EXPLAINING THE SALIENCY MAP SPARSITY OF ADVERSARIALLY-TRAINED NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding why deep neural networks make a given prediction is of great importance for their safe deployment. In computer vision, saliency maps, which highlight the image region most influential for a prediction, remain a widely-used form of explanation. An empirical observation is the apparent sparsity of gradient saliency maps of adversarially-trained neural networks. In this paper, we propose a theoretical explanation of this phenomenon for two-layer ReLU networks. We build on the established equivalence of adversarial training to the minimization of the empirical risk with weight-decay penalization and an added adversarial total variation term – valid for certain loss functions. As the number of data points and neurons grows and the regularization parameters are sent to zero at appropriate rates, we prove that minimizers converge to a Bayes classifier with minimal gradient and Barron norm. Sparsity appears since for adversarial training with ℓ_∞-attacks the gradient norm is anisotropic and favors axis-aligned / sparse gradients. We illustrate our theoretical findings experimentally by evaluating the gradient ℓ_1-norm and thresholded sparsity of naturally versus adversarially trained models.

1 INTRODUCTION

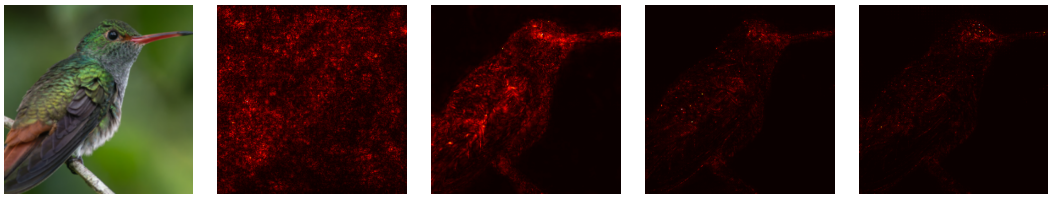
Despite the immense success of deep neural networks, two fundamental challenges continue to limit their deployment in high-stakes domains: their vulnerability to small, imperceptible adversarial perturbations (Goodfellow et al., 2015), and the "black-box" nature of their decision-making. Addressing the former, *adversarial training* (Madry et al., 2018), which mitigates the problem by incorporating adversarial examples in the training, has emerged as one of the most successful defenses and maintained its relevance even in the era of large language models (Xhonneux et al., 2024).

To address the latter, a wide range of methods have been proposed in recent years under the umbrella of *Explainable AI (XAI)* (Linardatos et al., 2020). Among them, local feature importance methods aim to identify the input features that are the most influential for a model's decision at a given point. In computer vision, a prominent example is given by saliency maps based on the gradient $\nabla f_\theta(x)$ of the model with respect to the input. The most important features can then be easily visualized as an image overlay, see Figure 1. Motivated by the observation that plain input gradients are often noisy and visually hard to interpret, a number of extension have been proposed, notably gradient times input (Shrikumar et al., 2017), integrated gradient (Sundararajan et al., 2017), and SmoothGrad (Smilkov et al., 2017). A line of work also developed around specific architecture, such as GradCAM for convolutional neural networks (Selvaraju et al., 2017).

Previous work has shown empirically that adversarial training not only increases robustness of the trained model but also **leads to sparse, perceptually aligned, and hence more interpretable gradient saliency maps** (Zhang & Zhu, 2019; Etmann et al., 2019; Chalasani et al., 2020); see Figure 1. This observation is surprising, since adversarial training was introduced to improve robustness against worst-case perturbations, with no explicit objective related to interpretability, or even the gradient structure of the predictor.

In this paper, we provide a theoretical explanation for this phenomenon. The starting point of our journey is to notice that adversarial training (with the ℓ_1-loss) can be reformulated as empirical risk minimization with an additional total variation (TV) term depending on the adversarial budget ε (Bungert et al., 2023). Focusing on two-layer neural networks, we show that, in the limit of infinite width and vanishing ε, the limiting optimization problem reduces to finding a Bayes classifier which

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Original (b) Natural training (c) $\varepsilon = 0.5/255$ (d) $\varepsilon = 1/255$ (e) $\varepsilon = 2/255$

Figure 1: Comparison of saliency maps for a naturally trained ResNet50 on ImageNet versus adversarially trained models across varying ℓ_∞ perturbation budgets ε. As predicted by Theorem 1, with decreasing adversarial budget ε the gradient saliency maps converge to a sparse one whereas natural training (corresponding to $\varepsilon = 0$) produces a noisy saliency map.

minimizes a balance of the average dual norm of the classifier's input gradients and a measure of its complexity. In particular, for the ℓ_∞-bounded perturbations typically used in adversarial training, this yields an ℓ_1-norm penalty of the input gradients. Informally, adversarial training in this regime selects a Bayes classifier that balances accuracy, complexity, and average ℓ_1-norm of its gradients. Since ℓ_1-type regularization is well-known to promote sparsity, this provides a mathematical explanation for the sparse saliency maps observed in adversarially trained models. We illustrate this result in a simple setting in Figure 2, see Appendix F for more details.

Our main technical tool is Gamma-convergence (see Braides (2002) for an introduction), which is key in showing that the empirical risk minimizers converge to minimizers of the limit problem. Existing results which we review in Section 4 are obtained in an unrealistic non-parametric setting with a population risk, whereas in practice one uses neural networks trained on finite datasets. In this work, we bridge this gap by embedding our problem into the homogeneous Barron space (Park et al., 2023) and deriving generalization bounds and quantitative approximation results to transition from the idealized and continuous theory to the empirical and parametric setting.

Our main contributions are as follows:

- Theorem 1 states that as the number of data points and the width of the neural networks go to infinity and the adversarial budget and the weight decay coefficient go to zero at specific coupled rates, every sequence of empirical minimizers admits a subsequence that converges to a Bayes classifier with optimal balance of its Barron semi-norm with a TV functional of the form $\int \|\nabla f(x)\|_\star \, \mathrm{d}\mu(x)$, where $\|\cdot\|_\star$ is the dual norm to the one restricting adversarial attacks. From this, we deduce an explanation for the observed sparsity of saliency maps.
- We support our theoretical findings with numerical experiments, demonstrating that adversarial training systematically reduces the average ℓ_1-norm of the input gradients. Furthermore, by evaluating thresholded gradient sparsity, we provide a quantitative empirical validation of the observations of Zhang & Zhu (2019); Etmann et al. (2019) and confirm that this implicit ℓ_1-regularization translates to the sparse saliency maps.

2 Setting

We study a binary classification problem, where the distribution of the training pairs $(x, y) \in \Omega \times \{0, 1\}$ is described by some probability measure μ on $\Omega \times \{0, 1\}$. In our setting, we assume that $\Omega \subset \mathbb{R}^d$ is a bounded domain with Lipschitz boundary. We define $\mu_i := \mu(\cdot \times \{i\})$ for $i \in \{0, 1\}$, so $\frac{\mu_i}{\mu_i(\Omega)}$ is the probability distribution of the data points corresponding to the label i.

Two-Layer Neural Networks. In this work, we consider two-layer neural networks with m hidden neurons of the form

$$f_m(x) := b_0 + \sum_{i=1}^{m} a_i \sigma(w_i \cdot x + b_i), \qquad x \in \mathbb{R}^d, \tag{1}$$

for $\theta := (a, W, b) \in \mathbb{R}^m \times \mathbb{R}^{m \times d} \times \mathbb{R}^{m+1} =: \Theta_m$, where w_i is the i-th row of the matrix W and σ denotes the ReLU activation function defined as $\sigma(t) := \max\{0, t\}$. The matrix W collects

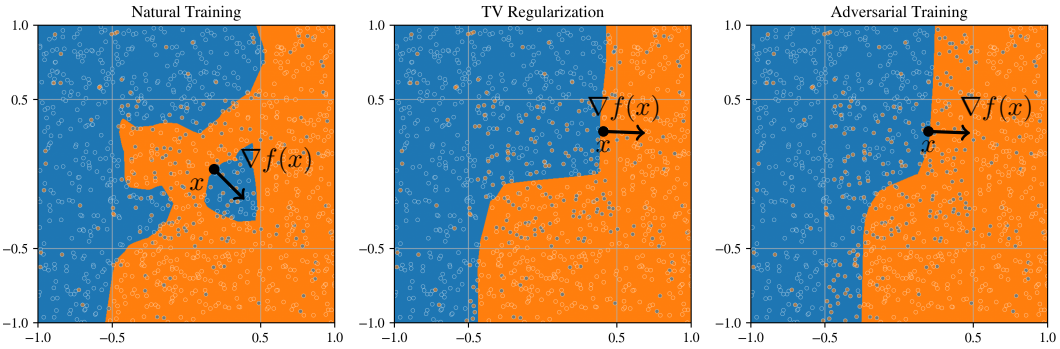


Figure 2: Decision boundaries and training data for a data distribution with non-uniquely determined Bayes classifier in the strip $[-0.5, 0.5]^2$: Natural training (left) has a highly oscillatory decision boundary arising from fitting the training data. Nonlocal TV regularization (center, analyzed here) and adversarial training (right) have a decision boundary with small anisotropic perimeter and exhibit approximate gradient sparsity, as indicated by the axis-parallel parts of the decision boundary. See Appendix F for more details.

the weights of the hidden layer, while $(b_1, \dots, b_m)$ forms the bias vector. Additionally, a denotes the output weights and b_0 acts as the output bias. Often, when training neural networks, the weight decay regularizer $R_{WD}(\theta) \coloneqq \frac{1}{2}\sum_{i=1}^m \big(a_i^2 + \|w_i\|_2^2\big)$ is added to the objective, where $\|\cdot\|_2$ denotes the Euclidean norm. Importantly, this functional does not control the bias vector b, which is typical.

Barron space. Taking the infinite width limit (Bach, 2017; E et al., 2022) of such neural networks, one ends up with functions of the form

$$f_{\pi,y}(x) \coloneqq y + \int_{\mathbb{R}\times\mathbb{R}^d\times\mathbb{R}} a\big(\sigma(w\cdot x + b) - \sigma(b)\big)\,\mathrm{d}\pi(a,w,b), \qquad x \in \mathbb{R}^d, \tag{2}$$

where $y \in \mathbb{R}$ captures the function's value at the origin, and π is a probability measure on $\mathbb{R}\times\mathbb{R}^d\times\mathbb{R}$. With a slight abuse of notation, the counterpart of the weight decay regularizer is given by

$$R_{WD}(\pi) \coloneqq \frac{1}{2}\int_{\mathbb{R}\times\mathbb{R}^d\times\mathbb{R}} \big(a^2 + \|w\|_2^2\big)\,\mathrm{d}\pi(a,w,b).$$

For a given function f written as (2), its representation via a measure π is not unique and hence $R_{WD}(\pi)$ is not well-suited to measure the complexity of the function. Instead, we define the Barron semi-norm $[f]_{\mathcal{B}}$ as the infimum of the weight decay regularizer over all possible representations,

$$[f]_{\mathcal{B}(\Omega)} \coloneqq \inf_{f_{\pi,y}=f \text{ on } \Omega} R_{WD}(\pi).$$

The *homogeneous Barron space* $\mathcal{B}(\Omega)$ is then defined as the set of all functions that admit a representation $f = f_{\pi,y}$ with a finite Barron semi-norm,

$$\mathcal{B}(\Omega) \coloneqq \{f\colon \Omega \to \mathbb{R} \,|\, [f]_{\mathcal{B}(\Omega)} < \infty\}.$$

For more details on this space and the previous definitions, we refer to Appendix A, Park et al. (2023, Appendix C) and Ongie et al. (2020).

Remark 2.1. Neural networks of the form (1) are contained in $\mathcal{B}(\Omega)$. Setting $y = b_0 + \sum_{i=1}^m a_i\sigma(b_i)$ and defining the probability measure $\pi = \frac{1}{m}\sum_{i=1}^m \delta_{\sqrt{m}(a_i,w_i,b_i)}$ yields $f_{\pi,y} = f_m$. Consequently, we get that $[f_m]_{\mathcal{B}} \le R_{WD}(\pi) = R_{WD}(\theta)$.

Adversarial Training. *Adversarial training* aims to minimize the adversarial population risk which, in our binary classification setting, is given by

$$\inf_{f\in\mathcal{H}} \int_{\Omega\times\{0,1\}} \sup_{\tilde{x}\in B_\varepsilon(x)} \ell(f(\tilde{x}), y)\,\mathrm{d}\mu(x,y), \tag{3}$$

where $\ell : \mathbb{R}\times\{0,1\}$ is a loss-function, $\varepsilon > 0$ is the adversarial budget, and minimization takes place over a certain class of functions $\mathcal{H}$. Here and in the following, the balls $B_\varepsilon(x)$ are taken with respect to an *arbitrary* norm $\|\cdot\|$ on $\mathbb{R}^d$ (not necessarily the Euclidean norm) and by $\|\cdot\|_\star$, we denote its dual norm, defined via $\|y\|_\star \coloneqq \sup_{\|x\|\le 1}\langle y, x\rangle$. In practice, the norm considered is often the ℓ_∞-norm (Madry et al., 2018; Croce et al., 2021), leading to the ℓ_1-norm as the dual norm.

Total Variation Regularization. Let $\mathcal{H}$ be an hypothesis class of functions from Ω to $\mathbb{R}$. As is typical, we define the *population risk* as

$$\mathcal{R}(f) \coloneqq \int_{\Omega\times\{0,1\}} \ell(f(x),y)\,\mathrm{d}\mu(x,y), \qquad f\in\mathcal{H}. \tag{4}$$

Before introducing our primary empirical objective, we introduce the following motivating population-level problem

$$\inf_{f\in\mathcal{H}} \mathcal{R}(f) + \varepsilon\,\mathrm{TV}_\varepsilon(f;\mu)\,, \tag{5}$$

where $\mathrm{TV}_\varepsilon(f;\mu)$ is a *nonlocal total variation* given by

$$\mathrm{TV}_\varepsilon(f;\mu) \coloneqq \int_\Omega \frac{\sup_{B_\varepsilon(x)\cap\Omega} f - f(x)}{\varepsilon}\mu_0(x)\,\mathrm{d}x + \int_\Omega \frac{f(x) - \inf_{B_\varepsilon(x)\cap\Omega} f}{\varepsilon}\mu_1(x)\,\mathrm{d}x. \tag{6}$$

Note that the objective in (5) is very similar to the TRADES method (Zhang et al., 2019) and both coincide with adversarial training for the ℓ_1-loss (Bungert et al., 2023). Intuitively, the total variation TV_ε is closely connected to a nonlocal, ε-dependent surface content of the decision boundary, and quantifies the adversarial robustness of a classifier (Bungert & Stinson, 2024). One should think of (5) as **loosely equivalent to population-level adversarial training** with budget ε, with strict mathematical equivalence for specific choices of loss functions and hypothesis classes. Our motivation for working with (5) instead of standard adversarial training (3) is that the former is amenable to our asymptotic analysis thanks to deep mathematical properties of the nonlocal total variation TV_ε.

Key optimization problem. In our setting, we are given $(X_1,Y_1),\dots,(X_n,Y_n)\sim\mu$, n i.i.d. samples, and restrict our attention to $f=f_\theta$, where f_θ with $\theta\in\Theta_{m_n}$ belongs to the class of two-layer neural networks with ReLU activation and $m_n\in\mathbb{N}$ neurons (1). The population risk (4) is approximated, as usual, by the *empirical risk*

$$\widehat{\mathcal{R}}_n(f) \coloneqq \frac{1}{n}\sum_{i=1}^n \ell(f(X_i),Y_i)\,, \tag{7}$$

whereas the empirical counterpart of $\mathrm{TV}_\varepsilon(f;\mu)$ is given by

$$\widehat{\mathrm{TV}}_{\varepsilon_n,n}(f;\mu) \coloneqq \frac{1}{n}\left(\sum_{i\,:\,Y_i=0} \frac{\sup_{B_{\varepsilon_n}(X_i)\cap\Omega} f - f(X_i)}{\varepsilon_n} + \sum_{i\,:\,Y_i=1} \frac{f(X_i) - \inf_{B_{\varepsilon_n}(X_i)\cap\Omega} f}{\varepsilon_n}\right). \tag{8}$$

With these definitions in hand, the natural finite-sample version of (5) is

$$\inf_{\theta\in\Theta_{m_n}} \widehat{\mathcal{R}}_n(f_\theta) + \varepsilon_n\widehat{\mathrm{TV}}_{\varepsilon_n,n}(f_\theta;\mu)\,.$$

Minimization problems over a space of neural networks usually do not possess solutions since parameters might have to blow up to fit the data. Therefore, we add the usual weight regularization term $R_{WD}(\theta)$ defined previously which ensures existence as well as a mathematical compactness property in the Barron space $\mathcal{B}(\Omega)$, making sure that sequences of minimizers $(\theta_n)_{n\in\mathbb{N}}$ of problem (9) below do not diverge. We want to stress that, while the weight decay term is usually not written explicitly as part of the objective in the practice of adversarial training, it is often implicitly present in the optimization procedure.[1] Letting the adversarial budget $\varepsilon_n>0$ and the weight decay parameter $\lambda_n>0$ depend on n, we arrive at the following optimization problem, which is at the core of our study:

$$\inf_{\theta\in\Theta_{m_n}} \widehat{\mathcal{R}}_n(f_\theta) + \varepsilon_n\widehat{\mathrm{TV}}_{\varepsilon_n,n}(f_\theta;\mu) + \lambda_n R_{WD}(\theta)\,. \tag{9}$$

In the same spirit as the previous paragraph, one should think of (9) as loosely equivalent to finite-sample adversarial training with budget ε_n and weight decay λ_n. We show in Figures 2, 4 and 5 for the toy dataset and in Figures 6 and 7 for CIFAR-10 that empirically they produce very similar results, with even a slight advantage for (9) in terms of robust accuracy and sparsity on CIFAR-10.

[1] See, e.g., the adversarial training code of the Madry lab: `https://github.com/MadryLab/cifar10_challenge`

3 MAIN RESULT

In this section, we state our main result: under suitable scalings of the parameters $\varepsilon_n, \lambda_n \to 0$ and $m_n \to \infty$, a sequence of solutions to (9) converges in a weak sense to a Bayes classifier with specific properties. To make this precise, we first define

$$\mathrm{TV}(f;\mu) \coloneqq \int_\Omega \|\nabla f(x)\|_\star (\mu_0(x) + \mu_1(x))\,\mathrm{d}x$$

for $f \in \mathcal{B}$, which arises as limit of TV_ε defined in (6) as $\varepsilon \to 0$. We note that this expression is well-defined for Barron functions, since they are Lipschitz-continuous (Lemma A.1) and in particular differentiable almost everywhere by Rademacher's theorem. We additionally define the functional

$$F(f) \coloneqq \begin{cases} \mathrm{TV}(f;\mu) + \alpha[f]_{\mathcal{B}}, & f \in \mathcal{B} \text{ and } \mathcal{R}(f) = \inf\limits_{\mathcal{B}} \mathcal{R}, \\ \infty, & \text{otherwise.} \end{cases} \tag{10}$$

In the whole paper, we shall assume the existence of a Barron function $f^\star \in \mathcal{B}$ that minimizes the population risk $\mathcal{R}$, i.e., $\mathcal{R}(f^\star) = \inf_{\mathcal{B}} \mathcal{R}$. Otherwise, $F \equiv \infty$ and there is nothing to show.

