Generalised Lipschitz Regularisation Equals Distributional Robustness

Zac Cranko [*1] **Zhan Shi** [*2] **Xinhua Zhang** [2] **Richard Nock** [3] **Simon Kornblith** [3]

Abstract

The problem of adversarial examples has highlighted the need for a theory of regularisation that is general enough to apply to exotic function classes, such as universal approximators. In response, we have been able to significantly sharpen existing results regarding the relationship between distributional robustness and regularisation, when defined with a transportation cost uncertainty set. The theory allows us to characterise the conditions under which the distributional robustness equals a Lipschitz-regularised model, and to *tightly* quantify, for the first time, the slackness under very mild assumptions. As a theoretical application we show a new result explicating the connection between adversarial learning and distributional robustness. We then give new results for how to achieve Lipschitz regularisation of kernel classifiers, which are demonstrated experimentally.

1. Introduction

When learning a statistical model, it is rare that one has complete access to the distribution. More often it is the case that one approximates the risk minimisation by an empirical risk, using sequence of samples from the distribution. In practice this can be problematic — particularly when the curse of dimensionality is in full force — to *a*) know with certainty that one has enough samples, and *b*) guarantee good performance away from the data. Both of these two problems can, in effect, be cast as problems of ensuring generalisation. A remedy for both of these problems has been proposed in the form of a modification to the risk minimisation framework, wherein we integrate a certain amount of distrust of the distribution. This distrust results in a guarantee of worst case performance if it turns out later that the distribution was specified imprecisely, improving generalisation.

[*]Equal contribution [1]Universität Tübingen, Tübingen, Germany [2]University of Illinois at Chicago, IL, USA [3]Google Brain. Correspondence to: Xinhua Zhang <zhangx@uic.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

In order to make this notion of distrust concrete, we introduce some mathematical notation. The set of Borel probability measures on an outcome space Ω is $\mathfrak{P}(\Omega)$. A loss function is a mapping $f : \Omega \to \bar{\mathbb{R}}$ so that $f(\omega)$ is the loss incurred with some prediction under the outcome $\omega \in \Omega$. For example, if $\Omega = X \times Y$ then $f_v(x, y) = (v(x) - y)^2$ could be a loss function for regression or classification with some classifier $v : X \to Y$. For a distribution $\mu \in \mathfrak{P}(\Omega)$ we replace the objective in the classical risk minimisation $\min_v \mathrm{E}_\mu[f_v]$ with the *robust Bayes risk*:

$$\sup_{\nu \in \mathrm{B}_c(\mu, r)} \mathrm{E}_\nu[f] \tag{rB}$$

where $\mathrm{B}_c(\mu, r) \subseteq \mathfrak{P}(\Omega)$ is a set containing μ, called the *uncertainty set* (viz. Berger, 1993; Vidakovic, 2000, Grünwald & Dawid, 2004, §4). It is in this way that we introduce distrust into the classical risk minimisation, by instead minimising the worst case risk over a set of distributions.

It is sometimes the case that for an uncertainty set, $\mathrm{B}_c(\mu, r) \subseteq \mathfrak{P}(\Omega)$, there is a function, $r \operatorname{lip}_c : \bar{\mathbb{R}}^\Omega \to \bar{\mathbb{R}}_{\geq 0}$ (not necessarily the usual Lipschitz constant), so that

$$\sup_{\nu \in \mathrm{B}_c(\mu, r)} \mathrm{E}_\nu[f] \leq \mathrm{E}_\mu[f] + r \operatorname{lip}_c(f). \tag{L}$$

Results like (L) have been studied in the literature, however these usually make onerous assumptions on the structure of the loss function/model class (Shafieezadeh-Abadeh et al., 2019; Blanchet et al., 2019) or on the cost function underpinning the uncertainty set (Kuhn et al., 2019). Thus ruling out application to many common machine learning and statistical techniques. Therefore, in §3, our first major contribution is to revisit such a result using a new proof technique that relies on the difference-convex optimization literature to strictly generalise and improve upon several well-known related results (summarised in Table 1). In particular, a major novelty of our approach lies with the characterisation of when (L) holds as an equality, and when the bound is tight. These are quite involved and are as important as the inequality (L) itself.

In practice, however, the evaluation of Lipschitz constant is NP-hard for neural networks (Scaman & Virmaux, 2018), compelling approximations of it, or the explicit engineering of Lipschitz layers and analysing the resulting expressiveness in specific cases (e.g., ∞-norm, Anil et al., 2019). By

Table 1: Comparison of results related to (L). Assumptions listed in boldface are the weakest.

Reference	(L)	f	c	μ	X
(Shafieezadeh-Abadeh et al., 2019, Thm. 14)	$=$	convex Lipschitz margin loss with linear classifier	norm	empirical dist.	$\mathbb{R}^d$
(Kuhn et al., 2019, Thm. 5)	$\leq$	**upper semicontinuous**	norm	empirical dist.	$\mathbb{R}^d$
(Kuhn et al., 2019, Thm. 10)	$=$	**convex, Lipschitz**	norm	empirical dist.	$\mathbb{R}^d$
(Gao & Kleywegt, 2016, Cor. 2 (iv))	$\leq$	similar to generalised Lipschitz	**p-metric**	empirical dist.	$\mathbb{R}^d$
Theorem 1 (this paper)	$\leq$ $=$	- **convex, generalised Lipschitz**	- **convex, k-positively homogeneous**	**probability measure**	**separable Banach space**

comparison, kernel machines have a reproducing kernel Hilbert space (RKHS) encompassing a family of models that are universal (Micchelli et al., 2006). Our second major contribution, in §4, is to show that product kernels, such as Gaussian kernels, have a Lipzchitz constant that can be efficiently approximated and optimised with high probability. By using the Nyström approximation (Williams & Seeger, 2000; Drineas & Mahoney, 2005). we show that an ϵ approximation error requires only $\mathrm{O}(1/\epsilon^2)$ samples. Such a sampling-based approach also leads to a single convex constraint, making it scalable to large sample sizes, even with an interior-point solver (§5). As our experiments show, this method achieves higher robustness than state of the art (Cisse et al., 2017; Anil et al., 2019).

2. Preliminaries

Let $\bar{\mathbb{R}} \stackrel{\text{def}}{=} [-\infty, \infty]$ and $\bar{\mathbb{R}}_{\geq 0} \stackrel{\text{def}}{=} [0, \infty]$, with similar notations for the real numbers. Let $[n]$ denote the set $\{1, \ldots, n\}$ for $n \in \mathbb{N}$. Unless otherwise specified, X, Y, Ω are topological outcome spaces. Often X will be used when there is some linear structure so that $\Omega = X \times Y$ may be interpreted as the classical outcome space for classification problems (cf. Vapnik, 2000). In particular, in all cases X and Y can be taken to be $\mathbb{R}^d$ and $\{1, \ldots, k\}$ respectively.

The Dirac measure at some point $\omega \in \Omega$ is $\delta_\omega \in \mathfrak{P}(\Omega)$, and the set of Borel mappings $X \to Y$ is $\mathcal{L}_0(X, Y)$. For $\mu \in \mathfrak{P}(\Omega)$, denote by $\mathcal{L}_p(\Omega, \mu)$ the Lebesgue space of functions $f \in \mathcal{L}_0(\Omega, \mathbb{R})$ satisfying $\left(\int |f(\omega)|^p \mu(\mathrm{d}\omega)\right)^{1/p} < \infty$ for $p \geq 1$. The continuous real functions on Ω are collected in $\mathrm{C}(\Omega)$. In many of our subsequent formulas it is more convenient to write an expectation directly as an integral: $\mathrm{E}_\mu[f] = \int f \,\mathrm{d}\mu \stackrel{\text{def}}{=} \int f(\omega)\mu(\mathrm{d}\omega)$.

