Lipschitz Bounds and Provably Robust Training by Laplacian Smoothing

Vishaal Krishnan
Mechanical Engineering Department
University of California Riverside
vishaalk@ucr.edu

Abed AlRahman Al Makdah
Electrical & Computer Engineering Department
University of California Riverside
aalmakdah@engr.ucr.edu

Fabio Pasqualetti
Mechanical Engineering Department
University of California Riverside
fabiopas@engr.ucr.edu

Abstract

In this work we propose a graph-based learning framework to train models with provable robustness to adversarial perturbations. In contrast to regularization-based approaches, we formulate the adversarially robust learning problem as one of loss minimization with a Lipschitz constraint, and show that the saddle point of the associated Lagrangian is characterized by a Poisson equation with weighted Laplace operator. Further, the weighting for the Laplace operator is given by the Lagrange multiplier for the Lipschitz constraint, which modulates the sensitivity of the minimizer to perturbations. We then design a provably robust training scheme using graph-based discretization of the input space and a primal-dual algorithm to converge to the Lagrangian's saddle point. Our analysis establishes a novel connection between elliptic operators with constraint-enforced weighting and adversarial learning. We also study the complementary problem of improving the robustness of minimizers with a margin on their loss, formulated as a loss-constrained minimization problem of the Lipschitz constant. We propose a technique to obtain robustified minimizers, and evaluate fundamental Lipschitz lower bounds by approaching Lipschitz constant minimization via a sequence of gradient p-norm minimization problems. Ultimately, our results show that, for a desired nominal performance, there exists a fundamental lower bound on the sensitivity to adversarial perturbations that depends only on the loss function and the data distribution, and that improvements in robustness beyond this bound can only be made at the expense of nominal performance. Our training schemes provably achieve these bounds both under constraints on performance and robustness.

1 Introduction

Sensitivity to adversarial perturbations is one of the main limitations of data-driven models, and a hurdle to their deployment in safety-critical applications. Improving adversarial robustness requires adjusting the worst-case sensitivity of the data-driven input-output map, which is characterized by its Lipschitz constant. Training under a Lipschitz regularization or constraint is therefore a natural way of improving adversarial robustness, which has led to many works on the subject [1, 2]. Yet, a fundamental understanding of the limitations of this approach, as well as a general framework for training models that are provably robust to adversarial perturbations, remain critically lacking.

Preprint. Under review.

Motivated by this need, we consider the problem of adversarially robust learning, formulated as a loss minimization problem with a Lipschitz constraint:

$$\inf_{f \in \mathrm{Lip}(\mathbb{X};\mathbb{Y})} \underbrace{\mathbb{E}_{(x,y)\sim\sigma}\left[\ell\left(f(x), y\right)\right]}_{\triangleq L_\sigma(f)}, \qquad \text{s.t. } \mathrm{lip}(f) \leq \alpha, \tag{1}$$

where $\mathbb{X}$ and $\mathbb{Y}$ are the input and output spaces equipped with distance functions, ℓ is the loss function for the learning problem, σ the data-generating distribution and the search space is the space $\mathrm{Lip}(\mathbb{X};\mathbb{Y})$ of Lipschitz-continuous maps from $\mathbb{X}$ to $\mathbb{Y}$ with an upper bound α on the Lipschitz constant. This class of problems includes, for instance, the problem of image classification with a constraint on the Lipschitz constant of the classifier. In this case, x denotes an image, y a probability vector over the space of labels and σ captures the relation between images and labels. In (1), we do not restrict our attention to any finite-dimensional subspace of $\mathrm{Lip}(\mathbb{X};\mathbb{Y})$, as done when a particular machine learning model is chosen (for instance, neural network, where the dimension of the search space is specified by the network structure). Instead, we focus on the infinite-dimensional learning problem to derive insights and fundamental bounds for the underlying adversarial learning problem. Finally, imposing a hard constraint on the Lipschitz constant (as opposed to a regularization term) allows us to provide hard guarantees on the robustness of the minimizer to adversarial perturbations.

Contributions. In this paper we characterize fundamental robustness bounds for machine learning algorithms, and design provably robust training schemes. Our approach creates, to the best of our knowledge, a novel and useful bridge between the nascent theory of provably robust learning and the classic theories of elliptic operators, partial differential equations, and numerical integration. The technical contributions of this paper are twofold. First, in Section 2 we consider Problem (1) of designing a data-driven map to minimize the loss function, with a desired bound on the map's Lipschitz constant. Under assumptions on strict convexity of the loss function and compactness of the input and output spaces, we show that the problem has a unique minimizer and characterize the saddle point of the corresponding Lagrangian for the problem as the (weak) solution to a Poisson partial differential equation involving a weighted Laplace operator, with the weighting given by the Lagrange multiplier for the constraint. This result provides key insights into the nature of the optimal data-driven map satisfying robustness constraints. We then design a provably robust training scheme based on a graph discretization of the domain to numerically solve for the minimizer of the problem.

Second, we consider the problem of minimizing the Lipschitz constant of a data-driven map with a guaranteed bound (margin) on its loss. We show that the Lipschitz constant is tightly and inversely related to the loss, thereby revealing a fundamental tradeoff between the robustness of a data-driven map and its performance. This result implies that the Lipschitz contant of any data-driven algorithm achieving a desired level of performance has a fundamental lower bound that depends only on the loss function ℓ and the data-generating distribution σ, which constitutes a fundamental lower bound to benchmark any training algorithm and learning problem. We also provide a training scheme for further improving the robustness of a minimizer with a margin on the loss, by using a graph-based iterative procedure that involves solving a series of p-Poisson equations, decsribed in Section 3.

Related work. Motivated by real-world incidents and empirical studies [3], the issue of robustness of data-driven models to adversarial perturbations has received extensive attention in the last years [4–7]. When perturbations are chosen carefully, early studies [8] have shown that small input variations can cause large prediction errors in otherwise highly accurate neural networks. Several frameworks exist to design robust data-driven models, including regularization [1], adversarial training [9], distributionally robust optimization [10] and training under Lipschitz constraints. Of the above, the latter approach is particularly attractive, as it results in trained models with certified robustness.

The study of robustness of the class of neural network models has particularly drawn a lot of attention [11–17]. Many works [18–21] explore, in particular, the problem of training networks with Lipschitz constraints, and related issues. The complementary problem of estimating the Lipschitz constant of a trained neural network is also a crucial part of providing robustness certificates for trained models, and avoiding the danger of deploying unsafe models under a false sense of security. Recent works [22–24] have focused on deriving upper bounds on the Lipschitz constant of neural networks. While these certificates and training schemes provide a way of estimating and improving robustness of a certain class of data-driven models, they fall short in providing insight into the fundamental robustness bounds for the underlying learning problem and the means to exploit them in design.

Furthermore, recent works also point towards fundamental tradeoffs between accuracy and robustness of data-driven models [25–28] in various settings and training frameworks. The connection of adversarial robustness to model complexity and generalization, and the existence (or non-existence) of fundamental tradeoffs between them is another important problem that has received attention [29–34], and is the subject of ongoing debate. This paper builds and extends upon these early studies.

Notation. We introduce here some useful notation. We use $|\cdot|$ to denote the Euclidean norm in $\mathbb{R}^d$, for any $d \in \mathbb{N}$ (when $d = 1$, this denotes the absolute value) and more generally the Hilbert-Schmidt (H-S) norm in finite dimensions. We use $\|\cdot\|$ for function space norms. For maps f between high-dimensional spaces, we often require the notation $\| \, |f| \, \|$, which specifies the function space norm of $|f|$ (which is in turn the function that evaluates to the H-S norm of the map f at any point in its domain). For $\mathbb{X} \subset \mathbb{R}^{\dim(\mathbb{X})}$, we denote by $(\mathbb{X}, \mu)$ the set $\mathbb{X}$ with an underlying measure μ. We denote by $\mathcal{F}(\mathbb{X}; \mathbb{Y})$ a class $\mathcal{F}$ (placeholder for the particular spaces mentioned below) of maps from $\mathbb{X}$ to $\mathbb{Y}$. We denote by $L^p(\mathbb{X}, \mu)$ the space of p-integrable (measurable) functions on $\mathbb{X}$, where the integration is carried out with the underlying measure μ (the Lebesgue measure is implied when μ is not specified), and by $W^{1,p}(\mathbb{X}, \mu)$ the space of p-integrable (measurable) functions with p-integrable (measurable) derivatives. When generalized to the space of maps, as in $f \in L^p((\mathbb{X}, \mu); \mathbb{Y})$, we mean $|f| \in L^p(\mathbb{X}, \mu)$. Also, for $f \in W^p((\mathbb{X}, \mu); \mathbb{Y})$, we mean $|f| \in L^p(\mathbb{X}, \mu)$ and $|\nabla f| \in L^p(\mathbb{X}, \mu)$.

2 Lipschitz-constrained loss minimization and provably robust training

In this section we study and solve the Lipschitz constrained loss minimization problem (1). We start by specifying the setting for Problem (1). Let $\mathbb{X} \subset \mathbb{R}^{\dim(\mathbb{X})}$ and $\mathbb{Y} \subset \mathbb{R}^{\dim(\mathbb{Y})}$ be convex and compact, σ an absolutely continuous probability measure on $\mathbb{X} \times \mathbb{Y}$ with (absolutely continuous) marginal μ supported on $\mathbb{X}$ and conditional π. Let the loss function $\ell : \mathbb{Y} \times \mathbb{Y} \to \mathbb{R}_{\geq 0}$ be strictly convex and Lipschitz continuous. The Lipschitz constraint on the maps in (1) is a global constraint involving every pair of points in the domain $\mathbb{X}$. To obtain a tractable formulation, we equivalently rewrite the Lipschitz constraint as a bound on the norm of the gradient in the domain $\mathbb{X}$. The space of Lipschitz continuous maps $\mathrm{Lip}(\mathbb{X}; \mathbb{Y})$ is also the Sobolev space $W^{1,\infty}((\mathbb{X}, \mu); \mathbb{Y})$ of essentially bounded (measurable) maps with essentially bounded (measurable) gradients, that is, $\mathrm{Lip}(\mathbb{X}; \mathbb{Y}) = W^{1,\infty}((\mathbb{X}, \mu); \mathbb{Y})$.[1] The Lipschitz constant of a map $f \in \mathrm{Lip}(\mathbb{X}; \mathbb{Y})$ is $\mathrm{lip}(f) = \| |\nabla f| \|_{L^\infty((\mathbb{X},\mu);\mathbb{Y})}$ (the $W^{1,\infty}$-seminorm of f). We refer the reader to our supplementary material or [35] for a discussion of these notions.

Using the above definitions, the Lipschitz constrained loss minimization problem (1) becomes

$$\inf_{f \in W^{1,\infty}((\mathbb{X},\mu);\mathbb{Y})} \left\{ L_\sigma(f), \qquad \text{s.t.} \;\; \| |\nabla f| \|_{L^\infty(\mathbb{X},\mu)} \leq \alpha \right\}. \tag{2}$$

To see the role of the Lipschitz constant in the sensitivity of the loss to adversarial perturbations, first notice that adversarial perturbations can be written as the perturbations on the joint distribution σ generated by a map T that perturbs the inputs $x \in \mathbb{X}$ while preserving the outputs $y \in \mathbb{Y}$ [8]. In compact form, the class of adversarial perturbations can be written as:

$$\mathcal{T} = \{ T \;|\; T(x, y) = (T_1(x, y) \;,\; y), \;\text{s.t.}\;\; T_1(x, y) \in B_\delta(x) \cap \mathbb{X} \},$$

where $B_\delta(x)$ is the open ball in $\mathbb{R}^{\dim(\mathbb{X})}$ of radius $\delta > 0$ and centered at x. Defining the sensitivity as the worst-case increase of the loss L_σ following an adversarial perturbation $T \in \mathcal{T}$ for any σ, we get[2] that it is modulated by L^∞-norm of the gradient $\nabla_1 \ell \cdot \nabla f$ (precisely, $\| |\nabla_1 \ell \cdot \nabla f| \|_{L^\infty(\mathbb{X}\times\mathbb{Y},\sigma)}$)[3] and whose upper bound is determined by the Lipschitz constant:

$$\underbrace{\| |\nabla_1 \ell \cdot \nabla f| \|_{L^\infty(\mathbb{X}\times\mathbb{Y},\sigma)}}_{\text{sensitivity of } L \text{ to adv. perturbation}} \leq \underbrace{\| |\nabla_1 \ell| \|_{L^\infty(\mathbb{X}\times\mathbb{Y},\sigma)}}_{\text{Lipschitz constant of } \ell} \cdot \underbrace{\| |\nabla f| \|_{L^\infty(\mathbb{X},\mu)}}_{\text{Lipschitz constant of } f}. \tag{3}$$

Problem 2 is convex (owing to the strict convexity of the loss L_σ[4] and the convexity of the constraint). Thus, we can expect to obtain a (unique) minimizer from the saddle point of the corresponding Lagrangian. With $G_f(x) = \frac{1}{2}\left(|\nabla f(x)|^2 - \alpha^2\right)$, we can reformulate the Lipschitz constraint as

[1] We let μ be the underlying measure on $\mathbb{X}$, since the input data is generated from μ on the support $\mathbb{X}$.