Minimizers of the functional F are optimal classifiers (i.e., minimizers of the population risk in (7)) that additionally achieve the minimal value of the sum of the total variation $\mathrm{TV}(f;\mu)$ and the Barron semi-norm $\alpha[f]_{\mathcal{B}}$. Hence, adversarial training acts as selection procedure to pick a robust Bayes classifier as the adversarial budget shrinks. In a nutshell, Theorem 1 below states that the limit of solutions to (9) is a minimizer of the functional F.

We make the following assumptions on the loss function ℓ in (7).

Assumption 1. We assume that the loss function $\ell : \mathbb{R} \times \{0,1\} \to [0,\infty)$ is measurable, bounded and Lipschitz continuous in the first argument. Moreover, we assume that there exist constants $c, C > 0$ such that almost surely (w.r.t. the samples) $\widehat{\mathcal{R}}_n(f) - \inf \mathcal{R} \geq c$ for all but finitely many $n \in \mathbb{N}$ whenever $f \geq C$ on Ω or $f \leq -C$ on Ω.

The first part of Assumption 1 is satisfied for, e.g., clipped versions of the cross-entropy, hinge loss and ℓ_1-loss. The second part demands that trivial constant classifiers be strictly suboptimal which is true, e.g., for $\ell(\tilde{y}, y) \coloneqq |\tilde{y} - y|$. Then, for f such that $f \geq 3$, we immediately get $\widehat{\mathcal{R}}_n(f) \geq 2$ while $\inf \mathcal{R} \leq \mathcal{R}(1) \leq 1$ yielding $\widehat{\mathcal{R}}_n(f) - \inf \mathcal{R} \geq 1$. Similar arguments hold for $f \leq -2$ and the statement obviously remains true if the loss function is clipped at a sufficiently large value.

Additionally, we need to make assumptions about the data distribution μ (cf. Siethoff, 2025) which enforce the class-conditionals to be sufficiently regular and the data density to be strictly positive.

Assumption 2. We assume that the measures μ_0 and μ_1 are absolutely continuous with respect to the Lebesgue measure. By slight abuse of notation, we denote their densities by μ_0 and μ_1 as well and assume that $\mu_i \in C(\overline{\Omega})$ for $i \in \{0,1\}$. Moreover, we assume that $\mu_1 + \mu_0 > 0$ on $\overline{\Omega}$.

With these assumptions in place, we can state our main result.

Theorem 1. *Let Assumptions 1 and 2 be satisfied. Moreover, let $(\varepsilon_n)_{n\in\mathbb{N}}$ be a sequence of positive numbers converging to zero and $\lambda_n > 0$, $m_n \in \mathbb{N}$ chosen such that*

$$\lim_{n\to\infty} \left(\frac{1}{\varepsilon_n} \sqrt{\frac{\log(n)}{n}} + \frac{1}{\varepsilon_n \sqrt{m_n}} + \frac{1}{\lambda_n \sqrt{n}} \right) = 0, \qquad \alpha \coloneqq \lim_{n\to\infty} \frac{\lambda_n}{\varepsilon_n} \in (0,\infty). \tag{11}$$

Let θ_n be a solution to (9) *and $f_n = f_{\theta_n}$ be the associated neural network. Then, almost surely, the sequence (f_n) possesses a subsequence that converges in $L^1(\Omega)$ to a minimizer of F.*

Note that a similar result was shown by Bungert & Stinson (2024) and quantified by Morris & Murray (2026), however, in a non-parametric population risk setting where the Barron norm $[f]_{\mathcal{B}}$, that we need to control the parameters of the neural networks, does not appear.

Remark 3.1. We note that there exist many hyperparameter configurations satisfying the scaling assumptions. One may, for example, choose $m_n = n$, and $\lambda_n = \varepsilon_n = n^{-\frac{1}{3}}$, leading to $\alpha = 1$.

Remark 3.2. While Theorem 1 is stated for minimizers of (9), it is actually sufficient to consider quasi-minimizers as long as the gap between the achieved objective and the optimal value vanishes sufficiently fast as $n \to \infty$. Precisely, Theorem 1 also holds for θ_n with $\mathcal{L}_n(\theta_n) - \inf_{\theta \in \Theta_{m_n}} \mathcal{L}_n(\theta) = o(\varepsilon_n)$, where $\mathcal{L}_n$ denotes the objective in (9):

$$\mathcal{L}_n(\theta) \coloneqq \widehat{\mathcal{R}}_n(f_\theta) + \varepsilon_n \widehat{\mathrm{TV}}_{\varepsilon_n,n}(f_\theta; \mu) + \lambda_n R_{WD}(\theta).$$

Connection to Sparse Saliency Maps. The standard choice for the norm in adversarial training is the ℓ_∞-norm with the ℓ_1-norm as the corresponding dual norm. In this particular setting, Theorem 1 states that the minimizers of (9) converge (up to a subsequence) to an optimal classifier minimizing

$$\int_\Omega \|\nabla f(x)\|_1 (\mu_0(x) + \mu_1(x)) \,\mathrm{d}x + \alpha [f]_{\mathcal{B}}.$$

To interpret the effect of this regularizer, it is instructive to first consider a linear classifier $f(x) = w \cdot x + b$. In this regime, the gradient is constant with respect to the input ($\nabla f(x) = w$), and the integral reduces to $\|w\|_1$. Furthermore, its Barron semi-norm is $[f]_{\mathcal{B}} = \|w\|_2$ and the regularizer reduces to the elastic-net-type regularization $\|w\|_1 + \alpha \|w\|_2$ which is known to be sparsity-promoting, e.g., if the loss function is the squared ℓ_2-loss (Zou & Hastie, 2005).
In the general non-linear case, the objective imposes an *average* ℓ_1-penalty of the gradient weighted by the data distribution plus the Barron semi-norm. The former can be interpreted as group $\ell_{1,1}$-norm where a group is the array $\nabla f(x)$ for a data point x. Hence, the regularizer promotes group-sparsity in the sense that for many data points the array $\nabla f(x)$ is the zero array and for the remaining data points (the ones close to or on the decision boundary) it is sparse; see Figure 2 for an illustration.

Limitations and Extensions.

LIMITING CASES FOR α We note that we restrict our analysis to the case, where α is positive but finite, meaning that weight decay and adversarial regularization are asymptotically equivalent. On a technical level, this enables us to argue that sequences of minimizers of (9) are compact. While the case $\alpha = \infty$, corresponding to $\lambda_n \gg \varepsilon_n$ essentially boils down to the setting considered in Park et al. (2023) without the total variation regularization, the case $\alpha = 0$ presents technical challenges. Specifically, in Proposition 5.2 one would need to approximate general BV functions with neural networks and not only Barron functions. Bounding the approximating sequence's Barron semi-norm in this case is expected to require a dimension-dependent scaling of ε_n relative to n.

MULTICLASS CLASSIFICATION In a classification problem with more than two different classes, which means that labels can take values in $\{1, \dots, C\}$ for some $C \in \mathbb{N}$, we need to consider functions $f : \Omega \to \mathbb{R}^C$, and the nonlocal total variation takes the form (García Trillos et al., 2026)

$$\mathrm{TV}_\varepsilon(f; \mu) \coloneqq \frac{1}{\varepsilon} \sum_{c=1}^{C} \int_\Omega \left(f_c(x) - \inf_{B_\varepsilon(x) \cap \Omega} f_c \right) \mu_c(x) \,\mathrm{d}x,$$

where μ_c describes the distribution of data points x with label c and f_c denotes the c-th component of the vector-valued function f.
The limit functional in this case is $\mathrm{TV}(f; \mu) \coloneqq \sum_{c=1}^{C} \int_\Omega \|\nabla f_c(x)\|_\star \mu_c(x) \,\mathrm{d}x$ for sufficiently regular functions f. We expect our theory to extend to this setting, provided that the neural network approximation result (Lemma A.3) and the bounds for the Rademacher complexity (Lemmas A.4 and A.5) are generalized to vector-valued Barron functions.

GENERAL ACTIVATION FUNCTIONS The specific structure of the ReLU activation function enters at different occasions into our analysis, for example, when proving lower semi-continuity of the Barron semi-norm (Lemma A.2). However, we believe that the theory can be extended to general Lipschitz-continuous activation functions if one is willing to include the bias term into the regularizer which is favorable for showing compactness in the Barron space.

DEEP NETWORKS We only consider neural networks with one hidden layer, and these can be embedded into the Barron space. For deeper feedforward architectures such an analytical framework does not exist to this date. A notable exception are deep ResNets since they can be embedded into the flow-induced Barron space (E et al., 2022) which supports approximation by neural networks and Rademacher complexity bounds. We expect that our results can be generalized to such architectures.

ORIGINAL ADVERSARIAL TRAINING Our objective in (9) only coincides with the "original" adversarial training of Madry et al. (2018) for the zero-one loss (Bungert et al., 2023). For other loss functions, (9) is a variant of adversarial training that is interesting in its own right, just like the TRADES method (Zhang et al., 2019). As mentioned before, the results achieved by this method (see Figures 2, 4, 6 and 7) are very similar to the results produced by adversarial training, even though we use the cross-entropy loss in these experiments. Extending our analysis to Madry's adversarial training seems challenging, but we expect the corresponding regularizer to have the similar form $\sum_{c=1}^{C} \int_\Omega \|\nabla_x \ell(u(x), c)\|_\star \mu_c(x)\, \mathrm{d}x$ (cf. Siethoff, 2025; Finlay & Oberman, 2021).

4 RELATED WORK

Adversarial Robustness and Interpretability A growing body of literature highlights the connection between a model's robustness and the interpretability of its saliency maps. Early observations by Tsipras et al. (2019) noted a tradeoff between adversarial and standard accuracy, while also revealing that the saliency maps of adversarially trained models (especially with ℓ_∞-restricted attacks) align well with perceptually relevant features. This was further supported by Zhang & Zhu (2019), who empirically show that adversarially trained CNNs produce sparser saliency maps that focus more strongly on object shapes and contours. Theoretical justifications have been proposed by Etmann et al. (2019); Chalasani et al. (2020) for linear classifiers. More recently, analyzing linear classifiers as well, Bhusal et al. (2026) observed that while adversarial training enhances the sparsity and input-stability of saliency maps it can compromise output-side stability.

Gradient Regularization Ross & Doshi-Velez (2018) demonstrated that direct input-gradient regularization not only improves adversarial robustness but also naturally improves model interpretability. Building on this, Gong et al. (2024) proposed replacing fixed-budget adversarial training with a penalty-based approach, utilizing a regularizer to induce desired properties in the gradient with respect to the input. The TRADES method (Zhang et al., 2019) involves regularization with a nonlocal loss-gradient as modification of adversarial training (Madry et al., 2018).

Theoretical Foundations Bungert et al. (2023) formalized the equivalence of adversarial training to regularization with the nonlocal total variation functional TV_ε (see (5)) for the ℓ_1-loss and gave an existence proof. Subsequently, Bungert & Stinson (2024) established Gamma-convergence of this functional. Siethoff (2025) proved existence of solutions for (5) with general loss functions. Furthermore, Yeh et al. (2019) and García Trillos & García Trillos (2022) both provide an upper bound for the adversarial risk in terms of a regularization involving the dual[2] norm of the gradient with respect to the input. Finally, Park et al. (2023) leveraged Gamma-convergence to analyze the training of two-layer neural networks with ReLU activation with weight decay, operating in a similar setting to ours.

5 SKETCH OF PROOF

In this section we provide an overview of the proof of Theorem 1. We start by defining

$$F_n(\theta) \coloneqq \frac{\widehat{\mathcal{R}}_n(f_\theta) - \inf_{\mathcal{B}} \mathcal{R}}{\varepsilon_n} + \widehat{\mathrm{TV}}_{\varepsilon_n, n}(f_\theta; \mu) + \frac{\lambda_n}{\varepsilon_n} R_{WD}(\theta)\,,$$

and observe that the minimizers of this functional coincide with the solutions to (9). We divided the objective in (9) by ε_n for otherwise one would only be able to prove that its minimizers converge to Bayes classifiers, without recovering the selection principle encoded in the functional F from (10).

To prove Theorem 1, we will show that these functionals Gamma-converge (Braides, 2002) to the functional F defined in (11) (see Theorem 2 for the precise statement). Gamma-convergence is a notion of convergence that together with a corresponding compactness result—that we provide in Proposition 5.3 for our case—implies that minimizers of the functionals F_n converge to minimizers of the limit functional F. Formally, establishing Gamma-convergence requires two complementary conditions: the $\liminf$ inequality and the $\limsup$ inequality.

The $\liminf$ implies that, if we have a sequence of minimizers of F_n that is to known to converge to some limit, this limit has a sufficiently small value of F.

[2] dual with respect to the norm used to constrain the adversarial perturbations

Proposition 5.1. *Let Assumptions 1 and 2 be satisfied. For $n \in \mathbb{N}$ let $\theta_n \in \Theta_{m_n}$ such that $f_n := f_{\theta_n}$ converges to a function $f \in L^1(\Omega)$ in $L^1(\Omega)$. Moreover, let $(\varepsilon_n)_{n\in\mathbb{N}}$ and $(\lambda_n)_{n\in\mathbb{N}}$ be sequences of positive numbers converging to zero satisfying*

$$\limsup_{n\to\infty} \frac{1}{\varepsilon_n}\sqrt{\frac{\log(n)}{n}} < \infty, \qquad \lim_{n\to\infty} \frac{1}{\lambda_n\sqrt{n}} = 0, \qquad \alpha := \lim_{n\to\infty} \frac{\lambda_n}{\varepsilon_n} \in (0,\infty).$$

Then it holds $F(f) \le \liminf_{n\to\infty} F_n(\theta_n)$ almost surely (with respect to the drawing of samples).

The proof of this result, which we carry out in Appendix E, combines Rademacher-based generalization bounds for the empirical risk established in Appendix B with a corresponding inequality for the regularizing term (empirical nonlocal total variation and weight decay) which we prove in Appendix D. This analysis is achieved by combining the properties of the population level functional TV_ε (Appendix C) with uniform concentration bounds for its empirical version (Appendix B).

The $\limsup$ inequality complements the $\liminf$ inequality by constructing a sequence that achieves the bound established there. This will be used to show that the limit of a sequence of minimizers of F_n is actually a minimizer of F.

Proposition 5.2. *Let Assumption 2 be satisfied. Let $f \in L^1(\Omega)$, $(\varepsilon_n)_{n\in\mathbb{N}}$ and $(\lambda_n)_{n\in\mathbb{N}}$ be sequences of positive numbers converging to zero satisfying*

$$\lim_{n\to\infty} \frac{1}{\varepsilon_n}\sqrt{\frac{\log(n)}{n}} = 0, \qquad \lim_{n\to\infty} \frac{1}{\varepsilon_n\sqrt{m_n}} = 0, \qquad \alpha := \lim_{n\to\infty} \frac{\lambda_n}{\varepsilon_n} \in (0,\infty).$$

Then, there exists a sequence (θ_n) with $\theta_n \in \Theta_{m_n}$ such that the sequence $f_n := f_{\theta_n}$ converges to f in $L^1(\Omega)$ and $\limsup_{n\to\infty} F_n(\theta_n) \le F(f)$ almost surely (with respect to the drawing of samples). Such a sequence (f_n) is called recovery sequence for f.

As for the $\liminf$ inequality, we combine generalization bounds and the results for the population level nonlocal total variation to prove this result (see Appendix E). Additionally, since $F(f) = \infty$ otherwise, we can assume that f is a Barron function, and use this to obtain a sequence of neural networks approximating f with dimension-independent rates (Lemma A.3).

Finally, as already mentioned, to enable us to extract a convergent subsequence from a sequence of minimizers of F_n, we require a compactness result.

Proposition 5.3. *Let Assumption 1 be satisfied and assume that $\alpha := \lim_{n\to\infty} \lambda_n/\varepsilon_n \in (0,\infty)$. For $n \in \mathbb{N}$ let $\theta_n \in \Theta_{m_n}$ such that $\limsup_{n\to\infty} F_n(\theta_n) < \infty$. Then, almost surely (with respect to the samples) there exists a subsequence θ_{n_j} and a function $f \in \mathcal{B}$ such that $f_{\theta_{n_j}} \to f$ in $L^1(\Omega)$.*

With these ingredients established, the proof of Theorem 1 is a standard argument from the theory of Gamma-convergence. For the sake of completeness, we carry out the explicit steps below.

Proof of Theorem 1. For $n \in \mathbb{N}$, let θ_n be a solution to (9) and denote the induced neural network by $f_n = f_{\theta_n}$. As we have mentioned, this implies that θ_n minimizes F_n. By $f^\star$ we denote an approximate minimizer of F, i.e. $f^\star \in \mathcal{B}$ with $\mathcal{R}(f^\star) = \inf_{\mathcal{B}} \mathcal{R}$ such that

$$\mathrm{TV}(f^\star;\mu) + \alpha[f^\star]_{\mathcal{B}} \le \inf F + \delta$$

for $\delta > 0$. Moreover, let $(f_n^\star)$ be a recovery sequence for $f^\star$ according to Proposition 5.2, i.e.,

$$\alpha[f^\star]_{\mathcal{B}} + \mathrm{TV}(f^\star;\mu) = F(f^\star) \ge \limsup_{n\to\infty} F_n(f_n^\star).$$

almost surely. Since f_n minimizes the functional F_n, this implies

$$\limsup_{n\to\infty} F_n(f_n) \le \limsup_{n\to\infty} F_n(f_n^\star) \le \mathrm{TV}(f^\star;\mu) + \alpha[f^\star]_{\mathcal{B}} < \infty.$$

Therefore, we can apply Proposition 5.3 and obtain the existence of a function $f \in \mathcal{B}$ and a subsequence (f_{n_j}) such that $f_{n_j} \to f$ in $L^1(\Omega)$. We can conclude

$$F(f) \le \liminf_{j\to\infty} F_{n_j}(f_{n_j}) \le \limsup_{j\to\infty} F_{n_j}(f_{n_j}) \le \mathrm{TV}(f^\star;\mu) + \alpha[f^\star]_{\mathcal{B}} \le \inf F + \delta$$

almost surely from Proposition 5.1. Since $\delta > 0$ is arbitrary, this implies that f minimizes F. □

6 Numerical Illustrations

To empirically validate our findings beyond simple two-layer ReLU networks, we compare ResNet18 models trained on clean CIFAR-10 images to those that were adversarially trained on the CIFAR-10 dataset (Krizhevsky & Hinton, 2009), using different adversarial budgets. We report our results in the top row of Figure 3, plotting the mean over 5 random seeds as well extremal values. From left to right, we plot the accuracy achieved on clean images, the robust accuracy for images attacked with Projected Gradient Descent (PGD) (Madry et al., 2018) with 50 steps and a fixed adversarial budget of $8/255$, the ℓ_1-norms of the gradient averaged over the test set, and the average sparsity of these gradients thresholded at different levels. Here, sparsity refers to the ratio of components in the gradient whose magnitude lies below the threshold.