For two measures $\mu, \nu \in \mathfrak{P}(\Omega)$ the set of (μ, ν)-*couplings* is $\Pi(\mu, \nu) \subseteq \mathfrak{P}(\Omega \times \Omega)$ where $\pi \in \Pi(\mu, \nu)$ if and only if the marginals of π are μ and ν. For a *coupling function* $c : \Omega \times \Omega \to \bar{\mathbb{R}}$, the *c-transportation cost* of $\mu, \nu \in \mathfrak{P}(\Omega)$ is $\mathrm{cost}_c(\mu, \nu) \stackrel{\text{def}}{=} \inf_{\pi \in \Pi(\mu,\nu)} \int c \,\mathrm{d}\pi$. The *c-transportation cost* ball of radius $r \geq 0$ centred at $\mu \in \mathfrak{P}(\Omega)$ is $\mathrm{B}_c(\mu, r) \stackrel{\text{def}}{=} \{\nu \in \mathfrak{P}(\Omega) \mid \mathrm{cost}_c(\mu, \nu) \leq r\}$, and serves as our *uncertainty set*. The the *least c-Lipschitz constant* (cf. Cranko et al., 2019) of a function $f : X \to \bar{\mathbb{R}}$ is the number $\mathrm{lip}_c(f) \stackrel{\text{def}}{=} \inf \Lambda_c(f)$, where

$$\Lambda_c(f) \stackrel{\text{def}}{=} \{\lambda \geq 0 \mid \forall_{x,y \in X} : |f(x) - f(y)| \leq \lambda c(x, y)\}.$$

Thus when (X, d) is a metric space, $\mathrm{lip}_d(f)$ agrees with the usual Lipschitz notion. When c maps $X \to \bar{\mathbb{R}}$, for example when c is a norm, we let $c(x, y) \stackrel{\text{def}}{=} c(x - y)$ for all $x, y \in X$.

A function $f : X \to \mathbb{R}$ is called *k-positively homogeneous* if, for all $a > 0$, there is $f(ax) = a^k f(x)$ for all $x \in X$. Throughout we always assume $k \geq 1$.

To a function $f : X \to \bar{\mathbb{R}}$ we associate another function $\overline{\mathrm{co}}\, f : X \to \bar{\mathbb{R}}$, called the *convex envelope* of f, defined to be the greatest closed convex function that minorises f. The quantity $\rho(f) \stackrel{\text{def}}{=} \sup_{x \in X}(f(x) - \overline{\mathrm{co}}\, f(x))$ was first suggested by Aubin & Ekeland (1976) to quantify the lack of convexity of a function f, and has since shown to be of considerable interest for, among other things, bounding the duality gap in nonconvex optimisation (cf. Lemaréchal & Renaud, 2001; Udell & Boyd, 2016; Askari et al., 2019; Kerdreux et al., 2019). In particular, observe

$$\rho(f) = 0 \iff f = \overline{\mathrm{co}}\, f \iff f \text{ is closed convex.}$$

While it may seem like somewhat of an intractable quantity, $\rho(f)$ can be estimated in principle, details of which are included in the supplementary material (Supplement B). Complete proofs of all technical results are relegated to the supplementary material.

3. Distributional robustness

In this section we present our major result regarding identities of the form (L).

Theorem 1. *Suppose X is a separable Banach space and fix $\mu \in \mathfrak{P}(X)$. Suppose $c : X \to \bar{\mathbb{R}}_{\geq 0}$ is closed convex, k-positively homogeneous, and $f \in \mathcal{L}_1(X, \mu)$ is upper semicontinuous with $\mathrm{lip}_c(f) < \infty$. Then for all $r \geq 0$, there exists $\Delta_{f,c,r}(\mu) \geq 0$ so that*

$$\sup_{\nu \in \mathrm{B}_c(\mu, r)} \int f \,\mathrm{d}\nu + \Delta_{f,c,r}(\mu) = \int f \,\mathrm{d}\mu + r \,\mathrm{lip}_c(f), \quad (1)$$

Table 2: Comparison of results related to Theorem 2. Assumptions listed in boldface are the weakest, and assumptions in red are prohibitive.

Reference	Result	f	c	μ	X
(Staib & Jegelka, 2017, Prop. 3.1)	$\leq$	unclear	**p-metric**	unclear	**metric space**
(Shafieezadeh-Abadeh et al., 2019, Thm. 12)	$\leq$ $=$	Lipschitz margin loss with linear classifier additional strong regularity condition	norm	empirical dist.	$\mathbb{R}^d$
(Gao & Kleywegt, 2016, Cor. 2 (ii))	$=$	**concave**	**p-metric**	empirical dist.	convex subset of $\mathbb{R}^d$
Theorem 2 (this paper)	$\leq$ $=$	**measurable** **continuous**	norm	**probability measure non-atomic, compact support**	**separable Banach space**

and

$$\Delta_{f,c,r}(\mu) \leq r \operatorname{lip}_c(f) - \max\{0, r \operatorname{lip}_c(\overline{\operatorname{co}} f) - \mathrm{E}_\mu[f - \overline{\operatorname{co}} f]\}. \quad (2)$$

Observe that when f is closed convex, (2) implies $\Delta_{f,c,r}(\mu) = 0$.

A summary of the results Theorem 1 improves upon is presented in Table 1 and a more detailed discussion follows in the supplementary material (Supplement A).

Proposition 1. *Suppose X is a separable Banach space. Suppose $c : X \to \bar{\mathbb{R}}_{\geq 0}$ satisfies the conditions of Theorem 1, and $f \in \bigcap_{\mu \in \mathfrak{P}(X_0)} \mathcal{L}_1(X, \mu)$ is upper semicontinuous, has $\operatorname{lip}_c(f) < \infty$, and attains its maximum on $X_0 \subseteq X$. Then for all $r \geq 0$*

$$\sup_{\mu \in \mathfrak{P}(X_0)} \Delta_{f,c,r}(\mu) = r \operatorname{lip}_c(f) - \max\Big\{0, r \operatorname{lip}_c(\overline{\operatorname{co}} f) - \rho(f)\Big\}.$$

Remark 1. Proposition 1 shows that for any compact subset $X_0 \subseteq \mathbb{R}^d$ (such as the set of d-dimensional images, $[0,1]^d$) the bound (1) is tight with respect to the set of distributions supported here, for any upper semicontinuous $f \in \bigcap_{\mu \in \mathfrak{P}(X_0)} \mathcal{L}(X, \mu)$.

It is the first time to our knowledge that the slackness (2) has been characterised tightly. Remark 3 (in §A.1) discusses a similar way to construct such a bound from some existing results in the literature, and compares it to Theorem 2.

3.1. Adversarial learning

Szegedy et al. (2014) observe that deep neural networks, trained for image classification using empirical risk minimisation, exhibit a curious behaviour whereby an image, $x \in \mathbb{R}^d$, and a small, imperceptible amount of noise, $\delta_x \in \mathbb{R}^d$, may found so that the network classifies x and $x + \delta_x$ differently. Imagining that the troublesome noise vector is sought by an adversary seeking to defeat the classifier, such pairs have come to be known as *adversarial examples* (Moosavi Dezfooli et al., 2017; Goodfellow et al., 2015; Kurakin et al., 2017).

The closed $c : X \to \bar{\mathbb{R}}$ ball of radius $r \geq 0$, centred at $x \in X$ is denoted $\mathrm{B}_c(x, r) \stackrel{\text{def}}{=} \{y \in X \mid c(x - y) \leq r\}$. Let X be a linear space and Y a topological space. Fix $\mu \in \mathfrak{P}(X \times Y)$. The following objective has been proposed as a means of learning classifiers that are robust to adversarial examples (viz. Madry et al., 2018; Shaham et al., 2018; Carlini & Wagner, 2017; Cisse et al., 2017)

$$\int \sup_{\delta \in \mathrm{B}_c(0,r)} f(x + \delta, y) \mu(\mathrm{d}x \times \mathrm{d}y), \quad (3)$$

where $f : X \times Y \to \bar{\mathbb{R}}$ is the loss of some classifier.

