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	$\boldsymbol{p}$**-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**	$\boldsymbol{p}$**-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):

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

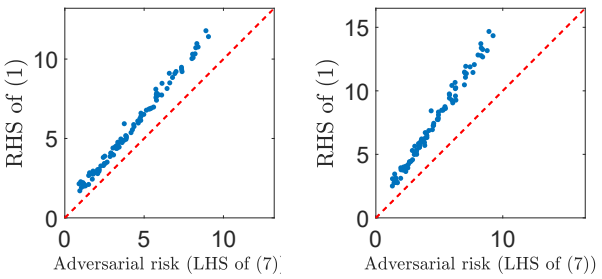
For Gaussian kernels, $g(z) = \max\{\sigma^{-1}, 1\} z$. For exponential and inverse kernels, $g(z) = z$ (Bietti & Mairal, 2019). Bietti et al. (2019) justified that the RKHS norm of a neural network may serve as a surrogate for Lipschitz regularisation. But the quality of such an approximation, i.e., 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

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

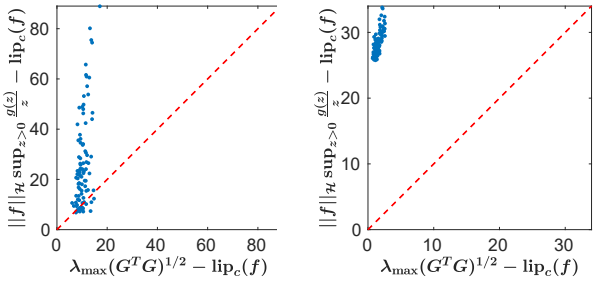
where $\lambda_{\max}$ evaluates the maximum eigenvalue, and $G \stackrel{\text{def}}{=} (g_1, \ldots, g_d)$. The "matrix" is only a notation because each column is a function in $\mathcal{H}$, and obviously the (i, j)-th entry of $G^\top G$ is $\langle g_i, g_j \rangle_{\mathcal{H}}$.

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

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: $|f(x) - f(x')| \leq \|f\|_{\mathcal{H}} \cdot \|k(x,\cdot) - k(x',\cdot)\|_{\mathcal{H}}$ and $\frac{\|k(x,\cdot)-k(x',\cdot)\|_{\mathcal{H}}}{\|x-x'\|_2} \leq \sup_{z>0} \frac{g_z}{z}$. 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 $\lambda_{\max}(G^\top G) \leq L^2 + \epsilon$, intuitively we can enforce $\lambda_{\max}(\tilde{G}^\top \tilde{G}) \leq L^2$, 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

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

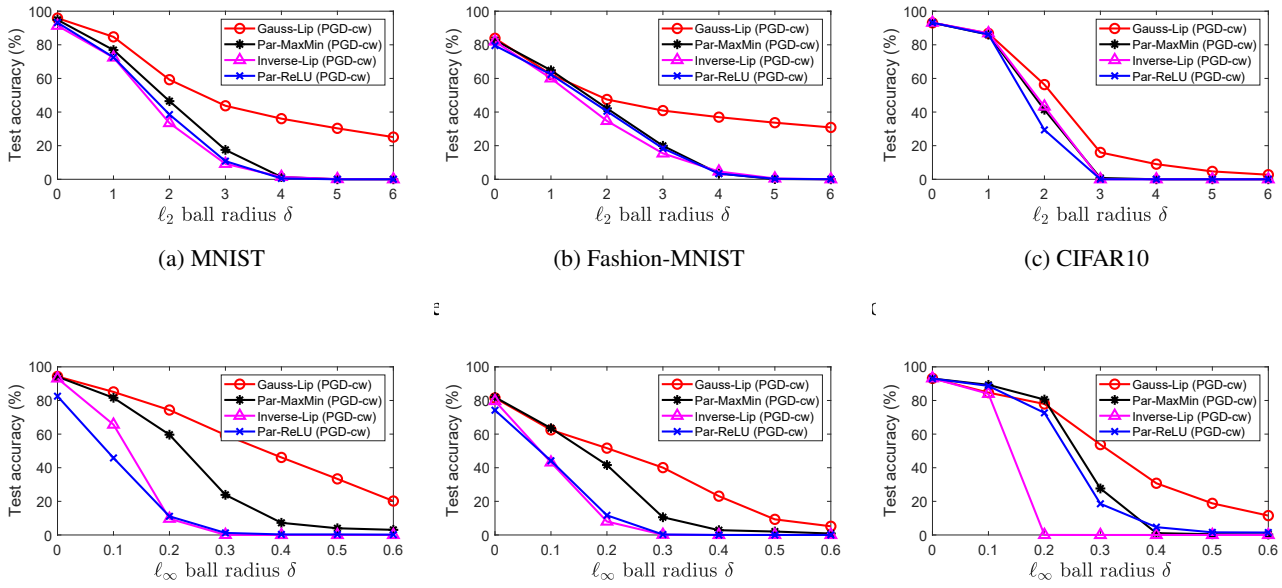
$$\text{where} \quad G_c \stackrel{\text{def}}{=} (g_1^c, \cdots, g_d^c).$$

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 2-norm and ∞-norm norms. The improved robustness of



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

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

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

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 reduces the number of samples for both algorithms. In fact,

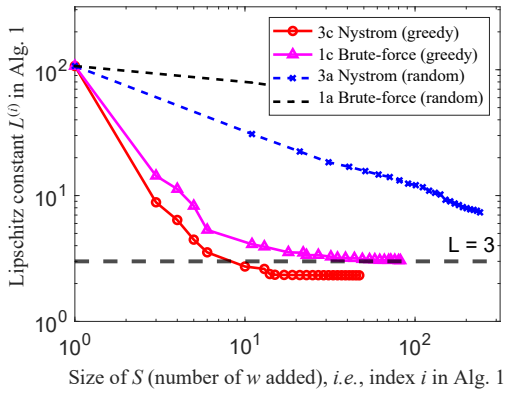


Figure 7: Comparison of efficiency in enforcing Lipschitz constant by various methods.

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.

Acknowledgements We thank the reviewers for their constructive comments. This work is supported by NSF grant RI:1910146.

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.

Supplementary Material

All code and data are available anonymously, with no tracing, at

https://github.com/learndeep2019/DRobust.

A. Discussion of Theorems 1 and 2

The reason we are interested in studying identities like (L) in full generality is to demonstrate that these relationships, which have been studied in particular specific cases by a number of authors (cf. Tables. 1 and 2) have a simple common structure. In this manner our goal is to contribute to the understanding of distributional robustness and regularisation directly, rather than the specific application articulated in the adversarial robustness literature. In particular, our choice of a separable Banach space for X is primarily motivated by the work of Blanchet & Murthy (2019), wherein the authors consider a Polish space. When X is a Polish space equipped with a linear structure (so that we can exploit identities from convex analysis), this makes X a separable Fréchet space. Our analysis is only restricted to the Banach setting only by our use of the generalised Euler identity (Yang & Wei, 2008, Thm. 3.2), however we feel that this restriction is elementary.

A.1. Results related to Theorem 1

There are a number of similar results concerning identities of the form (L) and these are summarised in Table 1; the result column refers to the relationship shown in (L). The assumptions necessary to show only inequality in Theorem 1 are substantially weaker than the complete statement of the theorem (this is shown in the first paragraph of the proof on p.) and so we don't include them in table. The weakest assumptions are highlighed with bold text, and any onerous assumptions are highlighted with bold red text. In all cases our result is a strict generalisation, and no other works cited observe our slackness bound using the lack of convexity parameter. The closest result to our slackness bound is not noted in — but can be derived from — the work of Kuhn et al. (2019), which mention in Remark 3.

Remark 3. A similar slackness bound to (2) can be derived from Kuhn et al. (2019, Thms. 5,10), who show (under additional assumptions)

$$\sup_{\nu \in \mathrm{B}_{\|\cdot\|}(\mu, r)} \int f \,\mathrm{d}\mu \leq \int f \,\mathrm{d}\mu + r \operatorname{lip}_{\|\cdot\|}(f)$$

and

$$\sup_{\nu \in \mathrm{B}_{\|\cdot\|}(\mu, r)} \int \overline{\mathrm{co}}\, f \,\mathrm{d}\mu = \int \overline{\mathrm{co}}\, f \,\mathrm{d}\mu + r \operatorname{lip}_{\|\cdot\|}(\overline{\mathrm{co}}\, f),$$

which, together with the observation $\overline{\mathrm{co}}\, f \leq f$, implies the slackness bound

$$\forall_{\mu \in \mathfrak{P}(X)} : \; \Delta_{f, \|\cdot\|, r}(\mu) \leq r \Big(\operatorname{lip}_{\|\cdot\|}(f) - \operatorname{lip}_{\|\cdot\|}(\overline{\mathrm{co}}\, f) \Big) + \rho(f). \tag{A.1}$$

However, (A.1) neither enjoys the same tightness guarantee as (2) (as demonstrated by Example 1), nor is stated with our level of generality.

Example 1. Let $I \stackrel{\text{def}}{=} [-r_0/2, r_0/2] \subseteq \mathbb{R}$ be an interval defined for some $r_0 > 0$. Let $f(x) \stackrel{\text{def}}{=} 1 - (2x/r_0)^2$ for $x \in I$ and $f(x) = 0$ for all other points x. Then f is upper semicontinuous, $\overline{\mathrm{co}}\, f \equiv 0$, $\rho(f) = 1$. Then

$$\forall_{\mu \in \mathfrak{P}(I)} : \; \operatorname{cost}_{|\cdot|}(\mu, \delta_0) = \int_I |x| \mu(\mathrm{d}x) \leq r_0,$$

and $\delta_0 \in \mathrm{B}_c(\mu, r_0)$ for all $\mu \in \mathfrak{P}(I)$. The left hand side of (A.1) at any $\mu \in \mathfrak{P}(I)$ is

$$\begin{aligned} \int f \,\mathrm{d}\mu + r_0 \operatorname{lip}_c(f) - \sup_{\nu \in \mathrm{B}_c(\mu, r_0)} \int f \,\mathrm{d}\nu &= \int f \,\mathrm{d}\mu + r_0 \operatorname{lip}_c(f) - 1 \\ &\leq 1 + r_0 \operatorname{lip}_c(f) - 1, \\ &= r_0 \operatorname{lip}_c(f), \end{aligned}$$

while the right hand side of (A.1) is

$$\rho(f) + r_0(\operatorname{lip}_c(f) - \operatorname{lip}_c(\overline{\operatorname{co}}\, f)) = 1 + r_0 \operatorname{lip}_c(f).$$

This shows that

$$\sup_{\mu \in \mathfrak{P}(I)} \Delta_{f,\|\cdot\|,r}(\mu) < \rho(f) + r_0(\operatorname{lip}_c(f) - \operatorname{lip}_c(\overline{\operatorname{co}}\, f)).$$

Then, by the intermediate value theorem, there exists $0 \leq r < r_0$ so that the bound (A.1) is not tight in the same way as (2).

B. Technical results on distributional robustness

For a topological vector space X we denote by X^* its topological dual. These are in a duality with the pairing $\langle \cdot, \cdot \rangle : X \times X^* \to \mathbb{R}$. The weakest topology on X so that X^* is its topological dual is denoted $\sigma(X, X^*)$. The continuous real functions on a topological space Ω are collected in $\mathrm{C}(\Omega)$, and the subset of these that are bounded is $\mathrm{C_b}(\Omega)$. For a measure $\mu \in \mathfrak{P}(X)$ and a Borel mapping $f : X \to Y$, the push-forward measure is denoted $f_{\#}\mu \in \mathfrak{P}(Y)$ where $f_{\#}\mu(A) \stackrel{\text{def}}{=} \mu(f^{-1}(A))$ for every Borel $A \subseteq Y$.

The *ϵ-subdifferential* of a convex function $f : X \to \bar{\mathbb{R}}$ at a point $x \in X$ is

$$\partial_\epsilon f(x) \stackrel{\text{def}}{=} \{x^* \in X^* \mid \forall_{y \in X} :\ \langle y - x, x^* \rangle - \epsilon \leq f(y) - f(x)\},$$

where $\epsilon \geq 0$. The *Moreau–Rockafellar subdifferential* is $\partial f(x) \stackrel{\text{def}}{=} \partial_0 f(x)$ and satisfies $\partial f(x) = \bigcap_{\epsilon > 0} \partial_\epsilon f(x)$. The Legendre–Fenchel conjugate of a function $f : X \to \bar{\mathbb{R}}$ is the function $f^* : X^* \to \bar{\mathbb{R}}$ defined by

$$\forall_{x^* \in X^*} :\ f^*(x^*) \stackrel{\text{def}}{=} \sup_{x \in X} (\langle x, x^* \rangle - f(x)),$$

and satisfies the following Fenchel–Young rule when f is closed convex

$$\forall_{x \in f^{-1}(\mathbb{R})} \forall_{x^* \in \partial_\epsilon f(x)} :\ f(x) + f^*(x^*) - \langle x, x^* \rangle \leq \epsilon. \tag{B.1}$$

Finally the domains are $\operatorname{dom} \partial f \stackrel{\text{def}}{=} \{x \in X \mid \partial f(x) \neq \emptyset\}$ and $\operatorname{dom} \partial_\epsilon f \stackrel{\text{def}}{=} \{x \in X \mid \partial_\epsilon f(x) \neq \emptyset\}$.

A coupling function $c : X \times X \to \bar{\mathbb{R}}$ has an associated conjugacy operation with

$$f^c(x) \stackrel{\text{def}}{=} \sup_{y \in X} (f(y) - c(x, y)),$$

for any function $f : X \to \bar{\mathbb{R}}$. The *indicator function* of a set $A \subseteq X$ is $\iota_A(x) \stackrel{\text{def}}{=} 0$ for $x \in A$ and $\iota_A(x) \stackrel{\text{def}}{=} \infty$ for $x \notin A$.

