty} \alpha(s) \left(\frac{\operatorname{diam}(C_i)}{2} \right)^s : A \subset \bigcup_{i=0}^{\infty} C_i, \operatorname{diam}(C_i) < \delta \right\},$$

where

$$\alpha(s) \coloneqq \frac{\pi^{\frac{s}{2}}}{\Gamma(\frac{s}{2}+1)}$$

for the gamma function $\Gamma(s) \coloneqq \int_0^\infty e^{-x} x^{s-1} \, \mathrm{d}x$.

Definition 2.27 (Reduced boundary)**.** Let A be an $\mathcal{L}^d$-measurable set of finite perimeter. We call the *reduced boundary* $\partial^* A$ the collection of all points $x \in \mathbb{R}^d$ such that the limit

$$\nu_A(x) \coloneqq \lim_{r \searrow 0} \frac{D\chi_A(B_r(x))}{\mathcal{L}^d(D\chi_A(B_r(x)))}$$

exists in $\mathbb{R}^d$ and satisfies $|\nu_A(x)| = 1$.

Proofs for the following example and remarks about the reduced boundary can be found in [34].

Remark 2.28.

- As a simple example we consider a square $A \subset \mathbb{R}^2$ with edges parallel to the coordinate axes. Then, $\nu_A(x)$ exists for every $x \in \partial A$. However, $|\nu_A(x)| = 1$ if and only if x is not a vertex of A. Thus, the reduced boundary $\partial^* A$ is equal to the classical boundary ∂A minus the four vertices.
- If $A \subset \mathbb{R}^d$ is an open set with C^1 boundary, then $\partial^* A$ coincides with ∂A.
- Let $A, B \subset \mathbb{R}^d$. If $\chi_A = \chi_B$ for $\mathcal{L}^d$-almost every $x \in \mathbb{R}^d$, then $\partial^* A = \partial^* B$. In particular, we have $\partial^* A = \partial^* A^1$.

The central result connecting the total variation of a characteristic function the Hausdorff measure, is the structure theorem for sets of finite perimeter, a proof of which can be found in [25, Theorem 5.15].

Theorem 2.29 (Structure theorem for sets of finite perimeter)**.** *Let $A \subset \mathbb{R}^d$ be an $\mathcal{L}^d$-measurable set. Then the total variation measure $|D\chi_A|$ coincides with the $(d-1)$-dimensional Hausdorff measure restricted to the reduced boundary of A, i.e.,*

$$|D\chi_A|(\,\cdot\,) = \mathcal{H}^{d-1}(\,\cdot \cap \partial^* A).$$

In particular, this theorem implies that for a set $A \subset \mathbb{R}^d$ with C^1 boundary, the perimeter coincides with the $(d-1)$-dimensional Hausdorff measure of ∂A

$$\operatorname{Per}(A; \mathbb{R}^d) = \mathcal{H}^{d-1}(\partial A).$$

Having built the necessary mathematical framework, we now turn to the formulation and analysis of existence problems for different adversarial training models.

Chapter 3

Existence of Minimizers for Adversarial Training Models

In this chapter we will prove the existence of minimizers for different adversarial training models for a fixed adversarial budget $\varepsilon > 0$.

3.1 Regularized Problem with General Loss Function

We study the total variation regularized problem (1.5) and begin by precisely stating the minimizing problem and the assumption under which we are going to prove existence of solutions.

Let $\mathcal{X}$ be a separable metric space and $\mathfrak{B}(\mathcal{X})$ its associated Borel σ-algebra. The distribution of training pairs is given by a probability measure $\mu \in \mathcal{P}(\mathcal{X} \times \{0,1\})$ and we further define the data distribution $\rho \in \mathcal{P}(\mathcal{X})$ as the push-forward measure $\rho \coloneqq \pi_{1\sharp}\mu$, where $\pi_1 : \mathcal{X} \times \{0,1\} \to \mathcal{X}$, $(x,y) \mapsto x$ is the projection on the first coordinate. The data distribution can also be split into the conditional distributions $\rho = \rho_0 + \rho_1$, where $\rho_i(\,\cdot\,) \coloneqq \mu(\,\cdot\, \times \{i\})$ for $i \in \{0,1\}$. Note that ρ is a probability measure on $\mathcal{X}$ since

$$\rho(\mathcal{X}) = \rho_0(\mathcal{X}) + \rho_1(\mathcal{X}) = \mu(\mathcal{X} \times \{0\}) + \mu(\mathcal{X} \times \{1\}) = \mu(\mathcal{X} \times \{0,1\}) = 1.$$

We further assume the existence of a reference measure ν, the significance of which will be explained shortly.

Assumption 3.1. There exists a σ-finite measure $\nu \in \mathcal{M}(\mathcal{X})$ such that

(i) $\rho \ll \nu$;

(ii) $\{x \in \mathcal{X} : d(x, \operatorname{supp}\rho) < \varepsilon\} \subset \operatorname{supp}\nu$;

(iii) $\nu(B_\varepsilon(x)) < \infty$ for ρ-almost every $x \in \mathcal{X}$.

The goal of this chapter will be to prove the existence of solutions result for the total variation regularized adversarial training problem with a general loss-function in the data term in the hypothesis space of soft classifiers

$$\mathcal{H} := \{u \in L^\infty(\mathcal{X}; \nu) : 0 \leq u \leq 1\}.$$

Theorem 3.2. *Let $l : [0,1] \times \mathcal{Y} \to [0,\infty]$ be a loss function which is continuous and convex in the first variable. Then, under Assumption 3.1*

$$\min_{u \in \mathcal{H}} \mathbb{E}_{(x,y)\sim\mu} \left[l(u(x), y) \right] + \varepsilon \mathrm{TV}_\varepsilon(u; \mu)$$

admits a solution.

The investigation of this problem is motivated by previous results, which we will restate here.

3.1.1 Motivation

Following the argumentation by Bungert et al. in [13], we will see that (1.3) is a natural generalization of the robust optimization problem (1.2) for the 0-1-loss

$$\min_{u \in \widetilde{\mathcal{H}}} \mathbb{E}_{(x,y)\sim\mu} \left[\sup_{\tilde{x} \in B_\varepsilon(x)} |u(\tilde{x}) - y| \right], \tag{3.1}$$

where $\widetilde{\mathcal{H}} := \{u : \mathcal{X} \to [0,1] \text{ measurable}\}$.

Remark 3.3. In [13] the proofs for this result and its supporting lemma Lemma 3.5 were provided under slightly stricter assumptions for the reference measure ν than in Assumption 3.1. Instead of Assumption 3.1(iii), ν was assumed to be locally doubling, allowing the Lebesgue differentiation theorem to be used. It was since found out that these statements continue to hold under our weaker assumptions, retaining the equivalence of the minimization problems in our setting [12]. The same will apply to the proof of Lemma 3.14 in Section 3.3.

The following lemma, which is easily verified by straightforward computations states that one can rewrite the functional in (3.1) as the sum of a data term and a total variation type regularizer.

Lemma 3.4. *For any $u \in \widetilde{\mathcal{H}}$ it holds*

$$\mathbb{E}_{(x,y)\sim\mu} \left[\sup_{\tilde{x} \in B_\varepsilon(x)} |u(\tilde{x}) - y| \right] = \mathbb{E}_{(x,y)\sim\mu}[|u(x) - y|] + \widetilde{\mathrm{TV}}_\varepsilon(u; \mu), \tag{3.2}$$

where

$$\widetilde{\mathrm{TV}}_\varepsilon(u; \mu) := \int_{\mathcal{X}} \frac{\sup_{B_\varepsilon(x)} u - u(x)}{\varepsilon} \, \mathrm{d}\rho_0(x) + \int_{\mathcal{X}} \frac{u(x) - \inf_{B_\varepsilon(x)} u}{\varepsilon} \, \mathrm{d}\rho_1(x).$$

The data term $\mathbb{E}_{(x,y)\sim\mu}[|u(x)-y|]$ is minimized by the so-called *Bayes classifier* u^* which is susceptible to adversarial attacks. The total variation makes the optimization problem robust by penalizing data points for which the classifier u varies strongly on the ε-ball, i.e., data points which lie close to the 'boundary' between the two data sets.

In order to turn minimization of (3.2) in $\widetilde{\mathcal{H}}$ into a well-defined problem on $L^\infty(\mathcal{X};\nu)$—and thus rendering it analytically advantageous—we replace the supremum with the ν-essential supremum, which, in fact, does not change the underlying problem.

Lemma 3.5. *Under Assumption 3.1, for any $u \in L^\infty(\mathcal{X};\nu)$ there exists $u^* \in L^\infty(\mathcal{X};\nu)$ such that $u = u^*$ holds ν-almost everywhere and*

$$\nu\text{-TV}_\varepsilon(u;\mu) = \widetilde{\text{TV}}_\varepsilon(u^*;\mu),$$

where

$$\nu\text{-TV}_\varepsilon(u;\mu) \coloneqq \inf_{\substack{v\in L^\infty(\mathcal{X};\nu)\\ u=v\ \nu\text{-a.e.}}} \widetilde{\text{TV}}_\varepsilon(v;\mu).$$

Since one trivially has $\text{TV}_\varepsilon(u;\mu) \le \widetilde{\text{TV}}_\varepsilon(u;\mu)$, the previous lemma implies that the minimal values of

$$\min_{u\in\mathcal{H}} \mathbb{E}_{(x,y)\sim\mu}[|u(x)-y|] + \text{TV}_\varepsilon(u;\mu) \quad \text{and} \quad \min_{u\in\widetilde{\mathcal{H}}} \mathbb{E}_{(x,y)\sim\mu}[|u(x)-y|] + \widetilde{\text{TV}}_\varepsilon(u;\mu)$$

coincide and that the minimizers agree up to a set of measure zero, implying that problem (3.1) is equivalent to

$$\min_{u\in\mathcal{H}} \mathbb{E}_{(x,y)\sim\mu}[|u(x)-y|] + \text{TV}_\varepsilon(u;\mu).$$

We generalize this problem to (1.5), by considering a larger class of loss functions in the data term, while retaining the regularizing term for which equivalency holds.

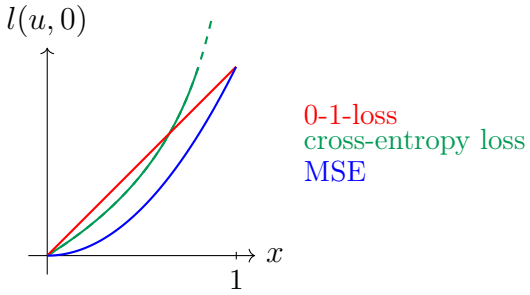

Figure 3.1: Different loss functions for the 0 class.

The assumptions made for the loss function in Theorem 3.2 clearly include problem (3.1), since the 0-1-loss is continuous and convex. Furthermore, these assumptions are satisfied by a plethora of loss functions used in adversarial training and classification problems [49], first and foremost by the prevalent *cross-entropy loss*

$$CE(u(x),y) \coloneqq -(1-y)\log(1-u(x)) - y\log(u(x)),$$

but also by the mean squared error loss, which are depicted together with the 0-1-loss in Figure 3.1 for the 0-class.

3.1.2 Proof of Existence of Minimizers

We turn to the proof of Theorem 3.2 which will follow the *direct method in the calculus of variations*.

Lemma 3.6 (Direct method in the calculus of variations)**.** *Let $J : K \to [-\infty, \infty]$ be a functional over some topological space K. Then the minimizing problem*

$$\min_{u \in K} J(u)$$

admits a solution if the following is satisfied:

(i) *J is bounded from below, i.e., there exists $L \in \mathbb{R}$ such that $J(u) \geq L \quad \forall u \in K$.*

(ii) *There exists a topology T on K such that K is sequentially compact in T.*

(iii) *J is lower semicontinuous on K with respect to the topology T, i.e., if we have $(u_k)_{k \in \mathbb{N}} \subset K$ such that $u_k \xrightarrow{T} u$ for some $u \in K$, it holds that*

$$J(u) \leq \liminf_{k \to \infty} J(u_k).$$

Proof. Since $J(u) > -\infty$ there exists a minimizing sequence $(u_k)_{k \in \mathbb{N}}$ in K, such that $\lim_{k \to \infty} J(u_k) = \inf_{u \in K} J(u)$. By the sequential compactness of K, the minimizing sequence has s subsequence $(u_{k_j})_{j \in \mathbb{N}} \subset K$ converging to some $u^* \in K$ with respect to T. With the lower semicontinuity of J it follows that

$$J(u^*) \leq \liminf_{j \to \infty} J(u_{k_j}) \leq \lim_{k \to \infty} J(u_k) = \inf_{u \in K} J(u).$$

□

We begin by showing a simple lower bound of problem (1.5): Clearly,

$$F : \mathcal{H} \to [-\infty, \infty], \; F(u) \coloneqq \mathbb{E}_{(x,y) \sim \mu}[l(u(x), y)] + \varepsilon \mathrm{TV}_\varepsilon(u; \mu)$$

satisfies $F(u) \geq 0$ for all $u \in \mathcal{H}$. The data term is bounded from below by 0, since the loss function is non-negative. The non-local total variation is bounded from below because $\rho \ll \nu$ and it holds that

$$\nu\text{-}\operatorname*{ess\,sup}_{B_\varepsilon(x)} u \geq u(x) \quad \text{and} \quad \nu\text{-}\operatorname*{ess\,inf}_{B_\varepsilon(x)} u \leq u(x) \quad \nu\text{-a.e.}$$

In the next step we are going to prove sequential compactness of $\mathcal{H}$ with respect to the weak-* topology. By Theorem 2.3 the separability of $\mathcal{X}$ implies that $L^1(\mathcal{X};\nu)$ is separable. Since ν is a σ-finite measure, $L^\infty(\mathcal{X};\nu)$ can be identified with the dual of $L^1(\mathcal{X};\nu)$ and, by [20, Theorem 5.1], we have that the unit ball $B_{L^\infty(\mathcal{X};\nu)}$ is metrizable in the weak-* topology. Therefore, by [39, Theorem 28.2], weak-* sequential compactness is equivalent to weak-* compactness on $\mathcal{H} \subset B_{L^\infty(\mathcal{X};\nu)}$, and we may use these concepts interchangeably.

Lemma 3.7 (Compactness)**.** *The hypothesis space $\mathcal{H}$ is compact with respect to the weak-* topology on $L^\infty(\mathcal{X};\nu)$.*

Proof. For all $u \in \mathcal{H}$ we have $\|u\|_{L^\infty} = \operatorname{ess\,sup}_{x\in\mathcal{X}} u(x) \leq 1$. Therefore, $\mathcal{H}$ is contained in the closed unit ball $B_{L^\infty(\mathcal{X};\nu)}$. Since ν is σ-finite by assumption, $L^\infty(\mathcal{X};\nu)$ is the dual of the Banach space $L^1(\mathcal{X};\nu)$, and it follows by Banach–Alaoglu's theorem that we have weak-* precompactness of $\mathcal{H}$. To see that $\mathcal{H}$ is weak-* closed, let $(u_k)_{k\in\mathbb{N}}$ be a weak-* convergent sequence in $\mathcal{H}$ with the limit $u \in L^\infty(\mathcal{X};\nu)$, i.e.,

$$\lim_{k\to\infty} \int_{\mathcal{X}} u_k \phi \,\mathrm{d}\nu = \int_{\mathcal{X}} u\phi \,\mathrm{d}\nu \qquad \forall \phi \in L^1(\mathcal{X};\nu).$$

Suppose $u \notin \mathcal{H}$. Then there exists a set $A \subset \mathcal{X}$ with $\nu(A) > 0$ such that we have $u(x) < 0$ or $u(x) > 1$ for all $x \in A$. Let us consider the case of $u(x) > 1$. Since $\chi_A \in L^1(\mathcal{X};\nu)$ and $u_k \leq 1$ ν-almost everywhere, it holds

$$\nu(A) < \int_A u \,\mathrm{d}\nu = \int_{\mathcal{X}} \chi_A u \,\mathrm{d}\nu = \lim_{k\to\infty} \int_{\mathcal{X}} \chi_A u_k \,\mathrm{d}\nu \leq \int_{\mathcal{X}} \chi_A \,\mathrm{d}\nu = \nu(A). \quad \lightning$$

Similarly, we arrive at a contradiction in the other case. Hence, $u \in \mathcal{H}$ and $\mathcal{H}$ is weak-* compact. □

In the third step we prove that the the functional F in (1.5) is lower semicontinuous with respect to the weak-* topology. Since the sum of lower semicontinuous functions is lower semicontinuous, we will prove it separately for the data term and the non-local total variation. The proof for the non-local total variation was originally provided in [13] for slightly different assumptions for the reference measure ν (cf. Remark 3.3). In [12] the statement was presented with our assumptions, the proof of which we shall reproduce here for the sake of completeness.

Lemma 3.8. *The functional $u \mapsto \mathrm{TV}_\varepsilon(u;\mu)$ is lower semi-continuous with respect to the weak-* topology on $\mathcal{H}$.*

Proof. Let $(u_k)_{k\in\mathbb{N}} \subset \mathcal{H}$ such that $u_k \rightharpoonup^* u$. Using the definition of weak-* convergence and the fact that for the Radon–Nikodym derivative it holds $\frac{d\rho_i}{d\nu} \in L^1(\mathcal{X};\nu)$, we trivially have

$$\int_{\mathcal{X}} u \,\mathrm{d}\rho_i = \lim_{k\to\infty} \int_{\mathcal{X}} u_k \,\mathrm{d}\rho_i \quad \text{for} \quad i \in \{0,1\}.$$

Hence, it suffices to study the simplified functional

$$J(u) := \int_{\mathcal{X}} \nu\text{-}\operatorname*{ess\,sup}_{B_\varepsilon(x)} u \,\mathrm{d}\rho_0(x).$$

The proof for the ρ_1-part containing the essential infimum will follow with a similar argument.

We claim that for every $x \in \mathcal{X}$, it holds

$$m_0(x) := \nu\text{-}\operatorname*{ess\,sup}_{B_\varepsilon(x)} u \leq \liminf_{k\to\infty} \nu\text{-}\operatorname*{ess\,sup}_{B_\varepsilon(x)} u_k. \tag{3.3}$$

If $m_0 = 0$ for every $x \in \mathcal{X}$, we are done since $0 \leq u_k \leq 1$ ν-almost everywhere. Assume there exists $x \in \mathcal{X}$ such that $m_0(x) > 0$. Using this, as well as Assumption 3.1, for any $0 < \gamma < m_0(x)$, the set

$$C := \{y \in B_\varepsilon(x) : u(y) \geq m_0(x) - \gamma\}$$

satisfies $0 < \nu(C) \leq \nu(B_\varepsilon(x)) < \infty$. Furthermore, using the weak-* convergence $u_k \rightharpoonup^* u$ together with $\chi_C \in L^1(\mathcal{X};\nu)$, it follows

$$\nu(C)(m_0(x) - \gamma) \leq \int_C u \,\mathrm{d}\nu = \lim_{k\to\infty} \int_C u_k \,\mathrm{d}\nu \leq \nu(C) \liminf_{k\to\infty} \nu\text{-}\operatorname*{ess\,sup}_{B_\varepsilon(x)} u_k.$$

Canceling $\nu(C)$ and sending $\gamma \to 0$ afterwards proves the claim. We conclude the proof of lower semicontinuity of J by employing (3.3) and Fatou's lemma

$$\begin{aligned} J(u) &= \int_{\mathcal{X}} \nu\text{-}\operatorname*{ess\,sup}_{B_\varepsilon(x)} u \,\mathrm{d}\rho_0(x) \leq \int_{\mathcal{X}} \liminf_{k\to\infty} \nu\text{-}\operatorname*{ess\,sup}_{B_\varepsilon(x)} u_k \,\mathrm{d}\rho_0(x) \\ &\leq \liminf_{k\to\infty} \int_{\mathcal{X}} \nu\text{-}\operatorname*{ess\,sup}_{B_\varepsilon(x)} u_k \,\mathrm{d}\rho_0(x) = \liminf_{k\to\infty} J(u_k). \end{aligned}$$

□

To proof the weak-* lower semicontinuity of the data term we will need the following simple relation between weak and weak-* convergence.