Theorem 2. *Suppose (X, c_0) is a separable Banach space. Fix $\mu \in \mathfrak{P}(X)$ and for $r \geq 0$ let $R_\mu(r) \stackrel{\text{def}}{=} \{g \in \mathcal{L}_0(X, \mathbb{R}_{\geq 0}) \mid \int g \,\mathrm{d}\mu \leq r\}$. Then for $f \in \mathcal{L}_0(\Omega, \bar{\mathbb{R}})$ and $r \geq 0$ there is*

$$\sup_{g \in R_\mu(r)} \int \mu(\mathrm{d}\omega) \sup_{\omega' \in \mathrm{B}_{c_0}(\omega, g(\omega))} f(\omega') \leq \sup_{\nu \in \mathrm{B}_{c_0}(\mu, r)} \int f \,\mathrm{d}\nu, \quad (4)$$

If f is continuous and μ is non-atomically concentrated with compact support, then (4) *is an equality.*

Remark 2. By observing the constant function $g_r \equiv r$ is included in the set $R_\mu(r)$, it's easy to see that the adversarial risk (3) is upper bounded as follows

$$(3) = \int \sup_{\omega' \in \mathrm{B}_c(\omega, r)} f(\omega') \mu(\mathrm{d}\omega) \leq \sup_{g \in R_\mu(r)} \int \mu(\mathrm{d}\omega) \sup_{\omega' \in \mathrm{B}_c(\omega, g(\omega))} f(\omega'), \quad (5)$$

where, in the equality, we extend c_0 to a metric c on $X \times Y$ in the same way as (B.6).

Theorem 2 generalises and subsumes a number of existing results to relate the adversarial risk minimisation (3) to the

distributionally robust risk in Theorem 1. A discussion and summary of the improvements made by Theorem 2 on other comparable results is presented in §3.2, with a table that is similar to Table 1.

A simulation is in place demonstrating that the sum of the gaps from Theorems 1 and 2 and Equation (5) is relatively low. We randomly generated 100 Gaussian kernel classifiers $f = \sum_{i=1}^{100} \gamma_i k(x^i, \cdot)$, where x^i was sampled from the MNIST dataset and γ_i sampled uniformly from $[-2, 2]$. The bandwidth was set to the median of pairwise distances. In Figure 3, the x-axis is the adversarial risk (LHS of (5), i.e., (3)) where the perturbation δ is bounded in an ℓ_p ball and computed by projected gradient descent (PGD). The y-axis is the Lipschitz regularised empirical risk (RHS of (1)). The scattered dots lie closely to the diagonal, demonstrating that the above bounds are tight in practice.

3.2. Results related to Theorem 2

Similarly to Theorem 1, Theorem 2 improves upon a number of existing results in the literature. These are listed in Table 2. The majority of other results mentioned are are formulated with respect to an empirical distribution, that is, an average of Dirac masses. Of course any finite set is compact, and so these empirical distributions satisfy the concentration assumption. Staib & Jegelka (2017, Prop. 3.1) also state an equality result, but this is in the setting of an ∞-Wasserstein ball, which is a much more exotic object (viz. Champion et al., 2008) and is not obvious how it relates to the other results, so we choose to omit it from Table 2.

4. Lipschitz regularisation for kernel methods

Theorems 1 and 2 open up a new path to optimising the adversarial risk (3) by Lipschitz regularisation (RHS of (1)). In general, however, it is still hard to compute the Lipschitz constant for a nonlinear model (Scaman & Virmaux, 2018). Interestingly, we will show that for some types of kernels, this can be done efficiently on functions in its RKHS, which is rich enough to approximate continuous functions on a bounded domain (Micchelli et al., 2006). Thanks to the connections between kernel method and deep learning, this technique also potentially benefits the latter. For example, ℓ_1-regularised neural networks are compactly contained in the RKHS of multi-layer inverse kernels $k(x, y) = (2 - x^\top y)^{-1}$ with $\|x\|_2 \leq 1$ and $\|y\|_2 \leq 1$ (Zhang et al., 2016, Lem. 1 & Thm. 1) and (Shalev-Shwartz et al., 2011; Zhang et al., 2017), and possibly Gaussian kernels $k(x, y) = \exp(\frac{-1}{2\sigma^2} \|x-y\|^2)$ (Shalev-Shwartz et al., 2011, §5).

Consider a Mercer's kernel k on a convex domain $X \subseteq \mathbb{R}^d$, with the corresponding RKHS denoted as $\mathcal{H}$. The standard kernel method seeks a discriminant function f from $\mathcal{H}$ with the conventional form of finite kernel expansion $f(x) = \frac{1}{l} \sum_{a=1}^{l} \gamma_a k(x^a, \cdot)$, such that the regularised empirical risk can be minimised with the standard (hinge) loss and RKHS norm. We start with real-valued f for univariate output such as binary classification, and later extend it to multiclass.

Our goal here is to additionally enforce, while retaining a **convex** optimisation in $\boldsymbol{\gamma} \stackrel{\text{def}}{=} \{\gamma_a\}$, that the Lipschitz constant of f falls below a prescribed threshold $L > 0$, which is equivalent to $\sup_{x \in X} \|\nabla f(x)\|_2 \leq L$ thanks to the convexity of X. A quick but primitive solution is to piggyback on the standard RKHS norm constraint $\|f\|_{\mathcal{H}} \leq C$, in view that it already induces an upper bound on $\|\nabla f(x)\|_2$ as shown in Example 3.23 of Shafieezadeh-Abadeh et al. (2019):

$$\sup_{x \in X} \|\nabla f(x)\|_2 \;\leq\; \|f\|_{\mathcal{H}} \sup_{z > 0} \tfrac{1}{z} g(z), \tag{6}$$
$$\text{where } g(z) \geq \sup_{x, x' \in X : \|x - x'\|_2 = z} \|k(x, \cdot) - k(x', \cdot)\|_{\mathcal{H}}.$$

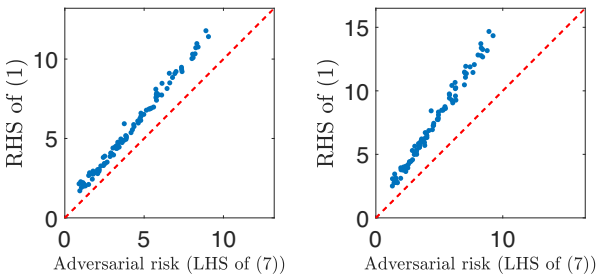
For Gaussian kernels, $g(z) = \max\{\sigma^{-1}, 1\} z$. For exponential and inverse kernels, $g(z) = z$ (Bietti & Mairal, 2019). Bietti et al. (2019) justified that the RKHS norm of a neural network may serve as a surrogate for Lipschitz regularisation. But the quality of such an approximation, i.e., the gap in (6), can be loose as we will see later in Figure 4. Besides, C and L are supposed to be independent parameters.

How can we tighten the approximation? A natural idea is to directly bound the gradient norm at n random locations $\{w^s\}_{s=1}^n$ sampled i.i.d. from X, an approach adopted by Arbel et al. (2018, Appendix D). These obviously result in convex constraints on $\boldsymbol{\gamma}$. But how many samples are needed to ensure $\|\nabla f(x)\|_2 \leq L + \epsilon$ for *all* $x \in X$? Unfortunately, as shown in §C.1, n may have to grow exponentially by $1/\epsilon^d$ for a d-dimensional space. Therefore we seek a more efficient approach by first slightly relaxing $\|\nabla f(x)\|_2$. Let $g_j(x) \stackrel{\text{def}}{=} \partial^j f(x)$ be the partial derivative with respect to the j-th coordinate of x, and $\partial^{i,j} k(x, y)$ be the partial derivative to x_i and y_j. i or j being 0 means no derivative. Assuming $\sup_{x \in X} k(x, x) = 1$ and $g_j \in \mathcal{H}$ (true for various kernels considered by Assumptions 1 and 2 below), we get a bound

$$\begin{aligned} \sup_{x \in X} \|\nabla f(x)\|_2^2 \;&=\; \sup_{x \in X} \sum\nolimits_{j=1}^{d} \langle g_j, k(x, \cdot) \rangle_{\mathcal{H}}^2 \\ &\leq \sup_{\phi : \|\phi\|_{\mathcal{H}} = 1} \sum\nolimits_{j=1}^{d} \langle g_j, \phi \rangle_{\mathcal{H}}^2 \\ &= \lambda_{\max}(G^\top G), \end{aligned} \tag{7}$$

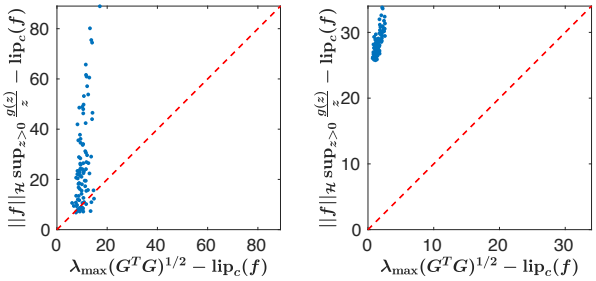
where $\lambda_{\max}$ evaluates the maximum eigenvalue, and $G \stackrel{\text{def}}{=} (g_1, \ldots, g_d)$. The "matrix" is only a notation because each column is a function in $\mathcal{H}$, and obviously the (i, j)-th entry of $G^\top G$ is $\langle g_i, g_j \rangle_{\mathcal{H}}$.

