ESAIM: COCV 32 (2026) 56
https://doi.org/10.1051/cocv/2026040

**ESAIM: Control, Optimisation and Calculus of Variations**
www.esaim-cocv.org

# IT BEGINS WITH A BOUNDARY: A GEOMETRIC VIEW ON PROBABILISTICALLY ROBUST LEARNING

Leon Bungert[1], Nicolás García Trillos[2,*], Matt Jacobs[3], Daniel McKenzie[4], Đorđe Nikolić[5] and Qingsong Wang[6]

**Abstract.** Although deep neural networks have achieved super-human performance on many classification tasks, they often exhibit a worrying lack of robustness towards adversarially generated examples. Thus, considerable effort has been invested into reformulating standard Risk Minimization (RM) into an adversarially robust framework. Recently, attention has shifted towards approaches which interpolate between the robustness offered by adversarial training and the higher clean accuracy and faster training times of RM. In this paper, we take a fresh and geometric view on one such method – Probabilistically Robust Learning (PRL) [A. Robey *et al. International Conference on Machine Learning*. PMLR (2022) 18667–18686]. We propose a mathematical framework for understanding PRL, which allows us to identify geometric pathologies in its original formulation and to introduce a family of probabilistic nonlocal perimeter functionals to rectify them. We prove existence of solutions to the original and modified problems using relaxation methods in the calculus of variations and also study properties, as well as local limits, of the introduced perimeters. We also clarify, through a suitable Γ-convergence analysis, the way in which the original and modified PRL models interpolate between risk minimization and adversarial training.

**Mathematics Subject Classification.** 49J45, 49Q10.

Received November 24, 2024. Accepted May 14, 2026.

## 1. Introduction

The fragility of deep neural network (DNN) based classifiers in the face of adversarial examples [1–4] and distributional shifts [5, 6] is by now nearly as familiar as their success stories. In light of this, a multitude of works (see Sections 1.5) propose replacing standard Risk Minimization (RM) [7] with a more robust alternative,

*Keywords and phrases:* Probabilistically robust learning, adversarial training, perimeter minimization, regularized empirical risk minimization, gamma convergence.

[1] Institute of Mathematics & Center for Artifical Intelligence and Data Science (CAIDAS), University of Würzburg, Emil-Fischer-Str. 40, 97074, Würzburg, Germany.
[2] Department of Statistics, University of Wisconsin-Madison, 1300 University Avenue, Madison, WI 53706, USA.
[3] Department of Mathematics, University of California Santa Barbara, South Hall, Room 6607, Santa Barbara, CA 93106-3080, USA.
[4] Department of Applied Mathematics and Statistics, Colorado School of Mines, 1500 Illinois Street, Golden, CO 80401, USA.
[5] Department of Mathematics and Statistics, Lederle Graduate Research Tower, 1654, University of Massachusetts Amherst, 710 N. Pleasant Street, Amherst, MA 01003-9305, USA.
[6] Halıcıoğlu Data Science Institute, University of California San Diego, 3234 Matthews Lane, La Jolla, CA 92093, USA.

* Corresponding author: garciatrillo@wisc.edu

with adversarial training (AT) [3, 8] being a leading example. Unfortunately, there is no free lunch: robust classifiers frequently exhibit degraded performance on clean data and significantly longer training times [9]. Consequently, identifying frameworks which balance performance and robustness is of pressing interest to the Machine Learning (ML) community, and over the past several years a few such frameworks have been proposed [10–12]. From the theoretical perspective, it is crucial that the mechanisms by which such frameworks balance these competing aims are understood.

In this paper we revisit the Probabilistically Robust Learning (PRL) framework introduced in [10] and discuss it, analyze it, and extend it using a geometric perspective. This perspective reveals certain subtle and paradoxical aspect of PRL: It can have solutions which are very pathological classifiers, *e.g.*, see Figure 1. Building on this observation one can conclude that solutions of PRL do not necessarily interpolate between standard risk minimization and AT. This interpolating property was one of the main theoretical motivations for designing the PRL model in [10]; see Section 4 for an extensive discussion on this.

Fortunately, our geometric perspective suggests a natural remedy which leads to an interpretation of the modified PRL as regularized RM, where a new notion of nonlocal length (or perimeter) of decision boundaries acts as a regularizer; this notion of perimeter (see (1.9) for its definition) is of interest in its own right and in Section 3 we will discuss some of its properties. The interpretation of PRL as perimeter-regularized RM leads us to further generalizations and allows us to provide a novel view of the Conditional Value at Risk (CVaR) relaxation of PRL proposed in [10]. We provide conditions that guarantee the existence of solutions to the optimization problems determining our models as well as the original PRL models, and clarify, through rigorous analysis using the notion of $\Gamma$-convergence, the sense in which the original PRL and the modified PRL models interpolate between adversarial training and standard risk minimization. Our existence proofs exploit interesting structural properties of our functionals that allow us to circumvent the apparent incompatibility between the topologies under which our functionals are lower semicontinuous and under which the compactness of minimizing sequences can be guaranteed.

### 1.1. From empirical risk minimization to robustness

Given an input space $\mathcal{X}$, an output space $\mathcal{Y}$, a probability measure $\mu \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$, a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, and a hypothesis class $\mathcal{H}$ (*i.e.*, a class of measurable functions from $\mathcal{X}$ into $\mathcal{Y}$), the standard risk minimization problem is

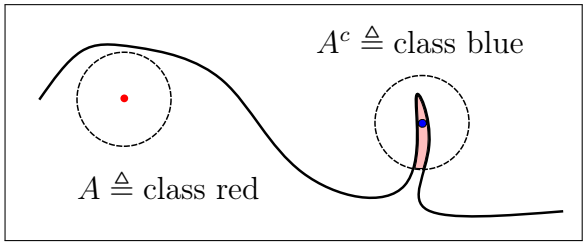
$$\inf_{h \in \mathcal{H}} \mathbb{E}_{(x,y)\sim\mu} \left[ \ell(h(x), y) \right]. \tag{RM}$$

To train classifiers which are robust against adversarial attacks, **(author?)** [3, 8] suggested adversarial training:

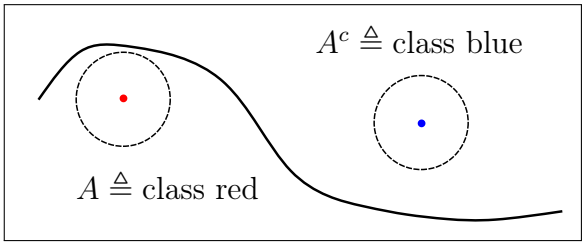
$$\inf_{h \in \mathcal{H}} \mathbb{E}_{(x,y)\sim\mu} \left[ \sup_{x' \in B_\varepsilon(x)} \ell(h(x'), y) \right]. \tag{AT}$$

In the above, $\mathcal{X}$ is assumed to have the structure of a metric space and $B_\varepsilon(x)$ for $\varepsilon > 0$ denotes the (open or closed) ball of radius $\varepsilon$ around $x$, although for convenience we will consider open balls whenever is needed.

The recent work [10] offered an alternative to adversarial training that attempts to reduce the (in general) large trade-off between accuracy and robustness inherent in (AT); see [9, 10] for discussion. Instead of requiring classifiers to be robust to *all* possible attacks around a point $x$ – as enforced through the supremum in (AT) – one may consider a less stringent notion of robustness, only requiring classifiers to be robust to $100 \times (1-p)\%$ of attacks drawn from a certain distribution $\mathfrak{m}_x$ centered at $x$. For this, the authors replace the supremum in (AT) by the so-called $p$-ess sup operator for some value $p \in [0, 1)$. To define this operator, consider a probability




(A) This pathological classifier minimizes the original PRL risk *but not* the modified one.




(B) This classifier minimizes the original PRL risk *and* the modified one.

FIGURE 1. Even for the simple task of classifying two data points one can easily construct a pathological solution (Fig. 1a) of PRL, observing that all but the small red set of perturbations of the *misclassified* blue point are correctly classified as blue. Both classifiers depicted in Figure 1a and 1b have zero PRL loss. In Proposition B.1 in the Appendix, we prove that the modified PRL model studied in this paper selects classifiers that rule out situations like the one considered here.

distribution $\mathfrak{m}$ and a function $f$ and let

$$p\text{-}\operatorname*{ess\,sup}_{x'\sim\mathfrak{m}} f(x') := \inf\left\{t \in \mathbb{R} \,:\, \mathbb{P}_{x'\sim\mathfrak{m}}\left[f(x') > t\right] \leq p\right\}.$$

In the mathematical finance or economics literature the $p$-ess sup operator is better known as the value at risk (VaR) of a random variable at level $p$. It is the smallest value $t \in \mathbb{R}$ such that the probability that a randomly chosen point $x' \sim \mathfrak{m}$ satisfies $f(x') > t$ is smaller than $p$. This notion reduces to the usual essential supremum of $f$ with respect to $\mathfrak{m}$ if $p = 0$, which motivates the name $p$-ess sup. In [10], it was thus suggested to replace (AT) with the *probabilistically robust learning* problem

$$\inf_{h\in\mathcal{H}} \mathbb{E}_{(x,y)\sim\mu}\left[p\text{-}\operatorname*{ess\,sup}_{x'\sim\mathfrak{m}_x} \ell(h(x'),y)\right], \tag{PRL}$$

where $\{\mathfrak{m}_x\}_{x\in\mathcal{X}}$ is a suitable family of probability distributions, one for each $x$ in the data space, interpreted as user-chosen distributions "centered" around the different $x$. The prototypical example to keep in mind for $\mathcal{X} = \mathbb{R}^d$ is the uniform distribution over the $\varepsilon$-ball around $x$, *i.e.*, $\mathfrak{m}_x := \mathrm{Unif}(B_\varepsilon(x))$, which is particularly relevant when dealing with adversarial attacks on image classifiers. We will provide further theoretical discussion on the choice of the measures $\mathfrak{m}_x$ in Section 4.

## 1.2. Modifying PRL for binary classification

To understand (PRL) better, we focus first (and mostly) on the binary classification problem (*i.e.*, $\mathcal{Y} = \{0,1\}$) using indicator functions of admissible sets (*i.e.*, $\mathcal{H} := \{\mathbf{1}_A \,:\, A \in \mathcal{A}\}$). For now, one can think of $\mathcal{A} \subset 2^{\mathcal{X}}$ as a suitable $\sigma$-algebra which fits to the problem. Note that we identify the two expressions $\mathbf{1}_A(x) = \mathbf{1}_{x\in A}$. We focus on the standard 0-1 loss $\ell(\tilde{y},y) = \mathbf{1}_{\tilde{y}\neq y}$, which equals one if $y \neq \tilde{y}$ and zero otherwise. In this scenario, the standard risk minimization problem (RM) reduces to

$$\inf_{A\in\mathcal{A}}\left\{\mathrm{R}_{\mathrm{std}}(A) := \mathbb{E}_{(x,y)\sim\mu}\left[y\mathbf{1}_{x\in A^c} + (1-y)\mathbf{1}_{x\in A}\right]\right\}, \tag{1.1}$$

and we refer to its minimizers as *Bayes classifiers* and to its minimum value as *Bayes risk*.

Similarly, (AT) can be rewritten as

$$\inf_{A\in\mathcal{A}}\left\{ \mathrm{R}_{\mathrm{adv}}(A) := \mathbb{E}_{(x,y)\sim\mu}\left[ y\mathbf{1}_{x\in(A^c)^{\oplus\varepsilon}} + (1-y)\mathbf{1}_{x\in A^{\oplus\varepsilon}}\right]\right\}, \tag{1.2}$$

where for a set $A\in\mathcal{A}$ its fattening by $\varepsilon$-balls is defined as $A^{\oplus\varepsilon} := \bigcup_{x\in A} B_\varepsilon(x)$. As for (PRL), it reduces to

$$\inf_{A\in\mathcal{A}}\left\{ \mathrm{R}_{\mathrm{prob}}(A) := \mathbb{E}_{(x,y)\sim\mu}\left[ y\mathbf{1}_{\mathfrak{m}_x(A^c)>p} + (1-y)\mathbf{1}_{\mathfrak{m}_x(A)>p}\right]\right\}, \tag{1.3}$$

where, in comparison to (1.2), the fattenings $A^{\oplus\varepsilon}$ and $(A^c)^{\oplus\varepsilon}$ are replaced by the set of all points $x\in\mathcal{X}$ for which the probability that a neighboring point sampled from $\mathfrak{m}_x$ lies inside $A$, or $A^c$, respectively, is larger than $p$. From this definition one can see that PRL may lead to counter-intuitive classifiers like the ones observed in Figure 1a. Indeed, taking $\mathfrak{m}_x$ to be the uniform measure over $B_\varepsilon(x)$, we see that the red-shaded spiky region therein has such a small volume that the blue point with label $y=0$ lies in the set $\{x\in\mathcal{X} : \mathfrak{m}_x(A)\le p\}$. Hence, the second term in (1.3) is zero and the spike adds zero cost to the overall energy. At an abstract level, the issue with (1.3) is that sets $\{x\in\mathcal{X} : \mathfrak{m}_x(A^c)>p\}$ and $\{x\in\mathcal{X} : \mathfrak{m}_x(A)>p\}$ are not fattenings, *i.e.*, supersets, of $A^c$ and $A$. Speaking the language of classification, it means that points which are incorrectly classified incur zero loss if most of their perturbations are correctly classified.

In order to obtain a model which does not exhibit this counter-intuitive behavior we modify (1.3) slightly, thereby obtaining a model which has a geometric regularization interpretation and can be embedded in a rich class of models that allows to interpolate between risk minimization and adversarial training. To motivate our modified PRL risk, we first observe that the original adversarial training risk $\mathrm{R}_{\mathrm{adv}}$ admits the following decomposition:

$$\mathrm{R}_{\mathrm{adv}}(A) = \mathrm{R}_{\mathrm{std}}(A) + \mathrm{Per}_{\mathrm{adv}}(A),$$

where $\mathrm{Per}_{\mathrm{adv}}(A)$ is the non-negative functional

$$\mathrm{Per}_{\mathrm{adv}}(A) := \int_{A^c}\sup_{x'\in B_\varepsilon(x)}\mathbf{1}_A(x')\,\mathrm{d}\rho_0(x) + \int_A\sup_{x'\in B_\varepsilon(x)}\mathbf{1}_{A^c}(x')\,\mathrm{d}\rho_1(x)$$

and $\rho_i(\bullet) := \mu(\bullet\times\{i\})$, *i.e.*, the weighted conditional distribution of the points with label $i\in\{0,1\}$. We refer the interested reader to [13], where a detailed discussion on this decomposition, results on existence of minimizers for adversarial training, and a discussion on the interpretation of the functional $\mathrm{Per}_{\mathrm{adv}}$ as a nonlocal *perimeter* are presented. The two terms in the decomposition of $\mathrm{R}_{\mathrm{adv}}$ can be interpreted as follows: 1) the term $\mathrm{R}_{\mathrm{std}}(A)$ is the standard risk and captures the contribution to the overall adversarial risk by misclassified points; 2) $\mathrm{Per}_{\mathrm{adv}}(A)$ captures the contribution of the points that, although classified correctly by the binary classifier $\mathbf{1}_A$, are within distance $\varepsilon$ from the decision boundary and thus can be perturbed by the adversary to make the classifier's output differ from the correct label. Motivated by the above decomposition for $\mathrm{R}_{\mathrm{adv}}$, by substituting the suprema in the definition of $\mathrm{Per}_{\mathrm{adv}}$ with a softer penalty, the following modified PRL model becomes natural:

$$\inf_{A\in\mathcal{A}}\left\{ \mathrm{R}_{\text{m-prob}}(A) := \mathrm{R}_{\mathrm{std}}(A) + \mathrm{Per}_{\mathrm{prob}}(A)\right\}, \tag{1.4}$$

where $\mathrm{Per}_{\mathrm{prob}}$ is the non-negative functional

$$\mathrm{Per}_{\mathrm{prob}}(A) := \int_{A^c}\mathbf{1}_{\mathfrak{m}_x(A)>p}\,\mathrm{d}\rho_0(x) + \int_A\mathbf{1}_{\mathfrak{m}_x(A^c)>p}\,\mathrm{d}\rho_1(x). \tag{1.5}$$

The notation that we have chosen for the functional $\mathrm{Per}_{\mathrm{prob}}$ suggests a perimeter interpretation, and we will justify this choice shortly. With this interpretation, it becomes clear that the modified PRL model (abbreviated m-PRL in Sect. 6) (1.4) has the structure of a regularized risk minimization problem where the regularization term penalizes the size of the boundary of a set. As discussed earlier, this interpretation also holds for AT.

A simple computation (see Prop. 2.1) allows us to rewrite $\mathrm{R}_{\text{m-prob}}$ as

$$\mathrm{R}_{\text{m-prob}}(A) = \mathbb{E}_{(x,y)\sim\mu}\left[y\mathbf{1}_{x\in A^c \vee \mathfrak{m}_x(A^c)>p} + (1-y)\mathbf{1}_{x\in A\vee \mathfrak{m}_x(A)>p}\right]. \tag{1.6}$$

From this rewriting it is apparent that a point $x$ with label $y$ contributes to the modified PRL risk if it is either non-robust in a probabilistic sense *or* if it is misclassified. For instance, if $y = 0$, the loss is one when either $\mathfrak{m}_x(A) > p$ or when $x \in A$. Note that the latter condition is not captured by the original PRL formulation.

### 1.3. A larger class of PRL models for binary classification

For theoretical and computational reasons that we will soon discuss, it is convenient to generalize problems (1.3) and (1.4) further. Precisely, given an arbitrary function $\Psi : [0,1] \to [0,1]$ we define

$$\mathrm{R}_{\mathrm{prob}\,\Psi}(A) := \int_{\mathcal{X}} \Psi\left(\mathfrak{m}_x(A)\right)\, \mathrm{d}\rho_0(x) + \int_{\mathcal{X}} \Psi\left(\mathfrak{m}_x(A^c)\right)\, \mathrm{d}\rho_1(x), \tag{1.7}$$

as well as

$$\mathrm{R}_{\text{m-prob}\,\Psi}(A) := \mathrm{R}_{\mathrm{std}}(A) + \mathrm{Per}_{\mathrm{prob}\,\Psi}(A), \tag{1.8}$$

where

$$\mathrm{Per}_{\mathrm{prob}\,\Psi}(A) := \int_{A^c} \Psi\left(\mathfrak{m}_x(A)\right)\, \mathrm{d}\rho_0(x) + \int_{A} \Psi\left(\mathfrak{m}_x(A^c)\right)\, \mathrm{d}\rho_1(x). \tag{1.9}$$

The functionals $\mathrm{R}_{\mathrm{prob}\,\Psi}$ and $\mathrm{R}_{\text{m-prob}\,\Psi}$ are indeed generalizations of $\mathrm{R}_{\mathrm{prob}}$ and $\mathrm{R}_{\text{m-prob}}$, respectively. This can be seen by taking $\Psi$ to be $\Psi(t) := \mathbf{1}_{t>p}$, as can be easily verified. Another important choice for $\Psi$ that we will discuss extensively in the sequel is the function $\Psi_p(t) := \min\{\frac{t}{p}, 1\}$, which is the smallest concave function larger than $\mathbf{1}_{t>p}$. Later, in Proposition 2.6 and in Section 5, we will discuss the connection between the CVaR relaxation of PRL introduced for computational convenience in [10] (see more discussion in Appendix C) and the optimization of the risks $\mathrm{R}_{\mathrm{prob}\,\Psi}$ and $\mathrm{R}_{\text{m-prob}\,\Psi}$ for the choice $\Psi = \Psi_p$. From a theoretical perspective, we will establish several structural properties satisfied by the functionals $\mathrm{R}_{\mathrm{prob}\,\Psi}$ and $\mathrm{R}_{\text{m-prob}\,\Psi}$ when $\Psi$ is concave. In particular, while we are not able to prove existence of minimizers (in a large class of sets $\mathcal{A}$) for the original binary problem (1.3) nor for our modification (1.4), we will be able to do it for the minimization of $\mathrm{R}_{\mathrm{prob}\,\Psi}$ and $\mathrm{R}_{\text{m-prob}\,\Psi}$ when $\Psi$ is concave, *e.g.*, when $\Psi = \Psi_p$. Without concavity the problems are degenerate because of the hard thresholding induced by hard classifiers and we will only be able to establish existence of solutions within the enlarged family of "soft" classifiers. A detailed discussion on this is presented in Section 2 below.

### 1.4. Informal statements of main results

This paper investigates three main questions related to PRL. First, we study the existence of minimizers to PRL and to modified PRL. Second, we explore different geometric interpretations of the functionals $\mathrm{Per}_{\mathrm{prob}\,\Psi}$. In particular, we establish two types of results that characterize these functionals as *perimeters*, thereby further suggesting how the models studied in this paper for robust training of classifiers are closely related to geometric regularization methods that explicitly penalize the size of decision boundaries of classifiers. Finally, we present rigorous analysis that describes, in detail, the way in which PRL and modified PRL interpolate between adversarial training and standard risk minimization.

### *1.4.1. Existence of minimizers of PRL models*

In [10], an explicit solution to (PRL) is presented for the case where the hypothesis class consists of linear classifiers and $\mu$ is a mixture of normal distributions. To the best of our knowledge, existence of solutions to (PRL) for general data distributions or hypothesis classes – even for the binary problem (1.3) – has not been proved so far.

Our first main results assert the existence of minimizers of the risks $\mathrm{R}_{\text{m-prob}_\Psi}$ and $\mathrm{R}_{\text{prob}_\Psi}$ for suitable choices of $\Psi$. We impose no restrictions on the data distribution but consider families of classifiers satisfying certain closure relations discussed precisely in later sections.

**Informal Theorem 1.1** (Existence of hard classifiers)**.** For every concave and non-decreasing function $\Psi{:}[0,1] \to [0,1]$ there exists a solution to the problem

$$\min_{A \subset \mathcal{X}} \mathrm{R}_{\text{m-prob}_\Psi}(A).$$

**Informal Theorem 1.2** (Existence of hard classifiers)**.** For every concave and non-decreasing function $\Psi{:}[0,1] \to [0,1]$ there exists a solution to the problem

$$\min_{A \subset \mathcal{X}} \mathrm{R}_{\text{prob}_\Psi}(A).$$

Note that these existence results do not cover the problem (1.4) since the function $\Psi(t) = \mathbf{1}_{t>p}$ is not concave. They do include, however, the CVaR model from Proposition 2.6 below since this model is realized by the choice $\Psi_p(t) := \min\{t/p, 1\}$, which is a concave and non-decreasing function. We believe that our proof strategy, which takes advantage of some interesting structural properties of the functionals $\mathrm{R}_{\text{m-prob}_\Psi}$ and $\mathrm{R}_{\text{prob}_\Psi}$, is of interest in its own right and is captured by our Theorem 2.15 in Section 2.4 below. Precise statements of our results can be found in Section 2.2.

We also present existence results for more general, non-concave functions $\Psi$, including the choice $\Psi(t) = \mathbf{1}_{t>p}$ giving rise to the original PRL model and its modified version introduced here. These results, however, apply when we optimize over "soft" classifiers instead of "hard" ones, meaning we optimize over functions $u : \mathcal{X} \to [0,1]$ in a suitably closed hypothesis class $\mathcal{H}$ instead of characteristic functions $u = \mathbf{1}_A$ of sets $A$. To make sense of our statements, we first need to replace the functional $\mathrm{Per}_{\text{prob}_\Psi}(A)$ of a set $A$ by a suitable regularization functional defined according to

$$\begin{aligned} \mathrm{J}_{\text{prob}_\Psi}(u) := &\int_{\mathcal{X}} (1-u(x))\, \Psi\left(\mathbb{E}_{x' \sim \mathfrak{m}_x}\left[u(x')\right]\right) \mathrm{d}\rho_0(x) \\ &+ \int_{\mathcal{X}} u(x) \Psi\left(\mathbb{E}_{x' \sim \mathfrak{m}_x}\left[1-u(x')\right]\right) \mathrm{d}\rho_1(x). \end{aligned} \tag{1.10}$$

A straightforward computation reveals that $\mathrm{J}_{\text{prob}_\Psi}(\mathbf{1}_A) = \mathrm{Per}_{\text{prob}_\Psi}(A)$.