Lemma 3.9. *Let $(u_k)_{k\in\mathbb{N}} \subset L^\infty(\mathcal{X};\nu)$ be a weak-* convergent sequence with the limit $u \in L^\infty(\mathcal{X};\nu)$. Then $u_k \rightharpoonup u$ in the weak topology on $L^2(\mathcal{X};\rho)$.*

Proof. By assumption we have

$$\lim_{k\to\infty} \int_{\mathcal{X}} u_k \phi \,\mathrm{d}\nu = \int_{\mathcal{X}} u\phi \,\mathrm{d}\nu \quad \forall \phi \in L^1(\mathcal{X};\nu). \tag{3.4}$$

Since ρ is absolutely continuous with respect to ν, one easily has $u, u_k \in L^\infty(\mathcal{X};\rho)$, and we argue that $u_k \rightharpoonup^* u$ in $L^\infty(\mathcal{X};\rho)$:
With the Radon-Nykodyn derivative $\frac{\mathrm{d}\rho}{\mathrm{d}\nu}$, it holds for any $v \in L^1(\mathcal{X};\rho)$ that

$$\int_{\mathcal{X}} |v| \frac{\mathrm{d}\rho}{\mathrm{d}\nu} \,\mathrm{d}\nu = \int_{\mathcal{X}} |v| \,\mathrm{d}\rho < \infty.$$

Hence, $\tilde{\phi} \coloneqq v\frac{\mathrm{d}\rho}{\mathrm{d}\nu} \in L^1(\mathcal{X};\nu)$ and we have

$$\lim_{k\to\infty}\int_{\mathcal{X}} u_k v \,\mathrm{d}\rho = \lim_{k\to\infty}\int_{\mathcal{X}} u_k\tilde{\phi}\,\mathrm{d}\nu \overset{(3.4)}{=} \int_{\mathcal{X}} u\tilde{\phi}\,\mathrm{d}\nu = \int_{\mathcal{X}} uv\,\mathrm{d}\rho.$$

Since $v \in L^1(\mathcal{X};\rho)$ was arbitrary, we have weak-* convergence of u_k in $L^\infty(\mathcal{X};\rho)$. We further have $u, u_k \in L^2(\mathcal{X};\rho)$ and $L^2(\mathcal{X};\rho) \subset L^1(\mathcal{X};\rho)$, by the fact that $(\mathcal{X};\rho)$ is a finite measure space. Therefore, it immediately follows that

$$\lim_{k\to\infty}\int_{\mathcal{X}} u_k\phi\,\mathrm{d}\rho = \int_{\mathcal{X}} u\phi\,\mathrm{d}\rho \quad \forall\phi \in L^2(\mathcal{X};\rho),$$

i.e., $u_k \to u$ weakly in $L^2(\mathcal{X};\rho)$. □

Lemma 3.10. *Let $l : [0,1]\times\mathcal{Y} \to [0,\infty]$ be continuous and convex in the first variable. Then the functional $u \mapsto \mathbb{E}_{(x,y)\sim\mu}[l(u(x),y)]$ is lower semicontinuous with respect to the weak-* topology on $\mathcal{H}$.*

Proof. By the law of total expectation we can split the data term into the conditional distributions

$$\begin{aligned}\mathbb{E}_{(x,y)\sim\mu}[l(u(x),y)] &= \int_{\mathcal{X}\times\{0,1\}} l(u(x),y)\,\mathrm{d}\mu(x,y)\\ &= \int_{\mathcal{X}} l(u(x),0)\,\mathrm{d}\rho_0(x) + \int_{\mathcal{X}} l(u(x),1)\,\mathrm{d}\rho_1(x)\end{aligned}$$

and show lower semicontinuity for each summand separately. Let $i \in \{0,1\}$ be fixed. Then, by assumption, $l_i : [0,1] \to [0,\infty]$, $l_i(x) \coloneqq l(x,i)$ is a convex function and it immediately follows that $J(u) \coloneqq \int_{\mathcal{X}} l(u(x),i)\,\mathrm{d}\rho_i(x)$ is a convex functional, since

$$\begin{aligned}\int_{\mathcal{X}} l_i(\lambda u + (1-\lambda)v)\,\mathrm{d}\rho_i(x) &\le \int_{\mathcal{X}} \lambda l_i(u) + (1-\lambda)l_i(v)\,\mathrm{d}\rho_i(x)\\ &= \lambda\int_{\mathcal{X}} l_i(u)\,\mathrm{d}\rho_i(x) + (1-\lambda)\int_{\mathcal{X}} l_i(v)\,\mathrm{d}\rho_i(x)\end{aligned}$$

holds for all $u, v \in \mathcal{H}$ and $\lambda \in [0,1]$. Hence, by Theorem 2.13, weak lower semicontinuity of J is equivalent to strong lower semicontinuity it the Hilbert space $L^2(\mathcal{X};\rho)$. Let $u_k \to u$ be a sequence converging in the strong topology on $L^2(\mathcal{X};\rho)$ and extract a subsequence $(u_{k_l})_{l\in\mathbb{N}}$ such that

$$\lim_{l\to\infty}\int_{\mathcal{X}} l(u_{k_l}(x),i)\,\mathrm{d}\rho_i(x) = \liminf_{k\to\infty}\int_{\mathcal{X}} l(u_k(x),i)\,\mathrm{d}\rho_i(x). \tag{3.5}$$

By extracting a further subsequence (not relabeled) for which pointwise convergence holds almost everywhere, it follows with the continuity of l, Fatou's lemma and (3.5)

$$\begin{aligned}\int_{\mathcal{X}} l(u(x),i)\,\mathrm{d}\rho_i(x) &= \int_{\mathcal{X}} l(\lim_{l\to\infty} u_{k_l}(x),i)\,\mathrm{d}\rho_i(x) \le \liminf_{l\to\infty}\int_{\mathcal{X}} l(u_{k_l}(x),i)\,\mathrm{d}\rho_i(x)\\ &= \liminf_{k\to\infty}\int_{\mathcal{X}} l(u_k(x),i)\,\mathrm{d}\rho_i(x).\end{aligned}$$

Therefore, J is lower semicontinuous with respect to the weak topology in $L^2(\mathcal{X};\rho)$, and since weak-* convergence in $L^\infty(\mathcal{X};\nu)$ implies weak convergence in $L^2(\mathcal{X};\rho)$ (see Lemma 3.9), we have proved weak-* lower semicontinuity in $\mathcal{H}$. □

By the direct method, this concludes the proof of Theorem 3.2.

3.2 Multiclass Case

The result for the existence of solutions to the binary classification problem is easily generalized to the classification problem with M classes. The training data is given by a probability measure $\mu \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ with $\mathcal{Y} \coloneqq \{1, \dots, M\}$, and as in the binary case we define the conditional distributions $\rho_i(\,\cdot\,) \coloneqq \mu(\,\cdot\, \times \{i\})$. The hypothesis space

$$\mathcal{H} \coloneqq \left\{ u \in [L^\infty(X;\nu)]^M : 0 \le u_i \le 1, \sum_{i=1}^{M} u_i = 1 \right\} \tag{3.6}$$

consists of vector-valued L^∞-functions with positive components summing to 1. We interpret each component $u^i(x)$ of these soft classifiers as a probability that the data point x belongs to the class $i \in \{1, \dots, M\}$ and one may construct a hard classifier by assigning to x the class which corresponding component of u has a maximal entry.

As in the previous section, our aim is to show existence of solutions to the total variation regularized problem.

Theorem 3.11. *Let $l : [0,1]^M \times \mathcal{Y} \to [0,\infty]$ be a loss function which is continuous and convex for fixed $y \in \mathcal{Y}$. Then, under Assumption 3.1*

$$\min_{u \in \mathcal{H}} \mathbb{E}_{(x,y)\sim\mu} \left[l^M(u(x), y) \right] + \varepsilon \mathrm{TV}_\varepsilon^M(u;\mu) \tag{3.7}$$

admits a solution.

The multiclass non-local total variation is defined as

$$\mathrm{TV}_\varepsilon^M(u;\mu) \coloneqq \frac{1}{\varepsilon} \sum_{i=1}^{M} \int_{\mathcal{X}} u^i(x) - \nu\text{-}\operatorname*{ess\,inf}_{\tilde{x} \in B_\varepsilon(x)} u^i(\tilde{x}) \, \mathrm{d}\rho_i(x),$$

and we assume the multiclass loss function

$$l^M : [0,1]^M \times \mathcal{Y} \to [0,\infty],\ l^M(u(x), y) \coloneqq l_y(u(x))$$

to be a continuous and convex loss l_y on the convex set $[0,1]^M$ for each label $y \in \mathcal{Y}$. This general case enables the use of a different loss in each class, which, for example, allows for the possibility to weigh the classes differently. Furthermore, each l_y may depend on the whole classification vector u, allowing the loss for a fixed class $l_y(u(x))$ to differ from

$l_y(v(x))$, even when $u_y(x) = v_y(x)$. However, in the case of the multiclass cross-entropy loss

$$CE(u(x), y) \coloneqq -\log(u^y(x)),$$

for example, this simplifies to the same logarithmic loss in each class y, which only depends on the y-th component of the probability vector u.

We proceed to show the existence of minimizers analogously to the binary case and begin with the compactness result on $\mathcal{H}$.

Lemma 3.12. *For every sequence of functions $(u_k)_{k\in\mathbb{N}} \subset \mathcal{H}$ there exists a converging subsequence with limit $u \in \mathcal{H}$.*

Proof. By the compactness argument in Chapter 3.1, for every component u_k^i of u_k there exits a subsequence such that u_k^i weakly-* converges to $u^i \in L^\infty(\mathcal{X}; \nu)$ and by successively taking subsequences we find a subsequence such that convergence holds for all components. By the same reasoning as in Lemma 3.7, we have $u_i \geq 0$ which only leaves to show that the components of the weak-* limit u sum to 1 ν-almost everywhere. Suppose there exists a set $A \subset \mathcal{X}$ with positive measure such that $\sum_{i=1}^M u^i(x) < 1$ for all $x \in A$. Then, by weak-* convergence and $\chi_A \in L^\infty(\mathcal{X}; \nu)$ we have

$$\begin{aligned}\nu(A) &> \int_{\mathcal{X}} \sum_{i=1}^M u^i(x)\chi_A \,\mathrm{d}\nu = \sum_{i=1}^M \int_{\mathcal{X}} u^i(x)\chi_A \,\mathrm{d}\nu = \sum_{i=1}^M \lim_{k\to\infty} \int_{\mathcal{X}} u_k^i(x)\chi_A \,\mathrm{d}\nu \\ &= \lim_{k\to\infty} \int_{\mathcal{X}} \sum_{i=1}^M u_k^i(x)\chi_A \,\mathrm{d}\nu = \nu(A), \quad \lightning\end{aligned}$$

and similarly, one arrives at a contradiction in the case $\sum_{i=1}^M u^i(x) > 1$, proving the lemma. □

The weak-* lower semicontinuity of the functional in (3.7) directly follows from the binary case, as well. The proof for the weak-* lower semicontinuity of the Bayes risk works analogously to the proof of Lemma 3.10, since the data term can be split into the different classes

$$\begin{aligned}\mathbb{E}_{(x,y)\sim\mu}[l^M(u(x), y)] &= \int_{\mathcal{X}\times\{0,\dots,M\}} l(u(x), y) \,\mathrm{d}\mu(x, y) \\ &= \int_{\mathcal{X}} l_1(u(x)) \,\mathrm{d}\rho_1(x) + \cdots + \int_{\mathcal{X}} l_M(u(x)) \,\mathrm{d}\rho_M(x),\end{aligned}$$

and since $\mathrm{TV}_\varepsilon^M(u; \mu)$ is simply the sum of terms which coincide with the second term of the binary non-local total variation, the proof of Lemma 3.8 is easily transferred to the multiclass case. Hence, we have existence of solutions to (3.7) via the direct method, proving Theorem 3.11.

3.3 Original Optimization Problem

We return to the original robust optimization problem (1.2) and combine ideas form [13] and Section 3.1 to show the existence of minimizers in the hypothesis space

$$\mathcal{H} := \{u \in L^\infty(\mathcal{X}; \nu) : 0 \leq u \leq 1\}.$$

Theorem 3.13. *Let $l : [0,1] \times \mathcal{Y} \to [0,\infty]$ be a loss function which is continuous, convex and monotone in the first variable. Then, under Assumption 3.1*

$$\min_{u \in \mathcal{H}} \mathbb{E}_{(x,y)\sim\mu} \left[\sup_{\tilde{x} \in B_\varepsilon(x)} l(u(\tilde{x}), y) \right]$$

admits a solution.

Similarly to the argumentation in Section 3.1.1, the following lemma shows that one can replace the supremum in (1.2) by the ν-essential supremum. Therefore, we proceed to show the existence of minimizer to the equivalent problem

$$\min_{u \in \mathcal{H}} \mathbb{E}_{(x,y)\sim\mu} \left[\nu\text{-}\operatorname*{ess\,sup}_{\tilde{x} \in B\varepsilon(x)} l(u(\tilde{x}), y) \right]. \tag{3.8}$$

Lemma 3.14. *Under Assumption 3.1 for any Borel measurable function $u \in L^\infty(\mathcal{X}; \nu)$ there exists $u^* : \mathcal{X} \to \mathbb{R}$ such that $u = u^*$ holds ν-almost everywhere and*

$$\sup_{B_\varepsilon(x)} u^* = \nu\text{-}\operatorname*{ess\,sup}_{B_\varepsilon(x)} u^*, \qquad \inf_{B_\varepsilon(x)} u^* = \nu\text{-}\operatorname*{ess\,inf}_{B_\varepsilon(x)} u^* \quad \forall x \in \operatorname{supp} \rho.$$

Furthermore, if $l : [0,1] \times Y \to \mathbb{R}_0^+$ is continuous and monotone in the first variable it holds

$$\sup_{B_\varepsilon(x)} l(u^*, i) = \nu\text{-}\operatorname*{ess\,sup}_{B_\varepsilon(x)} l(u^*, i) \quad \forall x \in \operatorname{supp} \rho.$$

Proof. The first part of this lemma was proved in [13, Lemma 3.21] for slightly stricter assumptions on ν, but continues to hold in our setting (cf. Remark 3.3). To show the second part, let u^* be as described and let l be non-increasing in the first variable. Then we have for all $x \in \mathcal{X}$

$$l(u^*(\tilde{x}), i) \leq l(\inf_{B_\varepsilon(x)} u^*, i) \quad \forall \tilde{x} \in B_\varepsilon(x)$$

and therefore,

$$\sup_{B_\varepsilon(x)} l(u^*, i) \leq l(\inf_{B_\varepsilon(x)} u^*, i). \tag{3.9}$$

On the other hand, by definition of the infimum, there exists a sequence $(\tilde{x}_k)_{k\in\mathbb{N}} \subset B_\varepsilon(x)$ with $\lim_{k\to\infty} u^*(\tilde{x}_k) = \inf_{B_\varepsilon(x)} u^*$, and by continuity it holds

$$\sup_{B_\varepsilon(x)} l(u^*, i) \geq \lim_{k\to\infty} l(u^*(\tilde{x}_k), i) = l(\lim_{k\to\infty} u^*(\tilde{x}_k), i) = l(\inf_{B_\varepsilon(x)} u^*, i). \tag{3.10}$$

Hence, by combining (3.9) and (3.10), we have equality and by the choice of u^* it holds

$$\sup_{B_\varepsilon(x)} l(u^*, i) = l(\inf_{B_\varepsilon(x)} u^*, i) = l(\nu\text{-}\operatorname{ess\,inf}_{B_\varepsilon(x)} u^*, i) = \nu\text{-}\operatorname{ess\,sup}_{B_\varepsilon(x)} l(u^*, i).$$

The last equality holds by the same argumentation as for the supremum.

In the case of a non-decreasing loss function l, one similarly shows the two inequalities

$$\sup_{B_\varepsilon(x)} l(u^*, i) \leq l(\sup_{B_\varepsilon(x)} u^*, i), \quad \sup_{B_\varepsilon(x)} l(u^*, i) \geq l(\sup_{B_\varepsilon(x)} u^*, i),$$

such that, by the choice of u^*, it holds

$$\sup_{B_\varepsilon(x)} l(u^*, i) = l(\sup_{B_\varepsilon(x)} u^*, i) = l(\nu\text{-}\operatorname{ess\,sup}_{B_\varepsilon(x)} u^*, i) = \nu\text{-}\operatorname{ess\,sup}_{B_\varepsilon(x)} l(u^*, i),$$

proving the lemma. □

Remark 3.15. The assumption that the loss function should be monotone is reasonable and already satisfied by our continuous and convex loss functions. Since we only consider hard labels, i.e., the labels are either 0 or 1 and can not take values in between, the loss function splits into the different classes and since the loss is 0 for correctly classified points, the convexity and non-negativity of the loss function, implies monotonicity. Furthermore, a loss function which is non-monotone for a fixed class would imply the seemingly unreasonable property that the loss can decrease if one deviates further from the correct label. See, for example Figure 3.2, depicting the cross-entropy loss, for which $CE(u(x), 0) = -\log(1 - u(x))$ is non-decreasing on $[0, 1]$ and $CE(u(x), 1) = -\log(u(x))$ is non-increasing on $[0, 1]$.

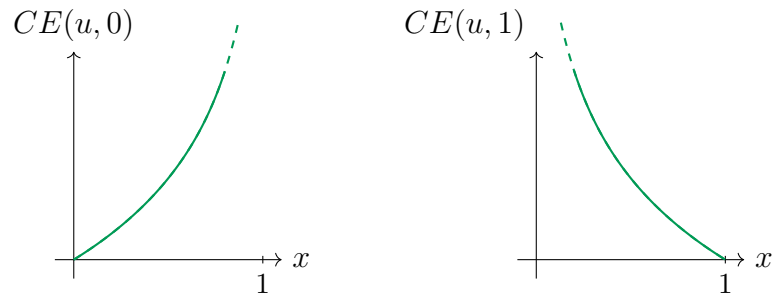

Figure 3.2: The cross-entropy loss for a fixed class $y \in \mathcal{Y}$.