Why does $\lambda_{\max}(G^\top G)$ tend to provide a *lower* (i.e., tighter) approximation of the Lipschitz constant than (6)? To gain some intuition, note that the latter takes **two**



(a) $\|\delta\|_2 \le 3$ (b) $\|\delta\|_\infty \le 0.3$

Figure 3: Empirical evaluation of the sum of the gaps from Theorems 1 and 2. The Lipschitz constants $\sup_{x\in X} \|\nabla f(x)\|_q$ (left: $p = 2$, right: $p = \infty$, $1/p + 1/q = 1$) were estimated by BFGS.



(a) 5-layer inverse kernel (b) Gaussian kernel

Figure 4: Comparison of $\lambda_{\max}(G^\top G)$ and the RHS of (6), as upper bounds for the Lipschitz constant. Smaller values are tighter. We sampled 100 functions in the same way as in Figure 3.

steps of relaxation: $|f(x) - f(x')| \le \|f\|_{\mathcal{H}} \cdot \|k(x,\cdot) - k(x',\cdot)\|_{\mathcal{H}}$ and $\frac{\|k(x,\cdot)-k(x',\cdot)\|_{\mathcal{H}}}{\|x-x'\|_2} \le \sup_{z>0} \frac{g_z}{z}$. They attain equality at potentially very different (x, x') pairs, and the former depends on f while the latter does not. In contrast, our bound in (7) only relaxes once, leveraging the efficiently approximable partial derivatives g_j in §4.1 and capturing the correlations across different coordinates j by the eigenvalue.

An empirical comparison is further shown in Figure 4, where $\lambda_{\max}(G^\top G)$ was computed from (9) derived below, and the landmarks $\{w^s\}$ consisted of the whole training set; drawing more samples led to little difference. The gap is smaller when the bandwidth σ is larger, making functions smoother. To be fair, both Figure 3 and Figure 4 set σ to the median of pairwise distances, a common practice.

Such a positive result motivated us to develop refined algorithms to address the only remaining obstacle to leveraging $\lambda_{\max}(G^\top G)$: a computational strategy. Interestingly, it is readily approximable in both theory and practice. Indeed, the role of g_j can be approximated by its Nyström approximation $\tilde{g}_j \in \mathbb{R}^d$ (Williams & Seeger, 2000; Drineas & Mahoney, 2005) with $K \stackrel{\text{def}}{=} [k(w^i, w^{i'})]_{i,i'}$ and $Z \stackrel{\text{def}}{=} (k(w^1,\cdot), k(w^2,\cdot), \ldots, k(w^n,\cdot))$:

$$\begin{aligned}\tilde{g}_j &\stackrel{\text{def}}{=} K^{-1/2}(g_j(w^1), \ldots, g_j(w^n))^\top \\ &= (Z^\top Z)^{-1/2} Z^\top g_j \end{aligned} \tag{8}$$

because $g_j(w^i) = \langle g_j, k(w^i,\cdot)\rangle_{\mathcal{H}}$. Then to ensure $\lambda_{\max}(G^\top G) \le L^2 + \epsilon$, intuitively we can enforce $\lambda_{\max}(\tilde{G}^\top \tilde{G}) \le L^2$, where $\tilde{G} \stackrel{\text{def}}{=} (\tilde{g}_1, \ldots, \tilde{g}_d)$. It retains the convexity in the constraint on $\boldsymbol{\gamma}$. However, to guarantee ϵ error, the number of samples (n) required is generally *exponential* (Barron, 1994). Fortunately, we will next show that n can be reduced to *polynomial* for quite a general class of kernels that possess some decomposed structure.

4.1. A Nyström approximation for product kernels

A number of kernels factor multiplicatively over the coordinates, such as periodic kernels (MacKay, 1998), Gaussian kernels, and Laplacian kernels. Let us consider $k(x,y) = \prod_{j=1}^d k_0(x_j, y_j)$ where $X = (X_0)^d$ and k_0 is a base kernel on an interval X_0. Let the RKHS of k_0 be $\mathcal{H}_0$, and let μ_0 be a finite Borel measure with $\mathrm{supp}[\mu_0] = X_0$. Periodic kernels have $k_0(x_j, y_j) = \exp\big(-\sin\big(\frac{\pi}{v}(x_j - y_j)\big)^2/(2\sigma^2)\big)$.

We stress that product kernels can induce very rich function spaces. For example, Gaussian kernel is universal (Micchelli et al., 2006), meaning that its RKHS is *dense* in the space of continuous functions in the ℓ_∞ norm over any bounded domain. Also note that the factorization of kernel k does *not* imply a function $f \in \mathcal{H}$ must factor as $\prod_j f_j(x_j)$.

The key benefit of this decomposition of k is that the derivative $\partial^{0,1}k(x,y)$ can be written as $\partial^{0,1}k_0(x_1,y_1)\prod_{j=2}^d k_0(x_j,y_j)$. Since $k_0(x_j,y_j)$ can be easily dealt with, approximation will be needed *only* for $\partial^{0,1}k_0(x_1,y_1)$. Applying this idea to $g_j = \frac{1}{l}\sum_{a=1}^l \gamma_a \partial^{0,j}k(x^a,\cdot)$, we can derive

$$l^2 \|g_1\|_{\mathcal{H}}^2 = \sum_{a,b=1}^l \Big[\gamma_a\gamma_b M_{a,b} \prod_{j=2}^d k_0(x_j^a, x_j^b)\Big], \tag{9}$$

where $M_{a,b} \stackrel{\text{def}}{=} \big\langle \partial^{0,1}k_0(x_1^a,\cdot), \partial^{0,1}k_0(x_1^b,\cdot)\big\rangle_{\mathcal{H}_0}$,

$l^2 \langle g_1, g_2\rangle_{\mathcal{H}} =$

$$\sum_{a,b=1}^l \Big[\gamma_a\gamma_b \partial^{0,1}k_0(x_1^a,x_1^b)\partial^{0,2}k_0(x_2^b,x_2^a)\prod_{j=3}^d k_0(x_j^a,x_j^b)\Big].$$

So the off-diagonal entries of $G^\top G$ can be computed exactly. But this is not the case for the diagonal entries because $M_{a,b}$ is not equal to $\partial^{1,1}k_0(x_1^a,x_1^b)$. This differs from the $(\frac{\partial}{\partial x_1}f(x))^2$ used in Arbel et al. (2018), which can be computed with more ease via $\big\langle f, [\partial^{1,0}k(x,\cdot)\otimes\partial^{1,0}k(x,\cdot)]f\big\rangle_{\mathcal{H}}$. Now it is natural to apply Nyström approximation to $M_{a,b}$ in the diagonal, using samples $\{w_1^1,\ldots,w_1^n\}$ from μ_0:

$$M_{a,b} \approx \partial^{0,1}k_0(x_1^a,\cdot)^\top Z_1 (Z_1^\top Z_1)^{-1} Z_1^\top \partial^{0,1}k_0(x_1^b,\cdot), \tag{10}$$

where $Z_1 \stackrel{\text{def}}{=} (k_0(w_1^1,\cdot),\ldots,k_0(w_1^n,\cdot))$. Note

$$Z_1^\top \partial^{0,1}k_0(x_1^a,\cdot) = (\partial^{0,1}k_0(x_1^a,w_1^1),\ldots,\partial^{0,1}k_0(x_1^a,w_1^n))^\top,$$

and similarly for $Z_1^\top \partial^{0,1} k_0(x_1^b, \cdot)$. Denote this approximation of $G^\top G$ as $\tilde{P}_G$. Clearly, $\lambda_{\max}(\tilde{P}_G) \leq L^2$ is a convex constraint on $\boldsymbol{\gamma}$, based on i.i.d. samples $\{w_j^s \mid s \in [n], j \in [d]\}$ from μ_0.