**Informal Theorem 1.3** (Existence of soft classifiers)**.** For every lower semicontinuous function $\Psi : [0,1] \to [0,1]$ and every hypothesis class $\mathcal{H}$ which is closed in a suitable sense there exists a solution to the problem

$$\min_{u \in \mathcal{H}} \mathbb{E}_{(x,y)\sim\mu}\left[|u(x) - y|\right] + \mathrm{J}_{\text{prob}_\Psi}(u).$$

**Informal Theorem 1.4** (Existence of soft classifiers)**.** For every lower semicontinuous function $\Psi : [0,1] \to [0,1]$ and every hypothesis class $\mathcal{H}$ which is closed in a suitable sense there exists a solution to the problem

$$\min_{u\in\mathcal{H}} \int_{\mathcal{X}} \Psi\left(\mathbb{E}_{x'\sim\mathfrak{m}_x}\left[u(x')\right]\right) \mathrm{d}\rho_0(x) + \int_{\mathcal{X}} \Psi\left(\mathbb{E}_{x'\sim\mathfrak{m}_x}\left[1-u(x')\right]\right) \mathrm{d}\rho_1(x).$$

Note that the objective functions in Informal Theorems 3 and 4 coincide with $\mathrm{R}_{\text{m-prob}_\Psi}$ and $\mathrm{R}_{\text{prob}_\Psi}$, respectively, when $u$ is an indicator function. When $\Psi$ is non-decreasing and concave, and $\mathcal{H}$ is suitably chosen, these optimization problems are continuous and exact relaxations of $\inf_{A\in\mathcal{A}} \mathrm{R}_{\text{m-prob}_\Psi}$ and $\inf_{A\in\mathcal{A}} \mathrm{R}_{\text{prob}_\Psi}$, respectively. Other hypothesis classes $\mathcal{H}$ of interest that are admissible in Informal Theorem 1.3 and 1.4 include functions which are parameterized by neural networks, as discussed in Example 2.13.

### *1.4.2. Interpretation of* $\mathrm{Per}_{\text{prob}_\Psi}$ *as a perimeter*

As suggested by our discussion in earlier sections, the functional $\mathrm{Per}_{\text{prob}_\Psi}$ can be interpreted as a perimeter functional. We can justify this interpretation from two different perspectives.

First, we establish the following structural properties for the functional $\mathrm{Per}_{\text{prob}_\Psi}$ when $\Psi$ is non-decreasing and concave, the exact same assumptions needed in our existence results discussed in Section 1.4.1. These properties allow us to view $\mathrm{Per}_{\text{prob}_\Psi}$ as a generalized perimeter.

**Informal Theorem 1.5** (Submodularity)**.** Provided $\Psi$ is concave, non-decreasing, and $\Psi(0) = 0$ the functional $\mathrm{Per}_{\text{prob}_\Psi}$ is a generalized perimeter. In particular,

$$\mathrm{Per}_{\text{prob}_\Psi}(\emptyset) = \mathrm{Per}_{\text{prob}_\Psi}(\mathcal{X}) = 0$$

and $\mathrm{Per}_{\text{prob}_\Psi}$ is submodular, *i.e.*,

$$\mathrm{Per}_{\text{prob}_\Psi}(A\cup B) + \mathrm{Per}_{\text{prob}_\Psi}(A\cap B) \le \mathrm{Per}_{\text{prob}_\Psi}(A) + \mathrm{Per}_{\text{prob}_\Psi}(B) \quad \forall A, B \in \mathcal{A}.$$

As a generalized perimeter, $\mathrm{Per}_{\text{prob}_\Psi}$ admits a convex extension (its total variation) that is denoted by $\mathrm{TV}_{\text{prob}_\Psi}$ and that we define precisely in (2.11). Interestingly, $\mathrm{Per}_{\text{prob}_\Psi}$ has a second extension, the functional $\mathrm{J}_{\text{prob}_\Psi}$ defined in (1.10), which is sequentially lower-semicontinuous (although it fails to be convex) w.r.t. the weak topology where we can guarantee precompactness of minimizing sequences and is always greater than or equal to $\mathrm{TV}_{\text{prob}_\Psi}$. We use this ordering in our proof strategy for the existence of optimal hard classifiers. We also note that, while non-convex, the extension $\mathrm{J}_{\text{prob}_\Psi}$ has a relatively simple explicit expression that makes it useful for computations, while the same is not true for $\mathrm{TV}_{\text{prob}_\Psi}$, which instead has an impractical expression. being defined through a coarea formula.

When $\Psi$ is more general, we can still give a perimeter interpretation to the functional $\mathrm{Per}_{\text{prob}_\Psi}$, at least in an asymptotic sense. For this purpose, we assume that $\mathfrak{m}_x$ localizes to a point mass at $x$ as certain scaling parameter shrinks. A rigorous counterpart to the next informal result can be found in Theorem 3.7.

**Informal Theorem 1.6** (Local asymptotics)**.** Let $\mathcal{X} = \mathbb{R}^d$. If $\mathfrak{m}_x \equiv \mathfrak{m}_{x,\varepsilon}$ for $\varepsilon > 0$ and $\mathfrak{m}_{x,\varepsilon}$ converges to the Dirac delta $\delta_x$ in a suitable way as $\varepsilon \to 0$ (*e.g.*, take $\mathfrak{m}_{x,\varepsilon} = \mathrm{Unif}(B_\varepsilon(x))$), then the rescaled probabilistic perimeters $\frac{1}{\varepsilon}\mathrm{Per}_{\text{prob}_\Psi}(A)$ of a smooth set $A$ converge, as $\varepsilon \to 0$, to the weighted local perimeter

$$\mathrm{Per}(A) := \int_{\partial A} f(x, n(x))\, \mathrm{d}\mathcal{H}^{d-1},$$

where $n(x)$ denotes the outer unit normal at $x \in \partial A$ and $f$ is a suitable weight function, depending on the distributions $\mathfrak{m}_{x,\varepsilon}$, and $\rho_i$ for $i \in \{0,1\}$. $\mathcal{H}^{d-1}$ denotes the $(d-1)$-dimensional Hausdorff measure in $\mathbb{R}^d$.

### *1.4.3. Interpolation properties of PRL and modified PRL*

Our last main results describe the relation between adversarial training, RM, and the PRL and modified PRL models, at least in the agnostic setting where the family of admissible binary classifiers consists of all Borel measurable sets. We focus on the energies $\mathrm{R}_{\text{m-prob}\,\Psi_p}$ and $\mathrm{R}_{\text{prob}\,\Psi_p}$ as $p$ tends to zero or $\infty$ for the choice $\Psi_p(t) = \min\{t/p, 1\}$.

**Informal Theorem 1.7** (Convergence to minimum adversarial training)**.** Let $\Psi_p(t) := \min\{t/p, 1\}$. Then the minimum value of $\mathrm{R}_{\text{m-prob}\,\Psi_p}$ converges to the minimum value of the adversarial training problem (1.2) as $p \to 0$. If, in addition, for every $x$ the measure $\mathfrak{m}_x$ dominates the data measures $\rho_0$ and $\rho_1$ restricted to the support of $\mathfrak{m}_x$, then minimizers of $\mathrm{R}_{\text{m-prob}\,\Psi_p}$ converge to minimizers of (1.2) as $p \to 0$.

Analogous statements hold for the minimization of $\mathrm{R}_{\text{prob}\,\Psi_p}$.

While the convergence of the minimal risks holds under fairly general assumptions, we emphasize that the convergence of minimizers holds provided we impose additional conditions on the measures $\{\mathfrak{m}_x\}_{x\in\mathcal{X}}$, as our examples in Section 4 illustrate.

For large values of $p$ we have the following result.

**Informal Theorem 1.8** (Convergence to minimum standard risk)**.** Let $\Psi_p(t) := \min\{t/p, 1\}$. Then the minimum value of $\mathrm{R}_{\text{m-prob}\,\Psi_p}$ converges to the minimum value of the RM problem (1.2) as $p \to \infty$. Also, minimizers of $\mathrm{R}_{\text{m-prob}\,\Psi_p}$ converge to minimizers of (1.2) as $p \to \infty$.

We emphasize that the above statement holds for $\mathrm{R}_{\text{m-prob}\,\Psi_p}$ but not for $\mathrm{R}_{\text{prob}\,\Psi_p}$. In fact, there is no value or limiting value of $p$ that makes $\mathrm{R}_{\text{prob}\,\Psi_p}$ into the standard risk functional. This means that the family of functionals $\{\mathrm{R}_{\text{prob}\,\Psi_p}\}_p$ does not actually interpolate between AT and RM, while the family of functionals $\{\mathrm{R}_{\text{m-prob}\,\Psi_p}\}_p$ does. The latter model has the additional advantage of being interpretable as a geometric regularization model.

## 1.5. Related work

Adversarial training was developed in [3, 8] as an approach to produce networks that are less sensitive to adversarial attacks. [14] reduced its computational complexity by reusing gradients from the backpropagation when training neural networks. [15] showed that training with noise perturbations followed by a single signed gradient ascent (FGSM) step can be on par with adversarial training while being much cheaper. This approach was picked up and improved upon in [16] based on gradient alignment. Different authors also investigated test-time robustification of pretrained classifiers using randomized smoothing [17] or geometric / gradient-based approaches [18, 19]. While some of the previous models use a combination of random perturbations and gradient-based adversarial attacks to robustify classifiers, [10] proposed probabilistically robust learning, which is entirely based on random perturbations. PRL aims to interpolate between clean and adversarial accuracy and enjoys the favorable sample complexity of vanilla empirical risk minimization; see also [20] for more insights on this issue. Connections between adversarial training and local perimeter regularization as well as mean curvature flows of decision boundaries were explored in [12, 21] and recently rigorously tied in [22, 23]. Furthermore, in [23] it was proved that solutions to adversarial training converge to distinguished minimizers of standard risk minimization as $\varepsilon \to 0$ which was later quantified and extended to the probabilistically robust setting in [24]. An interpretation of adversarial training as gradient flow is given in [25].

Our work is in line with a series of papers [13, 26–30] that explore the existence of solutions to adversarial training problems in different settings. These existence proofs involve dealing with different kinds of measurability issues, depending on whether open or closed balls $B_\varepsilon(x)$ are used in the attack model. For open balls one can work with the Borel $\sigma$-algebra $\mathcal{A} = \mathfrak{B}(\mathcal{X})$ [13], whereas closed balls require the use of the universal $\sigma$-algebra to make sure that $A^{\oplus\varepsilon}$ is measurable [26, 27, 30]. Recently, these results were improved in [29] where it was proved that, for the case of multi-class classification, even for the closed ball model Borel measurable classifiers exist, although they are not necessarily indicator functions of measurable sets. Moreover, [29] showed

that for all but countably many values of the adversarial budget $\varepsilon > 0$ the open and the closed ball models have the same minimal value.

### 1.6. Outline

The rest of the paper is organized as follows. In Section 2 we make the setting for our variational problems precise and then present a series of results on the existence of minimizers of these problems. In particular, in Sections 2.2 and 2.3 we make precise our Informal Theorems 1.1 and 1.2, and in Section 2.4 present their proofs. Section 3 is devoted to the study of the functional $\mathrm{Per}_{\mathrm{prob}\,\Psi}$ and its interpretation as a perimeter. There, we make the Informal Theorems 1.5 and 1.6 precise and present their proofs. Section 4 is devoted to the interpolation properties of the different PRL models. There we make our Informal Theorems 1.7 and 1.8 precise using the notion of $\Gamma$-convergence. In Section 5, we introduce a modified PRL model for general supervised learning problems beyond binary classification and discuss the connection between the CVar relaxation of PRL in [10] and our geometric framework. In Section 6 we present some numerical results of an implementation of our models in real data settings. We also comment on a gap between theory and practice that arises when using empirical surrogates of $\mathfrak{m}_x$ and $\mu$. We wrap up the paper in Section 7 with some conclusions and perspectives for future research.

## 2. Existence of probabilistically robust classifiers

### 2.1. Preliminaries

In this section we discuss a few other reformulations for the energies discussed in the introduction. These reformulations provide some additional context and connections to the literature that we will later use in our discussion.

To start, we first present a reformulation of the probabilistic risk $\mathrm{R}_{\text{m-prob}}$ as the expected maximum of the sample-wise standard risk and the probabilistically robust risk from [10].

**Proposition 2.1.** *For the* 0-1 *loss* $\ell(\tilde{y}, y) := \mathbf{1}_{\tilde{y}\neq y}$ *we can rewrite the probabilistic risk* $\mathrm{R}_{\text{m-prob}}$ *as in* (1.6)*. In that same setting,* $\mathrm{R}_{\text{m-prob}}$ *can also be written as the expectation of a sample-wise maximum of the standard loss and the loss suggested in [10], that is:*

$$\mathrm{R}_{\text{m-prob}}(A) = \mathbb{E}_{(x,y)\sim\mu}\left[\max\left\{\ell(\mathbf{1}_A(x), y), p\text{-}\operatorname*{ess\,sup}_{x'\sim\mathfrak{m}_x} \ell(\mathbf{1}_A(x'), y)\right\}\right]. \tag{2.1}$$

**Remark 2.2.** This result will be key for adapting the functional $\mathrm{R}_{\text{m-prob}}$ to the setting of multi-class classification as well as to general hypothesis classes and loss functions. See Section 5

**Remark 2.3.** In its original expression, $\mathrm{R}_{\text{m-prob}}$ is written as the sum of the standard risk and a nonlocal perimeter functional. This *probabilistic perimeter*, which we will discuss in more detail in Section 3, can be rewritten as:

$$\mathrm{Per}_{\mathrm{prob}}(A) = \rho_0\left(\{x \in A^c \,:\, \mathfrak{m}_x(A) > p\}\right) + \rho_1\left(\{x \in A \,:\, \mathfrak{m}_x(A^c) > p\}\right).$$

Note that $\mathrm{Per}_{\mathrm{prob}}(A)$ counts the proportion of correctly classified points $x$ for which more than $100 \times p\,\%$ of their neighbors, sampled from $\mathfrak{m}_x$, constitute an attack. From this it should be intuitively clear that $\mathrm{Per}_{\mathrm{prob}}(A)$ is a boundary term.

**Remark 2.4.** The interpretation of the statement of this proposition in the light of Figure 1 is clear: Only if a point $x$ is correctly classified – meaning $\mathbf{1}_{\mathbf{1}_A(x)\neq y} = 0$ – the probabilistically robust regularization kicks in through the second term in the maximum. Points which are incorrectly classified will always be penalized

even if most attacks correct the label, *i.e.*, if $\mathbf{1}_{\mathbb{P}_{x'\sim\mathfrak{m}_x}[\mathbf{1}_A(x')\neq y]>p} = 0$. Thus, minimizing $\mathrm{R}_{\text{m-prob}}$ instead of $\mathrm{R}_{\text{prob}}$ corrects the pathology of the original PRL formulation.

*Proof of Proposition 2.1.* Disintegrating the risk functional $\mathrm{R}_{\text{std}}$ we can write

$$\mathrm{R}_{\text{std}}(A) = \int_{\mathcal{X}} \mathbf{1}_{x\in A}\,\mathrm{d}\rho_0(x) + \mathbf{1}_{x\in A^c}\,\mathrm{d}\rho_1(x).$$

Also,

$$\mathrm{Per}_{\text{prob}}(A) = \int_{\mathcal{X}} \mathbf{1}_{x\in A\vee \mathfrak{m}_x(A)>p} - \mathbf{1}_{x\in A}\,\mathrm{d}\rho_0(x) + \int_{\mathcal{X}} \mathbf{1}_{x\in A^c\vee \mathfrak{m}_x(A^c)>p} - \mathbf{1}_{x\in A^c}\,\mathrm{d}\rho_1(x).$$

Adding the above two identities we deduce (1.6). Now that (1.6) has been established we see that

$$\begin{aligned}&\mathrm{R}_{\text{m-prob}}(A)\\ &= \int_{\mathcal{X}} \mathbf{1}_{x\in A\vee \mathfrak{m}_x(A)>p}\,\mathrm{d}\rho_0(x) + \mathbf{1}_{x\in A^c\vee \mathfrak{m}_x(A^c)>p}\,\mathrm{d}\rho_1(x)\\ &= \int_{\mathcal{X}} \max\left\{\mathbf{1}_{x\in A}, \mathbf{1}_{\mathfrak{m}_x(A)>p}\right\}\,\mathrm{d}\rho_0(x) + \int_{\mathcal{X}} \max\left\{\mathbf{1}_{x\in A^c}, \mathbf{1}_{\mathfrak{m}_x(A^c)>p}\right\}\,\mathrm{d}\rho_1(x),\end{aligned}$$

where we used the fact that the indicator function of the union of two sets equals the maximum of the two indicator functions. (2.1) follows immediately. □

In the same spirit as in Proposition 2.1, we can rewrite $\mathrm{R}_{\text{m-prob}_\Psi}$ for general $\Psi$ as follows.

**Proposition 2.5.** *For the* 0-1 *loss* $\ell(\tilde{y}, y) := \mathbf{1}_{\tilde{y}\neq y}$ *and for any function* $\Psi : [0,1] \to [0,1]$ *we can rewrite the probabilistic risk* $\mathrm{R}_{\text{m-prob}_\Psi}$ *as a sample-wise maximum in the following way:*

$$\mathrm{R}_{\text{m-prob}_\Psi}(A) = \mathbb{E}_{(x,y)\sim\mu}\left[\max\left\{\ell(\mathbf{1}_A(x), y), \Psi\left(\mathbb{E}_{x'\sim\mathfrak{m}_x}\left[\ell(\mathbf{1}_A(x'), y)\right]\right)\right\}\right].$$

Besides the choice $\Psi(t) = \mathbf{1}_{t>p}$ another interesting choice is the smallest concave function which lies above it, *i.e.*, $\Psi(t) = \min\{t/p, 1\}$. For this choice we can connect the probabilistic risk with the conditional value at risk (CVaR) [31] which was suggested in [10] as a computationally feasible replacement for the $p$-ess sup operator in problem (PRL). Its precise definition will be important later and can be found in (5.2) in Section 5.

**Proposition 2.6.** *For the* 0-1 *loss* $\ell(\tilde{y}, y) := \mathbf{1}_{\tilde{y}\neq y}$ *and for* $\Psi_p(t) := \min\{t/p, 1\}$ *with* $p \in (0,1)$ *it holds that*

$$\mathrm{R}_{\text{m-prob}_{\Psi_p}}(A) = \mathbb{E}_{(x,y)\sim\mu}\left[\max\left\{\ell(\mathbf{1}_A(x), y), \mathrm{CVaR}_p\left(\ell(\mathbf{1}_A(\cdot), y); \mathfrak{m}_x\right)\right\}\right].$$

We proceed to prove Proposition 2.5.

*Proof of Proposition 2.5.* The proof is similar to that of Proposition 2.1 after noting that for $\Psi : [0,1] \to [0,1]$

$$\mathbf{1}_{x\in A} + \mathbf{1}_{x\in A^c}\Psi\left(\mathfrak{m}_x(A)\right) = \max\left\{\mathbf{1}_{x\in A}, \Psi\left(\mathfrak{m}_x(A)\right)\right\},$$

which can easily be shown by checking cases. Then:

$$\mathrm{R}_{\text{m-prob}_\Psi}(A) = \int_{\mathcal{X}} \mathbf{1}_{x\in A}\,\mathrm{d}\rho_0(x) + \int_{\mathcal{X}} \mathbf{1}_{x\in A^c}\,\mathrm{d}\rho_1(x)$$

$$
\begin{aligned}
&\quad + \int_{\mathcal{X}} \mathbf{1}_{x\in A^c} \Psi\left(\mathfrak{m}_x(A)\right) \,\mathrm{d}\rho_0(x) + \int_{\mathcal{X}} \mathbf{1}_{x\in A} \Psi\left(\mathfrak{m}_x(A^c)\right) \,\mathrm{d}\rho_1(x) \\
&= \int_{\mathcal{X}} \left[\mathbf{1}_{x\in A} + \mathbf{1}_{x\in A^c} \Psi\left(\mathfrak{m}_x(A)\right)\right] \,\mathrm{d}\rho_0(x) \\
&\quad + \int_{\mathcal{X}} \left[\mathbf{1}_{x\in A^c} + \mathbf{1}_{x\in A} \Psi\left(\mathfrak{m}_x(A^c)\right)\right] \,\mathrm{d}\rho_1(x) \\
&= \int_{\mathcal{X}} \max\left\{\mathbf{1}_{x\in A}, \Psi\left(\mathfrak{m}_x(A)\right)\right\} \,\mathrm{d}\rho_0(x) \\
&\quad + \int_{\mathcal{X}} \max\left\{\mathbf{1}_{x\in A^c}, \Psi\left(\mathfrak{m}_x(A^c)\right)\right\} \,\mathrm{d}\rho_1(x).
\end{aligned}
$$

□

We postpone the proof of Proposition 2.6, which discusses another reformulation of $\mathrm{R}_{\text{m-prob}\,\Psi}$ in terms of the so-called conditional value at risk (CVaR), to Section 5, since Proposition 2.6 will not be used in the remainder of this section.

## 2.2. Model family for hard classifiers

We now begin the discussion of existence of solutions to the minimization problems associated with the risks $\mathrm{R}_{\text{prob}\,\Psi}$ and $\mathrm{R}_{\text{m-prob}\,\Psi}$ defined in (1.7) and (1.8). As was discussed in the introduction, these families of risks include $\mathrm{R}_{\text{prob}}$ and $\mathrm{R}_{\text{m-prob}}$ (for $\Psi(t) := \mathbf{1}_{t>p}$) but also a relaxation of the adversarial risk $\mathrm{R}_{\text{adv}}$ (for $\Psi(t) := \mathbf{1}_{t>0}$).

We start our discussion by making our setting precise.

**Assumption 2.7.** We let $\mathcal{X}$ be a set and $\mathcal{A} \subset 2^{\mathcal{X}}$ be a $\sigma$-algebra. We assume that:

- $(\mathcal{X}\times\mathcal{Y}, \mathcal{A}\otimes 2^{\{0,1\}}, \mu)$ is a probability space;
- $(\mathcal{X}, \mathcal{A}, \rho)$ is a probability space, where we define $\rho(\bullet) := \mu(\bullet\times\{0,1\})$;
- $\{\mathfrak{m}_x\}_{x\in\mathcal{X}}$ is a family such that $(\mathcal{X}, \mathcal{A}, \mathfrak{m}_x)$ is a probability space for $\rho$-almost every $x\in\mathcal{X}$.

The measure $\rho$ equals the first marginal of $\mu$ and models the distribution of all data, irrespective of their label. It can be decomposed as

$$\rho := \rho_0 + \rho_1, \tag{2.2}$$

where $\rho_0, \rho_1$ are as in Section 1.2. We also define the probability measures

$$\sigma(A) := \int_{\mathcal{X}} \mathfrak{m}_x(A) \,\mathrm{d}\rho(x), \qquad A\in\mathcal{A}, \tag{2.3}$$

and

$$\nu(A) := \frac{1}{2}\sigma(A) + \frac{1}{2}\rho(A), \qquad A\in\mathcal{A}. \tag{2.4}$$

The measure $\sigma$ is the convolution of $\rho$ with the family of probability measures $\{\mathfrak{m}_x\}_{x\in\mathcal{X}}$ and can be interpreted as a noisy observation model for the clean data distribution $\rho$. $\nu$, on the other hand, is the equal-weight mixture of the measures $\sigma$ and $\rho$. In the sequel, we use $\mathrm{supp}(\rho)$ to denote the support of the probability measure $\rho$.

By construction, we have the following properties:

$$\sigma(A) = 0 \implies \mathfrak{m}_x(A) = 0 \text{ for } \rho\text{-almost every } x\in\mathcal{X}, \tag{2.5}$$

$$\nu(A) = 0 \implies \Big[\rho(A) = 0 \quad \text{and} \quad \mathfrak{m}_x(A) = 0 \text{ for } \rho\text{-almost every } x \in \mathcal{X}\Big], \tag{2.6}$$

$$\rho(A) = 0 \implies \Big[\rho_0(A) = 0 \quad \text{and} \quad \rho_1(A) = 0\Big]. \tag{2.7}$$

A simple example for $\mathcal{X} = \mathbb{R}^d$ is $\rho := \frac{1}{N}\sum_{i=1}^N \delta_{x_i}$ and $\mathfrak{m}_x := \mathrm{Unif}(B_\varepsilon(x))$, in which case

$$\nu = \frac{1}{2N}\sum_{i=1}^N \Big(\mathrm{Unif}(B_\varepsilon(x_i)) + \delta_{x_i}\Big)$$

is a sum of absolutely continuous measures on balls centered at $x_i$ and the empirical measure of the points $x_i$.

Using the measure $\nu$ we will work with the weak-* topology of $L^\infty(\mathcal{X};\nu)$ which is the dual space of $L^1(\mathcal{X};\nu)$ since $\nu$ is *a fortiori* a $\sigma$-finite measure [32], IV.8.3, Theorem 5.

