CERTIFYING DISTRIBUTIONAL ROBUSTNESS USING LIPSCHITZ REGULARISATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Distributional robust risk (DRR) minimisation has arisen as a flexible and effective framework for machine learning. Approximate solutions based on dualisation have become particularly favorable in addressing the semi-infinite optimisation, and they also provide a certificate of the robustness for the worst-case population loss. However existing methods are restricted to either linear models or very small perturbations, and cannot find the globally optimal solution for restricted nonlinear models such as kernel machines. In this paper we resolve these limitations by upper bounding DRRs with an empirical risk regularised by the Lipschitz constant of the model, including deep neural networks and kernel methods. As an application, we showed that it also provides a certificate for adversarial training, and global solutions can be achieved on product kernel machines in polynomial time.

1 INTRODUCTION

Regularised risk minimisation has been the workhorse of learning nonlinear hypotheses such as deep neural networks and kernel machines. Recently, *distributional robust risk* (DRR) minimization has emerged as a promising instance with marked efficacy and flexibility. Instead of perturbing the observed data points, DRRs consider perturbations to the *empirical distribution*, constituting an *ambiguity set* $\mathcal{P}$ that lives in the space of data distributions. Let Ω be an outcome space with (true) distribution μ, e.g., the joint space of input and output. Given a loss function ℓ, a model f suffers a loss value $\ell_f(\omega)$ over an outcome ω, and the risk of f under μ is $\mathrm{risk}_\ell(f, \mu) := \mathrm{E}_\mu[\ell_f]$. In DRR minimisation, a model f is sought that minimises the expectation of loss ℓ over an ambiguity set $\mathcal{P}$, i.e., that minimises $\sup_{\nu \in \mathcal{P}} \mathrm{risk}_\ell(f, \nu)$ (Delage & Ye, 2010; Goh & Sim, 2010; Wiesemann et al., 2014). The ambiguity sets can be constructed by moment matching (Bhattacharyya et al., 2005; Farnia & Tse, 2016), divergence balls (Ben-Tal et al., 2013; Duchi et al., 2016; Hu & Hong, 2016), or Wasserstein distance balls (Kantorovitch, 1958). In this work we focus on the last due to its favorable properties in statistics and computation, along with extensive applications in DRR (Esfahani & Kuhn, 2018; Gao & Kleywegt, 2016; Zhao & Guan, 2018).

Despite the generality of DRR, its computational efficiency remains a challenge, since the supremum is over a (typically) uncountably infinite dimension space. Tractable equivalent convex programs can be derived only for a limited range of loss functions along with *linear* hypothesis spaces (Blanchet et al., 2016; El Ghaoui & Lebret, 1997; Shafieezadeh-Abadeh et al., 2015; 2017; Xu et al., 2009a;b). Although Shafieezadeh-Abadeh et al. (2017) developed lifted variants for reproducing kernel Hilbert spaces (RKHS) to accommodate nonlinear hypotheses, the perturbation was applied to $\Phi(\omega)$, where Φ is the implicit feature map. This still falls short of robustness with respect to distributions over Ω.

A more promising technique for optimizing DRRs over *nonlinear* hypothesis spaces—including deep neural networks and kernel machines—is by dualising it to a form that is amenable to (approximate) optimisation (Mohajerin Esfahani & Kuhn, 2018). The fundamental strong duality result was established independently by Blanchet & Murthy (2019) and Gao & Kleywegt (2016), and has been applied to various tasks such as specification of regularisation parameter (Blanchet et al., 2016), design of transport cost (Blanchet et al., 2017), and selection of ambiguity region size for optimal confidence interval (Duchi et al., 2016). In particular, Sinha et al. (2018) used it to construct an efficiently computable *certificate* on the level of robustness for the worst-case population loss. However, these methods are still subject to marked restrictions when applied to smooth nonlinear models. First, Sinha et al. (2018) restricts the perturbation to be small, which despite the common interest of imperceptible perturbations, leaves unaddressed the equally interesting regimes of medium to large

perturbations; see discussions in Openreview (2018). Moreover, although the global solvability of the inner Lagrangian penalty problem (robust surrogate loss) can be ensured by small enough perturbation, there is no practical procedure to compute the threshold. Finally, the overall optimisation of the nonlinear model is still subject to nonconvexity, precluding tractable global solutions for restricted but still general classes of nonlinear models such as kernel methods.

The first goal of this work, therefore, is to develop a novel certificate on distributional robustness that dispenses with these restrictions (§3). Specifically, we will leverage the McShane-Whitney extension theorem (McShane, 1934; Whitney, 1934) to upper bound DRRs by the empirical risk regularized with the Lipschitz constant of the model f, while additionally accounting for the underlying transport cost and the loss ℓ. The result vastly generalises the vector norm regularisation in *linear* binary classification (Shafieezadeh-Abadeh et al., 2017, Thm. 3.11) to nonlinear models and *extended real-valued* cost functions that encode constraints, along with an arbitrary metric space of labels that is general enough for multiclass problems. Appealing to any magnitude of perturbation, it also enjoys improved computational efficiency compared with the robust surrogate loss in Sinha et al. (2018).

A particularly effective domain to apply this new certificate is adversarial learning (Szegedy et al., 2014), where models are trained to be resilient to malicious distortions on the data. Although Lipschitz regularisation has been a popular recipe for robustness (e.g., Anil et al., 2019; Cisse et al., 2017; Farnia et al., 2019; Gouk et al., 2018; Huster et al., 2018; Scaman & Virmaux, 2018) and generalisation accuracy (Miyato et al., 2018; Yoshida & Miyato, 2017), it remains a heuristic and therefore our second major contribution is to reveal in §3.1 that adversarial risks (Goodfellow et al., 2015; Madry et al., 2018; Shaham et al., 2018) can be bounded by a DRR. Such a rigorous justification has hitherto been restricted to logistic loss (Suggala et al., 2019, Thm. 9), and a similar tightness result has been established only for linear models (Shafieezadeh-Abadeh et al., 2017, Thm. 3.20). As a result, our new certificate amounts to a new bound on the worst-case risk under attacks, complementing the existing certificates (Raghunathan et al., 2018; Tsuzuku et al., 2018; Weng et al., 2018; Wong & Kolter, 2018; Wong et al., 2018) with a more computationally efficient approach. It further achieved state-of-the-art accuracy under a range of attacks on standard benchmark datasets (§5).

In practice, however, the evaluation of Lipschitz constant L is NP-hard for neural networks (Scaman & Virmaux, 2018), compelling approximations of it, or explicit engineering of layers to respect Lipschitz while analyzing the expressiveness in specific cases (e.g., ℓ_∞ norm in Anil et al. (2019)). We, instead, pursue a new path and explore the following question: does there exist a hypothesis space which: **a)** is expressive enough in modeling; **b)** allows the *exact* value of L to be computed *efficiently*; **c)** enforcing the Lipschitz constant leads to a *convex* constraint that renders efficient optimisation.

Interestingly, kernel machines satisfy all these requirements for some kernels. For example, Gaussian kernels are universal, whose RKHS can approximate any continuous function on a compact set in a *uniform* sense (Micchelli et al., 2006). The RKHS of multi-layer inverse kernels compactly encompasses ℓ_1-regularized neural networks (Shalev-Shwartz et al., 2011), degrading the generalisation performance by only a polynomial constant (Zhang et al., 2016; 2017). Similar results have been conjectured for Gaussian kernels (Shalev-Shwartz et al., 2011). Our third contribution proves that **b)** can be achieved for *product kernels* such as Gaussian kernels with high probability by using the Nyström approximation (Drineas & Mahoney, 2005; Williams & Seeger, 2000), and ε approximation error of L requires only $O(1/\varepsilon^2)$ samples (§4). Empirically this approximation is also effective for non-product kernels like inverse kernels. Such a sampling based approach also leads to a *single* convex constraint, making it scalable to 60k examples with even an interior-point solver (§5). The convenience in evaluating L renders our certificate of DRR even more favorable than those based on robust surrogate losses (Blanchet & Murthy, 2019; Gao & Kleywegt, 2016; Sinha et al., 2018).

2 Preliminaries

The vast majority of our technical results and proofs are deferred to Appendix A, for which theorem-like statements are numbered to be consistent.

The *extended real line* is $\bar{\mathbb{R}} := [-\infty, +\infty]$. For topological spaces X, Y, the Borel subsets are $\mathscr{B}(X)$, and the Borel probability measures are $\mathfrak{P}(X)$. The *universal sigma algebra* is $\mathscr{U}(X) := \bigcap_{\mu \in \mathfrak{P}(X)} \mathscr{B}_\mu(X)$ where $\mathscr{B}_\mu(X)$ is the completion of the Borel sets with respect to $\mu \in \mathfrak{P}(X)$. Let $\mathrm{M}(X, Y)$ denote the Borel measurable mappings $X \to Y$. We say a function is *universally measurable* if it is measurable as a mapping $(X, \mathscr{U}(X)) \to (Y, \mathscr{B}(Y))$. For $f \in \mathrm{M}(X, Y)$ and a

measure $\mu \in \mathfrak{P}(X)$, the *push forward of* μ *by* f is $f_{\#}\mu \in \mathfrak{P}(Y)$ with $f_{\#}\mu := \mu \circ f^{-1}$. Universally measurable functions can be integrated with a Borel measure because for $\mu \in \mathfrak{P}(X)$, $f : X \to \bar{\mathbb{R}}$ is universally measurable if and only if there exists a unique $f_\mu \in \mathrm{M}(X, \bar{\mathbb{R}})$ with $f(x) = f_\mu(x)$ for μ-almost every x (Bertsekas & Shreve, 1978, Lem. 7.27, p. 173), and so we let $\int f \,\mathrm{d}\mu := \int f_\mu \,\mathrm{d}\mu$.

Let $c \in \mathrm{M}(\Omega \times \Omega, \bar{\mathbb{R}})$. The *c-transportation cost* of $\mu, \nu \in \mathfrak{P}(\Omega)$, and *c-transportation cost ball* of radius $r \geq 0$, centred at $\mu \in \mathfrak{P}(\Omega)$ are respectively

$$\mathrm{cost}_c(\mu, \nu) := \inf \left\{ \int c \,\mathrm{d}\pi : \pi \in \Pi(\mu, \nu) \right\} \quad \text{and} \quad \mathrm{cost}_c(\mu, r) := \{\nu \in \mathfrak{P}(\Omega) \mid \mathrm{cost}_c(\mu, \nu) \leq r\}.$$

A function $f : (\Omega, d) \to \bar{\mathbb{R}}$ (on a metric space) is *L-Lipschitz* if there exists $L > 0$ such that

$$\forall \omega_1, \omega_2 \in \mathrm{dom}\, f : \; |f(\omega_1) - f(\omega_2)| \leq L d(\omega_1, \omega_2).$$

The smallest such L is called the *Lipschitz constant* and is denoted as $\mathrm{lip}(f) := \sup_{x,y \in \mathrm{dom}(f)} |f(x) - f(y)|/d(x, y)$, with $0/0 := 0$. For $f : X \times Y \to \bar{\mathbb{R}}$ define its *X-Lipschitz constant*, $\mathrm{lip}_X(f) := \sup_{(x,y) \in \mathrm{dom}(f)} \mathrm{lip}(f(\,\cdot\,, y))$. Finally, denote $[n] := \{1, 2, \ldots, n\}$.

3 Certificate for Distributional Robustness

While an elegant concept, the DRR suffers from a lack of tractability. That is, in order to effectively minimise it, we first need to be able to compute or estimate it. When the loss function is convex with respect to the input space this is straight-forward, however in general approximations are necessary. Our first contribution is an upper bound for it.