When $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ is minorised by an affine function, there is (cf. Hiriart-Urruty & Lemaréchal, 2010, Prop. X.1.5.4; Benoist & Hiriart-Urruty, 1996)

$$\overline{\operatorname{co}}\, f(x) = \inf \left\{ \sum_{i \in [n+1]} \alpha_i f(x_i) \;\middle|\; (\alpha_1, \ldots, \alpha_{n+1}) \in \Delta^n, (x_i)_{i \in [n+1]} \subseteq \mathbb{R}^d, \sum_{i \in [n+1]} \alpha_i x_i = x \right\}$$

for all $x \in \mathbb{R}^d$, where $\Delta^n \stackrel{\text{def}}{=} \{(\alpha_1, \ldots, \alpha_{n+1}) \in \mathbb{R}^n_{\geq 0} \mid \sum_{i \in [n+1]} \alpha_i = 1\}$. Consequentially it is well known that $\rho(f)$ can be computed via

$$\rho(f) = \sup_{\substack{(\alpha_1, \ldots, \alpha_{n+1}) \in \Delta^{n+1} \\ (x_1, \ldots, x_{n+1}) \in (\mathbb{R}^d)^{n+1}}} \left(f\left(\sum_{i \in [n+1]} \alpha_i x_i \right) - \sum_{i \in [n+1]} \alpha_i f(x_i) \right).$$

B.1. Proof of Theorem 1 and other technical results

Lemma 1 ((Blanchet & Murthy (2019, Thm. 1))). *suppose Ω is a Polish space and fix $\mu \in \mathfrak{P}(\Omega)$. Let $c : \Omega \times \Omega \to \bar{\mathbb{R}}_{\geq 0}$ be lower semicontinuous with $c(\omega, \omega) = 0$ for all $\omega \in \Omega$, and $f : \Omega \to \mathbb{R}$ is upper semicontinuous. Then for all $r \geq 0$ there is*

$$\sup_{\nu \in \mathrm{B}_c(\mu, r)} \int f \,\mathrm{d}\nu = \inf_{\lambda \geq 0} \left(\lambda r + \int f^{\lambda c} \,\mathrm{d}\mu \right). \tag{B.2}$$

Duality results like Lemma 1 have been the basis of a number of recent theoretical efforts in the theory of adversarial learning (Sinha et al., 2018; Gao & Kleywegt, 2016; Blanchet et al., 2019; Shafieezadeh-Abadeh et al., 2019), the results of Blanchet & Murthy (2019) being the most general to date. The necessity for such duality results like Lemma 1 is because while the supremum on the left hand side of (B.2) is over a (usually) infinite dimensional space, the right hand side only involves only a finite dimensional optimisation. The generalised conjugate in (B.2) also hides an optimisation, but when the outcome space Ω is finite dimensional, this too is a finite dimensional problem.

We also require the following result of Yang & Wei (2008) to exploit the structure of k-homogenous functions.

Lemma 2 ((Yang & Wei (2008, Thm. 3.2))). *Suppose X is a Banach space and $c : X \to \bar{\mathbb{R}}$ is convex, k-positively homogeneous for $k > 0$, and lower semicontinuous. Then for every $x \in \operatorname{dom} \partial c$ there is*

$$\forall_{x^* \in \partial c(x)} : \; c(x) = k^{-1} \langle x, x^* \rangle.$$

The following lemma is sometimes stated a consequence of, or in the proof of, the McShane–Whitney extension theorem (McShane, 1934; Whitney, 1934), but it is immediate to observe.

Lemma 3. *Let X be a set. Assume $c : X \times X \to \bar{\mathbb{R}}_{\geq 0}$ satisfies $c(x, x) = 0$ for all $x \in X$, $f : X \to \mathbb{R}$. Then*

$$1 \geq \mathrm{lip}_c(f) \iff \forall_{y \in X} : \; f(y) = \sup_{x \in X} (f(x) - c(x, y)).$$

Proof. Suppose $1 \geq \mathrm{lip}_c(f)$. Fix $y_0 \in X$. Then

$$\forall_{x \in X} : \; f(x) - c(x, y_0) \leq f(y_0),$$

with equality when $x = y_0$. Next suppose

$$\forall_{y \in X} : \; f(y) = \sup_{x \in X} (f(x) - c(x, y)),$$

then

$$\begin{aligned} \forall_{x,y \in X} : \; f(y) \geq f(x) - c(x, y) &\iff \forall_{x,y \in X} : \; f(x) - f(y) \leq c(x, y) \\ &\iff 1 \geq \mathrm{lip}_c(f), \end{aligned}$$

as claimed. □

Lemma 4. *Suppose X is a locally convex Hausdorff topological vector space and $c : X \to \bar{\mathbb{R}}_{\geq 0}$ satisfies $c(0) = 0$, and $f : X \to \mathbb{R}$ is convex. Then*

$$1 \geq \mathrm{lip}_c(f) \iff \forall_{\epsilon \geq 0} : \; \partial_\epsilon f(X) \subseteq \partial_\epsilon c(0).$$

Proof. Assume $1 \geq \mathrm{lip}_c(f)$. Then $f(x) - f(y) \leq c(x - y)$ for all $x, y \in X$. Fix $\epsilon \geq 0$, $x \in X$ and suppose $x^* \in \partial_\epsilon f(x)$. Then

$$\begin{aligned} \forall_{y \in X} : \; \langle y - x, x^* \rangle - \epsilon \leq f(y) - f(x) &\leq c(y - x) \\ \iff \forall_{y \in X} : \; \langle y, x^* \rangle - \epsilon \leq f(y + x) - f(x) &\leq c(y) - c(0), \end{aligned}$$

because $c(0) = 0$. This shows $x^* \in \partial_\epsilon c(0)$.

Next assume $\partial_\epsilon f(x) \subseteq \partial_\epsilon c(0)$ for all $\epsilon \geq 0$ and $x \in X$. Because f is not extended-real valued, it is continuous on all of X (via Zălinescu, 2002, Cor. 2.2.10) and $\partial f(x)$ is nonempty for all $x \in X$ (via Zălinescu, 2002, Thm. 2.4.9). Fix an arbitrary $x \in X$. Then $\emptyset \neq \partial f(x) \subseteq \partial c(0)$, and

$$\begin{aligned} &\exists_{x^* \in \partial f(x)} \forall_{y \in X} : \ f(x) - f(y) \leq \langle x - y, x^* \rangle \\ &\qquad \implies \forall_{y \in X} : \ f(x) - f(y) \leq \langle x - y, x^* \rangle \leq c(x - y), \end{aligned} \tag{B.1}$$

where the implication is because $x^* \in \partial c(0)$ and $c(0) = 0$. Since the choice of x in (B.1) was arbitrary, the proof is complete. □

Lemma 5. *Suppose X is a Banach space and $c : X \to \bar{\mathbb{R}}_{\geq 0}$ is convex, k-positively homogeneous. Then (i) $c^* \geq \iota_{\frac{1}{k}\partial c(0)}$, and (ii) $c^*(x^*) = \infty$ for any $x^* \notin \partial c(0)$.*

Proof. Fix an arbitrary $x \in X$. Then, for $\epsilon \geq 0$, there is $x^* \in \partial_\epsilon c(x)$ if and only if

$$\begin{aligned} \langle y - x, x^* \rangle \leq c(y) - c(x) + \epsilon &\iff \langle y - x, x^* \rangle \leq c(y) - c(x) + \epsilon \\ &\iff \langle y, x^* \rangle - \underbrace{\langle x, x^* \rangle}_{kc(x)} \leq c(y) - c(x) + \epsilon \\ &\iff \langle y, x^* \rangle \leq c(y) + (k-1)c(x) + \epsilon, \end{aligned}$$

holds for every $y \in X$. Then, so long as $k \geq 1$, we have $\partial_\epsilon c(x) = \partial_{(k-1)c(x)+\epsilon} c(0) \supseteq \partial_\epsilon c(0)$. Setting $\epsilon = 0$ we find

$$\forall_{x \in \operatorname{dom}(\partial c)} : \ \partial c(x) \supseteq \partial c(0). \tag{B.1}$$

Fix an arbitrary $x_0^* \in X^*$. Then because c is convex and real-valued, $\operatorname{dom} \partial c = X$ and

$$\begin{aligned} c^*(x_0^*) &= \sup_{x \in \operatorname{dom}(\partial c)} \left(\langle x, x_0^* \rangle - c(x) \right) \\ &\overset{\text{L2}}{=} \sup_{x \in \operatorname{dom}(\partial c)} \sup_{x^* \in \partial c(x)} \left(\langle x, x_0^* \rangle - k^{-1} \langle x, x^* \rangle \right) \\ &\overset{\text{(B.1)}}{\geq} \sup_{x \in \operatorname{dom}(\partial c)} \sup_{x^* \in \partial c(0)} \left(\langle x, x_0^* \rangle - k^{-1} \langle x, x^* \rangle \right) \\ &= \sup_{x \in \operatorname{dom}(\partial c)} \sup_{x^* \in \partial c(0)} \left\langle x, x_0^* - k^{-1} x^* \right\rangle \\ &\geq \sup_{x \in \operatorname{dom}(\partial c)} f(x, x_0^*), \end{aligned} \tag{B.2}$$

where

$$f(x, x^*) \overset{\text{def}}{=} \begin{cases} 0 & kx^* \in \partial c(0) \\ \langle x, x^* \rangle & kx^* \notin \partial c(0). \end{cases}$$

If $kx_0^* \notin \partial c(0)$ then there is $x_0 \in X$ with

$$k\langle x_0, x_0^* \rangle > c(x_0) \implies \infty > \langle x_0, x_0^* \rangle > \frac{1}{k} c(x_0) \geq 0,$$

and $x_0 \in \operatorname{dom} f$. Therefore for any $x_0^* \notin \partial c(0)$,

$$\sup_{x \in \operatorname{dom}(\partial c)} f(x, x_0^*) = \sup_{x \in \operatorname{dom}(c)} f(x, x_0^*) \geq \sup_{a > 0} a \langle x_0, x_0^* \rangle = \infty. \tag{B.3}$$

In the first equality we used the fact that $\operatorname{cl} \operatorname{dom}(\partial c) = \operatorname{cl} \operatorname{dom}(c)$. This shows

$$c^*(x_0^*) \overset{\text{(B.2)}}{\geq} \sup_{x \in \operatorname{dom}(\partial c)} f(x, x_0^*) \overset{\text{(B.3)}}{=} \iota_{\frac{1}{k}\partial c(0)},$$

and proves (i).

Suppose $x_0^* \notin \partial c(0)$. Then there exists $y \in X$ so that $\langle y, x_0^* \rangle > c(y)$. Let $a_0 \overset{\text{def}}{=} \sqrt[p-1]{p}$. Then $a_0 > 0$, $\frac{a_0^p}{a_0 k} = 1$, and

$$\begin{aligned}\langle y, x_0^* \rangle > c(y) &\iff \langle y, x_0^* \rangle > \frac{a_0^k}{a_0 p} c(y) \\ &\iff \langle a_0 y, k x_0^* \rangle > a_0^k c(y) \\ &\iff \langle a_0 y, k x_0^* \rangle > c(a_0 y),\end{aligned}$$

where in the last line we used the k-positive homogeneity of c. This shows that $k x_0^* \notin \partial c(0)$. Using (i) we obtain

$$x_0^* \notin \partial c(0) \implies k x_0^* \notin \partial c(0) \implies \iota_{\partial c(0)}(x_0^*) = \infty \overset{\text{L5 (i)}}{\implies} c^*(x_0^*) = \infty,$$

which completes the proof of (ii). □

Lemma 6. *Assume X is a Banach space. Suppose X is a Banach space and $c : X \to \bar{\mathbb{R}}$ is convex, k-positively homogeneous, and lower semicontinuous. Then there is*

$$\forall_{y \in X} : \sup_{x \in X} \Big(f(x) - c(x - y) \Big) = \begin{cases} f(y) & 1 \geq \operatorname{lip}_c(f) \\ \infty & \textit{otherwise}. \end{cases}$$

Proof. Fix an arbitrary $y_0 \in X$. From Lemma 4 we know

$$1 \geq \operatorname{lip}_c(f) \iff \forall_{\epsilon \geq 0} : \ \partial_\epsilon f(X) \subseteq \partial_\epsilon c(0).$$

Assume $\partial_\epsilon f(X) \subseteq \partial_\epsilon c(0)$ for all $\epsilon \geq 0$. Consequentially $\partial_\epsilon f(y_0) \subseteq \partial_\epsilon c(0) = \partial_\epsilon c(\,\cdot - y_0)(y_0)$ for every $\epsilon \geq 0$. From the usual difference-convex global ϵ-subdifferential condition (Hiriart-Urruty, 1989, Thm. 4.4) it follows that

$$\inf_{x \in X} \Big(c(x - y_0) - f(x) \Big) = \underbrace{c(y_0 - y_0)}_{0} - f(y_0) = -f(y_0),$$

where we note that $c(y_0 - y_0) = c(0) = 0$ because c is sublinear.

Assume $\partial_\epsilon f(X) \not\subseteq \partial_\epsilon c(0)$ for some $\epsilon \geq 0$. By hypothesis there exists $\epsilon_0 \geq 0$, $x_0 \in X$, and $x_0^* \in X^*$ with

$$x_0^* \in \partial_{\epsilon_0} f(x_0) \quad \text{and} \quad x_0^* \notin \partial_{\epsilon_0} c(0).$$

Using the Toland (1979) duality formula (viz. Hiriart-Urruty, 1986, Cor. 2.3) and the usual calculus rules for the Fenchel conjugate (e.g. Zălinescu, 2002, Thm. 2.3.1) we have

$$\begin{aligned}\inf_{x \in X} \Big(c(x - y_0) - f(x) \Big) &= \inf_{x^* \in X^*} \Big(f^*(x^*) - (c(\,\cdot - y_0))^*(x^*) \Big) \\ &= \inf_{x^* \in X^*} \Big(f^*(x^*) - c^*(x^*) + \langle y_0, x^* \rangle \Big) \\ &\leq f^*(x_0^*) - c^*(x_0^*) + \langle y_0, x_0^* \rangle \\ &\overset{\text{(B.1)}}{\leq} \epsilon_0 + \langle x_0, x_0^* \rangle - f(x_0) - c^*(x_0^*) + \langle y_0, x_0^* \rangle \\ &= \underbrace{\epsilon_0 + \langle x_0 + y_0, x_0^* \rangle - f(x_0)}_{< \infty} - c^*(x_0^*), \end{aligned} \tag{B.1}$$

where the second inequality is because $x_0^* \in \partial_{\epsilon_0} f(x_0)$.