**Definition 2.8.** Under Assumption 2.7 we say that a sequence of functions $(u_n)_{n\in\mathbb{N}} \subset L^\infty(\mathcal{X};\nu)$ converges to $u \in L^\infty(\mathcal{X};\nu)$ in the weak-* sense (written $u_n \overset{*}{\rightharpoonup} u$) as $n \to \infty$ if

$$\lim_{n\to\infty}\int_{\mathcal{X}} u_n\varphi \,\mathrm{d}\nu = \int_{\mathcal{X}} u\varphi \,\mathrm{d}\nu \qquad \forall \varphi \in L^1(\mathcal{X};\nu). \tag{2.8}$$

We will use this topology to prove existence of minimizers of the probabilistic risks $\mathrm{R}_{\mathrm{prob}\,\Psi}$ and $\mathrm{R}_{\text{m-prob}\,\Psi}$. Furthermore, we use this same topology for proving that our modified PRL using the function $\Psi_p$ converges to adversarial training as $p \to 0$, at least when the family of measures $\{\mathfrak{m}_x\}_x$ satisfies a suitable assumption; see the discussion in Section 4.

The following theorem establishes existence of minimizers of the risk $\mathrm{R}_{\text{m-prob}\,\Psi}$ for concave and non-decreasing functions $\Psi$. The existence result for hard classifiers is obtained using a non-standard abstract strategy, rather than the usual direct method in the calculus of variations. We take this approach since it is difficult to establish lower semicontinuity for an extension of $\mathrm{Per}_{\mathrm{prob}\,\Psi}$ to soft classifiers that behaves well with respect to thresholding operations. Instead, we introduce two different relaxations; one which is lower semicontinuous, and one which is defined through a coarea formula and hence behaves well with respect to thresholding; our result then follows from an ordering relation on the two relaxations if $\Psi$ is concave and non-decreasing.

We arrive at the rigorous versions of Informal Theorem 1.1 and Informal Theorem 1.2.

**Theorem 2.9.** *Suppose $\Psi : [0,1] \to [0,1]$ is concave and non-decreasing, and that Assumption 2.7 holds. Then, there exists a solution to the problem*

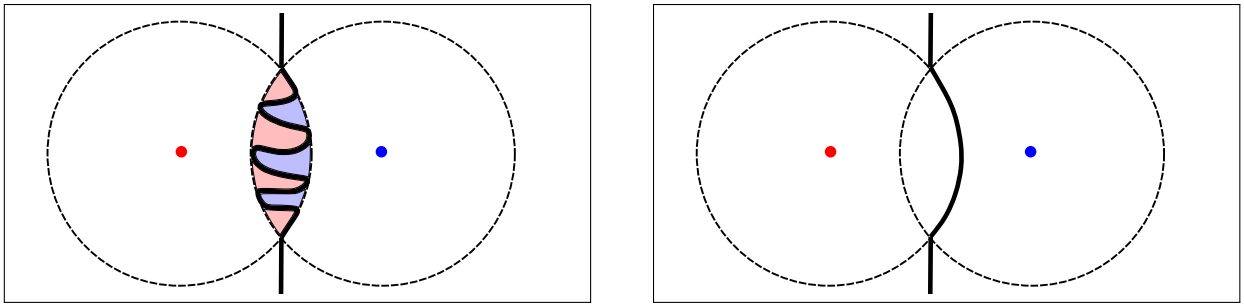
$$\inf_{A\in\mathcal{A}} \mathrm{R}_{\text{m-prob}\,\Psi}(A). \tag{2.9}$$

**Theorem 2.10.** *Suppose $\Psi : [0,1] \to [0,1]$ is concave and non-decreasing, and that Assumption 2.7 holds. Then there exists a solution to the problem*

$$\inf_{A\in\mathcal{A}} \mathrm{R}_{\mathrm{prob}\,\Psi}(A). \tag{2.10}$$

The proofs of Theorems 2.9 and 2.10 are provided in Section 2.4. The reason why an existence proof for (2.9) (or (2.10)) for non-concave functions $\Psi$ is currently not available is illustrated in the following example.

**Example 2.11** (Homogenizing solutions)**.** In this example we consider the situation of two data points (red and blue) which are so close that the $\varepsilon$-balls around them intersect. Furthermore, we assume that $p \in (0,1)$ is 0.9 times the ratio of the volume of the intersection to the volume of one $\varepsilon$-ball. If we pick $\Psi_p = \mathbf{1}_{t>p}$, there exist infinitely many minimizers of (2.9) with oscillatory boundaries in the intersection. One such minimizer is



(A) Homogenizing minimizer for $\Psi(t) := \mathbf{1}_{t>p}$. (B) Superlevel set which is not a minimizer.

FIGURE 2. Non-compactness of minimizing sequences.

depicted in Figure 2a. The minimizers can be arranged such that the ratios of the red or the blue volume to the volume of the intersection are in the interval $[0.4, 0.6]$. Consequently, all such sets have zero loss and are global minimizers. However, the set of such minimizers is not compact and has an accumulation point which is not a characteristic function, namely the function $u : \mathcal{X} \to [0, 1]$ with values $u = 1$ on the red ball without the blue ball, $u = 0$ on the blue ball without the red ball, and $u = 0.5$ on the intersection of the two balls. Hence, minimizing sequences of (2.9) are not compact. The issue of non-compactness of minimizing sequences can in fact also happen for concave functions $\Psi$ which is why in the proof of Theorem 2.9 we show that almost every superlevel set $\{u > t\}$ of a function $u$ which is an accumulation point of a minimizing sequence is a minimizer of (2.9). However, not even this strategy works in the non-concave case since the superlevel set $\{u > t\}$ *for all* $t \in [0, 1/2]$ is given by the set in Figure 2b which is not a minimizer of PRL.

## 2.3. Model family for soft classifiers

Note that, besides the reasons given in Example 2.11, an existence result similar to Theorem 2.9 for the non-concave function $\Psi(t) = \mathbf{1}_{t>p}$ is currently not available since our relaxation techniques rely on concavity of $\Psi$. However, in this section we shall state an existence theorem for "soft classifiers" which is valid for very general functions $\Psi$, including $\Psi(t) = \mathbf{1}_{t>p}$, since no relaxation is needed.

Such soft classifiers are particularly relevant since they include neural network based models with `Softmax` activation in the last layer which are used in practice. A suitable regularization functional for soft classifiers is given by $\mathrm{J}_{\mathrm{prob}_\Psi}$ defined in (1.10). For convenience we repeat its definition. Given an $\mathcal{A}$-measurable function $u : \mathcal{X} \to [0, 1]$ we let

$$\begin{aligned}\mathrm{J}_{\mathrm{prob}_\Psi}(u) := &\int_{\mathcal{X}} (1 - u(x))\, \Psi\left(\mathbb{E}_{x' \sim \mathfrak{m}_x}\left[u(x')\right]\right) \mathrm{d}\rho_0(x) \\ &+ \int_{\mathcal{X}} u(x) \Psi\left(\mathbb{E}_{x' \sim \mathfrak{m}_x}\left[1 - u(x')\right]\right) \mathrm{d}\rho_1(x)\end{aligned}$$

which satisfies $\mathrm{J}_{\mathrm{prob}_\Psi}(\mathbf{1}_A) = \mathrm{Per}_{\mathrm{prob}_\Psi}(A)$ for every choice of $\Psi$. Hence, it is a natural generalization of the perimeter to soft classifiers and one could call $\mathrm{J}_{\mathrm{prob}_\Psi}$ a total variation. However, it is neither positively 1-homogeneous nor convex so this name would be misleading. Instead, for the proof of Theorem 2.9 we shall also work with the *total variation functional* $\mathrm{TV}_{\mathrm{prob}_\Psi}$ defined as

$$\mathrm{TV}_{\mathrm{prob}_\Psi}(u) := \int_0^1 \mathrm{Per}_{\mathrm{prob}_\Psi}(\{u > t\})\, \mathrm{d}t, \tag{2.11}$$

for all $\mathcal{A}$-measurable functions $u : \mathcal{X} \to [0,1]$. Note that also the total variation satisfies $\mathrm{TV}_{\mathrm{prob}_\Psi}(\mathbf{1}_A) = \mathrm{Per}_{\mathrm{prob}_\Psi}(A)$. Both $\mathrm{J}_{\mathrm{prob}_\Psi}$ and $\mathrm{TV}_{\mathrm{prob}_\Psi}$ are of paramount importance for the proof of Theorem 2.9.

The next theorem is the rigorous version of Informal Theorem 1.3 and asserts existence of soft classifiers for the regularized risk minimization using $\mathrm{J}_{\mathrm{prob}_\Psi}$ for very general functions $\Psi$ and hypothesis classes $\mathcal{H}$. It is sufficient that $\Psi$ is lower semicontinuous, which is satisfied by every continuous function and also by $\Psi(t) = \mathbf{1}_{t>p}$ for $p \in [0,1]$. Furthermore, the existence theorem is valid for all hypotheses classes which are closed in the weak-* topology of $L^\infty(\mathcal{X};\nu)$.

**Theorem 2.12** (Existence of soft classifiers for modified PRL)**.** *Under Assumption 2.7, for every lower semicontinuous function $\Psi : [0,1] \to [0,1]$, and whenever $\mathcal{H}$ is a weak-* closed hypothesis class of $\mathcal{A}$-measurable functions $u : \mathcal{X} \to [0,1]$ in the sense of Definition 2.8, there exists a solution to the problem*

$$\inf_{u\in\mathcal{H}} \mathbb{E}_{(x,y)\sim\mu}\left[|u(x)-y|\right] + \mathrm{J}_{\mathrm{prob}_\Psi}(u). \tag{2.12}$$

**Example 2.13** (Hypothesis classes)**.** The hypothesis class consisting of measurable sets, *i.e.*, $\mathcal{H} = \{\mathbf{1}_A : A \in \mathcal{A}\}$ is not weak-* closed which is why Theorem 2.9 needs stronger assumptions on $\Psi$ than Theorem 2.12. Let us instead consider three interesting hypothesis classes of weak-* closed (in fact, even compact) classifiers for which Theorem 2.12 applies.

1. The simplest such class $\mathcal{H}$ is the class of *all* $\mathcal{A}$-measurable soft classifiers $u : \mathcal{X} \to [0,1]$ which could be referred to as *agnostic* classifiers since they are not parametrized. This class is a bounded subset of $L^\infty(\mathcal{X};\nu)$ and therefore, by the Banach–Alaoglu theorem, it is weak-* compact.
2. An example with more practical relevance is the class of (feedforward or residual) neural networks defined on the unit cube $\mathcal{X} := [-1,1]^d$ with uniformly bounded parameters

$$\mathcal{H} := \Big\{ \Phi_L \circ \cdots \circ \Phi_1 : [-1,1]^d \to [0,1] \ : \Phi_l(\bullet) = A_l \bullet + \sigma_l(W_l \bullet + b_l), \\ \|(A_l, W_l, b_l)\| \le C \ \forall l \in \{1,\dots,L\} \Big\},$$

   where we assume that the activations $\sigma_l : \mathbb{R} \to \mathbb{R}$ are continuous. Note that the boundedness of the weights cannot be relaxed. To see this, consider the (very simplistic) neural network $u_n(x) = \tanh(w_n x)$ for $x \in [-1,1]$ and $w_n \in \mathbb{R}$. For $w_n \to \infty$ it is easy to see that $u_n$ converges to $u(x) := \mathrm{sign}(x)$ which does not lie in the same hypothesis class.
   To argue why neural networks with bounded parameters and continuous activation functions are weak-* compact, let $(u_n)_{n\in\mathbb{N}} \subset \mathcal{H}$ be a sequence. Thanks to finite-dimensional compactness a subsequence of the associated parameters converge to some limiting parameters. The continuity of the activations implies that the associated neural networks converge (uniformly in the space of continuous functions on the unit cube $[-1,1]^d$) to a limiting neural network $u \in \mathcal{H}$. In particular, the convergence is true in the weak-* sense, which shows that $\mathcal{H}$ is weak-* compact.
3. Finally, one can also consider the class of hard linear classifiers on $\mathbb{R}^d$. Letting $\theta(t) := \mathbf{1}_{t>0}$ denote the Heaviside function, this class is given by

$$\mathcal{H} := \big\{\theta(w\cdot x + b) \ : \ w \in \mathbb{R}^d,\ |w| = 1,\ b \in [-\infty,\infty]\big\},$$

   where one interprets $u(x) := \theta(w\cdot x + b)$ as $u \equiv 1$ if $b = \infty$ and $u \equiv 0$ if $b = -\infty$. If the distributions $\rho_0$, $\rho_1$, and $\mathfrak{m}_x$ are are such that $\nu$ defined in (2.4) has a density with respect to the Lebesgue measure, then $\mathcal{H}$ has the desired closedness property. A sufficient condition for this to hold is that $\rho_0$, $\rho_1$, and $\mathfrak{m}_x$ have densities with respect to the Lebesgue measure.
   Let us prove the closedness under the above conditions. If $(u_n)_{n\in\mathbb{N}} \subset \mathcal{H}$ is a sequence of linear classifiers, thanks to finite-dimensional compactness a subsequence (which we do not relabel) of the associated

parameters $(w_n, b_n)$ converges to $w \in \mathbb{R}^d$ with $|w| = 1$ and $b \in [-\infty, \infty]$. For simplicity we only consider the case where $b \neq \pm\infty$. In this case one can define the half-spaces

$$A_n := \{x \in \mathbb{R}^d \,:\, w_n \cdot x + b > 0\},$$
$$A := \{x \in \mathbb{R}^d \,:\, w \cdot x + b > 0\},$$

upon which $u_n$ and $u$ are supported. Then for any $\phi \in L^1(\mathbb{R}^d; \nu)$ it holds

$$\left|\int_{\mathbb{R}^d} (u_n - u)\phi \,\mathrm{d}\nu\right| = \left|\int_{A_n} \phi \,\mathrm{d}\nu - \int_A \phi \,\mathrm{d}\nu\right| \leq \int_{A_n \triangle A} |\phi| \,\mathrm{d}\nu$$

where we used the symmetric difference $A_n \triangle A := (A_n \setminus A) \cup (A \setminus A_n)$. Note that this set is either a double cone (if $w_n \neq w$) or a strip of width $|b_n - b|$ (if $w_n = w$).
Since $\phi \in L^1(\mathbb{R}^d; \nu)$ and $\nu$ is a probability measure, for every $\varepsilon > 0$ there exists a compact set $K \subset \mathbb{R}^d$ such that $\int_{\mathbb{R}^d \setminus K} |\phi| \,\mathrm{d}\nu < \varepsilon$. Using this, we can compute

$$\left|\int_{\mathbb{R}^d} (u_n - u)\phi \,\mathrm{d}\nu\right| \leq \int_{A_n \triangle A} |\phi| \,\mathrm{d}\nu \leq \int_{(A_n \triangle A) \cap K} |\phi| \,\mathrm{d}\nu + \varepsilon.$$

Using that $\nu$ has a density with respect to the Lebesgue measure $\mathcal{L}^d$ and using also that $\mathcal{L}^d(A_n \triangle A \cap K) \to 0$ as $n \to \infty$, we obtain

$$\lim_{n\to\infty} \left|\int_{\mathbb{R}^d} (u_n - u)\phi \,\mathrm{d}\nu\right| \leq \varepsilon$$

and since $\varepsilon > 0$ was arbitrary we get

$$\lim_{n\to\infty} \int_{\mathbb{R}^d} (u_n - u)\phi \,\mathrm{d}\nu = 0,$$

which implies the weak-* convergence of $u_n$ to $u \in \mathcal{H}$ and hence the weak-* compactness of $\mathcal{H}$.
Note that for general measures $\nu$ the above argument fails. For instance,the sequence of linear classifiers $u_n(x) = \mathbf{1}_{x_1 > -1/n}$ has the natural limit $u(x) = \mathbf{1}_{x_1 > 0}$. However, if $\nu = \delta_0$ then $\int_{\mathbb{R}^d} u_n \,\mathrm{d}\nu = 1$ for all $n \in \mathbb{N}$ but $\int_{\mathbb{R}^d} u \,\mathrm{d}\nu = 0$, meaning that $u$ is not the weak-* limit of $u_n$.

Similarly, we can state a precise version of Informal Theorem 1.4.

**Theorem 2.14** (Existence of soft classifiers for original PRL)**.** *Let $\sigma$ be the probability measure defined in (2.3). Under Assumption 2.7, for every lower semicontinuous function $\Psi : [0,1] \to [0,1]$, and whenever $\mathcal{H}$ is a hypothesis class of $\mathcal{A}$-measurable functions $u : \mathcal{X} \to [0,1]$ that is closed in the weak* topology of $L^\infty(\mathcal{X}; \sigma)$, there exists a solution to the problem*

$$\inf_{u \in \mathcal{H}} \int_{\mathcal{X}} \Psi\left(\mathbb{E}_{x' \sim \mathfrak{m}_x}\left[u(x')\right]\right) \mathrm{d}\rho_0(x) + \int_{\mathcal{X}} \Psi\left(\mathbb{E}_{x' \sim \mathfrak{m}_x}\left[1 - u(x')\right]\right) \mathrm{d}\rho_1(x). \tag{2.13}$$

We observe that the above theorem applies to all of the hypothesis classes discussed in Example 2.13. Indeed, this is a consequence of the fact that, given the definitions of $\sigma$ and $\nu$, if a set is closed in the weak$^*$ topology of $L^\infty(\mathcal{X}; \nu)$ then it is also closed in the weak$^*$ topology of $L^\infty(\mathcal{X}; \sigma)$.

### 2.4. Existence proofs

In this section we will prove all theorems stated in the previous section using the direct method of calculus of variations together with non-standard relaxation arguments. For this we first outline our proof strategy in a very abstract way by using the quadruple

$$\mathfrak{Q} := (R,\, S,\, T,\, \mathcal{T}), \tag{2.14}$$

where $R : \mathcal{A} \to [0,\infty]$ is a functional defined on $\mathcal{A}$-measurable sets, $S, T : \mathcal{U} \to [0,\infty]$ are proper functionals defined on the unit ball $\mathcal{U} := \{u : \mathcal{X} \to [0,1]\ \mathcal{A}\text{-measurable}\}$ of the $\mathcal{A}$-measurable functions, and $\mathcal{T}$ is a topology on the unit ball. The key ingredients and properties of $\mathfrak{Q}$ are the following:

(D1) *Compactness*: $\mathcal{U}$ is compact with respect to $\mathcal{T}$;

(D2) *Lower semicontinuity*: For all sequences of $\mathcal{A}$-measurable functions $(u_n)_{n\in\mathbb{N}} \subset \mathcal{U}$ converging to $u \in \mathcal{U}$ in the topology $\mathcal{T}$ it holds

$$S(u) \le \liminf_{n\to\infty} S(u_n);$$

(D3) *Consistency*: It holds for all $A \in \mathcal{A}$ that

$$R(A) = S(\mathbf{1}_A).$$

(D4) *Coarea formula*: For all $u \in \mathcal{U}$ it holds

$$T(u) = \int_0^1 R(\{u \ge t\})\,\mathrm{d}t,$$

and in particular $R(A) = T(\mathbf{1}_A)$ for all $A \in \mathcal{A}$;

(D5) *Ordering*: For all $u \in \mathcal{U}$ it holds

$$T(u) \le S(u).$$

Under these conditions we can prove the following meta-theorem.

**Meta Theorem 2.15.** Assuming desiderata 2.4 adn 2.4 there exists

$$u \in \operatorname{argmin}_{u\in\mathcal{U}} S(u). \tag{2.15}$$

Assuming desiderata 2.4 to 2.4, for Lebesgue almost every $t \in [0,1]$ the set $A_t := \{u \ge t\}$ satisfies

$$A_t \in \operatorname{argmin}_{A\in\mathcal{A}} R(A). \tag{2.16}$$

*Proof.* Since $S$ is proper, there exists a minimizing sequence $\{u_n\}_{n\in\mathbb{N}} \subset \mathcal{U}$ for (2.15). By assumption, $\{u_n\}_{n\in\mathbb{N}}$ is precompact in the topology $\mathcal{T}$. Therefore, up to a subsequence that we do not relabel, we can assume that $u_n \to_{\mathcal{T}} u$ for some $u : \mathcal{X} \to [0,1]$ which solves (2.15) by lower semicontinuity of $S$.

Defining $A_t := \{u \ge t\}$ for $t \in [0,1]$ it holds

$$\inf_{A\in\mathcal{A}} R(A) = \int_0^1 \inf_{A\in\mathcal{A}} R(A)\,\mathrm{d}t \le \int_0^1 R(A_t)\,\mathrm{d}t = T(u) \le S(u) \le S(\mathbf{1}_A) = R(A) \quad \forall A \in \mathcal{A},$$

and therefore, after taking the infimum over $A \in \mathcal{A}$, we get

$$\int_0^1 R(A_t)\,\mathrm{d}t = \inf_{A\in\mathcal{A}} R(A). \tag{2.17}$$

Let us assume that for a subset $N \subset [0,1]$ of positive Lebesgue measure it holds

$$\inf_{A\in\mathcal{A}} R(A) < R(A_t) \quad \forall t \in N.$$

Integrating this inequality and using (2.17) we would then have

$$\inf_{A\in\mathcal{A}} R(A) < \int_0^1 R(A_t)\,\mathrm{d}t = \inf_{A\in\mathcal{A}} R(A),$$

which is a contradiction. Hence we have proved that for Lebesgue almost every $t \in [0,1]$ the set $A_t$ solves (2.16). □

**Remark 2.16.** We would like to point out that the assumptions that go into Theorem 2.15 are more general than what is typically found in the literature. In particular, $S$ and $R$ are *not* directly connected through a generalized coarea formula (*cf.* [33, 34])but only indirectly through the functional $T$ which satisfies $T \leq S$. The more standard setting of $S = T$, however, would not suffice for our purposes.

To use this theorem we need to specify $\mathfrak{Q}$ and verify the desiderata. We consider the following choices:

$$\begin{aligned} R(A) &:= \mathrm{R}_{\text{m-prob}\,\Psi}(A), \quad A \in \mathcal{A}, \\ S(u) &:= \mathbb{E}_{(x,y)\sim\mu}\left[|u(x)-y|\right] + \mathrm{J}_{\text{prob}\,\Psi}(u), \quad u \in \mathcal{U}, \\ T(u) &:= \mathbb{E}_{(x,y)\sim\mu}\left[|u(x)-y|\right] + \mathrm{TV}_{\text{prob}\,\Psi}(u), \quad u \in \mathcal{U}, \\ \mathcal{T} &:= \text{weak-* topology of } L^\infty(\mathcal{X};\nu), \end{aligned}$$

where $\nu$ was defined in (2.4). We note that desiderata 2.4 follows from the Banach–Alaoglu theorem. By definition of the functionals $\mathrm{J}_{\text{prob}\,\Psi}$ and $\mathrm{TV}_{\text{prob}\,\Psi}$ in (1.10) and (2.11) the desiderata 2.4 and 2.4 are satisfied, as well. It remains to prove the lower semicontinuity desiderata 2.4 and the ordering 2.4.

We start with the following lemma:

**Lemma 2.17.** *Under Assumption 2.7, let $\nu$ be a measure on $\mathcal{X}$ which satisfies $\mathfrak{m}_x \ll \nu$ for $\rho$-almost every $x \in \mathcal{X}$, and let $(u_n)_{n\in\mathbb{N}} \subset L^\infty(\mathcal{X};\nu)$ satisfy $u_n \overset{*}{\rightharpoonup} u$ in the sense of Definition 2.8. Then it holds*

$$\lim_{n\to\infty} \mathbb{E}_{x'\sim\mathfrak{m}_x}\left[u_n(x')\right] = \mathbb{E}_{x'\sim\mathfrak{m}_x}\left[u(x')\right] \qquad \text{for } \rho\text{-almost every } x \in \mathcal{X}.$$

*Proof.* Using the Radon–Nikodým theorem, the weak-* convergence implies that for $\rho$-almost every $x \in \mathcal{X}$ it holds

$$\begin{aligned} \lim_{n\to\infty} \mathbb{E}_{x'\sim\mathfrak{m}_x}\left[u_n(x')\right] &= \lim_{n\to\infty} \int_{\mathcal{X}} u_n(x')\,\mathrm{d}\mathfrak{m}_x(x') = \lim_{n\to\infty} \int_{\mathcal{X}} u_n(x') \frac{\mathrm{d}\mathfrak{m}_x}{\mathrm{d}\nu}(x')\,\mathrm{d}\nu(x') \\ &= \int_{\mathcal{X}} u(x') \frac{\mathrm{d}\mathfrak{m}_x}{\mathrm{d}\nu}(x')\,\mathrm{d}\nu(x') = \int_{\mathcal{X}} u(x')\,\mathrm{d}\mathfrak{m}_x(x') = \mathbb{E}_{x'\sim\mathfrak{m}_x}\left[u(x')\right] \end{aligned}$$

since $\frac{\mathrm{d}\mathfrak{m}_x}{\mathrm{d}\nu} \in L^1(\mathcal{X};\nu)$. □

Using Lemma 2.17 it is relatively straightforward to prove lower semicontinuity of $\mathrm{J}_{\text{prob}\,\Psi}$ if $\Psi$ is a continuous function. If, however, $\Psi$ is only lower semicontinuous, *e.g.*, $\Psi(t) = \mathbf{1}_{t>p}$, we will employ a suitable approximation from below of $\Psi$ by continuous functions.