Theorem 1. *Let (Ω, d) be a Polish space. Assume $\ell_f : \Omega \to \mathbb{R}$ is continuous. Then*

$$\sup_{\nu \in \mathrm{cost}_d(\mu, r)} \mathrm{risk}_\ell(f, \nu) \leq r \,\mathrm{lip}(\ell_f) + \mathrm{risk}_\ell(f, \mu), \tag{1}$$

where lip *is measured with respect to d. If additionally if ℓ_f is convex, then* (1) *is an equality.*

It follows from the proof of Theorem 1 that in the particular case of $\Omega := X \times Y$, where $(X, |\cdot|)$ is a normed linear space and Y is discrete, i.e. a typical setting for classification, the upper bound

$$\sup_{\nu \in \mathrm{cost}_d(\mu, r)} \mathrm{risk}_\ell(f, \nu) \leq r \,\mathrm{lip}_X(\ell_f) + \mathrm{risk}_\ell(f, \mu), \tag{2}$$

is attained for $d(\omega, \omega') := |\omega_X - \omega'_X| + \infty \,[\![\omega_Y \neq \omega'_Y]\!]$, where the subscripts refer to the components of ω, ω' belonging to X and Y. The *Iverson bracket* $[\![\,\cdot\,]\!]$ returns 1 if its argument is true and 0 else. Again (2) is an equality if $x \mapsto \ell_f(x, y)$ is convex for all $y \in Y$ (using Lemma 2 in Appendix A).

The approach here is not new, namely using strong duality to establish a tractable formulation of the DRR. However, previous approaches suffer from onerous conditions on the metric (Sinha et al., 2018), the model class, or family of distributions (Shafieezadeh-Abadeh et al., 2017). Theorem 1 subsumes many of these results with a great deal more generality, applying to an arbitrary family of models, loss functions, and outcome spaces.

3.1 Distributional robustness as adversarial robustness

We next show how Theorem 1 can be useful for adversarial learning. The following objective function has been proposed to build a robust classifier f (often) on an input–label space $X \times Y = \Omega$:

$$\int \mu(\mathrm{d}\omega) \max_{\varepsilon \in B} \ell_f(\omega + \varepsilon), \tag{3}$$

where $B \subseteq X \times \{0\}$ is typically built from a ℓ_p ball. We refer to (3) as the *adversarial risk* (Goodfellow et al., 2015; Madry et al., 2018; Shaham et al., 2018).

For a metric space (Ω, d), let $R_\mu(r) := \left\{ g \in \mathrm{M}(\Omega, \mathbb{R}_{\geq 0}) \mid \int g \,\mathrm{d}\mu \leq r \right\}$ be a set of radius functions, and $B(\omega, r) := \{\omega' \in \Omega \mid d(\omega, \omega') \leq r\}$ denote the ball of radius r centred at ω.

Theorem 2. *Let $(\Omega, |\cdot|)$ be a separable Banach space with induced metric d. Fix $\mu \in \mathfrak{P}(\Omega)$, and $\ell_f \in \mathrm{M}(\Omega, \mathbb{R})$. Then it holds for all $r \geq 0$ and $g \in R_\mu(r)$ that $\sup_{\omega' \in B(\,\cdot\,, g(\,\cdot\,))} \ell_f(\omega')$ is universally measurable, and there is the upper bound*

$$\sup_{g \in R_\mu(r)} \int \mu(\mathrm{d}\omega) \sup_{\omega' \in B(\omega, g(\omega))} \ell_f(\omega') \leq \sup_{\nu \in \mathrm{cost}_d(\mu, r)} \mathrm{risk}_\ell(f, \nu). \tag{4}$$

This can in turn be upperbounded by Theorem 1. Furthermore if Ω is compact with μ non-atomic and ℓ_f continuous, then (4) *is an equality. We refer to the left-hand side of* (4) *as the* variable-radius risk.

Because $R_\mu(r)$ will always contain the (constant) function identically equal to r, clearly

$$(3) = \int \mu(\mathrm{d}\omega) \sup_{\omega' \in B(\omega,r)} \ell_f(\omega') \leq \sup_{g \in R_\mu(r)} \int \mu(\mathrm{d}\omega) \sup_{\omega' \in B(\omega,g(\omega))} \ell_f(\omega') \leq \sup_{\nu \in \mathrm{cost}_d(\mu,r)} \mathrm{risk}_\ell(f,\nu). \quad (5)$$

That is, (5) shows the adversarial risk (3) is upper bounded by the left-hand side of (4), which by Theorem 2 is upper bounded by the DRR.

Both the adversarial risk and the variable-radius risk imply an uncertainty set over a collection of adversaries that may perturb the data. Figure 1 shows the practical difference between the kinds of adversaries in these uncertainty sets.

Remark 1 (tightness of Theorem 2). *With additional conditions it is possible to strengthen* (5) *to an equality. Shafieezadeh-Abadeh et al. (2017, Thm. 3.20) achieve equality in* (5) *for an empirical distribution* $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{(x_i,y_i)}$ *in the setting of linear classifiers with* $\Omega := \mathbb{R}^n \times \{-1,+1\}$*, and a metric on* Ω *such that* $d(\omega,\omega') = \infty$ *whenever* ω,ω' *differ in label.*

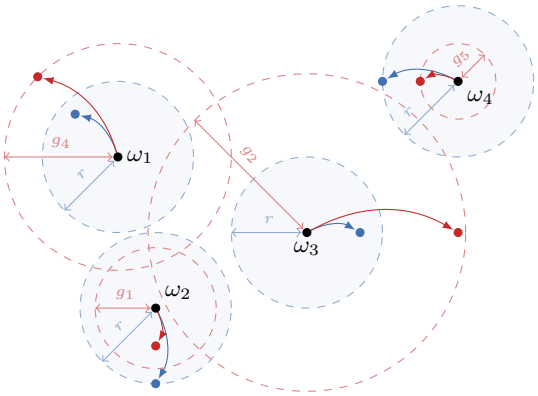


Figure 1: The blue set of perturbations is realisable with the classical adversarial risk (3), but the red is only available in the variable-radius risk.

In general, it is hard to characterize the tightness of the upper bounds in Theorem 1 and 2. So we resorted to an empirical approach, demonstrating that the **sum of the gaps** in these two theorems *and* in (5) is relatively low. We randomly generated 100 Gaussian kernel classifiers $f = \sum_{i=1}^{100} \gamma_i k(x^i, \cdot)$, with x^i sampled from the MNIST dataset and γ_i sampled uniformly from $[-2, 2]$. The bandwith was set to the median of pairwise distances. In Figure 2, the x-axis is the adversarial risk in (3), computed by PGD. The y-axis is the Lipschitz regularised empirical risk — right-hand side (RHS) of (1). The scattered dots lie closely to the diagonal, demonstrating that the above bounds are tight in practice.

4 Provable Lipschitz Regularisation for Kernel Methods

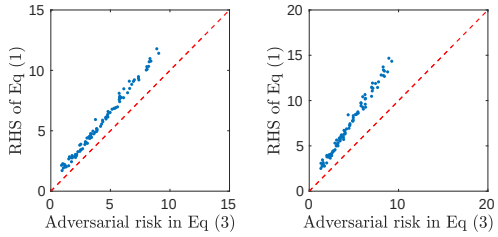
Theorems 1 and 2 open up a new path to optimising the adversarial risk (3) by Lipschitz regularisation (right-hand side of (1)), where the upper bounding relationship is established through DRR. In general, however, it is still hard to compute the Lipschitz constant for a nonlinear model. However, we will show that for some types of kernels, this can be done efficiently on functions in its RKHS. Thanks to the known connections between kernel method and deep learning, this technique will also potentially benefit the latter. For example, ℓ_1-regularized 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, Lemma 1 and Theorem 1) and (Shalev-Shwartz et al., 2011; Zhang et al., 2017), and even possibly Gaussian kernels $k(x,y) = \exp(-\|x-y\|^2/(2\sigma^2))$ (Shalev-Shwartz et al., 2011, §5).

Let us 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 $\boldsymbol{\gamma} := \{\gamma_a\}$, 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. (2017),

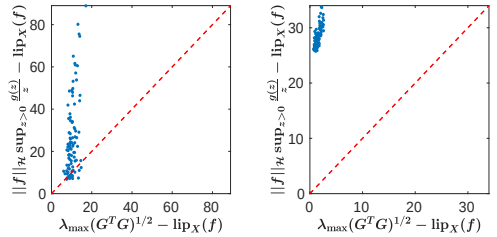
$$\sup_{x \in X} \|\nabla f(x)\|_2 \leq \|f\|_{\mathcal{H}} \sup_{z>0} z^{-1} g(z), \quad \text{where} \quad g(z) \geq \sup_{x,x' \in X : \|x-x'\|_2 = z} \|k(x,\cdot) - k(x',\cdot)\|_{\mathcal{H}}. \quad (6)$$

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 3. Besides, C and L are independent parameters.



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

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



(a) 5-layer inverse kernel (b) Gaussian kernel (σ=3)

Figure 3: Comparison of $\lambda_{\max}(G^\top G)$ and the RHS of (6), as upper bounds for the Lipschitz constant. Smaller values are tighter. 100 functions sampled in the same way as in Figure 2.

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. These are obviously convex constraints on $\boldsymbol{\gamma}$. But how many samples are needed in order to ensure $\|\nabla f(x)\|_2 \leq L + \varepsilon$ for *all* $x \in X$? Unfortunately, as shown in Appendix A.1, n may have to grow exponentially by $1/\varepsilon^d$ for a d-dimensional space. Therefore we seek a more efficient approach by first slightly relaxing $\|\nabla f(x)\|_2$. Let $g_j(x) := \partial^j f(x)$ 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 new bound

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

where $\lambda_{\max}$ evaluates the maximum eigenvalue, and $G := (g_1, \ldots, g_d)$. 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 $\langle g_i, g_j\rangle_{\mathcal{H}}$. Interestingly, $\lambda_{\max}(G^\top G)$ delivers *significantly lower* (i.e., tighter) value in approximating the Lipschitz constant $\sup_{x\in X} \|\nabla f(x)\|_2$, compared with $\|f\|_{\mathcal{H}} \max_{z>0} \frac{g(z)}{z}$ from (6). Figure 3 compared these two approximants, with $\lambda_{\max}(G^\top G)$ computed from (9) derived below, with the landmarks $\{w^s\}$ consisting of all training examples; drawing more samples led to little difference.

Such a positive result motivated us to develop refined algorithms to address the only remaining obstacle to leveraging $\lambda_{\max}(G^\top G)$: no analytic form for computation. Interestingly, it is readily approximable in both theory and practice. Indeed, the role of g_j can be approximated by $\tilde{g}_j$, where $\tilde{g}_j \in \mathbb{R}^n$ is the Nyström approximation (Drineas & Mahoney, 2005; Williams & Seeger, 2000):

$$\tilde{g}_j := K^{-1/2}(g_j(w^1), \ldots, g_j(w^n))^\top = (Z^\top Z)^{-1/2} Z^\top g_j \quad (\text{noting } g_j(w^1) = \langle g_j, k(w^1,\cdot)\rangle_{\mathcal{H}}) \quad (8)$$

$$\text{where} \quad K := [k(w^i, w^{i'})]_{i,i'}, \quad Z := (k(w^1,\cdot), k(w^2,\cdot), \ldots, k(w^n,\cdot)), \quad \tilde{G} := (\tilde{g}_1, \ldots, \tilde{g}_d).$$

So to ensure $\lambda_{\max}(G^\top G) \leq L^2 + \varepsilon$, intuitively we can resort to enforcing $\lambda_{\max}(\tilde{G}^\top \tilde{G}) \leq L^2$, which also retains the convexity in the constraint in $\boldsymbol{\gamma}$. 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 COORDINATE-WISE 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. We will 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 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 $k_0(x_j, y_j) = \exp\left(-\sin\left(\frac{\pi}{v}(x_j - y_j)\right)^2/(2\sigma^2)\right)$.