We turn to the proof of weak-* lower semicontinuity which combines the ideas in the proofs of Lemma 3.10 and Lemma 3.8.

Lemma 3.16. *Let $l : [0, 1] \times \mathcal{Y} \to [0, \infty]$ be continuous and convex in the first variable. Then the functional*

$$u \mapsto \mathbb{E}_{(x,y)\sim\mu}\left[\nu\text{-}\operatorname{ess\,sup}_{\tilde{x}\in B_\varepsilon(x)} l(u(\tilde{x}), y)\right]$$

is weak- lower semicontinuous.*

Proof. As in the proof of Lemma 3.10 we split the functional according to the two classes $i \in \{0, 1\}$ and show lower semicontinuity of

$$J(u) \coloneqq \int_{\mathcal{X}} \nu\text{-}\operatorname*{ess\,sup}_{\tilde{x} \in B_\varepsilon(x)} l(u(\tilde{x}), i) \, \mathrm{d}\rho_i(x).$$

Since

$$\nu\text{-}\operatorname*{ess\,sup}_{\tilde{x} \in B_\varepsilon(x)} [\lambda l(u(\tilde{x}), i) + (1-\lambda) l(v(\tilde{x}), i)] \leq \lambda \nu\text{-}\operatorname*{ess\,sup}_{\tilde{x} \in B_\varepsilon(x)} l(u(\tilde{x}), i) + (1-\lambda) \nu\text{-}\operatorname*{ess\,sup}_{\tilde{x} \in B_\varepsilon(x)} l(v(\tilde{x}), i)$$

holds for all $u, v \in \mathcal{H}$ and $\lambda \in [0, 1]$, the functional J is convex and we proceed to show weak-* lower semicontinuity by verifying lower semicontinuity in the strong topology on $L^2(\mathcal{X}; \nu)$ (cf. proof of Lemma 3.10). Let $u_k \to u$ be a sequence converging in the strong topology on $L^2(\mathcal{X}; \nu)$ and extract a subsequence $(u_{k_l})_{l \in \mathbb{N}}$ such that

$$\lim_{l \to \infty} \int_{\mathcal{X}} \nu\text{-}\operatorname*{ess\,sup}_{\tilde{x} \in B_\varepsilon(x)} l(u_{k_l}(\tilde{x}), i) \, \mathrm{d}\rho_i(x) = \liminf_{k \to \infty} \int_{\mathcal{X}} \nu\text{-}\operatorname*{ess\,sup}_{\tilde{x} \in B_\varepsilon(x)} l(u_k(\tilde{x}), i) \, \mathrm{d}\rho_i(x). \tag{3.11}$$

By extracting a further subsequence (not relabeled) for which pointwise convergence holds almost everywhere, one has for all $x \in \mathcal{X}$

$$m_0(x) \coloneqq \nu\text{-}\operatorname*{ess\,sup}_{\tilde{x} \in B_\varepsilon(x)} l(u(\tilde{x}), i) \leq \liminf_{l \to \infty} \nu\text{-}\operatorname*{ess\,sup}_{\tilde{x} \in B_\varepsilon(x)} l(u_{k_l}(\tilde{x}), i). \tag{3.12}$$

The proof of this claim works similarly to the proof of (3.3): If $m_0(x) = 0$ for all $x \in \mathcal{X}$, we are done. Otherwise, it holds for all $0 < \gamma < m_0(x)$ that there exists $x \in \mathcal{X}$ such that for

$$C \coloneqq \{\tilde{x} \in B_\varepsilon(x) : l(u(\tilde{x}), i) \geq m_0(x) - \gamma\}$$

one has $0 < \nu(C) < \infty$. By continuity of l and Fatou's lemma we have

$$\begin{aligned} \nu(C)(m_0(x) - \gamma) &\leq \int_C l(u(x), i) \, \mathrm{d}\nu \leq \liminf_{l \to \infty} \int_C l(u_{k_l}(x), i) \, \mathrm{d}\nu \\ &\leq \nu(C) \liminf_{l \to \infty} \nu\text{-}\operatorname*{ess\,sup}_{\tilde{x} \in B_\varepsilon(x)} l(u_{k_l}(\tilde{x}), i). \end{aligned}$$

Canceling $\nu(C)$ and sending γ to 0 proves the claim. The lower semicontinuity follows by Fatou's lemma and (3.11):

$$\begin{aligned} \int_{\mathcal{X}} \nu\text{-}\operatorname*{ess\,sup}_{\tilde{x} \in B_\varepsilon(x)} l(u(\tilde{x}), i) \, \mathrm{d}\rho_i(x) &\leq \liminf_{l \to \infty} \int_{\mathcal{X}} \nu\text{-}\operatorname*{ess\,sup}_{\tilde{x} \in B_\varepsilon(x)} l(u_{k_l}(\tilde{x}), i) \, \mathrm{d}\rho_i(x) \\ &= \liminf_{k \to \infty} \int_{\mathcal{X}} \nu\text{-}\operatorname*{ess\,sup}_{\tilde{x} \in B_\varepsilon(x)} l(u_k(\tilde{x}), i) \, \mathrm{d}\rho_i(x) \end{aligned}$$

□

Since weak-* compactness of $\mathcal{H}$ was shown in Chapter 3.1, the lower semicontinuity implies the existence of solutions to (3.8) by the direct method, proving Theorem 3.13.

Chapter 4

Asymptotics of the Non-local Weighted Perimeter and Total Variation

The main goal of this chapter is to show the Γ-convergence and equi-coerciveness of the non-local weighted total variation (1.4) as the adversarial budget ε tends to 0. In Chapter 5 we will apply these results to derive conclusions about the asymptotic behavior of the regularized optimization problem (1.5). As a intermediate step—and an interesting problem in its own right, as it relates to optimization with hard classifiers—we begin by proving Γ-convergence and equi-coerciveness of the non-local weighted perimeter. Although these problems have been studied extensively in [14], we wish to present straightforward proofs under altered assumptions for the data distributions. In particular, we will work with continuous densities and drop the assumption that they are functions of bounded variation.

The setup for this chapter arises as a special case from Chapter 3. As the metric space $\mathcal{X}$, representing the input data, we consider an open and bounded Lipschitz domain $\Omega \subset \mathbb{R}^d$ and the data distribution is given by the two probability measures $\rho_0, \rho_1 \in \mathcal{P}(\Omega)$, which we assume to be absolutely continuous with respect to the d-dimensional Lebesgue measure. Throughout this chapter, ρ_i will sometimes refer to the density of the measure ρ_i, meaning that $\mathrm{d}\rho_i(x) = \rho_i \,\mathrm{d}x$ and $\rho_i \in L^1(\Omega)$ for $i \in \{0, 1\}$. The use will be clear from the context.

In this setting, the non-local total variation is of the form

$$\mathrm{TV}_\varepsilon(u; \boldsymbol{\rho}) \coloneqq \frac{1}{\varepsilon} \int_\Omega \operatorname*{ess\,sup}_{B_\varepsilon(x)\cap\Omega} u - u(x) \,\mathrm{d}\rho_0(x) + \frac{1}{\varepsilon} \int_\Omega u(x) - \operatorname*{ess\,inf}_{B_\varepsilon(x)\cap\Omega} u \,\mathrm{d}\rho_1(x),$$

where $\boldsymbol{\rho} \coloneqq (\rho_0, \rho_1)$ denotes the dependency on the data distribution.

For a measurable set $A \subset \Omega$ the non-local weighted perimeter is defined as

$$\mathrm{Per}_\varepsilon(A; \boldsymbol{\rho}) \coloneqq \frac{1}{\varepsilon} \int_\Omega \operatorname*{ess\,sup}_{B_\varepsilon(x)\cap\Omega} \chi_A - \chi_A(x) \,\mathrm{d}\rho_0(x) + \frac{1}{\varepsilon} \int_\Omega \chi_A(x) - \operatorname*{ess\,inf}_{B_\varepsilon(x)\cap\Omega} \chi_A \,\mathrm{d}\rho_1(x). \tag{4.1}$$

Further, we denote by

$$\mathrm{Per}_\varepsilon^0(A;\boldsymbol{\rho}) \coloneqq \frac{1}{\varepsilon}\int_\Omega \operatorname*{ess\,sup}_{B_\varepsilon(x)\cap\Omega} \chi_A - \chi_A(x)\,\mathrm{d}\rho_0(x),$$
$$\mathrm{Per}_\varepsilon^1(A;\boldsymbol{\rho}) \coloneqq \frac{1}{\varepsilon}\int_\Omega \chi_A(x) - \operatorname*{ess\,inf}_{B_\varepsilon(x)\cap\Omega} \chi_A\,\mathrm{d}\rho_1(x)$$

the *outer* and *inner perimeter*, respectively. Analogously to the non-local total variation, this notion of a perimeter naturally arises from the robust optimization problem with 0-1-loss (3.1), when restricted to the hypothesis space of hard classifiers (see [13] for details). Furthermore, these functional are closely linked by the coarea formula in the non-local case

$$\mathrm{TV}_\varepsilon(u;\boldsymbol{\rho}) = \int_{\mathbb{R}} \mathrm{Per}_\varepsilon(\{u \geq t\};\boldsymbol{\rho})\,\mathrm{d}t,$$

a proof of which can be found in [13, Proposition 3.13]. Since the measure of the set of values t such that level-sets $\{u = t\}$ have positive mass is zero, the coarea formula also holds with a strict inequality

$$\mathrm{TV}_\varepsilon(u;\boldsymbol{\rho}) = \int_{\mathbb{R}} \mathrm{Per}_\varepsilon(\{u > t\};\boldsymbol{\rho})\,\mathrm{d}t. \tag{4.2}$$

We turn to the investigation of the asymptotic behavior of the non-local perimeter.

4.1 Asymptotics of the Perimeter

We show a compactness result for the perimeter, which will be necessary to prove the asymptotic result of the total variation in the next section. In the second part of this section, we will prove the following Γ-convergence result for the non-local perimeter:

Theorem 4.1 (Γ-convergence of the perimeter)**.** *Let $\rho_0, \rho_1 \in C(\overline{\Omega})$. Then it holds*

$$\mathrm{Per}_\varepsilon(\,\cdot\,;\boldsymbol{\rho}) \xrightarrow{\Gamma} \mathrm{Per}(\,\cdot\,;\boldsymbol{\rho}), \tag{4.3}$$

in the strong $L^1(\Omega)$ topology with the Γ-limit given by

$$\mathrm{Per}(A;\boldsymbol{\rho}) \coloneqq \begin{cases} \int_{\partial^* A\cap\Omega} \rho_0 + \rho_1\,\mathrm{d}\mathcal{H}^{d-1}, & \chi_A \in BV(\Omega) \\ \infty, & \textit{else.} \end{cases}$$

With the structure theorem for sets of finite perimeter it is easy to see that the Γ-limit coincides with the weighted perimeter defined in (2.11) for the weight $\rho_0 + \rho_1$. The compactness result together with the Γ-convergence of the perimeter were employed in [14] to analyze the asymptotic behavior of (3.1) for hard classifiers.

As a preparatory step for all results in this section, we prove a reformulation of the non-local perimeter of a measurable set A, showing that it is uniquely determined by A^1,

the points of density 1 in A. From here on out, we denote by A^c the relative complement of A in Ω, i.e., $A^c \coloneqq (\mathbb{R}^d \setminus A) \cap \Omega$.

Lemma 4.2. *The perimeter of a measurable set $A \subset \Omega$ admits the following equivalent representation:*

$$\mathrm{Per}_\varepsilon(A;\boldsymbol{\rho}) = \frac{1}{\varepsilon}\rho_0(\{x \in (A^1)^c : d(x, A^1) < \varepsilon\}) + \frac{1}{\varepsilon}\rho_1(\{x \in A^1 : d(x, (A^c)^1) < \varepsilon\})$$

Proof. For a set A we define the *essential distance*

$$\text{ess-}d(x, A) \coloneqq \operatorname*{ess\,inf}_{y\in A}|x - y|,$$

and claim that the following two equalities hold

$$\operatorname*{ess\,sup}_{B_\varepsilon(x)\cap\Omega} \chi_A - \chi_A(x) = \chi_{\{x\in A^c:\text{ess-}d(x,A)<\varepsilon\}} \quad \rho_0\text{-a.e.} \tag{4.4}$$

and

$$\text{ess-}d(x, A) = d(x, A^1). \tag{4.5}$$

Combining these two claims with the Lebesgue density theorem (Theorem 2.7) will give us

$$\begin{aligned}\int_\Omega \operatorname*{ess\,sup}_{B_\varepsilon(x)\cap\Omega} -\chi_A(x)\,\mathrm{d}\rho_0(x) &= \int_\Omega \chi_{\{x\in A^c:d(x,A^1)<\varepsilon\}}\,\mathrm{d}\rho_0(x)\\ &= \int_\Omega \chi_{\{x\in (A^1)^c:d(x,A^1)<\varepsilon\}}\,\mathrm{d}\rho_0(x).\end{aligned}$$

We only provide a detailed proof for the outer perimeter and note that for the inner perimeter one shows

$$\chi_A(x) - \operatorname*{ess\,inf}_{B_\varepsilon(x)\cap\Omega} \chi_A = \chi_{\{x\in A:\text{ess-}d(x,A^c)<\varepsilon\}} \quad \rho_1\text{-a.e.},$$

analogously to (4.4), and applies (4.5) to A^c.

We begin by proving (4.4) by cases. If the right hand side of (4.4) is equal to 1 we have ess-$d(x, A) < \varepsilon$. Therefore, by the Lebesgue density theorem, there exists $\tilde{x} \in B_\varepsilon(x) \cap \Omega$ with

$$\lim_{r\searrow 0} \frac{\mathcal{L}^d(A \cap B_r(\tilde{x}))}{\mathcal{L}^d(B_r(\tilde{x}))} = 1,$$

and since $B_\varepsilon(x) \cap \Omega$ is open, for some $\tilde{r} > 0$ we have $B_{\tilde{r}}(\tilde{x}) \subset B_\varepsilon(x) \cap \Omega$ and $\mathcal{L}^d(A \cap B_{\tilde{r}}(\tilde{x})) > 0$ which implies $\mathcal{L}^d(A \cap B_\varepsilon(x)) > 0$. Then, $\operatorname{ess\,sup}_{B_\varepsilon(x)\cap\Omega} \chi_A = 1$ and the left hand side of (4.4) also equals 1. If the right hand side equals 0, we have $x \in A$ or ess-$d(x, A) \geq \varepsilon$. If ess-$d(x, A) \geq \varepsilon$, then it trivially holds that

$$\operatorname*{ess\,sup}_{B_\varepsilon(x)\cap\Omega} \chi_A = \chi_A(x) = 0.$$

In the case of $x \in A$, we have

$$\operatorname*{ess\,sup}_{B_\varepsilon(x)\cap\Omega} \chi_A - \chi_A(x) = 0 \quad \rho_0\text{-a.e.},$$

since by the Lebesgue density theorem $\operatorname{ess\,sup}_{B_\varepsilon(y)\cap\Omega} \chi_A = 1$ for ρ_0-almost every $y \in A$, proving the claim.

To verify the equality in (4.5) first note that by the Lebesgue density theorem we have

$$\text{ess-}d(x, A) = \operatorname*{ess\,inf}_{y\in A}|x-y| = \operatorname*{ess\,inf}_{y\in A^1}|x-y| \geq \inf_{y\in A^1}|x-y| = d(x, A^1).$$

Further, let $\varepsilon, \delta > 0$ and let $y \in A^1$ such that $|y - x| < d(x, A^1) + \varepsilon$. Then the set

$$\{z \in A : |z - x| < d(x, A^1) + \varepsilon + \delta\}$$

has positive measure, since it contains $B_\delta(y) \cap A$, which in turn is a set with positive measure because $y \in A^1$. By the definition of the essential infimum it follows

$$\text{ess-}d(x, A) = \operatorname*{ess\,inf}_{y\in A}|x-y| = \sup\{b : |\{y \in A : |x-y| < b\}| = 0\} \leq d(x, A^1) + \varepsilon + \delta.$$

Sending ε and δ to 0 completes the proof. □

4.1.1 A Compactness Result

To prove the compactness result for sequences of sets with bounded non-local perimeter, we will need the following lemma, stating that BV is closed under multiplication with L^∞-functions.

Lemma 4.3. *Let $\Omega \subset \mathbb{R}^d$ be an open and bounded Lipschitz domain, let $u \in W^{1,1}(\Omega) \cap L^\infty(\Omega)$ and $v \in BV(\Omega) \cap L^\infty$ be non-negative. Then it holds $uv \in BV(\Omega)$ and*

$$\mathrm{TV}(uv) \leq \int_\Omega |\nabla u| v \,\mathrm{d}x + \|u\|_{L^\infty} \mathrm{TV}(v).$$

Proof. Define $\tilde{v} \in BV(\Omega)$ as

$$\tilde{v} \coloneqq \begin{cases} v, & \text{for } x \in \Omega \\ 0, & \text{for } x \in \mathbb{R}^d \setminus \Omega. \end{cases}$$

and for $\varepsilon > 0$ let $v_\varepsilon \coloneqq \tilde{v} * \phi_\varepsilon$ be a mollification of $\tilde{v}$ with the standard mollifier ϕ_ε. Then, by [1, Remark 3.22], we have $v_\varepsilon \to v$ in $L^1(\Omega)$ such that $\mathrm{TV}(v_\varepsilon) \to \mathrm{TV}(v)$. Furthermore, it follows with Young's convolution inequality (see, e.g., [31, Theorem 1.2.10])

$$\|v_\varepsilon\|_{L^\infty} = \|\tilde{v} * \phi_\varepsilon\|_{L^\infty} \leq \|\tilde{v}\|_{L^\infty} \|\phi_\varepsilon\|_{L^1} = \|\tilde{v}\|_{L^\infty} = \|v\|_{L^\infty}. \tag{4.6}$$

By Hölder's inequality it holds

$$\|uv_\varepsilon - uv\|_{L^1} \leq \|u\|_{L^\infty} \|v_\varepsilon - v\|_{L^1} \to 0$$

and we employ the lower semi-continuity of the total variation and the product rule for the weak derivative (see, e.g., [25, Theorem 1] on p.261) to see

$$\begin{aligned}\mathrm{TV}(uv) \leq \liminf_{\varepsilon\to 0} \mathrm{TV}(uv_\varepsilon) &= \liminf_{\varepsilon\to 0} \int_\Omega |\nabla(uv_\varepsilon)| \,\mathrm{d}x \\ &\leq \liminf_{\varepsilon\to 0} \left(\int_\Omega |\nabla u||v_\varepsilon| \,\mathrm{d}x + \int_\Omega |\nabla v_\varepsilon||u| \,\mathrm{d}x \right) \\ &\leq \liminf_{\varepsilon\to 0} \left(\int_\Omega |\nabla u| v_\varepsilon \,\mathrm{d}x + \|u\|_{L^\infty} \mathrm{TV}(v_\varepsilon) \right).\end{aligned}$$

The second term converges to $\|u\|_{L^\infty} \mathrm{TV}(v)$ as $\varepsilon \to 0$ and the first term can be estimated with the reverse Fatou lemma since $|\nabla u| v_\varepsilon$ is non-negative and bounded by the integrable function $|\nabla u| v$

$$\liminf_{\varepsilon\to 0} \int_\Omega |\nabla u| v_\varepsilon \,\mathrm{d}x \leq \limsup_{\varepsilon\to 0} \int_\Omega |\nabla u| v_\varepsilon \,\mathrm{d}x \leq \int_\Omega \limsup_{\varepsilon\to 0} |\nabla u| v_\varepsilon \,\mathrm{d}x = \int_\Omega |\nabla u| v \,\mathrm{d}x.$$

This proves the statement. □

We are now ready to prove the compactness result for the perimeter.