**Proposition 2.18** (Lower semicontinuity of $\mathrm{J}_{\mathrm{prob}_\Psi}$)**.** *Under Assumption 2.7 let $(u_n)_{n\in\mathbb{N}} \subset L^\infty(\mathcal{X};\nu)$ be a sequence of functions with values in $[0,1]$ satisfying $u_n \overset{*}{\rightharpoonup} u$ in the sense of Definition 2.8, and let $\Psi : [0,1] \to [0,1]$ be lower semicontinuous. Then $0 \le u \le 1$ holds $\nu$-almost everywhere and furthermore*

$$\mathrm{J}_{\mathrm{prob}_\Psi}(u) \le \liminf_{n\to\infty} \mathrm{J}_{\mathrm{prob}_\Psi}(u_n).$$

*Proof.* First we show that $0 \le u \le 1$. By the weak-* lower semicontinuity of the $L^\infty$-norm we get $u \le 1$ from the fact that $0 \le u_n \le 1$. To show that $u \ge 0$ we assume that on a measurable set $N$ with $\nu(N) > 0$ it holds $u < 0$. Then from the weak-* convergence and the fact that $u_n \ge 0$ we obtain

$$0 > \int_{\mathcal{X}} u\mathbf{1}_N \,\mathrm{d}\nu = \lim_{n\to\infty} \int_{\mathcal{X}} u_n \mathbf{1}_N \,\mathrm{d}\nu \ge 0$$

which is a contradiction. Therefore, $u \ge 0$ holds $\nu$-almost everywhere.

Since both terms in the definition of $\mathrm{J}_{\mathrm{prob}_\Psi}$ are dealt with symmetrically, we assume without loss of generality and for an easier notation that $\rho_1 = 0$ and rewrite $\mathrm{J}_{\mathrm{prob}_\Psi}$ as

$$\mathrm{J}_{\mathrm{prob}_\Psi}(u) = \int_{\mathcal{X}} (1-u(x))\, \Psi\left(\mathbb{E}_{x'\sim\mathfrak{m}_x}\left[u(x')\right]\right) \mathrm{d}\rho_0(x).$$

Since $\Psi$ is lower semicontinuous there exists a sequence of continuous functions $\Psi_\delta : [0,1] \to [0,1]$ which converge to $\Psi$ in the pointwise sense as $\delta \to 0$ and satisfy $\Psi_\delta \le \Psi$. For instance, the functions

$$\Psi_\delta(t) := \inf_{s\in[0,1]} \Psi(s) + \frac{1}{\delta}\left|s-t\right|, \qquad t \in [0,1],$$

suffice. Lemma 2.17 implies that $\mathbb{E}_{x'\sim\mathfrak{m}_x}\left[u_n(x')\right] \to \mathbb{E}_{x'\sim\mathfrak{m}_x}\left[u(x')\right]$ for $\rho$-almost every $x$ as $n \to \infty$. Consequently, since $\Psi_\delta$ is continuous, we get $\Psi_\delta\left(\mathbb{E}_{x'\sim\mathfrak{m}_x}\left[u_n(x')\right]\right) \to \Psi_\delta\left(\mathbb{E}_{x'\sim\mathfrak{m}_x}\left[u(x')\right]\right)$ for $\rho$-almost every $x$ as $n \to \infty$. Since $0 \le u_n \le 1$ and hence $\Psi_\delta\left(\mathbb{E}_{x'\sim\mathfrak{m}_x}\left[u_n(x')\right]\right)$ is uniformly bounded, the convergence even holds true in $L^1(\mathcal{X};\rho)$. Taking into account (2.7), this implies convergence in $L^1(\mathcal{X};\rho_0)$ and therefore

$$\begin{aligned} &\lim_{n\to\infty} \int_{\mathcal{X}} (1-u_n(x))\, \Psi_\delta\left(\mathbb{E}_{x'\sim\mathfrak{m}_x}\left[u_n(x')\right]\right) \mathrm{d}\rho_0(x) \\ &= \int_{\mathcal{X}} (1-u(x))\, \Psi_\delta\left(\mathbb{E}_{x'\sim\mathfrak{m}_x}\left[u(x')\right]\right) \mathrm{d}\rho_0(x). \end{aligned} \tag{2.18}$$

Next we would like to use the Fatou lemma to take the limit as $\delta \to 0$ on both sides. For this we notice that the sequence of functions

$$f_\delta(x) := (1-u_n(x))\, \Psi_\delta\left(\mathbb{E}_{x'\sim\mathfrak{m}_x}\left[u_n(x')\right]\right)$$

converges to $(1-u_n(x))\, \Psi\left(\mathbb{E}_{x'\sim\mathfrak{m}_x}\left[u_n(x')\right]\right)$ pointwise as $\delta \to 0$ and satisfies the bounds $f_\delta \ge 0$. Thanks to the non-negativity we can apply the standard Fatou lemma. Using $\Psi_\delta \le \Psi$ and (2.18) we get

$$\begin{aligned} \mathrm{J}_{\mathrm{prob}_\Psi}(u) &= \int_{\mathcal{X}} (1-u(x))\, \Psi\left(\mathbb{E}_{x'\sim\mathfrak{m}_x}\left[u(x')\right]\right) \mathrm{d}\rho_0(x) \\ &\le \liminf_{\delta\to 0} \int_{\mathcal{X}} (1-u(x))\, \Psi_\delta\left(\mathbb{E}_{x'\sim\mathfrak{m}_x}\left[u(x')\right]\right) \mathrm{d}\rho_0(x) \\ &= \liminf_{\delta\to 0} \lim_{n\to\infty} \int_{\mathcal{X}} (1-u_n(x))\, \Psi_\delta\left(\mathbb{E}_{x'\sim\mathfrak{m}_x}\left[u_n(x')\right]\right) \mathrm{d}\rho_0(x) \end{aligned}$$

$$\leq \liminf_{n\to\infty} \int_{\mathcal{X}} (1-u_n(x))\, \Psi\left(\mathbb{E}_{x'\sim\mathfrak{m}_x}\left[u_n(x')\right]\right) \mathrm{d}\rho_0(x)$$
$$= \liminf_{n\to\infty} \mathrm{J}_{\mathrm{prob}_\Psi}(u_n).$$

□

We recall the definition of the total variation in (2.11) for which we now prove the following ordering with respect to $\mathrm{J}_{\mathrm{prob}_\Psi}$ defined in (1.10) in case that $\Psi$ is concave and non-decreasing.

**Proposition 2.19.** *If $\Psi$ is concave and non-decreasing, it holds for every $\mathcal{A}$-measurable function $u : \mathcal{X} \to [0,1]$ that*

$$\mathrm{TV}_{\mathrm{prob}_\Psi}(u) \leq \mathrm{J}_{\mathrm{prob}_\Psi}(u).$$

*Proof.* As in the proof of Proposition 2.18 we assume without loss of generality that $\rho_1 = 0$. We compute

$$\begin{aligned}
\mathrm{TV}_{\mathrm{prob}_\Psi}(u) &= \int_0^1 \mathrm{Per}_{\mathrm{prob}_\Psi}(\{u \geq t\})\, \mathrm{d}t \\
&= \int_{\mathcal{X}} \int_0^1 \mathbf{1}_{u(x)<t} \Psi\left(\mathbb{P}_{x'\sim\mathfrak{m}_x}\left[u(x') \geq t\right]\right) \mathrm{d}t\, \mathrm{d}\rho_0(x) \\
&= \int_{\mathcal{X}} \int_0^1 \mathbf{1}_{u(x)<t} \Psi\left(\mathbb{E}_{x'\sim\mathfrak{m}_x}\left[\mathbf{1}_{\{u\geq t\}}(x')\right]\right) \mathrm{d}t\, \mathrm{d}\rho_0(x) \\
&= \int_{\mathcal{X}} \int_{u(x)}^1 \Psi\left(\mathbb{E}_{x'\sim\mathfrak{m}_x}\left[\mathbf{1}_{\{u\geq t\}}(x')\right]\right) \mathrm{d}t\, \mathrm{d}\rho_0(x). \qquad (2.19)
\end{aligned}$$

Since $\Psi$ is concave and non-decreasing, we get from (2.19) and Jensen's inequality that

$$\begin{aligned}
\mathrm{TV}_{\mathrm{prob}_\Psi}(u) &\leq \int_{\mathcal{X}} (1-u(x))\, \Psi\left(\frac{1}{1-u(x)} \int_{u(x)}^1 \mathbb{E}_{x'\sim\mathfrak{m}_x}\left[\mathbf{1}_{\{u\geq t\}}(x')\right] \mathrm{d}t\right) \mathrm{d}\rho_0(x) \\
&= \int_{\mathcal{X}} (1-u(x))\, \Psi\left(\frac{1}{1-u(x)} \mathbb{E}_{x'\sim\mathfrak{m}_x}\left[\int_{u(x)}^1 \mathbf{1}_{\{u\geq t\}}(x')\, \mathrm{d}t\right]\right) \mathrm{d}\rho_0(x) \\
&= \int_{\mathcal{X}} (1-u(x))\, \Psi\left(\frac{1}{1-u(x)} \mathbb{E}_{x'\sim\mathfrak{m}_x}\left[\left(u(x')-u(x)\right)_+\right]\right) \mathrm{d}\rho_0(x) \\
&\leq \int_{\mathcal{X}} (1-u(x))\, \Psi\left(\frac{1}{1-u(x)} \mathbb{E}_{x'\sim\mathfrak{m}_x}\left[u(x')(1-u(x))\right]\right) \mathrm{d}\rho_0(x) \\
&= \int_{\mathcal{X}} (1-u(x))\, \Psi\left(\mathbb{E}_{x'\sim\mathfrak{m}_x}\left[u(x')\right]\right) \mathrm{d}\rho_0(x) = \mathrm{J}_{\mathrm{prob}_\Psi}(u).
\end{aligned}$$

□

A remarkable consequence of this lower bound and the lower semicontinuity of $\mathrm{J}_{\mathrm{prob}_\Psi}$ from Proposition 2.18 is the following proof of lower semicontinuity of $\mathrm{TV}_{\mathrm{prob}_\Psi}$ for sequences of characteristic functions that doesn't use any abstract functional analysis machinery).

**Corollary 2.20.** *If $\Psi$ is concave and non-decreasing, then for any sequence of measurable sets $(A_n) \subset \mathcal{A}$ such that $\mathbf{1}_{A_n} \overset{*}{\rightharpoonup} u$ in $L^\infty(\mathcal{X};\nu)$ it holds*

$$\mathrm{TV}_{\mathrm{prob}\,\Psi}(u) \leq \liminf_{n\to\infty} \mathrm{TV}_{\mathrm{prob}\,\Psi}(\mathbf{1}_{A_n}).$$

*Proof.* The result follows from Propositions 2.18 and 2.19 and the following computation

$$\begin{aligned}\mathrm{TV}_{\mathrm{prob}\,\Psi}(u) &\leq \mathrm{J}_{\mathrm{prob}\,\Psi}(u) \leq \liminf_{n\to\infty} \mathrm{J}_{\mathrm{prob}\,\Psi}(\mathbf{1}_{A_n}) = \liminf_{n\to\infty} \mathrm{Per}_{\mathrm{prob}\,\Psi}(A_n)\\ &= \liminf_{n\to\infty} \mathrm{TV}_{\mathrm{prob}\,\Psi}(\mathbf{1}_{A_n}).\end{aligned}$$

□

Now we are ready to prove Theorem 2.9 and 2.12 as special cases of the meta-theorem in the beginning of this section.

*Proof of Theorem 2.9* . The result follows from Theorem 2.15 noticing that desiderata 2.4 follows from Proposition 2.19 and desiderata 2.4 follows from Proposition 2.18 together with the trivial lower semicontinuity of the standard risk $u \mapsto \mathbb{E}_{(x,y)\sim\mu}\left[|u(x)-y|\right]$ since $\rho \ll \nu$. □

We remark that the main issue with proving the existence result of Theorem 2.9 for non-concave functions $\Psi$ is that the ordering desiderata 2.4 is violated, which was essential for the proof of Theorem 2.15. However, if we already consider the relaxed problem (2.12) of optimizing over soft classifiers instead of characteristic functions and regularizing with $\mathrm{J}_{\mathrm{prob}\,\Psi}$, no ordering is necessary and we obtain existence for very general functions $\Psi$.

*Proof of Theorem 2.12.* The result is the first part of Theorem 2.15 for the choices made in the proof of Theorem 2.9. □

**Remark 2.21.** From the proof of Theorem 2.15 it follows that when $\Psi$ is concave and non-decreasing, then

$$\min_{A\in\mathcal{A}} \mathrm{R}_{\text{m-prob}\,\Psi}(A) = \min_{u\in\mathcal{U}} \mathrm{S}_{\text{m-prob}\,\Psi}(u),$$

where

$$\mathrm{S}_{\text{m-prob}\,\Psi}(u) := \mathbb{E}_{(x,y)\sim\mu}\left[|u(x)-y|\right] + \mathrm{J}_{\mathrm{prob}\,\Psi}(u). \tag{2.20}$$

Next we discuss the existence results for the original PRL model.

*Proof of Theorem 2.10.* We apply Theorem 2.15 with the choices $R = \mathrm{R}_{\mathrm{prob}\,\Psi}$ and $S = \mathrm{S}_{\mathrm{prob}\,\Psi}$, where $\mathrm{S}_{\mathrm{prob}\,\Psi}$ is defined over soft classifiers $u : \mathcal{X} \to [0,1]$ according to:

$$\mathrm{S}_{\mathrm{prob}\,\Psi}(u) := \int_{\mathcal{X}} \Psi\left(\mathbb{E}_{x'\sim\mathfrak{m}_x}\left[u(x')\right]\right) \mathrm{d}\rho_0(x) + \int_{\mathcal{X}} \Psi\left(\mathbb{E}_{x'\sim\mathfrak{m}_x}\left[1-u(x')\right]\right) \mathrm{d}\rho_1(x), \tag{2.21}$$

which by definition satisfies $\mathrm{S}_{\mathrm{prob}\,\Psi}(\mathbf{1}_A) = \mathrm{R}_{\mathrm{prob}\,\Psi}(A)$. We also consider the functional

$$T(u) := \int_0^1 R(\{u \geq t\})\,\mathrm{d}t.$$

As topology $\mathcal{T}$ we choose the weak-* topology of $L^\infty(\mathcal{X};\sigma)$ where the probability measure $\sigma$ is as in (2.3).

It is straightforward to check that the quadruple $\mathfrak{Q} := (R, S, T, \mathcal{T})$ satisfies all five requirements for Theorem 2.15 to be applicable: (1) is clear, (2) is proved as Proposition 2.18 using Lemma 2.17 with $\sigma$ instead of $\nu$, (3)

and (4) are trivial consequences of the definitions, and (5) is an easy consequence of the layer cake representation and Jensen's inequality for concave $\Psi$. Note that (5) is only true if $\Psi$ is concave. □

*Proof of Theorem 2.14.* The result is the first part of Theorem 2.15 for the choices made in the proof of Theorem 2.10. □

**Remark 2.22.** As in Remark 2.21, it is easy to verify that when $\Psi$ is concave and non-decreasing, then

$$\min_{A\in\mathcal{A}} \mathrm{R}_{\mathrm{prob}_\Psi}(A) = \min_{u\in\mathcal{U}} \mathrm{S}_{\mathrm{prob}_\Psi}(u).$$

## 3. The functional $\mathrm{Per}_{\mathrm{prob}_\Psi}$ as a perimeter

In this section we shall discuss the interpretation of the functional $\mathrm{Per}_{\mathrm{prob}_\Psi}$ defined in (1.9) as a *perimeter*. We do this in two ways.

First, we focus on the case where $\Psi$ is concave and non-decreasing and prove that $\mathrm{Per}_{\mathrm{prob}_\Psi}$ is a *submodular functional*. If, in addition, $\Psi$ is assumed to satisfy $\Psi(0) = 0$, then

$$\mathrm{Per}_{\mathrm{prob}_\Psi}(\mathcal{X}) = \mathrm{Per}_{\mathrm{prob}_\Psi}(\emptyset) = 0.$$

Following [35], for $\Psi$ satisfying these properties one can interpret $\mathrm{Per}_{\mathrm{prob}_\Psi}$ as a generalized perimeter, *i.e.*, a functional that can be used to measure the "size" of the boundary of a set. This discussion is summarized in the next proposition.

**Proposition 3.1.** *If $\Psi(0) = 0$, then $\mathrm{Per}_{\mathrm{prob}_\Psi}(\mathcal{X}) = \mathrm{Per}_{\mathrm{prob}_\Psi}(\emptyset) = 0$. If $\Psi$ is concave and non-decreasing, then the functional $\mathrm{Per}_{\mathrm{prob}_\Psi}$ is submodular, meaning that*

$$\mathrm{Per}_{\mathrm{prob}_\Psi}(A\cup B) + \mathrm{Per}_{\mathrm{prob}_\Psi}(A\cap B) \le \mathrm{Per}_{\mathrm{prob}_\Psi}(A) + \mathrm{Per}_{\mathrm{prob}_\Psi}(B) \quad \forall A, B \in \mathcal{A}.$$

**Example 3.2.** For $\Psi(t) = t$ our perimeter reduces to the perimeter on the *random walk space* $(\mathcal{X}, \mathfrak{m})$, introduced in [36]: $\mathrm{Per}_{\mathrm{prob}_\Psi}(A) = \int_{\mathcal{X}\setminus A}\int_A \mathrm{d}\mathfrak{m}_x\,\mathrm{d}\rho_0(x) + \int_A\int_{\mathcal{X}\setminus A}\mathrm{d}\mathfrak{m}_x\,\mathrm{d}\rho_1(x)$.

For proving Proposition 3.1 we need the following lemma.

**Lemma 3.3.** *Let $\Psi : [0,\infty) \to \mathbb{R}$ be a concave and non-decreasing function, and let $0 \le a \le b \le b' \le a'$ be real numbers with $a + a' \le b + b'$. Then*

$$\Psi(a) + \Psi(a') \le \Psi(b) + \Psi(b').$$

*Proof.* Let $a, a', b, b'$ be as stated. Since $\Psi$ is concave and finite, it satisfies the fundamental theorem of calculus and thus it is possible to write

$$\Psi(s) = \Psi(a) + \int_a^s \Psi'(r)\,\mathrm{d}r, \quad s \ge a$$

for a function $\Psi'$ that is non-increasing and non-negative. It follows that

$$\Psi(b) - \Psi(a) = \int_a^b \Psi'(r)\,\mathrm{d}r \ge (b-a)\Psi'(b) \ge (a'-b')\Psi'(b) \ge \int_{b'}^{a'} \Psi'(r)\,\mathrm{d}r = \Psi(a') - \Psi(b'),$$

which is precisely what we wanted to show. □

With the above preliminary results in hand, we are ready to prove Proposition 3.1.

*Proof of Proposition 3.1.* First, the fact that $\mathrm{Per}_{\mathrm{prob}_\Psi}(\mathcal{X}) = \mathrm{Per}_{\mathrm{prob}_\Psi}(\emptyset) = 0$ if $\Psi(0) = 0$ is easy to see from the definition of $\mathrm{Per}_{\mathrm{prob}_\Psi}$. Second, we trivially have

$$\mathfrak{m}_x(A \cup B) + \mathfrak{m}_x(A \cap B) \leq \mathfrak{m}_x(A) + \mathfrak{m}_x(B).$$

Define

$$\begin{aligned} a' &:= \mathfrak{m}_x(A \cup B), & b' &:= \mathfrak{m}_x(B) \\ b &:= \mathfrak{m}_x(A), & a &:= \mathfrak{m}_x(A \cap B); \end{aligned}$$

without the loss of generality we can assume that $b$ and $b'$ defined above satisfy $b \leq b'$, for otherwise we can simply swap these labels. We can then use Lemma 3.3 to conclude that:

$$\Psi(\mathfrak{m}_x(A \cup B)) + \Psi(\mathfrak{m}_x(A \cap B)) \leq \Psi(\mathfrak{m}_x(A)) + \Psi(\mathfrak{m}_x(B)). \tag{3.1}$$

The submodularity follows directly once we have verified the following pointwise identity:

$$\begin{aligned} &\mathbf{1}_{x \in (A \cup B)^c} \Psi(\mathfrak{m}_x(A \cup B)) + \mathbf{1}_{x \in (A \cap B)^c} \Psi(\mathfrak{m}_x(A \cap B)) \\ &\leq \mathbf{1}_{x \in A^c} \Psi(\mathfrak{m}_x(A)) + \mathbf{1}_{x \in B^c} \Psi(\mathfrak{m}_x(B)). \end{aligned} \tag{3.2}$$

To do this we consider two complementary cases:

**Case 1,** $x \in (A \cup B)^c$**:** This is equivalent to $x \in A^c \cap B^c$. Furthermore, since $(A \cup B)^c \subset (A \cap B)^c$ we also have that $x \in (A \cap B)^c$. Hence, all indicator functions in (3.2) take the value one and (3.2) is the same as (3.1), which we have already verified.

**Case 2,** $x \in A \cup B$**:** In this case the first indicator function on the left hand side of (3.2) is zero.

**Case 2.1,** $x \in A \cap B$**:** In this subcase all indicator functions are equal to zero and the inequality is trivially satisfied.

**Case 2.2,** $x \in A \cup B \setminus (A \cap B)$**:** Without loss of generality we can assume that $x \in A \setminus B = A \cap B^c$. In this case only the second indicator function on the left hand side and the second one on the right hand side of (3.2) take the value one and the inequality reduces to the trivial inequality

$$\Psi(\mathfrak{m}_x(A \cap B)) \leq \Psi(\mathfrak{m}_x(B))$$

which is true since $\Psi$ is non-decreasing and $A \cap B \subset B$. □

**Remark 3.4.** From the submodularity of $\mathrm{Per}_{\mathrm{prob}_\Psi}$ one can show that the functional $\mathrm{TV}_{\mathrm{prob}_\Psi}$ defined in (2.11) is convex. Indeed, this can be proved following the proof of Proposition 3.4 in [33], where, actually, we do not need to assume the lower semicontinuity of the functional *a priori*.

Next, we consider rather general $\Psi$ and show that $\mathrm{Per}_{\mathrm{prob}_\Psi}$ is related to a standard *local* perimeter when $\mathcal{X} = \mathbb{R}^d$ and the probability measure $\mathfrak{m}_x$ localizes to a Dirac delta $\delta_x$; for the case of adversarial training such a connection was proved in [23], where the authors utilized the notion of $\Gamma$-convergence of functionals. We take a first step in this direction by proving that for sufficiently smooth sets the probabilistic perimeter converges to a local one if the family of probability distributions $\mathfrak{m}_x$ localizes suitably. For example, one could think of $\mathfrak{m}_x := \mathrm{Unif}(B_\varepsilon(x))$, which converges to a point mass at $x$ if $\varepsilon \to 0$. To make our setting precise, we pose the following general assumption:

**Assumption 3.5.** We assume that $\mathcal{X} = \mathbb{R}^d$, $\Psi(0) = 0$, $\Psi$ is Borel measurable and bounded, and $\rho_1, \rho_0$ have continuous densities with respect to the Lebesgue measure which we shall also denote as $\rho_1, \rho_0$. Furthermore,

we assume that there is $\varepsilon > 0$ and a measurable function $K : \mathcal{X} \times \mathbb{R}^d \to [0, \infty)$ such that for every $x \in \mathbb{R}^d$ we have the representation

$$\mathrm{d}\mathfrak{m}_x(x') = \varepsilon^{-d} K\left(x, \frac{x' - x}{\varepsilon}\right) \mathrm{d}x'.$$

We also assume that for every $x \in \mathcal{X}$ we have $K(x, \bullet) \in L^1(\mathbb{R}^d)$, $\int_{\mathbb{R}^d} K(x, z)\,\mathrm{d}z = 1$, $K(x, z) = 0$ if $|z| > 1$, and that for every $z \in \mathbb{R}^d$ the mapping $x \mapsto K(x, z)$ is continuous.