We have assumed $x_0^* \notin \partial_\epsilon c(0) \supseteq \partial c(0)$. Because c convex k-positively homogeneous, $c^*(x_0^*) = \infty$ (via Lemma 5 (ii)). Then (B.1) yields

$$\inf_{x \in X} \Big(c(x - y_0) - f(x) \Big) \leq -\infty,$$

which completes the proof. □

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 $\operatorname{lip}_c(f) < \infty$. Then for all $r \geq 0$, there exists $\Delta_{f,c,r}(\mu) \geq 0$ so that*

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

and

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

Proof. (1): Since c is k-positively homogeneous, there is $c(x, x) = c(x - x) = c(0) = 0$ for all $x \in X$. Therefore we can apply Lemma 1 and Lemma 3 to obtain

$$\begin{aligned} \sup_{\nu \in \mathrm{B}_c(\mu, r)} \int f \,\mathrm{d}\nu &\overset{\text{L1}}{=} \inf_{\lambda \geq 0} \left(r\lambda + \int f^{\lambda c} \,\mathrm{d}\mu \right) \\ &\leq \inf_{\lambda \geq \operatorname{lip}_c(f)} \left(r\lambda + \int f^{\lambda c} \,\mathrm{d}\mu \right) \\ &\overset{\text{L3}}{=} r \operatorname{lip}_c(f) + \int f \,\mathrm{d}\mu, \end{aligned} \tag{B.2}$$

and therefore $\Delta_{f,c,r}(\mu) \geq 0$.

(2): Observing that $\overline{\operatorname{co}} f \leq f$, from Lemma 6 we find for all $x \in X$

$$\begin{aligned} &\sup_{\lambda \in [0,\infty)} (f(x) - f^{\lambda c}(x) - r\lambda) \\ &\quad = \sup_{\lambda \in [0,\infty)} (f(x) - \sup_{y \in X} (f(y) - \lambda c(x - y)) - r\lambda) \\ &\quad = \sup_{\lambda \in [0,\infty)} \inf_{y \in X} (f(x) - f(y) + \lambda c(x - y) - r\lambda) \\ &\quad \leq \sup_{\lambda \in [0,\infty)} \inf_{y \in X} (f(x) - \overline{\operatorname{co}} f(y) + \lambda c(x - y) - \lambda r) \\ &\quad \overset{\text{L6}}{=} \sup_{\lambda \in [0,\infty)} \begin{cases} f(x) - \overline{\operatorname{co}} f(x) - \lambda r & \operatorname{lip}_c(\overline{\operatorname{co}} f) \leq \lambda \\ -\infty & \operatorname{lip}_c(\overline{\operatorname{co}} f) > \lambda \end{cases} \\ &\quad = f(x) - \overline{\operatorname{co}} f(x) - r \operatorname{lip}_c(\overline{\operatorname{co}} f). \end{aligned} \tag{B.3}$$

Similarly, for all $x \in X$ there is

$$\begin{aligned} \sup_{\lambda \in [0,\infty)} \Big(f(x) - f^{\lambda c}(x) - r\lambda \Big) &\leq \sup_{\lambda \in [0,\infty)} \Big(f(x) - f^{\lambda c}(x) \Big) + \sup_{\lambda \in [0,\infty)} \Big(-r\lambda \Big) \\ &= \sup_{\lambda \in [0,\infty)} \Big(f(x) - f^{\lambda c}(x) \Big) \\ &= \sup_{\lambda \in [0,\infty)} \inf_{y \in X} \Big(f(x) - f(y) + \lambda c(x - y) \Big) \\ &\leq \inf_{y \in X} \sup_{\lambda \in [0,\infty)} \Big(f(x) - f(y) + \lambda c(x - y) \Big) \\ &= \inf_{y \in X} \begin{cases} \infty & c(x - y) > 0 \\ 0 & c(x - y) = 0 \end{cases} \\ &= 0. \end{aligned} \tag{B.4}$$

Together, (B.3) and (B.4) show

$$\int \sup_{\lambda\in[0,\infty)} (f - f^{\lambda c} - r\lambda)\,\mathrm{d}\mu \\ \leq \min\Big\{\int (f - \overline{\mathrm{co}}\, f)\,\mathrm{d}\mu - r\,\mathrm{lip}_c(\overline{\mathrm{co}}\, f), 0\Big\}. \tag{B.5}$$

Then

$$\begin{aligned}
\Delta_{f,c,r}(\mu) &= \left(r\,\mathrm{lip}_c(f) + \int f\,\mathrm{d}\mu\right) - \sup_{\nu\in \mathrm{B}_c(\mu,r)} \int f\,\mathrm{d}\nu \\
&\overset{\text{(B.2)}}{=} \left(r\,\mathrm{lip}_c(f) + \int f\,\mathrm{d}\mu\right) - \inf_{\lambda\in[0,\infty)} \left(r\lambda - \int f^{\lambda c}\,\mathrm{d}\mu\right) \\
&= r\,\mathrm{lip}_c(f) + \sup_{\lambda\in[0,\infty)} \int \big(f - f^{\lambda c} - \lambda r\big)\,\mathrm{d}\mu \\
&\leq r\,\mathrm{lip}_c(f) + \int \sup_{\lambda\in[0,\infty)} \big(f - f^{\lambda c} - \lambda r\big)\,\mathrm{d}\mu \\
&\overset{\text{(B.5)}}{\leq} r\,\mathrm{lip}_c(f) + \min\Big\{\int (f - \overline{\mathrm{co}}\, f)\,\mathrm{d}\mu - r\,\mathrm{lip}_c(\overline{\mathrm{co}}\, f), 0\Big\},
\end{aligned}$$

which implies (2). □

The extension of Theorem 1 for robust classification in the absence of label noise is straight-forward.

Corollary 1. *Assume X is a separable Banach space and Y is a topological space. Fix $\mu \in \mathfrak{P}(X \times Y)$. Assume $c : (X \times Y) \times (X \times Y) \to \bar{\mathbb{R}}$ satisfies*

$$c((x, y), (x', y')) = \begin{cases} c_0(x - x') & y = y' \\ \infty & y \neq y', \end{cases} \tag{B.6}$$

where $c_0 : X \to \bar{\mathbb{R}}$ satisfies the conditions of Theorem 1, and $f \in \mathcal{L}_1(X \times Y, \mu)$ is upper semicontinuous and has $\mathrm{lip}_c(f) < \infty$. Then for all $r \geq 0$ there is (1) *and* (2)*, where the closed convex hull is interpreted $\overline{\mathrm{co}}(f)(x, y) \overset{\text{def}}{=} \overline{\mathrm{co}}(f(\,\cdot\,, y))(x)$.*

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 $\mathrm{lip}_c(f) < \infty$, and attains its maximum on $X_0 \subseteq X$. Then for all $r \geq 0$*

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

Proof. Let $x_0 \in X_0$ be a point at which $f(x_0) = \sup f(X_0)$. Then $\mathrm{cost}_c(\delta_{x_0}, \delta_{x_0}) = 0 \leq r$, and $\sup_{\nu\in \mathrm{B}_c(\delta_{x_0}, r)} \int f\,\mathrm{d}\nu = f(x_0)$. Therefore

$$\Delta_{f,c,r}(\delta_{x_0}) = r\,\mathrm{lip}_c(f) + f(x_0) - f(x_0) = r\,\mathrm{lip}_c(f). \tag{B.2}$$

And so we have

$$\begin{aligned}
r\,\mathrm{lip}_c(f) &\overset{\text{(B.2)}}{\leq} \sup_{\mu\in\mathfrak{P}(X_0)} \Delta_{f,c,r}(\mu) \\
&\overset{\text{TI}}{\leq} r\,\mathrm{lip}_c(f) - \max\Big\{r\,\mathrm{lip}_c(\overline{\mathrm{co}}\, f) - \rho(f), 0\Big\} \\
&\leq r\,\mathrm{lip}_c(f),
\end{aligned}$$

which implies the claim. □

B.2. Proof of Theorem 2

Lemma 7 will be used to show an equality result in Theorem 2.

Lemma 7. *Assume (Ω, c) is a compact Polish space and $\mu \in \mathfrak{P}(\Omega)$ is non-atomic. For $r > 0$ and $\nu^\star \in \mathrm{B}_c(\mu, r)$ there is a sequence $(f_i)_{i\in\mathbb{N}} \subseteq A_\mu(r) \stackrel{\text{def}}{=} \left\{ f \in \mathcal{L}_0(\Omega, \Omega) \;\middle|\; \int c \,\mathrm{d}(\mathrm{Id}, f)_{\#}\mu \leq r \right\}$ with $(f_i)_{\#}\mu$ converging at $\nu^\star$ in $\sigma(\mathfrak{P}(\Omega), \mathrm{C}(\Omega))$.*

Proof. Let $P(\mu, \nu) \stackrel{\text{def}}{=} \{f \in \mathcal{L}_0(X, X) \mid f_{\#}\mu = \nu\}$. Since μ is non-atomic and c is continuous we have (via Pratelli, 2007, Thm. B)

$$\forall_{\nu\in\mathfrak{P}(\Omega)} : \inf_{f\in P(\mu,\nu)} \int c \,\mathrm{d}(\mathrm{Id}, f)_{\#}\mu = \mathrm{cost}_c(\mu, \nu).$$

Let $r^\star \stackrel{\text{def}}{=} \mathrm{cost}_c(\mu, \nu^\star)$, obviously $r^\star \leq r$. Assume $r^\star > 0$, otherwise the lemma is trivial. Fix a sequence $(\epsilon_k)_{k\in\mathbb{N}} \subseteq (0, r^\star)$ with $\epsilon_k \to 0$. For $u \geq 0$ let $\nu(u) \stackrel{\text{def}}{=} \mu + u(\nu^\star - \mu)$. Then

$$\mathrm{cost}_c(\mu, \nu(0)) = 0 \quad \text{and} \quad \mathrm{cost}_c(\mu, \nu(1)) = r^\star,$$

and because cost_c metrises the $\sigma(\mathfrak{P}(\Omega), \mathrm{C}(\Omega))$-topology on $\mathfrak{P}(\Omega)$ (Villani, 2009, Cor. 6.13), the mapping $u \mapsto \mathrm{cost}_c(\mu, \nu(u))$ is $\sigma(\mathfrak{P}(\Omega), \mathrm{C}(\Omega))$-continuous. Then by the intermediate value theorem for every $k \in \mathbb{N}$ there is some $u_k > 0$ with $\mathrm{cost}_c(\mu, \nu(u_k)) = r^\star - \epsilon_k$, forming a sequence $(u_k)_{k\in\mathbb{N}} \subseteq [0, 1]$. Then for every k there is a sequence $(f_{jk})_{j\in\mathbb{N}} \subseteq P(\mu, \nu(u_k))$ so that $(f_{jk})_{\#}\mu \to \nu(k)$ in $\sigma(\mathfrak{P}(\Omega), \mathrm{C}(\Omega))$ and

$$\begin{aligned}
\lim_{j\in\mathbb{N}} \int c \,\mathrm{d}(\mathrm{Id}, f_{jk})_{\#}\mu &= \inf_{f\in P(\mu,\nu(k))} \int c \,\mathrm{d}(\mathrm{Id}, f_k)_{\#}\mu \\
&= \mathrm{cost}_c(\mu, \nu(k)) \\
&= r^\star - \epsilon_k.
\end{aligned}$$

Therefore for every $k \in \mathbb{N}$ there exists $j_k \geq 0$ so that for every $j \geq j_k$

$$\int c \,\mathrm{d}(\mathrm{Id}, f_{jk})_{\#}\mu \leq r^\star. \tag{B.2}$$

Let us pass directly to this subsequence of $(f_{jk})_{j\in\mathbb{N}}$ for every $k \in \mathbb{N}$ so that (B.2) holds for all $j, k \in \mathbb{N}$. Next by construction we have $\nu(u_k) \to \nu^\star$. Therefore $(f_{jk})_{j,k\in N}$ has a subsequence in k so that $(f_{jk})_{\#}\mu \to \nu^\star$ in in $\sigma(\mathfrak{P}(\Omega), \mathrm{C}(\Omega))$. By ensuring (B.2) is satisfied, the sequences $(f_{jk})_{j\in\mathbb{N}} \subseteq A_\mu(r)$ for every $k \in \mathbb{N}$. □

We can now prove our main result Theorem 2. When (X, c) is a normed space, the closed ball of radius $r \geq 0$, centred at $x \in X$ is denoted $\mathrm{B}_c(x, r) \stackrel{\text{def}}{=} \{y \in X \mid c(x - y) \leq r\}$.

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*

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

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

Proof. For convenience of notation let $c \stackrel{\text{def}}{=} c_0$.

When $r = 0$, the set $R_\mu(r)$ consists of the set of functions g which are 0 μ-almost everywhere, in which case $\mathrm{B}_c(x, g(x)) = \{0\}$ for μ-almost all $x \in X$. Thus (5) is equal to $\int f(x)\mu(\mathrm{d}x)$. Since c is a norm, $c(0) = 0$, and by a similar argument there is equality with the right hand side. We now complete the proof for the cases where $r > 0$.