The key benefit of this decomposition 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 $g_1 = \frac{1}{l} \sum_{a=1}^l \gamma_a \partial^{0,1} k(x^a, \cdot)$, we can derive

$$\|g_1\|_{\mathcal{H}}^2 = l^{-2} \sum\nolimits_{a,b=1}^l \gamma_a \gamma_b \left\langle \partial^{0,1} k_0(x_1^a, \cdot), \partial^{0,1} k_0(x_1^b, \cdot)\right\rangle_{\mathcal{H}_0} \prod\nolimits_{j=2}^d k_0(x_j^a, x_j^b), \quad (9)$$

$$\langle g_1, g_2\rangle_{\mathcal{H}} = l^{-2} \sum\nolimits_{a,b=1}^l \gamma_a \gamma_b \partial^{0,1} k_0(x_1^a, x_1^b) \partial^{0,1} k_0(x_2^b, x_2^a) \prod\nolimits_{j=3}^d k_0(x_j^a, x_j^b).$$

So the off-diagonal entries of $G^\top G$ can be computed *exactly*. To approximate the diagonal, we sample $\{w_1^1, \ldots, w_1^n\}$ from μ_0, set $Z_1 = (k_0(w_1^1, \cdot), \ldots, k_0(w_1^n, \cdot))$, and apply Nyström approximation:

$$\left\langle \partial^{0,1} k_0(x_1^a, \cdot), \partial^{0,1} k_0(x_1^b, \cdot) \right\rangle_{\mathcal{H}_0} \approx \partial^{0,1} k_0(x_1^a, \cdot)^\top Z_1 \cdot (Z_1^\top Z_1)^{-1} \cdot Z_1^\top \partial^{0,1} k_0(x_1^b, \cdot) \tag{10}$$

$$\text{where} \quad Z_1^\top \partial^{0,1} k_0(x_1^b, \cdot) = (\partial^{0,1} k_0(x_1^b, w_1^1), \ldots, \partial^{0,1} k_0(x_1^b, w_1^n))^\top, \tag{11}$$

and analogously for $Z_1^\top \partial^{0,1} k_0(x_1^b, \cdot)$. We will denote this approximation of $G^\top G$ as $\tilde{P}_G$. Clearly, $\lambda_{\max}(\tilde{P}_G) \le L^2$ is a convex constraint on $\boldsymbol{\gamma}$, based on i.i.d. samples $\{w_j^s : s \in [n], j \in [d]\}$ from μ_0. It is now important to analyze how many samples w_j^s are needed, such that

$$\lambda_{\max}(\tilde{P}_G) \le L^2 \qquad \Longrightarrow \qquad \lambda_{\max}(G^\top G) \le L^2 + \varepsilon \quad \text{with high probability.}$$

4.2 General sample complexity and assumptions on the product kernel

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, Chapter 4) that the integral operator for k_0 and μ_0 is defined by

$$T_{k_0} \;=\; I \circ S : L_2(X_0, \mu_0) \to L_2(X_0, \mu_0)$$

$$\text{where } \; S \;:\; L_2(X_0, \mu_0) \to \mathcal{C}(X_0), \quad (Sf)(x) \;=\; \int k_0(x, y) f(y) d\mu_0(y), \quad f \in L_2(X_0, \mu_0),$$

and I: $\mathcal{C}(X_0) \hookrightarrow L_2(X_0; \mu_0)$ 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 $L_2(X_0, \mu_0)$ and $\{\lambda_j\}_{j \in J}$ with $\lambda_1 \ge \lambda_2 \ge \ldots > 0$ such that $T_{k_0} f = \sum_{j \in J} \lambda_j \langle f, \tilde{e}_j \rangle_{L_2(X_0; \mu_0)} \tilde{e}_j$ for all $f \in L_2(X_0, \mu_0)$. It is easy to see that $\varphi_j := \sqrt{\lambda_j} e_j$ is an orthonormal basis of $\mathcal{H}_0$ (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 $\varepsilon > 0$, there exists $N_\varepsilon \in \mathbb{N}$ 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 $\Phi_m := (\varphi_1, \ldots, \varphi_m)$,*

$$N_\varepsilon := \inf_m \Big\{ \left\| \partial^{0,1} k_0(x, \cdot) - \Phi_m \Phi_m^\top \partial^{0,1} k_0(x, \cdot) \right\|_{\mathcal{H}_0} < \varepsilon \text{ for all } x \in X_0 \text{ and}$$
$$\left\| k_0(x, \cdot) - \Phi_m \Phi_m^\top k_0(x, \cdot) \right\|_{\mathcal{H}_0} < \varepsilon \text{ for all } x \in X_0 \Big\} < \infty.$$

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

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

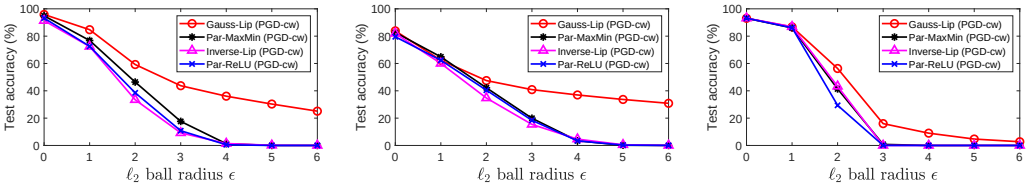
$$M_\varepsilon := \sup_{x \in X_0} \max_{j \in [N_\varepsilon]} \left| \left\langle \partial^{0,1} k_0(x, \cdot), e_j \right\rangle_{\mathcal{H}_0} \right| < \infty, \qquad \text{and} \qquad Q_\varepsilon := \sup_{x \in X_0} \max_{j \in [N_\varepsilon]} |e_j(x)| < \infty.$$

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 $\lambda_{\max}(\tilde{P}_G) \le L^2$, the Lipschitz condition $\lambda_{\max}(G^\top G) \le L^2 + \varepsilon$ is met with probability $1 - \delta$, as long as $n \ge \tilde{\Theta}\big(\frac{1}{\varepsilon^2} N_\varepsilon^2 M_\varepsilon^2 Q_\varepsilon^2 \log \frac{d N_\varepsilon}{\delta}\big)$, almost independent of d. Here $\tilde{\Theta}$ hides all poly-log terms.*

Satisfaction of Assumptions. In Appendix A.4 and A.5, we will show that for periodic kernel and Gaussian kernel, Assumptions 1 and 2 hold true with $\tilde{O}(1)$ values of N_ε, M_ε, and Q_ε. It remains open whether non-product kernels such as inverse kernel also enjoy this polynomial sample complexity. Appendix A.6 suggests that the complexity is *quasi-polynomial* for inverse kernels.

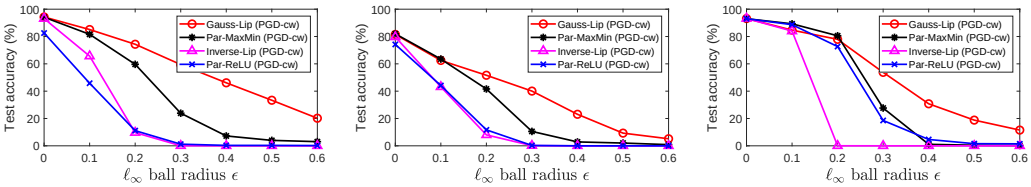
5 Experimental Results

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 defense algorithms under effective attacks.



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

Figure 4: Test accuracy under PGD attacks on the C&W approximation with ℓ_2 norm bound



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

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

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 $[0, 1]$. They were used as the input for training all the competing algorithms and were subject to attack.

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 ℓ_2 or ℓ_∞ ball. To evaluate robustness, we scaled the perturbation bound ε from 0.1 to 0.6 for ℓ_∞ norm, and from 1 to 6 for ℓ_2 norm (when $\varepsilon = 6$, the average coordinate magnitude $|\delta_i|$ is 0.214).

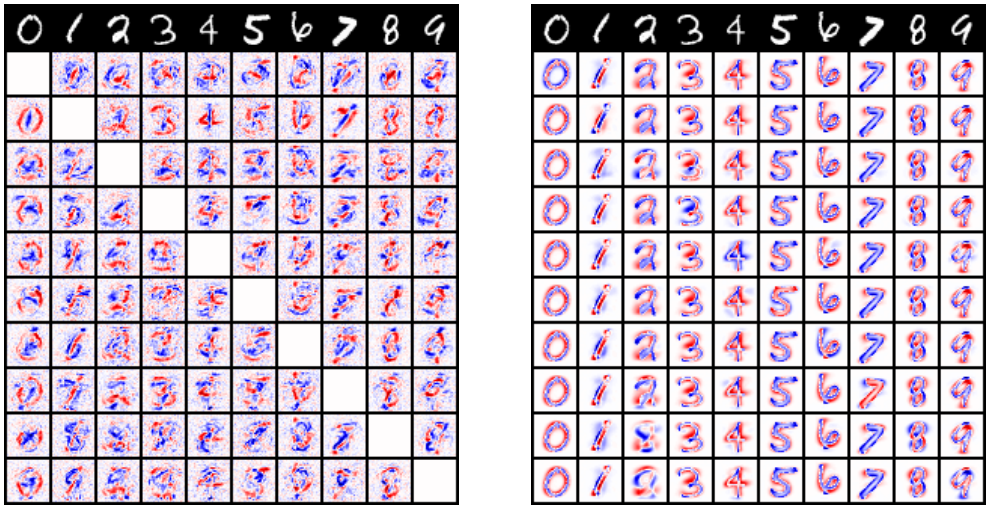
Algorithms. We compared four training algorithms. The Parseval network orthonormalises the weight matrices to enforce the Lipschitz constant (Cisse et al., 2017). We used three hidden layers of 1024 units and ReLU activation (Par-ReLU). Also considered is the Parseval network with MaxMin activations (Par-MaxMin), which enjoys much improved robustness (Anil et al., 2019). Both algorithms can be customised for ℓ_2 or ℓ_∞ attacks, and were trained under the corresponding norms. Using multi-class hinge loss, they constitute strong baselines for adversarial learning.

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 Eq (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 Appendix A.9. 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 $f^\tau := \sum_a \alpha_a^\tau k(x^a, \cdot)$ for each class $\tau \in [10]$, based on which a large supply 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 $\{f^\tau\}$ to a real-valued loss is typically at most 1, it suffices to bound the Lipschitz constant of $x \mapsto (f^1(x), \ldots, f^{10}(x))^\top$ by

$$\max_x \lambda_{\max}(G(x)G(x)^\top) \leq \max_{\|\varphi\|_{\mathcal{H}}=1} \lambda_{\max}\Big(\sum\nolimits_{c=1}^{10} G_c^\top \varphi\varphi^\top G_c\Big) \leq L^2, \tag{12}$$

where $G(x) := [\nabla f^1(x), \cdots, \nabla f^{10}(x)] = [G_1^\top k(x, \cdot), \ldots, G_{10}^\top k(x, \cdot)]$ with $G_c := (g_1^c, \ldots, g_d^c)$.



(a) ℓ_2 trained Par-MaxMin (b) Gauss-Lip with $\sigma = 2$

Figure 6: Gradients of targeted cross-entropy loss with respect to input images. One image per class (0-9) was sampled randomly from the test set, shown in the first row. The following rows show the gradient with different class targets. Red and blue stand for positive and negative pixel values, resp.

The last term in (12) can be approximated using the same technique as in the binary case. Furthermore, the principle can be extended to ℓ_∞ attacks, whose details are relegated to Appendix A.10.