In the more difficult case when $\Psi$ is not necessarily continuous, we shall need some additional assumptions on the kernel.

**Assumption 3.6.** For our extra assumptions on the kernel, we shall require that for every $z \in \mathbb{R}^d$ the mapping $x \mapsto K(x, z)$ is $C^1$ and for every $x \in \mathbb{R}^d$ the mapping $z \mapsto K(x, z)$ is lower semicontinuous. Furthermore, for all $x \in \mathbb{R}^d$, $t \in (-1, 1)$, and $n \in S^{d-1}$, the Radon transform of $K$ with respect to the $z$-variable

$$\mathcal{R}(K(x, \cdot))(n, t) := \int_{\{z \cdot n = t\}} K(x, z)\,\mathrm{d}\mathcal{H}^{d-1}(z)$$

is strictly positive.

The following theorem derives the asymptotics of the probabilistic perimeter as $\varepsilon \to 0$ and is the rigorous counterpart of Informal Theorem 1.6.

**Theorem 3.7.** *Let $\mathcal{X} = \mathbb{R}^d$. Under Assumption 3.5, if $A$ has a compact $C^{1,1}$ boundary and either $\Psi$ is continuous or $K$ satisfies the additional Assumption 3.6, then*

$$\lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \operatorname{Per}_{\operatorname{prob}_\Psi}(A) = \int_{\partial A} \sigma_{0,\Psi}\left[x, n(x)\right] \rho_0(x) + \sigma_{1,\Psi}\left[x, n(x)\right] \rho_1(x)\,\mathrm{d}\mathcal{H}^{d-1}(x), \tag{3.3}$$

*where we let $n(x)$ denote the normal to $\partial A$ at a point $x \in \partial A$, and for any vector $v \in \mathbb{R}^d$ we define*

$$\sigma_{0,\Psi}\left[x, v\right] := \int_0^1 \Psi\left(\int_{\{z \cdot v \le -t\}} K(x, z)\,\mathrm{d}z\right) \mathrm{d}t, \quad \sigma_{1,\Psi}\left[x, v\right] := \int_0^1 \Psi\left(\int_{\{z \cdot v \ge t\}} K(x, z)\,\mathrm{d}z\right) \mathrm{d}t.$$

**Remark 3.8.** If $K$ is radially symmetric and independent of $x \in \mathcal{X}$, then $\sigma_{0,\Psi} = \sigma_{1,\Psi} =: \sigma_\Psi$ is just a constant. E.g., for $K(x, z) := |B_1(0)|^{-1} \mathbf{1}_{|z| \le 1}$ and $\Psi(t) = \mathbf{1}_{t > p}$ it is trivial that for $p = 0$ we have $\sigma_\Psi = 1$. However, for $p \ge \frac{1}{2}$ one easily sees $\sigma_\Psi = 0$, hence the limiting perimeter equals zero and there is no regularization effect. Using the function $\Psi(t) = \min\{t/p, 1\}$ corrects this degeneracy.

Notably, for radially symmetric $K$ the limiting perimeter in (3.3) coincides, provided $\sigma_\Psi > 0$, with the one derived for adversarial training (problem (1.2)) in [23], although they considered more general (potentially discontinuous) densities $\rho_i$. In particular, our result indicates that for very small adversarial budgets the regularization effect of both probabilistically robust learning and adversarial training is dominated by the perimeter in (3.3). While Theorem 3.7 already completes half of the proof (namely the limsup inequality) of $\Gamma$-convergence of $\frac{1}{\varepsilon} \operatorname{Per}_{\operatorname{prob}_\Psi}$ to the limiting perimeter, the remaining liminf inequality is beyond the scope of this paper. Proving that the convergence (3.3) does not only hold for sufficiently smooth sets as assumed in Theorem 3.7 but even in the sense of $\Gamma$-convergence is an extremely important topic for future work since only $\Gamma$-convergence allows to deduce from the convergence of the perimeters that also the solutions of probabilistically robust learning converge to certain regular Bayes classifiers as $\varepsilon \to 0$, see [23], Section 4.2. The exploration of this is left for future work.

*Proof of Theorem 3.7.* UnderAssumption 3.5, a simple change of variables shows

$$\begin{aligned}\mathrm{Per}_{\mathrm{prob}\,\Psi}(A) = &\int_{A^c} \Psi\left(\int_{\mathbb{R}^d} \mathbf{1}_A(x+\varepsilon z)K(x,z)\,\mathrm{d}z\right)\rho_0(x)\,\mathrm{d}x\\ &+\int_A \Psi\left(\int_{\mathbb{R}^d} \mathbf{1}_{A^c}(x+\varepsilon z)K(x,z)\,\mathrm{d}z\right)\rho_1(x)\,\mathrm{d}x.\end{aligned}$$

Let $\tau:\mathbb{R}^d\to\mathbb{R}$ be the signed distance function to $\partial A$ such that $\tau(x)\le 0$ for $x\in A$. Using $\tau$, we can rewrite the previous line as

$$\begin{aligned}\mathrm{Per}_{\mathrm{prob}\,\Psi}(A) = &\int_{\{\tau(x)\ge 0\}} \Psi\left(\int_{\{\tau(x+\varepsilon z)\le 0\}} K(x,z)\,\mathrm{d}z\right)\rho_0(x)\,\mathrm{d}x\\ &+\int_{\{\tau(x)\le 0\}} \Psi\left(\int_{\{\tau(x+\varepsilon z)\ge 0\}} K(x,z)\,\mathrm{d}z\right)\rho_1(x)\,\mathrm{d}x.\end{aligned}$$

Recalling that $K(x,z)=0$ whenever $|z|>1$, it follows that

$$\begin{aligned}\mathrm{Per}_{\mathrm{prob}\,\Psi}(A) = &\int_{\{0\le\tau(x)\le\varepsilon\}} \Psi\left(\int_{\{\tau(x+\varepsilon z)\le 0\}} K(x,z)\,\mathrm{d}z\right)\rho_0(x)\,\mathrm{d}x\\ &+\int_{\{-\varepsilon\le\tau(x)\le 0\}} \Psi\left(\int_{\{\tau(x+\varepsilon z)\ge 0\}} K(x,z)\,\mathrm{d}z\right)\rho_1(x)\,\mathrm{d}x.\end{aligned}$$

In the rest of what follows, we will focus on proving that

$$\lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\int_{\{0\le\tau(x)\le\varepsilon\}} \Psi\left(\int_{\{\tau(x+\varepsilon z)\le 0\}} K(x,z)\,\mathrm{d}z\right)\rho_0(x)\,\mathrm{d}x = \int_{\partial A}\sigma_{0,\Psi}\left[x,n(x)\right]\rho_0(x)\,\mathrm{d}\mathcal{H}^{d-1}(x),$$

the argument for showing that

$$\lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\int_{\{-\varepsilon\le\tau(x)\le 0\}} \Psi\left(\int_{\{\tau(x+\varepsilon z)\ge 0\}} K(x,z)\,\mathrm{d}z\right)\rho_1(x)\,\mathrm{d}x = \int_{\partial A}\sigma_{1,\Psi}\left[x,n(x)\right]\rho_1(x)\,\mathrm{d}\mathcal{H}^{d-1}(x)$$

will be identical.

Since $\partial A$ is $C^{1,1}$, there exists some $\varepsilon_0>0$ such that for all $\varepsilon<\varepsilon_0$ the map $T_\varepsilon(y,t):\partial A\times[-1,1]\to\{x\in\mathbb{R}^d : 0\le\tau(x)\le\varepsilon\}$ given by $T_\varepsilon(y,t)=y+tn(y)$ is a bijection and $\varepsilon^{-1}\det(DT_\varepsilon)$ converges uniformly to 1 as $\varepsilon\to 0$. Using this change of variables, we may write

$$\begin{aligned}&\frac{1}{\varepsilon}\int_{\{0\le\tau(x)\le\varepsilon\}} \Psi\left(\int_{\{\tau(x+\varepsilon z)\le 0\}} K(x,z)\,\mathrm{d}z\right)\rho_0(x)\,\mathrm{d}x\\ &=\int_{\partial A}\int_0^1 \frac{1}{\varepsilon}\det(DT_\varepsilon(y,t))\Psi\left(a_\varepsilon(y,t)\right)\rho_0(y+\varepsilon tn(y))\,\mathrm{d}t\,\mathrm{d}\mathcal{H}^{d-1}(y),\end{aligned}$$

where we abbreviate

$$a_\varepsilon(y,t) := \int_{\{\tau(y+\varepsilon(tn(y)+z))\le 0\}} K(y+\varepsilon tn(y),z)\,\mathrm{d}z$$

Since we know that $\lim_{\varepsilon\to 0}\frac{\det(DT_\varepsilon(y,t))}{\varepsilon}=1$ and $\lim_{\varepsilon\to 0}\rho_0(y+\varepsilon t n(y))=\rho_0(y)$ pointwise almost everywhere, the main difficulty lies in passing to the limit in the term involving $\Psi$. For this we shall first prove convergence of $a_\varepsilon(y,t)$ to

$$a(y,t):=\int_{\{z\cdot n(y)\le -t\}} K(y,z)\,\mathrm{d}z$$

Since $\nabla\tau(y)=n(y)$ for any $y\in\partial A$, we have the expansion

$$\tau(y+\varepsilon(tn(y)+z))=\varepsilon(t+z\cdot n(y))+O(\varepsilon^2).$$

It now follows from our assumptions on $K$ that for all $y\in\partial A$ and all $t\in[0,1]$

$$\lim_{\varepsilon\to 0} a_\varepsilon(y,t)=a(y,t).$$

If $\Psi$ is continuous, the result now follows from dominated convergence. If $\Psi$ is not continuous, then we must work harder.

Since $\Psi\in L^1([0,1])\cap L^\infty([0,1])$, we can always find a sequence of smooth functions $\Psi_n$ such that $\|\Psi_n\|_{L^\infty([0,1])}\le\|\Psi\|_{L^\infty([0,1])}$ and $\Psi_n$ converges pointwise almost everywhere to $\Psi$ on $[0,1]$. If we can show that

$$\lim_{n\to\infty}\limsup_{\varepsilon\to 0}\left|\int_{\partial A}\int_0^1\big(\Psi(a_\varepsilon(y,t))-\Psi_n(a_\varepsilon(y,t))\big)\,\mathrm{d}t\,\mathrm{d}\mathcal{H}^{d-1}(y)\right|=0,$$

and

$$\lim_{n\to\infty}\left|\int_{\partial A}\int_0^1\big(\Psi(a(t,y))-\Psi_n(a(t,y))\big)\,\mathrm{d}t\,\mathrm{d}\mathcal{H}^{d-1}(y)\right|=0,$$

then we can safely approximate $\Psi$ by $\Psi_n$ and use the previous argument where $\Psi$ was continuous to conclude the result. Hence, it remains to show that the above integrals vanish.

Viewing $a_\varepsilon$ and $a$ as maps from $\partial A\times[0,1]\to[0,1]$ we can construct measures $\mu_\varepsilon,\mu$ on $[0,1]$ *via* pushforwards by setting

$$\int_0^1 g(s)\,\mathrm{d}\mu_\varepsilon(s):=\int_0^1\int_{\partial A} g(a_\varepsilon(y,t))\,\mathrm{d}t\,\mathrm{d}\mathcal{H}^{d-1}(y)$$

$$\int_0^1 g(s)\,\mathrm{d}\mu(s):=\int_0^1\int_{\partial A} g(a(y,t))\,\mathrm{d}t\,\mathrm{d}\mathcal{H}^{d-1}(y)$$

for any bounded continuous function $g:[0,1]\to\mathbb{R}$. Using these definitions, we can write

$$\left|\int_{\partial A}\int_0^1\big(\Psi(a_\varepsilon(t,y))-\Psi_n(a_\varepsilon(t,y))\big)\,\mathrm{d}t\,\mathrm{d}\mathcal{H}^{d-1}(y)\right|=\left|\int_0^1\big(\Psi(s)-\Psi_n(s)\big)\,\mathrm{d}\mu_\varepsilon(s)\right|,$$

and

$$\left|\int_{\partial A}\int_0^1\big(\Psi(a(t,y))-\Psi_n(a(t,y))\big)\,\mathrm{d}t\,\mathrm{d}\mathcal{H}^{d-1}(y)\right|=\left|\int_0^1\big(\Psi(s)-\Psi_n(s)\big)\,\mathrm{d}\mu(s)\right|.$$

If we can show that the measures $\mu_\varepsilon$ are uniformly absolutely continuous with respect to the Lebesgue measure on $[0,1]$, namely, that

$$\lim_{|E|\to 0} \limsup_{\varepsilon\to 0} \mu_\varepsilon(E) = 0,$$

then the pointwise almost everywhere convergence of $\Psi_n$ to $\Psi$ along with the control $\|\Psi_n\|_{L^\infty([0,1]} \leq \|\Psi\|_{L^\infty([0,1]}$ will imply that

$$\lim_{n\to\infty} \limsup_{\varepsilon\to 0} \left| \int_0^1 \big(\Psi(t') - \Psi_n(t')\big)\, \mathrm{d}\mu_\varepsilon(t') \right| = 0,$$

and for free it will give us

$$\lim_{n\to\infty} \left| \int_0^1 \big(\Psi(t') - \Psi_n(t')\big)\, \mathrm{d}\mu(t') \right| = 0,$$

since $\mu$ is the distributional limit of the $\mu_\varepsilon$ (from the pointwise convergence of $a_\varepsilon$ to $a$).

To prove that the $\mu_\varepsilon$ are uniformly absolutely continuous with respect to the Lebesgue measure on $[0,1]$, we will show that $|\partial_t a_\varepsilon|$ is strictly bounded away from 0 whenever $t$ is bounded away from 1. To do this, we shall need the additional Assumption 3.6 on our kernel $K$. Let us first assume that $K$ is $C^1$ in both variables. Changing variables $z' = z + tn(y)$ we can write

$$a_\varepsilon(y,t) = \int_{\{\tau(y+\varepsilon z')\leq 0\}} K(y+\varepsilon t n(y), z' - tn(y))\, \mathrm{d}z',$$

and hence

$$\begin{aligned}\partial_t a_\varepsilon(y,t) = &\int_{\{\tau(y+\varepsilon z')\leq 0\}} \varepsilon \nabla_y K(y+\varepsilon t n(y), z' - tn(y)) \cdot n(y)\, \mathrm{d}z' \\ &- \int_{\{\tau(y+\varepsilon z')\leq 0\}} \nabla_{z'} K(y+\varepsilon t n(y), z' - tn(y)) \cdot n(y)\, \mathrm{d}z'.\end{aligned}$$

Since the second term is the complete derivative with respect to $z'$, we can integrate by parts to obtain

$$\begin{aligned}\partial_t a_\varepsilon(y,t) = &\int_{\{\tau(y+\varepsilon z')\leq 0\}} \varepsilon \nabla_y K\big(y+\varepsilon t n(y), z' - tn(y)\big) \cdot n(y)\, \mathrm{d}z' \\ &- \int_{\{\tau(y+\varepsilon z')=0\}} K\big(y+\varepsilon t n(y), z' - tn(y)\big) n(y) \cdot \nabla\tau(y+\varepsilon z')\, \mathrm{d}\mathcal{H}^{d-1}(z')\end{aligned}$$

Using the expansion $\nabla\tau(y+\varepsilon z) = n(y) + O(\varepsilon)$ and the smoothness properties of $K$, we see that

$$\partial_t a_\varepsilon(y,t) = -\int_{\{\tau(y+\varepsilon z')=0\}} K(y, z' - tn(y))\, \mathrm{d}\mathcal{H}^{d-1}(z') + O(\varepsilon)$$

where we note that the constant in the big O bound does not depend on the differentiability of $K$ with respect to the $z$ variable. Hence, after approximating $K$ with $C^1$ kernels, we can assume the above control holds for any kernel $K$ satisfying both Assumptions 3.5 and 3.6.

Thus, for each fixed $y \in \partial A, t \in [0,1]$ we have

$$\liminf_{\varepsilon\to 0} |\partial_t a_\varepsilon(t,y)| \geq \int_{\{z'\cdot n(y)=0\}} K(y, z' - tn(y))\,\mathrm{d}\mathcal{H}^{d-1}(z') = \int_{\{z\cdot n(y)=-t\}} K(y,z)\,\mathrm{d}\mathcal{H}^{d-1}(z),$$

where we have used the lower semicontinuity of $K$ with respect to $z$. We can also now recognize that the right hand side is the Radon transform $\mathcal{R}(K(y,\cdot))(-t, n(y))$. From our additional Assumption 3.6, we know that $\mathcal{R}(K(y,\cdot))(t, n(y)) > 0$ for each fixed $y \in \partial A$, $t \in (-1,1)$ and $n \in S^{d-1}$. Fix some $\delta > 0$. Since the set $\partial A \times [0, 1-\delta] \times S^{d-1}$ is compact, it follows that there exists some $c_\delta > 0$ such that $\mathcal{R}(K(y,\cdot))(-t, n(y)) \geq c_\delta$ for all $y \in \partial A, n(y) \in S^{d-1}, t \in [0, 1-\delta]$. Therefore,

$$\liminf_{\varepsilon\to 0} |\partial_t a_\varepsilon(t,y)| \geq c_\delta$$

for all $y \in \partial A$ and $t \in [0, 1-\delta]$. Thus, for all $\varepsilon > 0$ sufficiently small we have

$$|\partial_t a_\varepsilon(t,y)| \geq c_\delta/2$$

for all $y \in \partial A$ and $t \in [0, 1-\delta]$.

Now we are ready to show that $\mu_\varepsilon$ is uniformly absolutely continuous with respect to the Lebesgue measure on $[0,1]$. Let $\mathcal{L}_{[a,b]}$ denote the Lebesgue measure on the interval $[a,b]$. From our work above, for all $\varepsilon > 0$ sufficiently small, we have the inequality

$$\mu_\varepsilon = a_{\varepsilon\,\#}\big(\mathcal{H}^{d-1}_{\partial A} \otimes \mathcal{L}_{[0,1]}\big) \leq \frac{2\mathcal{H}^{d-1}(\partial A)}{c_\delta}\mathcal{L}_{[0,1]} + a_{\varepsilon\,\#}\big(\mathcal{H}^{d-1}_{\partial A} \otimes \mathcal{L}_{[1-\delta,1]}\big).$$

Thus, for any measurable $E \subset [0,1]$, we have

$$\begin{aligned}\mu_\varepsilon(E) \leq& 2|E|\frac{\mathcal{H}^{d-1}(\partial A)}{c_\delta} + \int_{a_\varepsilon^{-1}(E)\cap(\partial A\times[1-\delta,1])} \mathrm{d}t\,\mathrm{d}\mathcal{H}^{d-1}(y)\\ \leq& 2|E|\frac{\mathcal{H}^{d-1}(\partial A)}{c_\delta} + \delta\mathcal{H}^{d-1}(\partial A).\end{aligned}$$

Hence, for all $\delta > 0$,

$$\lim_{|E|\to 0}\limsup_{\varepsilon\to 0} \mu_\varepsilon(E) \leq \delta\mathcal{H}^{d-1}(\partial A),$$

which implies that $\lim_{|E|\to 0}\limsup_{\varepsilon\to 0}\mu_\varepsilon(E) = 0$. This completes the proof.

□

## 4. Interpolation properties of probabilistically robust learning

In this section we discuss the limiting behavior of the energies $\mathrm{R}_{\text{m-prob}\,\Psi_p}$ and $\mathrm{R}_{\text{prob}\,\Psi_p}$, for $\Psi_p(t) := \min\{t/p, 1\}$, as $p \to 0$ or as $p$ grows. This limiting behavior is studied in the sense of $\Gamma$-convergence, which, in particular, is closely connected to the convergence of minimizers of the sequence of functionals. We start by recalling the definition of this notion of convergence for functionals defined over arbitrary metric spaces.

**Definition 4.1.** Given a topological space $(\mathcal{U}, \tau)$ and a sequence of functionals $F_n : \mathcal{U} \to [0,\infty]$, and another functional $F : \mathcal{U} \to [0,\infty]$, we say that the sequence of functionals $F_n$ $\Gamma$-converges toward $F$ (with respect to the topology $\tau$) if the following two conditions hold:

1. **Liminf inequality:** For any $u \in \mathcal{U}$ and any sequence $\{u_n\}_{n\in\mathbb{N}}$ converging (in the $\tau$ topology) toward $u$ we have:
$$\liminf_{n\to\infty} F_n(u_n) \geq F(u).$$
2. **Limsup inequality:** For every $u \in \mathcal{U}$ there exists a sequence $\{u_n\}_{n\in\mathbb{N}}$ converging (in the $\tau$ topology) toward $u$ such that:
$$\limsup_{n\to\infty} F_n(u_n) \leq F(u).$$

The relevance of the notion of Γ-convergence becomes apparent when an additional compactness property holds.

**Proposition 4.2** (Fundamental Theorem of Γ-convergence; see [37] and [38])**.** *Let $(\mathcal{U},\tau)$ be a topological space. Let $F_n : \mathcal{U} \to [0,\infty]$ be a sequence of functionals Γ-converging toward a functional $F : \mathcal{U} \to [0,\infty]$ that is not identically equal to $+\infty$. Suppose, in addition, that the sequence of functionals $\{F_n\}_{n\in\mathbb{N}}$ satisfies the following compactness property: any sequence $\{u_n\}_{n\in\mathbb{N}}$ satisfying*
$$\sup_{n\in\mathbb{N}} F_n(u_n) < \infty$$
*is precompact in the topology $\tau$, i.e., every subsequence of $\{u_n\}_{n\in\mathbb{N}}$ has a further subsequence that converges to an element of $\mathcal{U}$.*

*Then*
$$\lim_{n\to\infty} \inf_{u\in\mathcal{U}} F_n(u) = \inf_{u\in\mathcal{U}} F(u).$$
*Moreover, if $u_n$ for every $n \in \mathbb{N}$ is a minimizer of $F_n$, then any cluster point of $\{u_n\}_{n\in\mathbb{N}}$ is a minimizer of $F$.*

For our analysis in this section it will be useful to define $\Psi_0(t) := \mathbf{1}_{t>0}$ for $t \geq 0$, which is a concave and non-decreasing function in its domain. The probabilistic perimeter associated to this function is
$$\begin{aligned}\mathrm{Per}_{\mathrm{prob}\,\Psi_0}(A) &= \int_{A^c} \mathbf{1}_{\mathfrak{m}_x(A)>0}\,\mathrm{d}\rho_0(x) + \int_A \mathbf{1}_{\mathfrak{m}_x(A^c)>0}\,\mathrm{d}\rho_1(x)\\ &= \int_{A^c} \mathfrak{m}_x\text{-}\operatorname{ess\,sup} \mathbf{1}_A\,\mathrm{d}\rho_0(x) + \int_A \mathfrak{m}_x\text{-}\operatorname{ess\,sup} \mathbf{1}_{A^c}\,\mathrm{d}\rho_1(x),\end{aligned}$$
and the corresponding probabilistic risk (1.8) can be written as
$$\begin{aligned}\mathrm{R}_{\text{m-prob}\,\Psi_0}(A) = &\int_{\mathcal{X}} (\mathbf{1}_{A^c}(x)\cdot \mathfrak{m}_x\text{-}\operatorname{ess\,sup} \mathbf{1}_A + \mathbf{1}_A(x))\,\mathrm{d}\rho_0(x)\\ &+ \int_{\mathcal{X}} (\mathbf{1}_A(x)\cdot \mathfrak{m}_x\text{-}\operatorname{ess\,sup} \mathbf{1}_{A^c} + \mathbf{1}_{A^c}(x))\,\mathrm{d}\rho_1(x).\end{aligned}$$
Likewise, the risk $\mathrm{R}_{\mathrm{prob}\,\Psi_0}$ from (1.7) takes the form
$$\mathrm{R}_{\mathrm{prob}\,\Psi_0}(A) = \int_{\mathcal{X}} \mathfrak{m}_x\text{-}\operatorname{ess\,sup} \mathbf{1}_A\,\mathrm{d}\rho_0(x) + \int_{\mathcal{X}} \mathfrak{m}_x\text{-}\operatorname{ess\,sup} \mathbf{1}_A^c\,\mathrm{d}\rho_1(x).$$

We start by proving $\Gamma$-convergence of $\mathrm{R}_{\text{m-prob}\,\Psi_p}$ toward $\mathrm{R}_{\text{m-prob}\,\Psi_0}$ as $p \to 0$ in the weak-* topology of $L^\infty(\mathcal{X};\nu)$.