Inequality: For $g \in R_\mu(r)$, let $\Gamma_g : X \to 2^X$ denote the set-valued mapping with $\Gamma_g(x) \stackrel{\text{def}}{=} \mathrm{B}_c(x, g(x))$. Let $\mathcal{L}_0(X, \Gamma_g)$ denote the set of Borel $a : X \to X$ so that $a(x) \in \Gamma_g(x)$ for μ-almost all $x \in X$. Let $A_\mu(r) \stackrel{\text{def}}{=} \bigcup_{g\in\mathbb{R}_\mu(r)} \mathcal{L}_0(X, \Gamma_g)$. Clearly for every $a \in A_\mu(r)$ there is

$$r \geq \int c(x, a(x)) \,\mathrm{d}\mu = \int c \,\mathrm{d}(\mathrm{Id}, a)_{\#}\mu,$$

which shows $\{a_{\#}\mu \mid a \in A_\mu(r)\} \subseteq \mathrm{B}_c(\mu, r)$. Then if there is equality in (B.3), we have

$$\begin{aligned}\sup_{g\in R_\mu(r)} \int \sup_{x'\in \Gamma_g(x)} f(x) &= \sup_{g\in R_\mu(r)} \sup_{a\in\mathcal{L}_0(X,\Gamma_g)} \int f \,\mathrm{d}a_{\#}\mu \\ &= \sup_{a\in A_\mu(r)} \int f \,\mathrm{d}a_{\#}\mu \\ &\leq \sup_{\nu\in \mathrm{B}_c(\mu,r)} \int f \,\mathrm{d}\nu,\end{aligned} \tag{B.3}$$

which proves the inequality.

To complete the proof we will now justify the exchange of integration and supremum in (B.3). The set $\mathcal{L}_0(X, \Gamma_g)$ is trivially decomposable (Giner, 2009, see the remark at the bottom of p. 323, Def. 2.1). By assumption f is Borel measurable. Since f is measurable, any decomposable subset of $\mathcal{L}_0(X, X)$ is f-decomposable (Giner, 2009, Prop. 5.3) and f-linked (Giner, 2009, Prop. 3.7 (i)). Giner (2009, Thm. 6.1 (c)) therefore allows us to exchange integration and supremum in (B.3).

Equality: Under the additional assumptions there exists $\nu^\star \in \mathfrak{P}(\Omega)$ with (via Blanchet & Murthy, 2019, Prop. 2)

$$\int f \,\mathrm{d}\nu^\star = \sup_{\nu\in \mathrm{B}_c(\mu,r)} \int f \,\mathrm{d}\nu.$$

The compact subset where μ is concentrated and non-atomic is a Polish space with the Banach metric. Therefore using Lemma 7 there is a sequence $(f_i)_{i\in\mathbb{N}} \subseteq A_\mu(r)$ so that

$$\lim_{i\in\mathbb{N}} \int f_i \,\mathrm{d}\mu = \int f \,\mathrm{d}\nu^\star = \sup_{\nu\in \mathrm{B}_c(\mu,r)} \int f \,\mathrm{d}\nu,$$

proving the desired equality. □

C. Proofs and additional results on the Lipschitz regularisation of kernel methods

C.1. Random sampling requires exponential cost

The most natural idea of leveraging the samples is to add the constraints $\|g(w^s)\| \leq L$. For Gaussian kernel, we may sample from $\mathcal{N}(\mathbf{0}, \sigma^2 I)$ while for inverse kernel we may sample uniformly from B. This leads to our training objective:

$$\min_{f\in\mathcal{H}} \quad \frac{1}{l}\sum_{i=1}^{l} \mathrm{loss}(f(x^i), y^i) + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2 \qquad s.t. \qquad \|g(w^s)\| \leq L, \quad \forall s \in [n].$$

Unfortunately, this method may require $O(\frac{1}{\epsilon^d})$ samples to guarantee $\sum_j \|g_j\|_{\mathcal{H}}^2 \leq L^2 + \epsilon$ w.h.p. This is illustrated in Figure 8, where k is the polynomial kernel with degree 2 whose domain X is the unit ball B, and $f(x) = \frac{1}{2}(v^\top x)^2$. We seek to test whether the gradient $g(x) = (v^\top x)v$ has norm bounded by 1 for all $x \in B$, and we are only allowed to test whether $\|g(w^s)\| \leq 1$ for samples w^s that are drawn uniformly at random from B. This is equivalent to testing $\|v\| \leq 1$, and to achieve it at least one w^s must be from the ϵ ball around $v/\|v\|$ or $-v/\|v\|$, intersected with B. But the probability of hitting such a region decays exponentially with the dimensionality d.

The key insight from the above counter-example is that in fact $\|v\|$ can be easily computed by $\sum_{s=1}^{d}(v^\top \tilde{w}_s)^2$, where $\{\tilde{w}^s\}_{s=1}^d$ is the *orthonormal* basis computed from the Gram–Schmidt process on d random samples $\{w^s\}_{s=1}^d$ ($n = d$). With probability 1, n samples drawn uniformly from B must span $\mathbb{R}^d$ as long as $n \geq d$, i.e., $\mathrm{rank}(W) = d$ where $W = (w^1, \ldots, w^n)$. The Gram–Schmidt process can be effectively represented using a pseudo-inverse matrix (allowing $n > d$) as

$$\|v\|_2 = \left\|(W^\top W)^{-1/2} W^\top v\right\|_2,$$

where $(W^\top W)^{-1/2}$ is the square root of the pseudo-inverse of $W^\top W$. This is exactly the intuition underlying the Nyström approximation that we will leveraged.

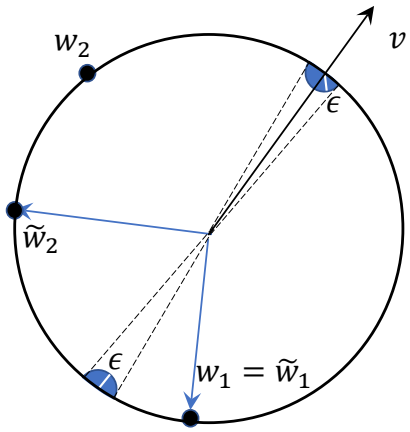


Figure 8: Suppose we use a polynomial kernel with degree 2, and $f(x) = \frac{1}{2}(v^\top x)^2$ for $x \in B$. Then $g(x) = (v^\top x)v$. If we want to test whether $\sup_{x\in B} \|g(x)\|_2 \le 1$ by evaluating $\|g(w)\|_2$ on w that is randomly sampled from B such as w_1 and w_2, we must sample within the ϵ balls around the intersection of B and the ray along v (both directions). See the blue shaded area. The problem, however, becomes trivial if we use the orthonormal basis $\{\tilde{w}_1, \tilde{w}_2\}$.

C.2. Spectrum of Kernels

Let k be a continuous kernel on a compact metric space X, and μ be a finite Borel measure on X with $supp[\mu] = X$. We will re-describe the following spectral properties in a more general way than in §4. Recall Steinwart & Christmann (2008, §4) that the integral operator for k and μ is defined by

$$\begin{aligned} T_k &= I_k \circ S_k : \mathcal{L}_2(X,\mu) \to \mathcal{L}_2(X,\mu) \\ \text{where } S_k &: L_2(X,\mu) \to \mathrm{C}(X), \quad (S_k f)(x) = \int k(x,y) f(y) d\mu(y), \quad f \in \mathcal{L}_2(X,\mu), \\ I_k &: \mathrm{C}(X) \hookrightarrow \mathcal{L}_2(X,\mu), \text{ inclusion operator.} \end{aligned}$$

By the spectral theorem, if T_k is compact, then there is an at most countable orthonormal set (ONS) $\{\tilde{e}_j\}_{j\in J}$ of $\mathcal{L}_2(X,\mu)$ and $\{\lambda_j\}_{j\in J}$ with $\lambda_1 \ge \lambda_2 \ge \ldots > 0$ such that

$$Tf = \sum_{j\in J} \lambda_j \langle f, \tilde{e}_j \rangle_{\mathcal{L}_2(X,\mu)} \tilde{e}_j, \qquad f \in \mathcal{L}_2(X,\mu).$$

In particular, we have $\langle \tilde{e}_i, \tilde{e}_j \rangle_{\mathcal{L}_2(X,\mu)} = \delta_{ij}$ (i.e., equals 1 if $i = j$, and 0 otherwise), and $T\tilde{e}_i = \lambda_i \tilde{e}_i$. Since $\tilde{e}_j$ is an equivalent class instead of a single function, we assign a set of continuous functions $e_j = \lambda_j^{-1} S_k \tilde{e}_j \in \mathrm{C}(X)$, which clearly satisfies

$$\langle e_i, e_j \rangle_{\mathcal{L}_2(X,\mu)} = \delta_{ij}, \quad Te_j = \lambda_j e_j.$$

We will call λ_j and e_j as eigenvalues and eigenfunctions respectively, and $\{e_j\}_{j\in J}$ clearly forms an ONS. By Mercer's theorem,

$$k(x,y) = \sum_{j\in J} \lambda_j e_j(x) e_j(y), \tag{C.1}$$

and all functions in $\mathcal{H}$ can be represented by $\sum_{j\in J} a_j e_j$ where $\{a_j/\sqrt{\lambda_j}\} \in \ell^2(J)$. The inner product in $\mathcal{H}$ is equivalent to $\left\langle \sum_{j\in J} a_j e_j, \sum_{j\in J} b_j e_j \right\rangle_{\mathcal{H}} = \sum_{j\in J} a_j b_j / \lambda_j$. Therefore it is easy to see that

$$\varphi_j \stackrel{\text{def}}{=} \sqrt{\lambda_j} e_j, \qquad j \in J$$

is an orthonormal basis of $\mathcal{H}$, with Moreover, for all $f \in \mathcal{H}$ with $f = \sum_{j\in J} a_j e_j$, we have $\langle f, e_j\rangle_{\mathcal{H}} = a_j/\lambda_j$, $\langle f, \varphi_j\rangle_{\mathcal{H}} = a_j/\sqrt{\lambda_j}$, and

$$f = \sum_j \langle f, \varphi_j\rangle_{\mathcal{H}}\, \varphi_j = \sum_j \sqrt{\lambda_j}\, \langle f, e_j\rangle_{\mathcal{H}}\, \varphi_j = \sum_j \lambda_j\, \langle f, e_j\rangle_{\mathcal{H}}\, e_j.$$

Most kernels used in machine learning are infinite dimensional, i.e., $J = \mathbb{N}$. For convenience, we define $\Phi_m \stackrel{\text{def}}{=} (\varphi_1, \ldots, \varphi_m)$ and $\Lambda_m = \operatorname{diag}(\lambda_1, \ldots, \lambda_m)$.

C.3. General sample complexity and assumptions on the product kernel

In this section, we first consider kernels k_0 with **scalar input**, i.e., $X_0 \subseteq \mathbb{R}$. Assume there is a measure μ_0 on X_0. This will serve as the basis for the more general product kernels in the form of $k(x, y) = \prod_{j=1}^d k_0(x_j, y_j)$ defined over X_0^d.

With Assumptions 1 and 2, we now state the formal version of Theorem 3 by first providing the sample complexity for approximating the partial derivatives. In the next subsection, we will examine how three different kernels satisfy/unsatisfy the Assumptions 1 and 2, and what the value of N_ϵ is. For each case, we will specify μ_0 on X_0, and the measure on X_0^d is trivially $\mu = \mu_0^d$.

Theorem 5. *Suppose $\{w^s\}_{s=1}^n$ are drawn iid from μ_0 on X_0, where μ_0 is the uniform distribution on $[-v/2, v/2]$ for periodic kernels or periodized Gaussian kernels. Let $Z \stackrel{\text{def}}{=} (k_0(w^1, \cdot), k_0(w^2, \cdot), \ldots, k_0(w^n, \cdot))$, and $g_1 = \frac{1}{l}\sum_{a=1}^l \gamma_a g_1^a$: $X_0^d \to \mathbb{R}$, where $\|\boldsymbol{\gamma}\|_\infty \le c_1$ and*

$$g_1^a(y) = \partial^{0,1} k(x^a, y) = h_1^a(y_1) \prod_{j=2}^d k_0(x_j^a, y_j) \quad \textit{with} \quad h_1^a(\cdot) \stackrel{\text{def}}{=} \partial^{0,1} k_0(x_1^a, \cdot).$$

Given $\epsilon \in (0, 1]$, let $\Phi_m = (\varphi_1, \ldots \varphi_m)$ where $m = N_\epsilon$. Then with probability $1 - \delta$, the following holds when the sample size $n = \max(N_\epsilon, \frac{5}{3\epsilon^2} N_\epsilon Q_\epsilon^2 \log \frac{2N_\epsilon}{\delta})$:

$$\|g_1\|_{\mathcal{H}}^2 \le \frac{1}{l^2}\boldsymbol{\gamma}^\top K_1 \boldsymbol{\gamma} + 3c_1\Big(1 + 2\sqrt{N_\epsilon} M_\epsilon\Big)\epsilon, \tag{C.2}$$

$$\textit{where} \quad (K_1)_{a,b} = (h_1^a)^\top Z (Z^\top Z)^{-1} Z^\top h_1^b \prod_{j=2}^d k_0(x_j^a, x_j^b).$$

Then we obtain the formal statement of sample complexity, as stated in the following corollary, by combining all the coordinates from Theorem 5.

Corollary 2. *Suppose all coordinates share the same set of samples $\{w^s\}_{s=1}^n$. Applying the results in* (C.2) *for coordinates from 1 to d and using the union bound, we have that with sample size $n = \max(N_\epsilon, \frac{5}{3\epsilon^2} N_\epsilon Q_\epsilon^2 \log \frac{2N_\epsilon}{\delta})$, the following holds with probability $1 - d\delta$,*

$$\lambda_{\max}(G^\top G) \le \lambda_{\max}(\tilde{P}_G) + 3c_1\Big(1 + 2\sqrt{N_\epsilon} M_\epsilon\Big)\epsilon. \tag{C.3}$$

Equivalently, if N_ϵ, M_ϵ and Q_ϵ are constants or poly-log terms of ϵ which we treat as constant, then to ensure $\lambda_{\max}(G^\top G) \le \lambda_{\max}(\tilde{P}_G) + \epsilon$ with probability $1 - \delta$, the sample size needs to be

$$n = \frac{15}{\epsilon^2} c_1^2 \Big(1 + 2\sqrt{N_\epsilon} M_\epsilon\Big)^2 N_\epsilon Q_\epsilon^2 \log \frac{2dN_\epsilon}{\delta}.$$

Remark 4. The first term on the right-hand side of (C.3) is explicitly upper bounded by L^2 in our training objective. In the case of Theorem 6, the values of Q_ϵ, N_ϵ, and M_ϵ lead to a $\tilde{O}(\frac{1}{\epsilon^2})$ sample complexity. If we further zoom into the dependence on the period v, then note that N_ϵ is almost a universal constant while $M_\epsilon = \frac{\sqrt{2}\pi}{v}(N_\epsilon - 1)$. So overall, n depends on v by $\frac{1}{v^2}$. This is not surprising because smaller period means higher frequency, hence more samples are needed.