The overall *convex* training procedure is summarised in Algorithm 1, where the goal is to train a kernel SVM with the additional constraint that the Lipschitz constant is at most L. More detailed formulations are available in §D. The three different ways to enforce the Lipschitz constant as discussed above correspond to options ① to ③. For practical efficiency, we greedily expand the Nyström landmark set S by locally maximizing the norm of the gradient at each iteration (step ⓑ). Figure 7 in §5.1 will show that the Nyström based algorithm is much more efficient than the brute-force counterpart, and the greedy approach significantly reduces the number of samples for both algorithms.

4.2. General sample complexity and assumptions

Finally, it is important to analyse how many samples w_j^s are needed, such that with high probability

$$\lambda_{\max}(\tilde{P}_G) \leq L^2 \implies \lambda_{\max}(G^\top G) \leq L^2 + \epsilon.$$

Fortunately, product kernels only require approximation bounds for each coordinate, making the sample complexity immune to the exponential growth in the dimensionality d. Specifically, we first consider base kernels k_0 with a scalar input, i.e., $X_0 \subseteq \mathbb{R}$. Recall from Steinwart & Christmann (2008, §4) that the integral operator for k_0 and μ_0 is $T_{k_0} \stackrel{\text{def}}{=} I \circ S_{k_0}$, where $S_{k_0} : \mathcal{L}_2(X_0, \mu_0) \to \mathrm{C}(X_0)$ operates according to $(S_{k_0} f)(x) \stackrel{\text{def}}{=} \int k_0(x, y) f(y) \mu_0(\mathrm{d}y)$ for all $f \in \mathcal{L}_2(X_0, \mu_0)$, and I: $\mathrm{C}(X_0) \hookrightarrow \mathcal{L}_2(X_0, \mu_0)$ is the inclusion operator. By the spectral theorem, if T_{k_0} is compact, then there is an at most countable orthonormal set $\{\tilde{e}_j\}_{j \in J}$ of $\mathcal{L}_2(X_0, \mu_0)$ and $\{\lambda_j\}_{j \in J}$ with $\lambda_1 \geq \lambda_2 \geq \ldots > 0$ such that $T_{k_0} f = \sum_{j \in J} \lambda_j \langle f, \tilde{e}_j \rangle_{\mathcal{L}_2(X_0, \mu_0)} \tilde{e}_j$ for all $f \in \mathcal{L}_2(X_0, \mu_0)$. It follows that $\varphi_j \stackrel{\text{def}}{=} \sqrt{\lambda_j} e_j$ is an orthonormal basis of $\mathcal{H}_0$ (cf. Steinwart & Christmann, 2008).

Our proof is built upon the following two assumptions on the base kernel. The first one asserts that fixing x, the energy of $k_0(x, \cdot)$ and $\partial^{0,1} k_0(x, \cdot)$ "concentrates" on the leading eigenfunctions.

Assumption 1. Suppose $k_0(x, x) = 1$ and $\partial^{0,1} k_0(x, \cdot) \in \mathcal{H}_0$ for all $x \in X_0$. For all $\epsilon > 0$, there exists $N_\epsilon \in \mathbb{N}$ such that the tail energy of $\partial^{0,1} k_0(x, \cdot)$ beyond the N_ϵ-th eigenpair is less than ϵ, uniformly for all $x \in X_0$. That is, denoting $\boldsymbol{\Phi}_m \stackrel{\text{def}}{=} (\varphi_1, \ldots, \varphi_m)$, $N_\epsilon < \infty$ is the smallest m such that

$$\forall_{x \in X_0} : \quad \left\| \partial^{0,1} k_0(x, \cdot) - \boldsymbol{\Phi}_m \boldsymbol{\Phi}_m^\top \partial^{0,1} k_0(x, \cdot) \right\|_{\mathcal{H}_0} < \epsilon$$
$$\text{and} \qquad \left\| k_0(x, \cdot) - \boldsymbol{\Phi}_m \boldsymbol{\Phi}_m^\top k_0(x, \cdot) \right\|_{\mathcal{H}_0} < \epsilon.$$

The second assumption asserts the smoothness and range of eigenfunctions *in a uniform sense*.

Algorithm 1 Training L-Lipschitz binary SVM

Randomly sample $S = \{w^1, \ldots, w^n\}$ from X.
for $i = 1, 2, \ldots$ **do**
  Train an SVM under one of the following constraints:
  - ① **Brute-force:** $\|\nabla f(w)\|_2^2 \leq L^2, \ \forall\, w \in S$
  - ② **Nyström holistic:** $\lambda_{\max}(\tilde{G}^\top \tilde{G}) \leq L^2$ in (8) by S
  - ③ **Nyström coordinate wise:** $\lambda_{\max}(\tilde{P}_G) \leq L^2$ in (10) by using S

  Let the trained SVM be $f^{(i)}$.
  Add a new w to S by one of the following methods:
  - ⓐ **Random:** randomly sample w from X.
  - ⓑ **Greedy:** find $\arg\max_{x \in X} \left\| \nabla f^{(i)}(x) \right\|$ (local optimisation) by L-BFGS with 10 random initialisations and add the distinct results

  Return if $L^{(i)} \stackrel{\text{def}}{=} \max_{x \in X} \left\| \nabla f^{(i)}(x) \right\|$ falls below L

Assumption 2. Under Assumption 1, $\{e_j(x) : j \in N_\epsilon\}$ is uniformed bounded over $x \in X_0$, and the RKHS inner product of $\partial^{0,1} k_0(x, \cdot)$ with $\{e_j : j \in N_\epsilon\}$ is also uniformly bounded over $x \in X_0$:

$$M_\epsilon \stackrel{\text{def}}{=} \sup_{x \in X_0} \max_{j \in [N_\epsilon]} \left| \left\langle \partial^{0,1} k_0(x, \cdot), e_j \right\rangle_{\mathcal{H}_0} \right| < \infty,$$
$$Q_\epsilon \stackrel{\text{def}}{=} \sup_{x \in X_0} \max_{j \in [N_\epsilon]} |e_j(x)| < \infty.$$

Theorem 3. *Suppose k_0, X_0, and μ_0 satisfy Assumptions 1 and 2. Let $\{w_j^s : s \in [n], j \in [d]\}$ be sampled i.i.d. from μ_0. Then for any f whose coordinate-wise Nyström approximation* (9) *and* (10) *satisfy $\lambda_{\max}(\tilde{P}_G) \leq L^2$, the Lipschitz condition $\lambda_{\max}(G^\top G) \leq L^2 + \epsilon$ is met with probability $1 - \delta$, as long as $n \geq \tilde{\Theta}\big(\frac{1}{\epsilon^2} N_\epsilon^2 M_\epsilon^2 Q_\epsilon^2 \log \frac{dN_\epsilon}{\delta}\big)$, almost independent of d. Here $\tilde{\Theta}$ hides all poly-log terms except those involving d. The proof is deferred to §C.3.*

The $\log d$ dependence on dimension d is interesting, but not surprising. After all, only the diagonal entries of $G^\top G$ need approximation, and the quantity of interest is its spectral norm, not Frobenious norm. Compared with the brute-force approach in Arbel et al. (2018) which costs exponential sample complexity, we manage to reduce it to $1/\epsilon^2$ by making two assumptions, which interestingly hold true for important classes of kernels.

Theorem 4. *Assumptions 1 and 2 hold for periodic kernel and Gaussian kernel with $\tilde{\mathrm{O}}(1)$ values of N_ϵ, M_ϵ, and Q_ϵ.*

The proof is in §C.4 and §C.5. It remains open whether non-product kernels such as inverse kernel also enjoy this polynomial sample complexity. §C.6 suggests that its complexity may be *quasi-polynomial*.

5. Experimental results

We studied the empirical robustness and accuracy of the proposed Lipschitz regularisation technique for adversarial training of kernel methods, under both Gaussian kernel and inverse kernel. Comparison will be made with state-of-the-art defence algorithms under effective attacks.