**Proposition 4.3** ($\Gamma$-convergence of modified PRL)**.** *For the functions $\Psi_p(t) := \min\{t/p, 1\}$ it holds that $\mathrm{Per}_{\text{prob}\,\Psi_p}$ $\Gamma$-converges to $\mathrm{Per}_{\text{prob}\,\Psi_0}$ and $\mathrm{R}_{\text{m-prob}\,\Psi_p}$ $\Gamma$-converges to $\mathrm{R}_{\text{m-prob}\,\Psi_0}$ as $p \to 0$ in the weak-* topology of $L^\infty(\mathcal{X};\nu)$, where $\nu$ is as in (2.4).*

*Proof.* Without loss of generality we assume $\rho_1 = 0$ for a more concise notation.

We start with the liminf inequality. Let $\{A_p\}_{p\in(0,1)}$ be a sequence of sets such that $\mathbf{1}_{A_p}$ converges to $\mathbf{1}_A$ as $p \to 0$ in the weak-* sense. Then Lemma 2.17 implies that

$$\lim_{p\to 0} \mathfrak{m}_x(A_p) = \lim_{p\to 0} \mathbb{E}_{x'\sim\mathfrak{m}_x}\left[\mathbf{1}_{A_p}(x')\right] = \mathbb{E}_{x'\sim\mathfrak{m}_x}\left[\mathbf{1}_A(x')\right] = \mathfrak{m}_x(A).$$

Next, we note that for $0 < p < l$ it holds $\Psi_p \geq \Psi_l$. Hence we get

$$\begin{aligned}\mathrm{Per}_{\text{prob}\,\Psi_p}(A_p) &\geq \int_{A_p^c} \Psi_l(\mathfrak{m}_x(A_p))\,\mathrm{d}\rho_0(x)\\ &= \int_{\mathcal{X}} (1-\mathbf{1}_{A_p})\Psi_l(\mathfrak{m}_x(A_p))\,\mathrm{d}\rho_0(x).\end{aligned}$$

As argued in the proof of Proposition 2.18, the functions $x \mapsto \Psi_l(\mathfrak{m}_x(A_p))$ converge to $x \mapsto \Psi_l(\mathfrak{m}_x(A))$ in $L^1(\mathcal{X};\rho_0)$ as $p \to 0$. Using this together with $\mathbf{1}_{A_p} \overset{*}{\rightharpoonup} \mathbf{1}_A$ in $L^\infty(\mathcal{X};\nu)$ and taking into account (2.6) and (2.7) we get

$$\liminf_{p\to 0} \mathrm{Per}_{\text{prob}\,\Psi_p}(A_p) \geq \int_{A^c} \Psi_l(\mathfrak{m}_x(A))\,\mathrm{d}\rho_0(x).$$

Since $l > 0$ was arbitrary we can let $l \to 0$ and use Fatou's lemma to obtain

$$\liminf_{p\to 0} \mathrm{Per}_{\text{prob}\,\Psi_p}(A_p) \geq \int_{A^c} \Psi_0(\mathfrak{m}_x(A))\,\mathrm{d}\rho_0(x) = \mathrm{Per}_{\text{prob}\,\Psi_0}(A).$$

For the limsup inequality we observe that, since $\Psi_p \leq \Psi_0$ for all $p \in (0,1)$, it holds $\mathrm{Per}_{\text{prob}\,\Psi_p}(A) \leq \mathrm{Per}_{\text{prob}\,\Psi_0}(A)$ and therefore the limsup inequality $\limsup_{p\to 0} \mathrm{Per}_{\text{prob}\,\Psi_p}(A) \leq \mathrm{Per}_{\text{prob}\,\Psi_0}(A)$ is true for the constant sequence equal to $A$.

Being continuous perturbations of the probabilistic perimeters with respect to the weak-* topology of $L^\infty(\mathcal{X};\nu)$, the probabilistic risks $\mathrm{R}_{\text{m-prob}\,\Psi_p}$ are easily seen to $\Gamma$-converge to $\mathrm{R}_{\text{m-prob}\,\Psi_0}$ as $p \to 0$; see the background section in [39]. □

**Remark 4.4.** With essentially the same proof as in Proposition 4.3, one can show that $\mathrm{S}_{\text{m-prob}\,\Psi_p}$ $\Gamma$-converges in the weak-* topology of $L^\infty(\mathcal{X};\nu)$ toward the functional $\mathrm{S}_{\text{m-prob}\,\Psi_0}$ as $p \to 0$. We recall that $\mathrm{S}_{\text{m-prob}\,\Psi}$ was introduced in (2.20) and corresponds to the l.s.c. (but non-convex) extension of $\mathrm{R}_{\text{m-prob}\,\Psi}$ to soft classifiers.

Next, we discuss the $\Gamma$-convergence of $\mathrm{R}_{\text{prob}\,\Psi_p}$ toward $\mathrm{R}_{\text{prob}\,\Psi_0}$. This limiting energy may in general be different from $\mathrm{R}_{\text{m-prob}\,\Psi_0}$.

**Proposition 4.5** ($\Gamma$-convergence of original PRL)**.** *For the functions $\Psi_p(t) := \min\{t/p, 1\}$ it holds that $\mathrm{R}_{\text{prob}\,\Psi_p}$ $\Gamma$-converges to $\mathrm{R}_{\text{prob}\,\Psi_0}$ as $p \to 0$ in the weak-* topology of $L^\infty(\mathcal{X};\sigma)$, where $\sigma$ is as in (2.3).*

*Furthermore, the $\Gamma$-convergence also holds with respect to the weak* topology of $L^\infty(\mathcal{X};\nu)$, where $\nu$ is as in (2.4).*

*Proof.* For simplicity, we again assume $\rho_1 = 0$. The proof starts verbatim as the proof of Proposition 4.3, using Lemma 2.17 with $\sigma$ instead of $\nu$. Then one gets for $0 < p < l$ that

$$\liminf_{p\to 0} \mathrm{R}_{\mathrm{prob}_{\Psi_p}}(A_p) \geq \int_{\mathcal{X}} \Psi_l\left(\mathfrak{m}_x\left(A\right)\right) \mathrm{d}\rho_0(x).$$

Taking also the limit $l \to 0$ one obtains

$$\liminf_{p\to 0} \mathrm{R}_{\mathrm{prob}_{\Psi_p}}(A_p) \geq \int_{\mathcal{X}} \Psi_0\left(\mathfrak{m}_x\left(A\right)\right) \mathrm{d}\rho_0(x).$$

The limsup inequality is again trivial, using that $\Psi_p \leq \Psi_0$ for all $p > 0$.

We point out that, since constant sequences work for the limsup inequality, the $\Gamma$-convergence also holds with respect to any stronger topology as well. In particular, the $\Gamma$-convergence holds under the weak-* topology of $L^\infty(\mathcal{X};\nu)$.

□

**Remark 4.6.** With essentially the same proof as in Proposition 4.5, one can prove that $\mathrm{S}_{\Psi_p}$ $\Gamma$-converges in the weak-* topology of $L^\infty(\mathcal{X};\sigma)$ (or of $L^\infty(\mathcal{X};\nu)$) toward the functional $\mathrm{S}_{\Psi_0}$ as $p \to 0$. We recall that $\mathrm{S}_\Psi$ was introduced in (2.21) and corresponds to the l.s.c. extension of $\mathrm{R}_{\mathrm{prob}_\Psi}$ to soft classifiers.

The next example shows that the functionals $\mathrm{R}_{\text{m-prob}_{\Psi_0}}$ and $\mathrm{R}_{\mathrm{prob}_{\Psi_0}}$, the $\Gamma$-limits of modified PRL and original PRL, respectively, may in general be different from each other.

**Example 4.7.** Consider the data distribution $\mu = \frac{1}{2}\delta_{(x^0,0)} + \frac{1}{2}\delta_{(x^1,1)}$ where $x^0 = (a,0), x^1 = (b,0) \in \mathbb{R}^2$ with $a < b$. Let furthermore $\mathfrak{m}_x := \mathrm{Unif}(B_\varepsilon(x))$ where $0 < \varepsilon < \frac{|b-a|}{2}$. Let $A := \{x = (x_1, x_2) \in \mathbb{R}^2 \,:\, x_1 > a + \frac{b-a}{2}\} \cup \{x^0\}$. Then it follows that

$$\mathrm{R}_{\text{m-prob}_{\Psi_0}}(A) = 1/2 > 0 = \mathrm{R}_{\mathrm{prob}_{\Psi_0}}(A).$$

While the energies $\mathrm{R}_{\text{m-prob}_{\Psi_0}}$, $\mathrm{R}_{\mathrm{prob}_{\Psi_0}}$, and $\mathrm{R}_{\mathrm{adv}}$ may in general be different, they are always ordered, with $\mathrm{R}_{\mathrm{adv}}$ being the largest. This is the content of the next proposition, which we state in more general terms.

**Proposition 4.8.** *Let $\Psi : [0,1] \to [0,1]$ be such that $\Psi(0) = 0$, and suppose that for every $x \in \mathcal{X}$ one has $\mathfrak{m}_x(\mathcal{X} \setminus B_\varepsilon(x)) = 0$. Then for every measurable $A$ the following holds:*

$$\mathrm{R}_{\mathrm{prob}_\Psi}(A) \leq \mathrm{R}_{\text{m-prob}_\Psi}(A) \leq \mathrm{R}_{\mathrm{adv}}(A). \tag{4.1}$$

*In particular, the above applies to $\Psi = \Psi_0$, recalling $\Psi_0(t) = \mathbf{1}_{t>0}$.*

*Proof.* It is straightforward to show that the following pointwise identity holds for any measurable set $A$:

$$\mathbf{1}_A(x) + \mathbf{1}_{A^c}(x) \cdot \sup_{\tilde{x}\in B_\varepsilon(x)} \mathbf{1}_A(\tilde{x}) \geq \mathbf{1}_A(x) + \mathbf{1}_{A^c}(x) \cdot \Psi(\mathfrak{m}_x(A)) \geq \Psi(\mathfrak{m}_x(A)).$$

Applying this identity to $A^c$ we get:

$$\mathbf{1}_{A^c}(x) + \mathbf{1}_A(x) \cdot \sup_{\tilde{x}\in B_\varepsilon(x)} \mathbf{1}_{A^c}(\tilde{x}) \geq \mathbf{1}_{A^c}(x) + \mathbf{1}_A(x) \cdot \Psi(\mathfrak{m}_x(A^c)) \geq \Psi(\mathfrak{m}_x(A^c)).$$

Integrating the first chain of inequalities with respect to $\rho_0$, the second with respect to $\rho_1$, and adding up the resulting expressions, we obtain (4.1).

□

Next, we specialize to the case where the admissible set of hard classifiers is $\mathcal{A} = \mathcal{B}(\mathcal{X})$, *i.e.*, the set of all Borel measurable subsets of $\mathcal{X}$ (in the machine learning literature this is an agnostic learning setting) and then show that, while $\mathrm{R}_{\text{m-prob}_{\Psi_0}}$ and $\mathrm{R}_{\text{prob}_{\Psi_0}}$ may be different, their minimal values coincide with that of the adversarial training value functional $\mathrm{R}_{\text{adv}}$ if we impose some reasonable assumptions on the family of measures $\{\mathfrak{m}_x\}_{x\in\mathcal{X}}$. Our result implies that, under Assumption 4.9 below, both PRL and modified PRL models are consistent with AT, provided consistency is interpreted as consistency in minimal risks.

**Assumption 4.9.** For every $x \in \operatorname{supp}(\rho)$ the following conditions hold:

1. $\mathfrak{m}_x(\mathcal{X} \setminus B_\varepsilon(x)) = 0$.
2. $B_\varepsilon(x) \subseteq \operatorname{supp}(\mathfrak{m}_x)$.
3. For every $x' \in \operatorname{supp}(\rho)$ the measure $\mathfrak{m}_{x'}\lfloor_{B_\varepsilon(x)}$ is absolutely continuous with respect to $\mathfrak{m}_x$.

**Remark 4.10.** Assumption 4.9 is very mild and in the Euclidean setting, when $\mathcal{X} = \mathbb{R}^d$, is satisfied by the simple model $\mathfrak{m}_x = \mathrm{Unif}(B_\varepsilon(x))$ .

**Theorem 4.11** (Consistency of optimal energies for PRL and modified PRL)**.** *Let $\mathcal{A} = \mathcal{B}(\mathcal{X})$ be the Borel $\sigma$-algebra over $\mathcal{X}$. Under Assumption 4.9,*

$$\lim_{p\to 0}\min_{A\in\mathcal{A}} \mathrm{R}_{\text{prob}_{\Psi_p}}(A) = \min_{A\in\mathcal{A}} \mathrm{R}_{\text{prob}_{\Psi_0}}(A) = \min_{A\in\mathcal{A}} \mathrm{R}_{\text{adv}}(A) = \min_{A\in\mathcal{A}} \mathrm{R}_{\text{m-prob}_{\Psi_0}}(A) = \lim_{p\to 0}\min_{A\in\mathcal{A}} \mathrm{R}_{\text{m-prob}_{\Psi_p}}(A),$$

*where we recall $\Psi_p(t) = \min\{t/p, 1\}$ and $\Psi_0(t) = \mathbf{1}_{t>0}$.*

*Proof.* Thanks to Remarks 4.4 and 4.6 and the Banach–Alaoglu theorem, we can apply the fundamental theorem of $\Gamma$-convergence (*i.e.*, Proposition 4.2) to conclude that

$$\lim_{p\to 0}\min_{u\in\mathcal{U}} \mathrm{S}_{\text{prob}_{\Psi_p}}(u) = \min_{u\in\mathcal{U}} \mathrm{S}_{\text{prob}_{\Psi_0}}(u)$$

and

$$\lim_{p\to 0}\min_{u\in\mathcal{U}} \mathrm{S}_{\text{m-prob}_{\Psi_p}}(u) = \min_{u\in\mathcal{U}} \mathrm{S}_{\text{m-prob}_{\Psi_0}}(u).$$

In addition, since $\Psi_p$ is concave and non-decreasing for every $p \geq 0$, we have, thanks to Remarks 2.21 and 2.22,

$$\min_{A\in\mathcal{A}} \mathrm{R}_{\text{prob}_{\Psi_p}}(A) = \min_{u\in\mathcal{U}} \mathrm{S}_{\text{prob}_{\Psi_p}}(u)$$

as well as

$$\min_{A\in\mathcal{A}} \mathrm{R}_{\text{m-prob}_{\Psi_p}}(A) = \min_{u\in\mathcal{U}} \mathrm{S}_{\text{m-prob}_{\Psi_p}}(u).$$

Therefore

$$\lim_{p\to 0}\min_{A\in\mathcal{A}} \mathrm{R}_{\text{prob}_{\Psi_p}}(A) = \min_{A\in\mathcal{A}} \mathrm{R}_{\text{prob}_{\Psi_0}}(A)$$

and

$$\lim_{p\to 0}\min_{A\in\mathcal{A}} \mathrm{R}_{\text{m-prob}_{\Psi_p}}(A) = \min_{A\in\mathcal{A}} \mathrm{R}_{\text{m-prob}_{\Psi_0}}(A).$$

On the other hand, thanks to Proposition 4.8 applied to $\Psi = \Psi_0$, we have

$$\min_{A\in\mathcal{A}} \mathrm{R}_{\mathrm{adv}}(A) \geq \min_{A\in\mathcal{A}} \mathrm{R}_{\text{m-prob}_{\Psi_0}}(A) \geq \min_{A\in\mathcal{A}} \mathrm{R}_{\mathrm{prob}_{\Psi_0}}(A).$$

Thus it suffices to prove that

$$\min_{A\in\mathcal{A}} \mathrm{R}_{\mathrm{prob}_{\Psi_0}}(A) \geq \min_{A\in\mathcal{A}} \mathrm{R}_{\mathrm{adv}}(A).$$

To see this, let $A$ be an arbitrary measurable subset of $\mathcal{X}$. We can then proceed as in the proof of [13], Lemma 3.8 and use the fact that

$$\{x \in \mathcal{X} \text{ s.t. } \mathrm{dist}(x, \mathrm{supp}(\rho)) < \varepsilon\} \subseteq \mathrm{supp}(\sigma),$$

which follows from (2) in Assumption 4.9, to conclude that we can find a measurable set $A'$ which is $\sigma$-equivalent to $A$ (*i.e.*, $\sigma(A \triangle A') = 0$) and satisfies:

$$\sigma\text{- ess sup}_{\tilde{x}\in B_\varepsilon(x)} \mathbf{1}_A(\tilde{x}) = \sup_{\tilde{x}\in B_\varepsilon(x)} \mathbf{1}_{A'}(\tilde{x}), \quad \sigma\text{- ess sup}_{\tilde{x}\in B_\varepsilon(x)} \mathbf{1}_{A^c}(\tilde{x}) = \sup_{\tilde{x}\in B_\varepsilon(x)} \mathbf{1}_{A'^c}(\tilde{x})$$

for $\rho$-a.e. $x \in \mathcal{X}$. Thanks to (2.5), we deduce that

$$\mathfrak{m}_x(A \triangle A') = 0, \quad \rho\text{-a.e. } x \in \mathcal{X}.$$

It follows that for $\rho$ almost every $x \in \mathcal{X}$ we have

$$\sigma\text{- ess sup}_{\tilde{x}\in B_\varepsilon(x)} \mathbf{1}_A(\tilde{x}) = \sup_{\tilde{x}\in B_\varepsilon(x)} \mathbf{1}_{A'}(\tilde{x}) \geq \mathfrak{m}_x\text{- ess sup } \mathbf{1}_{A'}(\tilde{x}) = \mathfrak{m}_x\text{- ess sup } \mathbf{1}_A(\tilde{x}).$$

as well as

$$\sigma\text{- ess sup}_{\tilde{x}\in B_\varepsilon(x)} \mathbf{1}_{A^c}(\tilde{x}) \geq \mathfrak{m}_x\text{- ess sup } \mathbf{1}_{A^c}(\tilde{x}).$$

Since $\sigma$ is absolutely continuous with respect to $\mathfrak{m}_x$ for $\rho$ a.e. $x \in \mathcal{X}$ (thanks to Asm. 4.9), we deduce that

$$\mathfrak{m}_x\text{- ess sup } \mathbf{1}_{A^c}(\tilde{x}) \geq \sigma\text{- ess sup}_{\tilde{x}\in B_\varepsilon(x)} \mathbf{1}_{A^c}(\tilde{x}), \quad \mathfrak{m}_x\text{- ess sup } \mathbf{1}_A(\tilde{x}) \geq \sigma\text{- ess sup}_{\tilde{x}\in B_\varepsilon(x)} \mathbf{1}_A(\tilde{x})$$

for $\rho$-a.e. $x \in \mathcal{X}$. Combining the above, we deduce that for $\rho$-a.e. $x \in \mathcal{X}$

$$\mathfrak{m}_x\text{- ess sup } \mathbf{1}_{A^c}(\tilde{x}) = \sup_{\tilde{x}\in B_\varepsilon(x)} \mathbf{1}_{A'^c}(\tilde{x}), \quad \mathfrak{m}_x\text{- ess sup } \mathbf{1}_A(\tilde{x}) = \sup_{\tilde{x}\in B_\varepsilon(x)} \mathbf{1}_{A'}(\tilde{x}).$$

From this fact, we deduce that

$$\begin{aligned}
\mathrm{R}_{\mathrm{prob}_{\Psi_0}}(A) &= \int_{\mathcal{X}} \mathfrak{m}_x\text{- ess sup}_{\tilde{x}\in B_\varepsilon(x)} \mathbf{1}_A \,\mathrm{d}\rho_0(x) + \int_{\mathcal{X}} \mathfrak{m}_x\text{- ess sup}_{\tilde{x}\in B_\varepsilon(x)} \mathbf{1}_{A^c} \,\mathrm{d}\rho_1(x) \\
&= \int_{\mathcal{X}} \sup_{\tilde{x}\in B_\varepsilon(x)} \mathbf{1}_{A'}(\tilde{x}) \,\mathrm{d}\rho_0(x) + \int_{\mathcal{X}} \sup_{\tilde{x}\in B_\varepsilon(x)} \mathbf{1}_{A'^c}(\tilde{x}) \,\mathrm{d}\rho_1(x) \\
&= \mathrm{R}_{\mathrm{adv}}(A') \\
&\geq \min_{\tilde{A}\in\mathcal{A}} \mathrm{R}_{\mathrm{adv}}(\tilde{A}).
\end{aligned}$$

Taking the infimum over $A$ on the left hand side we obtain the desired result.

□

While the above result states that both PRL and modified PRL models recover the AT model from the perspective of consistency of their *optimal energies* (at least in the agnostic setting, where we can prove existence of minimizers to AT; see [13]), the next example illustrates that *minimizers* of the PRL energies may not converge to AT minimizers as the parameter $p$ goes to zero. Studying the behavior of minimizers is important, since, after all, PRL models and AT are used to produce robust classifiers by minimizing energy functionals. We start with an example.

**Example 4.12.** Let $\mathcal{X} = \mathbb{R}$ and for every $x \in \mathbb{R}$ let $\mathfrak{m}_x$ be the uniform measure over $B_\varepsilon(x)$, the open Euclidean ball of radius $\varepsilon$ and center $x$. Consider the probability distribution $\mu$ over $\mathcal{X} \times \{0,1\}$ given by

$$\mu = \frac{2}{5}\delta_{(x^1,0)} + \frac{1}{5}\delta_{(x^2,0)} + \frac{2}{5}\delta_{(x^3,1)},$$

where $x^1 < x^2 < x^3$, $|x^1 - x^2| < \varepsilon$, and $B_\varepsilon(x^2) \cap B_\varepsilon(x^3) \neq \emptyset$, $B_\varepsilon(x^1) \cap B_\varepsilon(x^3) = \emptyset$.

It is straightforward to show that the optimal adversarial risk $\mathrm{R}^*_{\mathrm{adv}}$ in this situation is $1/5$ and that $A'$ achieves this adversarial risk if and only if $B_\varepsilon(x^3) \subseteq A'$ and $B_\varepsilon(x^1) \subseteq A'^c$. On the other hand, the set $\tilde{A} = B_\varepsilon(x^3) \cup \{x^2\}$ satisfies $\mathrm{R}_{\text{m-prob}\,\Psi_0}(\tilde{A}) = 1/5$. In particular, thanks to Proposition 4.11, $\tilde{A}$ is a minimizer of both $\mathrm{R}_{\text{m-prob}\,\Psi_0}$ and $\mathrm{R}_{\text{prob}\,\Psi_0}$ (*i.e.*, the $\Gamma$-limits of the PRL models as $p \to 0$). Note, however, that there does not exist a set $A' \sim_\nu \tilde{A}$ that is a minimizer of $\mathrm{R}_{\mathrm{adv}}$. This is because for $A'$ to be a minimizer of $\mathrm{R}_{\mathrm{adv}}$ we would need to exclude $x^2$ from $\tilde{A}$, but it is impossible to achieve this while maintaining the condition $\tilde{A} \sim_\nu A'$, since $\rho$ assigns mass strictly larger than zero to $\{x^2\}$.

From the above example it is clear that in general we cannot expect that an arbitrary minimizer of $\mathrm{R}_{\text{prob}\,\Psi_0}$ or $\mathrm{R}_{\text{m-prob}\,\Psi_0}$ can be turned into a solution to adversarial training by modifying it within a set of $\nu$ measure zero, in particular by only modifying the classifier outside of the observed data (the support of the distribution $\rho$) and their admissible perturbations. In other words, this means that there may be solutions to PRL or modified PRL for $p \approx 0$ that, putting measure theoretic technicalities aside, aren't solutions to AT. In this sense, PRL or its modified version may not necessarily recover the AT model when $p \to 0$. Fortunately, it is possible to avoid the issue described above by imposing one extra assumption on the family of distributions $\{\mathfrak{m}_x\}_{x\in\mathcal{X}}$.

**Assumption 4.13.** Suppose that for every $x \in \mathrm{supp}(\rho)$ the measure $\rho\lfloor_{B_\varepsilon(x)}$, the restriction of $\rho$ to the ball $B_\varepsilon(x)$, is absolutely continuous with respect to $\mathfrak{m}_x$.

**Remark 4.14.** If for $x \in \mathrm{supp}(\rho)$ we choose $\mathfrak{m}_x$ to be $\mathfrak{m}_x = \frac{1}{2}\,\mathrm{Unif}(B_\varepsilon(x)) + \frac{1}{2\rho(B_\varepsilon(x))}\rho\lfloor_{B_\varepsilon(x)}$, then the family $\{\mathfrak{m}_x\}_{x\in\mathcal{X}}$ satisfies Assumptions 4.9 and 4.13.