[2] See Supplementary Material for a proof.

[3] We use $\nabla_1 \ell$ to denote the gradient of ℓ with respect to its first argument.

[4] See supplementary material for a proof.

$G_f \leq 0$ μ–a.e. in $\mathbb{X}$[5]. Since $f \in W^{1,\infty}((\mathbb{X},\mu);\mathbb{Y})$, the constraint function G_f belongs to the space $L^\infty(\mathbb{X},\mu)$. Correspondingly, the Lagrange multiplier for the constraint $G_f \leq 0$ (μ–a.e. in $\mathbb{X}$) is non-negative[6] and belongs to the dual space of $L^\infty(\mathbb{X},\mu)$, which we denote as $\lambda \in L^\infty(\mathbb{X},\mu)^*_{\geq 0}$. The Lagrangian $\mathcal{L}_\sigma : W^{1,\infty}((\mathbb{X},\mu);\mathbb{Y}) \times L^\infty(\mathbb{X},\mu)^*_{\geq 0}$ for Problem (2) is then given by:

$$\mathcal{L}_\sigma(f,\lambda) = L_\sigma(f) + \lambda\left(G_f\right). \tag{4}$$

Theorem 2.1. ***(Lipschitz constrained loss minimization)*** *Problem (2) has a unique global minimizer $f^* \in W^{1,\infty}((\mathbb{X},\mu);\mathbb{Y})$. The Lagrangian $\mathcal{L}_\sigma$ has a unique saddle point $(f^*,\lambda^*) \in W^{1,\infty}((\mathbb{X},\mu);\mathbb{Y}) \times L^1(\mathbb{X},\mu)_{\geq 0}$. Moreover, (f^*,λ^*) satisfies the first-order optimality conditions:*

1. *Stationarity: The saddle point (f^*,λ^*) is a weak solution of the Poisson equation,*

$$-\frac{1}{\mu}\nabla\cdot\left(\mu\lambda^*\nabla f^*\right) + g_{f^*} = 0 \quad \textit{in } \mathbb{X}, \qquad \mu\lambda^*\nabla f^*\cdot\mathbf{n} = 0 \quad \textit{on } \partial\mathbb{X}, \tag{5}$$

   where $g_{f^*}(x) = \mathbb{E}_{y\sim\pi(y\,|\,x)}\left[\nabla_1 \ell(f^*(x),y)\right]$ and $\mathbf{n}$ is the outward normal to the boundary $\partial\mathbb{X}$.

2. *Feasibility: $|\nabla f^*| \leq \alpha$ and $\lambda^* \geq 0$, $\mu -$ a.e. in $\mathbb{X}$.*

3. *Complementary slackness: $\lambda^*\left(|\nabla f^*| - \alpha\right) = 0$, $\mu -$ a.e. in $\mathbb{X}$.*

Some comments on Theorem 2.1 are in order. In the absence of the constraint in (2) (that is, $\alpha = \infty$), the stationarity condition is characterized by $\mathbb{E}_{y\sim\pi(y\,|\,x)}\left[\nabla_1\ell(f^*_{\text{unc}}(x),y)\right] = 0$, where $f^*_{\text{unc}}(x),y)$ is the unconstrained minimizer of the loss functional. The saddle point of $\mathcal{L}_\sigma$ is characterized by the Poisson equation (5), which encodes the stationarity condition for the Lagrangian. The Neumann boundary condition in (5) results from the fact that we do not enforce a boundary constraint on the map in the loss minimization problem (2). The λ^*-weighted Laplace operator, $\frac{1}{\mu}\nabla\cdot(\mu\lambda^*\nabla)$, is responsible for locally enforcing the Lipschitz constraint and regularizing (smoothing) the minimizer. Moreover, the Lagrange multiplier satisfies $\lambda^* \in L^1(\mathbb{X},\mu)_{\geq 0}$, and is therefore integrable (this is stronger regularity than in the definition $\lambda \in L^\infty(\mathbb{X},\mu)^*_{\geq 0}$). It follows from the feasibility condition in Theorem 2.1 that the minimizer (provably) satisfies the Lipschitz bound (in contrast to Lipschitz regularization-based approaches to adversarial learning). From the complementary slackness condition in Theorem 2.1, smoothing is enforced only when the constraint is active: when the constraint is inactive in a region $D \subset \mathbb{X}$ of non-zero measure (that is, $|\nabla f^*(x)| < \alpha$ for $x \in D$ and $\mu(D) > 0$), the Lagrange multiplier satisfies $\lambda^* = 0$ (μ-a.e. in D) and smoothing is not enforced.

The fact that the saddle point of the Lagrangian $\mathcal{L}_\sigma$ in (4) satisfies the Lipschitz bound forms the basis for the design of a provably robust training scheme, which we obtain through a discretization of Problem (2) over a graph. To this end, we select n points $\{X_i\}_{i=1}^n$, $X_i \in \mathbb{X}$, via i.i.d. sampling of the distribution μ (in practice, we sample uniformly i.i.d. from the input dataset, that defines the empirical marginal measure $\widehat{\mu}$). With the discretization points $\{X_i\}_{i=1}^n$ as the (embedding of) vertices, we construct an undirected, weighted, connected graph $\mathcal{G} = (\mathcal{V},\mathcal{E},W)$, with vertex set $\mathcal{V} = \{1,\ldots,n\}$, edge set $\mathcal{E} = \mathcal{V}\times\mathcal{V}$, and weighted adjacency matrix $W = [w_{ij}]_{i,j=1}^n$.

We assume the availability of a labeled dataset $D = \{(x_i,y_i)\}_{i=1}^N$ consisting of $N > n$ i.i.d. samples of σ, and define a partition $\mathcal{W} = \{\mathcal{W}_i\}_{i=1}^n$ of the dataset D as follows:

$$\mathcal{W}_i = \{(x,y) \in D \mid |x - X_i| \leq |x - X_j| \ \forall\, j \in \mathcal{V}\setminus\{i\}\}. \tag{6}$$

We then assign weights $\theta_{ij} = N^{-1}$ to the samples $\xi_j = (x_j,y_j) \in \mathcal{W}_i$ (a different weighing scheme may affect generalization and performance of our model; we leave this for future research). Finally, we write the discrete (empirical) Lipschitz constrained loss minimization problem over the graph $\mathcal{G}$ as follows (this minimization problem can be viewed as the discretized version of (2) over $\mathcal{G}$):

$$\min_{\substack{\mathbf{v}=(v_1,\ldots,v_n)\\ v_i\in\mathbb{R}^{\dim(\mathbb{Y})}}} \left\{\sum_{i\in\mathcal{V}}\left(\sum_{j\in\mathcal{W}_i}\theta_{ij}\ell(v_i,y_j)\right), \quad \text{s.t.} \quad |v_r - v_s| \leq \alpha\,|X_r - X_s|\,, \quad \forall\,(r,s)\in\mathcal{E}\right\}. \tag{7}$$

[5] The constraint violation set is of zero measure, that is, $\mu\left(\{x \in \mathbb{X} \mid G_f(x) > 0\}\right) = 0$.

[6] Any $\lambda \in L^\infty(\mathbb{X},\mu)^*$ is also a bounded, finitely additive (absolutely continuous) measure on $\mathbb{X}$.

We note that the above constrained minimization problem (7) is convex (strictly convex objective function with convex constraints) and the corresponding Lagrangian is given by:

$$\mathcal{L}_{\mathcal{G}}(\mathbf{v}, \Lambda) = \sum_{i \in \mathcal{V}} \left[\sum_{s \in \mathcal{W}_i} \theta_{is} \ell(v_i, y_s) + \frac{1}{2} \sum_{j \in \mathcal{V}} \lambda_{ij} w_{ij} \left(|v_i - v_j|^2 - \alpha \, |X_i - X_j|^2 \right) \right], \tag{8}$$

where $\Lambda = [\lambda_{ij}]_{i,j=1}^n$ is the matrix of Lagrange multiplier for the pairwise Lipschitz constraints. Define a primal-dual dynamics for the Lagrangian $\mathcal{L}_{\mathcal{G}}(\mathbf{v}, \Lambda)$ with time-step sequence $\{h(k)\}_{k \in \mathbb{N}}$:

$$\begin{aligned} \mathbf{v}(k+1) &= \mathbf{v}(k) - h(k) \, \nabla_{\mathbf{v}} \mathcal{L}_{\mathcal{G}} \left(\mathbf{v}(k), \Lambda(k) \right), \\ \Lambda(k+1) &= \max\{0 \, , \, \Lambda(k) + h(k) \, \nabla_{\Lambda} \mathcal{L}_{\mathcal{G}} \left(\mathbf{v}(k), \Lambda(k) \right) \}. \end{aligned} \tag{9}$$

The primal dynamics is a discretized heat flow over the graph $\mathcal{G}$ with a weighted Laplacian, where $\nabla_{\mathbf{v}} \mathcal{L}_{\mathcal{G}} \left(\mathbf{v}(k), \Lambda(k) \right) = \left(\Delta(\Lambda, W) \otimes I_{\dim(\mathbb{Y})} \right) \mathbf{v} + \theta \cdot \nabla_1 \ell(\mathbf{v}, \mathbf{y})$, and $\Delta(\Lambda, W)$ is the $\Lambda \circ W$-weighted Laplacian of the graph $\mathcal{G}$ (where $\circ$ denotes the Hadamard or entry-wise product of matrices). The convergence of the solution $\{(\mathbf{v}(k), \Lambda(k))\}_{k \in \mathbb{N}}$ of the primal-dual dynamics (9) to the saddle point of the Lagrangian $\mathcal{L}_{\mathcal{G}}$ follows [36] from the convexity of Problem (7).

As the size of the dataset N and the size of graph n increase, the solution to Problem (7) approaches the solution to Problem (2), under certain mild conditions. In particular, by the Glivenko-Cantelli Theorem [37], the empirical measure $\widehat{\sigma}_N = \frac{1}{N} \sum_{i=1}^N \delta_{(x_i, y_i)}$ converges uniformly and almost surely to the distribution σ in the limit for $N \to \infty$, and so does $\widehat{\mu}_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i} \to \mu$ as $n \to \infty$, where δ here denotes the Dirac measure. Further, the convergence as $n \to \infty$ (higher model complexity) and $N \to \infty$ (larger dataset) of the minimizer of the (empirical) discrete minimization problem (7) to the infinite-dimensional problem (2) is modulated by the weights θ (which govern the convergence of the empirical loss) and w (which governs the convergence of the graph Laplacian to the Laplace operator on the domain [38]).

We conclude this section with an illustrative example. Consider a dataset of 10000 i.i.d. samples (x_i, y_i), with $x_i \in [0, 1]^2$ and $y_i \in \{[1 \ 0]^\mathsf{T}, [0 \ 1]^\mathsf{T}\}$, taken uniformly from the distribution σ in Fig. 1(a), where $y_i = [1 \ 0]^\mathsf{T}$ if x_i belongs to a white cell and $y_i = [0 \ 1]^\mathsf{T}$ if x_i belongs to a black cell. We randomly select n nodes in $[0, 1]^2$, with $n = \{125, 200, 500\}$, construct a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ by connecting each node to its 10 nearest neighbors, and compute the solution $\mathbf{v}^*$ to (9) for different values of the Lipschitz constant α. Then, we generate a testing set of 2000 i.i.d. samples from σ, associate them with the closest node, and evaluate the classification confidence of $\mathbf{v}^*$. In particular, if the testing sample $\bar{x}_i$ is closest to the i-th node and $v_i^* = [p_1 \ p_2]^\mathsf{T}$, then $\bar{x}_i$ is classified as $[1 \ 0]^\mathsf{T}$ with confidence p_1 if $p_1 > p_2$, and as $[0 \ 1]^\mathsf{T}$ with confidence $1 - p_1$ if $p_1 < p_2$. Fig. 1(b)-(h) shows the Voronoi cells associated with the nodes $\mathcal{V}$, where each cell is colored on a gray scale using the first entries of v_i^* (darker colors indicate higher confidence in classifying the samples in a cell as $[0 \ 1]^\mathsf{T}$, while lighter colors indicate higher confidence in classifying the samples in a cell as $[1 \ 0]^\mathsf{T}$). It can be seen that the classification confidence increases with the number of nodes and the Lipschitz bound, at the expenses of a higher model complexity and sensitivity to adversarial perturbations. This trend is also visible in Fig.1(i), where the classification confidence increases with the Lipschitz bound until it saturates for the classifier with highest confidence given the training set and discretization points.