Theorem 4.4 (Compactness)**.** *Let $\rho_0, \rho_1 \in C(\overline{\Omega})$ such that $\operatorname{ess\,inf}_\Omega(\rho_0 + \rho_1) > 0$. Then, for any sequence of positive numbers $(\varepsilon_k)_{k\in\mathbb{N}}$ with $\varepsilon_k \to 0$ and a collection of sets A_k with $\limsup_{k\to\infty} \mathrm{Per}_{\varepsilon_k}(A_k; \boldsymbol{\rho}) < \infty$, we have that up to a subsequence (not relabeled)*

$$\chi_{A_k} \to \chi_A \text{ in } L^1(\Omega) \quad \text{and} \quad \mathrm{Per}(A) < \infty.$$

Remark 4.5. The difference to the compactness result, proved in [14] is that here—and, in fact, for all results in this chapter—we assume the densities to be continuous on $\overline{\Omega}$. This allows us to drop the assumption that $\rho_0, \rho_1 \in L^\infty(\Omega)$, since the densities are trivially bounded, and although continuity on $\overline{\Omega}$ does not imply bounded variation of the densities, we can also drop the assumption that $\rho_0, \rho_1 \in BV(\Omega)$ by employing an approximation argument for the open superlevel sets

$$\{\rho_i > t\} \coloneqq \{x \in \Omega : \rho_i(x) > t\}$$

of the densities.

Proof of Theorem 4.4. We begin by defining the two sequences of functions $(u_k)_{k\in\mathbb{N}}$ and $(v_k)_{k\in\mathbb{N}}$ by $u_k, v_k : \mathbb{R}^d \to [0,1]$,

$$u_k \coloneqq \max\left\{1 - \frac{d(x, A_k^1)}{\varepsilon_k}, 0\right\}, \quad v_k \coloneqq \min\left\{\frac{d(x, (A_k^c)^1)}{\varepsilon_k}, 1\right\}.$$

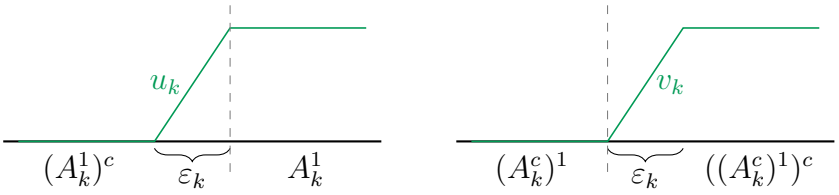

Figure 4.1: A 1-dimensional sketch of the functions u_k and v_k.

As these functions are constant except for a ε_k-strip outside A_k^1 (and $(A_k^c)^1$ respectively) and the gradient of the distance function has norm 1 almost everywhere outside the 0-level set (see, e.g., [26, Theorem 3.14]), we have

$$|\nabla u_k| = \frac{1}{\varepsilon_k}\chi_{\{x\in\mathbb{R}^d:0<d(x,A_k^1)<\varepsilon_k\}} \quad \text{and} \quad |\nabla v_k| = \frac{1}{\varepsilon_k}\chi_{\{x\in\mathbb{R}^d:0<d(x,(A_k^c)^1)<\varepsilon_k\}} \quad \mathcal{L}^d\text{-a.e.}$$

It is easy to see that

$$\rho_0(\{x\in\Omega : 0<d(x,A_k^1)<\varepsilon_k\}) \le \rho_0(\{x\in\Omega\setminus A_k^1 : d(x,A_k^1)<\varepsilon_k\}),$$

and with the Lebesgue density theorem

$$\begin{aligned}\rho_1(\{x\in\Omega : 0<d(x,(A_k^c)^1)<\varepsilon_k\}) &\le \rho_1(\{x\in((A_k^c)^1)^c : d(x,(A_k^c)^1)<\varepsilon_k\})\\ &= \rho_1(\{x\in A^1 : d(x,(A_k^c)^1)<\varepsilon_k\}),\end{aligned}$$

which by Lemma 4.2 implies

$$\int_\Omega |\nabla u_k|\rho_0\,\mathrm{d}x \le \mathrm{Per}_k^0(A_k;\boldsymbol{\rho}) \quad \text{and} \quad \int_\Omega |\nabla v_k|\rho_1\,\mathrm{d}x \le \mathrm{Per}_k^1(A_k;\boldsymbol{\rho}). \tag{4.7}$$

By the hypothesis $\operatorname{ess\,inf}_\Omega(\rho_0+\rho_1)>0$, we have $\rho_0(x)+\rho_1(x)>c_\rho$ for all $x\in\Omega$ and for some constant $c_\rho>0$. Hence, for $0<\tilde{\delta}<c_\rho$ the open sets

$$\{\rho_i>\tilde{\delta}\},\quad i=0,1$$

cover Ω, i.e.,

$$\Omega = \{\rho_0>\tilde{\delta}\}\cup\{\rho_1>\tilde{\delta}\}.$$

We want to smoothly approximate these sets to obtain a covering of Ω by sets of finite perimeter. By the Tietze extension theorem (see, e.g., [39, Theorem 35.1]) there exist continuous extensions $\tilde{\rho}_i:\widetilde{\Omega}\to\mathbb{R}$ of ρ_i on the "thickened" set $\widetilde{\Omega}:=\{x\in\mathbb{R}^d : d(x,\Omega)<1\}$. Since open sets can be approximated from the inside by sets with smooth boundary in the Hausdorff distance (see, e.g., [21, Proposition 8.2.1]), we find sets

$$\widetilde{\Omega}^i \subset \{x\in\widetilde{\Omega} : \tilde{\rho}_i(x)>\delta\},\quad i=0,1$$

with smooth boundary which, by continuity of $\tilde{\rho}_i$ and the fact the boundary of $\widetilde{\Omega}$ has positive distance to Ω, can be chosen in a way that $\Omega\subset\widetilde{\Omega}^0\cup\widetilde{\Omega}^1$. Then the sets $\Omega^i:=\widetilde{\Omega}^i\cap\Omega$

have finite perimeter in Ω and additionally, we have $u_k, v_k \in W^{1,1}(\Omega)$ since $0 \leq u_k, v_k \leq 1$ and $|\nabla u_k|, |\nabla v_k| \leq 1/\varepsilon_k$. Therefore, we may apply Lemma 4.3 to $u_k \chi_{\Omega^0}$:

$$\begin{aligned} \mathrm{TV}(u_k \chi_{\Omega^0}) &\leq \int_\Omega |\nabla u_k| \chi_{\Omega^0} \,\mathrm{d}x + \|u_k\|_{L^\infty} \mathrm{TV}(\chi_{\Omega^0}) \\ &\leq \frac{1}{\delta} \int_\Omega |\nabla u_k| \rho_0 \,\mathrm{d}x + \mathrm{Per}(\Omega^0) \end{aligned}$$

Together with (4.7) this implies

$$\limsup_{k\to\infty} \mathrm{TV}(u_k \chi_{\Omega^0}) \leq \frac{1}{\delta} \limsup_{k\to\infty} \mathrm{Per}_k^0(A_k; \boldsymbol{\rho}) + \mathrm{Per}(\Omega^0) < \infty, \tag{4.8}$$

and it follows with compactness in BV (Theorem 2.23) that, up to a subsequence, $u_k \chi_{\Omega^0} \xrightarrow{L^1(\Omega)} u \in BV(\Omega)$. By a similar argument we have $v_k \chi_{\Omega^1} \xrightarrow{L^1(\Omega)} v \in BV(\Omega)$. Furthermore, the following computation shows that χ_{A_k} and u_k have the same limit in $L^1(\Omega^0)$

$$\begin{aligned} \delta \int_{\Omega^0} |u_k - \chi_{A_k}| \,\mathrm{d}x &= \delta \int_{\Omega^0} u_k - \chi_{A_k^1} \,\mathrm{d}x \\ &= \delta \int_{\Omega^0} \left(1 - \frac{d(x, A_k^1)}{\varepsilon_k}\right) \chi_{\{x\in\Omega : 0 < d(x,A_k^1) < \varepsilon_k\}} \,\mathrm{d}x \\ &\leq \int_{\Omega^0} \chi_{\{x\in\Omega : 0 < d(x,A_k^1) < \varepsilon_k\}} \rho_0 \,\mathrm{d}x \\ &\leq \varepsilon_k \mathrm{Per}_{\varepsilon_k}^0(A_k; \boldsymbol{\rho}) \longrightarrow 0 \quad \text{as } k \to \infty, \end{aligned} \tag{4.9}$$

i.e., $\chi_{A_k} \to u$ in $L^1(\Omega^0)$, and similarly we get $\chi_{A_k} \to v$ in $L^1(\Omega^1)$. Up to a subsequence, we have pointwise convergence $\mathcal{L}^d$-almost everywhere, u and v map to $\{0, 1\}$ $\mathcal{L}^d$-almost everywhere and thus, $u = \chi_U$ and $v = \chi_V$ for sets $U \subset \Omega^0$ and $V \subset \Omega^1$. Define $A := U \cup V$. As both functions χ_U and χ_V are the limit of χ_{A_k}, they necessarily coincide on the set $\Omega^0 \cap \Omega^1$ and therefore,

$$\int_\Omega |\chi_{A_k} - \chi_A| \,\mathrm{d}x \leq \int_{\Omega^0} |\chi_{A_k} - \chi_U| \,\mathrm{d}x + \int_{\Omega^1} |\chi_{A_k} - \chi_V| \,\mathrm{d}x \longrightarrow 0 \quad \text{as } k \to \infty, \tag{4.10}$$

as claimed. The only thing left to show is that A has finite perimeter. Using the lower semicontinuity of the total variation, we see that U and V have finite perimeter

$$\begin{aligned} \mathrm{Per}(U) = \mathrm{TV}(\chi_U) &\leq \liminf_{k\to\infty} \mathrm{TV}(u_k \chi_{\Omega^0}) < \infty, \\ \mathrm{Per}(V) = \mathrm{TV}(\chi_V) &\leq \liminf_{k\to\infty} \mathrm{TV}(v_k \chi_{\Omega^1}) < \infty, \end{aligned}$$

implying the same for $A = U \cup V$. □

4.1.2 Γ-convergence

The Liminf-inequality

We turn to the proof of the Γ-convergence result for the perimeter, and begin with the liminf-inequality, whose proof is greatly simplified by the continuity assumption for the densities. Again, an approximation argument is needed, since we do not require the densities to be BV-functions.

Theorem 4.6 (Liminf-inequality for the perimeter)**.** *Let $\rho_0, \rho_1 \in C(\overline{\Omega})$ and let $(\varepsilon_k)_{k\in\mathbb{N}}$ be a sequence of positive numbers with $\varepsilon_k \to 0$. Consider a sequence of sets $(A_k)_{k\in\mathbb{N}}$ in Ω with $\chi_{A_k} \to \chi_A$ in $L^1(\Omega)$. Then,*

$$\mathrm{Per}(A;\boldsymbol{\rho}) \leq \liminf_{k\to\infty} \mathrm{Per}_{\varepsilon_k}(A_k;\boldsymbol{\rho}).$$

Proof. We focus on showing the claim for the outer perimeter:

$$\liminf_{k\to\infty} \mathrm{Per}^0_{\varepsilon_k}(A_k;\boldsymbol{\rho}) \geq \int_{\partial^* A\cap\Omega} \rho_0 \,\mathrm{d}\mathcal{H}^{d-1}$$

An analogous argument will apply to the inner perimeter. Let the sequence $(u_k)_{k\in\mathbb{N}}$ and the constant c_ρ be defined as in the proof of Theorem 4.4. Then, with the same reasoning as before, we have

$$\mathrm{Per}^0_{\varepsilon_k}(A_k;\boldsymbol{\rho}) = \frac{1}{\varepsilon_k}\rho_0(\{x \in (A_k^1)^c : d(x, A_k^1) < \varepsilon_k\}) \geq \int_\Omega |\nabla u_k|\rho_0 \,\mathrm{d}x. \tag{4.11}$$

Let $\delta_n \to 0$ be a decreasing sequence with $0 < \delta_n < c_\rho$ for all $n \in \mathbb{N}$, and define the sets

$$\Omega^i_{\delta_n} := \{\rho_i > \delta_n\},$$

which are open, by the continuity of ρ_i. Hence, by [21, Proposition 8.2.1], for every $n \in \mathbb{N}$ there exists a sequence of open sets $(\Omega^i_{\delta_n})_l$ with smooth boundary such that $(\Omega^i_{\delta_n})_l \subset (\Omega^i_{\delta_n})_{l+1}$ for all $l \in \mathbb{N}$ and

$$\bigcup_{l=1}^{\infty} (\Omega^i_{\delta_n})_l = \Omega^i_{\delta_n}. \tag{4.12}$$

By (4.8), the sequence $(u_k)_{k\in\mathbb{N}}$ is uniformly bounded in $BV((\Omega^0_{\delta_n})_l)$ for every $l, n \in \mathbb{N}$, and it follows by compactness in BV that, up to a subsequence, Du_k weakly-* converges in $(\Omega^0_{\delta_n})_l$. We identify the weak-* limit of Du_k as $D\chi_A$ since, by hypothesis and the computation in (4.9), we have $u_k \to \chi_A$ in $L^1(\Omega^0_\delta)$. Therefore, after taking the liminf on both sides of (4.11), we may apply Reshetnyak's lower semi-continuity theorem (see, e.g., [44, Theorem 1.7])

$$\liminf_{k\to\infty} \mathrm{Per}^0_{\varepsilon_k}(A_k;\boldsymbol{\rho}) \geq \liminf_{k\to\infty} \int_{(\Omega^0_{\delta_n})_l} |\nabla u_k|\rho_0 \,\mathrm{d}x \geq \int_{(\Omega^0_{\delta_n})_l} \rho_0 \,\mathrm{d}|D\chi_A|. \tag{4.13}$$

Since this holds for all $l \in \mathbb{N}$, by (4.12) and the continuity of measure from below (see, e.g., [32, Theorem 1.4.9]) we have

$$\liminf_{k\to\infty} \mathrm{Per}^0_{\varepsilon_k}(A_k;\boldsymbol{\rho}) \geq \lim_{l\to\infty} \int_{(\Omega^0_{\delta_n})_l} \rho_0 \,\mathrm{d}|D\chi_A| = \int_{\Omega^0_{\delta_n}} \rho_0 \,\mathrm{d}|D\chi_A|,$$

and similarly for $n \in \mathbb{N}$

$$\liminf_{k\to\infty} \mathrm{Per}^0_{\varepsilon_k}(A_k;\boldsymbol{\rho}) \geq \lim_{n\to\infty} \int_{\Omega^0_{\delta_n}} \rho_0 \,\mathrm{d}|D\chi_A| = \int_{\{\rho_0>0\}} \rho_0 \,\mathrm{d}|D\chi_A| = \int_{\Omega} \rho_0 \,\mathrm{d}|D\chi_A|,$$

by the fact that

$$\Omega^0_{\delta_n} \subset \Omega^0_{\delta_{n+1}} \ \forall n \in \mathbb{N} \quad \text{and} \quad \bigcup_{n=1}^{\infty} \Omega^0_{\delta_n} = \{\rho_0 > 0\}.$$

The statement follows with the structure theorem for sets of finite perimeter (Theorem 2.29)

$$\liminf_{k\to\infty} \mathrm{Per}^0_{\varepsilon_k}(A_k;\boldsymbol{\rho}) \geq \int_{\Omega} \rho_0 \,\mathrm{d}|D\chi_A| = \int_{\partial^* A\cap\Omega} \rho_0 \,\mathrm{d}\mathcal{H}^{d-1}.$$

□

The Limsup-inequality

To prove the limsup-inequality for the perimeter we will employ the density argument presented in Remark 2.17. In the first step we will show that sets of finite perimeter can be approximated by open sets with smooth boundary. To this end, we require the statement, that the level sets of smooth functions are smooth manifolds, which is a corollary of Sard's theorem.

Lemma 4.7 (Sard's Theorem)**.** *Let $\Omega \subset \mathbb{R}^d$ be an open set, let $u \in C^\infty(\Omega)$, and let*

$$\Sigma := \{x \in \Omega : \nabla u(x) = 0\}.$$

Then $\mathcal{L}^1(u(\Sigma)) = 0$.

Proof. See, e.g., [33, Theorem 13.42]. □

Lemma 4.8. *Let $\Omega \subset \mathbb{R}^d$ be an open set and let $u \in C^\infty(\Omega)$. Then for $\mathcal{L}^1$-almost every $t \in \mathbb{R}$ the sets $\{x \in \Omega : u(x) = t\}$ are C^∞-manifolds.*

Proof. For $\Sigma := \{x \in \Omega : \nabla u(x) = 0\}$ it holds by Sard's theorem $\mathcal{L}^1(u(\Sigma)) = 0$. Hence, if $t \in \mathbb{R} \setminus u(\Sigma)$, then by the implicit function theorem (see, e.g., [5, 21.11]) it holds that $\{x \in \Omega : u(x) = t\}$ is a C^∞-manifold. □

Additionally, we will need the following fundamental convergence result for sequences of real numbers:

Lemma 4.9. *If $(a_k)_{k\in\mathbb{N}}$ and $(b_k)_{k\in\mathbb{N}}$ are sequences such that*

(i) $\liminf_{k\to\infty} a_k \geq a$,

(ii) $\liminf_{k\to\infty} b_k \geq b$,

(iii) $\limsup_{k\to\infty}(a_k + b_k) \leq a + b$

for some $a, b \in \mathbb{R}$. Then $(a_k)_{k\in\mathbb{N}}$ converges to a and $(b_k)_{k\in\mathbb{N}}$ converges to b.