Remark 5. Corollary 2 postulates that all coordinates share the same set of samples $\{w^s\}_{s=1}^n$. When coordinates differ in their domains, we can draw different sets of samples for them. The sample complexity hence grows by d times as we only use a weak union bound. More refined analysis could save us a factor of d as these sets of samples are independent of each other.

Proof of Theorem 5. Let $\epsilon' \stackrel{\text{def}}{=} (1 + 2\sqrt{m}M_\epsilon)\epsilon$. Since

$$\left\langle g_1^a, g_1^b \right\rangle_{\mathcal{H}} = \left\langle h_1^a, h_1^b \right\rangle_{\mathcal{H}_0} \prod_{j=2}^{d} k_0(x_j^a, x_j^b)$$

and $\left|k_0(x_j^a, x_j^b)\right| \le 1$, it suffices to show that for all $a, b \in [l]$,

$$\left| \left\langle h_1^a, h_1^b \right\rangle_{\mathcal{H}_0} - (h_1^a)^\top Z(Z^\top Z)^{-1} Z^\top h_1^b \right| \le 3\epsilon'.$$

Towards this end, it is sufficient to show that for any $h(\cdot) = \theta_x \partial^{0,1} k_0(x, \cdot) + \theta_y \partial^{0,1} k_0(y, \cdot)$ where $x, y \in X_0$ and $|\theta_x| + |\theta_y| \le 1$, we have

$$\left| h^\top Z(Z^\top Z)^{-1} Z^\top h - \|h\|_{\mathcal{H}_0}^2 \right| \le \epsilon'. \tag{C.4}$$

This is because, if so, then

$$\begin{aligned}
&\left| \left\langle h_1^a, h_1^b \right\rangle_{\mathcal{H}_0} - (h_1^a)^\top Z(Z^\top Z)^{-1} Z^\top h_1^b \right| \\
&\quad = \Big| \frac{1}{2} \Big(\left\| h_1^a + h_1^b \right\|_{\mathcal{H}_0}^2 - \|h_1^a\|_{\mathcal{H}_0}^2 - \left\| h_1^b \right\|_{\mathcal{H}_0}^2 \Big) \\
&\qquad - \frac{1}{2} \Big[(h_1^a + h_1^b)^\top Z(Z^\top Z)^{-1} Z^\top (h_1^a + h_1^b) \\
&\qquad - (h_1^a)^\top Z(Z^\top Z)^{-1} Z^\top h_1^a - (h_1^b)^\top Z(Z^\top Z)^{-1} Z^\top h_1^b \Big] \Big| \\
&\quad \le \frac{1}{2}(4\epsilon' + \epsilon' + \epsilon') \\
&\quad = 3\epsilon'.
\end{aligned}$$

The rest of the proof is devoted to (C.4). Since $n \ge m$, the SVD of $\Lambda_m^{-1/2} \Phi_m^\top Z$ can be written as $U \Sigma V^\top$, where $UU^\top = U^\top U = V^\top V = I_m$ (m-by-m identity matrix), and $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_m)$. Define

$$\boldsymbol{\alpha} = n^{-1/2} V U^\top \Lambda_m^{-1/2} \Phi_m^\top h.$$

Consider the optimization problem $o(\boldsymbol{\alpha}) \stackrel{\text{def}}{=} \frac{1}{2} \|Z\boldsymbol{\alpha} - h\|_{\mathcal{H}_0}^2$. It is easy to see that its minimal objective value is $o^* \stackrel{\text{def}}{=} \frac{1}{2}\|h\|_{\mathcal{H}_0}^2 - \frac{1}{2} h^\top Z(Z^\top Z)^{-1} Z^\top h$. So

$$0 \le 2o^* = \|h\|_{\mathcal{H}_0}^2 - h^\top Z(Z^\top Z)^{-1} Z^\top h \le 2o(\boldsymbol{\alpha}).$$

Therefore to prove (C.4), it suffices to bound $o(\boldsymbol{\alpha}) = \|Z\boldsymbol{\alpha} - h\|_{\mathcal{H}_0}$. Since $\sqrt{n}\Phi_m \Lambda^{1/2} U V^\top \boldsymbol{\alpha} = \Phi_m \Phi_m^\top h$, we can decompose $\|Z\boldsymbol{\alpha} - h\|_{\mathcal{H}_0}$ by

$$\begin{aligned}
\|Z\boldsymbol{\alpha} - h\|_{\mathcal{H}_0} \le{}& \left\| (Z - \Phi_m \Phi_m^\top Z)\boldsymbol{\alpha} \right\|_{\mathcal{H}_0} \\
&+ \left\| (\Phi_m \Phi_m^\top Z - \sqrt{n}\Phi_m \Lambda_m^{1/2} U V^\top)\boldsymbol{\alpha} \right\|_{\mathcal{H}_0} \\
&+ \left\| \Phi_m \Phi_m^\top h - h \right\|_{\mathcal{H}_0}.
\end{aligned} \tag{C.5}$$

The last term $\left\|\Phi_m \Phi_m^\top h - h\right\|_{\mathcal{H}_0}$ is clearly below ϵ because by Assumption 1 and $m = N_\epsilon$

$$\begin{aligned}
\left\|\Phi_m \Phi_m^\top h - h\right\|_{\mathcal{H}_0} &\le |\theta_x| \left\| \Phi_m \Phi_m^\top \partial^{0,1} k_0(x, \cdot) - \partial^{0,1} k_0(x, \cdot) \right\|_{\mathcal{H}_0} \\
&\quad + |\theta_y| \left\| \Phi_m \Phi_m^\top \partial^{0,1} k_0(y, \cdot) - \partial^{0,1} k_0(y, \cdot) \right\|_{\mathcal{H}_0} \\
&\le (|\theta_x| + |\theta_y|)\epsilon \\
&\le \epsilon.
\end{aligned}$$

We will next bound the first two terms on the right-hand side of (C.5).

(i) By Assumption 1, $\left\|k_0(w^s,\cdot) - \Phi_m\Phi_m^\top k_0(w^s,\cdot)\right\|_{\mathcal{H}_0} \le \epsilon$, hence

$$\left\|(Z - \Phi_m\Phi_m^\top Z)\boldsymbol{\alpha}\right\|_{\mathcal{H}_0} \le \epsilon\sqrt{n}\left\|\boldsymbol{\alpha}\right\|_2.$$

To bound $\left\|\boldsymbol{\alpha}\right\|_2$, note all singular values of $VU^\top$ are 1, and so Assumption 2 implies that for all $i \in [m]$,

$$\begin{aligned}
\left|\lambda_j^{-1/2}\left\langle\varphi_j, h\right\rangle_{\mathcal{H}_0}\right| &= \left|\left\langle e_j, h\right\rangle_{\mathcal{H}_0}\right| \\
&= \left|\left\langle e_j, \theta_x\partial^{0,1}k_0(x,\cdot) + \theta_y\partial^{0,1}k_0(y,\cdot)\right\rangle_{\mathcal{H}_0}\right| \\
&\le \sup_{x\in X}\left|\left\langle e_j, \partial^{0,1}k(x,\cdot)\right\rangle_{\mathcal{H}_0}\right| \\
&\le M_\epsilon. \qquad \text{(C.6)}
\end{aligned}$$

As a result,

$$\left\|(Z - \Phi_m\Phi_m^\top Z)\boldsymbol{\alpha}_j\right\|_{\mathcal{H}_0} \le \epsilon n^{1/2}\cdot n^{-1/2}\left\|\Lambda_m^{-1/2}\Phi_m^\top h\right\| \le \epsilon\sqrt{m}M_\epsilon.$$

(ii) We first consider the concentration of the matrix

$$R \overset{\text{def}}{=} \frac{1}{n}\Lambda_m^{-1/2}\Phi_m^\top ZZ^\top\Phi_m\Lambda_m^{-1/2} \in \mathbb{R}^{m\times m}.$$

Clearly,

$$\mathop{\mathbb{E}}_{\{w_s\}}[R_{ij}] = \mathop{\mathbb{E}}_{\{w_s\}}\left[\frac{1}{n}\sum_{s=1}^n e_i(w_s)e_j(w_s)\right] = \int e_i(x)e_j(x)\,\mathrm{d}\mu(x) = \delta_{ij}.$$

By matrix Bernstein theorem (Tropp, 2015, Theorem 1.6.2), we have

$$\Pr\Big(\left\|R - I_m\right\|_{sp} \le \epsilon\Big) \ge 1-\delta$$

when $n \ge O(.)$. This is because

$$\left\|(e_1(x),\ldots,e_m(x))\right\|^2 \le mQ_\epsilon^2, \quad \left\|\mathbb{E}_{\{w_s\}}[RR^\top]\right\|_{sp} \le mQ_\epsilon^2/n,$$

and

$$\begin{aligned}
\Pr\Big(\left\|R - I_m\right\|_{sp} \le \epsilon\Big) &\ge 1 - 2m\exp\left(\frac{-\epsilon^2}{\frac{mQ_\epsilon^2}{n}\left(1+\frac{2}{3}\epsilon\right)}\right) \\
&\ge 1 - 2m\exp\left(\frac{-\epsilon^2}{\frac{5mQ_\epsilon^2}{3n}}\right) \\
&\ge 1-\delta,
\end{aligned}$$

where the last step is by the definition of n. Since $R = \frac{1}{n}U\Sigma^2U^\top$, this means with probability $1-\delta$, $\left\|\frac{1}{n}U\Sigma^2U^\top - I_m\right\|_{sp} \le \epsilon$. So for all $i \in [m]$,

$$\left|\frac{1}{n}\sigma_i^2 - 1\right| \le \epsilon \implies \left|\frac{1}{\sqrt{n}}\sigma_i - 1\right| < \epsilon\left|\frac{1}{\sqrt{n}}\sigma_i + 1\right|^{-1} \le \epsilon. \qquad \text{(C.7)}$$

Moreover, $\lambda_1 \leq 1$ since $k_0(x,x) = 1$. It then follows that

$$\begin{aligned}
&\left\|(\Phi_m\Phi_m^\top Z - \sqrt{n}\Phi_m\Lambda_m^{1/2}UV^\top)\boldsymbol{\alpha}\right\|_{\mathcal{H}_0} \\
=&\left\|\Phi_m\Lambda_m^{1/2}U\Sigma V^\top\frac{1}{\sqrt{n}}VU^\top\Lambda_m^{-1/2}\Phi_m^\top h - \sqrt{n}\Phi_m\Lambda_m^{1/2}UV^\top\frac{1}{\sqrt{n}}VU^\top\Lambda_m^{-1/2}\Phi_m^\top h\right\|_{\mathcal{H}_0} \\
=&\left\|\Lambda_m^{1/2}U\left(\frac{1}{\sqrt{n}}\Sigma - I_m\right)U^\top\Lambda_m^{-1/2}\Phi_m^\top h\right\|_2 \qquad (\text{because } \Phi_m^\top\Phi_m = I_m) \\
\leq&\sqrt{\lambda_1}\max_{i\in[m]}\left|\frac{1}{\sqrt{n}}\sigma_i - 1\right|\left\|\Lambda_m^{-1/2}\Phi_m^\top h\right\|_2 \\
\leq&\epsilon\sqrt{m}M_\epsilon \qquad (\text{by (C.7), (C.6), and } \lambda_1 \leq 1).
\end{aligned}$$

Combining (i) and (ii), we arrive at the desired bound in (C.2). □

Proof of Corollary 2. Since $\tilde{P}_G$ approximates $G^\top G$ only on the diagonal, $\tilde{P}_G - G^\top G$ is a diagonal matrix which we denote as $\mathrm{diag}(\delta_1, \ldots, \delta_d)$. Let $\mathbf{u} \in \mathbb{R}^d$ be the leading eigenvector of $\tilde{P}_G$. Then

$$\begin{aligned}
\lambda_{\max}(\tilde{P}_G) - \lambda_{\max}(G^\top G) &\leq \mathbf{u}^\top\tilde{P}_G\mathbf{u} - \mathbf{u}^\top G^\top G\mathbf{u} = \mathbf{u}^\top(\tilde{P}_G - G^\top G)\mathbf{u} = \sum_j \delta_j\mathbf{u}_j^2 \\
&\text{(by (C.2))} \quad \leq 3c_1\left(1 + 2\sqrt{N_\epsilon}M_\epsilon\right)\epsilon.
\end{aligned}$$

The proof is completed by applying the union bound and rewriting the results. □

C.4. Case 1: Checking Assumptions 1 and 2 on periodic kernels

Periodic kernels on $X_0 \stackrel{\text{def}}{=} \mathbb{R}$ are translation invariant, and can be written as $k_0(x,y) = \kappa(x-y)$ where $\kappa : \mathbb{R} \to \mathbb{R}$ is a) periodic with period v; b) even, with $\kappa(-t) = \kappa(t)$; and c) normalized with $\kappa(0) = 1$. A general treatment was given by (Williamson et al., 2001), and an example was given by David MacKay in (MacKay, 1998):

$$k_0(x,y) = \exp\left(-\frac{1}{2\sigma^2}\sin\left(\frac{\pi}{v}(x-y)\right)^2\right). \tag{C.8}$$

We define μ_0 to be a uniform distribution on $[-\frac{v}{2}, \frac{v}{2}]$, and let $\omega_0 = 2\pi/v$.