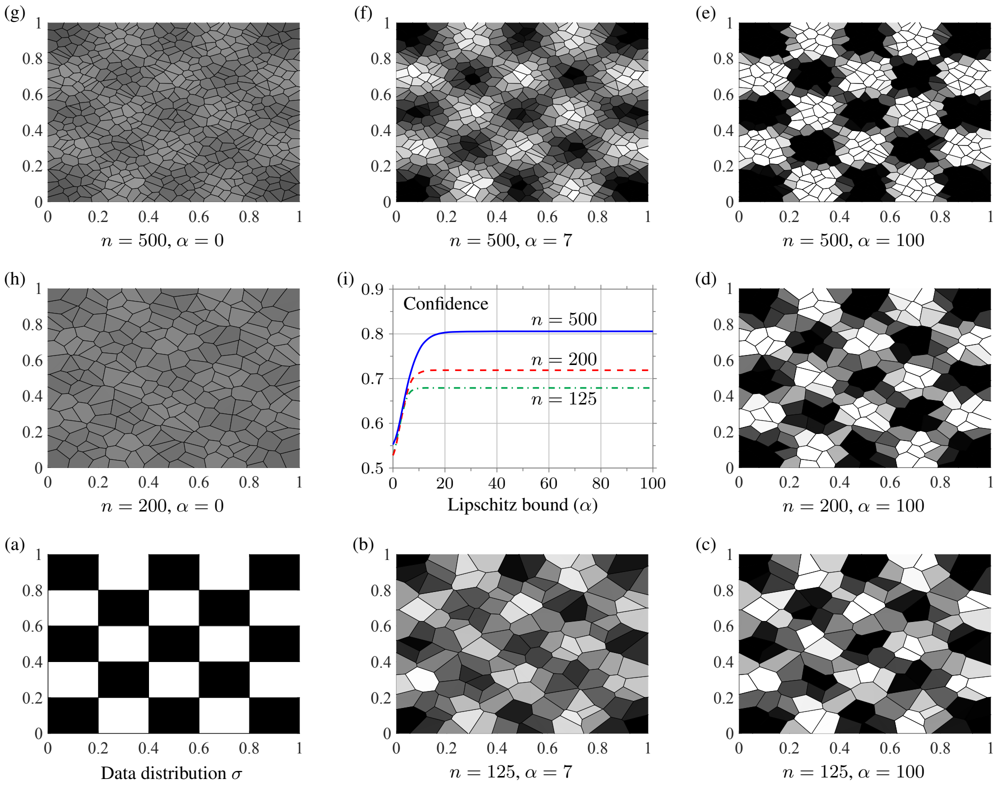
3 Robustification with loss margin and fundamental bound

In this section we study the problem of increasing the robustness of a minimizer with a margin on the loss. Let f^* be the minimizer of (1) with Lipschitz bound α, and let $J_\sigma^*(\alpha)$ be the optimal loss. We formulate and solve the following loss constrained Lipschitz constant minimization problem:

$$\inf_{f \in W^{1,\infty}((\mathbb{X},\mu);\mathbb{Y})} \left\{ \| |\nabla f| \|_{L^\infty((\mathbb{X},\mu);\mathbb{Y})}, \qquad \text{s.t. } L_\sigma(f) \leq J_\sigma^*(\alpha) + \epsilon \right\}. \tag{10}$$

Because the Lipschitz constant satisfies $\| |\nabla f| \|_{L^\infty((\mathbb{X},\mu);\mathbb{Y})} = \operatorname{ess\,sup} |\nabla f|$, Problem (10) has a $\min - \max$ (more precisely, an $\inf - \operatorname{ess\,sup}$) structure which is not amenable to tractable numerical schemes. We circumvent this hurdle by approaching problem (10) via a sequence of loss-constrained (convex) minimization problems involving the $W^{1,p}$-seminorm, for $p \in \mathbb{N}$, $p > 1$, given by:

$$\inf_{f \in W^{1,p}((\mathbb{X},\mu);\mathbb{Y})} \left\{ \| |\nabla f| \|_{L^p(\mathbb{X},\mu)}, \qquad \text{s.t. } L_\sigma(f) \leq J_\sigma^*(\alpha) + \epsilon \right\}. \tag{11}$$



$n = 500, \alpha = 0$

$n = 500, \alpha = 7$

$n = 500, \alpha = 100$

$n = 200, \alpha = 0$

$n = 200, \alpha = 100$

Data distribution σ

$n = 125, \alpha = 7$

$n = 125, \alpha = 100$

Figure 1: For the classification problem discussed in Section 2, this figure shows a tradeoff between the confidence of classification, the Lipschitz constant, and the complexity of the classifier designed using our algorithm (9). Increasing the Lipschitz constant of the classifier and its complexity also increases the confidence of classification, at the expenses of a higher sensitivity to perturbations.

$W^{1,p}$-seminorm minimization problems are typically formulated to obtain minimum Lipschitz extensions in semi-supervised learning [39–42]. A related problem is the one of $W^{1,p}$-seminorm regularized learning [43, 44]. Instead, we propose this approach, for the first time, to improve the robustness of minimizers to adversarial perturbations with a guaranteed margin on the loss.

Convexity of Problem (11) follows from the convexity of the $W^{1,p}$-seminorm in $W^{1,p}((\mathbb{X},\mu);\mathbb{Y})$ and the strict convexity of L_σ (which yields a convex constraint). The minimizers are obtained from the saddle points of the Lagrangian $\mathcal{H}^p_\sigma : W^{1,p}((\mathbb{X},\mu);\mathbb{Y}) \times \mathbb{R}_{\geq 0} \to \mathbb{R}$ for Problem (11), given by:

$$\mathcal{H}^p_\sigma(f,\kappa) = \frac{1}{p}\, \||\nabla f|\|^p_{L^p(\mathbb{X},\mu)} + \kappa \left(L_\sigma(f) - (J^*_\sigma(\alpha) + \epsilon)\right), \tag{12}$$

where we (equivalently) consider the p-th exponent $\||\nabla f|\|^p_{L^p(\mathbb{X},\mu)}$ of the $W^{1,p}$-seminorm in defining the Lagrangian. The saddle points of $\mathcal{H}^p_\sigma$ are now specified by a Poisson equation involving the p-Laplace operator,[7] as established in the following theorem:

Theorem 3.1. ***(Loss constrained $W^{1,p}$-seminorm minimization)*** *For every $p \in \mathbb{N}_{>1}$, there exists a global minimizer $f^{\epsilon,p} \in W^{1,p}((\mathbb{X},\mu);\mathbb{Y})$ for Problem (11). Also, there exists a saddle point $(f^{\epsilon,p},\kappa^{\epsilon,p}) \in W^{1,p}((\mathbb{X},\mu);\mathbb{Y}) \times \mathbb{R}_{\geq 0}$ of the Lagrangian $\mathcal{H}^p_\sigma$. Moreover, $(u,\kappa) \in W^{1,p}((\mathbb{X},\mu);\mathbb{Y}) \times \mathbb{R}_{\geq 0}$ is a saddle point of $\mathcal{H}^p_\sigma$ if and only if it satisfies the following first-order optimality conditions:*

1. *Stationarity: (u,κ) is a (weak) solution of the p-Poisson equation:*

$$-\Delta^\mu_p u + \kappa g_u = 0 \ \text{ in } \mathbb{X}, \qquad \mu \nabla u \cdot \mathbf{n} = 0 \ \text{ on } \partial\mathbb{X}, \tag{13}$$

where $g_u(x) = \mathbb{E}_{y\sim\pi(y \mid x)}\left[\nabla_1 \ell(u(x),y)\right]$ and Δ^μ_p is the p-Laplace operator on $(\mathbb{X},\mu)$.

[7] The p-Laplace operator is defined as $\Delta^\mu_p u = \frac{1}{\mu}\nabla \cdot \left(\mu |\nabla u|^{p-2} \nabla u\right)$.

2. *Feasibility:* $L_\sigma(u) \leq J^*_\sigma(\alpha) + \epsilon$ *and* $\kappa \geq 0$.

3. *Complementary slackness:* $\kappa \left(L_\sigma(f) - (J^*_\sigma(\alpha) + \epsilon)\right) = 0$.

With the characterization of the minimizers of (11) for every $p \in \mathbb{N},\ p > 1$ from Theorem 3.1, we now investigate whether the minimum value of (11) and its minimizers converge (as $p \to \infty$) to those of (10). The following theorem establishes that this is indeed the case, and that the minimum Lipschitz constant in (10) can be obtained as the limit of the sequence of minimum values of (11).

Theorem 3.2. ***(Limit as*** $p \to \infty$ ***and fundamental Lipschitz lower bound)*** *For any* $\epsilon > 0$*, it holds*

$$\lim_{p\to\infty} \min_{\substack{f\in W^{1,p}((\mathbb{X},\mu);\mathbb{Y})\\ L_\sigma(f)\leq J^*_\sigma(\alpha)+\epsilon}} |||\nabla f|||_{L^p(\mathbb{X},\mu)} = \min_{\substack{f\in W^{1,\infty}((\mathbb{X},\mu);\mathbb{Y})\\ L_\sigma(f)\leq J^*_\sigma(\alpha)+\epsilon}} |||\nabla f|||_{L^\infty((\mathbb{X},\mu);\mathbb{Y})}.$$

Moreover, as $p \to \infty$*, the sequence* $\{f^{\epsilon,p}\}_{p\in\mathbb{N}_{>1}}$ *of minimizers of Problem* (11) *converges uniformly to a (global) minimizer* $f^{\epsilon,\infty}$ *of* (10).

The facts that the saddle points of $\mathcal{H}^p_\sigma$ in (12) satisfy the bound on the loss (for every $p \in \mathbb{N}_{>1}$) for a given margin $\epsilon > 0$, and that the minimum value and minimizers of (11) converge in the limit $p \to \infty$ to those of (10), form the basis for the design of a robustification scheme. With the same graph structure and dataset partitioning as in Section 2, we write the discrete (empirical) loss-constrained $W^{1,p}$-seminorm minimization problem over the graph $\mathcal{G}$ as follows (this minimization problem can be viewed as the discretized version of (11) over the structure imposed by $\mathcal{G}$):

$$\min_{\substack{\mathbf{v}=(v_1,\dots,v_n)\\ v_i\in\mathbb{R}^{\dim(\mathbb{Y})}}} \left\{ \frac{1}{p}\sum_{i\in\mathcal{V}}\sum_{j\in\mathcal{N}_i} w_{ij}\left|v_i - v_j\right|^p, \qquad \text{s.t.} \sum_{i\in\mathcal{V}}\sum_{s\in\mathcal{W}_i} \theta_{is}\ell(v_i, y_s) \leq J^*_\sigma(\alpha) + \epsilon \right\}. \quad (14)$$

We note that the above constrained minimization problem (14) is convex (convex objective function with convex constraints), and that the corresponding Lagrangian is given by:

$$\mathcal{H}^p_\mathcal{G}(\mathbf{v},\kappa) = \sum_{i\in\mathcal{V}} \left[\frac{1}{p}\sum_{j\in\mathcal{N}_i} w_{ij}\left|v_i - v_j\right|^p + \kappa \sum_{s\in\mathcal{W}_i} \left(\theta_{is}\ell(v_i,y_s) - \frac{1}{n}(J^*_\sigma(\alpha)+\epsilon) \right) \right], \quad (15)$$

The saddle points of (15) can be obtained via a primal-dual algorithm similar to (9) in Section 2. We solve the (discrete) loss-constrained Lipschitz minimization problem using an iterative procedure that employs the primal-dual algorithm to converge to a saddle point of $\mathcal{H}^p_\mathcal{G}$ in (15) at every iteration step $p \in \mathbb{N}_{>1}$. We then use the saddle point of $\mathcal{H}^p_\mathcal{G}$ as the initialization for the iteration step $p + 1$.

Theorem 3.1 offers key insights on the fundamental tradeoff between robustness and nominal performance. From complementary slackness in Theorem 3.1, it follows that, for the saddle points $(f^{\epsilon,p}, \kappa^{\epsilon,p})$, either the Lagrange multiplier satisfies $\kappa^{\epsilon,p} = 0$ or the constraint is active ($f^{\epsilon,p}$ occurs at the boundary of the constraint and the loss is $L_\sigma(f^{\epsilon,p}) = J^*(\alpha) + \varepsilon$). If the Lagrange multiplier is zero, then the Poisson equation characterizing the Stationarity condition (13) reduces to the p-Laplace equation with a Neumann boundary condition, whose solution is a constant map (in the weak sense). However, in practically useful cases (for small values of α and ε) with a low optimal loss $J^*(\alpha)$, there will typically not exist a constant map satisfying the loss margin ε (unless the unconstrained minimizer f^*_{unc} is itself flat). This implies that the Lagrange multiplier κ is typically nonzero, that the minimizer $f^{\epsilon,p}$ occurs at the constraint boundary, and that the loss satisfies $L_\sigma((f^{\epsilon,p}) = J^*(\alpha)+\varepsilon$. Therefore, for every $p \in \mathbb{N}_{>1}$, the minimization problem (11) is typically dominated by the constraint, and the minimum value of the $W^{1,p}$-norm decreases monotonically with the loss margin. Thus, a fundamental tradeoff exists between performance and robustness.