Proof. By writing $a_k = (a_k + b_k) + (-b_k)$ and with (ii) and (iii) one has

$$\limsup_{k\to\infty} a_k \leq \limsup_{k\to\infty}(a_k + b_k) + \limsup_{k\to\infty} -b_k \leq a + b - \liminf_{k\to\infty} b_k \leq a + b - b = a.$$

It follows with (i) that

$$\lim_{k\to\infty} a_k = a.$$

Swapping the roles of a and b proves $b_k \to b$. □

When considering a set $A \subset \mathbb{R}^d$ of finite perimeter on $\mathbb{R}^d$, one can directly approximate A by smooth sets such that the perimeters converge (see, e.g., [33, Theorem 13.46]), but since we consider the perimeter on Ω, our proof will require a local result, following the ideas of [1, Remark 3.43]. We denote by

$$\mathrm{Per}(A; \boldsymbol{\rho}, S) \coloneqq \int_{\partial^* A \cap S} \rho_0 + \rho_1 \, \mathrm{d}\mathcal{H}^{d-1}$$

the perimeter on an arbitrary Borel set $S \subset \mathbb{R}^d$.

Lemma 4.10 (Density of smooth sets)**.** *Let $\rho_0, \rho_1 \in C(\overline{\Omega})$ and let $A \subset \Omega$ be a subset satisfying $\mathrm{Per}(A; \boldsymbol{\rho}) < \infty$. Then there exists a sequence $(A_k)_{k\in\mathbb{N}}$ of open sets in $\mathbb{R}^d$ with smooth boundary such that*

$$A_k \cap \Omega \to A \text{ in } L^1(\Omega) \quad \text{and} \quad \mathrm{Per}(A_k; \boldsymbol{\rho}, \overline{\Omega}) \to \mathrm{Per}(A; \boldsymbol{\rho}, \Omega).$$

Since, to avoid boundary issues, we do not approximate the set A by smooth sets in Ω, but by smooth subsets of $\mathbb{R}^d$ intersected with Ω (see Figure 4.2), we need to extend the set A to a set F such that the perimeter of F on $\partial\Omega$ vanishes. This is possible on extension domains (see [1, Definition 3.20]) and by [1, Proposition 3.21] sets with compact Lipschitz boundary are extension domains.

Proof of Lemma 4.10. Since Ω is a bounded set with Lipschitz boundary, it is an extension domain and there exists $u \in BV(\mathbb{R}^d)$ with compact support which is an extension of χ_A such that $\mathrm{TV}(u; \partial\Omega) = 0$ and $0 \leq u \leq 1$ (see [1, Remark 3.43]). Since the densities ρ_0, ρ_1 are bounded, this implies $\mathrm{TV}(u; \boldsymbol{\rho}, \partial\Omega) = 0$, and by the coarea formula we may choose

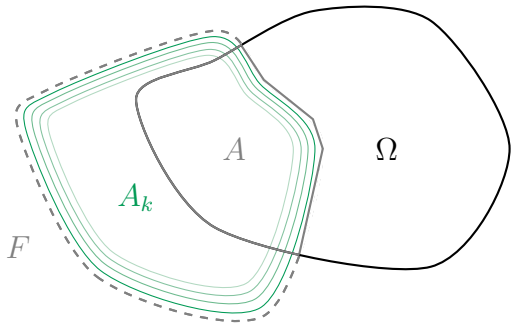

Figure 4.2: The set A, its extension F on $\mathbb{R}^d$ and the smooth approximation A_k of F.

$s \in (0,1)$ such that the superlevel set $F := \{u > s\}$ satisfies $\mathrm{Per}(F; \boldsymbol{\rho}, \partial\Omega) = 0$. For $\varepsilon > 0$ let $u_k := \phi_{\varepsilon_k} * \chi_F$ be a mollification of χ_F. Then, by [33, Remark 13.11] we know that $u_k \to \chi_F$ in $L^1(\mathbb{R}^d)$ and

$$\lim_{k\to\infty} \int_{\mathbb{R}^d} |\nabla u_k| \,\mathrm{d}x = \int_{\mathbb{R}^d} \mathrm{d}|D\chi_F|(x).$$

By careful inspection of the proof of [33, Theorem 13.9] one sees that this continues to hold in the weighted case with $\rho_0 + \rho_1 \in C(\overline{\Omega})$:

$$\lim_{k\to\infty} \int_{\mathbb{R}^d} |\nabla u_k|(\rho_0 + \rho_1) \,\mathrm{d}x = \int_{\mathbb{R}^d} \rho_0 + \rho_1 \,\mathrm{d}|D\chi_F|(x) = \mathrm{Per}(F; \boldsymbol{\rho}, \mathbb{R}^d) \tag{4.14}$$

For $t \in \mathbb{R}$ we define

$$F_k^t := \{x \in \mathbb{R}^d : u_k > t\}$$

and it follows from (4.14), the coarea formula, the fact that $0 \le u_{\varepsilon_k} \le 1$ and Fatou's lemma,

$$\begin{aligned}\mathrm{Per}(F; \boldsymbol{\rho}, \mathbb{R}^d) &= \lim_{k\to\infty} \int_{\mathbb{R}^d} |\nabla u_k|(\rho_0 + \rho_1) \,\mathrm{d}x = \lim_{k\to\infty} \int_0^1 \mathrm{Per}(F_k^t; \boldsymbol{\rho}, \mathbb{R}^d) \,\mathrm{d}t \\ &\ge \int_0^1 \liminf_{k\to\infty} \mathrm{Per}(F_k^t; \boldsymbol{\rho}, \mathbb{R}^d) \,\mathrm{d}t,\end{aligned}$$

which implies

$$\liminf_{k\to\infty} \mathrm{Per}(F_k^t; \boldsymbol{\rho}, \mathbb{R}^d) \le \mathrm{Per}(F; \boldsymbol{\rho}, \mathbb{R}^d)$$

for $\mathcal{L}^1$-almost every $t \in [0,1]$. Since $u_k \in C^\infty(\mathbb{R}^d)$, by Lemma 4.8 it holds that for all $k \in \mathbb{N}$ the the sets

$$\{x \in \mathbb{R}^d : u_k = t\},$$

which coincide with the boundaries of F_k^t, are C^∞-manifolds for $\mathcal{L}^1$-almost every $t \in [0,1]$. Hence, there exists a specific $t \in (0,1)$ such that ∂F_k^t is a C^∞-manifold for all $k \in \mathbb{N}$.

We define $(A_k)_{k\in\mathbb{N}}$ as a subsequence of $(F_k^t)_{k\in\mathbb{N}}$ for which the perimeter converges to the smallest cluster point, i.e.,

$$\lim_{k\to\infty} \operatorname{Per}(A_k;\boldsymbol{\rho},\mathbb{R}^d) = \liminf_{k\to\infty} \operatorname{Per}(F_k^t;\boldsymbol{\rho},\mathbb{R}^d) \leq \operatorname{Per}(F;\boldsymbol{\rho},\mathbb{R}^d). \tag{4.15}$$

Then we have $A_k \to F$ in $L^1(\mathbb{R}^d)$. To see this, note that

$$u_k - \chi_F > t, \quad x \in A_k \setminus F \quad \text{and} \quad \chi_F - u_k > 1-t, \quad x \in F \setminus A_k$$

and follow

$$\begin{aligned}
\int_{\mathbb{R}^d} |\chi_{A_k} - \chi_F| \,\mathrm{d}x &= \int_{A_k\setminus F} \mathrm{d}x + \int_{F\setminus A_k} \mathrm{d}x \\
&\leq \frac{1}{t}\int_{A_k\setminus F} u_k - \chi_F \,\mathrm{d}x + \frac{1}{1-t}\int_{F\setminus A_k} \chi_F - u_k \,\mathrm{d}x \\
&\leq \max\left\{\frac{1}{t}, \frac{1}{1-t}\right\}\left(\int_{A_k\setminus F} |u_k - \chi_F| \,\mathrm{d}x + \int_{F\setminus A_k} |\chi_F - u_k| \,\mathrm{d}x\right) \\
&\leq \max\left\{\frac{1}{t}, \frac{1}{1-t}\right\}\int_{\mathbb{R}^d} |u_k - \chi_F| \,\mathrm{d}x \longrightarrow 0.
\end{aligned} \tag{4.16}$$

Since $A = F \cap \Omega$, we have

$$A_k \cap \Omega \to A \text{ in } L^1(\Omega),$$

proving the first part of the lemma.

To proof the second claim, note that by lower semi-continuity (4.16) implies

$$\liminf_{k\to\infty} \operatorname{Per}(A_k;\boldsymbol{\rho},\mathbb{R}^d) \geq \operatorname{Per}(F;\boldsymbol{\rho},\mathbb{R}^d),$$

which combined with (4.15) gives convergence of the global perimeter

$$\lim_{k\to\infty} \operatorname{Per}(A_k;\boldsymbol{\rho},\mathbb{R}^d) = \operatorname{Per}(F;\boldsymbol{\rho},\mathbb{R}^d). \tag{4.17}$$

Again, by lower semi-continuity and (4.16), we also have

$$\liminf_{k\to\infty} \operatorname{Per}(A_k;\boldsymbol{\rho},\overline{\Omega}) \geq \operatorname{Per}(F;\boldsymbol{\rho},\overline{\Omega}) \quad \text{and} \quad \liminf_{k\to\infty} \operatorname{Per}(A_k;\boldsymbol{\rho},\mathbb{R}^d\setminus\overline{\Omega}) \geq \operatorname{Per}(F;\boldsymbol{\rho},\mathbb{R}^d\setminus\overline{\Omega}),$$

which together with (4.17) and Lemma 4.9 implies

$$\lim_{k\to\infty} \operatorname{Per}(A_k;\boldsymbol{\rho},\overline{\Omega}) = \operatorname{Per}(F;\boldsymbol{\rho},\overline{\Omega})$$

and since $\operatorname{Per}(F;\boldsymbol{\rho},\partial\Omega) = 0$, we conclude

$$\lim_{k\to\infty} \operatorname{Per}(A_k;\boldsymbol{\rho},\overline{\Omega}) = \operatorname{Per}(F;\boldsymbol{\rho},\Omega) = \operatorname{Per}(A;\boldsymbol{\rho},\Omega).$$

□

It follows immediately, with A, A_k defined as in the previous theorem, that

$$\limsup_{k\to\infty} \mathrm{Per}(A_k;\boldsymbol{\rho},\Omega) \leq \mathrm{Per}(A;\boldsymbol{\rho},\Omega),$$

and by lower semi-continuity

$$\liminf_{k\to\infty} \mathrm{Per}(A_k;\boldsymbol{\rho},\Omega) \geq \mathrm{Per}(A;\boldsymbol{\rho},\Omega).$$

Therefore, $\lim_{k\to\infty} \mathrm{Per}(A_k;\boldsymbol{\rho},\Omega) = \mathrm{Per}(A;\boldsymbol{\rho},\Omega)$ and since, by the previous theorem, sets of finite perimeter in Ω can be approximated by sets with smooth boundary intersected with Ω, we may use the density argument Remark 2.17 to reduce the proof of the limsup-inequality for the perimeter to the fact that, for sets with smooth boundary, the limsup-inequality already holds for a constant recovery sequence.

Theorem 4.11 (Limsup-inequality for the perimeter)**.** *Let $\rho_0, \rho_1 \in C(\overline{\Omega})$ and let $A \subset \mathbb{R}^d$ be a set with smooth boundary. Then hen, for any sequence of positive numbers $\varepsilon \to 0$ the following bound holds*

$$\limsup_{\varepsilon\to 0} \mathrm{Per}_\varepsilon(A;\boldsymbol{\rho}) \leq \mathrm{Per}(A;\boldsymbol{\rho}) = \int_{\partial^* A\cap\Omega} \rho_0 + \rho_1 \,\mathrm{d}\mathcal{H}^{d-1}.$$

While for a smooth set $A \in \mathbb{R}^d$ the set $A \cap \Omega$ might not have smooth boundary, the proof of this theorem relies on the fact that only the boundary of $A \cap \Omega$ lying inside Ω contributes to the perimeter, while the part of the boundary that coincides with $\partial\Omega$ does not (see Figure 4.3).

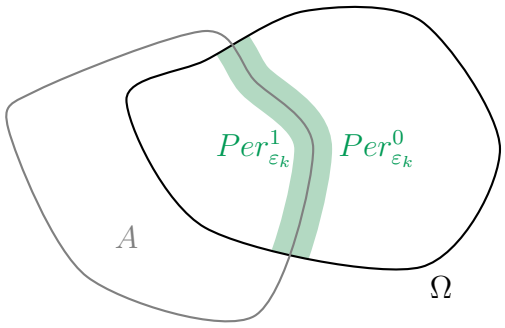

Figure 4.3: The outer and inner perimeter of a set A.

Also, in the proof we will employ the Taylor expansion representation of the determinant for matrices of the form $I + \varepsilon M$

$$\det(I + \varepsilon M) = 1 + \varepsilon \mathrm{tr}(M) + O(\varepsilon^2). \tag{4.18}$$

To verify this representation, let $e_1, \ldots, e_n$ be the standard basis vectors and let $M_1, \ldots, M_n$ denote the columns of M. Then, successively using the multi-linearity of

the determinant, we have

$$\begin{aligned}\det(I+\varepsilon M) &= \det(e_1+\varepsilon M_1|\dots|e_n+\varepsilon M_n)\\ &= \det(e_1|e_2+\varepsilon M_2|\dots|e_n+\varepsilon M_n)+\det(\varepsilon M_1|e_2+\varepsilon M_2|\dots|e_n+\varepsilon M_n)\\ &\vdots\\ &= \det(I)+\sum_{i=1}^{n}\det(e_1|\dots|e_{i-1}|\varepsilon M_i|e_{i+1}|\dots|e_n)+O(\varepsilon^2),\end{aligned}$$

where all remaining terms are determinants of matrices with at least two columns of the form εM_i, and therefore are of the order $O(\varepsilon^2)$. Expanding the determinant along the i-th row, we see

$$\det(e_1|\dots|e_{i-1}|\varepsilon M_i|e_{i+1}|\dots|e_n)=\varepsilon M_{i,i},$$

which shows

$$\det(I+\varepsilon M)=\det(I)+\sum_{i=1}^{n}\varepsilon M_{i,i}+O(\varepsilon^2)=1+\varepsilon\operatorname{tr}(M)+O(\varepsilon^2).$$

Proof of Theorem 4.11. We will show that the limsup-inequality holds for the outer perimeter

$$\limsup_{\varepsilon\to 0}\operatorname{Per}^0_\varepsilon(A;\boldsymbol{\rho})\le\int_{\partial^* A\cap\Omega}\rho_0\,\mathrm{d}\mathcal{H}^{d-1}.$$

The argument for $\operatorname{Per}^1_\varepsilon(\,\cdot\,;\boldsymbol{\rho})$ will be the same. Since A has a smooth boundary, there exists $\varepsilon_0>0$ such that for all $\varepsilon<\varepsilon_0$ the map

$$T_\varepsilon(y,t):(\partial A^1\cap\Omega)\times[0,1]\to\{x\in(A^1)^c:d(x,A^1\cap\Omega)<\varepsilon\},\ T_\varepsilon(y,t)=y+\varepsilon t n(y)$$

is a bijection, where $n(y)$ is the unit normal of ∂A^1 in y. Applying this change of variables allows us to write

$$\operatorname{Per}^0_\varepsilon(A;\boldsymbol{\rho})=\frac{1}{\varepsilon}\int_{\{x\in(A^1)^c:d(x,A^1\cap\Omega)<\varepsilon\}}\rho_0\,\mathrm{d}x=\frac{1}{\varepsilon}\int_{\partial A^1\cap\Omega}\int_0^1\rho_0(T_\varepsilon(y,t))|\det(DT_\varepsilon)|\,\mathrm{d}t\,\mathrm{d}\mathcal{H}^{d-1}.$$

Considering a local coordinate system in y with an orthonormal basis where the basis vectors $\{e_1,e_2,\dots,e_{d-1}\}$ span the tangent space $T_y(\partial A^1)$ and e_d coincides with $n(y)$, the Jacobian of T_ε has the following block matrix form

$$DT_\varepsilon=\begin{bmatrix} & & & 0\\ & I+\varepsilon t D_y n(y) & & \vdots\\ & & & 0\\ 0 & \dots & 0 & \varepsilon\end{bmatrix}$$

and we have $\det(DT_\varepsilon)=\varepsilon\cdot\det(I+\varepsilon t D_y n(y))$. Continuing, we use the Taylor expansion for the determinant (4.18) to find

$$\det(I+\varepsilon t D_y n(y))=1+\varepsilon\cdot\operatorname{tr}(tD_y n(y))+O(\varepsilon^2),$$

which uniformly converges to 1 for $\varepsilon \to 0$, since the derivative of the unit normal of the smooth manifold ∂A^1 is bounded. The statement follows with the reverse Fatou lemma, the continuity of ρ_0 and the fact that for open sets with smooth boundary one has $\partial A^1 = \partial A = \partial^* A$:

$$\begin{aligned}
\limsup_{\varepsilon\to 0} \operatorname{Per}_\varepsilon^0(A;\boldsymbol{\rho}) &= \limsup_{\varepsilon\to 0} \frac{1}{\varepsilon} \int_{\partial A^1\cap\Omega} \int_0^1 \rho_0(T_\varepsilon(y,t))|\det(DT_\varepsilon)|\,\mathrm{d}t\,\mathrm{d}\mathcal{H}^{d-1} \\
&= \limsup_{\varepsilon\to 0} \frac{1}{\varepsilon} \int_{\partial A^1\cap\Omega} \int_0^1 \rho_0(y+\varepsilon t n(y))\cdot|\varepsilon(1+\varepsilon\cdot\operatorname{tr}(tD_y n(y))+O(\varepsilon^2)|\,\mathrm{d}t\,\mathrm{d}\mathcal{H}^{d-1} \\
&\leq \int_{\partial A^1\cap\Omega} \int_0^1 \limsup_{\varepsilon\to 0} \rho_0(y+\varepsilon t n(y))\cdot|(1+\varepsilon\cdot\operatorname{tr}(tD_y n(y))+O(\varepsilon^2)|\,\mathrm{d}t\,\mathrm{d}\mathcal{H}^{d-1} \\
&= \int_{\partial A^1\cap\Omega} \int_0^1 \rho_0(y)\cdot 1\,\mathrm{d}t\,\mathrm{d}\mathcal{H}^{d-1} = \int_{\partial A^1\cap\Omega} \rho_0(y)\,\mathrm{d}\mathcal{H}^{d-1} \\
&= \int_{\partial^* A\cap\Omega} \rho_0(y)\,\mathrm{d}\mathcal{H}^{d-1}
\end{aligned}$$

Therefore, the limsup-inequality holds for sets with smooth boundary with the constant recovery sequence. □

By our density argument, this concludes the proof of the Γ-convergence result for the non-local perimeter (Theorem 4.1).

4.2 Asymptotics of the Total Variation

We use the insight of the previous section on the perimeter to study the asymptotics of the non-local total variation. Again, we begin by proving a compactness result for sequences of functions with bounded total variation and subsequently we will show the following Γ-convergence result:

Theorem 4.12 (Γ-convergence of the non-local total variation)**.** *Let $\rho_0, \rho_1 \in C(\overline{\Omega})$. Then it holds that*

$$\mathrm{TV}_\varepsilon(\,\cdot\,;\boldsymbol{\rho}) \xrightarrow{\Gamma} \mathrm{TV}(\,\cdot\,;\boldsymbol{\rho}) \tag{4.19}$$

in the strong $L^1(\Omega)$ topology.