Parameter selection. We used the same parameters as in Anil et al. (2019) for training Par-ReLU and Par-MaxMin. To defend against ℓ_2 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 ℓ_∞ 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 stacked layers.

Results. Figures 4 and 5 show how the test accuracy decays as an increasing amount of perturbation (ε) in ℓ_2 and ℓ_∞ 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 ℓ_2 and ℓ_∞ bounded PGD attacks with C&W loss. In contrast, Inverse-Lip only performs similarly to Par-ReLU. Interestingly, we noticed that ℓ_2 based Par-MaxMin are only slightly better than Par-ReLU under ℓ_2 attacks, although the former does perform significantly better under ℓ_∞ attacks.

For the sake of space, the results for cross-entropy PGD attacks are deferred to Figures 9 and 10 in Appendix A.11. 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 ℓ_2 and ℓ_∞ norms. The improved robustness of Gauss-Lip does not seem to be attributed to the obfuscated gradient (Athalye et al., 2018), because as shown Figures 4, 5, 9, 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 just one step.

Visualization. The gradient with respect to inputs is plotted in Figure 6 for ℓ_2 trained Par-MaxMin and Gauss-Lip. The i-th row and j-th column corresponds to the targeted attack of turning the original class j into a new class i, hence the gradient is on the cross-entropy loss with class i as the ground truth. These two figures also explained why Gauss-Lip is more robust than Par-MaxMin: the attacker can easily reduce the targeted cross-entropy loss by following the gradient as shown in Figure 6a, and hence successfully attack Par-MaxMin. In contrast, the gradient shown in Figure 6b does not provide much information on how to flip the class.

Conclusion. In this paper, we derived a new certificate for distributional robust risk minimization by using Lipschitz regularization. Application to adversarial learning based on kernel methods exhibited superior robustness, with provably polynomial sample complexity for product kernels. We will apply this function space to GANs to witness the difference between probability distributions, leading to a more stable training scheme as the inner level optimization becomes convex.

REFERENCES

Cem Anil, James Lucas, and Roger Grosse. Sorting out Lipschitz function approximation. In *International Conference on Machine Learning (ICML)*, 2019.

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

Andrew R. Barron. Approximation and estimation bounds for artificial neural networks. *Machine Learning*, 14(1):115–133, 1994.

Aharon Ben-Tal, Dick den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

Dimitri P Bertsekas and Steven E Shreve. *Stochastic optimal control: The discrete time case*. Academic Press, 1978.

C. Bhattacharyya, K. S. Pannagadatta, and A. J. Smola. A second order cone programming formulation for classifying missing data. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 153–160, 2005.

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

Alberto Bietti, Gregoire Mialon, Dexiong Chen, and Julien Mairal. A kernel perspective for regularizing deep neural networks. In *International Conference on Machine Learning (ICML)*, 2019.

Jose Blanchet and Karthyek Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 2019. doi: 10.1287/moor.2018.0936.

Jose Blanchet, Yang Kang, and Karthyek Murthy. Robust Wasserstein profile inference and applications to machine learning. *arXiv:1610.05627*, 2016.

Jose Blanchet, Yang Kang, Fan Zhang, and Karthyek Murthy. Data-driven optimal transport cost selection for distributionally robust optimization. *arXiv:1705.07152*, 2017.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39–57. IEEE, 2017.

D Aliprantis Charalambos and Kim C Border. *Infinite dimensional analysis: a hitchhiker's guide*. Springer, 2006.

Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. Parseval networks: Improving robustness to adversarial examples. In *International Conference on Machine Learning (ICML)*, 2017.

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

Erick Delage and Yinyu Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3):595–612, 2010.

P. Drineas and M. Mahoney. On the Nyström method for approximating a gram matrix for improved kernel-based learning. *Journal of Machine Learning Research (JMLR)*, 6:2153–2175, 2005.

John Duchi, Peter Glynn, and Hongseok Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *arXiv:1610.03425*, 2016.

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

Laurent El Ghaoui and Hervé Lebret. Robust solutions to least-squares problems with uncertain data. *SIAM J. Matrix Anal. Appl.*, 18(4):1035–1064, 1997.

Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1-2):115–166, 2018.

Farzan Farnia and David Tse. A minimax approach to supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Farzan Farnia, Jesse Zhang, and David Tse. Generalizable adversarial training via spectral normalization. In *International Conference on Learning Representations (ICLR)*, 2019.

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

Rui Gao and Anton J Kleywegt. Distributionally robust stochastic optimization with Wasserstein distance. *arXiv:1604.02199*, 2016.

Emmanuel Giner. Necessary and sufficient conditions for the interchange between infimum and the symbol of integration. *Set-Valued and Variational Analysis*, 17(4):321, 2009.

Joel Goh and Melvyn Sim. Distributionally robust optimization and its tractable approximations. *Operations Research*, 58(4):902–917, 2010.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv:1412.6572*, 2015.

Henry Gouk, Eibe Frank, Bernhard Pfahringer, and Michael Cree. Regularisation of neural networks by enforcing Lipschitz continuity. *arXiv:1804.04368*, 2018.

Zhaolin Hu and Jeff Liu Hong. Kullback-Leibler divergence constrained distributionally robust optimization, 2016. Available from Optimization Online.

Todd Huster, Cho-Yu Jason Chiang, and Ritu Chadha. Limitations of the Lipschitz constant as a defense against adversarial examples. *arXiv:1807.09705*, 2018.

L. Kantorovitch. On the translocation of masses. *Management Science*, 5(1):1–4, 1958.

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

John Lafferty and Guy Lebanon. Diffusion kernels on statistical manifolds. *Journal of Machine Learning Research (JMLR)*, 6:129–163, 01 2005.

Shao-Bo Lin, Xin Guo, and Ding-Xuan Zhou. Distributed learning with regularized least squares. *Journal of Machine Learning Research (JMLR)*, 18(92):1–31, 2017.

David J.C. MacKay. Introduction to gaussian processes. *NATO Adv Stud Inst Ser F Comput Syst Sci*, 168, 01 1998.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018.

E. J. McShane. Extension of range of functions. *Bull. Amer. Math. Soc.*, 40(12):837–842, 1934.

Charles A. Micchelli, Yuesheng Xu, and Haizhang Zhang. Universal kernels. *Journal of Machine Learning Research (JMLR)*, 2006.

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

Ha Quang Minh, Partha Niyogi, and Yuan Yao. Mercer's theorem, feature maps, and smoothing. In Gábor Lugosi and Hans Ulrich Simon (eds.), *Conference on Computational Learning Theory (COLT)*, pp. 154–168, 2006.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1):115–166, 2018.

Openreview. OpenReview comments on "Certifying some distributional robustness with principled adversarial training" by Aman Sinha, Hongseok Namkoong, and John Duchi. `https://openreview.net/forum?id=Hk6kPgZA-`, 2018.

Jean-Paul Penot. *Calculus without derivatives*, volume 266. Springer Science & Business Media, 2012.

Iosif Petrakis. Mcshane-Whitney extensions in constructive analysis. *arXiv preprint arXiv:1804.06757*, 2018.

Aldo Pratelli. 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, 2007.

Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2018.

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

Kevin Scaman and Aladin Virmaux. Lipschitz regularity of deep neural networks: analysis and efficient estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

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

Soroosh Shafieezadeh-Abadeh, Peyman Mohajerin Mohajerin Esfahani, and Daniel Kuhn. Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

Soroosh Shafieezadeh-Abadeh, Daniel Kuhn, and Peyman Mohajerin Esfahani. Regularization via mass transportation. *arXiv:1710.10016*, 2017.

Uri Shaham, Yutaro Yamada, and Sahand Negahban. Understanding adversarial training: Increasing local stability of supervised models through robust optimization. *Neurocomputing*, 307:195–204, 2018.

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

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

Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations (ICLR)*, 2018.

Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

Arun Sai Suggala, Adarsh Prasad, Vaishnavh Nagarajan, and Pradeep Ravikumar. Revisiting adversarial risk. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.

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

Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

Tsui-Wei Weng, Huan Zhang, Hongge Chen, Zhao Song, Cho-Jui Hsieh, Luca Daniel, Duane S. Boning, and Inderjit S. Dhillon. Towards fast computation of certified robustness for relu networks. In *International Conference on Machine Learning (ICML)*, 2018.

Hassler Whitney. Analytic extensions of differentiable functions defined in closed sets. *Transactions of the American Mathematical Society*, 36(1):63–89, 1934.

Wolfram Wiesemann, Daniel Kuhn, and Melvyn Sim. Distributionally robust convex optimization. *Operations Research*, 62(6):1358–1376, 2014.

C. K. I. Williams and M. Seeger. Using the Nyström method to speed up kernel machines. In T. G. Dietterich, S. Becker, and Z. Ghahramani (eds.), *Advances in Neural Information Processing Systems (NeurIPS)*, 2000.

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

Eric Wong and J. Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning (ICML)*, 2018.

Eric Wong, Frank Schmidt, Jan Hendrik Metzen, and J. Zico Kolter. Scaling provable adversarial defenses. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Huan Xu, Constantine Caramanis, and Shie Mannor. Robust regression and lasso. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2009a.

Huan Xu, Constantine Caramanis, and Shie Mannor. Robustness and regularization of support vector machines. *Journal of Machine Learning Research (JMLR)*, 10:1485–1510, 2009b.

Yuichi Yoshida and Takeru Miyato. Spectral norm regularization for improving the generalizability of deep learning. *arXiv:1705.10941*, 2017.

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

Yuchen Zhang, Percy Liang, and Martin Wainwright. Convexified convolutional neural networks. In *International Conference on Machine Learning (ICML)*, 2017.

Chaoyue Zhao and Yongpei Guan. Data-driven risk-averse stochastic optimization with Wasserstein metric. *Operations Research Letters*, 46(2):262–267, 2018.

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

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

Appendix

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

Algorithm 1: Training binary SVMs by enforcing Lipschitz constant L

1 Initialize the constraint set S by some random samples from X.
2 **for** $i = 1, 2, \ldots$ **do**
3 Train SVM using one of the following constraints:
① **Brute-force:** $\|\nabla f(w)\|_2^2 \leq L^2,\ \forall\, w \in S$
② **Nyström holistic:** $\lambda_{\max}(\tilde{G}^\top \tilde{G}) \leq L^2$ using S as the set $\{w^1, \ldots, w^n\}$ in Eq (8)
③ **Nyström coordinate wise:** $\lambda_{\max}(\tilde{P}_G) \leq L^2$ using S as the set $\{w^1, \ldots, w^n\}$ in Eq (10)
4 Let the trained SVM be $f^{(i)}$.
5 Find a new w to add to S by one of the following methods:
ⓐ **Random:** randomly sample w from X.
ⓑ **Greedy:** find $\arg\max_{x \in X} \left\|\nabla f^{(i)}(x)\right\|$ (approximately) by L-BFGS with 15 random initializations.
6 Record $L^{(i)} = \max_{x \in X} \left\|\nabla f^{(i)}(x)\right\|$.

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 15 random initializations 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 optimization (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 $\tilde{\varphi}(x^a) \in \mathbb{R}^{1000}$. Then $f(x)$ can be written as $\frac{1}{l}\sum_{a=1}^{l} \gamma_a \tilde{\varphi}(x^a)^\top \tilde{\varphi}(x)$. Defining $w = \frac{1}{l}\sum_{a=1}^{l} \gamma_a \tilde{\varphi}(x^a)$, we can equivalently optimize over w, and the RKHS norm bound on f can be equivalently imposed as the ℓ_2-norm bound on w.

To summarize, 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.

A Proofs of results

The following appendix contains the complete set of proofs and auxiliary results.