We observe the usual trade-off between accuracy and robustness. Interestingly, even though training with a small adversarial budget leads to comparable accuracy as standard training, these models exhibit a higher sparsity of their input-gradients and also a significantly lower ℓ_1-norm, in line with our theoretical results. The confidence intervals are hardly visible, indicating high robustness against the stochasticity of neural network training.

In the bottom row of Figure 3, we plot the same metrics for the ResNet50 models[3] of Salman et al. (2020), pretrained on the ImageNet dataset (Deng et al., 2009). Our previous observations remain true, indicating that our results scale to large datasets. See Appendix F for more experimental results.

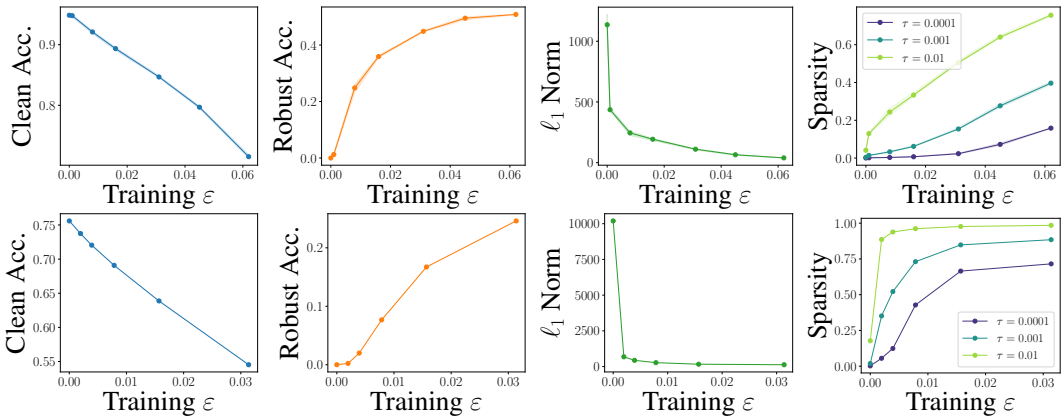


Figure 3: Accuracy and gradient metrics for ResNet18 on CIFAR-10 (top) and ResNet50 on ImageNet (bottom). From left to right, we display clean accuracy, robust accuracy, the average ℓ_1-norm, and the average sparsity (thresholded at different values) of the input gradients.

7 Conclusion

We provided a theoretical explanation for the empirical observation that gradient saliency maps of adversarially-trained neural networks tend to be sparser and more perceptually aligned than those of naturally-trained ones. For this we analyzed a total variation and weight decay regularized empirical risk minimization problem over the class of two-layer neural networks with ReLU activation, and proved that it selects a subclass of Bayes classifiers, consisting of those which minimize a balance of their average gradient ℓ_1-norm and their Barron norm. Hence, classifiers with axis-aligned gradients are favored which explains the sparsity of their gradient maps.

Our proofs involve generalization bounds, Monte Carlo approximation arguments, and Gamma-convergence techniques, and our numerical experiments confirm the theoretical results and indicate their generalizability. Our current analysis requires the weight decay regularization to be equally strong as the adversarial one and does not cover the original form of adversarial training, proposed by Madry et al. (2018). Future work includes investigating the connection between TV regularization and adversarial training theoretically and empirically as well as studying their influence on other saliency maps like integrated or smoothed gradients.

[3]Taken from `https://huggingface.co/madrylab/robust-imagenet-models`

AI USE STATEMENT

In this work we used generative AI tools for polishing the manuscript, discussing parts of the mathematical proofs and assisting with code to generate the numerical experiments. We have not used AI tools for generating the core research hypothesis or conducting literature reviews. We have reviewed all AI assisted work. Specifically, we verified all AI-discussed proof steps and tested the generated code for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Luigi Ambrosio, Nicola Fusco, and Diego Pallara. *Functions of bounded variation and free discontinuity problems*. Oxford university press, 2000.

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017.

Dipkamal Bhusal, Md Tanvirul Alam, and Nidhi Rastogi. Training for trustworthy saliency maps: Adversarial training meets feature-map smoothing, 2026.

Andrea Braides. *Gamma-convergence for Beginners*, volume 22 of *Oxford Lecture Series in Mathematics and their Applications*. Oxford University Press, 2002.

Leon Bungert and Kerrek Stinson. Gamma-convergence of a nonlocal perimeter arising in adversarial machine learning. *Calculus of Variations and Partial Differential Equations*, 63(5):114, 2024.

Leon Bungert, Nicolás García Trillos, and Ryan Murray. The geometry of adversarial training in binary classification. *Information and Inference: A Journal of the IMA*, 12(2):921–968, 2023.

Prasad Chalasani, Jiefeng Chen, Amrita Roy Chowdhury, Xi Wu, and Somesh Jha. Concise explanations of neural networks using adversarial training. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1383–1391. PMLR, 13–18 Jul 2020.

Antonin Chambolle, Stefano Lisini, and Luca Lussardi. A remark on the anisotropic outer minkowski content. *arXiv preprint arXiv:1203.5190*, 2012.

Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. IEEE, 2009.

Weinan E, Chao Ma, and Lei Wu. The Barron Space and the Flow-Induced Function Spaces for Neural Network Models. *Constructive Approximation*, 55(1):369–406, 2022. doi: 10.1007/s00365-021-09549-y.

Christian Etmann, Sebastian Lunz, Peter Maass, and Carola Schoenlieb. On the connection between adversarial robustness and saliency map interpretability. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1823–1832. PMLR, 09–15 Jun 2019.

Lawrence Craig Evans and Ronald F Gariepy. *Measure theory and fine properties of functions*, volume 5. CRC press, 1991.

Chris Finlay and Adam M Oberman. Scaleable input gradient regularization for adversarial robustness. *Machine Learning with Applications*, 3:100017, 2021.

Camilo Andrés García Trillos and Nicolás García Trillos. On the regularized risk of distributionally robust learning over deep neural networks. *Research in the Mathematical Sciences*, 9(3):54, 2022.

Nicolás García Trillos, Matt Jacobs, and Jakwang Kim. On the existence of solutions to adversarial training in multiclass classification. *European Journal of Applied Mathematics*, 37(1):3–23, 2026.

Shizhan Gong, Qi Dou, and Farzan Farnia. Structured gradient-based interpretations via norm-regularized adversarial training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11009–11018, June 2024.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4 (2):251–257, 1991. ISSN 0893-6080. doi: https://doi.org/10.1016/0893-6080(91)90009-T.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy*, 23(1), 2020.

Chao Ma, Stephan Wojtowytsch, Lei Wu, et al. Towards a mathematical understanding of neural network-based machine learning: what we know and what we don't. *arXiv preprint arXiv:2009.10713*, 2020.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

Rachel Morris and Ryan Murray. Uniform convergence of adversarially robust classifiers. *European Journal of Applied Mathematics*, 37(1):43–72, 2026.

Greg Ongie, Rebecca Willett, Daniel Soudry, and Nathan Srebro. A function space view of bounded norm infinite width relu nets: The multivariate case. In *International Conference on Learning Representations*, 2020.

Jiyoung Park, Ian Pelakh, and Stephan Wojtowytsch. Minimum norm interpolation by perceptra: Explicit regularization and implicit bias. *Advances in Neural Information Processing Systems*, 36:8714–8756, 2023.

Andrew Ross and Finale Doshi-Velez. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust imagenet models transfer better? In *Advances in Neural Information Processing Systems*, volume 33, pp. 3533–3545, 2020.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, 2017.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning Important Features Through Propagating Activation Differences. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3145–3153. PMLR, 06–11 Aug 2017.

Lennart Siethoff. Existence of solutions and asymptotic behavior of adversarial training with general loss functions. Master's thesis, University of Würzburg, 2025.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328. PMLR, 06–11 Aug 2017.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019.

Sophie Xhonneux, Alessandro Sordoni, Stephan Günnemann, Gauthier Gidel, and Leo Schwinn. Efficient adversarial training in llms with continuous attacks. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 1502–1530. Curran Associates, Inc., 2024. doi: 10.52202/079017-0048.

Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I. Inouye, and Pradeep K. Ravikumar. On the (In)fidelity and Sensitivity of Explanations. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pp. 7472–7482. PMLR, 2019.

Tianyuan Zhang and Zhanxing Zhu. Interpreting adversarially trained convolutional neural networks. In *International conference on machine learning*, pp. 7502–7511. PMLR, 2019.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 2005.

SUPPLEMENTARY MATERIAL

The supplementary material for our paper is organized into the following appendices. In particular, for readers familiar with Gamma-convergence we would like to point our Theorem 2 in Appendix E which, together with Proposition 5.3, implies our main result Theorem 1. However, for convenience, the proof of Theorem 1 given in Section 5 of the main paper is a direct one.

A The Homogeneous Barron Space **13**

B Generalization Bounds **19**

C Results for the Nonlocal Perimeter and Total Variation **21**

D Results for the Full Regularizer **26**

E Proofs of Propositions 5.1 to 5.3 **28**

F Additional Numerical Illustrations **30**

A THE HOMOGENEOUS BARRON SPACE

Recall our definition of the homogeneous Barron space $\mathcal{B}(\Omega)$ as all functions of the form

$$f_{\pi,y}(x) \coloneqq y + \int_{\mathbb{R}\times\mathbb{R}^d\times\mathbb{R}} a\big(\sigma(w\cdot x + b) - \sigma(b)\big)\,\mathrm{d}\pi(a,w,b), \qquad x \in \mathbb{R}^d,$$

such that $[f]_{\mathcal{B}(\Omega)} < \infty$, where $[f]_{\mathcal{B}(\Omega)} \coloneqq \inf_{f_{\pi,y}=f \text{ on } \Omega} R_{WD}(\pi)$ and

$$R_{WD}(\pi) \coloneqq \frac{1}{2}\int_{\mathbb{R}\times\mathbb{R}^d\times\mathbb{R}} \big(|a|^2 + \|w\|_2^2\big)\,\mathrm{d}\pi(a,w,b).$$

Furthermore, we recall that in our setting $\Omega \subset \mathbb{R}^d$ is assumed to be a bounded domain with Lipschitz boundary and that μ is a probability measure on $\Omega \times \{0,1\}$ describing the distribution of the training data. Note that the slightly non-standard representation of $f_{\pi,y}$ (see also (2)) is used to make sure that $R_{WD}(\pi) < \infty$ implies that the integral in (2) exists, see Park et al. (2023) for details.

One can easily see that every function in this space is Lipschitz continuous and the Lipschitz constant is bounded by the Barron semi-norm.

Lemma A.1. *Let $f \in \mathcal{B}(\Omega)$. Then,*

$$|f(x) - f(\tilde{x})| \le [f]_{\mathcal{B}}\|x - \tilde{x}\|_2$$

for $x, \tilde{x} \in \Omega$, i.e. f is Lipschitz continuous.

Proof. Let $x, \tilde{x} \in \Omega$ and let π, y such that $f = f_{\pi,y}$. Then, since the ReLU activation function is 1-Lipschitz,

$$\begin{aligned}
|f(x) - f(\tilde{x})| &= \left|\int a(\sigma(w\cdot x + b) - \sigma(w\cdot\tilde{x} + b))\,\mathrm{d}\pi(a,w,b)\right| \\
&\le \int |a||w\cdot x - w\cdot\tilde{x}|\,\mathrm{d}\pi(a,w,b) \\
&\le \int |a|\|w\|_2\,\mathrm{d}\pi(a,w,b)\|x - \tilde{x}\|_2 \\
&\le \|x - \tilde{x}\|_2 R_{WD}(\pi).
\end{aligned}$$

Taking the infimum over all representations $f_{\pi,y}$ yields the desired estimate. □

Since Ω is assumed to be bounded, the Lipschitz continuity of Barron functions implies the inclusion $\mathcal{B}(\Omega) \subset L^\infty(\Omega) \subset L^1(\Omega)$.

We note that the above proof holds for any Lipschitz continuous activation function σ with the only modification being that the Lipschitz constant becomes $\mathrm{Lip}(\sigma)[f]_{\mathcal{B}}$. In contrast, our proof of the following result–which established lower semi-continuity of the Barron semi-norm with respect to L^1 convergence–relies on the specific structure of the ReLU function.

Lemma A.2. *Let $(f_n) \subset \mathcal{B}(\Omega)$ and $f \in L^1(\Omega)$ such that $f_n \to f$ in $L^1(\Omega)$. Then,*

$$[f]_{\mathcal{B}} \le \liminf_{n\to\infty} [f_n]_{\mathcal{B}}.$$

Proof. By passing to a subsequence, we can assume that the liminf is a limit. Moreover, if this limit is infinite, the claimed inequality is trivially satisfied. Hence, we consider the case

$$\lim_{n\to\infty} [f_n]_{\mathcal{B}} =: M < \infty.$$

By definition of the Barron-semi norm, for any $n \in \mathbb{N}$, we can find a probability measure π_n over $\mathbb{R}\times\mathbb{R}^d\times\mathbb{R}$ and a real number y_n such that $f_n = f_{\pi_n,y_n}$ and $R_{WD}(\pi_n) < [f_n]_{\mathcal{B}} + \frac{1}{n}$. In particular, $R_{WD}(\pi_n)$ is uniformly bounded by some constant K. Moreover, defining the signed measure $\tilde{\pi}_n$ by $\mathrm{d}\tilde{\pi}_n(a,w,b) = a\|w\|_2\,\mathrm{d}\pi_n(a,w,b)$ leads to

$$\|\tilde{\pi}_n\|_{\mathrm{TV}} = \int_{\mathbb{R}^{d+2}} |a|\|w\|_2\,\mathrm{d}\pi_n(a,w,b) \le R_{WD}(\pi_n) \le K,$$

and using the fact that the ReLU functions is positively 1-homogeneous, we can compute

$$\begin{aligned} f_n(x) &= y_n + \int_{\mathbb{R}^{d+2}} a(\sigma(w \cdot x + b) - \sigma(b)) \, \mathrm{d}\pi_n(a, w, b) \\ &= y_n + \int_{\mathbb{R}^{d+2}} a\|w\|_2 \left(\sigma\left(\frac{w}{\|w\|_2} \cdot x + \frac{b}{\|w\|_2} \right) - \sigma\left(\frac{b}{\|w\|_2} \right) \right) \mathrm{d}\pi_n(a, w, b) \\ &= y_n + \int_{\mathbb{R}^{d+2}} \left(\sigma\left(\frac{w}{\|w\|_2} \cdot x + \frac{b}{\|w\|_2} \right) - \sigma\left(\frac{b}{\|w\|_2} \right) \right) \mathrm{d}\tilde{\pi}_n(a, w, b) \\ &= y_n + \int_{\mathbb{R}^{d+1}} (\sigma(\hat{w} \cdot x + \hat{b}) - \sigma(\hat{b})) \, \mathrm{d}(T_\sharp \tilde{\pi}_n)(\hat{w}, \hat{b}), \end{aligned}$$

where $T(a, w, b) \coloneqq (\frac{w}{\|w\|_2}, \frac{b}{\|w\|_2})$ for $w \neq 0$ and $T(a, w, b) = 0$ otherwise. Note that the signed measure $\nu_n \coloneqq T_\sharp \tilde{\pi}_n$ is supported on $\mathbb{S}^{d-1} \times \mathbb{R}$ and $\|\nu_n\|_{\mathrm{TV}} \leq \|\tilde{\pi}_n\|_{\mathrm{TV}} \leq K$. We additionally want to modify the measure ν_n such that it also has compact support in $\hat{b}$. To do so, we make use of the structure of the ReLU activation function and that the domain Ω is bounded. Let $R \coloneqq \sup_{x \in \Omega} \|x\|_2$ and define $\phi_x(w, b) \coloneqq \sigma(w \cdot x + b) - \sigma(b)$. We observe that for $|b| > R$, $w \in \mathbb{S}^{d-1}$ and $x \in \Omega$, we have $\phi_x(w, b) = \phi_x(w, \operatorname{sign}(b)R)$. Consequently, defining

$$P(w, b) \coloneqq \begin{cases} (w, R), & b > R, \\ (w, b), & -R \leq b \leq R, \\ (w, -R), & b < -R \end{cases}$$

and $\mu_n \coloneqq P_\sharp \nu_n$, we obtain

$$f_n(x) = y_n + \int_{\mathbb{S}^{d-1} \times [-R, R]} (\sigma(\hat{w} \cdot x + \hat{b}) - \sigma(\hat{b})) \, \mathrm{d}\mu_n(\hat{w}, \hat{b})$$

and $\|\mu_n\|_{\mathrm{TV}} \leq \|\nu_n\|_{\mathrm{TV}} \leq \mathbb{R}_{WD}(\pi_n) \leq K$. Note that we only need to integrate over $\mathbb{S}^{d-1} \times [-R, R]$ here, since the support of μ_n is contained in this compact set. Hence, by Ambrosio et al. (2000, Theorem 1.59), there exists a subsequence, that we again denote by (μ_n) that weakly* converges to a signed measure μ and

$$\|\mu\|_{\mathrm{TV}} \leq \liminf_{n \to \infty} \|\mu_n\|_{\mathrm{TV}} \leq \liminf_{n \to \infty} R_{WD}(\pi_n) = \lim_{n \to \infty} [f_n]_{\mathcal{B}} = M.$$