In Chapter 5, these results will provide insight into the minimizers of (1.5) in the limit as the adversarial budget tends to zero.

4.2.1 A Compactness Result

We state the compactness property for the total variation, mirroring the result in Theorem 4.4 for sequences of functions.

Theorem 4.13 (Compactness). *Let $\rho_0, \rho_1 \in C(\overline{\Omega})$ and let $(u_k)_{k\in\mathbb{N}} \subset L^1(\Omega)$ be a sequence of functions such that $\limsup_{k\to\infty} \mathrm{TV}_{\varepsilon_k}(u_k;\boldsymbol{\rho}) < \infty$ for any sequence $(\varepsilon_k)_{k\in\mathbb{N}}$ with $\lim_{k\to\infty}\varepsilon_k = 0$. Then there exists a subsequence $(u_k)_{k\in\mathbb{N}}$ (not relabeled) and a function $u \in L^1(\Omega)$ such that*

$$u_k \to u \text{ in } L^1(\Omega) \quad \text{and} \quad \mathrm{TV}(u;\boldsymbol{\rho}) < \infty.$$

The proof of this theorem is based on the ideas presented in [18, Corollary 3.2] and requires the compactness property for the perimeter (Theorem 4.4). Furthermore, the proof will use the fact that for a sequence of subsets of an bounded interval, there exists a subsequence of sets with non-empty intersection, if the measures of the sets admit an uniform lower bound.

Lemma 4.14. *For $a, b \in \mathbb{R}$ with $a < b$, let $(A_k)_{k\in\mathbb{N}} \subset [a,b]$ be a sequence of sets such that $\liminf_{k\to\infty} \mathcal{L}^d(A_k) \geq c > 0$. Then there exists a subsequence $(A_{k_l})_{l\in\mathbb{N}}$ such that*

$$\emptyset \neq \bigcap_{l=0}^{\infty} A_{k_l}.$$

Proof. Let $B_n := \bigcup_{k\geq n} A_k$. Then, B_n is a decreasing sequence with $\mathcal{L}^d(B_n) \geq \mathcal{L}^d(A_n)$ for all $n \in \mathbb{N}$ and we have, by the continuity of measure from above (see, e.g., [32, Theorem 1.4.9]),

$$\mathcal{L}^d\left(\bigcap_{n=1}^{\infty} B_n\right) = \lim_{n\to\infty} \mathcal{L}^d(B_n) \geq \lim_{n\to\infty} \mathcal{L}^d(A_n) \geq \liminf_{n\to\infty} \mathcal{L}^d(A_n) \geq c.$$

By definition, $\bigcap_{n=1}^{\infty} B_n$ is the set of points which lie in infinitely many A_k's and therefore, there exists a subsequence $(A_{k_l})_{l\in\mathbb{N}}$ with $\emptyset \neq \bigcap_{l=0}^{\infty} A_{k_l}$. □

We give an outline for the proof of Theorem 4.13: Using the given sequence $(u_k)_{k\in\mathbb{N}}$ to define a sequence of functions $u_k^{1/n}$ depending on n and k, we begin by showing convergence to some $u^{1/n}$ for $k \to \infty$. Next, we will show that $u_k^{1/n}$ converges to the given u_k for $n \to \infty$. Finally, we will show convergence of $u^{1/n} \to u$ and conclude with an iterated limit argument that the limits of $u^{1/n}$ and u_k coincide. This argument will be based on Lemma 4.16, for which we define the notion of uniform convergence for an iterated limit. The proof of the lemma can be found in [5, 14.15 Iterated Limit Theorem].

Definition 4.15 (Uniform convergence for iterated limits). For each $m \in \mathbb{N}$, let $Y_m = (x_{m,n})_{n\in\mathbb{N}}$ be a sequence in $\mathbb{R}$ which converges to y_m. We say that the sequences $\{Y_m : m \in \mathbb{N}\}$ are *uniformly convergent* if, for each $\varepsilon > 0$ there is a natural number $N(\varepsilon)$ such that if $n > N(\varepsilon)$, then $|x_{m,n} - y_m| < \varepsilon$ for all $m \in \mathbb{N}$.

Lemma 4.16 (Iterated limit theorem). *Suppose that the single limits*

$$y_m = \lim_{n\to\infty} x_{m,n}, \quad z_n = \lim_{m\to\infty} x_{m,n}, \quad m, n \in \mathbb{N}$$

exist and that the convergence of one of these collections is uniform. Then both iterated limits exists and are equal.

Proof of Theorem 4.13. For each $k, n \in \mathbb{N}$ define

$$u_k^{1/n} := \sum_{l \in \mathbb{N}} s_l \chi_{\{s_{l+1} \geq u_k > s_l\}}$$

where $s_l \in \left(\frac{1}{n}l, \frac{1}{n}(l+1)\right)$ is chosen in a way that the perimeter of $\{u_k > s_l\}$ are uniformly bounded

$$\mathrm{Per}_{\varepsilon_k}(\{u_k > s_l\}; \boldsymbol{\rho}) \leq n(1 + \limsup_{k \to \infty} \mathrm{TV}_{\varepsilon_k}(u_k; \boldsymbol{\rho})) \quad \forall k \in \mathbb{N}.$$

We proceed to show the existence of such an s_l up to a subsequence in k, from which convergence of $u_k^{1/n}$ for $k \to \infty$ will follow. Fix $l, n \in \mathbb{N}$ and let $A_k \subset \left(\frac{1}{n}l, \frac{1}{n}(l+1)\right)$ be the set of all points $s \in \left(\frac{1}{n}l, \frac{1}{n}(l+1)\right)$ which satisfy

$$\frac{1}{n}\mathrm{Per}_{\varepsilon_k}(\{u_k > s\}; \boldsymbol{\rho}) - 1 \leq \int_{\frac{1}{n}l}^{\frac{1}{n}(l+1)} \mathrm{Per}_{\varepsilon_k}(\{u_k > t\}; \boldsymbol{\rho})\, \mathrm{d}t.$$

Then there exists a $c > 0$ such that

$$\liminf_{k \to \infty} |A_k| \geq c.$$

To see this, define

$$I := \int_{\frac{1}{n}l}^{\frac{1}{n}(l+1)} \mathrm{Per}_{\varepsilon_k}(\{u_k > t\}; \boldsymbol{\rho})\, \mathrm{d}t,$$

and observe that on the relative complement $(A_k)^c := \left(\frac{1}{n}l, \frac{1}{n}(l+1)\right) \setminus A_k$ the inequality

$$\frac{1}{n}\mathrm{Per}_{\varepsilon_k}(\{u_k > t\}; \boldsymbol{\rho}) - 1 > I \tag{4.20}$$

holds. Integrating (4.20) over $(A_k)^c$ yields

$$\int_{(A_k)^c} \mathrm{Per}_{\varepsilon_k}(\{u_k > t\}; \boldsymbol{\rho})\, \mathrm{d}t > |(A_k)^c| \cdot n(I+1),$$

and since $\mathrm{Per}_{\varepsilon_k}(\{u_k > t\}; \boldsymbol{\rho}) \geq 0$ it follows

$$\begin{aligned} I &= \int_{A_k} \mathrm{Per}_{\varepsilon_k}(\{u_k > t\}; \boldsymbol{\rho})\, \mathrm{d}t + \int_{(A_k)^c} \mathrm{Per}_{\varepsilon_k}(\{u_k > t\}; \boldsymbol{\rho})\, \mathrm{d}t \\ &> 0 + |(A_k)^c| \cdot n(I+1) \\ &= \left(\frac{1}{n} - |A_k|\right) \cdot n(I+1). \end{aligned}$$

Solving for $|A_k|$ and taking the liminf, we have

$$\begin{aligned}\liminf_{k\to\infty}|A_k| &\geq \liminf_{k\to\infty}\frac{1}{n(I+1)}\\ &\geq \liminf_{k\to\infty}\frac{1}{n(\int_{\mathbb{R}}\operatorname{Per}_{\varepsilon_k}(\{u_k>t\};\boldsymbol{\rho})\,\mathrm{d}t+1)}\\ &= \frac{1}{n\left(\limsup_{k\to\infty}\operatorname{TV}_{\varepsilon_k}(u_k;\boldsymbol{\rho})+1\right)} =: c>0,\end{aligned}$$

by the assumption that $\limsup_{k\to\infty}\operatorname{TV}_{\varepsilon_k}(u_k;\boldsymbol{\rho})<\infty$. Therefore, Lemma 4.14 implies the existence of a subsequence $(A_k)_{k\in\mathbb{N}}$ with non-empty intersection

$$\emptyset\neq\bigcap_{k=1}^{\infty}A_k\subset\left(\frac{1}{n}l,\frac{1}{n}(l+1)\right),$$

and we may chose an $s_l\in\bigcap_{k=1}^{\infty}A_k$ for which

$$\begin{aligned}\limsup_{k\to\infty}\operatorname{Per}_{\varepsilon_k}(\{u_k>s_l\};\boldsymbol{\rho}) &\leq \limsup_{k\to\infty} n\,(I+1)\\ &\leq \limsup_{k\to\infty} n\left(\int_{\mathbb{R}}\operatorname{Per}_{\varepsilon_k}(\{u_k>t\};\boldsymbol{\rho})\,\mathrm{d}t+1\right)\\ &= n\left(\limsup_{k\to\infty}\operatorname{TV}_{\varepsilon_k}(u_k;\boldsymbol{\rho})+1\right)<\infty,\end{aligned}$$

where the first inequality holds by the definition of A_k. Since $l\in\mathbb{N}$ was arbitrary, this holds for all $l\in\mathbb{N}$ and therefore, the compactness result for sets of finite perimeter (Theorem 4.4) implies that for all $l\in\mathbb{N}$ there exists a subsequence in k such that the sets $\{u_k>s_l\}$ converge in $L^1(\Omega)$. This immediately implies L^1-convergence of the complements $\{u_k>s_{l+1}\}^c$ and hence, for all $l\in\mathbb{N}$ there exist a subsequence such that the sets

$$\{s_{l+1}\geq u_k>s_l\}=\{u_k>s_l\}\cap\{u_k>s_{l+1}\}^c$$

converge for $k\to\infty$. With a diagonal argument we can extract a single subsequence such that we have convergence for all $l\in\mathbb{N}$. Since for $s_l\in[1,\infty)$ the sets $\{s_{l+1}\geq u_k>s_l\}$ are empty, the series defining $u_k^{1/n}$ is actually a finite sum of converging characteristic functions multiplied by weights in $(0,1)$ and therefore, along this subsequence, we have

$$u_k^{1/n}\to u^{1/n}\quad\text{in } L^1(\Omega) \tag{4.21}$$

for some $u^{1/n}\in L^1(\Omega)$.

Next, we show that for fixed $k\in\mathbb{N}$ and $n\to\infty$ we also have convergence of $u_k^{1/n}$. In particular,

$$u_k^{1/n}\to u_k\quad\text{in}\quad L^1(\Omega). \tag{4.22}$$

To see this, note that for $x \in \{s_{l+1} \geq u_k > s_l\}$ we have $u_k(x) > s_l = u_k^{1/n}(x)$ and compute

$$\begin{aligned}
\lim_{n\to\infty} \int_\Omega \left| u_k^{1/n} - u_k \right| \mathrm{d}x &= \lim_{n\to\infty} \int_\Omega u_k - u_k^{1/n} \,\mathrm{d}x \\
&= \lim_{n\to\infty} \int_\Omega u_k - \sum_{l\in\mathbb{N}} s_l \chi_{\{s_{l+1}\geq u_k > s_l\}} \,\mathrm{d}x \\
&\leq \lim_{n\to\infty} \int_\Omega \sum_{l\in\mathbb{N}} s_{l+1}\chi_{\{s_{l+1}\geq u_k > s_l\}} - \sum_{l\in\mathbb{N}} s_l \chi_{\{s_{l+1}\geq u_k > s_l\}} \,\mathrm{d}x \\
&= \lim_{n\to\infty} \int_\Omega \sum_{l\in\mathbb{N}} (s_{l+1} - s_l)\chi_{\{s_{l+1}\geq u_k > s_l\}} \,\mathrm{d}x \\
&\leq \lim_{n\to\infty} \int_\Omega \sum_{l\in\mathbb{N}} \frac{2}{n} \chi_{\{s_{l+1}\geq u_k > s_l\}} \,\mathrm{d}x \\
&\leq \lim_{n\to\infty} \frac{2}{n} |\Omega| = 0.
\end{aligned} \tag{4.23}$$

Here the second inequality holds because $s_l > \frac{1}{n} l$ and $s_{l+1} < \frac{1}{n}(l+2)$; the last inequality holds since the sets $\{s_{l+1} \geq u_k > s_l\}$ are disjoint in l. This proves (4.22).

Furthermore, we have

$$u^{1/n} \to u \quad \text{in} \quad L^1(\Omega) \tag{4.24}$$

for some $u \in L^1(\Omega)$. Since Ω is bounded, it suffices to show convergence in $L^\infty(\Omega)$ which is established by the following

$$\left\| u^{1/n} - u^{1/m} \right\|_{L^\infty} = \left\| \lim_{k\to\infty} (u_k^{1/n} - u_k^{1/m}) \right\|_{L^\infty} \leq \lim_{k\to\infty} \left\| u_k^{1/n} - u_k^{1/m} \right\|_{L^\infty} \leq \frac{2}{\min\{m,n\}}.$$

Here the last inequality holds because for $x \in \{s_{l+1} \geq u_k > s_l\}$ adding the two inequalities

$$u_k(x) - u_k^{1/n}(x) \leq s_{l+1} - s_l \leq \frac{1}{n}(l+2) - \frac{1}{n} l = \frac{2}{n} \quad \text{and} \quad -u_k(x) + u_k^{1/m}(x) < 0$$

gives

$$u_k^{1/m}(x) - u_k^{1/n}(x) \leq \frac{2}{n},$$

and by swapping the roles of m and n we have

$$u_k^{1/n}(x) - u_k^{1/m}(x) \leq \frac{2}{m},$$

which jointly yield

$$\left| u_k^{1/n}(x) - u_k^{1/m}(x) \right| \leq \frac{2}{\min\{m,n\}} \quad \forall x \in \Omega.$$

Finally, by (4.21), (4.24) and (4.22) it holds for the iterated limits

$$\lim_{n\to\infty} \lim_{k\to\infty} \left\| u_k^{1/n} - u \right\|_{L^1} = \lim_{n\to\infty} \left\| u^{1/n} - u \right\|_{L^1} = 0, \tag{4.25}$$

$$\lim_{k\to\infty} \lim_{n\to\infty} \left\| u_k^{1/n} - u \right\|_{L^1} \overset{(*)}{=} \lim_{k\to\infty} \left\| u_k - u \right\|_{L^1}. \tag{4.26}$$

Since, by the computation in (4.23), we have

$$\left\| u_k^{1/n} - u_k \right\|_{L^1} \leq \frac{2}{n} |\Omega|$$

independently from k, the convergence in $(*)$ is uniform in the sense of iterated limits. Hence, the two iterated limits (4.25) and (4.26) coincide (see Lemma 4.16) and we have $u_k \to u$ in $L^1(\Omega)$, as desired. The fact that the limit satisfies $\mathrm{TV}(u;\boldsymbol{\rho}) < \infty$ will follow directly from the liminf-inequality, proved in Theorem 4.19. □

4.2.2 Γ-convergence

We turn to the proof of the Γ-convergence result for the non-local total variation, beginning with the limsup-inequality.

The Limsup-inequality

To prove the limsup-inequality for the non-local total variation, we employ a density argument once more, and give a straightforward proof of the limsup-inequality for smooth functions.

Theorem 4.17 (Limsup-inequality for the total variation)**.** *Let $\rho_0, \rho_1 \in C(\overline{\Omega})$ and let $u \in BV(\Omega)$. Then there exits a recovery sequence $(u_k)_{k\in\mathbb{N}} \subset BV(\Omega)$ with $u_k \to u$ in $L^1(\Omega)$ such that*

$$\limsup_{k\to\infty} \mathrm{TV}_{\varepsilon_k}(u_k;\boldsymbol{\rho}) \leq \mathrm{TV}(u;\boldsymbol{\rho})$$

for any sequence of positive numbers $(\varepsilon_k)_{k\in\mathbb{N}}$ with $\varepsilon_k \to 0$.

Proof. As is the proof of Lemma 4.10, we deduce from [33, Theorem 13.9] that smooth functions are dense in $BV(\Omega)$ with respect to the total variation with weight $\rho_0 + \rho_1$, and therefore, by Remark 2.17, it suffices to show that the limsup-inequality holds for smooth functions with the constant recovery sequence. Let $u \in C^\infty(\Omega)$. Since we can represent every point in $B_\varepsilon(x)$ as $x + \varepsilon' v$ for some unit vector v and $0 \leq \varepsilon' < \varepsilon$, we may write

$$\operatorname*{ess\,sup}_{\tilde{x}\in B_\varepsilon(x)\cap\Omega} u(\tilde{x}) = \operatorname*{ess\,sup}_{\varepsilon' v \in M(x)} u(x + \varepsilon' v),$$

where $M(x) \coloneqq \{\varepsilon' v : 0 \leq \varepsilon' < \varepsilon, \|v\| = 1, x + \varepsilon' v \in \Omega\}$, and compute the essential supremum of the Taylor expansion of u

$$\begin{aligned}
\operatorname*{ess\,sup}_{B_\varepsilon(x)\cap\Omega} u &= \operatorname*{ess\,sup}_{B_\varepsilon(x)\cap\Omega} (u(x) + \varepsilon' v \nabla u(x) + O(\varepsilon^2)) \\
&= u(x) + \operatorname*{ess\,sup}_{M(x)} \varepsilon' v \nabla u(x) + O(\varepsilon^2) \\
&\leq u(x) + \varepsilon |\nabla u(x)| + O(\varepsilon^2).
\end{aligned}$$

Here, the final inequality holds because $v\nabla u(x)$ is maximal when v is parallel to $\nabla u(x)$. With a similar argument we have

$$\operatorname*{ess\,inf}_{B_\varepsilon(x)\cap\Omega} u \geq u(x) - \varepsilon|\nabla u(x)| + O(\varepsilon^2),$$

and it follows with the reverse Fatou lemma

$$\begin{aligned}
\limsup_{k\to\infty} \mathrm{TV}_{\varepsilon_k}(u;\boldsymbol{\rho}) &= \limsup_{k\to\infty} \frac{1}{\varepsilon_k}\Bigg(\int_\Omega \operatorname*{ess\,sup}_{B_{\varepsilon_k}(x)\cap\Omega} u - u(x)\ \mathrm{d}\rho_0(x) \\
&\qquad\qquad + \int_\Omega u(x) - \operatorname*{ess\,inf}_{B_{\varepsilon_k}(x)\cap\Omega} u\ \mathrm{d}\rho_1(x)\Bigg) \\
&\leq \limsup_{k\to\infty} \frac{1}{\varepsilon_k}\left(\int_\Omega \varepsilon_k|\nabla u(x)| + O(\varepsilon_k^2)\,\mathrm{d}\rho_0(x) + \int_\Omega \varepsilon_k|\nabla u(x)| - O(\varepsilon_k^2)\,\mathrm{d}\rho_1(x)\right) \\
&= \limsup_{k\to\infty} \left(\int_\Omega |\nabla u(x)| + O(\varepsilon_k)\,\mathrm{d}\rho_0(x) + \int_\Omega |\nabla u(x)| - O(\varepsilon_k)\,\mathrm{d}\rho_1(x)\right) \\
&\leq \int_\Omega |\nabla u(x)| + \limsup_{k\to\infty} O(\varepsilon_k)\,\mathrm{d}\rho_0(x) + \int_\Omega |\nabla u(x)| + \limsup_{k\to\infty} O(\varepsilon_k)\,\mathrm{d}\rho_1(x) \\
&= \int_\Omega \rho_0 + \rho_1\,\mathrm{d}|Du|(x) = \mathrm{TV}(u;\boldsymbol{\rho}),
\end{aligned}$$

proving the limsup-inequality. □

The Liminf-inequality

The liminf-inequality for the total variation swiftly follows from the liminf-inequality for the perimeter after employing the following lemma, which states that, given a converging sequence of L^1-functions, one finds a subsequence such that the superlevel sets converge. The proof of this lemma follows the ideas of [33, Theorem 13.25].