Proofs for §3: certificate for distributional robustness

Lemma 1 McShane–Whitney (Petrakis, 2018, Thm. 2.1 (ii)). *Let (X, d) be a metric space and $f : X \to \mathbb{R}$. Then*

$$\forall \gamma \geq \operatorname{lip}(f) : \sup_{x \in X} \Big(f(x) - \gamma d(x_0, x) \Big) = f(x_0),$$

where lip *is measured with respect to d.*

Lemma 2 . *Let $(X, |\cdot|)$ be a normed linear space and $f : X \to \mathbb{R}$ is convex. Then for $x_0 \in X$, $\gamma > 0$*

$$\sup_{x \in X} \Big(f(x) - \gamma |x - x_0| \Big) = f(x_0) + \infty [\![\operatorname{lip}(f) > \gamma]\!],$$

where lip *is measured with respect to $|\cdot|$.*

Proof. Since f is closed proper convex, $f = f^{**}$ (Penot, 2012, Thm. 3.44, p. 214). Then, using a standard result (Charalambos & Border, 2006, Lem. 6.10, p. 231) we have

$$\forall x \in X: \; -\gamma|x| = -\gamma \sup_{|g|^* \leq 1} \langle g, x \rangle = \inf_{|g|^* \leq \gamma} -\langle g, x \rangle, \tag{13}$$

we have

$$\begin{aligned}
\sup_{x \in X} \Big(f(x) - \gamma|x - x_0| \Big) &= \sup_{x \in X} \Big(\sup_{x^* \in \operatorname{dom} f^*} \langle x^*, x \rangle - f^*(x) - \gamma|x - x_0| \Big) \\
&\overset{(13)}{=} \sup_{x \in X} \sup_{x^* \in \operatorname{dom} f^*} \Big(\inf_{|g^*|^* \leq \gamma} \langle x^*, x \rangle - f^*(x) - \langle g^*, x - x_0 \rangle \Big) \\
&= \sup_{x \in X} \sup_{x^* \in \operatorname{dom} f^*} \inf_{|g^*|^* \leq \gamma} \Big(\langle x^* - g^*, x \rangle + \langle g^*, x_0 \rangle - f^*(x^*) \Big) \\
&= \inf_{|g^*|^* \leq \gamma} \sup_{x^* \in \operatorname{dom} f^*} \sup_{x \in X} \Big(\langle x^* - g^*, x \rangle + \langle g^*, x_0 \rangle - f^*(x^*) \Big),
\end{aligned}$$

where the exchange the infimum and supremum follows from Penot (2012, Thm. 1.86, p. 59). Thus

$$\begin{aligned}
&\inf_{|g^*|^* \leq \gamma} \sup_{x^* \in \operatorname{dom} f^*} \sup_{x \in X} \Big(\langle x^* - g^*, x \rangle + \langle g^*, x_0 \rangle - f^*(x^*) \Big) \\
&\quad = \sup_{x^* \in \operatorname{dom} f^*} \inf_{|g^*|^* \leq \gamma} \Big(\infty(1 - [\![x^* = g^*]\!]) + \langle g^*, x_0 \rangle - f^*(x^*) \Big), \\
&\quad = \sup_{x^* \in \operatorname{dom} f^*} \Big(\infty\big(1 - [\![|x^*|^* \leq \gamma]\!]\big) + \langle x^*, x_0 \rangle - f^*(x^*) \Big) \\
&\quad = \sup_{x^* \in \operatorname{dom} f^*} \Big(\langle x^*, x_0 \rangle - f^*(x^*) \Big) + \infty \, [\![\exists g^* \in \operatorname{dom} f^* : \; |g^*|^* > \gamma]\!] \\
&\quad = f(x_0) + \infty \, [\![\operatorname{lip}(f) > \gamma]\!] \,,
\end{aligned}$$

as claimed. ■

Theorem 1. *Let (Ω, d) be a Polish space. Assume $\ell_f : \Omega \to \mathbb{R}$ is continuous. Then*

$$\sup_{\nu \in \operatorname{cost}_d(\mu, r)} \operatorname{risk}_\ell(f, \nu) \leq r \operatorname{lip}(\ell_f) + \operatorname{risk}_\ell(f, \mu),$$

where lip *is measured with respect to d. If additionally if ℓ_f is convex, then* (1) *is an equality.*

Proof. It is easily verified that d and ℓ_f satisfy assumptions 1 and 2 of Blanchet & Murthy (2019, Thm. 1) giving us

$$\begin{aligned}
\sup_{\nu \in \operatorname{cost}_d(\mu, r)} \operatorname{risk}_\ell(f, \nu) &= \inf_{\lambda \geq 0} \Big(\lambda r + \int_\Omega \ell_f^\lambda \, \mathrm{d}\mu \Big) \\
&= \inf_{\lambda \geq 0} \Big(\lambda r + \int_\Omega \mu(\mathrm{d}\omega) \sup_{\omega' \in \Omega} \Big(\ell_f(\omega') - \lambda d(\omega, \omega') \Big) \Big) \\
&\leq \inf_{\lambda \geq \operatorname{lip}(\ell_f)} \Big(\lambda r + \int_\Omega \ell_f \, \mathrm{d}\mu \Big) \\
&= r \operatorname{lip}(\ell_f) + \int_\Omega \ell_f \, \mathrm{d}\mu,
\end{aligned} \tag{14}$$

where the inequality follows from Lemma 1. To achieve equality in (1) (under the additional assumptions) apply Lemma 2 to (14).

■

A different upper bound in Theorem 1 for non-convex functions could be built using Lemma 2

$$\begin{aligned}
\sup_{x \in X} \Big(f(x) - \gamma|x - x_0| \Big) &= \sup_{x \in X} \Big(f^{**}(x) - \gamma|x - x_0| + (f - f^{**})(x) \Big) \\
&\leq \sup_{x \in X} \Big(f^{**}(x) - \gamma|x - x_0| \Big) + |f - f^{**}|_\infty \\
&= f^{**}(x_0) + |f - f^{**}|_\infty + \infty \, [\![\operatorname{lip}(f^{**}) > \gamma]\!] \,,
\end{aligned}$$

but it's not clear if this is tighter.

Lemma 3 . *Let (Ω, d) be a Polish space and assume $\mu \in \mathfrak{P}(\Omega)$ is non-atomic. Let $M(\mu,\nu) := \{t \in \mathrm{M}(\Omega,\Omega) \mid t_{\#}\mu = \nu\}$, and $\Lambda_d(r) := \left\{t \in \mathrm{M}(\Omega,\Omega) \mid \int d\,\mathrm{d}(\mathrm{id}\times t)_{\#}\mu \le r\right\}$. If there is $\nu^\star \in \mathfrak{P}(\Omega)$ with $\mathrm{risk}_\ell(f,\nu^\star) = \sup_{\nu\in\mathrm{cost}_d(\mu,r)} \mathrm{risk}_\ell(f,\nu)$, then*

$$\sup_{t\in\Lambda_d(r)} \int d(\omega, t(\omega))\mu(\mathrm{d}\omega) = \sup_{\nu\in\mathrm{cost}_d(\mu,r)} \mathrm{risk}_\ell(f,\nu). \tag{15}$$

Proof. Since μ is non-atomic and d is (trivially) continuous it follows (via Pratelli, 2007, Thm. B) that

$$\forall \nu' \in \mathfrak{P}(\Omega) : \inf_{t\in M(\mu,\nu')} \int d(\omega, t(\omega))\mu(\mathrm{d}\omega) = \mathrm{cost}_d(\mu,\nu'). \tag{16}$$

Let $r^\star := \mathrm{cost}_d(\mu,\nu^\star)$, and note that $r^\star \le r$. Fix a sequence $(\varepsilon_k)_{k\in\mathbb{N}} \subseteq (0, r^\star)$ with $\varepsilon_k \to 0$. For $c \ge 0$ let $\nu_c := \mu + c(\nu^\star - \mu)$. Since cost_d metrizes the weak topology on $\mathfrak{P}(\Omega)$ (Villani, 2008, Thm. 6.9, p. 96) the mapping $c \mapsto \mathrm{cost}_d(\mu,\nu_c)$ is continuous, with

$$\mathrm{cost}_d(\mu,\nu_0) = 0, \text{ and } \mathrm{cost}_d(\mu,\nu_1) = r^\star.$$

Thus for each k there exists some $c_k > 0$ with $\mathrm{cost}_d(\mu,\nu_{c_k}) = r^\star - \varepsilon_k$ by the intermediate value theorem.

It's clear that for each k there exists $(t_n^k)_{n\in\mathbb{N}} \subseteq M(\mu,\nu_{c_k})$ with

$$\lim_n \int d\,\mathrm{d}(\mathrm{id}, t_n^k)_{\#}\,\mathrm{d}\mu = \inf_{t\in M(\mu,\nu_{c_k})} \int d\,\mathrm{d}(\mathrm{id},t)_{\#}\mu \overset{(16)}{=} \mathrm{cost}_d(\mu,\nu_{c_k}) = r^\star - \varepsilon_k.$$

Thereore the sequences $(t_n^k)_{n\in\mathbb{N}}$ must eventually lie within $\Lambda_d(r)$ for every k. It is without loss of generality that we assume, by passing to a subsequence if necessary, that $(t_n^k)_{n\in\mathbb{N}} \subseteq \Lambda_d(r)$ for all k. Since $\mathrm{cost}_d(\nu^\star,\nu_{c_k}) \to 0$ (by a similar continuity argument) we have the weak convergence $\nu_{c_k} \rightharpoonup \nu$, and therefore $(t_n^k)_{\#}\mu \rightharpoonup \nu^\star$ as $n,k\to\infty$.

Thus there exists $(t_n^k)_{n\in\mathbb{N}} \subseteq \Lambda_d(r)$ with $(t_n^k)_{\#}\mu \rightharpoonup \nu^\star$ and

$$\int \ell_f\,\mathrm{d}\nu^\star = \lim_k \lim_n \int \ell_f\,\mathrm{d}(t_n^k)_{\#}\mu. \tag{17}$$

Therefore

$$\int \ell_f\,\mathrm{d}\nu^\star = \sup_{\nu\in\mathrm{cost}_d(\mu,r)} \mathrm{risk}_\ell(f,\nu) \ge \sup_{t\in\Lambda_d(r)} \int \ell_f\,\mathrm{d}t_{\#}\mu \ge \lim_k \lim_n \int \ell_f\,\mathrm{d}(t_n^k)_{\#}\mu \overset{(17)}{=} \int \ell_f\,\mathrm{d}\nu^\star,$$

giving equality in (15). ■

Theorem 2. *Let $(\Omega, |\cdot|)$ be a separable Banach space with induced metric d. Fix $\mu \in \mathfrak{P}(\Omega)$, and $\ell_f \in \mathrm{M}(\Omega,\mathbb{R})$. Then it holds for all $r \ge 0$ and $g \in R_\mu(r)$ that $\sup_{\omega'\in B(\,\cdot\,,g(\,\cdot\,))} \ell_f(\omega')$ is universally measurable, and there is the upper bound*

$$\sup_{g\in R_\mu(r)} \int \mu(\mathrm{d}\omega) \sup_{\omega'\in B(\omega,g(\omega))} \ell_f(\omega') \le \sup_{\nu\in\mathrm{cost}_d(\mu,r)} \mathrm{risk}_\ell(f,\nu).$$

This can in turn be upperbounded by Theorem 1. Furthermore if Ω is compact with μ non-atomic and ℓ_f continuous, then (4) *is an equality. We refer to the left-hand side of* (4) *as the* variable-radius risk.