We can also show that the sequence of numbers (y_n) is bounded and thus admits a converging subsequence, since

$$\begin{aligned} |y_n| \mathcal{L}^d(\Omega) &= \int_\Omega |y_n| \, \mathrm{d}x \leq \int_\Omega |f_n(x)| \, \mathrm{d}x + \int_\Omega \int_{\mathbb{R}^{d+2}} |a| \|w\|_2 \|x\|_2 \, \mathrm{d}\pi_n(a, w, b) \, \mathrm{d}x \\ &\leq \|f_n\|_{L^1(\Omega)} + R\mathcal{L}^d(\Omega) K, \end{aligned}$$

where $\mathcal{L}^d$ denotes the d-dimensional Lebesgue measure. The right-hand side is uniformly bounded since (f_n) converges in $L^1(\Omega)$. Therefore in the following, we assume that y_n converges to some $y \in \mathbb{R}$. Using the fact that ϕ_x is a test function for the weak* convergence of μ_n (ϕ_x is continuous and $\mathbb{S}^{d-1} \times [-R, R]$ is compact), we observe

$$\begin{aligned} \lim_{n \to \infty} f_n(x) &= \lim_{n \to \infty} \left(y_n + \int_{\mathbb{S}^{d-1} \times [-R, R]} (\sigma(\hat{w} \cdot x + \hat{b}) - \sigma(\hat{b})) \, \mathrm{d}\mu_n(\hat{w}, \hat{b}) \right) \\ &= y + \int_{\mathbb{S}^{d-1} \times [-R, R]} (\sigma(\hat{w} \cdot x + \hat{b}) - \sigma(\hat{b})) \, \mathrm{d}\mu(\hat{w}, \hat{b}) \end{aligned}$$

and since $f_n \to f$ in $L^1(\Omega)$, the right hand side must coincide with f almost everywhere on Ω. It remains to rewrite the latter integral as an integral with respect to some probability distribution π over $\mathbb{R} \times \mathbb{R}^d \times \mathbb{R}$. To do so, let $\mu = \mu^+ - \mu^-$ denote the Hahn-Jordan decomposition of μ and let $C \coloneqq \|\mu\|_{\mathrm{TV}} = \mu^+(\mathbb{R}^{d+1}) + \mu^-(\mathbb{R}^{d+1})$. Without loss of generality, we can assume that $C > 0$, since in the case $C = 0$, we have $\mu = 0$ and can define $\pi \coloneqq \delta_0$ with $R_{WD}(\delta_0) = 0$. For $C > 0$,

we define $S^{\pm}(\hat{w}, \hat{b}) \coloneqq (\pm\sqrt{C}, \sqrt{C}\hat{w}, \sqrt{C}\hat{b})$. Then, $\pi \coloneqq \frac{1}{C}(S^+_\sharp \mu^+ + S^-_\sharp \mu^-)$ defines a probability measure on $\mathbb{R} \times \mathbb{R}^d \times \mathbb{R}$ and

$$\begin{aligned}
&\int_{\mathbb{R}^{d+2}} a\phi_x(w,b)\,\mathrm{d}\pi(a,w,b)\\
&= \frac{1}{C}\int_{\mathbb{R}^{d+2}} a\phi_x(w,b)\,\mathrm{d}(S^+_\sharp\mu^+)(a,w,b) - \frac{1}{C}\int_{\mathbb{R}^{d+2}} a\phi_x(w,b)\,\mathrm{d}(S^-_\sharp\mu^-)(a,w,b)\\
&= \frac{1}{C}\int_{\mathbb{R}^{d+1}} \sqrt{C}\phi_x(\sqrt{C}\hat{w},\sqrt{C}\hat{b})\,\mathrm{d}\mu^+(\hat{w},\hat{b}) - \frac{1}{C}\int_{\mathbb{R}^{d+1}} \sqrt{C}\phi_x(\sqrt{C}\hat{w},\sqrt{C}\hat{b})\,\mathrm{d}\mu^-(\hat{w},\hat{b})\\
&= \int_{\mathbb{R}^{d+1}} \phi_x(\hat{w},\hat{b})\,\mathrm{d}(\mu^+ - \mu^-)(\hat{w},\hat{b})\\
&= \int_{\mathbb{R}^{d+1}} \phi_x(\hat{w},\hat{b})\,\mathrm{d}\mu(\hat{w},\hat{b}).
\end{aligned}$$

due to the positive 1-homogeneity of σ. Hence, we can conclude $f = f_{\pi,y}$ and $[f]_{\mathcal{B}} \leq R_{WD}(\pi) = \|\mu\|_{\mathrm{TV}} \leq \liminf_{n\to\infty}[f_n]_{\mathcal{B}}$. □

A fundamental property of Barron functions is that they can be approximated by neural networks with dimension-independent rates. For our application it is particularly useful to establish approximation in the L^∞-norm.

Lemma A.3. *Let $f \in \mathcal{B}$. Then, for every $m \in \mathbb{N}$, there exists a neural network f_m of the form* (1) *parametrized by θ_m such that*

$$\|f - f_m\|_{L^\infty(\Omega)} \leq \frac{C[f]_{\mathcal{B}}}{\sqrt{m}}$$

with some positive constant C and

$$R_{WD}(\theta_m) \leq [f]_{\mathcal{B}}.$$

The proof of this statement goes along the lines of Ma et al. (2020, Theorem 12), who consider the case $\Omega = [0,1]^d$ and non-homogeneous Barron spaces. For the sake of completeness, we adapt their proof to our setting.

Proof. Let π be a probability measure on $\mathbb{R}\times\mathbb{R}^d\times\mathbb{R}$ and $y \in \mathbb{R}$ such that $f = f_{y,\pi}$ and $R_{WD}(\pi) \leq [f]_{\mathcal{B}} + \delta$ for arbitrary positive δ. We want to normalize π in a way such that

$$|a| = \|w\|_2 \leq R_{WD}(\pi)$$

π-almost surely, which will happen in a two step procedure. First, we make sure that the product $|a|\|w\|_2$ is constant π-almost surely. To do so, let $C \coloneqq \int_{\mathbb{R}^{d+2}} |a|\|w\|_2\,\mathrm{d}\pi(a,w,b)$ and define the probability measure $\tilde{\pi}$ via $\mathrm{d}\tilde{\pi} = \frac{|a|\|w\|_2}{C}\,\mathrm{d}\pi$. We may assume without loss of generality that $C > 0$. The case $C = 0$ implies that f is a constant function, which can trivially be parametrized by a neural network of the form (1). Defining $S(a,w,b) \coloneqq (\frac{aC}{|a|\|w\|_2}, w, b)$, and considering the push-forward $\bar{\pi} = S_\sharp\tilde{\pi}$, we observe that for $h_{(a,w,b)}(x) \coloneqq a(\sigma(w\cdot x + b) - \sigma(b))$, we have

$$\begin{aligned}
\int_{\mathbb{R}^{d+2}} h_{(a,w,b)}(x)\,\mathrm{d}\bar{\pi}(a,w,b) &= \int_{\mathbb{R}^{d+2}} h_{S(a,w,b)}(x)\,\mathrm{d}\tilde{\pi}(a,w,b)\\
&= \int_{\mathbb{R}^{d+2}} \frac{aC}{|a|\|w\|_2}(\sigma(w\cdot x+b)-\sigma(b))\frac{|a|\|w\|_2}{C}\,\mathrm{d}\pi(a,w,b)\\
&= f(x) - f(0).
\end{aligned}$$

Hence, $\bar{\pi}$ represents f and we have $|a|\|w\|_2 = C$ $\bar{\pi}$-almost surely. Defining

$$T(a,w,b) \coloneqq \left(a\sqrt{\frac{\|w\|_2}{|a|}}, w\sqrt{\frac{|a|}{\|w\|_2}}, b\sqrt{\frac{|a|}{\|w\|_2}}\right),$$

one can show using the 1-homogeneity of the ReLU activation function that $\hat{\pi} \coloneqq T_\sharp\bar{\pi}$ also represents f and additionally satisfies that $|a|^2 = \|w\|_2^2 = C \leq R_{WD}(\pi)$ almost surely. As in the proof

of Lemma A.2, we can additionally construct the measure $\hat{\pi}$ such that its support w.r.t. the last variable is contained in some interval $[-K, K]$, where K depends on the domain Ω and on $[f]_{\mathcal{B}}$. For simplicity, we will from now on refer to $\hat{\pi}$ as π.

Using Shalev-Shwartz & Ben-David (2014, Lemma 26.2) yields

$$\mathbb{E}_{(a,w,b)\sim\pi^m}\left[\sup_{x\in\Omega}\left(f(x)-y-\frac{1}{m}\sum_{i=1}^{m}a_i(\sigma(w_i\cdot x+b_i)-\sigma(b_i)\right)\right]$$
$$\leq 2\mathbb{E}_{(a,w,b)\sim\pi^m}\mathbb{E}_{\xi}\left[\sup_{x\in\Omega}\frac{1}{m}\sum_{i=1}^{m}\xi_i a_i(\sigma(w_i\cdot x+b_i)-\sigma(b_i))\right],$$

where $\xi\in\mathbb{R}^m$ denotes a random vector whose components are independent Rademacher random variables. Moreover,

$$\begin{aligned}\mathbb{E}_{\xi}\left[\sup_{x\in\Omega}\frac{1}{m}\sum_{i=1}^{m}\xi_i a_i\left(\sigma(w_i\cdot x+b_i)-\sigma(b_i)\right)\right] &\leq \mathbb{E}_{\xi}\left[\sup_{x\in\Omega}\frac{1}{m}\sum_{i=1}^{m}\xi_i|a_i||w_i\cdot x|\right]\\ &= \frac{1}{m}\mathbb{E}_{\xi}\left[\sup_{x\in\Omega}x\cdot\sum_{i=1}^{m}\xi_i a_i w_i\right]\\ &\leq \frac{R}{m}\mathbb{E}_{\xi}\left[\left\|\sum_{i=1}^{m}\xi_i a_i w_i\right\|_2\right],\end{aligned}$$

where $R\coloneqq\sup_{x\in\Omega}\|x\|_2$. Defining $z_i\coloneqq a_i w_i\in\mathbb{R}^d$ yields $\|z_i\|_2\leq|a_i|\|w_i\|_2\leq[f]_{\mathcal{B}}+\delta$. Hence, we can further estimate

$$\begin{aligned}\frac{R}{m}\mathbb{E}_{\xi}\left[\left\|\sum_{i=1}^{m}\xi_i a_i w_i\right\|_2\right] &\leq \frac{R}{m}\sup_{\|z_i\|_2\leq[f]_{\mathcal{B}}+\delta}\mathbb{E}_{\xi}\left[\left\|\sum_{i=1}^{m}\xi_i z_i\right\|_2\right]\\ &\leq \frac{R([f]_{\mathcal{B}}+\delta)}{m}\sup_{\|z_i\|_2\leq 1}\mathbb{E}_{\xi}\left[\left\|\sum_{i=1}^{m}\xi_i z_i\right\|_2\right]\\ &\leq \frac{R([f]_{\mathcal{B}}+\delta)}{\sqrt{m}}.\end{aligned}$$

In the last step, we used Jensen's inequality together with the fact that the ξ_i are independent with expected value zero to bound the expectation by $\sqrt{m}$. Overall, we obtain

$$\mathbb{E}_{(a,w,b)\sim\pi^m}\left[\sup_{x\in\Omega}\left(f(x)-y-\frac{1}{m}\sum_{i=1}^{m}a_i(\sigma(w_i\cdot x+b_i)-\sigma(b_i)\right)\right]\leq\frac{2R([f]_{\mathcal{B}}+\delta)}{\sqrt{m}}.$$

This bound for the expectation means that there must exist a choice of parameters (a_i, w_i, b_i) for $i=1,\ldots,m$ such that $|a_i|^2=\|w_i\|_2^2\leq[f]_{\mathcal{B}}+\delta$ such that

$$\left\|f-b_0-\frac{1}{m}\sum_{i=1}^{m}a_i\sigma(w_i\cdot x+b_i)\right\|_{L^\infty(\Omega)}\leq\frac{C([f]_{\mathcal{B}}+\delta)}{\sqrt{m}}$$

with $b_0=y+\sum_{i=1}^{m}\sigma(b_i)$. To keep notation aligned with (1), the parameters of the neural network are given by $(\frac{a_i}{\sqrt{m}},\frac{w_i}{\sqrt{m}},\frac{b_i}{\sqrt{m}})$ and the resulting parametrization θ_m satisfies

$$R_{WD}(\theta_m)=\frac{1}{2}\sum_{I=1}^{m}\left(\frac{|a_i|^2}{m}+\frac{\|w_i\|_2^2}{m}\right)\leq[f]_{\mathcal{B}}+\delta.$$

Since the parameters lie in a compact subset of $\mathbb{R}^{d+2}$, when letting $\delta\to 0$, we can extract a convergent subsequence. The limit satisfied the above inequalities without δ, since the map $\mathbb{R}^{d+2}\ni(a,w,b)\mapsto a\sigma(w^T\cdot+b)\in L^\infty(\Omega)$ is continuous. □

For proving generalization bounds (see Appendix B), we must investigate the Rademacher complexity of the homogeneous Barron space. Specifically, we consider the subset of Barron functions with a uniform bound for the Barron semi-norm and for their value at zero,

$$\mathcal{F}_{Q,B} \coloneqq \{f \in \mathcal{B} \,|\, [f]_{\mathcal{B}} \leq Q, |f(0)| \leq B\}. \tag{12}$$

Similar to Park et al. (2023, Corollary D.3), we obtain the following estimate for the Rademacher complexity of this set.

Lemma A.4. *Let $Q, B \geq 0$. Then, there exists a constant C depending on Ω such that for all probability measures ρ that are supported on Ω, we have*

$$\mathrm{Rad}_{\rho,n}(\mathcal{F}_{Q,B}) \leq \frac{CQ}{\sqrt{n}} + \frac{B}{\sqrt{n}}.$$

Proof. For a fixed sample $S = \{x_i, \dots, x_n\}$ and independent Rademacher random variables $\xi_1, \dots, \xi_n$, we have

$$\begin{aligned}
\mathrm{Rad}_S(\mathcal{F}_{Q,B}) &= \mathbb{E}_\xi \left[\sup_{\substack{f \in \mathcal{F}_{Q,0}, \\ c \in [-B,B]}} \frac{1}{n} \sum_{i=1}^n \xi_i (f(x_i) + c) \right] \\
&\leq \mathbb{E}_\xi \left[\sup_{f \in \mathcal{F}_{Q,0}} \frac{1}{n} \sum_{i=1}^n \xi_i f(x_i) \right] + \mathbb{E}_\xi \left[\sup_{c \in [-B,B]} \frac{1}{n} c \sum_{i=1}^n \xi_i \right] \\
&\leq \mathrm{Rad}_S(\mathcal{F}_{Q,0}) + \mathbb{E}_\xi \left[B \left| \frac{1}{n} \sum_{i=1}^n \xi_i \right| \right] \\
&\leq \mathrm{Rad}_S(\mathcal{F}_{Q,0}) + \frac{B}{\sqrt{n}}.
\end{aligned}$$

In the last step, we have used that by Jensen's inequality,

$$\mathbb{E}_\xi \left[\left| \sum_{i=1}^n \xi_i \right| \right] \leq \sqrt{\mathbb{E}_\xi \left[\left| \sum_{i=1}^n \xi_i \right| \right]^2} = \sqrt{\mathbb{E}_\xi \left[\sum_{i=1}^n \xi_i^2 \right]} = \sqrt{n}.$$

Hence, taking the expectation over the data samples, we obtain

$$\begin{aligned}
\mathrm{Rad}_{\rho,n}(\mathcal{F}_{Q,B}) &\leq \mathrm{Rad}_{\rho,n}(\mathcal{F}_{Q,0}) + \frac{B}{\sqrt{n}} \\
&\leq \frac{CQ}{\sqrt{n}} + \frac{B}{\sqrt{n}}.
\end{aligned}$$

In the last step, we used the bound $\mathrm{Rad}_{\rho,n}(\mathcal{F}_{Q,0}) \leq \frac{CQ}{\sqrt{n}}$ which holds for data distributions ρ supported on Ω, where the constant C depends only on Ω. This bound follows readily by adapting the proof of Park et al. (2023, Lemma D.2). □

For our theory it is also important to obtain generalization bounds for the nonlocal total variation TV_ε. Due to its structure, we estimate the Rademacher complexity of the following class of functions

$$\begin{aligned}
\mathcal{G}_{Q,\varepsilon} &\coloneqq \{\Omega \ni x \mapsto \sup_{B_\varepsilon(x) \cap \Omega} f - f(x) \,|\, f \in \mathcal{F}_{Q,\infty}\} \\
&= \{\Omega \ni x \mapsto \sup_{B_\varepsilon(x) \cap \Omega} f - f(x) \,|\, f \in \mathcal{F}_{Q,0}\}
\end{aligned} \tag{13}$$

for $Q, \varepsilon > 0$, where

$$\mathcal{F}_{Q,\infty} \coloneqq \{f \in \mathcal{B}(\Omega) \,|\, [f]_{\mathcal{B}(\Omega)} \leq Q\}.$$

Lemma A.5. *There exists a constant $C > 0$ depending on Ω, such that for all $Q, B > 0$ and $\varepsilon \in (0, 1]$ and probability measures ρ supported on Ω, we have*

$$\mathrm{Rad}_{\rho,n}(\mathcal{G}_{Q,\varepsilon}) \leq \frac{CQ}{\sqrt{n}}.$$

To prove this statement, we first reduce the problem to functions represented by one neuron and exploit the fact that the supremum over a given ball can be computed explicitly for functions of this form. From there, we proceed exactly as in the proof of Park et al. (2023, Lemma D.2).

Proof. First, we consider a fixed sample $S = \{x_1, \dots, x_n\} \subset \Omega$ and let $R > 0$ such that $\Omega \subset B_R^2(0)$. Obviously,

$$\mathrm{Rad}_{\rho,n}(\mathcal{G}_{Q,\varepsilon}) \leq \mathrm{Rad}_{\rho,n}(S_{Q,\varepsilon}) + \mathrm{Rad}_{\rho,n}(\mathcal{F}_{Q.0}),$$

where $S_{Q,\varepsilon} = \{\Omega \ni x \mapsto \sup_{B_\varepsilon(x)\cap\Omega} f \,|\, f \in \mathcal{F}_{Q,0}\}$. Lemma A.4 provides a bound for the latter term, so it remains to bound the former.