Lemma 4.18. *Let $(u_k)_{k\in\mathbb{N}} \subset L^1(\Omega)$ be a sequence of functions with $u_k \to u$ in $L^1(\Omega)$. Then there exist a subsequence of $(u_k)_{k\in\mathbb{N}}$ such that the super-level sets $S_k^t := \{u_k > t\}$ converge to $S^t := \{u > t\}$ in $L^1(\Omega)$ for $\mathcal{L}^1$-almost every $t \in \mathbb{R}$.*

Proof. For every $x \in \Omega$ and $k \in \mathbb{N}$ we have

$$\int_{\mathbb{R}} |\chi_{S_k^t}(x) - \chi_{S^t}(x)|\,\mathrm{d}t = \int_{\min\{u_k(x),u(x)\}}^{\max\{u_k(x),u(x)\}} \mathrm{d}t = |u_k(x) - u(x)|.$$

Hence, by Fubini's theorem and the convergence assumption for $(u_k)_{k\in\mathbb{N}}$, we have

$$\int_{\mathbb{R}}\int_\Omega |\chi_{S_k^t}(x) - \chi_{S^t}(x)|\,\mathrm{d}x\,\mathrm{d}t = \int_\Omega |u_k(x) - u(x)|\,\mathrm{d}x \longrightarrow 0$$

as $k \to \infty$. Therefore,

$$\int_\Omega |\chi_{S_k^t}(x) - \chi_{S^t}(x)|\,\mathrm{d}x \longrightarrow 0$$

in $L^1(\mathbb{R})$, implying that there exists a subsequence of $(u_k)_{k\in\mathbb{N}}$ such that

$$\int_\Omega |\chi_{S_k^t}(x) - \chi_{S^t}(x)|\,\mathrm{d}x \longrightarrow 0$$

pointwise for $\mathcal{L}^1$-almost every $t \in \mathbb{R}$, i.e, $\chi_{S_k^t}(x) \to \chi_{S^t}(x)$ in $L^1(\Omega)$ for $\mathcal{L}^1$-almost every $t \in \mathbb{R}$. □

Equipped with this lemma, we are ready to proof the liminf-inequality for the total variation.

Theorem 4.19 (Liminf-inequality for the total variation)**.** *Let $\rho_0, \rho_1 \in C(\overline{\Omega})$ and let $(u_k)_{k\in\mathbb{N}} \subset L^1(\Omega)$ be a sequence of functions with $u_k \to u$ in $L^1(\Omega)$. Then, for any sequence of positive numbers $(\varepsilon_k)_{k\in\mathbb{N}}$ with $\varepsilon_k \to 0$, it holds*

$$\mathrm{TV}(u;\boldsymbol{\rho}) \le \liminf_{k\to\infty} \mathrm{TV}_{\varepsilon_k}(u_k;\boldsymbol{\rho}).$$

Proof. By the previous lemma there exists a subsequence $(u_k)_{k\in\mathbb{N}}$ such that the level sets $\{u_k > s\}$ converge to $\{u > s\}$ in $L^1(\Omega)$ for $\mathcal{L}^1$-almost every $s \in \mathbb{R}$. Hence, the statement directly follows from the coarea formula, Fatou's lemma and the liminf-inequality for the perimeter:

$$\begin{aligned}
\liminf_{k\to\infty} \mathrm{TV}_{\varepsilon_k}(u_k;\boldsymbol{\rho}) &= \liminf_{k\to\infty} \int_{\mathbb{R}} \mathrm{Per}_{\varepsilon_k}(\{u_k > s\};\boldsymbol{\rho})\,\mathrm{d}s \\
&\ge \int_{\mathbb{R}} \liminf_{k\to\infty} \mathrm{Per}_{\varepsilon_k}(\{u_k > s\};\boldsymbol{\rho})\,\mathrm{d}s \\
&\ge \int_{\mathbb{R}} \mathrm{Per}(\{u > s\};\boldsymbol{\rho})\,\mathrm{d}s = \mathrm{TV}(u;\boldsymbol{\rho})
\end{aligned}$$

□

This concludes the proof of Γ-convergence of the non-local total variation (Theorem 4.12), and we turn to the investigation of the asymptotic behavior of the adversarial training model (1.5).

Chapter 5

Asymptotic Behavior of Adversarial Training

5.1 Binary Case

In this chapter we study the asymptotic behavior of the total variation regularized optimization problem analyzed in Section 3.1 as the adversarial budget ε tends to 0. Since we wish to apply the Γ-convergence result of the total variation, we restrict the setup to the assumptions in Chapter 4: Let the metric space $\mathcal{X}$ be a bounded set $\Omega \subset \mathbb{R}^d$ with Lipschitz boundary. Let the data distributions $\rho_0, \rho_1 \in C(\overline{\Omega})$ satisfy $\operatorname{ess\,inf}_\Omega(\rho_0 + \rho_1) > 0$ and let $\mu \in \mathcal{M}(\Omega \times \{0,1\})$ be the measure characterized by $\mu(\,\cdot\, \times \{i\}) \coloneqq \rho_i(\,\cdot\,)$. We investigate the asymptotic behavior of the functional proposed in (1.5) with the hypothesis space $\mathcal{H} \coloneqq \{u \in L^\infty(\Omega) : 0 \leq u \leq 1\}$, and discuss the convergence of its minimizers.

In view of the previous results, mainly the existence of minimizers to this problem for fixed $\varepsilon > 0$ proved in Theorem 3.2, and the Γ-convergence result for the non-local total variation for continuous data distributions ρ_i in Section 4.2, it swiftly follows

$$\mathbb{E}_{(x,y)\sim\mu}[l(u(x),y)] + \varepsilon\, \mathrm{TV}_\varepsilon(u;\boldsymbol{\rho}) \xrightarrow{\Gamma} \mathbb{E}_{(x,y)\sim\mu}[l(u(x),y)] \tag{5.1}$$

in the weak-* topology on $L^1(\Omega)$. The liminf-inequality is clear, since the data term is weak-* lower semicontinuous and the non-local total variation is bounded below by 0. By a density argument, it suffices to show the limsup-inequality for smooth $u \in BV(\Omega)$, and by Theorem 4.17, we have $\limsup_{\varepsilon\to 0} \mathrm{TV}_\varepsilon(u;\boldsymbol{\rho}) = \mathrm{TV}(u;\boldsymbol{\rho}) < \infty$. Hence, $\limsup_{\varepsilon\to 0} \varepsilon \mathrm{TV}_\varepsilon(u;\boldsymbol{\rho}) = 0$, implying that the limsup-inequality holds for (5.1) with the constant recovery sequence.

Therefore, the minimizers of (1.5) converge, up to a subsequence, to the standard Bayes classifier, i.e., minimizers of $\mathbb{E}_{(x,y)\sim\mu}[l(u(x),y)]$. Since through this approach one completely loses the regularization effect of the total variation, we will study the rescaled

functional resulting from subtracting the Bayes risk and dividing by the adversarial budget

$$J_\varepsilon(u) := \frac{1}{\varepsilon}\left(\mathbb{E}_{(x,y)\sim\mu}[l(u(x),y)] - \inf_{v\in\mathcal{H}} \mathbb{E}_{(x,y)\sim\mu}[l(v(x),y)]\right) + \mathrm{TV}_\varepsilon(u;\boldsymbol{\rho}), \tag{5.2}$$

relating to Tikhonov regularization for inverse problems [7]. Clearly, the optimization problem

$$\min_{u\in\mathcal{H}} J_\varepsilon(u) \tag{5.3}$$

admits the same solutions as (1.5). Mirroring the approach taken in [14], we will see that, under Assumption 5.1, minimizers of (5.3) converge, up to a subsequence, to minimizers of

$$J(u) := \begin{cases} \mathrm{TV}(u;\boldsymbol{\rho}), & \text{if } u \in \arg\min_{v\in\mathcal{H}} \mathbb{E}_{(x,y)\sim\mu}[l(v(x),y)] \\ \infty, & \text{else,} \end{cases} \tag{5.4}$$

i.e., one obtains convergence to Bayes classifiers—recovering minimizers of the standard classification problem without an adversary—which additionally have minimal weighted total variation. The assumption needed to prove this convergence result, is that there exists a minimizer of J with a recovery sequence for which the data term converges sufficiently fast.

Assumption 5.1. There exists

$$u^* \in \arg\min\left\{\mathrm{TV}(u;\boldsymbol{\rho}) : u \in \arg\min_{v\in\mathcal{H}} \mathbb{E}_{(x,y)\sim\mu}[l(v(x),y)]\right\} \tag{5.5}$$

with $\mathrm{TV}(u^*;\boldsymbol{\rho}) < \infty$, which possesses a recovery sequence $(u_k^*)_{k\in\mathbb{N}}$ satisfying

$$\limsup_{k\to\infty} \frac{\mathbb{E}_{(x,y)\sim\mu}[l(u_k^*(x),y) - l(u^*(x),y)]}{\varepsilon_k} = 0. \tag{5.6}$$

Remark 5.2. For the 0-1-loss (5.6) is satisfied for the following speed of convergence of the recovery sequence

$$\|u_k^* - u^*\|_{L^1} = o(\varepsilon_k).$$

For our class of loss functions it is more difficult to deduce a sufficient speed of convergence. In Theorem 4.17, we have seen that for smooth u^* the constant sequence is a recovery sequence, which trivially satisfies (5.6).

Although, this Assumption 5.1 does not suffice to prove Γ-convergence $J_\varepsilon \xrightarrow{\Gamma} J$, it is in fact strong enough to show convergence of minimizers. We refer to [14] for a detailed discussion regarding how one might alter this assumption to obtain different results. Selecting this assumption, the main theorem we prove in this section, is the following conditional convergence result for the adversarial training problem in (1.5).

Theorem 5.3. *Let $(\varepsilon_k)_{k\in\mathbb{N}}$ be a sequence of positive numbers with $\varepsilon_k \to 0$, and for each $k \in \mathbb{N}$ let u_k be a minimizer of (1.5) for $\varepsilon = \varepsilon_k$. Then, under Assumption 5.1, the sequence of minimizers $(u_k)_{k\in\mathbb{N}}$ possesses a subsequence converging to a minimizer of J.*

As a preparatory lemma, we show the standard liminf-inequality for the rescaled functional, for which Assumption 5.1 is not yet necessary.

Lemma 5.4 (Liminf-inequality)**.** *Let $(u_k)_{k\in\mathbb{N}} \subset L^1(\Omega)$ such that $u_k \to u$ in $L^1(\Omega)$ and $u_k \to [0,1]$ for all $k \in \mathbb{N}$. Then, for any sequence of positive $(\varepsilon_k)_{k\in\mathbb{N}}$ numbers with $\varepsilon_k \to 0$, it holds*

$$J(u) \leq \liminf_{k\to\infty} J_{\varepsilon_k}(u_k).$$

Proof. First, we assume that

$$\alpha \coloneqq \mathbb{E}_{(x,y)\sim\mu}[l(u(x),y)] - \inf_{v\in\mathcal{H}} \mathbb{E}_{(x,y)\sim\mu}[l(v(x),y)] > 0,$$

implying $J(u) = \infty$. By passing to a subsequence $(u_k)_{k\in\mathbb{N}}$ we have pointwise convergence $\mathcal{L}^1$-almost everywhere, and it follows with the continuity of the loss function and Fatou's lemma

$$\mathbb{E}_{(x,y)\sim\mu}[l(u(x),y)] \leq \liminf_{k\to\infty} \mathbb{E}_{(x,y)\sim\mu}[l(u_k(x),y)],$$

which immediately implies

$$\alpha \leq \liminf_{k\to\infty} \mathbb{E}_{(x,y)\sim\mu}[l(u_k(x),y)] - \inf_{v\in\mathcal{H}} \mathbb{E}_{(x,y)\sim\mu}[l(v(x),y)].$$

Hence, it holds

$$\begin{aligned}\liminf_{k\to\infty} J_{\varepsilon_k}(u_k) &= \liminf_{k\to\infty} \frac{1}{\varepsilon_k}\left(\mathbb{E}_{(x,y)\sim\mu}[l(u_k(x),y)] - \inf_{v\in\mathcal{H}} \mathbb{E}_{(x,y)\sim\mu}[l(v(x),y)]\right) + \mathrm{TV}_{\varepsilon_k}(u_k;\boldsymbol{\rho}) \\ &\geq \alpha \liminf_{k\to\infty} \frac{1}{\varepsilon_k} = \infty = J(u).\end{aligned}$$

In the other case, if $\alpha = 0$ we use the liminf-inequality from Theorem 4.19 to find

$$\begin{aligned}\liminf_{k\to\infty} J_{\varepsilon_k}(u_k) &= \liminf_{k\to\infty} \frac{1}{\varepsilon_k}\left(\mathbb{E}_{(x,y)\sim\mu}[l(u_k(x),y)] - \inf_{v\in\mathcal{H}} \mathbb{E}_{(x,y)\sim\mu}[l(v(x),y)]\right) + \mathrm{TV}_{\varepsilon_k}(u_k;\boldsymbol{\rho}) \\ &\geq \mathrm{TV}(u;\boldsymbol{\rho}) = J(u),\end{aligned}$$

since the first term in $J_{\varepsilon_k}(u_k)$ is non-negative. □

It is easy to see that the limsup-inequality does not hold in general, unless one forces every Bayes classifier to possess a recovery sequence satisfying (5.6) (see [14] for details). Instead, we use the weaker Assumption 5.1 to show the following conditional compactness result:

Lemma 5.5 (Conditional compactness)**.** *Under Assumption 5.1, any sequence of solutions to (1.5) admits a subsequence converging in* $L^1(\Omega)$*.*

Proof. Let $(u_k)_{k\in\mathbb{N}}$ be a minimizing sequence to (1.5), and let $(u_k^*)_{k\in\mathbb{N}}$ be a recovery sequence for the total variation of the Bayes classifier u^* satisfying Assumption 5.1. Using the minimization property of u_k it holds

$$\mathbb{E}_{(x,y)\sim\mu}[l(u_k(x),y)] + \varepsilon_k \mathrm{TV}_{\varepsilon_k}(u_k;\boldsymbol{\rho}) \leq \mathbb{E}_{(x,y)\sim\mu}[l(u_k^*(x),y)] + \mathrm{TV}_{\varepsilon_k}(u_k^*;\boldsymbol{\rho}).$$

Subtracting the Bayes risk and rescaling by ε_k, we have

$$\begin{aligned}&\frac{\mathbb{E}_{(x,y)\sim\mu}[l(u_k(x),y)] - \mathbb{E}_{(x,y)\sim\mu}[l(u^*(x),y)]}{\varepsilon_k} + \mathrm{TV}_{\varepsilon_k}(u_k;\boldsymbol{\rho})\\ &\quad\leq \frac{\mathbb{E}_{(x,y)\sim\mu}[l(u_k^*(x),y)] - \mathbb{E}_{(x,y)\sim\mu}[l(u^*(x),y)]}{\varepsilon_k} + \mathrm{TV}_{\varepsilon_k}(u_k^*;\boldsymbol{\rho}).\end{aligned}$$

Since the leftmost term is non-negative, we can drop it and take the limsup on both sides of the inequality. By the convergence property (5.6), this gives

$$\begin{aligned}\limsup_{k\to\infty} \mathrm{TV}_{\varepsilon_k}(u_k;\boldsymbol{\rho}) &\leq \limsup_{k\to\infty} \frac{\mathbb{E}_{(x,y)\sim\mu}[l(u_k^*(x),y)] - \mathbb{E}_{(x,y)\sim\mu}[l(u^*(x),y)]}{\varepsilon_k} + \mathrm{TV}_{\varepsilon_k}(u_k^*;\boldsymbol{\rho})\\ &= \limsup_{k\to\infty} \mathrm{TV}_{\varepsilon_k}(u_k^*;\boldsymbol{\rho}) \leq \mathrm{TV}(u^*;\boldsymbol{\rho}) \leq \infty.\end{aligned}$$

The statement follows with Theorem 4.13. □

We conclude this chapter by combining Lemma 5.4 and Lemma 5.5 to prove the main result about the asymptotic behavior of the regularized optimization problem (1.5):

Proof of Theorem 5.3. By Lemma 5.5 there exists a subsequence $(u_k)_{k\in\mathbb{N}}$ converging in $L^1(\Omega)$ to some $u \in L^1(\Omega)$. Let $(u_k^*)_{k\in\mathbb{N}}$ denote a recovery sequence for u^* satisfying Assumption 5.1. By Lemma 5.4 and the fact that the limsup-inequality holds for $(u_k^*)_{k\in\mathbb{N}}$, we have

$$J(u) \leq \liminf_{k\to\infty} J_{\varepsilon_k}(u_k) \leq \limsup_{k\to\infty} J_{\varepsilon_k}(u_k^*) \leq J(u^*) = \mathrm{TV}(u^*;\boldsymbol{\rho}).$$

Since $\mathrm{TV}(u^*;\boldsymbol{\rho}) < \infty$, it holds $J(u) < \infty$ which, by definition, implies $J(u) = \mathrm{TV}(u;\boldsymbol{\rho})$. Therefore, we have

$$\mathrm{TV}(u;\boldsymbol{\rho}) = J(u) \leq \mathrm{TV}(u^*;\boldsymbol{\rho}),$$

and since u^* minimizes $\mathrm{TV}(\,\cdot\,;\boldsymbol{\rho})$, it follows that $u \in \arg\min_{v\in\mathcal{H}} J(v)$. □

In the following section, we generalize these findings to the multiclass setting, as the final result of this thesis.