Datasets We tested on three datasets: MNIST, Fashion-MNIST, and CIFAR10. The number of training/validation/test examples for the three datasets are 54k/6k/10k, 54k/6k/10k, 45k/5k/10k, respectively. Each image in MNIST and Fashion-MNIST is represented as a 784-dimensional feature vector, with each feature/pixel normalised to $[0, 1]$. For CIFAR10, we trained it on a residual network to obtain a 512-dimensional feature embedding, which were subsequently normalised to $[0, 1]$.

Attacks To evaluate the robustness of the trained model, we attacked them on test examples using the random initialized Projected Gradient Descent method with 100 steps (PGD, Madry et al., 2018) under two losses: cross-entropy and C&W loss (Carlini & Wagner, 2017). The perturbation δ was constrained in an 2-norm or ∞-norm ball. To evaluate robustness, we scaled the perturbation bound δ from 0.1 to 0.6 for ∞-norm norm, and from 1 to 6 for 2-norm norm (when $\delta = 6$, the average magnitude per coordinate is 0.214). We normalised gradient and fine-tuned the step size.

Algorithms We compared four training algorithms. The Parseval network orthonormalises the weight matrices to enforce the Lipschitz constant (Cisse et al., 2017). We used three hidden layers of 1024 units and ReLU activation (Par-ReLU). Also considered is the Parseval network with MaxMin activations (Par-MaxMin), which enjoys much improved robustness (Anil et al., 2019). Both algorithms can be customised for 2-norm or ∞-norm attacks, and were trained under the corresponding norms. Using multi-class hinge loss, they constitute strong baselines for adversarial learning. We followed the code from LNets with $\beta = 0.5$, which is equivalent to the first-order Bjorck algorithm. The final upper bound of Lipschitz constant computed from the learned weight matrices satisfied the orthogonality constraint as shown by Anil et al. (2019, Fig. 13).

Both Gaussian and inverse kernel machines applied Lipschitz regularisation by randomly and greedily selecting $\{w^s\}$, and they will be referred to as Gauss-Lip and Inverse-Lip, respectively. In practice, Gauss-Lip with the coordinate-wise Nyström approximation ($\lambda_{\max}(\tilde{P}_G)$ from (10)) can approximate $\lambda_{\max}(G^\top G)$ with a much smaller number of sample than if using the holistic approximation as in (8). Furthermore, we found an even more efficient approach. Inside the iterative training algorithm, we used L-BFGS to find the input that yields the steepest gradient under the current solution, and then added it to the set $\{w^s\}$ (which was initialized with 15 random points). Although L-BFGS is only a local solver, this greedy approach empirically reduces the number of samples by an order of magnitude. See the empirical convergence results in §5.1. Its theoretical analysis is left for future investigation. We also applied this greedy approach to Inverse-Lip.

Extending binary kernel machines to multiclass The standard kernel methods learn a discriminant function $f^c \stackrel{\text{def}}{=} \sum_a \gamma_a^c k(x^a, \cdot)$ for each class $c \in [10]$, based on which a large variety of multiclass classification losses can be applied, e.g., CS (Crammer & Singer, 2001) which was used in our experiment. Since the Lipschitz constant of the mapping from $\{f^c\}$ to a real-valued loss is typically at most 1, it suffices to bound the Lipschitz constant of $x \mapsto (f^1(x), \ldots, f^{10}(x))^\top$ via $\max_x \lambda_{\max}(G(x)G(x)^\top)$, where $G(x) \stackrel{\text{def}}{=} [\nabla f^1(x), \cdots, \nabla f^{10}(x)] = [\langle g_j^c, k(x, \cdot)\rangle_{\mathcal{H}}]_{j\in[d],c\in[10]}$. As $\|k(x, \cdot)\|_{\mathcal{H}} = 1$, we then enforce

$$\max_{\|\phi\|_{\mathcal{H}}=1} \lambda_{\max}\Big(\sum\nolimits_{c=1}^{10} G_c^\top \phi\phi^\top G_c\Big) \le L^2, \qquad (11)$$

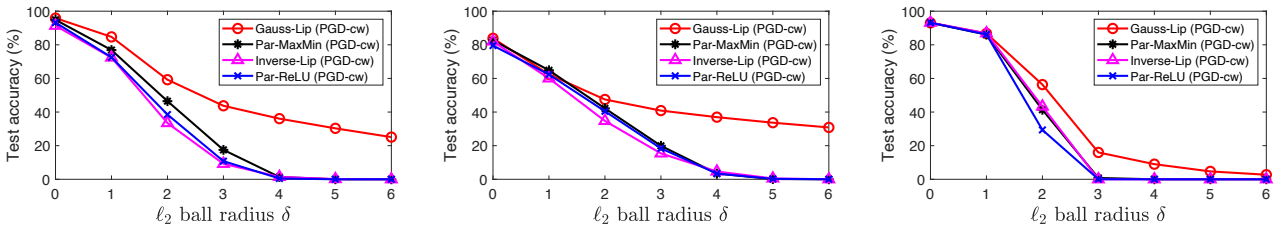
$$\text{where} \quad G_c \stackrel{\text{def}}{=} (g_1^c, \cdots, g_d^c).$$

The LHS of (11) is amenable to the same Nyström approximation as in the binary case. Further, the principle can be extended to ∞-norm attacks, whose details are in §D.1.

Parameter selection We used the same parameters as in Anil et al. (2019) for training Par-ReLU and Par-MaxMin. To defend against 2-norm attacks, we set $L = 100$ for all algorithms. Gauss-Lip achieved high accuracy and robustness on the validation set with bandwidth $\sigma = 1.5$ for FashionMNIST and CIFAR-10, and $\sigma = 2$ for MNIST. To defend against ∞-norm attacks, we set $L = 1000$ for all the four methods as in Anil et al. (2019). The best σ for Gauss-Lip is 1 for all datasets. Inverse-Lip used 5 layers.

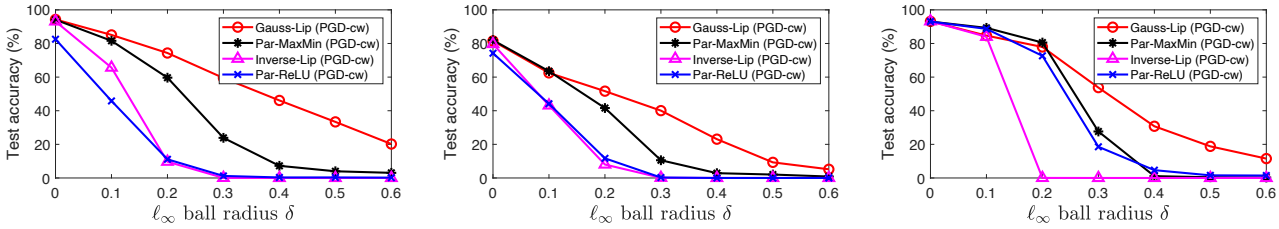
Results Figures. 5 and 6 show how the test accuracy decays as an increasing amount of perturbation (δ) in 2-norm and ∞-norm norm is added to the test images, respectively. Clearly Gauss-Lip achieves higher accuracy and robustness than Par-ReLU and Par-MaxMin on the three datasets, under both 2-norm and ∞-norm bounded PGD attacks with C&W loss. In contrast, Inverse-Lip only performs similarly to Par-ReLU. Interestingly, 2-norm based Par-MaxMin are only slightly better than Par-ReLU under 2-norm attacks, although the former does perform significantly better under ∞-norm attacks.

The results for cross-entropy PGD attacks are deferred to Figures. 9 and 10 in §E.1. Here cross-entropy PGD attackers find stronger attacks to Parseval networks but not to our kernel models. Our Gauss-Lip again significantly outperforms Par-MaxMin on all the three datasets and under both



(a) MNIST (b) Fashion-MNIST (c) CIFAR10

Figure 5: Test accuracy under PGD attacks on the C&W approximation with 2-norm norm bound



(a) MNIST (b) Fashion-MNIST (c) CIFAR10

Figure 6: Test accuracy under PGD attacks on the C&W approximation with ∞-norm norm bound

2-norm and ∞-norm norms. The improved robustness of Gauss-Lip does not seem to be attributed to the obfuscated (masked) gradient (Athalye et al., 2018), because as shown Figures. 5, 6, 9 and 10, increased distortion bound does increase attack success, and unbounded attacks drive the success rate to very low. In practice, we also observed that random sampling finds much weaker attacks, and taking 10 steps of PGD is much stronger than one step.