Proof. Measurability. First fix $\omega \in \Omega$. Because $B(\omega,r)$ is closed, the level sets

$$\mathrm{lev}_{>u}(\iota_{B(\omega,g(\omega))}) = \begin{cases} \Omega \setminus B(\omega,g(\omega)) & u \ge 0 \\ \Omega & u < 0. \end{cases}$$

are all Borel. Therefore $\iota_{B(\omega,g(\omega))}$ is Borel for each $\omega \in \Omega$. Let $d_\omega(\omega') := d(\omega,\omega')$. Now consider

$$\begin{aligned} \mathrm{lev}_{=0}(\iota_{B(\,\cdot\,,g(\,\cdot\,))}(\omega)) &= \{\omega' \in \Omega \mid d(\omega,\omega') \le g(\omega')\} \\ &= \{\omega' \in \Omega \mid 0 \le g(\omega') - d_\omega(\omega')\} \\ &= \mathrm{lev}_{\ge 0}(g - d_\omega). \end{aligned}$$

Since g and d_ω are Borel $\mathrm{lev}_{=0}(\iota_{B(\,\cdot\,,g(\,\cdot\,))}(\omega))$ is Borel. By a similar argument it's clear that $\mathrm{lev}_{>0}(\iota_{B(\,\cdot\,,g(\,\cdot\,))}(\omega)) = \mathrm{lev}_{<0}(g - d_\omega)$ is Borel too. Therefore $(\omega_1, \omega_2) \mapsto \iota_{B(\omega_1, g(\omega_1))}(\omega_2)$ is Borel.

Notice that

$$\forall \omega \in \Omega : \sup_{\omega' \in B(\omega, g(\omega))} \ell_f(\omega') = \sup_{\omega' \in \Omega} \ell_f(\omega') - \iota_{B(\omega, g(\omega))}(\omega'), \tag{18}$$

and so

$$\begin{aligned} &\mathrm{lev}_{>u}\left(\sup_{\omega' \in B(\,\cdot\,,g(\,\cdot\,))} \ell_f(\omega')\right) \\ &\qquad \overset{(18)}{=} \mathrm{lev}_{>u}\left(\sup_{\omega' \in \Omega} \ell_f(\omega') - \iota_{B(\,\cdot\,,g(\,\cdot\,))}(\omega')\right) \\ &\qquad = \mathrm{proj}_1\big\{(\omega_1, \omega_2) \in \Omega \times \Omega \mid \ell_f(\omega_2) - \iota_{B(\omega_1, g(\omega_1))}(\omega_2) > u\big\}. \end{aligned} \tag{19}$$

where $\mathrm{proj}_1(\omega_1, \omega_2) := \omega_1$. Since ℓ_f and $\iota_{B(\,\cdot\,,g(\,\cdot\,))}$ are Borel, the argument of the projection in (19) is Borel. The projection of a Borel set is universally universally measurable (via Bertsekas & Shreve, 1978, Prop. 7.39, Cor. 7.42.1), therefore $\sup_{\omega' \in B(\,\cdot\,,g(\,\cdot\,))} \ell_f(\omega')$ is universally measurable. □

Inequality (4). Let $\Gamma_g(\omega) := B(\omega, g(\omega))$ be the set-valued operator on Ω defined for the function $g : \Omega \to \mathbb{R}_{\geq 0}$ that specifies a radius at each point $\omega \in \Omega$. Let $\mathrm{M}_\mu(\Omega, \Gamma_g)$ denote the set of Borel mappings $\Omega \to \Omega$ that are selections of Γ_g μ-almost everywhere. Finally let

$$\Lambda_d(\mu, r) := \left\{ a \in \mathrm{M}(\Omega, \Omega) \mid \int d \,\mathrm{d}(\mathrm{id} \times a)_\# \mu \leq r \right\}.$$

To prove $\bigcup_{g \in R_\mu(r)} \mathrm{M}_\mu(\Omega, \Gamma_g) = \Lambda_d(\mu, r)$ suppose $a \in \bigcup_{g \in R_\mu(r)} \mathrm{M}_\mu(\Omega, \Gamma_g)$. Then $\int d \,\mathrm{d}a_\# \mu \leq r$, and this shows $a \in \Lambda_d(\mu, r)$, and $\bigcup_{g \in R_\mu(r)} \mathrm{M}_\mu(\Omega, \Gamma_g) \subseteq \Lambda_d(\mu, r)$. Let $a \in \Lambda_d(\mu, r)$ and $g_a(\omega) := d(\omega, a(\omega))$. Then $g_a \in R_\mu(r)$ and $a \in \mathrm{M}(\Omega, \Gamma_{g_a})$ trivially, thus $\bigcup_{g \in R_\mu(r)} \mathrm{M}_\mu(\Omega, \Gamma_g) \supseteq \Lambda_d(\mu, r)$. We have shown $\bigcup_{g \in R_\mu(r)} \mathrm{M}_\mu(\Omega, \Gamma_g) = \Lambda_d(\mu, r)$.

If $a \in \Lambda_d(\mu, r)$, then

$$\int d \,\mathrm{d}(\mathrm{id} \times a)_\# \mu \leq r \implies a_\# \mu \in \mathrm{cost}_d(\mu, r),$$

which shows $\{a_\# \mu \mid a \in \Lambda_d(\mu, r)\} \subseteq \mathrm{cost}_d(\mu, r)$.

Considering the right hand side of (4)

$$\begin{aligned} \sup_{g \in R_\mu(r)} \int \sup_{\omega' \in \Gamma_g(\omega)} \ell_f(\omega') \,\mathrm{d}\mu &= \sup_{g \in R_\mu(r)} \sup_{a \in \mathrm{M}_\mu(\Omega, \Gamma_g)} \int \ell_f \,\mathrm{d}a_\# \mu \qquad (*) \\ &= \sup_{a \in \bigcup_{g \in R_\mu(r)} \mathrm{M}_\mu(\Omega, \Gamma_g)} \int \ell_f \,\mathrm{d}a_\# \mu. \\ &= \sup_{a \in \Lambda_d(\mu, r)} \int \ell_f \,\mathrm{d}a_\# \mu. \end{aligned} \tag{20}$$

We have already proven $\{a_\# \mu \mid a \in \Lambda_d(\mu, r)\} \subseteq \mathrm{cost}_d(\mu, r)$. Therefore, in summary:

$$\sup_{g \in R_\mu(r)} \int \sup_{\omega' \in \Gamma_g(\omega)} \ell_f(\omega') \mu(\mathrm{d}\omega) \overset{(20)}{=} \sup_{a \in \Lambda_d(\mu, r)} \int \ell_f \,\mathrm{d}a_\# \mu \leq \sup_{\nu \in \mathrm{cost}_d(\mu, r)} \int \ell_f \,\mathrm{d}\nu,$$

which was to be proved.

To complete the proof we will now justify the exchange of integration and supremum in (*). The set $\mathrm{M}_\mu(\Omega, \Gamma_g)$ is trivially decomposable (Giner, 2009, see the remark at the bottom of p. 323, Def. 2.1). By assumption ℓ_f is Borel measurable. Since ℓ_f is measurable, any decomposable subset of $\mathrm{M}(\Omega, \Omega)$ is ℓ_f-decomposable (Giner, 2009, Prop. 5.3) and ℓ_f-linked (Giner, 2009, Prop. 3.7 (i)). Theorem 6.1 (c) of Giner (2009) allows us to exchange integration and supremum in (*). □

Equality (4). Let $\Phi_d(\mu, r) := \{\pi \in \bigcup_{\nu \in \mathfrak{P}(\Omega)} \Pi(\mu, \nu) \mid \int d \, \mathrm{d}\pi \leq r\}$. Under the additional assumptions, from Blanchet & Murthy (2019, Prop. 2) there exists $\pi^\star \in \Phi_d(\mu, r)$ with

$$\int \ell_f(\omega)\nu^\star(\mathrm{d}\omega) = \sup_{\nu \in \mathrm{cost}_d(\mu, r)} \int \ell_f \, \mathrm{d}\nu,$$

where $\nu^\star := \pi^\star(\,\cdot\,, \mathrm{d}\omega)$. The proof then follows from Lemma 3. □

This concludes the proof of Theorem 2. ■

Proofs and more results for §4: Kernel Approximation

A.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 $O(\frac{1}{\varepsilon^d})$ samples to guarantee $\sum_j \|g_j\|_{\mathcal{H}}^2 \leq L^2 + \varepsilon$ w.h.p. This is illustrated in Figure 7, 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.

A.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 from Chapter 4 of Steinwart & Christmann (2008) that the integral operator for k and μ is defined by

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

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 $L_2(X, \mu)$ and $\{\lambda_j\}_{j \in J}$ with $\lambda_1 \geq \lambda_2 \geq \ldots > 0$ such that

$$Tf = \sum_{j \in J} \lambda_j \langle f, \tilde{e}_j \rangle_{L_2(X;\mu)} \tilde{e}_j, \qquad f \in L_2(X, \mu).$$

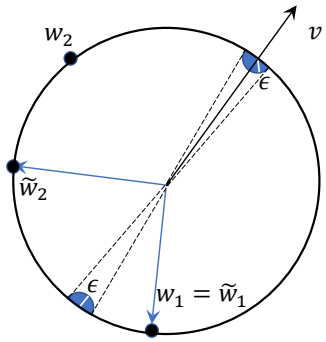


Figure 7: Suppose we use a polynomial kernel with degree 2, and $f(x) = \frac{1}{2}(v^\top x)^2$ for $x \in B$. Then $g(x) = (v^\top x)v$. If we want to test whether $\sup_{x\in B} \|g(x)\|_2 \leq 1$ by evaluating $\|g(w)\|_2$ on w that is randomly sampled from B such as w_1 and w_2, we must sample within 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 $\{\tilde{w}_1, \tilde{w}_2\}$.

In particular, we have $\langle \tilde{e}_i, \tilde{e}_j \rangle_{L_2(X;\mu)} = \delta_{ij}$ (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 $e_j = \lambda_j^{-1} S_k \tilde{e}_j \in \mathcal{C}(X)$, which clearly satisfies

$$\langle e_i, e_j \rangle_{L_2(X;\mu)} = \delta_{ij}, \quad Te_j = \lambda_j e_j.$$

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,

$$k(x, y) = \sum_{j\in J} \lambda_j e_j(x) e_j(y), \tag{21}$$

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

$$\varphi_j := \sqrt{\lambda_j} e_j, \qquad j \in J$$

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 $\Phi_m := (\varphi_1, \ldots, \varphi_m)$ and $\Lambda_m = \mathrm{diag}(\lambda_1, \ldots, \lambda_m)$.

A.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 4. *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 $Z :=$*

$(k_0(w^1,\cdot), k_0(w^2,\cdot), \ldots, k_0(w^n,\cdot))$, *and* $g_1 = \frac{1}{l}\sum_{a=1}^{l} \gamma_a g_1^a \colon X_0^d \to \mathbb{R}$, *where* $\|\boldsymbol{\gamma}\|_\infty \le c_1$ *and*

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

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

$$\|g_1\|_{\mathcal{H}}^2 \le \frac{1}{l^2} \boldsymbol{\gamma}^\top K_1 \boldsymbol{\gamma} + 3c_1 \Big(1 + 2\sqrt{N_\varepsilon} M_\varepsilon\Big)\varepsilon, \tag{22}$$

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

Then we obtain the formal statement of sample complexity, as stated in the following corollary, by combining all the coordinates from Theorem 4.