5.2 Multiclass Case

With the hypothesis space $\mathcal{H}$, the non-local multiclass total variation $\mathrm{TV}_\varepsilon^M(\,\cdot\,;\boldsymbol{\rho})$ and loss function l^M defined as in Section 3.2, the asymptotic result for the binary case easily generalizes to the optimization problem with M classes. We keep the setup from the binary case with $\rho_1,\dots,\rho_M \in C(\overline{\Omega})$ such that $\operatorname{ess\,inf}_\Omega\left(\sum_{i=1}^M \rho_i\right) > 0$ and define $\boldsymbol{\rho} \coloneqq (\rho_1,\dots,\rho_M)$. By the Γ-convergence of the binary total variation, we trivially have

$$\mathrm{TV}_\varepsilon^M(\,\cdot\,;\boldsymbol{\rho}) \xrightarrow{\Gamma} \mathrm{TV}^M(\,\cdot\,;\boldsymbol{\rho}), \tag{5.7}$$

where

$$\mathrm{TV}^M(u;\boldsymbol{\rho}) \coloneqq \sum_{i=1}^M \int_\Omega \rho_i \,\mathrm{d}|Du^i|(x).$$

Analogous to the previous section, for $\mu(\,\cdot\,\times\{i\}) \coloneqq \rho_i(\,\cdot\,)$, let

$$J_\varepsilon(u) \coloneqq \frac{1}{\varepsilon}\left(\mathbb{E}_{(x,y)\sim\mu}\left[l^M(u(x),y)\right] - \inf_{v\in\mathcal{H}} \mathbb{E}_{(x,y)\sim\mu}\left[l^M(v(x),y)\right]\right) + \mathrm{TV}_\varepsilon^M(u;\boldsymbol{\rho}) \tag{5.8}$$

and

$$J(u) \coloneqq \begin{cases} \mathrm{TV}^M(u;\boldsymbol{\rho}), & \text{if } u \in \arg\min_{v\in\mathcal{H}} \mathbb{E}_{(x,y)\sim\mu}[l^M(v(x),y)] \\ \infty, & \text{else.} \end{cases} \tag{5.9}$$

We state the assumption under which minimizers of the regularized multiclass optimization problem (3.7), admit subsequences converging to minimizers of J.

Assumption 5.6. There exists

$$u^* \in \arg\min\left\{\mathrm{TV}^M(u;\boldsymbol{\rho}) : u \in \arg\min_{v\in\mathcal{H}} \mathbb{E}_{(x,y)\sim\mu}\left[l^M(v(x),y)\right]\right\} \tag{5.10}$$

with $\mathrm{TV}^M(u^*;\boldsymbol{\rho}) < \infty$, which possesses a recovery sequence $(u_k^*)_{k\in\mathbb{N}}$ satisfying

$$\limsup_{k\to\infty} \frac{\mathbb{E}_{(x,y)\sim\mu}\left[l^M(u_k^*(x),y) - l^M(u^*(x),y)\right]}{\varepsilon_k} = 0. \tag{5.11}$$

Applying the same line of reasoning as in Theorem 5.3 yields the following asymptotic result.

Corollary 5.7. *Let $(\varepsilon_k)_{k\in\mathbb{N}}$ be a sequence of positive numbers with $\varepsilon_k \to 0$, and for each $k \in \mathbb{N}$ let u_k be a minimizer of (3.7) for $\varepsilon = \varepsilon_k$. Then, under Assumption 5.6, the sequence of minimizers $(u_k)_{k\in\mathbb{N}}$ possesses a subsequence converging to a minimizer of J.*

Chapter 6

Outlook

We conclude this thesis by presenting ideas on how our existence and asymptotic behavior results for adversarial training might be generalized further.

6.1 Existence of Solutions for Arbitrary Loss Functions

In Chapter 3 we proved existence of minimizers for different adversarial training models. In particular we have answered the question posed in [13] about the existence of minimizers to the original robust optimization problem (1.2) for continuous and convex loss functions. Although, this encompasses a large class of loss functions, for arbitrary loss functions the existence of minimizers is still an open and relevant question. Non-convex loss functions, for example, might be of interest, since Meunier et al. showed in [36] that convex loss cannnot be a consistent surrogate loss, i.e., a loss function which does not lead to a different minimization sequence than the 0-1-loss. Our argument in Lemma 3.16 does not generalize easily, since the convexity of the loss function was integral in turning the problem of weak-* lower semicontinuity into the analytically advantageous problem of strong lower semicontinuity.

Similarly, one might study the simpler regularized problem (1.5) for arbitrary loss functions. Here, the loss function only appears in the data term but the lower semicontinuity also heavily relies on convexity properties. The continuity assumption, on the other hand, is easily reduced to lower semicontinuity and it might be possible to slightly alter the proof and weaken this assumption further.

6.2 Asymptotics

6.2.1 Asymptotics for General Densities

In view of the Γ-convergence results in [14] and our results in Chapter 4, one might investigate the question of the largest class of densities ρ_0, ρ_1 for which Γ-convergence of the weighted total variation and perimeter hold. In [14] these results were shown for $\rho_0, \rho_1 \in BV(\Omega) \cap L^\infty(\Omega)$, while we worked with the assumption $\rho_0, \rho_1 \in C(\overline{\Omega})$. Combining these results, it might be possible to show Γ-convergence for densities which are the sum of a continuous and a BV function, i.e.,

$$\rho_i \coloneqq \rho_i^C + \rho_i^{BV}, \quad \rho_i^C \in C(\overline{\Omega}),\ \rho_i^{BV} \in BV(\Omega) \cap L^\infty(\Omega).$$

for $i \in \{0, 1\}$. The liminf-inequality easily follows from the existing results, since, by the additivity of the integral, the non-local and local total variation can be split into two parts only depending on (ρ_0^C, ρ_1^C) and $(\rho_0^{BV}, \rho_1^{BV})$, respectively:

$$\begin{aligned}
\liminf_{k\to\infty} \mathrm{TV}_\varepsilon(u_k; (\rho_0, \rho_1)) &= \liminf_{k\to\infty} \left(\mathrm{TV}_\varepsilon(u_k; (\rho_0^C, \rho_1^C)) + \mathrm{TV}_\varepsilon(u_k; (\rho_0^{BV}, \rho_1^{BV}))\right) \\
&\geq \liminf_{k\to\infty} \mathrm{TV}_\varepsilon(u_k; (\rho_0^C, \rho_1^C)) + \liminf_{k\to\infty} \mathrm{TV}_\varepsilon(u_k; (\rho_0^{BV}, \rho_1^{BV})) \\
&\geq \mathrm{TV}(u_k; (\rho_0^C, \rho_1^C)) + \mathrm{TV}(u_k; (\rho_0^{BV}, \rho_1^{BV})) \\
&= \mathrm{TV}(u_k; (\rho_0, \rho_1))
\end{aligned}$$

At this point it is unclear whether the limsup-inequality holds. Verifying the limsup-inequality would likely be more involved, as one would need to construct a recovery sequence, which recovers $\mathrm{TV}(u; (\rho_0^C, \rho_1^C))$ and $\mathrm{TV}(u; (\rho_0^{BV}, \rho_1^{BV}))$, simultaneously.

6.2.2 Asymptotics for Arbitrary Loss Functions

If one manages to show existence of solutions to the regularized problem (1.3) for arbitrary loss functions, it should also be possible to generalize the asymptotic behavior results from Chapter 5, as the convexity of the loss function does not influence the limiting behavior of the reformulated functional in (5.2), and the continuity is only necessary to show the lower semicontinuity of the data term, which would already have to be verified for the existence of minimizers in the non-local case. Alternatively, the investigation of the limiting behavior of problem (1.3) for arbitrary loss functions might still be of interest, even if establishing the existence of minimizers for a fixed adversarial budget proves challenging, since one can nonetheless derive statements about minimizers in the limit.

6.2.3 Asymptotics of the Robust Problem

Finally, we pose the question of the asymptotic behavior of the original robust optimization problem (1.2). For continuous and convex loss function we did show that the problem can be rewritten with the essential supremum with respect to a reference measure. Then, following the approach taken in [13], we can split the problem into a data term and a regularizer $R_\varepsilon(\,\cdot\,;\mu)$ depending on ε and μ

$$\mathbb{E}_{(x,y)\sim\mu}\left[\sup_{\tilde{x}\in B_\varepsilon(x)} l(u(\tilde{x}),y)\right] = \mathbb{E}_{(x,y)\sim\mu}\left[l(u(x),y)\right] + \varepsilon R_\varepsilon(u;\mu). \tag{6.1}$$

where

$$\begin{aligned} R_\varepsilon(u;\mu) \coloneqq \frac{1}{\varepsilon}\int_{\mathcal{X}} \nu\text{-}\operatorname*{ess\,sup}_{\tilde{x}\in B_\varepsilon(x)} l(u(\tilde{x}),0) - l(u(x),0)\,\mathrm{d}\rho_0(x) \\ + \frac{1}{\varepsilon}\int_{\mathcal{X}} \nu\text{-}\operatorname*{ess\,sup}_{\tilde{x}\in B_\varepsilon(x)} l(u(\tilde{x}),1) - l(u(x),1)\,\mathrm{d}\rho_1(x). \end{aligned}$$

Even when restricting the setup to continuous data distributions on Lipschitz domains in $\mathbb{R}^d$, as we did in Chapter 4, it is not clear how one would generalize our proofs, to show the compactness property and Γ-convergence for this non-local regularizer containing the loss function. Based on the findings in [27], one might expect

$$R(u;\mu) = \int_{\mathcal{X}} |\nabla l(u(x),0)|\,\mathrm{d}\rho_0(x) + \int_{\mathcal{X}} |\nabla l(u(x),1)|\,\mathrm{d}\rho_1(x)$$

as a reasonable Γ-limit for appropriate data distributions and loss functions.

Bibliography

[1] L. Ambrosio, N. Fusco, and D. Pallara. *Functions of Bounded Variation and Free Discontinuity Problems*. Oxford University PressOxford, Mar. 2000.

[2] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok. Synthesizing robust adversarial examples. In *International conference on machine learning*, pages 284–293. PMLR, 2018.

[3] P. Awasthi, N. Frank, and M. Mohri. On the existence of the adversarial bayes classifier. *Advances in Neural Information Processing Systems*, 34:2978–2990, 2021.

[4] T. Bai, J. Luo, J. Zhao, B. Wen, and Q. Wang. Recent advances in adversarial training for adversarial robustness. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, IJCAI-2021, page 4312–4321. International Joint Conferences on Artificial Intelligence Organization, Aug. 2021.

[5] R. Bartle. *The Elements of Real Analysis*. Bibliyografya ve İndeks. Wiley, 1976.

[6] H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer International Publishing, 2017.

[7] A. Belenkin, M. Hartz, and T. Schuster. A note on Γ-convergence of tikhonov functionals for nonlinear inverse problems, 2022.

[8] B. Biggio and F. Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, Dec. 2018.

[9] J. Bose, G. Gidel, H. Berard, A. Cianflone, P. Vincent, S. Lacoste-Julien, and W. Hamilton. Adversarial example games. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 8921–8934. Curran Associates, Inc., 2020.

[10] A. Braides. *Gamma-Convergence for Beginners*. Oxford University Press, July 2002.

[11] H. Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Springer New York, 2011.

[12] L. Bungert. PDEs in Data Science. Lecture notes, CIME Summer School "PDEs, Control and Deep Learning". July 22–26, 2024.

[13] L. Bungert, N. García Trillos, and R. Murray. The geometry of adversarial training in binary classification. *Information and Inference: A Journal of the IMA*, 12(2):921–968, Jan. 2023.

[14] L. Bungert and K. Stinson. Gamma-convergence of a nonlocal perimeter arising in adversarial machine learning. *Calculus of Variations and Partial Differential Equations*, 63(5), Apr. 2024.

[15] A. Cesaroni, S. Dipierro, M. Novaga, and E. Valdinoci. Minimizers for nonlocal perimeters of minkowski type. *Calculus of Variations and Partial Differential Equations*, 57:1–40, 2018.

[16] A. Cesaroni and M. Novaga. Isoperimetric problems for a nonlocal perimeter of minkowski type. *Geometric Flows*, 2(1):86–93, 2017.

[17] A. Chambolle, A. Giacomini, and L. Lussardi. Continuous limits of discrete perimeters. *ESAIM: Mathematical Modelling and Numerical Analysis*, 44(2):207–230, Dec. 2009.

[18] A. Chambolle, S. Lisini, and L. Lussardi. A remark on the anisotropic outer minkowski content. *Advances in Calculus of Variations*, 7(2):241–266, May 2013.

[19] R. Chen, I. C. Paschalidis, et al. Distributionally robust learning. *Foundations and Trends® in Optimization*, 4(1-2):1–243, 2020.

[20] J. B. Conway. *A Course in Functional Analysis*. Springer New York, 2007.

[21] D. Daners. *Domain Perturbation for Linear and Semi-Linear Boundary Value Problems*, page 1–81. Elsevier, 2008.

[22] E. De Giorgi. Sulla convergenza di alcune successioni d'integrali del tipo dell'area. *Ennio De Giorgi*, 414:64, 1975.

[23] E. De Giorgi and T. Franzoni. Su un tipo di convergenza variazionale. *Atti della Accademia Nazionale dei Lincei. Classe di Scienze Fisiche, Matematiche e Naturali. Rendiconti*, 58(6):842–850, 1975.

[24] Y. Deng, X. Zheng, T. Zhang, C. Chen, G. Lou, and M. Kim. An analysis of adversarial attacks and defenses on autonomous driving models. In *2020 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, page 1–10. IEEE, Mar. 2020.

[25] L. Evans. *Partial Differential Equations*. American Mathematical Society, Mar. 2010.

[26] L. C. Evans and R. F. Gariepy. *Measure Theory and Fine Properties of Functions, Revised Edition*. Chapman and Hall/CRC, Apr. 2015.

[27] C. Finlay and A. M. Oberman. Scaleable input gradient regularization for adversarial robustness. *Machine Learning with Applications*, 3:100017, Mar. 2021.

[28] N. García Trillos, M. Jacobs, and J. Kim. On the existence of solutions to adversarial training in multiclass classification. *European Journal of Applied Mathematics*, page 1–21, Dec. 2024.

[29] E. Giorgi and S. Spagnolo. Sulla convergenza degli integrali dell'energia per operatori ellitici del secondo ordine. *Bollettino della Unione Matematica Italiana. Series IV*, 8, 01 1973.

[30] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples, 2015.

[31] L. Grafakos. *Classical Fourier Analysis*. Springer New York, 2014.

[32] A. Khanfer. *Measure Theory and Integration*. Springer Nature Singapore, 2023.

[33] G. Leoni. *A First Course in Sobolev Spaces*. American Mathematical Society, July 2009.

[34] F. Maggi. *Sets of Finite Perimeter and Geometric Variational Problems: An Introduction to Geometric Measure Theory*. Cambridge University Press, Aug. 2012.

[35] P. Mattila. *Geometry of Sets and Measures in Euclidean Spaces: Fractals and Rectifiability*. Cambridge University Press, Apr. 1995.

[36] L. Meunier, M. Scetbon, R. B. Pinot, J. Atif, and Y. Chevaleyre. Mixed nash equilibria in the adversarial examples game. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7677–7687. PMLR, 18–24 Jul 2021.

[37] M. Miranda. Functions of bounded variation on "good" metric spaces. *Journal de Mathématiques Pures et Appliquées*, 82(8):975–1004, Aug. 2003.

[38] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. Deepfool: A simple and accurate method to fool deep neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2574–2582, 2016.

[39] J. Munkres. *Topology*. Featured Titles for Topology. Prentice Hall, Incorporated, 2000.

[40] A. Mądry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[41] A. Mądry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations*, Vancouver, BC, Canada, April 30 - May 3 2018. ICLR2018.

[42] R. C. Rogers. The calculus of several variables, 2011.

[43] H. Royden and P. Fitzpatrick. *Real Analysis*. Prentice Hall, 2010.

[44] D. Spector. Simple proofs of some results of Reshetnyak. *Proceedings of the American Mathematical Society*, 139(05):1681–1681, May 2011.

[45] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks, 2013.

[46] H. Taheri, R. Pedarsani, and C. Thrampoulidis. Asymptotic behavior of adversarial training in binary linear classification. In *2022 IEEE International Symposium on Information Theory (ISIT)*, pages 127–132. IEEE, 2022.

[47] N. G. Trillos and R. Murray. Adversarial classification: Necessary conditions and geometric flows. *Journal of Machine Learning Research*, 23(187):1–38, 2022.

[48] Y. Xie and X. Huo. Asymptotic behavior of adversarial training estimator under l_∞-perturbation. *Journal of the American Statistical Association*, (just-accepted):1–20, 2025.

[49] M. Zhao, L. Zhang, J. Ye, H. Lu, B. Yin, and X. Wang. Adversarial training: A survey. *arXiv preprint arXiv:2410.15042*, 2024.

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Versicherung zur selbständigen Leistungserbringung

Titel der Masterarbeit:

Existence of Solutions and Asymptotic Behavior of Adversarial Training with General Loss Functions

Thema bereitgestellt von (Titel, Vorname, Nachname, Lehrstuhl):
Prof. Dr. Leon Bungert

Eingereicht durch (Vorname, Nachname, Matrikel):
Lennart Siethoff 2300398

Ich versichere, dass ich die vorstehende schriftliche Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Die benutzte Literatur sowie sonstige Hilfsquellen sind vollständig angegeben. Wörtlich oder dem Sinne nach dem Schrifttum oder dem Internet entnommene Stellen sind unter Angabe der Quelle kenntlich gemacht.

Weitere Personen waren an der geistigen Leistung der vorliegenden Arbeit nicht beteiligt. Insbesondere habe ich nicht die Hilfe eines Ghostwriters oder einer Ghostwriting-Agentur in Anspruch genommen. Dritte haben von mir weder unmittelbar noch mittelbar Geld oder geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Arbeit stehen.

☐ Mit dem Prüfungsleiter bzw. der Prüfungsleiterin wurde abgestimmt, dass für die Erstellung der vorgelegten schriftlichen Arbeit Chatbots (insbesondere ChatGPT) bzw. allgemein solche Programme, die anstelle meiner Person die Aufgabenstellung der Prüfung bzw. Teile derselben bearbeiten könnten, entsprechend den Vorgaben der Prüfungsleiterin bzw. des Prüfungsleiters eingesetzt wurden. Die mittels Chatbots erstellten Passagen sind als solche gekennzeichnet.

Der Durchführung einer elektronischen Plagiatsprüfung stimme ich hiermit zu. Die eingereichte elektronische Fassung der Arbeit ist vollständig. Mir ist bewusst, dass nachträgliche Ergänzungen ausgeschlossen sind.

Die Arbeit wurde bisher keiner anderen Prüfungsbehörde vorgelegt und auch nicht veröffentlicht. Ich bin mir bewusst, dass eine unwahre Erklärung zur Versicherung der selbstständigen Leistungserbringung rechtliche Folgen haben kann.

Würzburg, 29.06.2025

Ort, Datum, Unterschrift