We conclude this section with an example. Consider the classification problem described in Section 2. Fig. 2 shows the properties of the minimizers to (14) for varying values of p and ε. It can be seen that, (i) as p increases, the minimum value of (14) converges to its supremum value, which, by Theorem 3.2, is smallest Lipschitz constant for a guaranteed loss margin ε (Fig. 2(a)), and (ii) the minimum Lipschitz constant associated with the loss-constrained minimization problem is a monotonically non-increasing function of the loss margin ε, and strictly decreasing for small values of α and ε (Fig. 2(b)). This curve describes a fundamental tradeoff between adversarial robustness and performance, and is entirely determined by the properties of the classification problem and not

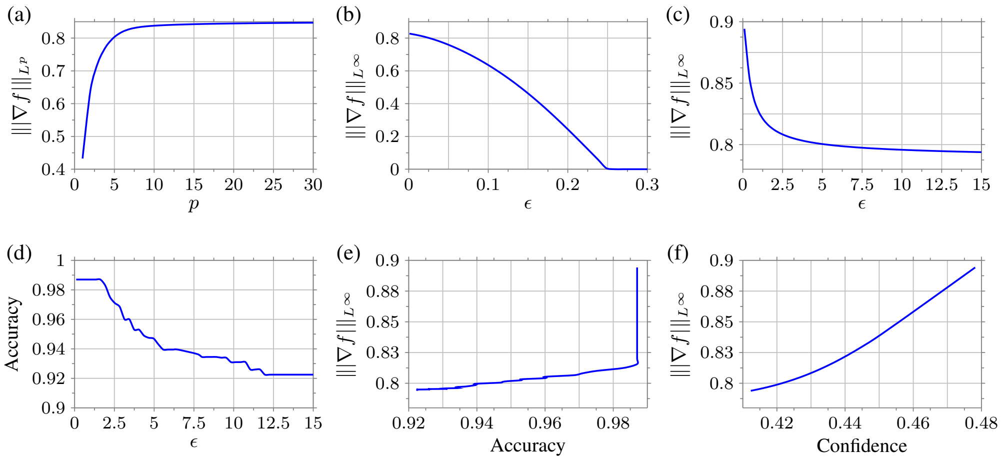


Figure 2: For the classification problem discussed in Section 2, (a) shows the convergence of the minimum values of (14) as p increases to the supremum value, which is the minimum Lipschitz constant. (b) shows the tradeoff between performance and robustness, seen by the monotonic decrease of the minimum Lipschitz constant as a function of the loss margin. Panels (c)-(f) show the relationships between Lipschitz constant, accuracy, and confidence for the standard MNIST dataset.

by the structure of the classifier. In Fig. 2(c)-(f) apply our algorithm to the standard MNIST dataset of handwritten digits [45]. As predicted by our theory, and in accordance with the results obtained in the other numerical example in Fig. 1(i), the classifier's Lipschitz constant (Fig. 2(c)) and accuracy (Fig. 2(d)) are decreasing functions of the classifier's loss margin, while the classification accuracy (Fig. 2(e)) and confidence (Fig. 2(f)) are directly proportional to the classifier's Lipschitz constant. This confirms the existence of a tradeoff between robustness and performance in general learning problems, and provides a limiting benchmark to compare other models and learning schemes.

4 Conclusion

In this paper we propose a novel framework to train models with provable robustness guarantees. At its core, our framework relies on formulating a provably robust learning problem as a (convex) Lipschitz constrained loss minimization problem, for which we characterize and compute the solution by graph-based discretization and discrete heat flows. Our analysis defines a link between the properties of elliptic operators and adversarial learning, which provides us with a new perspective and powerful tools to investigate robustness properties of the minimizers. Following a similar analysis, we also study the complementary problem of improving the robustness of a model under a margin on the loss. We show that the two notions are tightly related, and that improving robustness necessarily leads to the deterioration of the performance of the model (in typical regimes). This robustification problem, which can be solved using an iterative procedure based on discrete heat flows involving the p-Laplacian, leads to the characterization of a fundamental tradeoff between the robustness of a model and its loss, thereby extending and generalizing recent results relating robustness and performance in adversarial machine learning. We illustrate our results via academic and a standard benchmark.

The ideas presented in this paper are of broad interest to the machine learning community and potentially open up a number of research directions. For instance, quantifying the optimality gap of minimizers of (7) evaluated on (2), for finite values of n and N, under different Lipschitz bounds, interpolation, and graph structures, can be used as a formal framework to characterize the underlying fundamental relationships between model complexity, generalization, accuracy, and robustness.

References

[1] H. Gouk, E. Frank, B. Pfahringer, and M. Cree. Regularisation of neural networks by enforcing lipschitz continuity. *arXiv preprint arXiv:1804.04368*, 2018.

[2] C. Finlay, J. Calder, B. Abbasi, and A. Oberman. Lipschitz regularized deep neural networks generalize and are adversarially robust. *arXiv preprint arXiv:1808.09540*, 2018.

[3] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

[4] E. Wong and Z. Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning*, pages 5286–5295, 2018.

[5] A. Raghunathan, J. Steinhardt, and P. Liang. Certified defenses against adversarial examples. In *International Conference on Learning Representations*, Vancouver, Canada, May 2018.

[6] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry. Adversarial examples are not bugs, they are features. *arXiv preprint arXiv:1905.02175*, 2019.

[7] A. Fawzi, S. M. Dezfooli, and P. Frossard. Robustness of classifiers: from adversarial to random noise. In *Advances in Neural Information Processing Systems*, pages 1632–1640, 2016.

[8] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, Banff, Canada, Apr 2014.

[9] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017.

[10] D. Kuhn, P. M. Esfahani, V. A. Nguyen, and S. S. Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations Research & Management Science in the Age of Analytics*, pages 130–166. INFORMS, 2019.

[11] P. L. Bartlett, D. J. Foster, and M. J. Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, pages 6240–6249, 2017.

[12] O. Bastani, Y. Ioannou, L. Lampropoulos, D. Vytiniotis, A. Nori, and A. Criminisi. Measuring neural net robustness with constraints. In *Advances in Neural Information Processing Systems*, pages 2613–2621, 2016.

[13] L. Weng, H. Zhang, H. Chen, Z. Song, C. J. Hsieh, L. Daniel, D. Boning, and I. Dhillon. Towards fast computation of certified robustness for ReLU networks. In *International Conference on Machine Learning*, pages 5276–5285, 2018.

[14] S. Zheng, Y. Song, T. Leung, and I. Goodfellow. Improving the robustness of deep neural networks via stability training. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4480–4488, 2016.

[15] T. W. Weng, H. Zhang, P. Y. Chen, J. Yi, D. Su, Y. Gao, C. J. Hsieh, and L. Daniel. Evaluating the robustness of neural networks: An extreme value theory approach. *arXiv preprint arXiv:1801.10578*, 2018.

[16] H. Zhang, T. W. Weng, P. Y. Chen, C. J. Hsieh, and L. Daniel. Efficient neural network robustness certification with general activation functions. In *Advances in Neural Information Processing Systems*, pages 4939–4948, 2018.

[17] J. Sokolić, R. Giryes, G. Sapiro, and M. Rodrigues. Robust large margin deep neural networks. *IEEE Transactions on Signal Processing*, 65(16):4265–4280, 2017.

[18] P. Pauli, A. Koch, J. Berberich, and F. Allgöwer. Training robust neural networks using Lipschitz bounds. *arXiv preprint arXiv:2005.02929*, 2020.

[19] R. Balan, M. Singh, and D. Zou. Lipschitz properties for deep convolutional networks. *arXiv preprint arXiv:1701.05217*, 2017.

[20] C. Anil, J. Lucas, and R. Grosse. Sorting out Lipschitz function approximation. *arXiv preprint arXiv:1811.05381*, 2018.

[21] Q. Li, S. Haque, C. Anil, J. Lucas, R. Grosse, and J. H. Jacobsen. Preventing gradient attenuation in Lipschitz constrained convolutional networks. In *Advances in Neural Information Processing Systems*, pages 15364–15376, 2019.

[22] A. Virmaux and K. Scaman. Lipschitz regularity of deep neural networks: analysis and efficient estimation. In *Advances in Neural Information Processing Systems*, pages 3835–3844, 2018.

[23] M. Fazlyab, A. Robey, H. Hassani, M. Morari, and G. J. Pappas. Efficient and accurate estimation of Lipschitz constants for deep neural networks. In *Advances in Neural Information Processing Systems*, pages 11423–11434, 2019.

[24] P. L. Combettes and J. C. Pesquet. Lipschitz certificates for neural network structures driven by averaged activation operators. *arXiv preprint arXiv:1903.01014*, 2019.

[25] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, Ernest N. Morial Convention Center, NO, USA, May 2019.

[26] A. Javanmard, M. Soltanolkotabi, and H. Hassani. Precise tradeoffs in adversarial training for linear regression. *arXiv preprint arXiv:2002.10477*, 2020.

[27] A. A. A. Makdah, V. Katewa, and F. Pasqualetti. Accuracy prevents robustness in perception-based control. In *American Control Conference*, Denver, CO, USA, July 2020.

[28] A. A. A. Makdah, V. Katewa, and F. Pasqualetti. A fundamental performance limitation for adversarial classification. *IEEE Control Systems Letters*, 4(1):169–174, 2019.

[29] S. Gui, H. Wang, H. Yang, C. Yu, Z. Wang, and J. Liu. Model compression with adversarial robustness: A unified optimization framework. In *Advances in Neural Information Processing Systems*, pages 1283–1294, 2019.

[30] D. Stutz, M. Hein, and B. Schiele. Disentangling adversarial robustness and generalization. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6976–6987, 2019.

[31] S. Ye, K. Xu, S. Liu, H. Cheng, J. H. Lambrechts, H. Zhang, A. Zhou, K. Ma, Y. Wang, and X. Lin. Adversarial robustness vs. model compression, or both. In *IEEE International Conference on Computer Vision*, volume 2, pages 111–120, 2019.

[32] P. Nakkiran. Adversarial robustness may be at odds with simplicity. *arXiv preprint arXiv:1901.00532*, 2019.

[33] B. Neal, S. Mittal, A. Baratin, V. Tantia, M. Scicluna, S. L. Julien, and I. Mitliagkas. A modern take on the bias-variance tradeoff in neural networks. *arXiv preprint arXiv:1810.08591*, 2018.

[34] C. Louizos, M. Welling, and D. P. Kingma. Learning sparse neural networks through L_0 regularization. *arXiv preprint arXiv:1712.01312*, 2017.

[35] L.C. Evans. *Partial differential equations*. American Mathematical Society, 1998.

[36] K. Arrow, H. Azawa, L. Hurwicz, and H. Uzawa. *Studies in linear and non-linear programming*, volume 2. Stanford University Press, 1958.

[37] P. Billingsley. *Probability and measure*. John Wiley & Sons, 2008.

[38] M. Belkin and P. Niyogi. Towards a theoretical foundation for laplacian-based manifold methods. *Journal of Computer and System Sciences*, 74(8):1289–1308, 2008.

[39] A. El Alaoui, X. Cheng, A. Ramdas, M. Wainwright, and M. I. Jordan. Asymptotic behavior of ℓ_p-based Laplacian regularization in semi-supervised learning. In *Conference on Learning Theory*, pages 879–906, 2016.

[40] R. Kyng, A. Rao, S. Sachdeva, and D. A. Spielman. Algorithms for Lipschitz learning on graphs. In *Conference on Learning Theory*, pages 1190–1223, 2015.

[41] R. K. Ando and T. Zhang. Learning on graph with Laplacian regularization. In *Advances in Neural Information Processing Systems*, pages 25–32, 2007.

[42] J. Calder. Consistency of Lipschitz learning with infinite unlabeled data and finite labeled data. *SIAM Journal on Mathematics of Data Science*, 1(4):780–812, 2019.

[43] A. L. Bertozzi and A. Flenner. Diffuse interface models on graphs for classification of high dimensional data. *Multiscale Modeling and Simulation*, 10(3):1090–1118, 2012.

[44] E. Merkurjev, T. Kostic, and A. L. Bertozzi. An MBO scheme on graphs for classification and image processing. *SIAM Journal on Imaging Sciences*, 6(4):1903–1930, 2013.

[45] Y. LeCun, C. Cortes, and C. J. C. Burges. The MNIST database of handwritten digits. *URL: http://yann.lecun.com/exdb/mnist*, 1998.

[46] W. Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, Inc., 3 edition, 1964.

[47] J. F. Bonnans and A. Shapiro. *Perturbation analysis of optimization problems*. Springer, 2013.

[48] K. Yosida and E. Hewitt. Finitely additive measures. *Transactions of the American Mathematical Society*, 72(1):46–66, 1952.

A Mathematical preliminaries

We introduce some mathematical preliminaries related to function spaces useful in developing our results. In what follows, we let $\mathbb{X} \subset \mathbb{R}^{\dim(\mathbb{X})}$ and $\mathbb{Y} \subset \mathbb{R}^{\dim(\mathbb{Y})}$ be compact and convex.