For a function $f \in \mathcal{F}_{Q,0}$ and $\delta > 0$, we can find a representing probability measure π such that almost surely, we have $|a|^2 = \|w\|_2^2 \leq Q + \delta$ and $|b| \leq \|w\|_2 R$ (see the previous proofs). Then,

$$\begin{aligned}\sup_{y\in B_\varepsilon(x)\cap\Omega} f(y) &= \sup_{y\in B_\varepsilon(x)\cap\Omega} \int a(\sigma(w\cdot y + b) - \sigma(b))\,\mathrm{d}\pi(a,w,b) \\ &\leq \int \sup_{y\in B_\varepsilon(x)\cap\Omega} a(\sigma(w\cdot y + b) - \sigma(b))\,\mathrm{d}\pi(a,w,b).\end{aligned}$$

Therefore, with σ_i for $i = 1, \dots, n$ independent Rademacher random variables, we obtain

$$\begin{aligned}\sum_{i=1}^n \sigma_i \sup_{B_\varepsilon(x_i)\cap\Omega} f &\leq \int \sum_{i=1}^n \sigma_i \sup_{y\in B_\varepsilon(x_i)\cap\Omega} a(\sigma(w\cdot y + b) - \sigma(b))\,\mathrm{d}\pi(a,w,b) \\ &\leq \sup_{\substack{|a|^2=\|w\|_2^2\leq Q+\delta \\ |b|\leq\|w\|_2 R}} \sum_{i=1}^n \sigma_i \sup_{y\in B_\varepsilon(x_i)\cap\Omega} a(\sigma(w\cdot y + b) - \sigma(b)).\end{aligned}$$

In particular, taking the supremum over $f \in \mathcal{F}_{Q,0}$ yields

$$\sup_{f\in\mathcal{F}_{Q,0}} \sum_{i=1}^n \sigma_i \sup_{B_\varepsilon(x_i)\cap\Omega} f \leq \sup_{\substack{|a|^2=\|w\|_2^2\leq Q+\delta \\ |b|\leq\|w\|_2 R}} \sum_{i=1}^n \sigma_i \sup_{y\in B_\varepsilon(x_i)\cap\Omega} a(\sigma(w\cdot y + b) - \sigma(b)).$$

Hence, in the following we can restrict ourselves to functions of the form $x \mapsto a(\sigma(w\cdot x+b)-\sigma(b))$, where $|a|^2 = \|w\|_2^2 \leq Q + \delta$ and $|b| \leq \|w\|_2 R$. For such functions it is easy to see that due to the monotonicity of the ReLU function,

$$\sup_{y\in B_\varepsilon(x_i)\cap\Omega} a(\sigma(w\cdot y + b) - \sigma(b)) \leq a\sigma(w\cdot x_i \pm \varepsilon\|w\|_\star + b) - a\sigma(b),$$

where the sign in front of $\varepsilon\|w\|_\star$ is determined by the sign of a. Thus, using the bounds on the magnitudes of a and b, we obtain

$$\mathrm{Rad}_S(S_{Q,\varepsilon}) \leq \frac{1}{n}\mathbb{E}\left[\sup_{\substack{|a|^2=\|w\|_2^2\leq Q+\delta, \\ |b|\leq\|w\|_2 R}} \sum_{i=1}^n \sigma_i a\sigma\left(w\cdot x_i \pm \varepsilon\|w\|_\star + b\right)\right] + \frac{(Q+\delta)R}{\sqrt{n}}.$$

To bound the first term on the right hand side, we proceed exactly as in Step 3 of the proof of Park et al. (2023, Lemma D.2). First, we notice that the modified bias $b' \coloneqq b \pm \varepsilon\|w\|_\star$ satisfies $|b'| \leq \sqrt{Q+\delta}(R + c\varepsilon)$ for some constant c, since all norms on $\mathbb{R}^d$ are equivalent. Then, we define modified data points $\tilde{x}_i \coloneqq (x_i, R + c\varepsilon)$ and modified weights $\tilde{w} \coloneqq (w, \frac{b'}{R+c\varepsilon})$ which yields

$$\begin{aligned}&\tilde{w}\cdot\tilde{x}_i = w\cdot x_i + b', \\ &\|\tilde{x}_i\|_2 = \sqrt{\|x_i\|_2^2 + (R+c\varepsilon)^2} \leq \sqrt{2}(R + c\varepsilon), \\ &|a|^2 + \|\tilde{w}\|_2^2 \leq 2Q + \left(\frac{b'}{R+c\varepsilon}\right)^2 \leq 3(Q+\delta).\end{aligned}$$

Consequently,

$$\begin{aligned}
&\frac{1}{m}\mathbb{E}\left[\sup_{|a|^2=\|w\|_2^2\le Q+\delta,|b|\le\|w\|_2 R}\left|\sum_{i=1}^{n}\sigma_i a\sigma(w\cdot x_i+b')\right|\right]\\
&\quad\le\frac{1}{n}\mathbb{E}\left[\sup_{a^2+|\tilde{w}|^2\le 3(Q+\delta)}\left|\sum_{i=1}^{n}\sigma_i a\sigma(\tilde{w}\cdot\tilde{x}_i)\right|\right]\\
&\quad\le\frac{3(Q+\delta)}{n}\mathbb{E}\left[\sup_{|u|\le 1}\sum_{i=1}^{n}\sigma_i\sigma(u\cdot\tilde{x}_i)\right]\\
&\quad\le\frac{C(Q+\delta)}{\sqrt{n}}
\end{aligned}$$

for some constant $C>0$ depending on Ω. The last step follows from the Contraction lemma for Rademacher complexities (Shalev-Shwartz & Ben-David, 2014, Lemma 26.9) and a bound for the Rademacher complexity of linear functions (Shalev-Shwartz & Ben-David, 2014, Lemma 26.10). Since $\delta>0$ was arbitrary, the inequality also holds true without δ. Moreover, as the measure ρ is supported on Ω, we can take the expected value w.r.t. $S\sim\rho^n$ and obtain

$$\mathrm{Rad}_{\rho,n}(S_{Q,\varepsilon})\le\frac{CQ}{\sqrt{n}}.$$

□

Remark A.1. In context of TV_ε, we are rather interested in the Rademacher complexity of $\frac{1}{\varepsilon}\mathcal{G}_{Q,\varepsilon}$. Obviously, we immediately get

$$\mathrm{Rad}_{\rho,n}\left(\frac{1}{\varepsilon}\mathcal{G}_{Q,\varepsilon}\right)\le\frac{CQ}{\varepsilon\sqrt{n}}.$$

B Generalization Bounds

In this section, we prove generalization bounds for both the empirical risk (7) and the empirical nonlocal total variation (8). The results rely in particular on the Rademacher complexity bounds derived in the previous section. We start with a result on the empirical risk.

Lemma B.1. *Let $\ell\colon\mathbb{R}\times\mathbb{R}\to[0,\infty)$ be measurable, bounded and Lipschitz continuous in the first argument. Then, there exists a constant $C>0$ such that for all $Q,B\ge 0$ with probability at least $1-\delta$, we have*

$$\sup_{u\in\mathcal{F}_{Q,B}}\left(\mathcal{R}(u)-\widehat{\mathcal{R}}_n(u)\right)\le C\left(\frac{B+Q}{\sqrt{n}}+\sqrt{\frac{\log(2/\delta)}{n}}\right).$$

Proof. Since the loss function ℓ is bounded, measurable and Lipschitz continuous in the first argument, we have

$$\sup_{u\in\mathcal{F}_{Q,B}}\left(\mathcal{R}(u)-\widehat{\mathcal{R}}_n(u)\right)\le C\left(\mathrm{Rad}_{\rho_x,n}(\mathcal{F}_{Q,B})+\sqrt{\frac{\log(2/\delta)}{n}}\right)$$

with probability at least $1-\delta$ for some positive constant C depending on the loss function ℓ (see Shalev-Shwartz & Ben-David (2014, Theorem 26.5)). Combining this with the bound on the Rademacher complexity from Lemma A.4 yields the desired estimate. □

Now, we prove a similar result for the empirical nonlocal total variation $\widehat{\mathrm{TV}}_{\varepsilon,n}$. As observed when bounding the Rademacher complexity (Lemma A.5) the function's value at zero cancels out within the TV_ε functional. Consequently, we can establish the following result for the broader space $\mathcal{F}_{Q,\infty}$ rather than restricting it to $\mathcal{F}_{Q,B}$ as in the previous statement.

Lemma B.2. *Let $Q \geq 0$, $\delta \in (0,1)$ and $\varepsilon \in (0,1]$. Then, with probability at least $1-2\delta$, we have*

$$\sup_{f \in \mathcal{F}_{Q,\infty}} \left| \widehat{\mathrm{TV}}_{\varepsilon,n}(f;\mu) - \mathrm{TV}_{\varepsilon}(f;\mu) \right| \leq CQ \left(\frac{1}{\varepsilon\sqrt{n}} + \frac{\sqrt{\log(2/\delta)}}{\sqrt{2n}} \right)$$

for some constant $C > 0$.

Proof. We restrict our detailed analysis to the term corresponding to the label $y = 0$, as the argument for $y = 1$ is analogous. For a function $f \in \mathcal{F}_{Q,\infty}$, we define

$$h_f(x) \coloneqq \frac{1}{\varepsilon} \left(\sup_{B_\varepsilon(x) \cap \Omega} f - f(x) \right).$$

The quantity of interest is

$$\sup_{f \in \mathcal{F}_{Q,\infty}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(Y_i = 0) h_f(X_i) - \int_\Omega h_f(x) \mu_0(x) \, \mathrm{d}x \right|.$$

Letting $g_f(x,y) \coloneqq \mathbb{1}(y=0) h_f(x)$ and $\mathcal{G} \coloneqq \{g_f \mid f \in \mathcal{F}_{Q,\infty}\}$, we observe that

$$\mathbb{E}_{(X,y)\sim\mu}[g_f(X,Y)] = \int_\Omega h_f(x)\mu_0(x) \, \mathrm{d}x.$$

Defining the abbreviations $\widehat{\mathbb{E}}_{S_n}[g] \coloneqq \frac{1}{n}\sum_{i=1}^n g(X_i,Y_i)$ and $\mathbb{E}[g] \coloneqq \mathbb{E}_{(X,Y)\sim\mu}[g(X,Y)]$ for $g \in \mathcal{G}$, the target quantity can be written as $\sup_{g\in\mathcal{G}} |\widehat{\mathbb{E}}_{S_n}[g] - \mathbb{E}[g]|$. Since it is easier to handle this term without the absolute value, we additionally define the positive and negative deviations as

$$\Phi(S_n) \coloneqq \sup_{g\in\mathcal{G}} \left(\widehat{\mathbb{E}}_{S_n}[g] - \mathbb{E}[g] \right), \qquad \Psi(S_n) \coloneqq \sup_{g\in\mathcal{G}} \left(\mathbb{E}[g] - \widehat{\mathbb{E}}_{S_n}[g] \right).$$

By standard Rademacher symmetrization, the expected values with respect to $S_n \sim \mu^n$ of both $\Phi(S_n)$ and $\Psi(S_n)$ are upper bounded by $2\,\mathrm{Rad}_{n,\mu}(\mathcal{G})$. To relate this to the Rademacher complexity of $\mathcal{G}_{Q,\varepsilon}$ defined in (13), we condition on the labels Y and denote by N_0 the number of data points with label zero and $\tilde{X}_1, \ldots, \tilde{X}_{N_0}$ the corresponding x-components. Then,

$$\begin{aligned}
\mathrm{Rad}_{n,\mu}(\mathcal{G}) &= \mathbb{E}_{(X,Y)\sim\mu^n,\sigma} \left[\sup_{f\in\mathcal{F}_{Q,\infty}} \frac{1}{n} \sum_{i=1}^n \sigma_i \mathbb{1}(Y_i = 0) h_f(X_i) \right] \\
&= \mathbb{E}_Y \left[\frac{N_0}{n} \mathbb{E}_{X,\sigma|Y} \left[\sup_{f\in\mathcal{F}_{Q,\infty}} \frac{1}{N_0} \sum_{i=1}^{N_0} \sigma_i h_f(\tilde{X}_i) \right] \right] \\
&= \mathbb{E}_Y \left[\frac{N_0}{n} \frac{1}{\varepsilon} \mathrm{Rad}_{N_0, \frac{\mu_0}{\mu_0(\Omega)}}(\mathcal{G}_{Q,\varepsilon}) \right] \\
&\leq \frac{CQ}{n\varepsilon} \mathbb{E}_Y \left[\sqrt{N_0} \right].
\end{aligned}$$

We notice that $N_0 \sim \mathrm{Binomial}(n, \mu_0(\Omega))$ and consequently, using Jensen's inequality,

$$\mathbb{E}_Y[\sqrt{N_0}] \leq \sqrt{\mathbb{E}[N_0]} = \sqrt{n\mu_0(\Omega)}.$$

Overall, we arrive at

$$\mathbb{E}_{S_n}[\Phi(S_n)] \leq \frac{CQ}{\varepsilon\sqrt{n}}, \qquad \mathbb{E}[\Psi(S_n)] \leq \frac{CQ}{\varepsilon\sqrt{n}}.$$

Applying McDiarmid's inequality, which is possible since changing a single pair (X_i, Y_i) in the sample alters both $\Phi(S_n)$ and $\Psi(S_n)$ by at most $\frac{Q}{n}$ as all functions $f \in \mathcal{F}_{Q,\infty}$ are Q-Lipschitz, yields

$$\begin{aligned}
\mathbb{P}\left(\Phi(S_n) > \mathbb{E}[\Phi(S_n)] + t\right) &\leq \exp\left(-\frac{2nt^2}{Q^2}\right), \\
\mathbb{P}\left(\Psi(S_n) > \mathbb{E}[\Psi(S_n)] + t\right) &\leq \exp\left(-\frac{2nt^2}{Q^2}\right).
\end{aligned}$$

Taking a union bound and using the upper bounds for the expectations of $\Phi(S_n)$ and $\Psi(S_n)$, we obtain

$$\mathbb{P}\left(\sup_{g\in\mathcal{G}} |\widehat{\mathbb{E}}_{S_n}[g] - \mathbb{E}[g]| > \frac{CQ}{\varepsilon\sqrt{n}} + t\right) \leq 2\exp\left(-\frac{2nt^2}{Q^2}\right).$$

Setting the right hand side to equal δ leads to $t = \frac{Q\sqrt{\log(2/\delta)}}{\sqrt{2n}}$. Repeating the same argument for the μ_1-term concludes the proof. □

For our analysis, specifically for the $\liminf$ inequality for the regularizer in Proposition D.1, we must handle sequences of functions, where a uniform bound on the Barron semi-norm cannot be guaranteed. Instead, the Barron semi-norm may grow asymptotically like λ_n/ε_n where (ε_n) and (λ_n) are sequences of positive numbers converging to zero. Consequently, our next result establishes almost sure convergence of the empirical nonlocal total variation to its population version, under suitable scalings of these sequences.

Lemma B.3. *Let ε_n, λ_n be chosen such that*

$$\lim_{n\to\infty} \frac{\varepsilon_n\sqrt{\log(n)}}{\lambda_n\sqrt{n}} = 0, \qquad \lim_{n\to\infty} \frac{1}{\lambda_n\sqrt{n}} = 0.$$

Then, almost surely (w.r.t. the drawing of the samples) simultaneously for all sequences $(f_n) \subset \mathcal{B}(\Omega)$ that satisfy $\limsup_{n\to\infty} \frac{\lambda_n}{\varepsilon_n}[f_n]_{\mathcal{B}} < \infty$, we have

$$\lim_{n\to\infty} |\widehat{\mathrm{TV}}_{\varepsilon_n,n}(f_n;\mu) - \mathrm{TV}_{\varepsilon_n}(f_n;\mu)| = 0.$$

It is worth noticing that the first scaling condition for ε_n and λ_n is automatically satisfied, when $\lim_{n\to\infty} \frac{\lambda_n}{\varepsilon_n} \in (0,\infty]$, which is implied by our assumption on α in Theorem 1.

Proof. For an arbitrary but fixed $m \in \mathbb{N}$, we consider the sequence $Q_n^m = \frac{m\varepsilon_n}{\lambda_n}$. Setting $\delta = \frac{1}{n^2}$ in Lemma B.2 yields that with probability at least $1 - \frac{2}{n^2}$, we have

$$\sup_{f\in\mathcal{F}_{Q_n^m,\infty}} \left|\widehat{\mathrm{TV}}_{\varepsilon_n,n}(f;\mu) - \mathrm{TV}_{\varepsilon_n}(f;\mu)\right| \leq \frac{Cm\varepsilon_n}{\lambda_n}\left(\frac{1}{\varepsilon_n\sqrt{n}} + \frac{\sqrt{\log(n)}}{\sqrt{n}}\right)$$

for some constant C. Thus, from the Borel-Cantelli lemma, we can conclude that

$$\lim_{n\to\infty} \sup_{f\in\mathcal{F}_{Q_n^m,\infty}} |\widehat{\mathrm{TV}}_{\varepsilon_n,n}(f;\mu) - \mathrm{TV}_{\varepsilon_n}(f;\mu)| = 0$$

almost surely with an exceptional set of measure zero that may depend on m. Taking the union over all these sets yields a uniform set of measure zero, such that the above limit holds for all $m \in \mathbb{N}$. Since every sequence (f_n) with $\limsup_{n\to\infty} \frac{\lambda_n}{\varepsilon_n}[f_n]_{\mathcal{B}} < \infty$ satisfies $f_n \in \mathcal{F}_{Q_n^m,\infty}$ for some $m \in \mathbb{N}$, the above considerations prove the statement. □

C Results for the Nonlocal Perimeter and Total Variation

In this section we essentially extend the Gamma-convergence proof of the nonlocal total variation TV_ε for Euclidean norm balls, given by Bungert & Stinson (2024), to general norm balls. Note however that we impose stronger assumptions on the conditional densities μ_0 and μ_1 which strongly simplifies the arguments. Proving Gamma-convergence has two ingredients: The first one is the so-called liminf inequality, which demands that whenever $u_n \to u$ in $L^1(\Omega)$ we have

$$\mathrm{TV}(u;\mu) \leq \liminf_{n\to\infty} \mathrm{TV}_{\varepsilon_n}(u_n;\mu),$$

see Lemma C.5. The second ingredient is the limsup inequality which demands that for any $u \in L^1(\Omega)$ there exists a sequence $(u_n) \subset L^1(\Omega)$ (called recovery sequence) converging to u such that

$$\limsup_{n\to\infty} \mathrm{TV}_{\varepsilon_n}(u_n;\mu) \leq \mathrm{TV}(u;\mu).$$

Note that by virtue of the liminf inequality, equality holds above. Thanks to the regularity assumptions in Assumption 2 the proof of this result is pretty standard and follows from a Taylor expansion and a diagonal argument. In our scenario, where we are interested in minimizers of (9), we need a slightly weaker result, which just requires the limsup inequality to be true for Barron functions, see Lemma C.6.

Furthermore, because our subsequent analysis operates with L^1 functions rather than just Barron functions, we must slightly modify the definition of TV_ε. Specifically, the supremum and infimum must be replaced by essential supremum and infimum to account for the fact that L^1 functions are defined only up to sets of measure zero. This leads to

$$\mathrm{TV}_\varepsilon(f;\mu) \coloneqq \int_\Omega \frac{\operatorname{ess\,sup}_{B_\varepsilon(x)\cap\Omega} f - f(x)}{\varepsilon}\mu_0(x)\,\mathrm{d}x + \int_\Omega \frac{f(x) - \operatorname{ess\,inf}_{B_\varepsilon(x)\cap\Omega} f}{\varepsilon}\mu_1(x)\,\mathrm{d}x.$$

We note, however, that for continuous functions such as Barron functions, this generalized formulation coincides with our previous definition in (6).