Since κ is symmetric, we can simplify the Fourier transform of $\kappa(t)\delta_v(t)$, where $\delta_v(t) = 1$ if $t \in [-v/2, v/2]$, and 0 otherwise:

$$F(\omega) = \frac{1}{\sqrt{2\pi}}\int_{-v/2}^{v/2}\kappa(t)\cos(\omega t)\,\mathrm{d}t.$$

It is now easy to observe that thanks to periodicity and symmetry of κ, for all $j \in \mathbb{Z}$,

$$\begin{aligned}
&\frac{1}{v}\int_{-v/2}^{v/2}k_0(x,y)\cos(j\omega_0 y)\,\mathrm{d}y = \frac{1}{v}\int_{-v/2}^{v/2}\kappa(x-y)\cos(j\omega_0 y)\,\mathrm{d}y \\
=&\frac{1}{v}\int_{x-v/2}^{x+v/2}\kappa(z)\cos(j\omega_0(x-z))\,\mathrm{d}z \quad (\text{note } \cos(j\omega_0(x-z)) \text{ also has period } v) \\
=&\frac{1}{v}\int_{-v/2}^{v/2}\kappa(z)[\cos(j\omega_0 x)\cos(j\omega_0 z) + \sin(j\omega_0 x)\sin(j\omega_0 z))\,\mathrm{d}z \quad (\text{by periodicity}) \\
=&\frac{1}{v}\cos(j\omega_0 x)\int_{-v/2}^{v/2}\kappa(z)\cos(j\omega_0 z)\,\mathrm{d}z \quad (\text{by symmetry of } \kappa) \\
=&\frac{\sqrt{2\pi}}{v}F(j\omega_0)\cos(j\omega_0 x).
\end{aligned}$$

And similarly,

$$\frac{1}{v}\int_{-v/2}^{v/2} k_0(x,y)\sin(j\omega_0 y)\,\mathrm{d}y = \frac{\sqrt{2\pi}}{v}F(j\omega_0)\sin(j\omega_0 x).$$

Therefore the eigenfunctions of the integral operator T_k are

$$e_0(x) = 1, \qquad e_j(x) \stackrel{\text{def}}{=} \sqrt{2}\cos(j\omega_0 x), \qquad e_{-j}(x) \stackrel{\text{def}}{=} \sqrt{2}\sin(j\omega_0 x) \quad (j \geq 1)$$

and the eigenvalues are $\lambda_j = \frac{\sqrt{2\pi}}{v}F(j\omega_0)$ for all $j \in \mathbb{Z}$ with $\lambda_{-j} = \lambda_j$. An important property our proof will rely on is that

$$e_j'(x) = -j\omega_0 e_{-j}(x), \quad \text{for all } j \in \mathbb{Z}.$$

Applying Mercer's theorem in (C.1) and noting $\kappa(0) = 1$, we derive $\sum_{j\in\mathbb{Z}} \lambda_j = 1$.

Checking the Assumptions 1 and 2. The following theorem summarizes the assumptions and conclusions regarding the satisfaction of Assumptions 1 and 2. Again we focus on the case of $X \subseteq \mathbb{R}$.

Theorem 6. *Suppose the periodic kernel with period v has eigenvalues λ_j that satisfies*

$$\lambda_j(1+j)^2\max(1,j^2)(1+\delta(j\geq 1)) \leq c_6 \cdot c_4^{-j}, \quad \textit{for all } j \geq 0, \tag{C.9}$$

where $c_4 > 1$ and $c_6 > 0$ are universal constants. Then Assumption 1 holds with

$$N_\epsilon = 1 + 2\lfloor n_\epsilon \rfloor, \quad \textit{where} \quad n_\epsilon \stackrel{\text{def}}{=} \log_{c_4}\left(\frac{2.1c_6}{\epsilon^2}\max\left(1, \frac{v^2}{4\pi^2}\right)\right). \tag{C.10}$$

In addition, Assumption 2 holds with $Q_\epsilon = \sqrt{2}$ and $M_\epsilon = \frac{2\sqrt{2}\pi}{v}\lfloor n_\epsilon \rfloor = \frac{\sqrt{2}\pi}{v}(N_\epsilon - 1)$.

For example, if we set $v = \pi$ and $\sigma^2 = 1/2$ in the kernel in (C.8), elementary calculation shows that the condition (C.9) is satisfied with $c_4 = 2$ and $c_6 = 1.6$.

Proof of Theorem 6. First we show that $h(x) \stackrel{\text{def}}{=} \partial^{0,1}k_0(x_0,x)$ is in $\mathcal{H}_0$ for all $x_0 \in X_0$. Since $k_0(x_0,x) = \sum_{j\in\mathbb{Z}} \lambda_j e_j(x_0)e_j(x)$, we derive

$$h(x) = \sum_{j\in\mathbb{Z}} \lambda_j e_j(x_0)\partial^1 e_j(x) = \sum_{j\in\mathbb{Z}} \lambda_j e_j(x_0)(-j\omega_0 e_{-j}(x)) = \omega_0 \sum_{j\in\mathbb{Z}} \lambda_j j e_{-j}(x_0)e_j(x). \tag{C.11}$$

$h(x)$ is in $\mathcal{H}$ if the sequence $\lambda_j j e_{-j}(x_0)/\sqrt{\lambda_j}$ is square summable. This can be easily seen by (C.9):

$$\begin{aligned}\omega_0^{-2}\|h\|_{\mathcal{H}_0}^2 &= \sum_j \lambda_j j^2 e_{-j}^2(x_0) = \sum_{j\in\mathbb{Z}} \lambda_j j^2 e_{-j}^2(x_0) \\ &= \sum_{j\in\mathbb{Z}} \lambda_j j^2 e_{-j}^2(x_0) = \lambda_0 + 2\sum_{j\geq 1} j^2\lambda_j \leq \frac{2c_4c_5}{c_4-1}.\end{aligned}$$

Finally to derive N_ϵ, we reuse the orthonormal decomposition of $h(x)$ in (C.11). For a given set of j values A where $A \subseteq \mathbb{Z}$, we denote as Φ_A the "matrix" whose columns enumerate the φ_j over $j \in A$. Let us choose

$$A \stackrel{\text{def}}{=} \left\{ j : \lambda_j \max(1,j^2)(1+j^2)(1+\delta(j\geq 1)) \geq \min(1, w_0^{-2})\frac{\epsilon^2}{2.1} \right\}.$$

If $j \in A$, then $-j \in A$. Letting $\mathbb{N}_0 = \{0, 1, 2, \ldots\}$, we note $\sum_{j \in \mathbb{N}_0} \frac{1}{1+j^2} \leq 2.1$. So

$$\begin{aligned}
\left\|h - \Phi_A \Phi_A^\top h\right\|_{\mathcal{H}_0}^2 &= w_0^2 \sum_{j \in \mathbb{Z} \setminus A} \lambda_j j^2 e_{-j}^2(x_0) \\
&= w_0^2 \sum_{j \in \mathbb{N}_0 \setminus A} \lambda_j j^2 \left[(e_j^2(x) + e_{-j}^2(x))\delta(j \geq 1) + \delta(j = 0)\right] \\
&= w_0^2 \sum_{j \in \mathbb{N}_0 \setminus A} \lambda_j j^2 (1 + \delta(j \geq 1)) \\
&= w_0^2 \sum_{j \in \mathbb{N}_0 \setminus A} \left\{ \lambda_j j^2 (1 + j^2)(1 + \delta(j \geq 1)) \frac{1}{1 + j^2} \right\} \\
&\leq \frac{\epsilon^2}{2.1} \sum_{j \in \mathbb{N}_0} \frac{1}{1 + j^2} = \frac{\epsilon^2}{2.1} \sum_{j \in \mathbb{N}_0} \frac{1}{1 + j^2} \leq \epsilon^2.
\end{aligned}$$

Similarly, we can bound $\left\|k_0(x_0, \cdot) - \Phi_A \Phi_A^\top k_0(x_0, \cdot)\right\|_{\mathcal{H}_0}$ by

$$\begin{aligned}
&\left\|k_0(x_0, \cdot) - \Phi_A \Phi_A^\top k_0(x_0, \cdot)\right\|_{\mathcal{H}_0}^2 \\
&\quad = \sum_{j \in \mathbb{Z} \setminus A} \lambda_j e_j^2(x_0) \leq \sum_{j \in \mathbb{Z} \setminus A} \lambda_j \max(1, j^2) e_j^2(x_0) \\
&\quad = \sum_{j \in \mathbb{N}_0 \setminus A} \lambda_{\boldsymbol{\alpha}} \max(1, j^2) [\big(e_j^2(x) + e_{-j}^2(x)\big)\delta(j \geq 1) + \delta(j = 0)] \\
&\quad = \sum_{j \in \mathbb{N}_0 \setminus A} \left\{ \lambda_j \max(1, j^2)(1 + j^2)(1 + \delta(j \geq 1)) \frac{1}{1 + j^2} \right\} \\
&\quad \leq \frac{1}{2.1} \epsilon^2 \sum_{j \in \mathbb{N}_0} \frac{1}{1 + j^2} \\
&\quad \leq \epsilon^2.
\end{aligned}$$

To upper bound the cardinality of A, we consider the conditions for $j \notin A$. Thanks to the conditions in (C.9), we know that any j satisfying the following relationship cannot be in A:

$$c_6 \cdot c_4^{-|j|} < \min(1, w_0^{-2}) \frac{\epsilon^2}{2.1} \iff c_4^{-|j|} < \frac{1}{2.1 \cdot c_6} \min\left(1, \frac{4\pi^2}{v^2}\right) \epsilon^2.$$

So $A \subseteq \{j : |j| \leq n_\epsilon\}$, which yields the conclusion (C.10). Finally $Q_\epsilon \leq \sqrt{2}$, and to bound M_ϵ, we simply reuse (C.11). For any j with $|j| \leq n_\epsilon$,

$$\left|\langle h, e_j \rangle_{\mathcal{H}}\right| \leq \omega_0 \left|j e_{-j}(x_0)\right| \leq \frac{2\pi}{v} \sqrt{2} \lfloor n_\epsilon \rfloor = \frac{\sqrt{2}\pi}{v} (N_\epsilon - 1).$$

□

C.5. Case 2: Checking Assumptions 1 and 2 on Gaussian kernels

Gaussian kernels $k(x, y) = \exp(-\|x - y\|^2 / (2\sigma^2))$ are obviously product kernels with $k_0(x_1, y_1) = \kappa(x_1 - y_1) = \exp(-(x_1 - y_1)^2/(2\sigma^2))$. It is also translation invariant. The spectrum of Gaussian kernel k_0 on $\mathbb{R}$ is known; see, e.g., Chapter 4.3.1 of (Rasmussen & Williams, 2006) and Section 4 of (Zhu et al., 1998). Let μ be a Gaussian distribution $\mathcal{N}(0, \sigma^2)$. Setting $\epsilon^2 = \alpha^2 = (2\sigma^2)^{-1}$ in Eq 12 and 13 of (E Fasshauer, 2011), the eigenvalue and eigenfunctions are (for $j \geq 0$):

$$\begin{aligned}
\lambda_j &= c_0^{-j-1/2}, \quad \text{where} \quad c_0 = \frac{1}{2}(3 + \sqrt{5}) \\
e_j(x) &= \frac{5^{1/8}}{2^{j/2}} \exp\left(-\frac{\sqrt{5} - 1}{4} \frac{x^2}{\sigma^2}\right) \frac{1}{\sqrt{j!}} H_j\left(\sqrt[4]{1.25}\, \frac{x}{\sigma}\right),
\end{aligned}$$

where H_j is the Hermite polynomial of order j.

Although the eigenvalues decay exponentially fast, the eigenfunctions are not uniformly bounded in the L_∞ sense. Although the latter can be patched if we restrict x to a bounded set, the above closed-form of eigen-pairs will no longer hold, and the analysis will become rather challenging.

To resolve this issue, we resort to the period-ization technique proposed by (Williamson et al., 2001). Consider $\kappa(x) = \exp(-x^2/(2\sigma^2))$ when $x \in [-v/2, v/2]$, and then extend κ to $\mathbb{R}$ as a periodic function with period v. Again let μ be the uniform distribution on $[-v/2, v/2]$. As can be seen from the discriminant function $f = \frac{1}{l}\sum_{i=1}^{l} \gamma_i k(x^i, \cdot)$, as along as our training and test data both lie in $[-v/4, v/4]$, the modification of κ outside $[-v/2, v/2]$ does not effectively make any difference. Although the term $\partial^{0,1} k_0(x_1^a, w_1^1)$ in (10) may possibly evaluate κ outside $[-v/2, v/2]$, it is only used for testing the gradient norm bound of κ.

With this periodized Gaussian kernel, it is easy to see that $Q_\epsilon = \sqrt{2}$. If we standardize by $\sigma = 1$ and set $v = 5\pi$ as an example, it is not hard to see that (C.9) holds with $c_4 = 1.25$ and $c_6 = 50$. The expressions of N_ϵ and M_ϵ then follow from Theorem 6 directly.

C.6. Case 3: Checking Assumptions 1 and 2 on non-product kernels

The above analysis has been restricted to product kernels. But in practice, there are many useful kernels that are not decomposable. A prominent example is the inverse kernel: $k(x, y) = (2 - x^\top y)^{-1}$. In general, it is extremely challenging to analyze eigenfunctions, which are commonly *not* bounded (Zhou, 2002; Lafferty & Lebanon, 2005), i.e., $\sup_{i\to\infty} \sup_x |e_i(x)| = \infty$. The opposite was (incorrectly) claimed in Theorem 4 of Williamson et al. (2001) by citing an incorrect result in König (1986, p. 145), which was later corrected by Zhou (2002) and Steve Smale. Indeed, uniform boundedness is not known even for Gaussian kernels with uniform distribution on $[0, 1]^d$ (Lin et al., 2017), and Minh et al. (2006, Theorem 5) showed the unboundedness for Gaussian kernels with uniform distribution on the unit sphere when $d \geq 3$.