L^p and $W^{1,p}$ spaces. The space $L^p(\mathbb{X}, \mu)$ of p-integrable functions on $\mathbb{X}$ with respect to an underlying (absolutely continuous) probability measure $\mu \in \mathcal{P}(\mathbb{X})$, is defined as:

$$L^p(\mathbb{X}, \mu) = \left\{ f : \mathbb{X} \to \mathbb{R} \;\middle|\; f \text{ measurable },\; \int_{\mathbb{X}} |f|^p d\mu < \infty \right\}.$$

The Sobolev space $W^{1,p}(\mathbb{X}, \mu)$ is defined as:

$$W^{1,p}(\mathbb{X}, \mu) = \left\{ f \in L^p(\mathbb{X}, \mu) \;\middle|\; \int_{\mathbb{X}} |\nabla f|^p d\mu < \infty \right\}.$$

For $p = \infty$ in the above definitions, we get the space $L^\infty(\mathbb{X}, \mu)$ of essentially bounded measurable functions on $(\mathbb{X}, \mu)$ and the space $W^{1,\infty}(\mathbb{X}, \mu)$ of essentially bounded measurable functions with essentially bounded measurable gradients on $(\mathbb{X}, \mu)$.

Now, for $1 \leq p \leq \infty$, $L^p((\mathbb{X}, \mu); \mathbb{Y})$ is the space of measurable maps from $\mathbb{X}$ to $\mathbb{Y}$ such that $|f| \in L^p(\mathbb{X}, \mu)$ for any $f \in L^p((\mathbb{X}, \mu); \mathbb{Y})$, where $|\cdot|$ is the H-S norm in $\mathbb{Y}$. Moreover, $W^{1,p}((\mathbb{X}, \mu); \mathbb{Y})$ is the space of measurable maps such that $|f| \in L^p(\mathbb{X}, \mu)$ and $|\nabla f| \in L^p(\mathbb{X}, \mu)$ for any $f \in W^{1,p}((\mathbb{X}, \mu); \mathbb{Y})$.

Lipschitz-continuous maps. The space $\mathrm{Lip}(\mathbb{X}; \mathbb{Y})$ of Lipschitz-continuous maps from $\mathbb{X}$ to $\mathbb{Y}$ is such that for any $f \in \mathrm{Lip}(\mathbb{X}; \mathbb{Y})$, we have $|f(x_1) - f(x_2)| \leq \mathrm{lip}(f)\, |x_1 - x_2|$, where $\mathrm{lip}(f)$ is the Lipschitz constant of f. From Rademacher's theorem [35], every $f \in \mathrm{Lip}(\mathbb{X}; \mathbb{Y})$ is almost everywhere differentiable in $\mathbb{X}$ (with (a.e.) gradient ∇f, which is also its weak gradient). Further, $\| |\nabla f| \|_{L^\infty(\mathbb{X})} = \mathrm{lip}(f)$ and we get $\mathrm{Lip}(\mathbb{X}; \mathbb{Y}) = W^{1,\infty}(\mathbb{X}; \mathbb{Y})$.

B Robustness to adversarial perturbations and the Lipschitz constant

In this section, we establish the dependence of sensitivity to adversarial perturbations of the loss on the Lipschitz constant of the input-output map. Recall from (1) that the loss L_σ is given by:

$$L_\sigma(f) = \mathbb{E}_{(x,y)\sim\sigma}\left[\ell(f(x), y)\right].$$

Adversarial perturbations [8] are a subset of perturbations on the data-generating distribution σ generated by bounded maps T that perturb the inputs $x \in \mathbb{X}$ while preserving the outputs $y \in \mathbb{Y}$. We illustrate this for a classification problem: Let (x, y) be a true input-label pair in the (nominal) dataset and f be a classifier that locally assigns to an input $x \in \mathbb{X}$ the label $f(x) \in \mathbb{Y}$. Let r be a minimal perturbation on the input x, given a target label $y' \in \mathbb{Y}$, such that $f(x + r) = y'$ (where y' is typically chosen to be an incorrect label for x, that is, $y' \neq y$). Now, an adversarial perturbation for the classifier f is generated by the replacement of (x, y) by $(x + r, y)$ in the dataset. To formalize this, we define the class of maps:

$$\mathcal{T} = \{T \;|\; T(x, y) = (T_1(x, y)\;,\; y),\; \text{and}\;\; T_1(x, y) \in B_\delta(x) \cap \mathbb{X}\},$$

where $B_\delta(x)$ is the open ball in $\mathbb{R}^{\dim(\mathbb{X})}$ of radius $\delta > 0$ and centered at x. Now, adversarial perturbations on the data-generating distribution σ are a subset of perturbations generated by the class $\mathcal{T}$.

We first characterize the bound on the perturbation of the loss due to perturbations on σ generated by the class $\mathcal{T}$. The perturbation by $T \in \mathcal{T}$ of the probability measure σ yields the perturbed probability measure $T_\#\sigma$, where $T_\#\sigma$ is the pushforward of σ by the map T[8]. We note that the perturbation of

[8] Given a measurable map $T : \mathbb{Z} \to \mathbb{Z}'$ and a probability measure $\sigma \in \mathcal{P}(\mathbb{Z})$, we let $T_\#\sigma$ denote the pushforward of σ by the map T, where for any Borel measurable set $B \subset \mathbb{Z}'$ we have $T_\#\sigma(B) = \sigma(T^{-1}(B))$.

the loss $\left|L_{T_\# \sigma}(f) - L_\sigma(f)\right|$ satisfies:

$$\begin{aligned}
\left|L_{T_\# \sigma}(f) - L_\sigma(f)\right| &= \left|\mathbb{E}_{(x,y)\sim T_\# \sigma}\left[\ell(f(x),y)\right] - \mathbb{E}_{(x,y)\sim \sigma}\left[\ell(f(x),y)\right]\right| \\
&= \left|\int_{\mathbb{X}\times\mathbb{Y}} \ell(f(x),y) d\left(T_\# \sigma\right)(x,y) - \int_{\mathbb{X}\times\mathbb{Y}} \ell(f(x),y) d\sigma(x,y)\right| \\
&= \left|\int_{\mathbb{X}\times\mathbb{Y}} \left(\ell(f(T_1(x,y)),y) - \ell(f(x),y)\right) d\sigma(x,y)\right| \\
&\leq \operatorname{lip}(\ell)\operatorname{lip}(f)\left|\int_{\mathbb{X}} \left(T_1(x,y) - x\right) d\mu(x)\right| \\
&\leq \operatorname{lip}(\ell)\operatorname{lip}(f)\delta.
\end{aligned}$$

We next characterize the sensitivity of the loss for a given f to perturbations on the data-generating distribution generated by the class $\mathcal{T}$. Let a family of transport maps $T^h = (1-h)\operatorname{Id} + hT$ for some $T \in \mathcal{T}$ and $h \in [0,1]$ (with Id being the identity map), perturb the data-generating distribution σ as $\sigma^h = T^h_\# \sigma$. The (Gateaux) derivative of the loss along the family of adversarial perturbations T^h, is now given by:

$$\begin{aligned}
D^{(T)} L_\sigma(f) = \frac{d}{dh} L_{\sigma^h}(f)\bigg|_{h=0} &= \lim_{h\to 0} \frac{L_{\sigma^h}(f) - L_\sigma(f)}{h} \\
&= \lim_{h\to 0} \frac{1}{h} \int_{\mathbb{X}\times\mathbb{Y}} \left[\ell(f(T^h(x,y)),y) - \ell(f(x),y)\right] d\sigma(x,y).
\end{aligned}$$

We note that $\left|\frac{\ell(f(T^h(x,y)),y)-\ell(f(x),y)}{h}\right| \leq \operatorname{lip}(\ell)\left|\frac{f(T^h(x,y))-f(x)}{h}\right| \leq \operatorname{lip}(\ell)\operatorname{lip}(f)\frac{\left|T^h(x,y)-x\right|}{h} = \operatorname{lip}(\ell)\operatorname{lip}(f)\left|T_1(x,y) - x\right|$. It then follows from the Dominated Convergence Theorem [46] that:

$$\begin{aligned}
D^{(T)} L_\sigma(f) &= \int_{\mathbb{X}\times\mathbb{Y}} \langle \nabla_1 \ell(f(x),y) \cdot \nabla f(x)\ ,\ T_1(x,y) - x\rangle \, d\sigma(x,y) \\
&= \mathbb{E}_{(x,y)\sim\sigma}\left[\langle \nabla_1 \ell(f(x),y) \cdot \nabla f(x)\ ,\ T_1(x,y) - x\rangle\right].
\end{aligned}$$

We now define the sensitivity as the worst-case increase of the loss functional following an adversarial perturbation. That is, the sensitivity of the loss is the L^∞-norm (with respect to the measure σ) of the gradient $\nabla_1 \ell \cdot \nabla f$ (precisely, $|||\nabla_1 \ell \cdot \nabla f|||_{L^\infty(\mathbb{X}\times\mathbb{Y},\sigma)}$), which satisfies the bound:

$$\underbrace{|||\nabla_1 \ell \cdot \nabla f|||_{L^\infty(\mathbb{X}\times\mathbb{Y},\sigma)}}_{\text{sensitivity of } L \text{ to adv. perturbation}} \leq \underbrace{|||\nabla_1 \ell|||_{L^\infty(\mathbb{X}\times\mathbb{Y},\sigma)}}_{\text{Lipschitz constant of } \ell} \cdot \underbrace{|||\nabla f|||_{L^\infty(\mathbb{X},\mu)}}_{\text{Lipschitz constant of } f}$$

where μ is the marginal of σ over $\mathbb{X}$, and $|||\nabla f|||_{L^\infty(\mathbb{X},\mu)}$ is the Lipschitz constant of f over the support of μ.

We therefore get that the sensitivity of the loss functional to adversarial perturbations is indeed modulated by the Lipschitz constant of the input-output mapping. Thus, restricting the search space to the class of Lipschitz maps with a bound $\alpha \geq 0$ on the Lipschitz constant, as in the minimization problem (1), is convenient for analysis, and does not restrict the generality of the adversarially robust learning problem, and it allows us to obtain adversarially robust minimizers of the loss L_σ.

C The Lipschitz-constrained loss minimization problem (1) is convex

We recall that Problem (1) is given by:

$$\inf_{f\in \operatorname{Lip}(\mathbb{X},\mu)} \left\{ \underbrace{\mathbb{E}_{(x,y)\sim\sigma}\left[\ell\left(f(x),y\right)\right]}_{\triangleq L_\sigma(f)} \qquad \text{s.t. } \operatorname{lip}(f) \leq \alpha \right\},$$

where σ is an absolutely continuous probability measure on $\mathbb{X}\times\mathbb{Y}$ and the loss function $\ell : \mathbb{Y}\times\mathbb{Y} \to \mathbb{R}_{\geq 0}$ is strictly convex and Lipschitz continuous and $\alpha \geq 0$.

Firstly, we get that the loss L_σ in (1) is strictly convex. To see this, let $f_1, f_2 \in \mathrm{Lip}(\mathbb{X}, \mu)$ be such that $L_\sigma(f_1) < \infty$ and $L_\sigma(f_2) < \infty$. For $t \in [0, 1]$, we get from the convexity of $\mathrm{Lip}(\mathbb{X}, \mu)$ that $tf_1 + (1-t)f_2 \in \mathrm{Lip}(\mathbb{X}, \mu)$. Also, from the strict convexity of the loss function ℓ, we get:

$$\begin{aligned} L_\sigma(tf_1 + (1-t)f_2) &= \mathbb{E}_{(x,y)\sim\sigma}\left[\ell((tf_1 + (1-t)f_2)(x), y)\right] \\ &= \mathbb{E}_{(x,y)\sim\sigma}\left[\ell(tf_1(x) + (1-t)f_2(x), y)\right] \\ &\leq \mathbb{E}_{(x,y)\sim\sigma}\left[t\ell(f_1(x), y) + (1-t)\ell(f_2(x), y)\right] \\ &= t\mathbb{E}_{(x,y)\sim\sigma}\left[\ell(f_1(x), y)\right] + (1-t)\mathbb{E}_{(x,y)\sim\sigma}\left[\ell(f_2(x), y)\right] \\ &= tL_\sigma(f_1) + (1-t)L_\sigma(f_2). \end{aligned}$$

Moreover, the inequality is strict for $t \in (0, 1)$, from which it follows that the loss L_σ is strictly convex.

Now, let $f_1, f_2 \in \mathrm{Lip}(\mathbb{X}, \mu)$ such that $\mathrm{lip}(f_1) \leq \alpha$ and $\mathrm{lip}(f_2) \leq \alpha$. For the map $\lambda f_1 + (1-\lambda)f_2$, $\lambda \in [0, 1]$, and $x_1, x_2 \in \mathbb{X}$, it follows that:

$$\begin{aligned} &|(\lambda f_1 + (1-\lambda)f_2)(x_1) - (\lambda f_1 + (1-\lambda)f_2)(x_2)| \\ &\qquad = |\lambda\,(f_1(x_1) - f_1(x_2)) + (1-\lambda)(f_2(x_1) - f_2(x_2))| \\ &\qquad \leq \lambda\,|f_1(x_1) - f_1(x_2)| + (1-\lambda)\,|f_2(x_1) - f_2(x_2)| \\ &\qquad \leq \lambda \mathrm{lip}(f_1)\,|x_1 - x_2| + (1-\lambda)\mathrm{lip}(f_2)\,|x_1 - x_2| \\ &\qquad \leq \alpha\,|x_1 - x_2|\,, \end{aligned}$$

and we get $\mathrm{lip}(\lambda f_1 + (1-\lambda)f_2) \leq \alpha$. Therefore, the constraint in (1) is convex. From strict convexity of the loss L_σ and convexity of the constraint set $\{f \in \mathrm{Lip}((\mathbb{X}, \mu), \mathbb{Y}) \mid \mathrm{lip}(f) \leq \alpha\}$, we get that Problem (1) is convex.