With the aim of proving the liminf inequality, let us first use the definition of TV_ε in (6) to define a nonlocal perimeter through

$$\mathrm{Per}_\varepsilon(A;\mu) \coloneqq \mathrm{TV}_\varepsilon(\chi_A;\mu).$$

for measurable sets A, where χ_A denotes the characteristic function of the set A, taking the value 1 inside A and 0 outside A. Bungert et al. (2023) showed that Per_ε indeed qualifies as *perimeter*, being submodular and satisfying a coarea formula. Gamma-convergence of this perimeter was studied by Bungert & Stinson (2024) in the case where $\|\cdot\|$ is the Euclidean norm. Additionally, for abbreviation we denote

$$\mathrm{Per}^0_\varepsilon(A;\mu) \coloneqq \int_\Omega \frac{\operatorname{ess\,sup}_{B_\varepsilon(x)\cap\Omega} \chi_A - \chi_A(x)}{\varepsilon}\mu_0(x)\,\mathrm{d}x$$
$$\mathrm{Per}^1_\varepsilon(A;\mu) \coloneqq \int_\Omega \frac{\chi_A(x) - \operatorname{ess\,inf}_{B_\varepsilon(x)\cap\Omega} \chi_A}{\varepsilon}\mu_1(x)\,\mathrm{d}x$$

so that $\mathrm{Per}_\varepsilon(A;\mu) = \mathrm{Per}^0_\varepsilon(A;\mu) + \mathrm{Per}^1_\varepsilon(A;\mu)$.

First, we state a useful representation of the nonlocal perimeter, see also Chambolle et al. (2012). For this statement, we need to define the measure theoretic interior of a set $A \subset \mathbb{R}^d$. It is given by

$$A^1 \coloneqq \left\{ x \in \mathbb{R}^d \;\middle|\; \lim_{r\to 0} \frac{\mathcal{L}^d(B^2_r(x)\cap A)}{\mathcal{L}^d(B^2_r(x))} = 1 \right\}$$

where $B^2_r(x)$ denotes the open ball with radius r around x with respect to the Euclidean norm $\|\cdot\|_2$. We refer to Evans & Gariepy (1991) for more details.

Lemma C.1. *Let $A \subset \Omega$ be measurable. Then,*

$$\mathrm{Per}_\varepsilon(A;\mu) = \frac{1}{\varepsilon}\mu_0(\{x \in (A^1)^c \mid \mathrm{dist}(x, A^1) < \varepsilon\}) + \frac{1}{\varepsilon}\mu_1(\{x \in A^1 \mid \mathrm{dist}(x, (A^1)^c) < \varepsilon\}).$$

We also need an auxiliary result which follows from the product rule. For this, see Ambrosio et al. (2000) for the definition of the standard total variation $\mathrm{TV}(u)$ for a function of bounded variation $u \in BV(\Omega)$. Furthermore, $W^{1,1}(\Omega)$ denotes the Sobolev space of absolutely integrable functions whose weak derivatives are also absolutely integrable. It holds $W^{1,1}(\Omega) \subset BV(\Omega)$ and the inclusion is strict.

Lemma C.2. *Let $\Omega \subset \mathbb{R}^d$ be an open domain with Lipschitz boundary, $u \in W^{1,1}(\Omega) \cap L^\infty(\Omega)$ and $v \in BV(\Omega) \cap L^\infty(\Omega)$ be non-negative. Then, $uv \in BV(\Omega)$ with*

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

where $\beta > 0$ is a constant such that $\|\cdot\|_2 \le \beta\|\cdot\|_\star$.

Proof. First, let $\tilde{v} \in BV(\mathbb{R}^d)$ be an extension of v with compact support such that $0 \le \tilde{v} \le \|v\|_{L^\infty(\Omega)}$. Such a function exists, as there exists an extension operator E such that $Ev \in BV(\mathbb{R}^d)$

and Eu has compact support, since Ω is an extension domain. Subsequently, we can apply a Lipschitz-continuous truncation operator to guarantee the bounds, which preserves the compact support leads to $\tilde{v} \in BV(\mathbb{R}^d)$ (Ambrosio et al., 2000, Theorem 3.96). We then consider $v_\varepsilon \coloneqq \tilde{v} * \varphi_\varepsilon$, where φ_ε is a standard mollifier (i.e. $\varphi_\varepsilon = \varepsilon^{-d}\varphi(\frac{\cdot}{\varepsilon})$ with some $\varphi \in C_c^\infty(\mathcal{B}_1(0)), \varphi \geq 0$ and $\int \varphi = 1$). Then, $v_\varepsilon \to v$ in $L^1(\Omega)$ and $\mathrm{TV}(v_\varepsilon) \to \mathrm{TV}(v)$ (Ambrosio et al., 2000, Remark 3.22). Moreover, $v_\varepsilon \in C_c^\infty(\mathbb{R}^d)$ and

$$\|v_\varepsilon\|_{L^\infty(\mathbb{R}^d)} \leq \|\tilde{v}\|_{L^\infty(\mathbb{R}^d)} \int_{\mathbb{R}^d} \varphi(x)\,\mathrm{d}x = \|v\|_{L^\infty(\Omega)}$$

as well as

$$\lim_{\varepsilon\to 0} \|uv_\varepsilon - uv\|_{L^1(\Omega)} \leq \|u\|_{L^\infty(\Omega)} \lim_{\varepsilon\to 0} \|v_\varepsilon - v\|_{L^1(\Omega)} = 0.$$

Hence, we can use the product rule together with the lower semi-continuity of the TV functional to obtain

$$\begin{aligned}
\mathrm{TV}(uv) &\leq \liminf_{\varepsilon\to 0} \mathrm{TV}(uv_\varepsilon) = \liminf_{\varepsilon\to 0} \int_\Omega \|\nabla(uv_\varepsilon)\|_2\,\mathrm{d}x \\
&\leq \liminf_{\varepsilon\to 0} \beta \int_\Omega \|\nabla u\|_\star |v_\varepsilon|\,\mathrm{d}x + \int_\Omega |u| \|\nabla v_\varepsilon\|_2\,\mathrm{d}x \\
&\leq \liminf_{\varepsilon\to 0} \beta \int_\Omega \|\nabla u\|_\star |v_\varepsilon|\,\mathrm{d}x + \|u\|_{L^\infty(\Omega)}\,\mathrm{TV}(v_\varepsilon).
\end{aligned}$$

Finally, (after extracting a subsequence) v_ε converges to v pointwise almost everywhere on Ω and $\|\nabla u\|_\star v_\varepsilon$ is bounded by the integrable function $\|\nabla u\|_\star v$. Thus, the dominated convergence theorem yields

$$\lim_{\varepsilon\to 0} \int_\Omega \|\nabla u\|_\star v_\varepsilon\,\mathrm{d}x = \int_\Omega \|\nabla u\|_\star v\,\mathrm{d}x$$

concluding the proof. □

To prove the following liminf inequality for the perimeters—which will easily imply the desired result for the total variations—we proceed similar to Bungert & Stinson (2024); Siethoff (2025). We start with a compactness result.

Lemma C.3. *Let Assumption 2 be satisfied. Then, for any sequence of positive numbers (ε_k) converging to zero and measurable sets (A_k) with $\limsup_{k\to\infty} \mathrm{Per}_{\varepsilon_k}(A_k;\mu) < \infty$ there exists a subsequence (A_{n_j}) and a set A with $\mathrm{TV}(\chi_A) < \infty$ such that*

$$\chi_{A_k} \to \chi_A \text{ in } L^1(\Omega).$$

Proof. We define the sequences (u_k) and (v_k) by

$$u_k(x) \coloneqq \max\left\{1 - \frac{\mathrm{dist}(x, A_k^1)}{\varepsilon_k}, 0\right\} \qquad v_k(x) \coloneqq \min\left\{\frac{\mathrm{dist}(x, (A_k^1)^c)}{\varepsilon_k}, 1\right\}$$

for $x \in \mathbb{R}^d$. It holds that

$$\|\nabla u_k\|_\star = \frac{1}{\varepsilon_k}\chi_{\{x\in\mathbb{R}^d \,|\, 0<\mathrm{dist}(x,A_k^1)<\varepsilon_k\}}, \qquad \|\nabla v_k\|_\star = \frac{1}{\varepsilon_k}\chi_{\{x\in\mathbb{R}^d \,|\, 0<\mathrm{dist}(x,(A_k^1)^c)<\varepsilon_k\}}.$$

almost everywhere (see Chambolle et al. (2012, p. 9)). Therefore, we can deduce that

$$\mathrm{Per}^0_{\varepsilon_k}(A_k;\mu) \geq \int_\Omega \|\nabla u_k\|_\star \mu_0(x)\,\mathrm{d}x,$$

since $\{x \in \Omega \,|\, 0 < \mathrm{dist}(x, A_k^1) < \varepsilon_k\} \subset \{x \in (A_k^1)^c \,|\, \mathrm{dist}(x, A_k^1) < \varepsilon_k\}$. Analogously, we obtain

$$\mathrm{Per}^1_{\varepsilon_k}(A_k;\mu) \geq \int_\Omega \|\nabla v_k\|_\star\,\mathrm{d}\mu_1(x).$$

Since the densities μ_0 and μ_1 are continuous up to the boundary and $\mu_0 + \mu_1 > 0$ on $\overline{\Omega}$, these functions can be extended to functions $\tilde{\mu}_i \in C(\tilde{\Omega})$, where $\tilde{\Omega}$ is an open neighborhood of $\overline{\Omega}$ and additionally $\tilde{\mu}_0 + \tilde{\mu}_1 \geq \delta$ for some $\delta > 0$. We now consider, for $i \in \{0, 1\}$,

$$\tilde{\Omega}_i \coloneqq \left\{ x \in \tilde{\Omega} \,|\, \mu_i(x) > \frac{\delta}{3} \right\}$$

which defines open sets such that $\overline{\Omega} \subset \tilde{\Omega}_0 \cup \tilde{\Omega}_1$. In particular, there exists a partition of unity Ψ_0, Ψ_1 subordinate to this covering, meaning that Ψ_0, Ψ_1 are smooth functions satisfying $0 \leq \Psi_i \leq 1$, $\operatorname{supp} \Psi_i \subset \tilde{\Omega}_i$ and $\Psi_0 + \Psi_1 = 1$ on $\overline{\Omega}$. Lemma C.2 implies

$$\begin{aligned} \operatorname{TV}(u_k \Psi_0) &\leq \beta \int_\Omega \|\nabla u_k\|_\star \Psi_0 \,\mathrm{d}x + \|u_k\|_{L^\infty(\Omega)} \operatorname{TV}(\Psi_0) \\ &\leq \frac{3\beta}{\delta} \int_\Omega \|\nabla u_k\|_\star \mu_0 \,\mathrm{d}x + \operatorname{TV}(\Psi_0). \end{aligned}$$

Consequently,

$$\limsup_{k\to\infty} \operatorname{TV}(u_k \Psi_0) \leq \limsup \frac{3\beta}{\delta} \operatorname{Per}^0_{\varepsilon_k}(A_k; \mu) + \operatorname{TV}(\Psi_0) < \infty$$

and we can apply compactness in $BV(\Omega)$ (Ambrosio et al., 2000, Theorem 3.23) to extract a converging subsequence (without relabeling) such that $u_k \Psi_0$ converges in $L^1(\Omega)$ to some $u \in BV(\Omega)$. Similarly, we find that $v_k \Psi_1$ converges in $L^1(\Omega)$ to a function $v \in BV(\Omega)$. We finalize the proof by showing that χ_{A_k} converges to $u + v$. To see this, we first observe that

$$\begin{aligned} \frac{\delta}{3} \int_{\tilde{\Omega}_0 \cap \Omega} |u_k - \chi_{A_k}| \,\mathrm{d}x &= \int_{\tilde{\Omega}_0 \cap \Omega} \left(1 - \frac{\operatorname{dist}(x, A_k^1)}{\varepsilon_k}\right) \chi_{\{x \in \Omega \,|\, 0 < \operatorname{dist}(x, A_k^1 < \varepsilon_k\}}(x) \,\mathrm{d}x \\ &\leq \int_{\tilde{\Omega}_0 \cap \Omega} \chi_{\{x \in \Omega \,|\, 0 < \operatorname{dist}(x, A_k^1 < \varepsilon_k\}}(x) \mu_0(x) \,\mathrm{d}x \\ &\leq \varepsilon_k \operatorname{Per}^0_{\varepsilon_k}(A_k; \mu) \to 0, \quad \text{as } k \to \infty. \end{aligned}$$

Using the same ideas, we can argue that $\int_{\tilde{\Omega}_1 \cap \Omega} |v_k - \chi_{A_k}| \,\mathrm{d}x \to 0$ as $k \to \infty$. Hence, using that $\Psi_0 + \Psi_1 = 1$ on Ω, we obtain that

$$\int_\Omega |\Psi_0 u_k + \Psi_1 v_k - \chi_{A_k}| \,\mathrm{d}x \leq \int_{\tilde{\Omega}_0 \cap \Omega} \Psi_0 |u_k - \chi_{A_k}| \,\mathrm{d}x + \int_{\tilde{\Omega}_1 \cap \Omega} \Psi_1 |v_k - \chi_{A_k}| \,\mathrm{d}x$$

and the right hand side converges to zero as $k \to \infty$. Consequently, χ_{A_k} must have the same limit in $L^1(\Omega)$ as $\Psi_0 u_k + \Psi_1 v_k$ which is $u + v \in BV(\Omega)$. Since χ_{A_k} only takes the values 0 and 1, the limit must do so too, meaning that there exists a measurable set A such that $\chi_A = u + v$ and in particular, $\operatorname{TV}(\chi_A) < \infty$. □

We establish the $\liminf$ inequality using similar arguments.

Lemma C.4. *Let Assumption 2 be satisfied. Let (A_k) be a collection of measurable subsets of Ω and $A \subset \Omega$ measurable such that $\chi_{A_k} \to \chi_A$ in $L^1(\Omega)$. Then, for any sequence of positive numbers (ε_k) converging to zero it holds that*

$$\liminf_{k\to\infty} \operatorname{Per}_{\varepsilon_k}(A_k; \mu) \geq \operatorname{Per}(A; \mu),$$

where we define

$$\operatorname{Per}(A; \mu) \coloneqq \begin{cases} \displaystyle\int_{\partial^\star A \cap \Omega} (\mu_0 + \mu_1) \left\| \frac{\mathrm{d}D\chi_A}{\mathrm{d}|D\chi_A|} \right\|_\star \mathrm{d}\mathcal{H}^{d-1} & \text{if } \chi_A \in BV(\Omega), \\ +\infty & \text{else.} \end{cases}$$

To clarify our notation, we recall that for a function $u \in BV(\Omega)$ the distributional derivative can be represented by a finite $\mathbb{R}^d$-valued Radon measure Du (Ambrosio et al., 2000). Furthermore, we denote the associated total variation measure as $|Du|$ and $\mathcal{H}^{d-1}$ is the $(d-1)$-dimensional Hausdorff measure.

Proof. We first note that w.l.o.g. we can assume

$$\liminf_{k\to\infty} \mathrm{Per}_{\varepsilon_k}(\chi_{A_k};\mu) = \lim_{k\to\infty} \mathrm{Per}_{\varepsilon_k}(\chi_{A_k};\mu) < \infty$$

by extracting a subsequence and observing that the inequality is trivially satisfied if this limit is infinite. From Lemma C.3 we can conclude that $\chi_A \in BV(\Omega)$. We define the sequences (u_k) and (v_k) as in the previous proof via

$$u_k(x) \coloneqq \max\left\{1 - \frac{\mathrm{dist}(x, A_k^1)}{\varepsilon_k}, 0\right\} \qquad v_k(x) \coloneqq \min\left\{\frac{\mathrm{dist}(x, (A_k^1)^c)}{\varepsilon_k}, 1\right\}$$

for $x \in \mathbb{R}^d$. As before (see the proof of Lemma C.3) it holds that

$$\mathrm{Per}^0_{\varepsilon_k}(A_k;\mu) \geq \int_\Omega \|\nabla u_k\|_\star \mu_0(x)\,\mathrm{d}x.$$

Let (δ_n) be a sequence of positive real numbers that converge to zero as n approaches infinity. We define the sets

$$\Omega^i_{\delta_n} \coloneqq \{\mu_i > \delta_n\}$$

for $i \in \{0,1\}$, which are open thanks to Assumption 2. Since we assumed that the sequence $(\mathrm{Per}^0_{\varepsilon_k}(A_k;\mu))$ is bounded by some constant C, we can conclude

$$\begin{aligned}|Du_k|(\Omega^0_{\delta_n}) &= \int_{\Omega^0_{\delta_n}} \|\nabla u_k(x)\|_2\,\mathrm{d}x \leq \beta \int_{\Omega^0_{\delta_n}} \|\nabla u_k(x)\|_\star\,\mathrm{d}x \\ &\leq \frac{\beta}{\delta_n}\int_\Omega \|\nabla u_k(x)\|_\star \mu_0(x)\,\mathrm{d}x \leq \frac{\beta}{\delta_n}\mathrm{Per}^0_{\varepsilon_k}(A_k;\mu) \leq \frac{C\beta}{\delta_n}\end{aligned}$$

using the norm equivalence $\|\cdot\|_2 \leq \beta\|\cdot\|_\star$ and the fact that μ_0 is bounded from below by δ_n on $\Omega^0_{\delta_n}$. By Ambrosio et al. (2000, eq (3.11)), together with the convergence of u_k to χ_A in $L^1(\Omega^0_{\delta_n})$ which follows as in the proof of Lemma C.3, this implies the weak* convergence of Du_k to $D\chi_A$ on $\Omega^0_{\delta_n}$. Hence, applying the Reshetnyak lower semicontinuity theorem (see e.g. Ambrosio et al. (2000, Theorem 2.38)) yields

$$\begin{aligned}\liminf_{k\to\infty} \mathrm{Per}^0_{\varepsilon_k}(A_k;\mu) &\geq \liminf_{k\to\infty} \int_{\Omega^0_{\delta_n}} \|\nabla u_k\|_\star \mu_0\,\mathrm{d}x \\ &\geq \int_{\Omega^0_{\delta_n}} \left\|\frac{\mathrm{d}D\chi_A}{\mathrm{d}|D\chi_A|}\right\|_\star \mu_0\,\mathrm{d}|D\chi_A|.\end{aligned}$$

Since this holds for every $n \in \mathbb{N}$, we can take the limit $n \to \infty$ and use the dominated convergence theorem to obtain

$$\begin{aligned}\liminf_{k\to\infty} \mathrm{Per}^0_{\varepsilon_k}(A_k;\mu) &\geq \lim_{n\to\infty} \int_{\Omega^0_{\delta_n}} \left\|\frac{\mathrm{d}D\chi_A}{\mathrm{d}|D\chi_A|}\right\|_\star \mu_0\,\mathrm{d}|D\chi_A| \\ &= \int_\Omega \left\|\frac{\mathrm{d}D\chi_A}{\mathrm{d}|D\chi_A|}\right\|_\star \mu_0\,\mathrm{d}|D\chi_A| \\ &= \int_{\partial^\star A\cap\Omega} \left\|\frac{\mathrm{d}D\chi_A}{\mathrm{d}|D\chi_A|}\right\|_\star \mu_0\,\mathrm{d}\mathcal{H}^{d-1}.\end{aligned}$$

Repeating the arguments for v_k instead of u_k and μ_1 instead of μ_0 concludes the proof. □

In fact, a corresponding result for the nonlocal total variation follows directly from the following coarea formula (Bungert et al., 2023):

$$\mathrm{TV}_\varepsilon(f;\mu) = \int_{\mathbb{R}} \mathrm{Per}_\varepsilon(\{f \geq t\};\mu)\,\mathrm{d}t.$$

Lemma C.5. *Let Assumption 2 be satisfied. Let $(u_k) \subset BV(\Omega)$ and $u \in BV(\Omega)$ such that $u_k \to u$ in $L^1(\Omega)$. Then, for any sequence of positive numbers (ε_k) converging to zero, it holds that*

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

Proof. Without loss of generality, by passing to a subsequence if necessary, we can assume that $\liminf_{k\to\infty} \mathrm{TV}_{\varepsilon_k}(u_k;\mu) = \lim_{k\to\infty} \mathrm{TV}_{\varepsilon_k}(u_k;\mu)$. Moreover, there exists a subsequence (u_k) (without relabeling) such that $\{u_k > t\} \to \{u > t\}$ in $L^1(\Omega)$ for almost every $t \in \mathbb{R}$. Consequently, using the coarea formula, Fatou's lemma, and Lemma C.4 we obtain

$$\begin{aligned}\liminf_{k\to\infty} \mathrm{TV}_{\varepsilon_k}(u_k;\mu) &= \liminf_{k\to\infty} \int_{-\infty}^{\infty} \mathrm{Per}_{\varepsilon_k}(\{u_k > t\};\mu)\,\mathrm{d}t \\ &\geq \int_{-\infty}^{\infty} \liminf_{k\to\infty} \mathrm{Per}_{\varepsilon_k}(\{u_k > t\};\mu)\,\mathrm{d}t \\ &\geq \int_{-\infty}^{\infty} \mathrm{Per}(\{u > t\};\mu)\,\mathrm{d}t = \mathrm{TV}(u;\mu).\end{aligned}$$

□

Next, we prove that for Barron functions f, $\mathrm{TV}_\varepsilon(f;\mu)$ converges to $\mathrm{TV}(f;\mu)$ as ε tends to zero.