Obfuscated gradient To further illustrate the property of Gauss-Lip trained models, we visualised "large perturbation" adversarial examples with the 2-norm norm bounded by 8. Figure 11 in §E.2 shows the result of running PGD attack for 100 steps on Gauss-Lip trained model using (**targeted**) cross-entropy approximation. On a randomly sampled set of 10 images from MNIST, PGD successfully turned all of them into any target class by following the gradient. We further ran PGD on C&W approximation in Figure 12, and this **untargeted** attack succeeds on all 10 images. In both cases, the final images are quite consistent with human's perception.

5.1. Efficiency of enforcing Lipschitz constant

Figure 7 plots how fast the Lipschitz constant $L^{(i)}$ at iteration i is reduced by the variants 1a, 1c, 3a, and 3c in Algorithm 1, when more and more points w are added to the constraint set S. We used 400 random examples in the MNIST dataset (200 images of digit 1 and 0 each) and set $L = 3$ and RKHS norm $\|f\|_{\mathcal{H}} \leq \infty$ for all algorithms.

Clearly the Nyström algorithm is more efficient than the brute-force algorithm, and the greedy method significantly

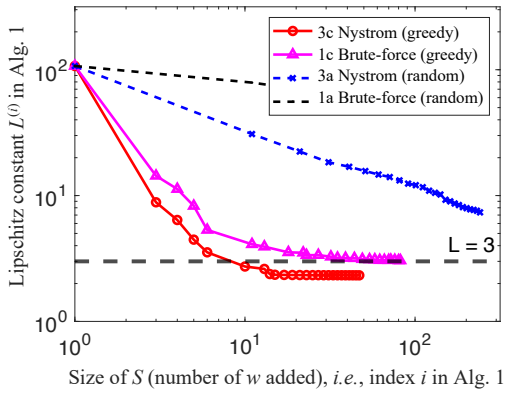


Figure 7: Comparison of efficiency in enforcing Lipschitz constant by various methods.

reduces the number of samples for both algorithms. In fact, Nyström with greedy selection (3c) eventually fell slightly below the pre-specified L, because of the gap in (7).

6. Conclusion

Risk minimisation can fail to be optimal when there is some misspecification of the distribution, such as when, as we always must, work with its empirical counterpart. Therefore we must turn to other techniques in order to ensure stability when learning a model. The robust Bayes framework provides a systematic approach to these problems, however it leaves open the choice as to which uncertainty set is most appropriate. We show that in many cases, the popular Lipschitz regularisation corresponds to robust Bayes with a transportation-cost-based uncertainty set.

References

Anil, C., Lucas, J., and Grosse, R. Sorting out Lipschitz function approximation. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th international conference on machine learning*, volume 97 of *Proceedings of machine learning research*, pp. 291–301, Long Beach, CA, USA, June 2019. PMLR.

Arbel, M., Sutherland, D. J., Binkowski, M., and Gretton, A. On gradient regularizers for MMD GANs. In *Advances in Neural Information Processing Systems (NIPS)*, 2018.

Askari, A., d'Aspremont, A., and El Ghaoui, L. Naive feature selection: sparsity in naive bayes. *arXiv:1905.09884 [cs, stat]*, May 2019. arXiv: 1905.09884.

Athalye, A., Carlini, N., and Wagner, D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning (ICML)*, 2018.

Aubin, J.-P. and Ekeland, I. Estimates of the duality gap in nonconvex optimization. *Mathematics of Operations Research*, 1(3):225–245, 1976. ISSN 0364765X, 15265471. doi: 10.1287/moor.1.3.225.

Barron, A. R. Approximation and estimation bounds for artificial neural networks. *Machine Learning*, 14:115–133, 1994.

Benoist, J. and Hiriart-Urruty, J.-B. What is the subdifferential of the closed convex hull of a function? *SIAM Journal on Mathematical Analysis*, 27(6):1661–1679, November 1996. ISSN 0036-1410, 1095-7154. doi: 10.1137/S0036141094265936.

Berger, J. O. *Statistical decision theory and Bayesian analysis*. Springer series in statistics. Springer-Verlag, New York, NY, USA, 2 edition, 1993. ISBN 0-387-96098-8. tex.mrclass: 62-02 (62A15 62Cxx) tex.mrnumber: 1234489.

Bietti, A. and Mairal, J. Group invariance, stability to deformations, and complexity of deep convolutional representations. *Journal of Machine Learning Research*, 20 (25):1–49, 2019.

Bietti, A., Mialon, G., Chen, D., and Mairal, J. A kernel perspective for regularizing deep neural networks. In *International Conference on Machine Learning (ICML)*, 2019.

Blanchet, J. and Murthy, K. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, May 2019. ISSN 0364-765X, 1526-5471. doi: 10.1287/moor.2018.0936.

Blanchet, J., Kang, Y., and Murthy, K. Robust Wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, 2019. ISSN 0021-9002. doi: 10.1017/jpr.2019.49.

Carlini, N. and Wagner, D. Towards Evaluating the Robustness of Neural Networks. In *IEEE Symposium on Security and Privacy*, pp. 39–57, San Jose, CA, USA, May 2017. IEEE. ISBN 978-1-5090-5533-3. doi: 10.1109/sp.2017.49.

Champion, T., De Pascale, L., and Juutinen, P. The ∞-Wasserstein distance: Local solutions and existence of optimal transport maps. *SIAM Journal on Mathematical Analysis*, 40(1):1–20, 2008. ISSN 0036-1410, 1095-7154. doi: 10.1137/07069938x.

Cisse, M., Bojanowski, P., Grave, E., Dauphin, Y., and Usunier, N. Parseval networks: Improving robustness to adversarial examples. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th international conference on machine learning*, volume 70, pp. 854–863, Sydney, Australia, August 2017. Proceedings of machine learning research.

Crammer, K. and Singer, Y. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of machine learning research*, 2(Dec):265–292, 2001.

Cranko, Z., Menon, A., Nock, R., Ong, C. S., Shi, Z., and Walder, C. Monge blunts Bayes: hardness results for adversarial training. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th international conference on machine learning*, volume 97, pp. 1406–1415, Long Beach, CA, USA, June 2019. Proceedings of machine learning research.

Drineas, P. and Mahoney, M. On the nystr om method for approximating a gram matrix for improved kernel-based learning. *JMLR*, 6:2153–2175, 2005.

E Fasshauer, G. Positive definite kernels: Past, present and future. *Dolomite Res. Notes Approx.*, 4, 01 2011.

Fasshauer, G. and McCourt, M. Stable evaluation of Gaussian radial basis function interpolants. *SIAM Journal on Scientific Computing*, 34(2):A737–A762, 2012.

Gao, R. and Kleywegt, A. J. Distributionally robust stochastic optimization with wasserstein distance. *arXiv:1604.02199 [math]*, July 2016. arXiv: 1604.02199.

Giner, E. Necessary and sufficient conditions for the interchange between infimum and the symbol of integration. *Set-Valued and Variational Analysis*, 17(4): 321–357, 2009. ISSN 1877-0533. doi: 10.1007/s11228-009-0119-y.

Goodfellow, I., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *International Conference on Learning Representations*, 2015.

Grünwald, P. D. and Dawid, A. P. Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. *The Annals of Statistics*, 32(4):1367–1433, August 2004. ISSN 0090-5364. doi: 10.1214/009053604000000553.

Hiriart-Urruty, J.-B. A general formula on the conjugate of the difference of functions. *Canadian Mathematical Bulletin*, 29(4):482–485, December 1986. ISSN 0008-4395, 1496-4287. doi: 10.4153/cmb-1986-076-7.

Hiriart-Urruty, J.-B. From convex optimization to nonconvex optimization. necessary and sufficient conditions for global optimality. In Clarke, F. H., Dem'yanov, V. F., and Giannessi, F. (eds.), *Nonsmooth Optimization and Related Topics*, pp. 219–239. Springer, Boston, MA, USA, 1989. ISBN 978-1-4757-6019-4. doi: 10.1007/978-1-4757-6019-4_13.