D Proof of Theorem 2.1 (Saddle point of Lagrangian $\mathcal{L}$)

(i) Derivative of loss function L w.r.t f. We have:

$$L_\sigma(f) = \mathbb{E}_{x\sim\mu}\left[\mathbb{E}_{y\sim\pi(y\,|\,x)}\left[\ell(f(x), y)\right]\right],$$

where μ is the marginal over $\mathbb{X}$ and π the conditional of the joint distribution $\sigma \in \mathcal{P}(\mathbb{X}\times\mathbb{Y})$. Let $\{f^\epsilon\}_{\epsilon\in[0,1]}$ be a family of maps from $\mathbb{X}$ to $\mathbb{Y}$ that is pointwise smooth (i.e., for any $x \in \mathbb{X}$, $F(\epsilon, x) = f^\epsilon(x)$ is smooth in ϵ). We now evaluate the derivative of the loss function L_σ w.r.t. the family $\{f^\epsilon\}_{\epsilon\in[0,1]}$, at $\epsilon = 0$, as follows:

$$\begin{aligned} \frac{dL_\sigma}{d\epsilon}(f^0) &= \lim_{\epsilon\to 0}\frac{L_\sigma(f^\epsilon) - L_\sigma(f^0)}{\epsilon} \\ &= \lim_{\epsilon\to 0}\frac{1}{\epsilon}\int_{\mathbb{X}}\left[\int_{\mathbb{Y}}\left(\ell(f^\epsilon(x), y) - \ell(f^0(x), y)\right) d\pi(y\mid x)\right] d\mu(x). \end{aligned}$$

We note that $\left|\frac{\ell(f^\epsilon(x),y)-\ell(f^0(x),y)}{\epsilon}\right| \leq \mathrm{lip}(\ell)\left|\frac{f^\epsilon(x)-f^0(x)}{\epsilon}\right| \leq \mathrm{lip}(\ell)\mathrm{lip}(F(\cdot, x))$, where $\mathrm{lip}(F(\cdot, x))$ is the Lipschitz constant of F as a function of ϵ at every $x \in \mathbb{X}$ (since $F(\cdot, x)$ is smooth in $[0, 1]$ for every $x \in \mathbb{X}$, it is also Lipschitz continuous). It then follows from the Dominated Convergence Theorem [46] that:

$$\begin{aligned} \frac{dL_\sigma}{d\epsilon}(f^0) &= \lim_{\epsilon\to 0}\frac{1}{\epsilon}\int_{\mathbb{X}}\left[\int_{\mathbb{Y}}\left(\ell(f^\epsilon(x), y) - \ell(f^0(x), y)\right) d\pi(y\mid x)\right] d\mu(x) \\ &= \int_{\mathbb{X}}\left[\int_{\mathbb{Y}}\lim_{\epsilon\to 0}\frac{1}{\epsilon}\left(\ell(f^\epsilon(x), y) - \ell(f^0(x), y)\right) d\pi(y\mid x)\right] d\mu(x) \\ &= \int_{\mathbb{X}}\left[\int_{\mathbb{Y}}\nabla_1\ell(f^0(x), y)\cdot\frac{\partial f^\epsilon}{\partial\epsilon}(x)\bigg|_{\epsilon=0} d\pi(y\mid x)\right] d\mu(x) \\ &= \int_{\mathbb{X}}\left[\int_{\mathbb{Y}}\nabla_1\ell(f^0(x), y)\, d\pi(y\mid x)\right]\cdot\frac{\partial f^\epsilon}{\partial\epsilon}(x)\bigg|_{\epsilon=0} d\mu(x) \\ &= \int_{\mathbb{X}}\frac{\partial\bar{L}}{\partial f}\cdot\frac{\partial f^\epsilon}{\partial\epsilon}\bigg|_{\epsilon=0} d\mu(x), \end{aligned}$$

where we denote by $\partial_f \bar{L}_\sigma = \frac{\partial \bar{L}_\sigma}{\partial f} = \int_{\mathbb{Y}} \nabla_1 \ell(f^0(x), y)\ d\pi(y \mid x)$ the functional derivative of $\bar{L}_\sigma$ w.r.t. f.

(ii) Minimizer of (2). The search space for Problem (2) is given by,

$$\mathcal{F} = \left\{ f \in W^{1,\infty}((\mathbb{X}, \mu), \mathbb{Y}) \mid \| |\nabla f| \|_{L^\infty(\mathbb{X},\mu)} \le \alpha \right\}.$$

We see that $\mathcal{F}$ is closed, convex and bounded. Boundedness of $\mathcal{F}$ follows from compactness of $\mathbb{Y}$ which implies that there exists an $M \in \mathbb{R}_{\ge 0}$ such that $\mathbb{Y} \subset B_M(\mathbf{0}_{\mathbb{Y}})$. It follows that for any $f \in \mathcal{F}$, we have $\| |f| \|_{L^\infty(\mathbb{X},\mu)} \le M$. Moreover, we have $\| |\nabla f| \|_{L^\infty(\mathbb{X},\mu)} \le \alpha$. Therefore, $\|f\|_{W^{1,\infty}((\mathbb{X},\mu),\mathbb{Y})} = \| |f| \|_{L^\infty(\mathbb{X},\mu)} + \| |\nabla f| \|_{L^\infty(\mathbb{X},\mu)} \le M + \alpha < \infty$ for any $f \in \mathcal{F}$.

The loss L_σ is strictly convex and lower semicontinuous (in fact, it is (Gateaux) differentiable as seen earlier for absolutely continuous σ, since ℓ is strictly convex and Lipschitz-continuous).

Let $\{f_n\}_{n\in\mathbb{N}}$ be a minimizing sequence in $\mathcal{F}$ for the loss L_σ, such that $f_n \in \mathcal{F}$ and $\lim_{n\to\infty} L_\sigma(f_n) = \inf_{f\in\mathcal{F}} L_\sigma(f)$. Clearly, the sequence $\{f_n\}_{n\in\mathbb{N}}$ is uniformly bounded since $\|f_n\|_{W^{1,\infty}((\mathbb{X},\mu),\mathbb{Y})} \le M + \alpha$. It is also uniformly equicontinuous, since $|f_n(x_1) - f_n(x_2)| \le \alpha |x_1 - x_2|$ for all $n \in \mathbb{N}$. Therefore, by the Arzelà-Ascoli Theorem [46], there exists a uniformly converging subsequence $\{f_{n_j}\}_{j\in\mathbb{N}}$, with the limit $f^* \in \mathcal{F}$. Furthermore, by the continuity of L_σ, we get $\lim_{j\to\infty} L_\sigma(f_{n_j}) = L_\sigma(f^*) = \min_{f\in\mathcal{F}} L_\sigma(f)$. By the strict convexity of the loss L_σ, we get that f^* is the unique global minimizer of L_σ.

Thus, Problem (2) has a unique global minimizer $f^* \in \left\{ f \in W^{1,\infty}((\mathbb{X},\mu),\mathbb{Y}) \mid \mathrm{lip}(f) \le \alpha \right\}$.

(iii) Saddle points of Lagrangian functional $\mathcal{L}_\sigma$. The constraint set is given by $\{f \in W^{1,\infty}((\mathbb{X},\mu),\mathbb{Y}) \mid \mathcal{G}(f) \in (-\infty, 0]\}$, where $\mathcal{G}(f) = \|G_f\|_{L^\infty(\mathbb{X},\mu)}$, and we have the constraint qualification:

$$0 \in \operatorname{int} \left\{ \mathcal{G}\left(W^{1,\infty}((\mathbb{X},\mu),\mathbb{Y}) \right) + [0,\infty) \right\},$$

where the operation $+$ denotes the Minkowski sum. This allows us to apply Theorem 3.6 in [47] to infer that the set of Lagrange multipliers corresponding to the (unique) minimizer f^* is a non-empty, convex, bounded and weakly$-^*$ compact subset of $L^\infty(\mathbb{X},\mu)^*_{\ge 0}$. Moreover, we note that $(-\infty, 0]$ is a closed convex cone, and it follows from Theorem 3.4-(iii) in [47] that for any Lagrange multiplier λ^*, the pair (f^*, λ^*) is a saddle point of the Lagrangian functional $\mathcal{L}_\sigma$. Uniqueness of λ^* again follows from the strict convexity of L_σ. We also have the *feasibility* condition $G_{f^*} \le 0$ (that is, $|\nabla f^*| \le \alpha$) and $\lambda^* \ge 0$ μ-a.e. in $\mathbb{X}$.

Now, the (Gateaux) derivative of the Lagrangian $\mathcal{L}_\sigma(f, \lambda) = L_\sigma(f) + \lambda(G_f)$ in $W^{1,\infty}((\mathbb{X},\mu),\mathbb{Y})$ along $V \in W^{1,\infty}((\mathbb{X},\mu),\mathbb{Y})$ is given by:

$$D_1^{(V)} \mathcal{L}_\sigma(f, \lambda) = \int_{\mathbb{X}} \partial_f \bar{L}_\sigma \cdot V\ d\mu + \int_{\mathbb{X}} \nabla f \cdot \nabla V\ d(\lambda\mu),$$

where $D_1^{(V)}$ denotes the directional derivative of the first argument along V and $\lambda\mu$ is an absolutely continuous measure (λ-weighting on the underlying measure μ. Recall that $\lambda \in L^\infty(\mathbb{X},\mu)^*_{\ge 0}$ is itself a bounded, finitely additive absolutely continuous measure). The above expression can be derived using a similar construction of a limit and the application of the Dominated Convergence Theorem as earlier in this section.

By the Minimax Theorem, we have $\mathcal{L}_\sigma(f^*, \lambda^*) = \inf_f \sup_\lambda \mathcal{L}_\sigma(f, \lambda) = \sup_\lambda \inf_f \mathcal{L}_\sigma(f, \lambda)$, where the infimum is taken over $W^{1,\infty}((\mathbb{X},\mu),\mathbb{Y})$ and the supremum over $\lambda \in L^\infty(\mathbb{X},\mu)^*_{\ge 0}$. We therefore have $\mathcal{L}_\sigma(f^*, \lambda^*) \ge \mathcal{L}_\sigma(f^*, 0)$, which yields the condition $\lambda^*(G_{f^*}) \ge 0$. Moreover, from feasibility, we have $G_{f^*} \le 0$ and $\lambda^* \ge 0$, which implies that $\lambda^*(G_{f^*}) \le 0$. This results in the *complementary slackness* condition $\lambda^*(G_{f^*}) = 0$. From the Minimax equality, we get that (f^*, λ^*) is also a critical point of $\mathcal{L}_\sigma$, that is, $D_1^{(V)} \mathcal{L}_\sigma(f^*, \lambda^*) = 0$, which implies that $\int_{\mathbb{X}} \partial_f \bar{L}_\sigma(f^*) \cdot V\ d\mu + \int_{\mathbb{X}} \nabla f^* \cdot \nabla V\ d(\lambda^*\mu) = 0$, which is the *stationarity* condition.