**Theorem 4.15.** *Suppose that Assumptions 4.9 and 4.13 hold. Suppose that for every $p > 0$ the set $A_p$ is a minimizer of $\mathrm{R}_{\text{m-prob}\,\Psi_p}$ over $\mathcal{A} = \mathcal{B}(\mathcal{X})$, and suppose that, as $p \to 0$, $A_p$ converges towards a set $A$ in the weak* topology of $L^\infty(\mathcal{X};\nu)$. Then there is a set $A'$ with $A' \sim_\nu A$ such that $A'$ is a minimizer of $\mathrm{R}_{\mathrm{adv}}$ over the class of all Borel measurable sets.*

*The above continues to be true if we substitute $\mathrm{R}_{\text{m-prob}\,\Psi_p}$ with $\mathrm{R}_{\text{prob}\,\Psi_p}$.*

*Proof.* The $\Gamma$-convergence in Proposition 4.3 implies that $A$ as in the statement must be a minimizer of $\mathrm{R}_{\text{m-prob}\,\Psi_0}$. Following the proof of Proposition 4.11 we know there exists a set $A'$ which is $\sigma$-equivalent to $A$ and satisfies $\mathrm{R}_{\text{m-prob}\,\Psi_0}(A) \geq \mathrm{R}_{\text{prob}\,\Psi_0}(A) = \mathrm{R}_{\mathrm{adv}}(A')$. Thanks to Proposition 4.11, it follows that $A'$ is a minimizer of $\mathrm{R}_{\mathrm{adv}}$. The desired result now follows from Assumption 4.13, since this assumption implies that $\nu$ is absolutely continuous with respect to $\sigma$ and as a consequence $A \sim_\nu A'$.

□

So far we have studied the behaviors of the energies $\mathrm{R}_{\text{m-prob}\,\Psi_p}$ and $\mathrm{R}_{\text{prob}\,\Psi_p}$ as $p \to 0$. To wrap up this section we discuss the limits of these energies as $p$ grows.

**Theorem 4.16.** *For the functions* $\Psi_p(t) := \min\{t/p, 1\}$ *it holds that* $\mathrm{R}_{\text{m-prob}_{\Psi_p}}$ $\Gamma$*-converges to*

$$\mathbb{E}_{(x,y)\sim\mu}\left[|\mathbf{1}_A(x) - y|\right] + \int_A \int_{A^c} \mathrm{d}\mathfrak{m}_x(\tilde{x})\,\mathrm{d}\rho_1(x) + \int_{A^c}\int_A \mathrm{d}\mathfrak{m}_x(\tilde{x})\,\mathrm{d}\rho_0(x)$$

*as* $p \to 1$*, and to* $\mathbb{E}_{(x,y)\sim\mu}\left[|\mathbf{1}_A(x) - y|\right]$ *as* $p \to \infty$*, in the weak-* topology of* $L^\infty(\mathcal{X};\nu)$*. We recall* $\nu$ *was introduced in* (2.4)*.*

*Proof.* This easily follows from the fact that, actually, the convergence of the energies is uniform over all measurable $A$ as $p \to 1$ and as $p \to \infty$. □

**Remark 4.17.** In the previous result we have highlighted the case $p = 1$, as this is the case where the modified PRL model coincides with a regularized risk minimization problem with nonlocal perimeter penalty of the type investigated in [36] in the context of random walk metric spaces.

**Remark 4.18.** In contrast to the modified PRL model, which recovers the standard risk minimization model as $p$ grows, the energy $\mathrm{R}_{\text{prob}_{\Psi_p}}$ is easily seen to converge uniformly toward the zero functional as $p \to \infty$. This means that the original PRL model does not interpolate between adversarial training and risk minimization.

## 5. PRL for general learning settings and the conditional value at risk

After extensively discussing the binary and 0-1 loss case in the previous sections, we shift our attention to training general hypotheses $h \in \mathcal{H}$ using general loss functions $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, even when $\mathcal{Y}$ is not binary. Motivated by Propositions 2.1 and 2.5 it is natural to consider the following probabilistically robust optimization problem

$$\inf_{h\in\mathcal{H}} \mathbb{E}_{(x,y)\sim\mu}\left[\max\left\{\ell(h(x),y), p\text{-}\operatorname*{ess\,sup}_{x'\sim\mathfrak{m}_x} \ell(h(x'),y)\right\}\right]. \tag{5.1}$$

Since the $p$-ess sup operator is notoriously hard to optimize, one shall replace it with the conditional value at risk (CVaR) which is convex and easier to optimize [10, 31]. For a function $f : \mathcal{X} \to \mathbb{R}$ and a probability distribution $\mathfrak{m}$ the CVaR at level $p \in (0,1)$ is defined as

$$\mathrm{CVaR}_p(f;\mathfrak{m}) := \inf_{\alpha\in\mathbb{R}} \alpha + \frac{\mathbb{E}_{x'\sim\mathfrak{m}_x}\left[(f(x')-\alpha)_+\right]}{p}. \tag{5.2}$$

It is easy to see that $p\text{-}\operatorname{ess\,sup}_{x'\sim\mathfrak{m}} f(x') \le \mathrm{CVaR}_p(f;\mathfrak{m})$. Using CVaR in place of the $p$-ess sup operator, a tractable version of (5.1) is

$$\inf_{h\in\mathcal{H}} \mathbb{E}_{(x,y)\sim\mu}\left[\max\left\{\ell(h(x),y), \mathrm{CVaR}_p(\ell(h(\bullet),y);\mathfrak{m}_x)\right\}\right]. \tag{5.3}$$

We emphasize that, if the loss function $\ell(\bullet,\bullet)$ is convex in its first argument, then (5.3) is a convex function of the hypothesis $h$. Furthermore, CVaR is positively 1-homogeneous and hence also (5.3) is positively 1-homogeneous in the loss function. So, taking the maximum of the samplewise CVaR and standard risk is consistent with a standard dimensionality analysis as both terms scale in the same way.

In the binary classification case we can prove Proposition 2.6 which states that the CVaR relaxation corresponds precisely to using the risk $\mathrm{R}_{\text{m-prob}_{\Psi_p}}$ with the piecewise linear and concave function $\Psi_p(t) = \min\{t/p, 1\}$ for which our theory from Section 2.2 applies.

*Proof of Proposition 2.6.* If $\ell$ is the 0-1-loss then the function $f := \ell(\mathbf{1}_A(\cdot), y)$ for $A \in \mathcal{A}$ can be written as $f = \mathbf{1}_S$ for set $S \in \{A, A^c\}$, depending on whether $y = 0$ or $y = 1$. So it suffices to deal with the case $f = \mathbf{1}_A$ where we can write

$$\mathrm{CVaR}_p(f; \mathfrak{m}_x) = \mathrm{CVaR}_p(\mathbf{1}_A; \mathfrak{m}_x) = \inf_{\alpha \in \mathbb{R}} \alpha + (1-\alpha)_+ \frac{\mathfrak{m}_x(A)}{p} + (-\alpha)_+ \frac{1 - \mathfrak{m}_x(A)}{p}.$$

Notice that the function

$$\zeta(\alpha) := \alpha + (1-\alpha)_+ \frac{\mathfrak{m}_x(A)}{p} + (-\alpha)_+ \frac{1 - \mathfrak{m}_x(A)}{p}, \quad \alpha \in \mathbb{R},$$

is continuous and piecewise linear with kinks at $\alpha = 0$ and $\alpha = 1$. Moreover, since $0 < p < 1$ it holds $\zeta(\alpha) \geq \zeta(1)$ for $\alpha > 1$, and $\zeta(\alpha) \geq \zeta(0)$ for $\alpha < 0$. Thus the minimum of $\zeta$ is attained at either $\alpha = 0$ or $\alpha = 1$. Therefore

$$\mathrm{CVaR}_p(\mathbf{1}_A; \mathfrak{m}_x) = \min\{\zeta(0), \zeta(1)\} = \min\left\{\frac{\mathfrak{m}_x(A)}{p}, 1\right\} = \Psi_p\left(\mathfrak{m}_x(A)\right).$$

This together with Proposition 2.5 implies the claim. □

## 6. Numerical experiments

In this section, we conduct a comparative analysis between the original formulation of PRL (denoted as "PRL" in Tab. 1) and our modification of PRL which is based on (5.3) (denoted as "m-PRL"). We build upon the code of [10] and our implementation is available on `GitHub`[1]. The algorithmic realization of (5.3) is a straightforward adaptation of their algorithm, which alternatingly minimizes the inner optimization problem that defines CVaR and the outer optimization to find a suitable classifier, see Algorithm 1 in Appendix C. Our experiments are conducted on MNIST and CIFAR-10 and to ensure a fair comparison we adhere to the hyperparameter settings described in [10], such that both the original and modified algorithms utilize the same set of hyperparameters for each specified value of $p$. The corresponding results for several baseline algorithms including empirical risk minimization and adversarial training can be found in [10].

### 6.1. Comparing accuracy and robustness

We report the following metrics, all computed on held-out test sets. **Clean accuracy** measures the fraction of test samples correctly classified without any perturbation. **Adversarial accuracy** measures the fraction of test samples correctly classified under PGD attacks [8] with $\ell_\infty$ perturbation budget $\varepsilon$ (0.3 for MNIST, 8/255 for CIFAR-10) and 10 attack iterations. **Augmented accuracy** is the average accuracy when each test sample is evaluated over 100 random perturbations sampled uniformly from the $\varepsilon$-ball. Finally, we report the **quantile accuracy** $\mathrm{ProbAcc}_p$ for different values of $p$, defined – as in [10], (6.1) – for a single data point $(x, y)$ by

$$\mathrm{ProbAcc}_p(x, y) = \mathbf{1}_{\mathbb{P}_{x' \sim \mathfrak{m}_x}[h(x') = y] > p}. \tag{6.1}$$

In practice, we use an empirical proportion, computed over 100 samples from $\mathfrak{m}_x$, in lieu of the true probability, as done in [10]. All quantities are averaged over three runs, and we perform model selection based on the best clean validation accuracy; see Appendix C.2 for more training details.

The results in Tables 1 demonstrate that the geometric modification does not compromise the accuracy of the original algorithm, and sometimes leads to improved performance, even as it provides stronger theoretical guarantees. Note that neither PRL nor m-PRL should be expected to match the adversarial robustness of classifiers trained with PGD attacks [8] or other worst-case optimization techniques. Instead, they shine with

[1] https://github.com/DanielMckenzie/Begins_with_a_boundary

TABLE 1. Accuracies [%] of the geometric and original algorithm for different values of $p$.

| Data | $p$ | Algorithm | Clean | Adv | Aug | ProbAcc(0.1) | ProbAcc(0.05) | ProbAcc(0.01) |
|---|---|---|---|---|---|---|---|---|
| MNIST | 0.01 | m-PRL | **99.20** | **12.19** | 99.04 | 98.18 | 97.69 | 96.38 |
| | | PRL | 99.19 | 10.76 | 98.90 | 97.94 | 97.38 | 95.67 |
| | 0.1 | m-PRL | 99.28 | **14.20** | 99.22 | 98.70 | 98.45 | 97.86 |
| | | PRL | **99.32** | 8.94 | 99.22 | 98.70 | 98.46 | 97.80 |
| | 0.3 | m-PRL | **99.29** | **3.02** | 99.21 | 98.76 | 98.53 | 97.95 |
| | | PRL | 99.27 | **3.02** | 99.22 | 98.77 | 98.55 | 98.01 |
| | 0.5 | m-PRL | **99.27** | **1.80** | 99.21 | 98.72 | 98.44 | 97.93 |
| | | PRL | 99.26 | 1.68 | 99.19 | 98.72 | 98.47 | 97.80 |
| CIFAR-10 | 0.01 | m-PRL | 80.65 | 0.15 | 78.13 | 73.44 | 72.13 | 68.80 |
| | | PRL | **81.73** | **0.24** | 79.16 | 74.61 | 73.19 | 69.96 |
| | 0.1 | m-PRL | 88.15 | 0.14 | 85.96 | 82.55 | 81.46 | 78.81 |
| | | PRL | **88.28** | **0.19** | 85.61 | 82.21 | 81.06 | 78.28 |
| | 0.3 | m-PRL | **90.43** | **11.80** | 88.70 | 85.17 | 83.93 | 80.93 |
| | | PRL | 89.97 | 7.20 | 88.62 | 85.07 | 83.75 | 80.87 |
| | 0.5 | m-PRL | **91.51** | 1.93 | 88.94 | 85.53 | 84.18 | 81.21 |
| | | PRL | 90.74 | **1.99** | 88.94 | 85.54 | 84.35 | 81.57 |

superior clean accuracy and easier training while maintaining probabilistic robustness, as well as a certain degree of adversarial robustness. This corroborates the findings of [10].

### 6.2. On the existence of pathological points

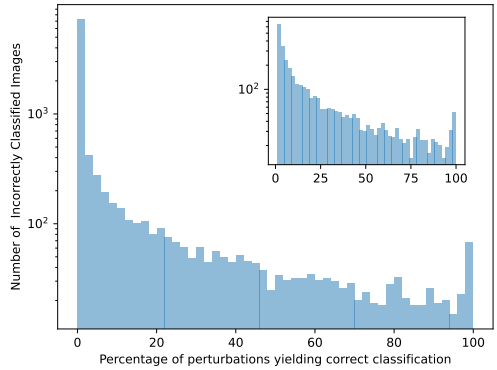
We say $(x, y)$ is a pathological data point if $x$ is incorrectly classified (according to the Bayes classifier) yet a large proportion of perturbations to $x$ yield a correct classification. See Figure 1 for a simple example. Although the discussion of Section 1 asserts that such points *can* arise for classifiers trained using PRL, it is interesting to verify whether this phenomenon occurs in practice.

To do so, we examine all misclassified CIFAR-10 images, in both test and train splits, for a model trained using PRL as described above. For each image $x$, we generate 100 samples $x' \sim \mathfrak{m}_x$ and compute the proportion of such samples which are correctly classified. As shown in Figure 3, while for the vast majority of such $x$, all 100 $x'$ remain incorrectly classified, images $x$ exist where an arbitrarily large proportion of the $x'$ are correctly classified. In short, pathological points do occur in practice.

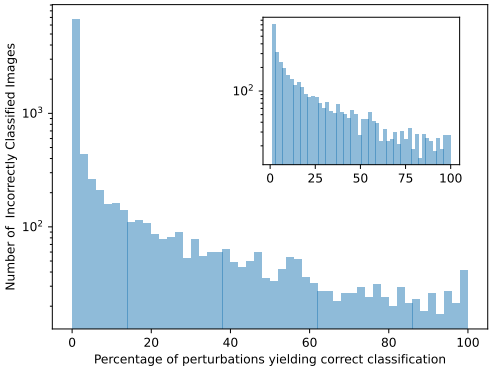
We repeated the same experiment, but with a model trained using m-PRL. Intriguingly, while m-PRL is guaranteed in theory to prevent pathological points (see Prop. B.1 in the Appendix), in practice they still occur, see Figure 3. We did not find strong empirical evidence that using m-PRL results in fewer pathological points in practice, see also Figures C.3 to C.2. We attribute this gap between theory and practice to the fact that m-PRL (as well as PRL) approximates the value of $\mathrm{CVaR}_p$, which involves a high-dimensional expectation, with an empirical average, see (5.3) and line 7 in Algorithm 1).

## 7. Discussion and Conclusion

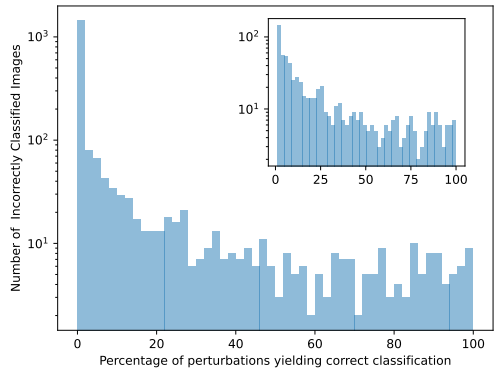
In this paper we considered probabilistically robust learning (PRL), originally proposed in [10]. We introduced a modification of the original PRL model that allowed us to address, at least theoretically, the possibility that certain solutions to the model possess pathological points as described in Figure 1. This modification has the appeal of being interpretable through the lens of regularized risk minimization, where the regularization terms take the form of non-local perimeters of interest in their own right. We discussed an asymptotic expansion for smooth decision boundaries to show that for small adversarial budgets these probabilistic perimeters induce the



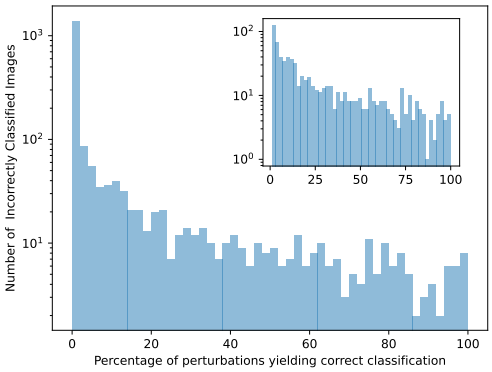

(A) $p = 0.01$, train, original PRL




(B) $p = 0.01$, train, m-PRL




(C) $p = 0.01$, test, original PRL




(D) $p = 0.01$, test, m-PRL

FIGURE 3. Histograms display the distribution of percentages of correctly classified perturbations among misclassified images for both original and m-PRL with parameter $p = 0.01$. The inner plot excludes the prevalent 0% case. The plots show that pathological data points – *i.e.* data points which are misclassified yet most perturbations to the data point are correctly classified – occur in real datasets.

same regularization effect as the original adversarial training model. For binary classification we proved existence of optimal hard classifiers and of very general classes of soft classifiers including neural networks in both the original and modified PRL settings. This was done through novel relaxation techniques taking advantage of the structure of our functionals. Finally, through rigorous $\Gamma$-convergence analysis we provided a detailed discussion on the relation between adversarial training, risk minimization, and all the PRL models, highlighting that the original PRL model fails to interpolate between risk minimization and adversarial training, contrary to claims in previous works. For general (not necessarily binary) problems we showed that the natural loss function to choose is the sample-wise maximum of the standard loss and conditional value at risk (CVaR).

One limitation of PRL is that it does not completely solve the accuracy vs. robustness trade-off, which remains a challenging problem. Furthermore, while the formal limit of PRL as $p \to 0$ is the worst-case adversarial problem, the algorithms for solving PRL exhibit limitations for very small values of $p$ (in the computation of $\mathrm{CVaR}_p$). Still, the results for moderately large values of $p$ are encouraging and future work should focus on understanding of this trade-off better.

The rich mathematical theory developed in this paper opens up new avenues for research, such as the explicit design of probabilistic regularizers for enforcing robustness of algorithms to certain perturbations of data and exploring the variational convergence of probabilistic perimeters.

## Funding

This material is based upon work supported by the National Science Foundation under Grant Number DMS 1641020 and was started during the summer of 2022 as part of the AMS-MRC program *Data Science at the Crossroads of Analysis, Geometry, and Topology*. NGT was supported by the NSF grants DMS-2005797 and DMS-2236447. MJ was supported by NSF grant DMS-2400641. LB acknowledges funding by the Federal Ministry of Research, Technology and Space of Germany (BMFTR) under grant agreement No. 01IS24072A (COMFORT) and by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 544579844 (GeoMAR).

## Data availability statement

The research code associated with this article is available in https://github.com/DanielMckenzie/Begins_with_a_boundary.

## References

[1] H.Q. Cai, Y. Lou, D. McKenzie and W. Yin, A zeroth-order block coordinate descent algorithm for huge-scale black-box optimization, in *International Conference on Machine Learning*. PMLR (2021) 1193–1203.

[2] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi and C.-J. Hsieh, ZOO: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models, in *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security* (2017) 15–26.

[3] I.J. Goodfellow, J. Shlens and C. Szegedy, Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014).

[4] Y. Qin, N. Carlini, G. Cottrell, I. Goodfellow and C. Raffel, Imperceptible, robust, and targeted adversarial examples for automatic speech recognition, in *International Conference on Machine Learning*. PMLR (2019) 5231–5240.

[5] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo, *et al.*, The many faces of robustness: A critical analysis of out-of-distribution generalization, in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021) 8340–8349.

[6] J. Quinoñero Candela, M. Sugiyama, A. Schwaighofer and N.D. Lawrence, *Dataset Shift in Machine Learning*. MIT Press (2008).

[7] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer Science & Business Media (1999).

[8] A. Madry, A. Makelov, L. Schmidt, D. Tsipras and A. Vladu, Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017).

[9] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner and A. Madry, Robustness may be at odds with accuracy, in *International Conference on Learning Representations* (2019).

[10] A. Robey, L. Chamon, G.J. Pappas and H. Hassani, Probabilistically robust learning: Balancing average and worst-case performance, in *International Conference on Machine Learning*. PMLR (2022) 18667–18686.

[11] B. Wang, B. Yuan, Z. Shi and S.J. Osher, EnResNet: ResNets ensemble *via* the Feynman–Kac formalism for adversarial defense and beyond. *SIAM J. Math. Data Sci.* **2** (2020) 559–582.

[12] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui and M. Jordan, Theoretically principled trade-off between robustness and accuracy, in *International Conference on Machine Learning*. PMLR (2019) 7472–7482.

[13] L. Bungert, N. García Trillos and R. Murray, The geometry of adversarial training in binary classification. *Inform. Inference* **12** (2023) 921–968.

[14] A. Shafahi, M. Najibi, M.Am. Ghiasi, Z. Xu, J. Dickerson, C. Studer, L.S. Davis, G. Taylor and T. Goldstein, Adversarial training for free! *Adv. Neural Inform. Process. Syst.* **32** (2019) 1–12.

[15] E. Wong, L. Rice and J. Zico Kolter, Fast is better than free: Revisiting adversarial training, in *International Conference on Learning Representations* (2020).

[16] M. Andriushchenko and N. Flammarion, Understanding and improving fast adversarial training. *Adv. Neural Inform. Process. Syst.* **33** (2020) 16048–16059.

[17] J. Cohen, E. Rosenfeld and Z. Kolter, Certified adversarial robustness via randomized smoothing, in *International Conference on Machine Learning*. PMLR (2019) 1310–1320.
[18] L. Schwinn, L. Bungert, A. Nguyen, R. Raab, F. Pulsmeyer, D. Precup, B. Eskofier and D. Zanca, Improving robustness against real-world and worst-case distribution shifts through decision region quantification, in *International Conference on Machine Learning*. PMLR (2022) 19434–19449.
[19] L. Schwinn, A. Nguyen, R. Raab, L. Bungert, D. Tenbrinck, D. Zanca, M. Burger and B. Eskofier, Identifying untrustworthy predictions in neural networks by geometric gradient analysis, in *Uncertainty in Artificial Intelligence*. PMLR (2021) 854–864.
[20] V. Raman, U. Subedi and A. Tewari, On proper learnability between average- and worst-case robustness. arXiv preprint arXiv:2211.05656 (2023).
[21] N. García Trillos and R. Murray, Adversarial classification: necessary conditions and geometric flows. *J. Mach. Learn. Res.* **23** (2022) 1–38.
[22] L. Bungert, T. Laux and K. Stinson, A mean curvature flow arising in adversarial training. *J. Math. Pures Appl.* **192** (2024) 103625.
[23] L. Bungert and K. Stinson, Gamma-convergence of a nonlocal perimeter arising in adversarial machine learning. *Calc. Var. Part. Diff. Equ.* **63** (2024) 114,.
[24] R. Morris and R. Murray, Uniform convergence of adversarially robust classifiers. *Eur. J. Appl. Math.* **37** (2026) 43–72.
[25] L. Weigand, T. Roith and M. Burger, Adversarial flows: A gradient flow characterization of adversarial attacks. *Eur. J. Appl. Math.* **37** (2026) 123–179.
[26] P. Awasthi, N.S. Frank and M. Mohri, On the existence of the adversarial Bayes classifier. *Adv. Neural Inform. Process. Syst.* **34** (2021) 2978–2990.
[27] P. Awasthi, N.S. Frank and M. Mohri, On the existence of the adversarial Bayes classifier (extended version). arXiv preprint arXiv:2112.01694 (2021).
[28] N.S Frank and J. Niles-Weed, Existence and minimax theorems for adversarial surrogate risks in binary classification. *J. Mach. Learn. Res.* **25** (2024) 1–41.
[29] N. García Trillos, M. Jacobs, and J. Kim, On the existence of solutions to adversarial training in multiclass classification. *Eur. J. Appl. Math.* **37** (2026) 3–23.
[30] M.S. Pydi and V. Jog, The many faces of adversarial risk. *Adv. Neural Inform. Process. Syst.* **34** (2021) 10000–10012.
[31] R.T. Rockafellar, S. Uryasev *et al.*, Optimization of conditional value-at-risk. *J. Risk* **2** (2000) 21–42.
[32] N. Dunford and J.T. Schwartz, *Linear Operators: General Theory*. Linear Operators. Interscience Publishers (1958).
[33] A. Chambolle, A. Giacomini and L. Lussardi, Continuous limits of discrete perimeters. *ESAIM Math. Model. Numer. Anal.* **44** (2010) 207–230.
[34] A. Visintin, Pattern evolution. *Ann. Scuola Normale Superiore Pisa-Classe Sci.* **17** (1990) 197–225.
[35] A. Chambolle, M. Morini and M. Ponsiglione, Nonlocal curvature flows. *Arch. Rational Mech. Anal.* **218** (2015) 1263–1329.
[36] J. M Mazón, M. Solera and J. Toledo, The total variation flow in metric random walk spaces. *Calc. Var. Part. Differ. Equ.* **59** (2020) 1–64.
[37] A. Braides, *Gamma-Convergence for Beginners*. Oxford University Press (2002).
[38] G. Dal Maso, *An Introduction to* Γ*-Convergence*. Birkhäuser, Boston (1993).
[39] A. Braides and L. Truskinovsky, Asymptotic expansions by Γ-convergence. *Continuum Mech. Thermodyn.* **20** (2008) 21–62.
[40] M.D. Zeiler, Adadelta: an adaptive learning rate method. arXiv preprint arXiv:1212.5701 (2012).
[41] K. He, X. Zhang, S. Ren and J. Sun, Deep residual learning for image recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016) 770–778.