Hiriart-Urruty, J.-B. and Lemaréchal, C. *Convex analysis and minimization algorithms II*. Springer-Verlag, Berlin, Germany, 2010. ISBN 978-3-642-08162-0. OCLC: 864385173.

Kerdreux, T., Colin, I., and d'Aspremont, A. An approximate Shapley-Folkman theorem. *arXiv:1712.08559 [math]*, July 2019. arXiv: 1712.08559.

König, H. *Eigenvalue Distribution of Compact Operators*. Birkhäuser, Basel, 1986.

Kuhn, D., Esfahani, P. M., Nguyen, V. A., and Shafieezadeh-Abadeh, S. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In Netessine, S., Shier, D., and Greenberg, H. J. (eds.), *Operations Research & Management Science in the Age of Analytics*, pp. 130–166. INFORMS, 2019. ISBN 978-0-9906153-3-0. doi: 10.1287/educ.2019.0198.

Kurakin, A., Goodfellow, I., and Bengio, S. Adversarial examples in the physical world. Technical report, 2017.

Lafferty, J. and Lebanon, G. Diffusion kernels on statistical manifolds. *Journal of Machine Learning Research*, 6: 129–163, 01 2005.

Lemaréchal, C. and Renaud, A. A geometric study of duality gaps, with applications. *Mathematical Programming*, 90 (3):399–427, May 2001. ISSN 0025-5610. doi: 10.1007/pl00011429.

Lin, S.-B., Guo, X., and Zhou, D.-X. Distributed learning with regularized least squares. *Journal of Machine Learning Research*, 18(92):1–31, 2017.

LNets. `https://github.com/cemanil/LNets`.

MacKay, D. J. C. Introduction to Gaussian processes. In Bishop, C. M. (ed.), *Neural Networks and Machine Learning*, pp. 133–165. Springer, Berlin, 1998.

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

McShane, E. J. Extension of range of functions. *Bulletin of the American Mathematical Society*, 40(12):837–842, 1934. doi: 10.1090/s0002-9904-1934-05978-0. tex.fjournal: Bulletin of the American Mathematical Society.

Micchelli, C. A., Xu, Y., and Zhang, H. Universal kernels. *Journal of Machine Learning Research*, 7:2651–2667, 2006.

Minh, H. Q. Some properties of Gaussian reproducing kernel Hilbert spaces and their implications for function approximation and learning theory. *Constructive Approximation*, 32(2):307–338, 2010.

Minh, H. Q., Niyogi, P., and Yao, Y. Mercer's theorem, feature maps, and smoothing. In Lugosi, G. and Simon, H. U. (eds.), *Conference on Computational Learning Theory (COLT)*, pp. 154–168, 2006.

Moosavi Dezfooli, S. M., Fawzi, A., Fawzi, O., and Frossard, P. Universal adversarial perturbations. In *Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9, Honolulu, HI, USA,, 2017. IEEE. ISBN 978-1-5386-0457-1. doi: 10.1109/Cvpr.2017.17. tex.address: New York tex.publisher: Ieee.

Pratelli, A. On the equality between Monge's infimum and Kantorovich's minimum in optimal mass transportation. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, 43(1):1–13, January 2007. ISSN 02460203. doi: 10.1016/j.anihpb.2005.12.001.

Rasmussen, C. E. and Williams, C. K. I. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006.

Scaman, K. and Virmaux, A. Lipschitz regularity of deep neural networks: analysis and efficient estimation. In *Proceedings of the 32nd international conference on neural information processing systems*, pp. 3839–3848, Montréal, Canada, 2018. Curran Associates Inc.

Schneider, H. An inequality for latent roots applied to determinants with dominant principal diagonal. *Journal of the London Mathematical Society*, s1-28(1):8–20, 1953.

Shafieezadeh-Abadeh, S., Kuhn, D., and Esfahani, P. M. Regularization via mass transportation. *Journal of Machine Learning Research*, 20(103):1–68, 2019.

Shaham, U., Yamada, Y., and Negahban, S. Understanding adversarial training: Increasing local stability of supervised models through robust optimization. *Neurocomputing*, 307:195–204, September 2018. ISSN 09252312. doi: 10.1016/j.neucom.2018.04.027.

Shalev-Shwartz, S., Shamir, O., and Sridharan, K. Learning kernel-based halfspaces with the 0-1 loss. *SIAM Journal on Computing*, 40(6):1623–1646, 2011.

Shi, Z.-C. and Wang, B.-Y. Bounds for the determinant, characteristic roots and condition number of certain types of matrices. *Acta Math. Sinica*, 15(3):326–341, 1965.

Sinha, A., Namkoong, H., and Duchi, J. Certifiable distributional robustness with principled adversarial training. In *International conference on learning representations*, 2018.

Staib, M. and Jegelka, S. Distributionally robust deep learning as a generalization of adversarial training. In *31st Conference on Neural Information Processing Systems*, Long Beach, CA, USA, 2017.

Steinwart, I. and Christmann, A. *Support vector machines*. Springer Science & Business Media, 2008.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *International conference on learning representations*, 2014.

Toland, J. F. A duality principle for non-convex optimisation and the calculus of variations. *Archive for Rational Mechanics and Analysis*, 71(1):41–61, May 1979. ISSN 0003-9527, 1432-0673. doi: 10.1007/bf00250669.

Tropp, J. A. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1-2):1–230, 2015.

Udell, M. and Boyd, S. Bounding duality gap for separable problems with linear constraints. *Computational Optimization and Applications*, 64(2):355–378, June 2016. ISSN 1573-2894. doi: 10.1007/s10589-015-9819-4.

Vapnik, V. N. *The nature of statistical learning theory*. Springer, New York, NY, USA, 2000. ISBN 978-1-4757-3264-1 978-1-4419-3160-3. OCLC: 864225872.

Vidakovic, B. Γ-Minimax: a paradigm for conservative robust bayesians. In Bickel, P., Diggle, P., Fienberg, S., Krickeberg, K., Olkin, I., Wermuth, N., Zeger, S., Insua, D. R., and Ruggeri, F. (eds.), *Robust Bayesian Analysis*, volume 152, pp. 241–259. Springer, New York, NY, USA, 2000. ISBN 978-0-387-98866-5 978-1-4612-1306-2. doi: 10.1007/978-1-4612-1306-2_13.

Villani, C. *Optimal transport: old and new*. Number 338 in Grundlehren der mathematischen Wissenschaften. Springer-Verlag, Berlin, Germany, 2009. ISBN 978-3-540-71049-3. OCLC: ocn244421231.

Whitney, H. Analytic extensions of differentiable functions defined in closed sets. *Transactions of the American Mathematical Society*, 36(1):63–89, 1934. doi: 10.1090/s0002-9947-1934-1501735-3.

Williams, C. K. I. and Seeger, M. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems (NIPS)*, 2000.

Williamson, R. C., Smola, A. J., and Schölkopf, B. Generalization bounds for regularization networks and support vector machines via entropy numbers of compact operators. *IEEE Transactions on Information Theory*, 47(6):2516–2532, 2001.

Yang, F. and Wei, Z. Generalized Euler identity for subdifferentials of homogeneous functions and applications. *Journal of Mathematical Analysis and Applications*, 337(1):516–523, 2008. ISSN 0022247X. doi: 10.1016/j.jmaa.2007.04.008.

Zhang, Y., Lee, J. D., and Jordan, M. I. ℓ_1-regularized neural networks are improperly learnable in polynomial time. In *International Conference on Machine Learning (ICML)*, 2016.

Zhang, Y., Liang, P., and Wainwright, M. Convexified convolutional neural networks. In *International Conference on Machine Learning (ICML)*, 2017.

Zhou, D.-X. The covering number in learning theory. *Journal of Complexity*, 18(3):739–767, 2002.

Zhu, H., Williams, C. K., Rohwer, R. J., and Morciniec, M. Gaussian regression and optimal finite dimensional linear models. In Bishop, C. M. (ed.), *Neural Networks and Machine Learning*. Springer-Verlag, Berlin, 1998.

Zălinescu, C. *Convex analysis in general vector spaces*. World Scientific, River Edge, NJ, USA, 2002. ISBN 978-981-238-067-8. OCLC: 845511462.