Corollary 1. *Suppose all coordinates share the same set of samples* $\{w^s\}_{s=1}^n$. *Applying the results in* (22) *for coordinates from 1 to* d *and using the union bound, we have that with sample size* $n = \max(N_\varepsilon, \frac{5}{3\varepsilon^2} N_\varepsilon Q_\varepsilon^2 \log \frac{2N_\varepsilon}{\delta})$, *the following holds with probability* $1 - d\delta$,

$$\lambda_{\max}(G^\top G) \le \lambda_{\max}(\tilde{P}_G) + 3c_1 \Big(1 + 2\sqrt{N_\varepsilon} M_\varepsilon\Big)\varepsilon. \tag{23}$$

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* $\lambda_{\max}(G^\top G) \le \lambda_{\max}(\tilde{P}_G) + \varepsilon$ *with probability* $1-\delta$, *the sample size needs to be*

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

Remark 2. *The first term on the right-hand side of* (23) *is explicitly upper bounded by* L^2 *in our training objective. In the case of Theorem 5, the values of* Q_ε, N_ε, *and* M_ε *lead to a* $\tilde{O}(\frac{1}{\varepsilon^2})$ *sample complexity. If we further zoom into the dependence on the period* v, *then note that* N_ε *is almost a universal constant while* $M_\varepsilon = \frac{\sqrt{2}\pi}{v}(N_\varepsilon - 1)$. *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 3. *Corollary 1 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 4. Let $\varepsilon' := (1 + 2\sqrt{m} M_\varepsilon)\varepsilon$. Since $\langle g_1^a, g_1^b \rangle_{\mathcal{H}} = \langle h_1^a, h_1^b \rangle_{\mathcal{H}_0} \prod_{j=2}^{d} k_0(x_j^a, x_j^b)$ and $\big|k_0(x_j^a, x_j^b)\big| \le 1$, it suffices to show that for all $a, b \in [l]$,

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

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

$$\Big| h^\top Z (Z^\top Z)^{-1} Z^\top h - \|h\|_{\mathcal{H}_0}^2 \Big| \le \varepsilon'. \tag{24}$$

This is because, if so, then

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

The rest of the proof is devoted to (24). Since $n \geq m$, 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 $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_m)$. Define

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

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

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

Therefore to prove (24), 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

$$\begin{aligned}\|Z\boldsymbol{\alpha} - h\|_{\mathcal{H}_0} \leq & \left\|(Z - \Phi_m\Phi_m^\top Z)\boldsymbol{\alpha}\right\|_{\mathcal{H}_0} + \left\|(\Phi_m\Phi_m^\top Z - \sqrt{n}\Phi_m\Lambda_m^{1/2}UV^\top)\boldsymbol{\alpha}\right\|_{\mathcal{H}_0} \\ & + \left\|\Phi_m\Phi_m^\top h - h\right\|_{\mathcal{H}_0}.\end{aligned} \tag{25}$$

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

$$\begin{aligned}&\left\|\Phi_m\Phi_m^\top h - h\right\|_{\mathcal{H}_0} \\ \leq & |\vartheta_x| \left\|\Phi_m\Phi_m^\top \partial^{0,1}k_0(x,\cdot) - \partial^{0,1}k_0(x,\cdot)\right\|_{\mathcal{H}_0} + |\vartheta_y| \left\|\Phi_m\Phi_m^\top \partial^{0,1}k_0(y,\cdot) - \partial^{0,1}k_0(y,\cdot)\right\|_{\mathcal{H}_0} \\ \leq & (|\vartheta_x| + |\vartheta_y|)\varepsilon \leq \varepsilon.\end{aligned}$$

We will next bound the first two terms on the right-hand side of (25).

(i) By Assumption 1, $\left\|k_0(w^s,\cdot) - \Phi_m\Phi_m^\top k_0(w^s,\cdot)\right\|_{\mathcal{H}_0} \leq \varepsilon$, hence $\left\|(Z - \Phi_m\Phi_m^\top Z)\boldsymbol{\alpha}\right\|_{\mathcal{H}_0} \leq \varepsilon\sqrt{n}\|\boldsymbol{\alpha}\|_2$. To bound $\|\boldsymbol{\alpha}\|_2$, note all singular values of $VU^\top$ are 1, and so Assumption 2 implies that for all $i \in [m]$,

$$\begin{aligned}\left|\lambda_j^{-1/2}\langle\varphi_j, h\rangle_{\mathcal{H}_0}\right| = \left|\langle e_j, h\rangle_{\mathcal{H}_0}\right| &= \left|\left\langle e_j, \vartheta_x\partial^{0,1}k_0(x,\cdot) + \vartheta_y\partial^{0,1}k_0(y,\cdot)\right\rangle_{\mathcal{H}_0}\right| \\ &\leq \sup_{x\in X}\left|\left\langle e_j, \partial^{0,1}k(x,\cdot)\right\rangle_{\mathcal{H}_0}\right| \leq M_\varepsilon.\end{aligned} \tag{26}$$

As a result,

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

(ii) We first consider the concentration of the matrix $R := \frac{1}{n}\Lambda_m^{-1/2}\Phi_m^\top Z Z^\top \Phi_m \Lambda_m^{-1/2} \in \mathbb{R}^{m\times m}$. 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 $\Pr\Big(\|R - I_m\|_{sp} \leq \varepsilon\Big) \geq 1 - \delta$ when $n \geq O(.)$. This is because $\|(e_1(x), \ldots, e_m(x))\|^2 \leq mQ_\varepsilon^2$, $\left\|\mathbb{E}_{\{w_s\}}[RR^\top]\right\|_{sp} \leq mQ_\varepsilon^2/n$, and

$$\Pr\Big(\|R - I_m\|_{sp} \leq \varepsilon\Big) \geq 1 - 2m\exp\left(\frac{-\varepsilon^2}{\frac{mQ_\varepsilon^2}{n}\left(1 + \frac{2}{3}\varepsilon\right)}\right) \geq 1 - 2m\exp\left(\frac{-\varepsilon^2}{\frac{5mQ_\varepsilon^2}{3n}}\right) \geq 1 - \delta,$$

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

$$\left|\frac{1}{n}\sigma_i^2 - 1\right| \leq \varepsilon \qquad \text{which implies} \qquad \left|\frac{1}{\sqrt{n}}\sigma_i - 1\right| < \varepsilon\left|\frac{1}{\sqrt{n}}\sigma_i + 1\right|^{-1} \leq \varepsilon. \tag{27}$$

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

$$\begin{aligned}
&\left\|(\Phi_m\Phi_m^\top Z - \sqrt{n}\Phi_m\Lambda_m^{1/2}UV^\top)\boldsymbol{\alpha}\right\|_{\mathcal{H}_0} \\
=&\left\|\Phi_m\Lambda_m^{1/2}U\Sigma V^\top\frac{1}{\sqrt{n}}VU^\top\Lambda_m^{-1/2}\Phi_m^\top h - \sqrt{n}\Phi_m\Lambda_m^{1/2}UV^\top\frac{1}{\sqrt{n}}VU^\top\Lambda_m^{-1/2}\Phi_m^\top h\right\|_{\mathcal{H}_0} \\
=&\left\|\Lambda_m^{1/2}U\left(\frac{1}{\sqrt{n}}\Sigma - I_m\right)U^\top\Lambda_m^{-1/2}\Phi_m^\top h\right\|_2 \qquad \text{(because } \Phi_m^\top\Phi_m = I_m\text{)} \\
\leq&\sqrt{\lambda_1}\max_{i\in[m]}\left|\frac{1}{\sqrt{n}}\sigma_i - 1\right|\left\|\Lambda_m^{-1/2}\Phi_m^\top h\right\|_2 \\
\leq&\varepsilon\sqrt{m}M_\varepsilon \qquad \text{(by (27), (26), and } \lambda_1 \leq 1\text{)}.
\end{aligned}$$

Combining (i) and (ii), we arrive at the desired bound in (22). ■

Proof of Corollary 1. 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

$$\begin{aligned}
\lambda_{\max}(\tilde{P}_G) - \lambda_{\max}(G^\top G) &\leq \mathbf{u}^\top\tilde{P}_G\mathbf{u} - \mathbf{u}^\top G^\top G\mathbf{u} = \mathbf{u}^\top(\tilde{P}_G - G^\top G)\mathbf{u} = \sum_j \delta_j\mathbf{u}_j^2 \\
\text{(by (22))} \quad &\leq 3c_1\left(1 + 2\sqrt{N_\varepsilon}M_\varepsilon\right)\varepsilon.
\end{aligned}$$

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

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

Periodic kernels on $X_0 := \mathbb{R}$ 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):

$$k_0(x,y) = \exp\left(-\frac{1}{2\sigma^2}\sin\left(\frac{\pi}{v}(x-y)\right)^2\right). \tag{28}$$

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 $j \in \mathbb{Z}$,

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

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

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

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 (21) 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 5. *Suppose the periodic kernel with period v has eigenvalues λ_j that satisfies*

$$\lambda_j (1+j)^2 \max(1, j^2)(1 + \delta(j \geq 1)) \leq c_6 \cdot c_4^{-j}, \quad \textit{for all } j \geq 0, \tag{29}$$

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

$$N_\varepsilon = 1 + 2\lfloor n_\varepsilon \rfloor, \quad \textit{where} \quad n_\varepsilon := \log_{c_4}\left(\frac{2.1 c_6}{\varepsilon^2} \max\left(1, \frac{v^2}{4\pi^2}\right)\right). \tag{30}$$

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

For example, if we set $v = \pi$ and $\sigma^2 = 1/2$ in the kernel in (28), elementary calculation shows that the condition (29) is satisfied with $c_4 = 2$ and $c_6 = 1.6$.

Proof of Theorem 5. First we show that $h(x) := \partial^{0,1} k_0(x_0, x)$ 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

$$h(x) = \sum_{j\in\mathbb{Z}} \lambda_j e_j(x_0) \partial^1 e_j(x) = \sum_{j\in\mathbb{Z}} \lambda_j e_j(x_0)(-j\omega_0 e_{-j}(x)) = \omega_0 \sum_{j\in\mathbb{Z}} \lambda_j j e_{-j}(x_0) e_j(x). \tag{31}$$

$h(x)$ 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 (29):

$$\begin{aligned}
\omega_0^{-2} \|h\|_{\mathcal{H}_0}^2 &= \sum_j \lambda_j j^2 e_{-j}^2(x_0) = \sum_{j\in\mathbb{Z}} \lambda_j j^2 e_{-j}^2(x_0) \\
&= \sum_{j\in\mathbb{Z}} \lambda_j j^2 e_{-j}^2(x_0) = \lambda_0 + 2\sum_{j\geq 1} j^2 \lambda_j \leq \frac{2c_4 c_5}{c_4 - 1}.
\end{aligned}$$

Finally to derive N_ε, we reuse the orthonormal decomposition of $h(x)$ in (31). 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

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

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

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

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

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

To upper bound the cardinality of A, we consider the conditions for $j \notin A$. Thanks to the conditions in (29), we know that any j satisfying the following relationship cannot be in A:

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

So $A \subseteq \{j : |j| \le n_\varepsilon\}$, which yields the conclusion (30). Finally $Q_\varepsilon \le \sqrt{2}$, and to bound M_ε, we simply reuse (31). For any j with $|j| \le n_\varepsilon$,

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

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

Gaussian kernels $k(x,y) = \exp(-\left\|x-y\right\|^2/(2\sigma^2))$ 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 $\varepsilon^2 = \alpha^2 = (2\sigma^2)^{-1}$ in Eq 12 and 13 of E Fasshauer (2011), the eigenvalue and eigenfunctions are (for $j \ge 0$):

$$\lambda_j = c_0^{-j-1/2}, \quad \text{where} \quad c_0 = \frac{1}{2}(3+\sqrt{5})$$

$$e_j(x) = \frac{5^{1/8}}{2^{j/2}}\exp\left(-\frac{\sqrt{5}-1}{4}\frac{x^2}{\sigma^2}\right)\frac{1}{\sqrt{j!}}H_j\left(\sqrt[4]{1.25}\,\frac{x}{\sigma}\right),$$

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 (11) 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 $Q_\varepsilon = \sqrt{2}$. If we standardize by $\sigma = 1$ and set $v = 5\pi$ as an example, it is not hard to see that (29) holds with $c_4 = 1.25$ and $c_6 = 50$. The expressions of N_ε and M_ε then follow from Theorem 5 directly.