(iv) Improved regularity of Lagrange multiplier λ^*. We can indeed establish stronger regularity for the Lagrange multiplier λ^*. We have that the Lagrange multipliers $\lambda^* \in L^\infty(\mathbb{X},\mu)^*_{\ge 0}$, which is a bounded, finitely additive measure absolutely continuous measure, is also a linear continuous functional on $L^\infty(\mathbb{X},\mu)$ and must therefore vanish on sets of μ-measure zero (i.e., $\lambda^*(A) = 0$ for $A \subset \mathbb{X}$ with $\mu(A) = 0$). Moreover, from Theorem 1.24 in [48], we can decompose $\lambda^* = \lambda^*_c + \lambda^*_p$,

where λ_c^* is a non-negative countably additive measure and λ_p^* is non-negative and purely finitely additive. By the Radon-Nikodym theorem, we get that there exists a function $h_c \in L^1(\mathbb{X}, \mu)$ such that the countably additive and absolutely continuous measure λ_c^* satisfies $d\lambda_c^* = h_c \, d\mu$. By substitution in the stationarity condition, we get $\int_{\mathbb{X}} \partial_f \bar{L}_\sigma \cdot V d\mu = -\int_{\mathbb{X}} \nabla f^* \cdot \nabla V \, d(\lambda_c^* \mu) - \int_{\mathbb{X}} \nabla f^* \cdot \nabla V \, d(\lambda_p^* \mu)$. We now consider a set $D_\delta = \{x \in \mathbb{X} \mid -\delta \leq G_{f^*}(x) \leq 0\}$, with $0 < \delta < \alpha^2$. By complementary slackness, we note that $\lambda^*(\mathbb{X} \setminus D_\delta) = 0$. Since λ_p^* is purely finitely additive, it implies that there must exist a collection of nonempty sets $\{E_n\}_{n\in\mathbb{N}}$ with $E_{n+1} \subset E_n$ and $\lim_{n\to\infty} E_n = \emptyset$, such that $\lim_{n\to\infty} \lambda_p^*(E_n) > 0$[9]. Since $\lambda^*(\mathbb{X} \setminus D_\delta) = 0$, we can suppose without loss of generality that $E_0 \subset D_\delta$. We also consider another collection of nonempty sets $\{E_n'\}_{n\in\mathbb{N}}$, with the same properties (with $E_0' \subset D_\delta$, $E_{n+1}' \subset E_n'$ and $\lim_{n\to\infty} E_n' = \emptyset$), such that $E_n \subset E_n'$ for all $n \in \mathbb{N}$. We note that for $x \in D_\delta$, we have $0 < \alpha^2 - \delta \leq |\nabla f^*(x)|^2 \leq \alpha^2$, which implies that ∇f^* does not vanish on E_n' for any $n \in \mathbb{N}$. We now consider a family of variations $V_n \in W^{1,\infty}(\mathbb{X}, \mu)$ for $n \in \mathbb{N}$ such that V_n and ∇V_n are supported in E_n', $\nabla f^* \cdot \nabla V_n \geq 0$ in E_n' and $\nabla f^* \cdot \nabla V_n \geq \epsilon$ in E_n (uniformly). The stationarity condition now yields, for $n \in \mathbb{N}$:

$$\begin{aligned} -\int_{E_n'} \partial_f \bar{L}_\sigma(f^*) \cdot V_n d\mu &= \int_{E_n'} (\nabla f^* \cdot \nabla V_n) \, h_c d\mu + \int_{E_n'} \nabla f^* \cdot \nabla V_n \, d(\lambda_p^* \mu) \\ &\geq \int_{E_n'} (\nabla f^* \cdot \nabla V_n) \, h_c d\mu + \epsilon \int_{E_n} d(\lambda_p^* \mu). \end{aligned}$$

In the limit $n \to 0$, we have $\lim_{n\to\infty} \int_{E_n'} \partial_f \bar{L}_\sigma(f^*) \cdot V_n d\mu = 0$ and $\lim_{n\to\infty} \int_{E_n'} (\nabla f^* \cdot \nabla V_n) \, h_c d\mu = 0$, which implies that $0 \leq \lim_{n\to\infty} \epsilon \int_{E_n} d(\lambda_p^* \mu) \leq 0$, and we get $\lim_{n\to\infty} \lambda_p^*(E_n) = 0$, i.e., the measure λ^* does not have a purely finitely additive component. Therefore, the measure λ^* is countably additive (and absolutely continuous) and possesses a Radon-Nikodym derivative w.r.t. μ, in $L^1(\mathbb{X}, \mu)$. For ease of notation, we henceforth let $\lambda^* \in L^1(\mathbb{X}, \mu)$ also denote its density function.

Since $\lambda^* \in L^1(\mathbb{X}, \mu)_{\geq 0}$ and $G_{f^*} \leq 0$ μ-a.e. in $\mathbb{X}$, we can now indeed state the complementary slackness condition as $\lambda^* \left(|\nabla f^*| - \alpha\right) = 0$ μ-a.e. in $\mathbb{X}$.

Moreover, the stationarity condition, under $\lambda^* \in L^1(\mathbb{X}, \mu)_{\geq 0}$ can now be expressed as:

$$\begin{aligned} 0 &= \int_{\mathbb{X}} \partial_f \bar{L}_\sigma(f^*) \cdot V \, d\mu + \int_{\mathbb{X}} \nabla f^* \cdot \nabla V \, \lambda^* \, d\mu \\ &= \int_{\mathbb{X}} \partial_f \bar{L}_\sigma(f^*) \cdot V \, d\mu - \int_{\mathbb{X}} \frac{1}{\mu} \nabla \cdot (\lambda^* \mu \nabla f^*) \cdot V \, d\mu + \int_{\partial\mathbb{X}} \lambda^* \nabla f^* \cdot \mathbf{n} V \mu \, dS, \end{aligned}$$

where we have used the Divergence Theorem to obtain the final equality, with S as the surface measure on $\partial\mathbb{X}$. As the above holds for any variation $V \in W^{1,\infty}((\mathbb{X}, \mu), \mathbb{Y})$, it must follow that $-\frac{1}{\mu} \nabla \cdot (\mu \lambda^* \nabla f^*) + \partial_f \bar{L}_\sigma(f^*) = 0$ μ-a.e. in $\mathbb{X}$ and $\lambda^* \mu \nabla f^* \cdot \mathbf{n} = 0$ on $\partial\mathbb{X}$, and if we do not suppose stronger regularity of the saddle point (f^*, λ^*), the equations must be hold weakly.

The above correspond to the necessary KKT conditions. Conversely, any solution pair (f^*, λ^*) which satisfies the above KKT conditions is a saddle point for the Lagrangian $\mathcal{L}_\sigma$ and is a solution to the original optimization problem.

E Proof of Theorem 3.1 (Saddle points of Lagrangian $\mathcal{H}$)

(i) Minimizers of (11). The search space for Problem (11) is given by:

$$\mathcal{F}_p = \left\{ f \in W^{1,p}((\mathbb{X}, \mu), \mathbb{Y}) \mid L_\sigma(f) \leq J_\sigma^*(\alpha) + \epsilon \right\}.$$

Let $\{u_n\}_{n\in\mathbb{N}}$ be a minimizing sequence in $\mathcal{F}_p$ for the $W^{1,p}$-seminorm, such that $u_n \in \mathcal{F}_p$ for all $n \in \mathbb{N}$ and $\lim_{n\to\infty} |||\nabla u_n|||_{L^p(\mathbb{X},\mu)} = \inf_{u\in\mathcal{F}_p} |||\nabla u|||_{L^p(\mathbb{X},\mu)}$. Since $f^* \in W^{1,\infty}((\mathbb{X}, \mu), \mathbb{Y})$, the minimizer of Problem (2) also belongs to $\mathcal{F}_p$, that is, $f^* \in \mathcal{F}_p$ and $inf_{u\in\mathcal{F}_p} |||\nabla u|||_{L^p(\mathbb{X},\mu)} \leq |||\nabla f^*|||_{L^p(\mathbb{X},\mu)} \leq \alpha$, we can choose the minimizing sequence to satisfy the bound $|||\nabla u_n|||_{L^p(\mathbb{X},\mu)} \leq$

[9] For a countably additive measure ν that is absolutely continuous w.r.t. the Lebesgue measure, and any collection of nonempty sets $\{E_n\}_{N\in\mathbb{N}}$ with $E_{n+1} \subset E_n$ and $\lim_{n\to\infty} E_n = \emptyset$, we have $\lim_{n\to\infty} \nu(E_n) = 0$ [48].

α. Similar to Section D, we now have the uniform bound $\|u_n\|_{W^{1,p}((\mathbb{X},\mu),\mathbb{Y})} \leq M + \alpha$ for all $n \in \mathbb{N}$. For $p > \dim(\mathbb{X})$, we have from Morrey's Inequality [35], for every $n \in \mathbb{N}$, that:

$$\begin{aligned}|u_n(x_1) - u_n(x_2)| &\leq \frac{2p\dim(\mathbb{X})}{p - \dim(\mathbb{X})}|x_1 - x_2|^{1-\frac{\dim(\mathbb{X})}{p}} |||\nabla u_n|||_{L^p(\mathbb{X},\mu)} \\ &\leq 2C\dim(\mathbb{X})(1 + \dim(\mathbb{X}))|x_1 - x_2|^{\frac{1}{1+\dim(\mathbb{X})}}\alpha,\end{aligned}$$

where $C = \max\left\{1, \operatorname{diam}(\mathbb{X})^{\frac{\dim(\mathbb{X})}{1+\dim(\mathbb{X})}}\right\}$. Thus, the sequence $\{u_n\}_{n\in\mathbb{N}}$ is also uniformly equicontinuous. Therefore, by the Arzelà-Ascoli Theorem, there exists a uniformly converging subsequence $\{u_{n_j}\}_{j\in\mathbb{N}}$ with limit $f^{\epsilon,p} \in \mathcal{F}_p$. Furthermore, by the continuity of the $W^{1,p}$-seminorm, we get that $\lim_{j\to\infty} |||\nabla u_{n_j}|||_{L^p(\mathbb{X},\mu)} = |||\nabla f^{\epsilon,p}|||_{L^p(\mathbb{X},\mu)} = \min_{f\in\mathcal{F}_p} |||\nabla f|||_{L^p(\mathbb{X},\mu)}$. By convexity of the $W^{1,p}$-seminorm, we get that $f^{\epsilon,p}$ is a global minimizer for Problem (11).

We therefore conclude that Problem (11) is guaranteed to have (atleast one) global minimizer $f^{\epsilon,p} \in \left\{f \in W^{1,p}((\mathbb{X},\mu),\mathbb{Y}) \mid L_\sigma(f) \leq J^*_\sigma(\alpha) + \epsilon\right\}$.

(ii) Saddle points of Lagrangian functional $\mathcal{H}_\sigma$. The constraint set is given by $\{f \in W^{1,p}((\mathbb{X},\mu),\mathbb{Y}) \mid \mathcal{G}(f) \leq 0\}$, where $\mathcal{G}(f) = L_\sigma(f) - (J^*_\sigma(\alpha) + \epsilon)$, and we have the constraint qualification:

$$0 \in \operatorname{int}\left\{\mathcal{G}\left(W^{1,p}((\mathbb{X},\mu),\mathbb{Y})\right) + [0,\infty)\right\},$$

where the operation $+$ denotes the Minkowski sum. This allows us to apply Theorem 3.6 in [47] to infer that the set of Lagrange multipliers corresponding to the minimizer $f^{\epsilon,p}$ is a non-empty, convex, bounded and weakly$-^*$ compact subset of $\mathbb{R}_{\geq 0}$. Moreover, we note that $(-\infty, 0]$ is a closed convex cone, and it follows from Theorem 3.4-(iii) in [47] that for any Lagrange multiplier $\kappa^{\epsilon,p}$, the pair $(f^{\epsilon,p}, \kappa^{\epsilon,p})$ is a saddle point of the Lagrangian functional $\mathcal{H}_\sigma$. We also have the *feasibility* condition $L_\sigma(f) \leq J^*_\sigma(\alpha) + \epsilon$.

Following a similar procedure as in Section D, we obtain the (Gateaux) derivative of the Lagrangian $\mathcal{H}_\sigma(f,\kappa) = \frac{1}{p}|||\nabla f|||^p_{L^p(\mathbb{X},\mu)} + \kappa\left(L_\sigma(f) - (J^*_\sigma(\alpha) + \epsilon)\right)$ in $W^{1,p}((\mathbb{X},\mu),\mathbb{Y})$ along $V \in W^{1,p}((\mathbb{X},\mu),\mathbb{Y})$ as:

$$D_1^{(V)}\mathcal{H}_\sigma(f,\kappa) = \int_{\mathbb{X}} |\nabla f|^{p-2}\nabla f \cdot \nabla V\, d\mu + \kappa\int_{\mathbb{X}} \partial_f \bar{L}_\sigma(f)\cdot V\, d\mu.$$

By the Minimax Theorem, we have $\mathcal{H}_\sigma(f^{\epsilon,p},\kappa^{\epsilon,p}) = \inf_f \sup_\kappa \mathcal{H}_\sigma(f,\kappa) = \sup_\kappa \inf_f \mathcal{H}_\sigma(f,\kappa)$, where the infimum is taken over $W^{1,p}((\mathbb{X},\mu),\mathbb{Y})$ and the supremum over $\mathbb{R}_{\geq 0}$. We therefore have $\mathcal{H}_\sigma(f^{\epsilon,p},\kappa^{\epsilon,p}) \geq \mathcal{H}_\sigma(f^{\epsilon,p},0)$, which yields the condition $\kappa^{\epsilon,p}\left(L_\sigma(f^{\epsilon,p}) - (J^*_\sigma(\alpha) + \epsilon)\right) \geq 0$. Moreover, from feasibility, we have $L_\sigma(f^{\epsilon,p}) \leq J^*_\sigma(\alpha) + \epsilon$ and $\kappa^{\epsilon,p} \geq 0$, which implies that $\kappa^{\epsilon,p}\left(L_\sigma(f^{\epsilon,p}) - (J^*_\sigma(\alpha) + \epsilon)\right) \leq 0$. This results in the *complementary slackness* condition $\kappa^{\epsilon,p}\left(L_\sigma(f^{\epsilon,p}) - (J^*_\sigma(\alpha) + \epsilon)\right) = 0$. From the Minimax equality, we get that $(f^{\epsilon,p},\kappa^{\epsilon,p})$ is also a critical point of $\mathcal{H}_\sigma$, that is $D_1^{(V)}\mathcal{H}_\sigma(f^{\epsilon,p},\kappa^{\epsilon,p}) = 0$ for any $V \in W^{1,p}((\mathbb{X},\mu),\mathbb{Y})$:

$$\begin{aligned}0 &= \int_{\mathbb{X}} |\nabla f^{\epsilon,p}|^{p-2}\nabla f^{\epsilon,p}\cdot\nabla V\, d\mu + \kappa^{\epsilon,p}\int_{\mathbb{X}} \partial_f \bar{L}_\sigma(f^{\epsilon,p})\cdot V\, d\mu \\ &= -\int_{\mathbb{X}} \frac{1}{\mu}\nabla\cdot\left(\mu|\nabla f^{\epsilon,p}|^{p-2}\nabla f^{\epsilon,p}\right)\cdot V\, d\mu + \int_{\partial\mathbb{X}} |\nabla f^{\epsilon,p}|^{p-2}\nabla f^{\epsilon,p}\cdot\mathbf{n}V\mu\, dS \\ &\qquad + \kappa^{\epsilon,p}\int_{\mathbb{X}} \partial_f\bar{L}_\sigma(f^{\epsilon,p})\cdot V\, d\mu,\end{aligned}$$

where we have used the Divergence Theorem to obtain the final equality, with S as the surface measure on $\partial\mathbb{X}$. This is the *stationarity* condition. As the above holds for any variation $V \in W^{1,p}((\mathbb{X},\mu);\mathbb{Y})$, it must follow that $-\frac{1}{\mu}\nabla\cdot\left(\mu|\nabla f^{\epsilon,p}|^{p-2}\nabla f^{\epsilon,p}\right) + \kappa^{\epsilon,p}\partial_f\bar{L}_\sigma(f^{\epsilon,p}) = 0$ μ-a.e. in $\mathbb{X}$ and $\mu\nabla f^{\epsilon,p}\cdot\mathbf{n} = 0$ on $\partial\mathbb{X}$, and if we do not suppose stronger regularity of $f^{\epsilon,p}$, the equations must be hold weakly.