Here we only present the limited results that we have obtained on the eigenvalues of the integral operator of inverse kernels with a uniform distribution on the unit ball. The analysis of eigenfunctions is left for future work. Specifically, in order to drive the eigenvalue λ_i below ϵ, i must be at least $d^{\lceil \log_2 \frac{1}{\epsilon} \rceil + 1}$. This is a quasi-quadratic bound if we view d and $1/\epsilon$ as two large variables.

It is quite straightforward to give an explicit characterization of the functions in $\mathcal{H}$. The Taylor expansion of z^{-1} at $z = 2$ is $\frac{1}{2}\sum_{i=0}^{\infty}(-\frac{1}{2})^i x^i$. Using the standard multi-index notation with $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d) \in (\mathbb{N} \cup \{0\})^d$, $|\boldsymbol{\alpha}| = \sum_{i=1}^d \alpha_i$, and $\mathbf{x}^{\boldsymbol{\alpha}} = x_1^{\alpha_1} \ldots x_d^{\alpha_d}$, we derive

$$\begin{aligned} k(\mathbf{x}, \mathbf{y}) &= \frac{1}{2 - \mathbf{x}^\top \mathbf{y}} \\ &= \frac{1}{2}\sum_{k=0}^{\infty}\left(-\frac{1}{2}\right)^k (-\mathbf{x}^\top \mathbf{y})^k \\ &= \sum_{k=0}^{\infty} 2^{-k-1} \sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|=k} C_{\boldsymbol{\alpha}}^k \mathbf{x}^{\boldsymbol{\alpha}} \mathbf{y}^{\boldsymbol{\alpha}} \\ &= \sum_{\boldsymbol{\alpha}} 2^{-|\boldsymbol{\alpha}|-1} C_{\boldsymbol{\alpha}}^{|\boldsymbol{\alpha}|} \mathbf{x}^{\boldsymbol{\alpha}} \mathbf{y}^{\boldsymbol{\alpha}}, \end{aligned}$$

where $C_{\boldsymbol{\alpha}}^k = \frac{k!}{\prod_{i=1}^d \alpha_i!}$. So we can read off the feature mapping for $\mathbf{x}$ as

$$\phi(\mathbf{x}) = \{w_{\boldsymbol{\alpha}} \mathbf{x}^{\boldsymbol{\alpha}} : \boldsymbol{\alpha}\}, \quad \text{where} \quad w_{\boldsymbol{\alpha}} = 2^{-\frac{1}{2}(|\boldsymbol{\alpha}|+1)} C_{\boldsymbol{\alpha}}^{|\boldsymbol{\alpha}|},$$

and the functions in $\mathcal{H}$ are

$$\mathcal{H} = \left\{ f = \sum_{\boldsymbol{\alpha}} \theta_{\boldsymbol{\alpha}} w_{\boldsymbol{\alpha}} \mathbf{x}^{\boldsymbol{\alpha}} : \|\boldsymbol{\theta}\|_{\ell_2} < \infty \right\}. \tag{C.12}$$

Note this is just an intuitive "derivation" while a rigorous proof for (C.12) can be constructed in analogy to that of Theorem 1 in Minh (2010).

C.7. Background of eigenvalues of a kernel

We now use (C.12) to find the eigenvalues of inverse kernel.

Now specializing to our inverse kernel case, let us endow a uniform distribution over the unit ball B: $p(x) = V_d^{-1}$ where $V_d = \pi^{d/2}\Gamma(\frac{d}{2}+1)^{-1}$ is the volume of B, with Γ being the Gamma function. Then λ is an eigenvalue of the kernel if there exists $f = \sum_{\boldsymbol{\alpha}} \theta_{\boldsymbol{\alpha}} w_{\boldsymbol{\alpha}} \mathbf{x}^{\boldsymbol{\alpha}}$ such that $\int_{\mathbf{y}\in B} k(\mathbf{x},\mathbf{y})p(\mathbf{y})f(\mathbf{y})\,\mathrm{d}\mathbf{y} = \lambda f(\mathbf{x})$. This translates to

$$V_d^{-1}\int_{\mathbf{y}\in B}\sum_{\boldsymbol{\alpha}} w_{\boldsymbol{\alpha}}^2 \mathbf{x}^{\boldsymbol{\alpha}}\mathbf{y}^{\boldsymbol{\alpha}} \sum_{\boldsymbol{\beta}} \theta_{\boldsymbol{\beta}} w_{\boldsymbol{\beta}} \mathbf{y}^{\boldsymbol{\beta}}\,\mathrm{d}\mathbf{y} = \lambda \sum_{\boldsymbol{\alpha}} \theta_{\boldsymbol{\alpha}} w_{\boldsymbol{\alpha}} \mathbf{x}^{\boldsymbol{\alpha}}, \qquad \forall\, \mathbf{x}\in B.$$

Since B is an open set, that means

$$w_{\boldsymbol{\alpha}} \sum_{\boldsymbol{\beta}} w_{\boldsymbol{\beta}} q_{\boldsymbol{\alpha}+\boldsymbol{\beta}} \theta_{\boldsymbol{\beta}} = \lambda \theta_{\boldsymbol{\alpha}}, \qquad \forall\, \boldsymbol{\alpha},$$

where

$$q_{\boldsymbol{\alpha}} = V_d^{-1}\int_{\mathbf{y}\in B} \mathbf{y}^{\boldsymbol{\alpha}}\,\mathrm{d}\mathbf{y} = \begin{cases} \frac{2\prod_{i=1}^d \Gamma\left(\frac{1}{2}\alpha_i+\frac{1}{2}\right)}{V_d\cdot(|\boldsymbol{\alpha}|+d)\cdot\Gamma\left(\frac{1}{2}|\boldsymbol{\alpha}|+\frac{d}{2}\right)} & \text{if all } \alpha_i \text{ are even} \\ 0 & \text{otherwise} \end{cases}.$$

In other words, λ is the eigenvalue of the infinite dimensional matrix $Q = [w_{\boldsymbol{\alpha}} w_{\boldsymbol{\beta}} q_{\boldsymbol{\alpha}+\boldsymbol{\beta}}]_{\boldsymbol{\alpha},\boldsymbol{\beta}}$,

C.8. Bounding the eigenvalues

To bound the eigenvalues of Q, we resort to the majorization results in matrix analysis. Since k is a PSD kernel, all its eigenvalues are nonnegative, and suppose they are sorted decreasingly as $\lambda_1 \ge \lambda_2 \ge \ldots$. Let the row corresponding to $\boldsymbol{\alpha}$ have ℓ_2 norm $r_{\boldsymbol{\alpha}}$, and let them be sorted as $r_{[1]} \ge r_{[2]} \ge \ldots$. Then by (Schneider, 1953; Shi & Wang, 1965), we have

$$\prod_{i=1}^n \lambda_i \;\le\; \prod_{i=1}^n r_{[i]}, \quad \forall\, n \ge 1.$$

So our strategy is to bound $r_{\boldsymbol{\alpha}}$ first. To start with, we decompose $q_{\boldsymbol{\alpha}+\boldsymbol{\beta}}$ into $q_{\boldsymbol{\alpha}}$ and $q_{\boldsymbol{\beta}}$ via Cauchy-Schwartz:

$$q_{\boldsymbol{\alpha}+\boldsymbol{\beta}}^2 = V_d^{-2}\left(\int_{\mathbf{y}\in B}\mathbf{y}^{\boldsymbol{\alpha}+\boldsymbol{\beta}}\,\mathrm{d}\mathbf{y}\right)^2 \le V_d^{-2}\int_{\mathbf{y}\in B}\mathbf{y}^{2\boldsymbol{\alpha}}\,\mathrm{d}\mathbf{y}\cdot\int_{\mathbf{y}\in B}\mathbf{y}^{2\boldsymbol{\beta}}\,\mathrm{d}\mathbf{y} = q_{2\boldsymbol{\alpha}}q_{2\boldsymbol{\beta}}.$$

To simplify notation, we consider without loss of generality that d is an even number, and denote the integer $b \stackrel{\text{def}}{=} d/2$. Now $V_d = \pi^b/b!$. Noting that there are $\left(\begin{array}{c} k+d-1 \\ k \end{array}\right)$ values of $\boldsymbol{\beta}$ such that $|\boldsymbol{\beta}| = k$, we can proceed by (fix below by changing $\left(\begin{array}{c} k+d \\ k \end{array}\right)$ into $\left(\begin{array}{c} k+d-1 \\ k \end{array}\right)$, or no need because the former upper bounds the latter)

$$\begin{aligned}
r_{\boldsymbol{\alpha}}^2 &= w_{\boldsymbol{\alpha}}^2 \sum_{\boldsymbol{\beta}} w_{\boldsymbol{\beta}}^2 q_{\boldsymbol{\alpha}+\boldsymbol{\beta}}^2 \le w_{\boldsymbol{\alpha}}^2 q_{2\boldsymbol{\alpha}} \sum_{\boldsymbol{\beta}} w_{\boldsymbol{\beta}}^2 q_{2\boldsymbol{\beta}} = w_{\boldsymbol{\alpha}}^2 q_{2\boldsymbol{\alpha}} \sum_{k=0}^\infty 2^{-k-1} \sum_{\boldsymbol{\beta}:|\boldsymbol{\beta}|=k} C_{\boldsymbol{\beta}}^k q_{2\boldsymbol{\beta}} \\
&\le w_{\boldsymbol{\alpha}}^2 q_{2\boldsymbol{\alpha}} \sum_{k=0}^\infty 2^{-k-1} \left(\begin{array}{c} k+d \\ d \end{array}\right) \max_{|\boldsymbol{\beta}|=k} C_{\boldsymbol{\beta}}^k q_{2\boldsymbol{\beta}} \\
&= w_{\boldsymbol{\alpha}}^2 q_{2\boldsymbol{\alpha}} \sum_{k=0}^\infty 2^{-k-1} \left(\begin{array}{c} k+d \\ d \end{array}\right) \max_{|\boldsymbol{\beta}|=k} \frac{k!}{\prod_{i=1}^d \beta_i!} \cdot \frac{2\prod_{i=1}^d \Gamma(\beta_i+\frac{1}{2})}{V_d\cdot(2k+d)\cdot\Gamma(k+\frac{d}{2})} \\
&= w_{\boldsymbol{\alpha}}^2 q_{2\boldsymbol{\alpha}} V_d^{-1} \sum_{k=0}^\infty 2^{-k} \left(\begin{array}{c} k+d \\ d \end{array}\right) \frac{k!}{(2k+d)\Gamma(k+\frac{d}{2})} \cdot \max_{|\boldsymbol{\beta}|=k} \prod_{i=1}^d \frac{\Gamma(\beta_i+\frac{1}{2})}{\beta_i!} \\
&< w_{\boldsymbol{\alpha}}^2 q_{2\boldsymbol{\alpha}} \cdot \frac{b!}{\pi^b d!} \cdot \sum_{k=0}^\infty 2^{-k-1} \frac{(k+d)!}{(k+b)!},
\end{aligned}$$

since $\Gamma(\beta_i + \frac{1}{2}) < \Gamma(\beta_i + 1) = \beta_i!$. The summation over k can be bounded by

$$\sum_{k=0}^{\infty} 2^{-k-1} \frac{(k+d)!}{(k+b)!} = \frac{1}{2} b! \left(2^d + \left(\begin{array}{c} d \\ b \end{array} \right) \right) \leq \frac{1}{2} \left(b! 2^d + 2^b \right) \leq b! 2^d,$$

where the first equality used the identity $\sum_{k=1}^{\infty} 2^{-k} \left(\begin{array}{c} d+k \\ b \end{array} \right) = 2^d$. Letting $l \stackrel{\text{def}}{=} |\boldsymbol{\alpha}|$, we can continue by

$$\begin{aligned} r_{\boldsymbol{\alpha}}^2 < w_{\boldsymbol{\alpha}}^2 q_{2\boldsymbol{\alpha}} \cdot \frac{b!}{\pi^b d!} b! 2^d &= 2^{-l-1} \frac{l!}{\prod_{i=1}^d \alpha_i!} \frac{2 \prod_{i=1}^d \Gamma\left(\alpha_i + \frac{1}{2}\right)}{V_d \cdot (2l+d) \cdot \Gamma(l+b)} \frac{(b!)^2 2^d}{\pi^b d!} \\ &\leq 2^{-l+d} \pi^{-2b} \frac{l! (b!)^3}{d! (l+b-1)! (2l+d)} \qquad (\text{since } \Gamma(\alpha_i + \tfrac{1}{2}) < \Gamma(\alpha_i + 1) = \alpha_i!) \\ &\leq 2^{-l+b-1} \pi^{-2b} \left(\begin{array}{c} l+b \\ l \end{array} \right)^{-1} \qquad (\text{since } \frac{(b!)^2}{d!} \leq 2^{-b}). \end{aligned}$$

This bound depends on $\boldsymbol{\alpha}$, not directly on $\boldsymbol{\alpha}$. Letting $n_l = \left(\begin{array}{c} l+d-1 \\ l \end{array} \right)$ and $N_L = \sum_{l=0}^{L} n_l = \left(\begin{array}{c} d+L \\ L \end{array} \right)$, it follows that

$$\begin{aligned} \sum_{l=0}^{L} l n_l = \sum_{l=1}^{L} \frac{l(l+d)!}{d! \cdot l!} &= (d+1) \sum_{l=1}^{L} \frac{(l+d)!}{(d+1)!(l-1)!} \\ &= (d+1) \sum_{l=1}^{L} \left(\begin{array}{c} l+d \\ d+1 \end{array} \right) = (d+1) \left(\begin{array}{c} L+d+1 \\ d+2 \end{array} \right). \end{aligned}$$

Now we can bound λ_{N_L} by

$$\begin{aligned} \lambda_{N_L}^{N_L} \leq \prod_{i=1}^{N_L} \lambda_i &\leq \prod_{l=0}^{L} \left(2^{-l+b-1} \pi^{-2b} \left(\begin{array}{c} l+b \\ l \end{array} \right)^{-1} \right)^{n_l} \\ \implies \quad \log \lambda_{N_L} &\leq N_L^{-1} \sum_{l=0}^{L} n_l \left(-(l-b+1) \log 2 - 2b \log \pi - \log \left(\begin{array}{c} l+b \\ l \end{array} \right) \right) \\ &\leq -N_L^{-1} \cdot \log 2 \cdot \sum_{l=0}^{L} l n_l \end{aligned}$$

since $\log 2 < 2 \log \pi$ as the coefficients of b

$$\begin{aligned} &= - \left(\begin{array}{c} d+L+1 \\ d+1 \end{array} \right)^{-1} \cdot \log 2 \cdot (d+1) \left(\begin{array}{c} d+L+1 \\ d+2 \end{array} \right) \\ &= -\frac{d+1}{d+2} L \log 2 \\ &\approx -L \log 2 \\ \implies \quad \lambda_{N_L} &\leq 2^{-L}. \end{aligned}$$

This means that the eigenvalue $\lambda_i \leq \epsilon$ provided that $i \geq N_L$ where $L = \left\lceil \log_2 \frac{1}{\epsilon} \right\rceil$. Since $N_L \leq d^{L+1}$, that means it suffices to choose i such that

$$i \geq d^{\left\lceil \log_2 \frac{1}{\epsilon} \right\rceil + 1}.$$

This is a quasi-polynomial bound. It seems tight because even in Gaussian RBF kernel, the eigenvalues follow the order of $\lambda_{\boldsymbol{\alpha}} = O(c^{-|\boldsymbol{\alpha}|})$ for some $c > 1$ (Fasshauer & McCourt, 2012, p.A742).