## Appendix A. List of functionals and notation

The following table summarizes all objective functions, perimeters, and regularization functionals used throughout this paper. We organize them by their domain: functionals for hard classifiers (sets $A \in \mathcal{A}$) and functionals for soft classifiers (functions $u : \mathcal{X} \to [0,1]$).

| Functional | Description | References |
|---|---|---|
| *Risk functionals for hard classifiers (sets $A \in \mathcal{A}$)* | | |
| $\mathrm{R}_{\mathrm{std}}(A)$ | Standard risk (risk minimization). | (1.1) |
| $\mathrm{R}_{\mathrm{adv}}(A)$ | Adversarial training risk; uses $\varepsilon$-fattening ($A^{\oplus\varepsilon}$). | (1.2) |
| $\mathrm{R}_{\mathrm{prob}}(A)$ | Original PRL risk from [10]; uses $p$-ess sup. | (1.3) |
| $\mathrm{R}_{\mathrm{prob}_\Psi}(A)$ | Generalized PRL risk for arbitrary $\Psi$. Recovers $\mathrm{R}_{\mathrm{prob}}$ when $\Psi(t) = \mathbf{1}_{t>p}$. | (1.7) |
| $\mathrm{R}_{\text{m-prob}}(A)$ | Modified PRL risk (abbrev. **m-PRL**): $\mathrm{R}_{\text{m-prob}}(A) = \mathrm{R}_{\mathrm{std}}(A) + \mathrm{Per}_{\mathrm{prob}}(A)$. | (1.4) |
| $\mathrm{R}_{\text{m-prob}_\Psi}(A)$ | Modified PRL with general $\Psi$: $\mathrm{R}_{\text{m-prob}_\Psi}(A) = \mathrm{R}_{\mathrm{std}}(A) + \mathrm{Per}_{\mathrm{prob}_\Psi}(A)$. | (1.8) |
| *Perimeter functionals for hard classifiers* | | |
| $\mathrm{Per}_{\mathrm{adv}}(A)$ | Adversarial perimeter; measures $\varepsilon$-boundary contribution. | Section 1.2 |
| $\mathrm{Per}_{\mathrm{prob}}(A)$ | Probabilistic perimeter with threshold $\Psi(t) = \mathbf{1}_{t>p}$. | (1.5) |
| $\mathrm{Per}_{\mathrm{prob}_\Psi}(A)$ | Probabilistic perimeter with general $\Psi$. | (1.9) |
| *Regularization functionals for soft classifiers ($u : \mathcal{X} \to [0,1]$)* | | |
| $\mathrm{J}_{\mathrm{prob}_\Psi}(u)$ | Nonlocal regularization for soft classifiers. Satisfies $\mathrm{J}_{\mathrm{prob}_\Psi}(\mathbf{1}_A) = \mathrm{Per}_{\mathrm{prob}_\Psi}(A)$. | (1.10) |
| $\mathrm{TV}_{\mathrm{prob}_\Psi}(u)$ | Total variation extension: $\mathrm{TV}_{\mathrm{prob}_\Psi}(u) = \int_0^1 \mathrm{Per}_{\mathrm{prob}_\Psi}(\{u > t\})\,\mathrm{d}t$. | (2.11) |
| $\mathrm{S}_{\mathrm{prob}_\Psi}(u)$ | Soft classifier risk. | (2.21) |
| $\mathrm{S}_{\text{m-prob}_\Psi}(u)$ | Modified soft classifier risk: $\mathrm{S}_{\text{m-prob}_\Psi}(u) = \mathbb{E}_{(x,y)\sim\mu}\left[\lvert u(x) - y\rvert\right] + \mathrm{J}_{\mathrm{prob}_\Psi}(u)$. | (2.20) |
| *Key choices of $\Psi$ and their properties* | | |
| $\Psi(t) = \mathbf{1}_{t>p}$ | Hard threshold at $p$; gives original PRL. Not concave. | (1.3) |
| $\Psi_p(t) = \min\{\frac{t}{p}, 1\}$ | CVaR relaxation; smallest concave function $\geq \mathbf{1}_{t>p}$. | Prop. 2.6 |
| $\Psi_0(t) = \mathbf{1}_{t>0}$ | Threshold at zero; used for analyzing Gamma-convergence. | Prop. 4.3 |

**Key relationships:**

- For hard classifiers: $\mathrm{R}_{\text{m-prob}_\Psi}(A) = \mathrm{R}_{\mathrm{std}}(A) + \mathrm{Per}_{\mathrm{prob}_\Psi}(A)$ (geometric regularization interpretation).
- Consistency: $\mathrm{J}_{\mathrm{prob}_\Psi}(\mathbf{1}_A) = \mathrm{TV}_{\mathrm{prob}_\Psi}(\mathbf{1}_A) = \mathrm{Per}_{\mathrm{prob}_\Psi}(A)$ for all $A \in \mathcal{A}$.
- Ordering: $\mathrm{TV}_{\mathrm{prob}_\Psi}(u) \leq \mathrm{J}_{\mathrm{prob}_\Psi}(u)$ when $\Psi$ is concave and non-decreasing.
- Limits: As $p \to 0$, $\mathrm{R}_{\text{m-prob}_{\Psi_p}} \to \mathrm{R}_{\mathrm{adv}}$; as $p \to \infty$, $\mathrm{R}_{\text{m-prob}_{\Psi_p}} \to \mathrm{R}_{\mathrm{std}}$.

## Appendix B. Pathological points in the modified PRL model

Contrary to the situation presented in Figure 1, where we describe the unintuitive behavior of a solution of the original PRL model, the modified PRL model possesses an interesting stability property that prevents its minimizers from having pathological points (recall the discussion in Sect. 6.2), a notion that we make more precise below. This property is independent of any specific details of the family $\{\mathfrak{m}_x\}_{x\in\mathcal{X}}$ and for example holds in the Euclidean setting when we set $\mathfrak{m}_x$ to be the uniform measure over the ball $B_\varepsilon(x)$.

**Proposition B.1.** *Let $\Psi : [0,1] \to [0,1]$ be such that $\Psi(0) = 0$ and $\Psi(1) = 1$. Under Assumption 4.9, if $A$ is a minimizer of $\mathrm{R}_{\text{m-prob}_\Psi}$, then for $\rho$-a.e. $x \in \mathcal{X}$ with $\frac{\mathrm{d}\rho_0}{\mathrm{d}\rho}(x) > \frac{\mathrm{d}\rho_1}{\mathrm{d}\rho}(x)$ we have: $\mathbf{1}_A(x) = 1$ implies $\mathfrak{m}_x\text{-}\operatorname{ess\,sup} \mathbf{1}_A = 1$. Likewise, for $\rho$-a.e. $x \in \mathcal{X}$ with $\frac{\mathrm{d}\rho_1}{\mathrm{d}\rho}(x) > \frac{\mathrm{d}\rho_0}{\mathrm{d}\rho}(x)$, $\mathbf{1}_{A^c}(x) = 1$ implies $\mathfrak{m}_x\text{-}\operatorname{ess\,sup} \mathbf{1}_{A^c} = 1$.*

*Proof.* Let us define the set of pathological points for the distribution $\rho_0$ as those which are misclassified (according to the Bayes classifier) even though, with probability one, their perturbations lead to a correct classification. For this let (recall $\rho_i \ll \rho := \rho_0 + \rho_1$ for $i \in \{0,1\}$)

$$N := \left\{ x \in \mathcal{X} \,:\, \frac{\mathrm{d}\rho_0}{\mathrm{d}\rho}(x) > \frac{\mathrm{d}\rho_1}{\mathrm{d}\rho}(x) \right\}$$

be the set of points which are strictly more likely to get the label zero than one, and define

$$\mathcal{P}_0 := \{ x \in A \cap N \,:\, \mathfrak{m}_x(A) = 0 \}\,.$$

So the points $x \in \mathcal{P}_0$ are assigned the label one although the Bayes classifier assigns the label zero and likewise all perturbations using $\mathfrak{m}_x$. Note that $\mathcal{P}_0 \subset A$ and define the competitor set $\tilde{A} := A \setminus \mathcal{P}_0$. Aiming for a contradiction, we will show that if $\rho(\mathcal{P}_0) > 0$, then $\tilde{A}$ would have a strictly smaller value of $\mathrm{R}_{\text{m-prob}_\Psi}$ than the minimizer $A$. We first observe that, thanks to Assumption 4.9 and the definition of $\mathcal{P}_0$, we have

$$\mathfrak{m}_x(A) = \mathfrak{m}_x(\tilde{A}) \qquad \forall x \in \mathcal{X}.$$

Hence, using also (1.8), the energy difference of $A$ and $\tilde{A}$ simplifies to

$$\begin{aligned}
&\mathrm{R}_{\text{m-prob}_\Psi}(A) - \mathrm{R}_{\text{m-prob}_\Psi}(\tilde{A}) \\
&= \rho_0(\mathcal{P}_0) - \rho_1(\mathcal{P}_0) - \int_{\mathcal{P}_0} \psi(\mathfrak{m}_x(A))\,\mathrm{d}\rho_0(x) + \int_{\mathcal{P}_0} \psi(\mathfrak{m}_x(A))\,\mathrm{d}\rho_1(x) \\
&= \rho_0(\mathcal{P}_0) - \rho_1(\mathcal{P}_0) \\
&= \int_{\mathcal{P}_0} \left( \frac{\mathrm{d}\rho_0}{\mathrm{d}\rho} - \frac{\mathrm{d}\rho_1}{\mathrm{d}\rho} \right) \mathrm{d}\rho > 0
\end{aligned}$$

where, for the second equality, we used that $\mathfrak{m}_x(A) = 0$ for all $x \in \mathcal{P}_0$ and that $\Psi(0) = 0$, and for the last inequality we used the definition of $\mathcal{P}_0$ and $N$. This is a contradiction to minimality and, hence, we have $\rho(\mathcal{P}_0) = 0$, which implies that, for $\rho$-almost every $x \in \mathcal{X}$ with $\frac{\mathrm{d}\rho_0}{\mathrm{d}\rho}(x) > \frac{\mathrm{d}\rho_1}{\mathrm{d}\rho}(x)$ and $x \in A$, we have $\mathfrak{m}_x(A) > 0$ and hence $\mathfrak{m}_x\text{-}\operatorname{ess\,sup} \mathbf{1}_A = 1$. The second statement is proved analogously. □

## Appendix C. Computational aspects of PRL

### C.1 Pseudocode for geometric probabilistically robust learning

In Algorithm 1 we provide a pseudocode for parametrized classifiers $f \equiv f_\theta : \mathcal{X} \to \mathcal{Y}$ based on stochastic gradient descent with batch size $B$. Furthermore, it involves a sample size of $M$ samples from a distribution $\mathfrak{m}_x$ around an input $x \in \mathcal{X}$, a learning rate $\eta_\alpha$ for the inner optimization in CVaR, and a learning rate $\eta$ for the parameter updates. The

pseudocode is a straightforward generalization of [10], Algorithm 1 and we implemented it in their code framework[2]. The code which can be used to reproduce our results is part of the supplementary material of this paper.

**Algorithm 1** Proposed algorithm for solving (5.3) for $p \in (0,1)$.

**for** minibatch $(x_j, y_j)_{j=1}^B$ **do**
  **for** $T$ steps **do** ▷ Approximate solution of inner problem
    Draw $x'_k \sim \mathfrak{m}_{x_j}$, $k = 1, \dots, M$
    $g_{\alpha_j} \leftarrow 1 - \frac{1}{pM} \sum_{k=1}^{M} \mathbf{1}\ell(f_\theta(x'_k), y_j) \geq \alpha_j$
    $\alpha_j \leftarrow \alpha_j - \eta_\alpha g_{\alpha_j}$
  **end for**
  $S_j \leftarrow \alpha_j + \frac{1}{pM} \sum_{k=1}^{M} (\ell(f_\theta(x'_k), y_j) - \alpha_j)_+$ ▷ Approximate value of $\mathrm{CVaR}_p$
  **if** $S_j > \ell(f_\theta(x_j), y_j))$ **then** ▷ If $\mathrm{CVaR}_p$ kicks in
    $g_j \leftarrow \frac{1}{pM} \sum_{k=1}^{M} \nabla_\theta \left(\ell(f_\theta(x'_k), y_j) - \alpha_j\right)_+$
  **else** ▷ If it doesn't
    $g_j \leftarrow \nabla_\theta \ell(f_\theta(x_j), y_j)$
  **end if**
  $g \leftarrow \frac{1}{B} \sum_{j=1}^{B} g_j$ ▷ Compute full $\theta$-gradient
  $v \leftarrow \mathrm{optimizer}(g)$ ▷ AdaDelta or SGD(+M)
  $\theta \leftarrow \theta - \eta v$ ▷ Update parameters
**end for**

### C.2 Training details

Hyperparameter values specific to Algorithm 1 used in training are presented in Table C.1. Following [10], we use AdaDelta [40] for MNIST experiments and SGD with momentum for CIFAR-10 experiments. The MNIST experiments use a CNN architecture with two convolutional layers (32 and 64 filters, size 3x3), two dropout layers (dropout probabilities 0.25, 0.5), and two fully connected layers (dimensions 9216 to 128 and 128 to 10). A ResNet-18 [41] is used in the CIFAR-10 experiments. The hyperparameter values used for these algorithms are contained in `hparams_registry.py` in the accompanying code.

[2] https://github.com/arobey1/advbench

TABLE C.1. Hyperparameters used for training. The probability distribution $\mathfrak{m}_x$ is always taken to be the uniform distribution over the ball $B_\varepsilon(x)$. Note that $p$ is called `beta` in the code. For consistency we always used the same hyperparameter values for the "Geometric" and "Original" versions.

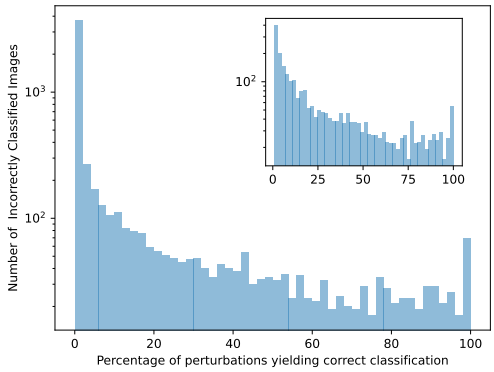
| **Data** | $p$ | $\varepsilon$ | $\eta_\alpha$ | M | T |
|---|---|---|---|---|---|
| **MNIST** | 0.01 | 0.3 | 0.1 | 20 | 5 |
| | 0.1 | 0.3 | 1.0 | 20 | 5 |
| | 0.3 | 0.3 | 1.0 | 20 | 5 |
| | 0.5 | 0.3 | 1.0 | 20 | 5 |
| **CIFAR-10** | 0.01 | 8/255 | 0.1 | 20 | 5 |
| | 0.1 | 8/255 | 1.0 | 20 | 5 |
| | 0.3 | 8/255 | 1.0 | 20 | 5 |
| | 0.5 | 8/255 | 1.0 | 20 | 5 |

### C.3 Computational resources used

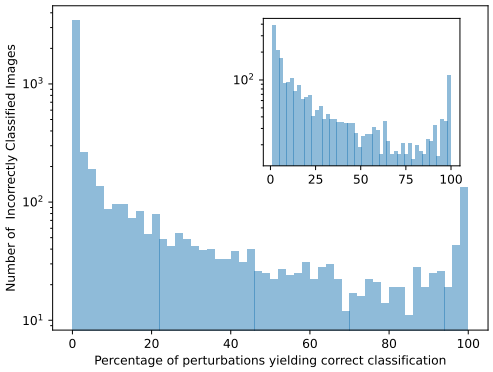
We performed the majority of the prototyping and some experimentation on a LambdaLabs Vector workstation equipped with 3 NVIDIA A6000 GPUs. We estimate that we used approximately 500 GPU-hours on this machine. We supplemented this with 550 GPU-hours of cloud compute – using the Lambda GPU cloud – predominately on instances equipped with a single A10 GPU.

### C.4 Further numerical experiments

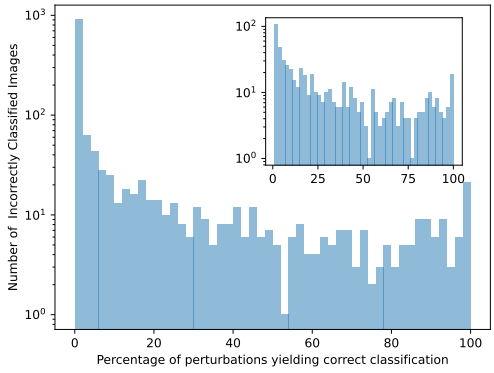
In Figures C.1 to C.3 we provide further numerical experiments on pathological points with parameters $p = 0.1, 0.3$, and $0.5$.



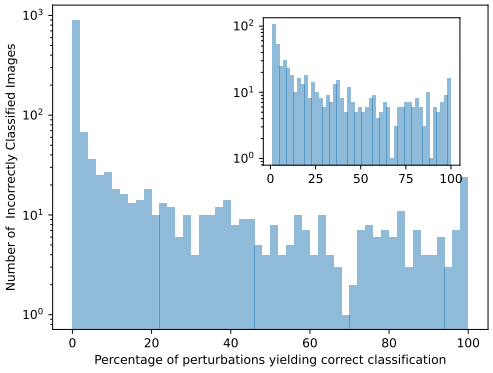

(A) $p = 0.1$, train, original PRL




(B) $p = 0.1$, train, m-PRL

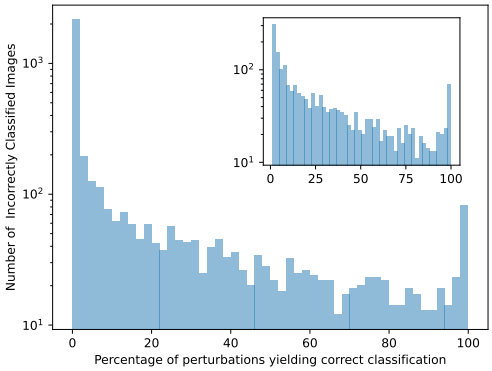



(C) $p = 0.1$, test, original PRL

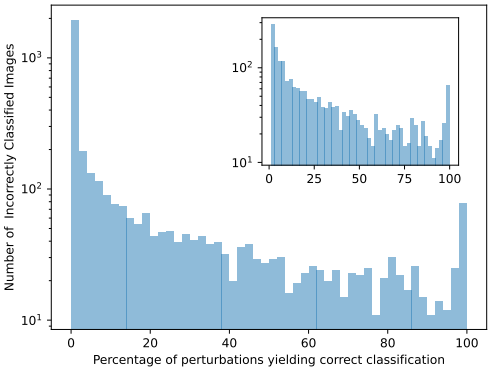



(D) $p = 0.1$, test, m-PRL

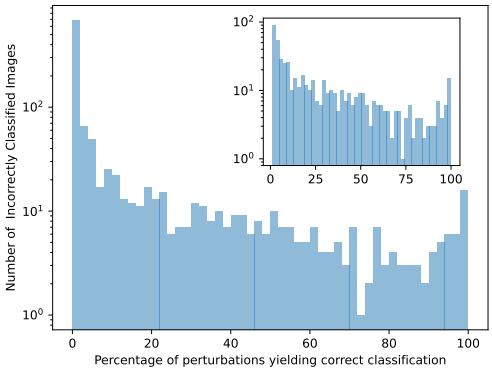
FIGURE C.1. Histograms showing the distribution of percentages of correctly classified perturbations among misclassified images for both original and m-PRL with parameter $p = 0.1$.



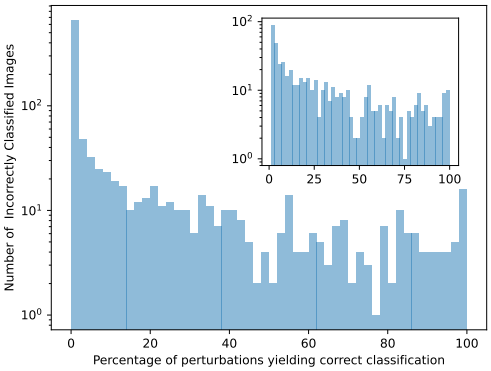

(A) $p = 0.3$, train, original PRL




(B) $p = 0.3$, train, m-PRL

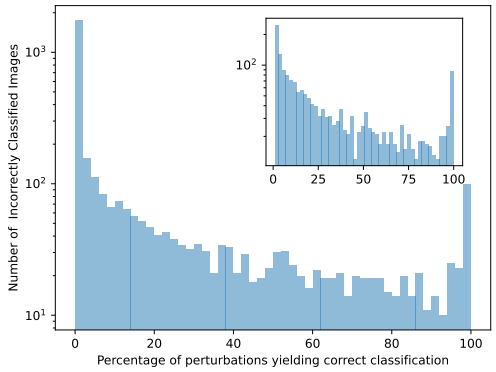



(C) $p = 0.3$, test, original PRL

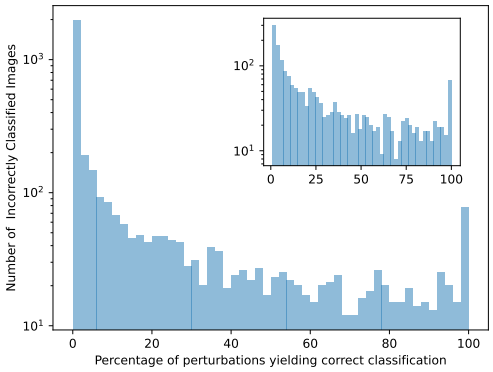



(D) $p = 0.3$, test, m-PRL

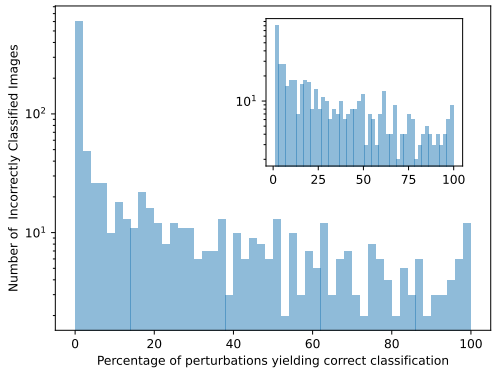
FIGURE C.2. Histograms showing the distribution of percentages of correctly classified perturbations among misclassified images for both original and m-PRL with parameter $p = 0.3$.



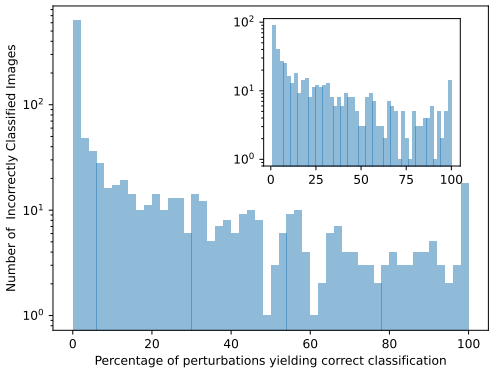

(A) $p = 0.5$, train, original PRL




(B) $p = 0.5$, train, m-PRL




(C) $p = 0.5$, test, original PRL




(D) $p = 0.5$, test, m-PRL

FIGURE C.3. Histograms showing the distribution of percentages of correctly classified perturbations among misclassified images for both original and m-PRL with parameter $p = 0.5$.