Lemma C.6. *Let $f \in \mathcal{B}(\Omega)$ and $(\varepsilon_n)_{n\in\mathbb{N}}$ a sequence of positive numbers converging to zero. Then,*

$$\lim_{n\to\infty} \mathrm{TV}_{\varepsilon_n}(f;\mu) = \mathrm{TV}(f;\mu).$$

Proof. Let $\varepsilon > 0$ arbitrary. By Lemma A.1 the function f is Lipschitz continuous, hence it is differentiable almost everywhere by Rademacher's theorem. Consequently, for almost every $x \in \Omega$ with $\mathrm{dist}(x, \partial\Omega) > \varepsilon$, we have

$$f(x + \varepsilon u) - f(x) = \varepsilon \nabla f(x) \cdot u + o_x(\varepsilon)$$

for $u \in B_1(0)$. In particular, taking the supremum with respect to u, we obtain

$$\operatorname*{ess\,sup}_{B_\varepsilon(x)\cap\Omega} f - f(x) = \varepsilon \|\nabla f(x)\|_\star + o_x(\varepsilon).$$

Taking the limit $\varepsilon \to 0$, we obtain that

$$\lim_{\varepsilon\to 0} \frac{1}{\varepsilon} \left(\operatorname*{ess\,sup}_{B_\varepsilon(x)\cap\Omega} f - f(x) \right) = \|\nabla f(x)\|_\star$$

for almost every $x \in \Omega$. Moreover, $\frac{1}{\varepsilon}\left(\operatorname{ess\,sup}_{B_\varepsilon(x)\cap\Omega} f - f(x)\right)$ is dominated by a constant which is integrable over Ω. Thus, with Lebesgue's dominated convergence theorem, we can conclude

$$\lim_{n\to\infty} \int_\Omega \frac{1}{\varepsilon_n} \left(\operatorname*{ess\,sup}_{B_{\varepsilon_n}(x)\cap\Omega} f - f(x) \right) \mu_0(x)\,\mathrm{d}x = \int_\Omega \|\nabla f(x)\|_\star \mu_0(x)\,\mathrm{d}x.$$

Repeating the same arguments for the μ_1-term yields the desired result. □

D Results for the Full Regularizer

In the previous section we studied the nonlocal total variation and proved the liminf inequality as well as a version of the limsup inequality. Next, we need analogous results for the regularizer in (9) which requires two key changes: First, we need to consider the empirical version of the nonlocal total variation and second the weight decay regularizer needs to be added to the functional to ensure existence of minimizers and compactness of sequences of minimizers. As already mentioned in Section 5, it is useful to additionally divide the entire functional by ε_n, leading to

$$R_n(\theta) \coloneqq \widehat{\mathrm{TV}}_{\varepsilon_n,n}(f_\theta;\mu) + \frac{\lambda_n}{\varepsilon_n} R_{WD}(\theta).$$

The liminf inequality for this regularizer functional is proved in Proposition D.1, and we refer to Proposition D.2 for the limsup inequality.

Proposition D.1. *Let Assumption 2 be satisfied. Let $(f_n) \subset \mathcal{B}(\Omega)$ such that $f_n \to f \in BV(\Omega)$ in $L^1(\Omega)$. Moreover, assume that*

$$\lim_{n\to\infty} \frac{\varepsilon_n}{\lambda_n}\sqrt{\frac{\log(n)}{n}} = 0, \qquad \lim_{n\to\infty} \frac{1}{\lambda_n\sqrt{n}} = 0, \qquad \alpha \coloneqq \lim_{n\to\infty} \frac{\lambda_n}{\varepsilon_n} \in [0,\infty).$$

Then,

$$\mathrm{TV}(f;\mu) + \alpha[f]_{\mathcal{B}} \le \liminf_{n\to\infty} \widehat{\mathrm{TV}}_{\varepsilon_n,n}(f_n;\mu) + \frac{\lambda_n}{\varepsilon_n}[f_n]_{\mathcal{B}}.$$

almost surely (with respect to the samples). In the case where $\alpha = 0$, we use the convention that $\alpha \cdot [f]_{\mathcal{B}} = 0$ even if $[f]_{\mathcal{B}} = \infty$.

Proof. Without loss of generality, by passing to a subsequence, we can assume that

$$\liminf_{n\to\infty} \widehat{\mathrm{TV}}_{\varepsilon_n,n}(f_n;\mu) + \frac{\lambda_n}{\varepsilon_n}[f_n]_{\mathcal{B}} = \lim_{n\to\infty} \widehat{\mathrm{TV}}_{\varepsilon_n,n}(f_n;\mu) + \frac{\lambda_n}{\varepsilon_n}[f_n]_{\mathcal{B}}.$$

Moreover, if $\limsup_{n\to\infty} \frac{\lambda_n}{\varepsilon_n}[f_n]_{\mathcal{B}} = \infty$, the claim becomes trivial by extracting another subsequence, where the limsup becomes a limit. Hence, in the following, we consider the case, where $\limsup_{n\to\infty} \frac{\lambda_n}{\varepsilon_n}[f_n]_{\mathcal{B}}$ is finite. By inserting zero, we can rewrite

$$\widehat{\mathrm{TV}}_{\varepsilon_n,n}(f_n;\mu) + \frac{\lambda_n}{\varepsilon_n}[f_n]_{\mathcal{B}} = \mathrm{TV}_{\varepsilon_n}(f_n;\mu) + \widehat{\mathrm{TV}}_{\varepsilon_n,n}(f_n;\mu) - \mathrm{TV}_{\varepsilon_n}(f_n;\mu) + \frac{\lambda_n}{\varepsilon_n}[f_n]_{\mathcal{B}}$$

and Lemma B.3 implies that the difference of the two middle terms converges to zero almost surely as $n \to \infty$. We can additionally use the lower semi-continuity of the Barron semi-norm and the liminf inequality from Lemma C.5 to conclude

$$\begin{aligned}\liminf_{n\to\infty} \widehat{\mathrm{TV}}_{\varepsilon_n,n}(f_n;\mu) + \frac{\lambda_n}{\varepsilon_n}[f_n]_{\mathcal{B}} &= \liminf_{n\to\infty} \mathrm{TV}_{\varepsilon_n}(f_n;\mu) + \frac{\lambda_n}{\varepsilon_n}[f_n]_{\mathcal{B}} \\ &\ge \liminf_{n\to\infty} \mathrm{TV}_{\varepsilon_n}(f_n;\mu) + \liminf_{n\to\infty} \frac{\lambda_n}{\varepsilon_n}[f_n]_{\mathcal{B}} \\ &\ge \mathrm{TV}(f;\mu) + \alpha[f]_{\mathcal{B}}.\end{aligned}$$

□

Remark D.1. In particular if the functions f_n are neural networks parametrized by $\theta_n \in \Theta_{m_n}$, then

$$\mathrm{TV}(f;\mu) + \alpha[f]_{\mathcal{B}} \le \liminf_{n\to\infty} \widehat{\mathrm{TV}}_{\varepsilon_n,n}(f_n;\mu) + \frac{\lambda_n}{\varepsilon_n} R_{WD}(\theta_n),$$

since $R_{WD}(\theta_n) \ge [f_n]_{\mathcal{B}}$ (see Remark 2.1).

From Lemma C.6, we can immediately deduce the following equality.

Lemma D.1. *For $f \in \mathcal{B}(\Omega)$ it holds that*

$$\limsup_{n\to\infty} \mathrm{TV}_{\varepsilon_n}(f;\mu) + \frac{\lambda_n}{\varepsilon_n}[f]_{\mathcal{B}} = \mathrm{TV}(f;\mu) + \alpha[f]_{\mathcal{B}},$$

where $\alpha \coloneqq \limsup_{n\to\infty} \frac{\lambda_n}{\varepsilon_n}$.

It remains to close the gap to obtaining a sequence of neural networks and considering the empirical version of the non-local total variation.

Proposition D.2. *Let $f \in \mathcal{B}(\Omega)$ and let*

$$\lim_{n\to\infty} \frac{1}{\varepsilon_n\sqrt{m_n}} = 0, \qquad \lim_{n\to\infty} \frac{1}{\varepsilon_n\sqrt{n}} = 0, \qquad \limsup_{n\to\infty} \frac{\lambda_n}{\varepsilon_n} =: \alpha.$$

Then, there exists a sequence of neural networks f_n with m_n neurons parametrized by $\theta_n \in \Theta_{m_n}$ such that $f_n \to f$ in $L^1(\Omega)$ and

$$\limsup_{n\to\infty} \widehat{\mathrm{TV}}_{\varepsilon_n,n}(f_n;\mu) + \frac{\lambda_n}{\varepsilon_n} R_{WD}(\theta_n) \le \mathrm{TV}(f;\mu) + \alpha[f]_{\mathcal{B}}.$$

almost surely (with respect to the samples).

Proof. From Lemma D.1, we infer that

$$\limsup_{n\to\infty} \mathrm{TV}_{\varepsilon_n}(f;\mu) + \frac{\lambda_n}{\varepsilon_n}[f]_{\mathcal{B}} = \mathrm{TV}(f;\mu) + \alpha[f]_{\mathcal{B}}.$$

We will approximate the function f by neural networks. Precisely, for $n \in \mathbb{N}$, there exists a neural network with m_n neurons parametrized by θ_n such that

$$\|f_n - f\|_{L^\infty(\Omega)} \le \frac{C}{\sqrt{m_n}}, \qquad R_{WD}(\theta_n) \le [f]_{\mathcal{B}(\Omega)},$$

where the constant C depends on the Barron semi-norm of f as well as the domain Ω, see Lemma A.3. Hence,

$$|\,\mathrm{TV}_{\varepsilon_n}(f;\mu) - \mathrm{TV}_{\varepsilon_n}(f_n;\mu)| \le \frac{C}{\varepsilon_n}\|f_n - f\|_{L^\infty(\Omega)} \le \frac{C}{\varepsilon_n\sqrt{m_n}}.$$

Moreover, we can use Lemma B.2 to obtain

$$|\widehat{\mathrm{TV}}_{\varepsilon_n,n}(f_n;\mu) - \mathrm{TV}_{\varepsilon_n}(f_n;\mu)| \le \frac{C\sqrt{\log(2/\delta)}}{\sqrt{n}} + \frac{C}{\varepsilon_n\sqrt{n}}$$

with probability at least $1 - 2\delta$. Hence, combining all the estimates with Lemma D.1 and using the Borel–Cantelli lemma, we obtain

$$\begin{aligned}\limsup_{n\to\infty} \widehat{\mathrm{TV}}_{\varepsilon_n,n}(f_n;\mu) + \frac{\lambda_n}{\varepsilon_n}R_{WD}(\theta_n) = \limsup_{n\to\infty} \Big(&\mathrm{TV}_{\varepsilon_n}(f;\mu) + \mathrm{TV}_{\varepsilon_n}(f_n;\mu) - \mathrm{TV}_{\varepsilon_n}(f;\mu) \\ &+ \widehat{\mathrm{TV}}_{\varepsilon_n,n}(f_n;\mu) - \mathrm{TV}_{\varepsilon_n}(f_n;\mu) + \frac{\lambda_n}{\varepsilon_n}R_{WD}(\theta_n)\Big) \\ &\le \mathrm{TV}(f;\mu) + \alpha[f]_{\mathcal{B}}\end{aligned}$$

almost surely. □

We note that our analysis benefits significantly from the homogeneous Barron space, which guarantees approximation rates independent of the ambient dimension d.

Remark D.2. The sequence (f_n) arising from the previous proof has the rate of convergence

$$\|f_n - f\|_{L^\infty(\Omega)} \le \frac{C}{\sqrt{m_n}}$$

for some constant C depending on f and Ω and satisfies $[f_n] \le R_{WD}(\theta_n) \le [f]_{\mathcal{B}}$.

E Proofs of Propositions 5.1 to 5.3

In this section, we carry out the proof of the main auxiliary results for proving Theorem 1. In the language of Gamma-convergence, Proposition 5.2 is the limsup inequality, Proposition 5.1 is the liminf equality, and together they imply the following Gamma-convergence theorem.

Theorem 2. *Under the conditions of Theorem 1 define the functionals*

$$G_n : L^1(\Omega) \to [0,\infty], \quad G_n(u) := \inf\{F_n(\theta) \,|\, \theta \in \Theta_{m_n} \text{ with } u = f_\theta\}, \qquad n \in \mathbb{N},$$

where f_θ for $\theta \in \Theta_{m_n}$ denotes the neural network with parameters θ. As $n \to \infty$ the functionals G_n Gamma-converge to F, defined in (10)*, in the $L^1(\Omega)$ topology.*

Proof. Notice that the limsup inequality for this functional follows directly from Proposition 5.2 since $G_n(u) \le F_n(\theta)$ if θ parametrizes u. For the liminf inequality one uses Proposition 5.1 with a parametrization θ_n of f_n that satisfies

$$F_n(\theta_n) \le G_n(f_n) + \frac{1}{n} \qquad \forall n \in \mathbb{N}.$$

□

We proceed with proving the $\liminf$ inequality Proposition 5.1.

Proof of Proposition 5.1. We first consider the case that $f \in \mathcal{B}$ and $\mathcal{R}(f) = \inf_{\mathcal{B}} \mathcal{R}$. In this case, by Proposition D.1,

$$\begin{aligned}\liminf_{n\to\infty} F_n(f_n) &\geq \liminf_{n\to\infty} \frac{\lambda_n}{\varepsilon_n} R_{WD}(\theta_n) + \widehat{\mathrm{TV}}_{\varepsilon_n,n}(f_n;\mu) \\ &\geq \liminf_{n\to\infty} \frac{\lambda_n}{\varepsilon_n} [f_n]_{\mathcal{B}} + \widehat{\mathrm{TV}}_{\varepsilon_n,n}(f_n;\mu) \\ &\geq \alpha[f]_{\mathcal{B}} + \mathrm{TV}(f;\mu)\end{aligned}$$

almost surely. In every other case, we need to prove that $\liminf_{n\to\infty} F_n(f_n) = \infty$. If $f \notin \mathcal{B}(\Omega)$, we know by Lemma A.2 that $\liminf_{n\to\infty} [f_n]_{\mathcal{B}} = \infty$ and consequently,

$$\liminf_{n\to\infty} F_n(f_n) \geq \liminf_{n\to\infty} \frac{\lambda_n}{\varepsilon_n} [f_n]_{\mathcal{B}} = \infty.$$

The last remaining case is $f \in \mathcal{B}$ with $\mathcal{R}(f) > \inf_{\mathcal{B}} \mathcal{R}$. In particular, there exists $\delta > 0$ such that $\mathcal{R}(f_n) - \inf_{\mathcal{B}} \mathcal{R} > \delta$ for n large enough. In this case, for convenience, we extract a subsequence such that the liminf under consideration becomes a limit. If then $\limsup_{n\to\infty} [f_n]_{\mathcal{B}} = \infty$, passing to the corresponding subsequence, the inequality follows trivially. Hence, we can assume that $\limsup_{n\to\infty} [f_n]_{\mathcal{B}} < \infty$, in particular, the Barron semi-norms are bounded by some constant $Q > 0$. Moreover, we can conclude that $f_n(0)$ needs to converge to $f(0)$ due to the convergence in L^1 and the fact that all the functions f_n and f are Lipschitz continuous with uniformly bounded Lipschitz constants. To see this, we assume the contrary, i.e. $|f_n(0) - f(0)| \geq \gamma > 0$ and conclude that there exists an open set U (independent of n) such that $|f_n - f| \geq \frac{\gamma}{2}$ on U, which contradicts the fact that f_n converges to f in $L^1(\Omega)$. Overall, there exists a constant $B > 0$ such that $f_n, f \in \mathcal{F}_{Q,B}$ as defined in (12). Therefore, invoking Lemma B.1, we obtain

$$\begin{aligned}F_n(f_n) &\geq \frac{\widehat{\mathcal{R}}_n(f_n) - \mathcal{R}(f_n)}{\varepsilon_n} + \frac{\mathcal{R}(f_n) - \inf_{\mathcal{B}} \mathcal{R}}{\varepsilon_n} \\ &\geq -\frac{C}{\varepsilon_n}\sqrt{\frac{\log(n)}{n}} + \frac{\delta}{\varepsilon_n},\end{aligned}$$

with probability at least $1 - \frac{1}{n^2}$. Since the first term is bounded, while the latter is unbounded, using the Borel–Cantelli lemma, we can conclude that

$$\liminf_{n\to\infty} F_n(f_n) = \infty$$

almost surely. □

The proof of the $\limsup$ inequality Proposition 5.2 uses similar ideas, combining Proposition D.2 with the generalization bound Lemma B.1