D. Algorithm for training a Lipschitz binary SVMs

The pseudo-code of training binary SVMs by enforcing Lipschitz constant is given in Algorithm 1.

Finding the exact $\arg\max_{x\in X} \left\|\nabla f^{(i)}(x)\right\|$ is intractable, so we used a local maximum found by L-BFGS with 10 random initialisations as the Lipschitz constant of the current solution $f^{(i)}$ ($L^{(i)}$ in step 6). The solution found by L-BFGS is also used as the new greedy point added in step 5b.

Furthermore, the kernel expansion $f(x) = \frac{1}{l}\sum_{a=1}^{l} \gamma_a k(x^a, \cdot)$ can lead to high cost in optimisation (our experiment used $l = 54000$), and therefore we used *another* Nyström approximation for the kernels. We randomly sampled 1000 landmark points, and based on them we computed the Nyström approximation for each $k(x^a, \cdot)$, denoted as $\tilde{\phi}(x^a) \in \mathbb{R}^{1000}$. Then $f(x)$ can be written as $\frac{1}{l}\sum_{a=1}^{l} \gamma_a \tilde{\phi}(x^a)^\top \tilde{\phi}(x)$. Defining $w = \frac{1}{l}\sum_{a=1}^{l} \gamma_a \tilde{\phi}(x^a)$, we can equivalently optimise over w, and the RKHS norm bound on f can be equivalently imposed as the ℓ_2-norm bound on w.

To summarise, Nyström approximation is used in two different places: one for approximating the kernel function, and one for computing $\|g_j\|_{\mathcal{H}}$ either holistically or coordinate wise. For the former, we randomly sampled 1000 landmark points; for the latter, we used greedy selection as option b in step 5 of Algorithm 1.

D.1. Detailed algorithm for multiclass classification

It is easy to extend Algorithm 1 to multiclass. For example, with MNIST dataset, we solve the following optimisation problem to defend ℓ_2 attacks:

$$\begin{aligned}
&\underset{\boldsymbol{\gamma}^1,\ldots,\boldsymbol{\gamma}^{10}}{\text{minimise}} \quad \sum_{i=1}^{n} \ell(F(x), \mathbf{y}), \quad \text{where} \quad F \stackrel{\text{def}}{=} \left[\sum_{i=1}^{n} \boldsymbol{\gamma}_i^1 k(x_i, \cdot); \ldots; \sum_{i=1}^{n} \boldsymbol{\gamma}_i^{10} k(x_i, \cdot)\right] \\
&\text{subject to} \quad \sup_{\|\phi\|_{\mathcal{H}}\le 1} \lambda_{\max}\left(\sum_{c=1}^{10} G_c^\top \phi\phi^\top G_c\right) \approx \sup_{\|v\|_2\le 1} \lambda_{\max}\left(\sum_{c=1}^{10} \tilde{G}_c^\top vv^\top \tilde{G}_c\right) \le L^2,
\end{aligned}$$

where $\ell(F(x), \mathbf{y})$ is the Crammer & Singer loss, and the constraint is derived from (11) by using its Nyström approximation $\tilde{G}_c = [\tilde{g}_1^c, \ldots, \tilde{g}_d^c]$, which depends on $\{\boldsymbol{\gamma}^1, \ldots, \boldsymbol{\gamma}^{10}\}$ linearly. Note that the constraint itself is a supremum problem:

$$\sup_{\|v\|_2\le 1} \lambda_{\max}\left(\sum_{c=1}^{10} \tilde{G}_c^\top vv^\top \tilde{G}_c\right) = \sup_{\|v\|_2\le 1, \|u\|_2\le 1} u^\top \left(\sum_{c=1}^{10} \tilde{G}_c^\top vv^\top \tilde{G}_c\right) u.$$

Since there is only one constraint, interior point algorithm is efficient. It requires the gradient of the constraint, which can be computed by Danskin's theorem. In particular, we alternates between updating v and u, until they converge to the optimal v_* and u_*. Finally, the derivative of the constraint with respect to $\{\boldsymbol{\gamma}^c\}$ can be calculated from $\sum_{c=1}^{10}(u_*^\top \tilde{G}_c^\top v_*)^2$, as a function of $\{\boldsymbol{\gamma}^c\}$.

To defend ∞-norm attacks, we need to enforce the ∞-norm of the Jacobian matrix:

$$\begin{aligned}
\sup_{x\in X} \left\| \left[g^1(x), \ldots, g^{10}(x)\right]^\top \right\|_\infty &= \sup_{x\in X} \max_{1\le c\le 10} \|g^c(x)\|_1 \\
&= \max_{1\le c\le 10} \sup_{x\in X} \|g^c(x)\|_1 \\
&\le \max_{1\le c\le 10} \sup_{\|\phi\|_2\le 1, \|u\|_\infty\le 1} u^\top \tilde{G}_c^\top \phi,
\end{aligned}$$

where the last inequality is due to

$$\sup_{x\in X} \|g(x)\|_1 = \sup_{x\in X} \sup_{\|u\|_\infty\le 1} u^\top g(x) \le \sup_{\|v\|_2\le 1, \|u\|_\infty\le 1} u^\top \tilde{G}^\top v.$$

Therefore, the overall optimisation problem for defense against ∞-norm attacks is

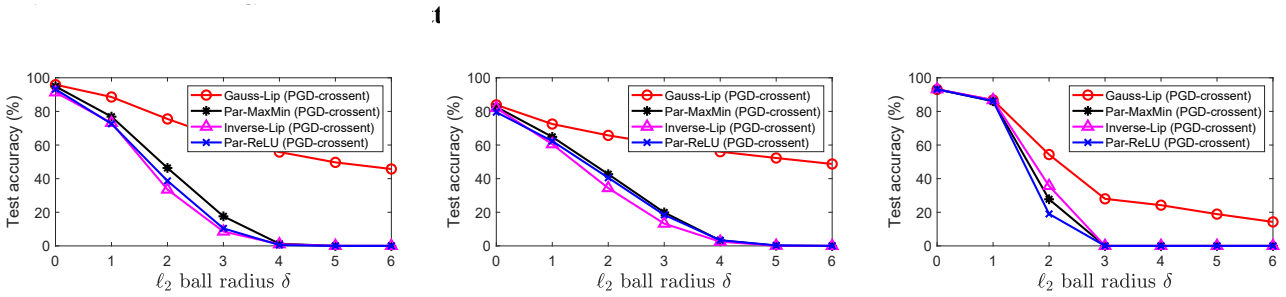
$$\begin{aligned}
&\underset{\boldsymbol{\gamma}^1,\ldots,\boldsymbol{\gamma}^{10}}{\text{minimise}} \quad \sum_{i=1}^{n} \ell(F(x), \mathbf{y}), \\
&\text{subject to} \quad \forall_{c\in[10]} : \sup_{\|v\|_2\le 1, \|u\|_\infty\le 1} u^\top \tilde{G}_c^\top v \le L
\end{aligned} \tag{D.1}$$

For each c, we alternatively update v and u in (D.1), converging to the optimal v_* and u_*. Finally, the derivative of $\sup_{\|v\|_2 \le 1, \|u\|_\infty \le 1} u^\top \tilde{G}_c^\top v$ with respect to $\boldsymbol{\gamma}^c$ can be calculated from $u_*^\top \tilde{G}_c^\top v_*$, as a function of $\boldsymbol{\gamma}^c$.

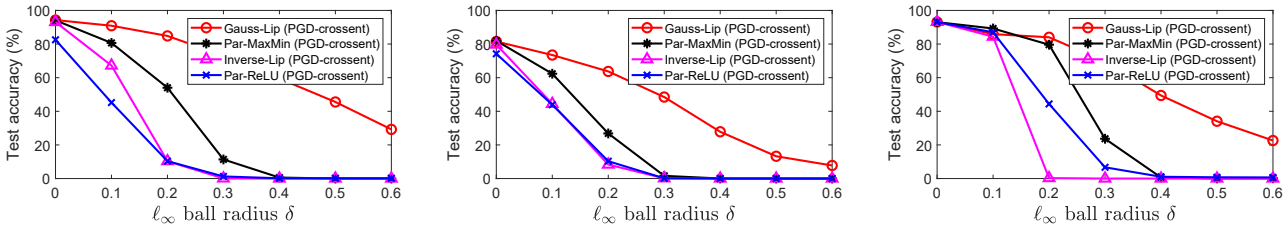
E. More experiments

All code and data are available anonymously, with no tracing, at

https://github.com/learndeep2019/DRobust.



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



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

Figure 10: Test accuracy under PGD attacks on cross-entropy approximation with ∞-norm bound

E.2. Visualization of attacks

In order to verify that the robustness of Gauss-Lip is not due to obfuscated gradient, we randomly sampled 10 images from MNIST, and ran **targeted** PGD for 100 steps with cross-entropy objective and the ℓ_2 norm upper bounded by 8. For example, in Figure 11, the row corresponding to class 4 tries to promote the likelihood of the target class 4. Naturally the diagonal is not meaningful, hence left empty. At the end of attack, PDG turned 89 out of 90 images into the target class by following the gradient of the defense model.

Please note that despite the commonality in using the cross-entropy objective, the setting of targeted attack in Figure 11 is not comparable to that in Figure 9, where to enable a batch test mode, an *untargeted* attacker was employed by increasing the cross-entropy loss of the correct class, i.e., decreasing the likelihood of the correct class. This is a common practice.

We further ran PGD for 100 steps on C&W approximation (an untargeted attack used in Figure 5), and the resulting images after every 10 iterations are shown in Figure 12. Here all 10 images were eventually turned into a different but untargeted class, and the final images are very realistic.

(a)

0	1	2	3	4	5	6	7	8	9
	0	0	0	0	0	0	0	0	0
0		1	1	1	1	1	1	1	1
2	2		2	2	2	2	2	2	2
3	3	3		3	3	3	3	3	3
4	4	4	4		4	4	4	4	4
5	5	5	5	5		5	5	5	5
6	6	6	6	6	6		6	6	6
7	7	7	7	7	7	7		7	7
8	8	8	8	8	8	8	8		8
9	9	9	9	9	9	9	9	9	

(b)

Figure 11: (a) perturbed images at the end of 100-step PGD attack using the (**targeted**) cross-entropy approximation. The top row shows 10 random images, one sampled from each class. The 10 rows below correspond to the target class. (b) classification on the perturbed image given by the trained Gauss-Lip. The left images are quite consistent with human's perception.

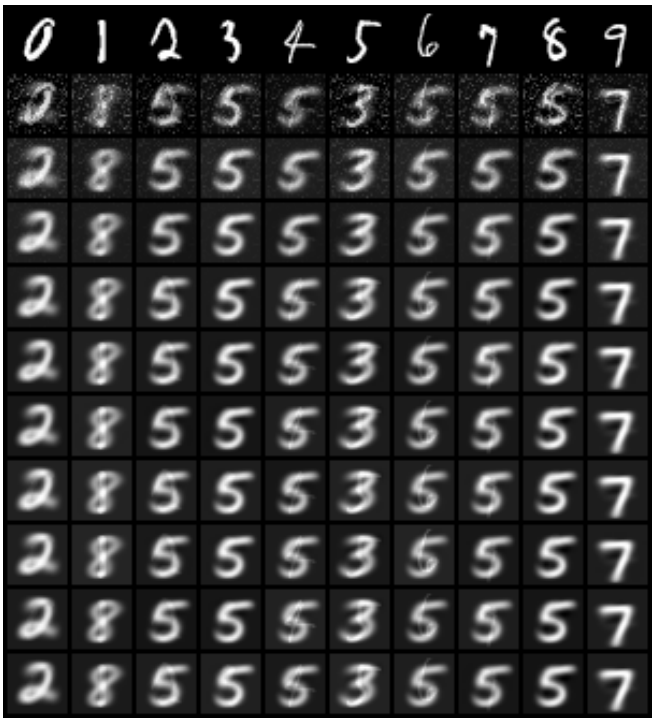


Figure 12: Perturbed images at the end of 100-step PGD attack using the (**untargeted**) C&W approximation. The top row shows 10 random images, one sampled from each class. The 10 rows below show the images after 10, 20, ..., 100 steps of PGD.