A.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 (Lafferty & Lebanon, 2005; Zhou, 2002), 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 $d^{\lceil \log_2 \frac{1}{\varepsilon} \rceil + 1}$. This is a quasi-quadratic bound if we view d and $1/\varepsilon$ 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

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

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

$$
\mathcal{H} = \left\{ f = \sum_{\boldsymbol{\alpha}} \vartheta_{\boldsymbol{\alpha}} w_{\boldsymbol{\alpha}} \mathbf{x}^{\boldsymbol{\alpha}} : \|\boldsymbol{\vartheta}\|_{\ell_2} < \infty \right\}. \tag{32}
$$

Note this is just an intuitive "derivation" while a rigorous proof for (32) can be constructed in analogy to that of Theorem 1 in Minh (2010).

A.7 Background of eigenvalues of a kernel

We now use (32) 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 $f = \sum_{\boldsymbol{\alpha}} \vartheta_{\boldsymbol{\alpha}} w_{\boldsymbol{\alpha}} \mathbf{x}^{\boldsymbol{\alpha}}$ 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

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

Since B is an open set, that means

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

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 $Q = [w_{\boldsymbol{\alpha}} w_{\boldsymbol{\beta}} q_{\boldsymbol{\alpha}+\boldsymbol{\beta}}]_{\boldsymbol{\alpha},\boldsymbol{\beta}}$,

A.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 $\lambda_1 \geq \lambda_2 \geq \ldots$. Let the row corresponding to $\boldsymbol{\alpha}$ have ℓ_2 norm $r_{\boldsymbol{\alpha}}$, and let them be sorted as $r_{[1]} \geq r_{[2]} \geq \ldots$. Then by Schneider (1953); Shi & Wang (1965), we have

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

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:

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

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

$$\begin{aligned}
r_{\boldsymbol{\alpha}}^2 &= w_{\boldsymbol{\alpha}}^2 \sum_{\boldsymbol{\beta}} w_{\boldsymbol{\beta}}^2 q_{\boldsymbol{\alpha}+\boldsymbol{\beta}}^2 \leq w_{\boldsymbol{\alpha}}^2 q_{2\boldsymbol{\alpha}} \sum_{\boldsymbol{\beta}} w_{\boldsymbol{\beta}}^2 q_{2\boldsymbol{\beta}} = w_{\boldsymbol{\alpha}}^2 q_{2\boldsymbol{\alpha}} \sum_{k=0}^{\infty} 2^{-k-1} \sum_{\boldsymbol{\beta}:|\boldsymbol{\beta}|=k} C_{\boldsymbol{\beta}}^k q_{2\boldsymbol{\beta}} \\
&\leq w_{\boldsymbol{\alpha}}^2 q_{2\boldsymbol{\alpha}} \sum_{k=0}^{\infty} 2^{-k-1} \begin{pmatrix} k+d \\ d \end{pmatrix} \max_{|\boldsymbol{\beta}|=k} C_{\boldsymbol{\beta}}^k q_{2\boldsymbol{\beta}} \\
&= w_{\boldsymbol{\alpha}}^2 q_{2\boldsymbol{\alpha}} \sum_{k=0}^{\infty} 2^{-k-1} \begin{pmatrix} k+d \\ d \end{pmatrix} \max_{|\boldsymbol{\beta}|=k} \frac{k!}{\prod_{i=1}^d \beta_i!} \cdot \frac{2\prod_{i=1}^d \Gamma(\beta_i+\frac{1}{2})}{V_d\cdot(2k+d)\cdot\Gamma(k+\frac{d}{2})} \\
&= w_{\boldsymbol{\alpha}}^2 q_{2\boldsymbol{\alpha}} V_d^{-1} \sum_{k=0}^{\infty} 2^{-k} \begin{pmatrix} k+d \\ d \end{pmatrix} \frac{k!}{(2k+d)\Gamma(k+\frac{d}{2})} \cdot \max_{|\boldsymbol{\beta}|=k} \prod_{i=1}^{d} \frac{\Gamma(\beta_i+\frac{1}{2})}{\beta_i!} \\
&< w_{\boldsymbol{\alpha}}^2 q_{2\boldsymbol{\alpha}} \cdot \frac{b!}{\pi^b d!} \cdot \sum_{k=0}^{\infty} 2^{-k-1} \frac{(k+d)!}{(k+b)!} \qquad (\text{since } \Gamma(\beta_i+\tfrac{1}{2}) < \Gamma(\beta_i+1) = \beta_i!).
\end{aligned}$$

The summation over k can be bounded by

$$\sum_{k=0}^{\infty} 2^{-k-1}\frac{(k+d)!}{(k+b)!} = \frac{1}{2}b!\bigg(2^d + \begin{pmatrix} d \\ b \end{pmatrix}\bigg) \leq \frac{1}{2}\big(b!2^d + 2^b\big) \leq b!2^d,$$

where the first equality used the identity $\sum_{k=1}^{\infty} 2^{-k}\begin{pmatrix} d+k \\ b \end{pmatrix} = 2^d$. Letting $l := |\boldsymbol{\alpha}|$, we can continue by

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

This bound depends on $\boldsymbol{\alpha}$, not directly on $\boldsymbol{\alpha}$. Letting $n_l = \binom{l+d-1}{l}$ and $N_L = \sum_{l=0}^{L} n_l = \binom{d+L}{L}$, it follows that

$$\sum_{l=0}^{L} l n_l = \sum_{l=1}^{L} \frac{l(l+d)!}{d! \cdot l!} = (d+1) \sum_{l=1}^{L} \frac{(l+d)!}{(d+1)!(l-1)!}$$
$$= (d+1) \sum_{l=1}^{L} \binom{l+d}{d+1} = (d+1) \binom{L+d+1}{d+2}.$$

Now we can bound λ_{N_L} by

$$\lambda_{N_L}^{N_L} \leq \prod_{i=1}^{N_L} \lambda_i \leq \prod_{l=0}^{L} \left(2^{-l+b-1} \pi^{-2b} \binom{l+b}{l}^{-1} \right)^{n_l}$$
$$\Rightarrow \quad \log \lambda_{N_L} \leq N_L^{-1} \sum_{l=0}^{L} n_l \left(-(l-b+1) \log 2 - 2b \log \pi - \log \binom{l+b}{l} \right)$$
$$\leq -N_L^{-1} \cdot \log 2 \cdot \sum_{l=0}^{L} l n_l \qquad (\text{since } \log 2 < 2 \log \pi \text{ as the coefficients of } b)$$
$$= -\binom{d+L+1}{d+1}^{-1} \cdot \log 2 \cdot (d+1) \binom{d+L+1}{d+2}$$
$$= -\frac{d+1}{d+2} L \log 2$$
$$\approx -L \log 2$$
$$\Rightarrow \quad \lambda_{N_L} \leq 2^{-L}.$$

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

$$i \geq d^{\lceil \log_2 \frac{1}{\varepsilon} \rceil + 1}.$$

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).

EXPERIMENTS

A.9 EFFICIENCY OF ENFORCING LIPSCHITZ CONSTANT BY DIFFERENT METHODS

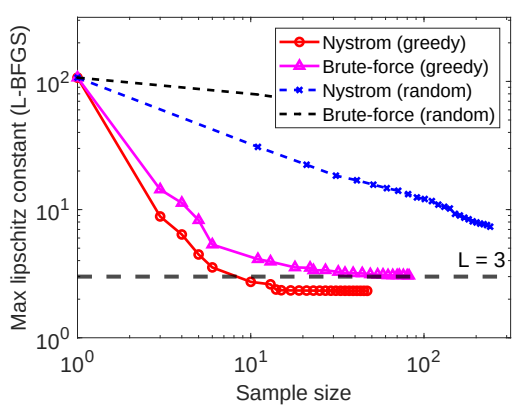


Figure 8: Comparison of efficiency in enforcing Lipschitz constant by various methods

The six different ways to train SVMs with Lipschitz regularisation are summarized in Algorithm 1. Figure 8 plots how fast the regularisation on gradient norm becomes effective when more and more points w are added to the constraint set. We call them "samples" although it is not so random in the greedy method, modulo the random initialization of BFGS within the greedy method. The horizontal axis is the loop index i in Algorithm 1, and the vertical axis is $L^{(i)}$ therein, which is the estimation of the Lipschitz constant of the current solution $f^{(i)}$. We used 400 random examples (200 images of digit 1 and 200 images of digit 0) in the MNIST dataset and set $L = 3$ and RKHS norm $\|f\|_{\mathcal{H}} \leq \infty$ for all algorithms. Inverse kernel is used, hence no results are shown for coordinate-wise Nyström.

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, Nyström with greedy selection eventually fell below the prespecified L, because of the gap in (7).

A.10 Extension to ℓ_∞-norm attacks for our kernel based method

We now extend our kernel based approach to ℓ_∞ norm ball attacks. Since most multiclass losses are 1-Lipschitz continuous with respect to ℓ_∞ norm on $(f^1(x), \ldots, f^{10}(x))$, we will seek

$$\sup_{x \in X} \sup_{\mathbf{u}: \|\mathbf{u}\|_\infty \leq 1} \left\| \left[g^1(x), \ldots, g^{10}(x) \right]^\top \mathbf{u} \right\|_\infty \leq L, \quad \text{where} \quad g^c(x) := \nabla f^c(x).$$

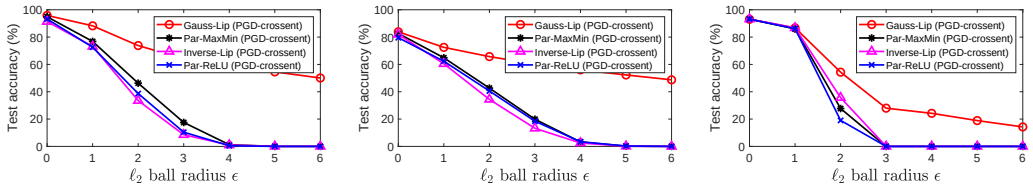
The left-hand side (LHS) can be bounded by

$$\text{LHS} = \sup_{x \in X} \max_{1 \leq c \leq 10} \|g^c(x)\|_1 \leq \max_{1 \leq c \leq 10} \sup_{\|\varphi\|_{\mathcal{H}} \leq 1} \left\| G_c^\top \varphi \right\|_1 .$$

Given the Nyström approximation $\tilde{G}_c$ of G_c, we can enforce the convex constraint of

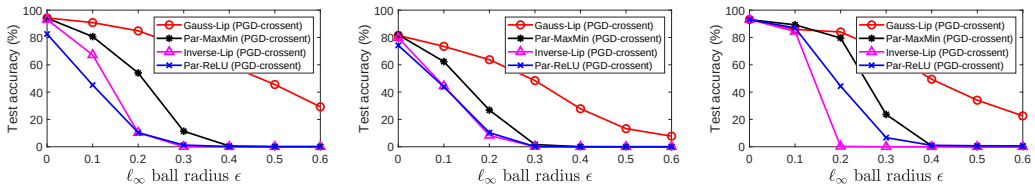
$$\max_{1 \leq c \leq 10} \sup_{\|\mathbf{v}\|_2 \leq 1} \left\| \tilde{G}_c^\top \mathbf{v} \right\|_1 \leq L.$$

A.11 More results on C&W attacks



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

Figure 9: Test accuracy under PGD attacks on cross-entropy approximation with ℓ_2 norm bound



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

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