Proof of Proposition 5.2. The only case of interest is $f \in \mathcal{B}(\Omega)$ such that $\mathcal{R}(f) = \inf_{\mathcal{B}} \mathcal{R}$. In every other case, we only need to show that $f \in L^1(\Omega)$ can be approximated by neural networks, because the inequality becomes trivial. The existence of an approximating sequence of neural networks follows directly from Hornik (1991). Hence, in the following we assume that $f \in \mathcal{B}$ with $\mathcal{R}(f) = \inf_{\mathcal{B}} \mathcal{R}$. We consider the recovery sequence of neural networks (f_n) with corresponding parameters θ_n for the regularizer term given by Proposition D.2. We note that by Remark D.2, we can further assume that

$$\|f_n - f\|_{L^\infty(\Omega)} \leq \frac{C}{\sqrt{m_n}} \qquad \text{and} \qquad [f_n]_{\mathcal{B}} \leq [f]_{\mathcal{B}}$$

for a positive constant C depending on f. Moreover,

$$\begin{aligned}F_n(f_n) &= \frac{\widehat{\mathcal{R}}_n(f_n) - \inf_{\mathcal{B}} \mathcal{R}}{\varepsilon_n} + \frac{\lambda_n}{\varepsilon_n} R_{WD}(\theta_n) + \widehat{\mathrm{TV}}_{\varepsilon_n,n}(f_n;\mu) \\ &= \frac{\widehat{\mathcal{R}}_n(f_n) - \mathcal{R}(f_n)}{\varepsilon_n} + \frac{\mathcal{R}(f_n) - \inf_{\mathcal{B}} \mathcal{R}}{\varepsilon_n} + \frac{\lambda_n}{\varepsilon_n} R_{WD}(\theta_n) + \widehat{\mathrm{TV}}_{\varepsilon_n,n}(f_n;\mu).\end{aligned}$$

The first term on the right hand side is bounded by $\frac{C}{\varepsilon_n}\sqrt{\frac{\log(n)}{n}}$ with probability at least $1-\frac{1}{n^2}$ due to Lemma B.1. For justifying the application of this result, notice that the Barron semi-norms $[f_n]_{\mathcal{B}}$ are uniformly bounded and hence all f_n and the limit f are Lipschitz continuous with a uniform Lipschitz-constant. Moreover, for n large enough, we must have $|f_n(0)| \leq |f(0)|+1$ as the contrary would contradict the convergence in $L^\infty(\Omega)$.

The second term can be bounded using the Lipschitz continuity of the loss function ℓ to obtain

$$\frac{\mathcal{R}(f_n) - \inf_{\mathcal{B}} \mathcal{R}}{\varepsilon_n} \leq \frac{C}{\varepsilon_n}\|f_n - f\|_{L^1(\Omega)} \leq \frac{C}{\varepsilon_n\sqrt{m_n}}$$

and the right hand side converges to zero due to our assumptions on the scaling of m_n and ε_n. Hence, using the Borel–Cantelli lemma together with Proposition D.2 yields

$$\limsup_{n\to\infty} F_n(f_n) \leq \alpha[f]_{\mathcal{B}} + \mathrm{TV}(f;\mu) = F(f)$$

almost surely. □

Finally, we prove the compactness result Proposition 5.3.

Proof of Proposition 5.3. For abbreviation, we denote $f_n := f_{\theta_n}$. Since $\alpha > 0$, from $\limsup_{n\to\infty} F_n(\theta_n)$ we can deduce that there exists a constant $C > 0$ such that $[f_n]_{\mathcal{B}} \leq R_{WD}(\theta_n) \leq C$. In particular, all functions f_n are Lipschitz continuous with a uniform bound for the Lipschitz constants. Proceeding by contradiction, suppose that $|f_{n_j}(0)| \to \infty$ along a subsequence (f_{n_j}). Combining this with the uniform Lipschitz bound implies that either $f_{n_j} \geq C$ on Ω or $f_{n_j} \leq -C$ for sufficiently large j. Consequently, Assumption 1 forces $\lim_{j\to\infty} F_{n_j}(f_{n_j}) = \infty$. This directly contradicts the hypothesis that $\limsup_{n\to\infty} F_n(f_n) < \infty$, thereby confirming the boundedness of $(f_n(0))$. Therefore, we can apply Park et al. (2023, Lemma C.3) and obtain the existence of a function $f \in \mathcal{B}$ and a subsequence (f_{n_j}) such that $f_{n_j} \to f$ in $L^1(\Omega)$. □

F Additional Numerical Illustrations

In this section, we provide additional details on the experiments presented already, and we also include some additional experiments. We focus on the datasets from the main data, i.e., the two-dimensional toy dataset from Figure 2—where Bayes classifiers are not unique and we can nicely illustrate the different decision boundaries induced by natural and total variation regularization / adversarial training—the CIFAR-10 dataset, as well as ImageNet.

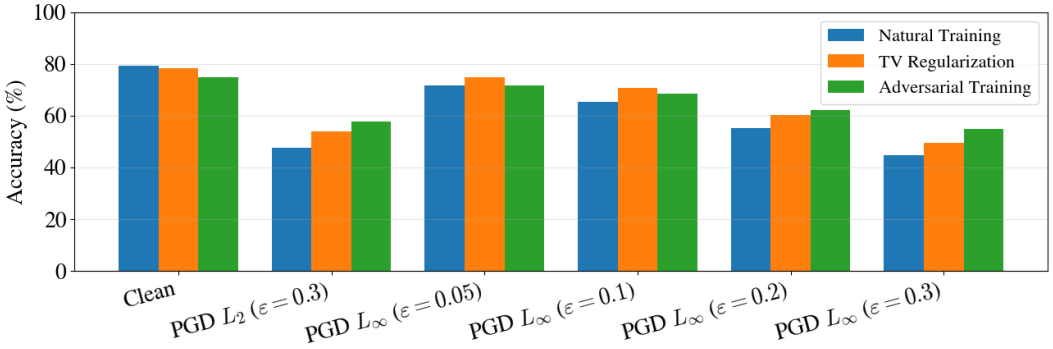


Figure 4: Clean and adversarial accuracies of the models from Figure 2, trained on the toy dataset using the architecture (1). We used $n = 1,000$ data points, $m_n = 2,000$ neurons, $\varepsilon_n = 0.24$, and $\lambda_n = 0.01 \cdot \varepsilon_n$. The Bayes (i.e., best achievable) accuracy for this dataset is 80%.

Toy Dataset. First, we revisit the toy example from Figure 2. Here, the probability distribution for the data points was chosen such that in the inner square $[-0.5, 0, 5]^2$ the probability of having label 0 or 1 is 50% each, making Bayes classifiers non-unique in this set. Furthermore, in the set

$[0.5, 1] \times [-1, 1] \cup [-0.5, 0.5] \times [-1, -0.5]$ the probability for a data point to have label 1 equals 90%. In the remaining area that probability is 10%. All Bayes classifiers equal 1 in the former set, 0 in the latter, and can take any number between 0 and 1 in the inner square $[-0.5, 0.5]^2$. Hence, this data distribution is particularly useful for evaluating how different training methods influence the decision boundary, particularly, in this set where it is undetermined by the data.

In Figure 4 we report the clean and adversarial accuracies of the three models from Figure 2. Note that, due to the overlap of the class conditional distributions, the best achievable accuracy on this dataset is 80%. Notably, adversarial training suffers most from the well-known accuracy-robustness trade-off on clean data, see also the rightmost panel in Figure 2. We observe that the classifier obtained by natural training has the highest clean accuracy, but as expected is most susceptible to all attacks considered here. The attacks themselves are standard; we consider one ℓ_2-attack and several ℓ_∞-attacks, all computed by projected gradient descent (PGD), see Madry et al. (2018). Moreover, TV regularization, i.e., problem (9), and standard adversarial training (Madry et al., 2018) produce comparable results. While TV regularization has better clean accuracy and is more robust to ℓ_2 and small ℓ_∞ attacks, adversarial training slightly dominates for larger adversarial budgets. Note that we used to cross-entropy loss for these results, hence the mathematical equivalence of TV regularization and adversarial training (Bungert et al., 2023) does not hold here.

Finally, in Figure 5 (which should be compared to Figure 2) we show the resulting decision boundaries when using ℓ_2-adversaries instead of ℓ_∞. Our theory does not predict gradient sparsity in this situation with ℓ_2-attacks, and indeed the resulting decision boundaries (just like the gradients) inside the undetermined set $[-0.5, 0.5]^2$ are not axis-parallel, but rather approximately diagonal. Nevertheless, the decision boundaries of TV regularization and adversarial training are better behaved than that of natural training, as a consequence of the implicit regularization with $\int_\Omega \|\nabla f\|_2(\mu_0(x) + \mu_1(x))\,\mathrm{d}x + \alpha[f]_{\mathcal{B}}$ which we proved in Theorem 1.

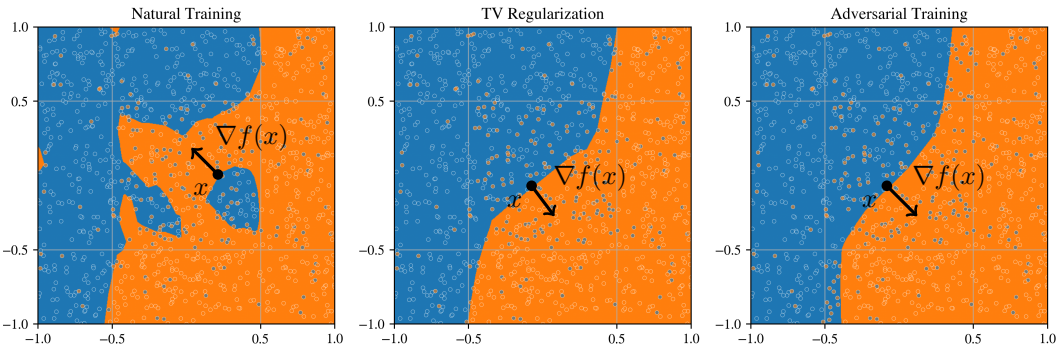


Figure 5: Decision boundaries, using ℓ_2-attacks instead of ℓ_∞-attacks: As expected, no gradient sparsity in the undetermined region $[-0.5, 0.5]^2$ is observed, but the decision boundaries of TV regularization and adversarial training still have much lower complexity and are shorter than the one of natural training.

CIFAR-2 and Two-Layer Neural Networks. We provide additional numerical illustrations obtained by training two layer neural networks on two classes extracted from the CIFAR-10 dataset. Specifically, we trained the two-layer neural network architecture from (1) with $4,096$ hidden neurons on the subset of CIFAR-10 that contains only the classes *frog* and *ship*. We do this to reduce the problem to a binary classification problem, precisely as analyzed in Theorem 1. As for the toy example, we utilize three different training methods: natural training, TV regularization, and adversarial training.

The results are displayed in Figure 6 where we compare the achieved accuracy for clean and adversarially perturbed inputs and Figure 7 for a thresholded sparsity comparison of the input gradients. It can be observed that adversarial training, as well as TV regularization both result in better robustness and higher thresholded gradient sparsity, with the latter being even more affected by TV regularization. Both methods are comparable in term of adversarial robustness and yield higher robustness against the considered attacks than the naturally trained model.

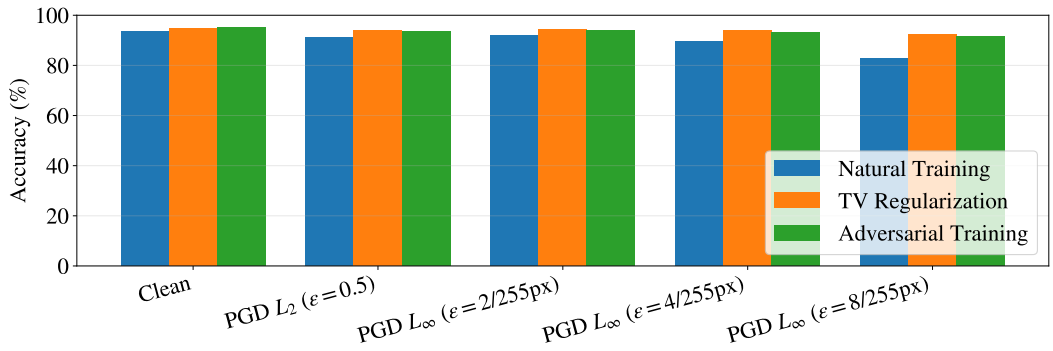


Figure 6: Clean and adversarial accuracies of the two-layer neural network models trained on two classes of CIFAR-10. We used $n = 10,000$, $m = 4,096$ neurons, $\varepsilon_n = 0.0587$, and $\lambda_n = 0.0055 \cdot \varepsilon_n$.

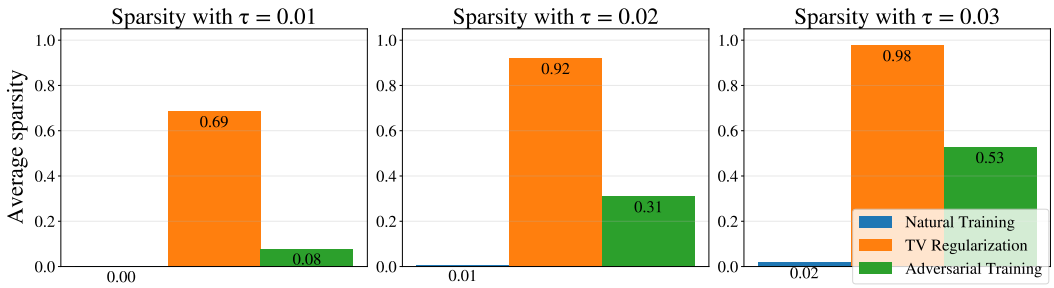


Figure 7: Sparsity of the saliency maps of the two-layer models. Larger value of the bar indicates more values below the threshold τ.

ImageNet. In addition to the thresholded sparsity comparison in Figure 3, we provide a histogram of the input gradient magnitudes to further analyze the networks' sensitivities. Unlike the spatial visualizations in Figure 1, here the gradient magnitude is the absolute gradient value of every individual color channel. Using the pre-trained models described in Section 6, Figure 8 allows for a direct comparison of the raw gradient distributions between a naturally trained ResNet50 and its adversarially trained counterpart ($\varepsilon = 1/255$) over a subset of the ImageNet validation dataset. As already expected from the thresholded sparsity results, the adversarially trained model's gradient values are much more concentrated around zero than those of the naturally trained model. In particular, this histograms reveals that the gradients are not strictly sparse, meaning that very few of the components are exactly zero, but rather that the majority of the gradient components have magnitudes extremely close to zero.

Although our theory gives explanations for the apparent sparsity of saliency maps when computing the gradient of the true class component (see our discussion on the multiclass case in Section 3) the empirical behavior of trained models for incorrect target classes remains of interest. To investigate this, we take the image from Figure 1 (true label: *hummingbird*) and compute the gradients for two different classes. Figure 9 visualizes these gradients with respect to the classes *goldfish* (top row) and *tennis ball* (bottom row). Similar to the true-class setting, the naturally trained model produces unstructured noise, while the adversarially trained models generate gradients that are sparser and more structured.

This observation aligns with the discussion of multiclass classification in Section 3 if the class conditionals μ_c for different classes $c \in \{1, \dots, C\}$ sufficient overlap. Otherwise, this remains an interesting topic for future research.

To generate the saliency maps in Figures 1 and 9, we first compute the gradient of the respective class logit with respect to the input. Specifically, we used the true class for Figure 1 and incorrect classes for Figure 9. We then extract the maximum absolute gradient magnitude across channels for

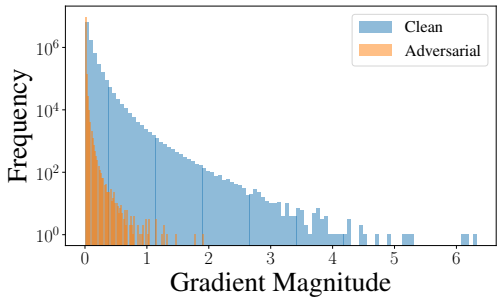
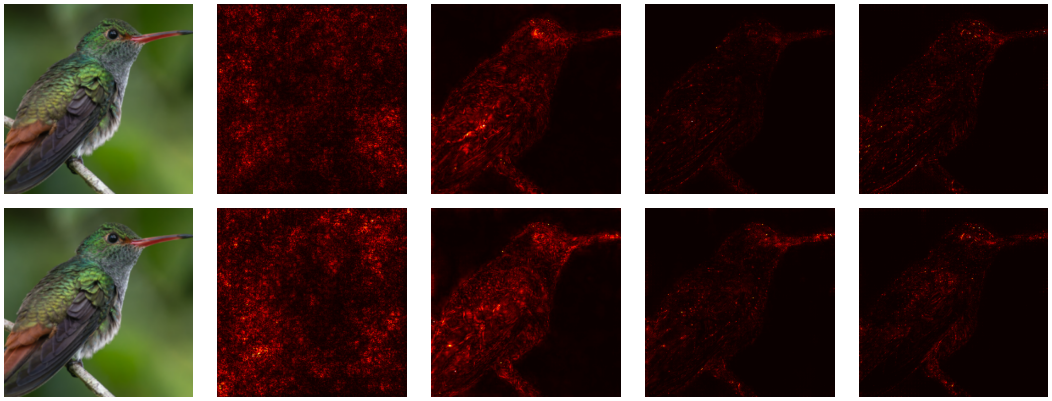


Figure 8: Distribution of input gradient magnitudes for naturally and adversarially trained ResNet50 models on the ImageNet dataset. Notably, the adversarially trained model exhibits a higher frequency of near-zero values.

.

each pixel and applying a global min-max normalization to scale the resulting spatial map into the $[0, 1]$ range for visualization.



(a) Original (b) Natural training (c) $\varepsilon = 0.5/255$ (d) $\varepsilon = 1/255$ (e) $\varepsilon = 2/255$

Figure 9: Off-target saliency maps across varying adversarial budgets. Left: The original image (ground truth: *hummingbird*). Right: Input gradients from a naturally trained ResNet50 and adversarially trained models with increasing budgets. The gradients are computed with respect to incorrect target classes rather than the ground truth: *goldfish* (top row) and *tennis ball* (bottom row).