The above correspond to the necessary KKT conditions. Conversely, any solution pair $(f^{\epsilon,p},\kappa^{\epsilon,p})$ which satisfies the above KKT conditions is a saddle point for the Lagrangian $\mathcal{H}_\sigma$ and is a solution to the original optimization problem.

F Proof of Theorem 3.2 (Convergence as $p \to \infty$)

(i) Monotonicity properties of $W^{1,p}((\mathbb{X},\mu);\mathbb{Y})$. We first note that for $p, q \in \mathbb{N}$, $1 < p < q$ and an $f \in W^{1,p}((\mathbb{X},\mu);\mathbb{Y})$, $|||f|||_{L^p(\mathbb{X},\mu)} \leq |||f|||_{L^q(\mathbb{X},\mu)}$ and $|||\nabla f|||_{L^p(\mathbb{X},\mu)} \leq |||\nabla f|||_{L^q(\mathbb{X},\mu)}$. It follows that $W^{1,q}((\mathbb{X},\mu);\mathbb{Y}) \subseteq W^{1,p}((\mathbb{X},\mu);\mathbb{Y})$. In particular, for any $p \in \mathbb{N}$, $p > 1$, we have $|||f|||_{L^p(\mathbb{X},\mu)} \leq |||f|||_{L^\infty(\mathbb{X},\mu)}$, $|||\nabla f|||_{L^p(\mathbb{X},\mu)} \leq |||\nabla f|||_{L^\infty(\mathbb{X},\mu)}$ and $W^{1,\infty}((\mathbb{X},\mu);\mathbb{Y}) \subseteq W^{1,p}((\mathbb{X},\mu);\mathbb{Y})$. It then follows that $\left\{f \in W^{1,q}((\mathbb{X},\mu);\mathbb{Y}) \mid L_\sigma(f) \leq \epsilon\right\} \subseteq \left\{f \in W^{1,p}((\mathbb{X},\mu);\mathbb{Y}) \mid L_\sigma(f) \leq \epsilon\right\}$ for $1 < p < q \leq \infty$.

(ii) Minimizers. From the strict convexity of L_σ, it follows that $\left\{f \in W^{1,p}((\mathbb{X},\mu);\mathbb{Y}) \mid L_\sigma(f) \leq \epsilon\right\}$ is closed and convex for any $1 < p \leq \infty$. Moreover, the semi-norm of $f \in W^{1,p}((\mathbb{X},\mu);\mathbb{Y})$, i.e., $|||\nabla f|||_{L^p(\mathbb{X},\mu)}$, is convex. The existence of global minimizers for the problem:

$$\inf_{f \in W^{1,p}((\mathbb{X},\mu);\mathbb{Y})} \left\{|||\nabla f|||_{L^p(\mathbb{X},\mu)}\,, \qquad \text{s.t. } \; L_\sigma(f) \leq J_\sigma^*(\alpha) + \epsilon\right\}$$

was established in Section E for every $\dim(\mathbb{X}) < p \leq \infty$ and $\epsilon > 0$.

(iii) Monotonicity of minimum value. From the existence of a global minimum value for any $\dim(\mathbb{X}) < p < \infty$, and the monotonicity properties of $W^{1,p}(\mathbb{X},\mu)$, we get for $\dim(\mathbb{X}) < p \leq q$:

$$\min_{\substack{f \in W^{1,p}((\mathbb{X},\mu);\mathbb{Y}) \\ L_\sigma(f) \leq J_\sigma^*(\alpha)+\epsilon}} |||\nabla f|||_{L^p(\mathbb{X},\mu)} \leq \min_{\substack{f \in W^{1,q}((\mathbb{X},\mu);\mathbb{Y}) \\ L_\sigma(f) \leq J_\sigma^*(\alpha)+\epsilon}} |||\nabla f|||_{L^q(\mathbb{X},\mu)}\,.$$

In particular, we get for any $p > \dim(\mathbb{X})$:

$$\min_{\substack{f \in W^{1,p}((\mathbb{X},\mu);\mathbb{Y}) \\ L_\sigma(f) \leq J_\sigma^*(\alpha)+\epsilon}} |||\nabla f|||_{L^p(\mathbb{X},\mu)} \leq \min_{\substack{f \in \mathrm{Lip}((\mathbb{X},\mu);\mathbb{Y}) \\ L_\sigma(f) \leq J_\sigma^*(\alpha)+\epsilon}} \mathrm{lip}\,(f)\,.$$

Therefore, by the convergence of bounded monotone sequences, we get:

$$\lim_{p \to \infty} \min_{\substack{f \in W^{1,p}((\mathbb{X},\mu);\mathbb{Y}) \\ L_\sigma(f) \leq J_\sigma^*(\alpha)+\epsilon}} |||\nabla f|||_{L^p(\mathbb{X},\mu)} \leq \min_{\substack{f \in \mathrm{Lip}((\mathbb{X},\mu);\mathbb{Y}) \\ L_\sigma(f) \leq J_\sigma^*(\alpha)+\epsilon}} \mathrm{lip}\,(f) = \bar{\alpha}(\epsilon).$$

(iv) Upper bound is indeed the supremum. We now consider the sequence of minimizers $\{f_\sigma^{\epsilon,p}\}_{p \in \mathbb{N}}$:

$$f_\sigma^{\epsilon,p} \in \arg \min_{\substack{f \in W^{1,p}((\mathbb{X},\mu);\mathbb{Y}) \\ L_\sigma(f) \leq J_\sigma^*(\alpha)+\epsilon}} |||\nabla f|||_{L^p(\mathbb{X},\mu)}\,.$$

Fixing a $p > \dim(\mathbb{X})$, from the monotonicity of minimum values and the compactness of $\mathbb{Y}$, we get that the sequence $\{f_\sigma^{\epsilon,q}\}_{q \geq p}$ is uniformly bounded in $W^{1,p}((\mathbb{X},\mu),\mathbb{Y})$ as $\|f_\sigma^{\epsilon,q}\|_{W^{1,p}((\mathbb{X},\mu),\mathbb{Y})} \leq M + \bar{\alpha}(\epsilon)$. Moreover, for $\dim(\mathbb{X}) < p \leq \infty$, we have from Morrey's inequality that:

$$\begin{aligned} |f_\sigma^{\epsilon,p}(x_1) - f_\sigma^{\epsilon,p}(x_2)| &\leq \frac{2p\dim(\mathbb{X})}{p - \dim(\mathbb{X})}\, |x_1 - x_2|^{1-\frac{\dim(\mathbb{X})}{p}}\, |||\nabla f_\sigma^{\epsilon,p}|||_{L^p(\mathbb{X},\mu)} \\ &\leq 2C\dim(\mathbb{X})\,(1+\dim(\mathbb{X}))\, |x_1 - x_2|^{\frac{1}{1+\dim(\mathbb{X})}}\, \bar{\alpha}(\epsilon), \end{aligned}$$

where $C = \max\left\{1, \mathrm{diam}(\mathbb{X})^{\frac{\dim(\mathbb{X})}{1+\dim(\mathbb{X})}}\right\}$. It follows from the above that the sequence $\{f_\sigma^{\epsilon,p}\}_{p \in \mathbb{N}, p > \dim(\mathbb{X})}$ is also uniformly equicontinuous. Therefore, by the Arzelà-Ascoli Theorem [46], there exists a subsequence $\left\{f_\sigma^{\epsilon,p_j}\right\}_{j \in \mathbb{N}}$ that converges uniformly to a Lipschitz continuous $f_\sigma^{\epsilon,\infty}$. Moreover, from the monotonicity of minimum values, it follows that the Lipschitz constant $\mathrm{lip}(f_\sigma^{\epsilon,\infty}) = |||\nabla f_\sigma^{\epsilon,\infty}|||_{L^\infty(\mathbb{X},\mu)} \leq \bar{\alpha}(\epsilon)$. We also have $\mathrm{lip}(f_\sigma^{\epsilon,\infty}) \geq \min_{\substack{f \in \mathrm{Lip}((\mathbb{X},\mu);\mathbb{Y}) \\ L_\sigma(f) \leq J_\sigma^*(\alpha)+\epsilon}} \mathrm{lip}\,(f) = \bar{\alpha}(\epsilon)$. Therefore, we have $\mathrm{lip}(f_\sigma^{\epsilon,\infty}) = \bar{\alpha}(\epsilon)$, and $\{f_\sigma^{\epsilon,p}\}_{p \in \mathbb{N}}$ converges uniformly (upto a subsequence) to a (global) minimizer $f^{\epsilon,\infty}$ of (10).

G Numerical Analysis of Classifier Robustness

In this section, we provide numerical analysis to quantify a classifier's robustness against data perturbation for the classification problem discussed in Section 2 and Fig. 1 of the manuscript. Using the same setup explained in Section 2, we design our classifiers by constructing a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $n = 500$ randomly selected nodes by connecting each node to its 10 nearest neighbors. We compute the solution $\mathbf{v}^*$ to (9) for different values of the Lipschitz constant $\alpha \in (0, 100]$. We generate a nominal testing set of 1000 i.i.d. samples from σ, associate them with the closest node, and evaluate the nominal classification confidence of $\mathbf{v}^*$. Then, we perturb each testing data sample with $\delta \in \mathbb{R}^2$ with $\|\delta\|_2 = 0.05$ in the direction perpendicular to the closest edge, associate each perturbed data point with the closest node and evaluate the perturbed classification confidence. To measure the sensitivity of the designed classifier, we compute the norm of the difference between the nominal and the perturbed confidence, then appoint the sensitivity measure to the maximum value across all the testing data points. Fig. G.3(a) shows the plot of the sensitivity for each classifier designed using different Lipschitz bound α, it can be seen that the sensitivity increases as we increase the Lipschitz bound up to $\alpha = 18$. Fig. G.3(b) shows the plot of the sensitivity for each classifier as a function of the classification confidence, we observe a tradeoff between classification performance and robustness to data perturbation seen by the monotonic increase of the sensitivity as a function of classification confidence, where improving classification performance comes at the expenses of robustness to data perturbation.

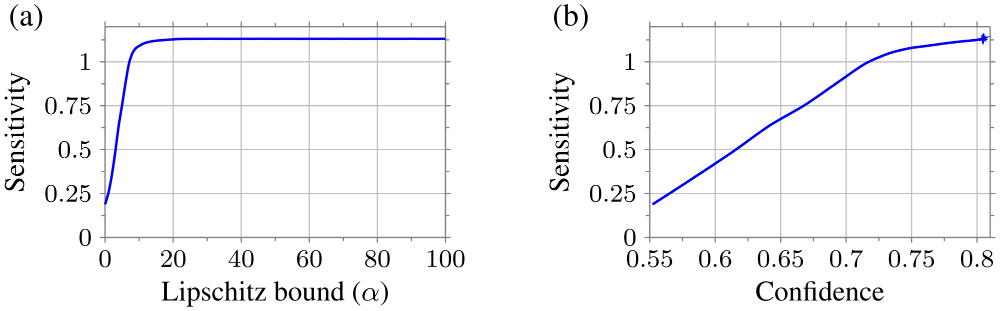


Figure G.3: For the classification problem discussed in Section 2 and Fig. 1 in the main manuscript, (a) shows the classifier's sensitivity to data perturbation as a function of the Lipschitz bound, the plot shows that sensitivity increases with the Lipschitz bound up to a certain value ($\alpha = 18$). (b) shows the tradeoff between performance and robustness, seen by the monotonic increase of the sensitivity as a function of classification confidence.