Supplementary Material

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

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

A. Discussion of Theorems 1 and 2

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

A.1. Results related to Theorem 1

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

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

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

and

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

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

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

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

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

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

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

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

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

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

This shows that

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

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

B. Technical results on distributional robustness

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

with equality when $x = y_0$. Next suppose

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

then

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

as claimed. □

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

where

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

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

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

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

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

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

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

and proves (i).

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

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

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

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

which completes the proof of (ii). □

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

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

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

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

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

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

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

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

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

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

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

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

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

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

which completes the proof. □

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

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

and

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

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

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

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

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

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

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

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

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

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

Then

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

which implies (2). □

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

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

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

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

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

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

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

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

And so we have

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

which implies the claim. □

B.2. Proof of Theorem 2

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

which proves the inequality.

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

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

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

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

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

proving the desired equality. □

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

C.1. Random sampling requires exponential cost

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

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

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

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

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

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

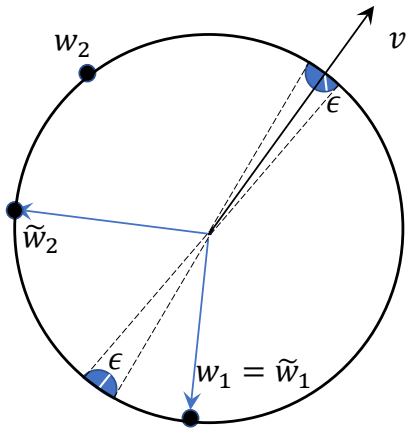


Figure 8: Suppose we use a polynomial kernel with degree 2, and $f(x) = \frac{1}{2}(v^\top x)^2$ for $x \in B$. Then $g(x) = (v^\top x)v$. If we want to test whether $\sup_{x\in B} \|g(x)\|_2 \le 1$ by evaluating $\|g(w)\|_2$ on w that is randomly sampled from B such as w_1 and w_2, we must sample within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

C.2. Spectrum of Kernels

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

This is because, if so, then

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

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

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

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

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

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

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

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

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

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

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

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

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

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

As a result,

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

(ii) We first consider the concentration of the matrix

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

Clearly,

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

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

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

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

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

and

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

And similarly,

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

Therefore the eigenfunctions of the integral operator T_k are

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

□

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

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

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

where H_j is the Hermite polynomial of order j.

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

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

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

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

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

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

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

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

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

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

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

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

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

C.7. Background of eigenvalues of a kernel

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

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

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

Since B is an open set, that means

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

where

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

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

C.8. Bounding the eigenvalues

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

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

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

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

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

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

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

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

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

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

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

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

Now we can bound λ_{N_L} by

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

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

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

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

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

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

D. Algorithm for training a Lipschitz binary SVMs

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

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

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

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

D.1. Detailed algorithm for multiclass classification

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

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

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

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

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

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

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

where the last inequality is due to

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

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

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

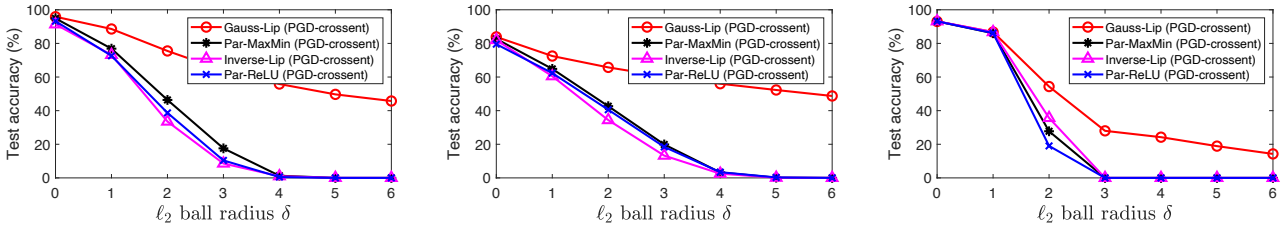
For each c, we alternatively update v and u in (D.1), converging to the optimal v_* and u_*. Finally, the derivative of $\sup_{\|v\|_2 \le 1, \|u\|_\infty \le 1} u^\top \tilde{G}_c^\top v$ with respect to $\boldsymbol{\gamma}^c$ can be calculated from $u_*^\top \tilde{G}_c^\top v_*$, as a function of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

E. More experiments

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

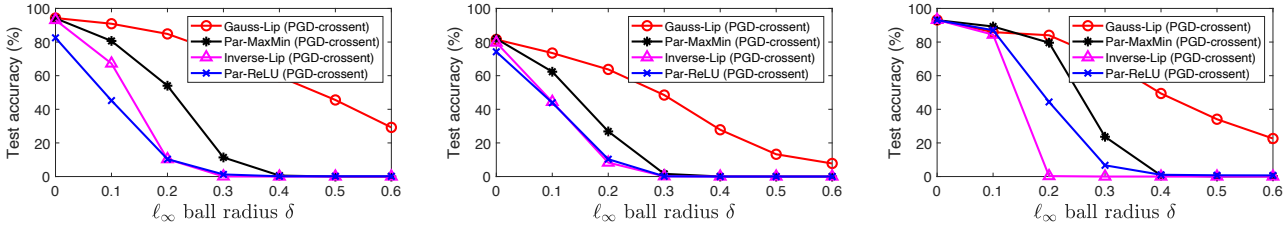
`https://github.com/learndeep2019/DRobust.`

E.1. More results on Cross-Entropy 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 ∞-norm bound

E.2. Visualization of attacks

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

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

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

(a)

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

(b)

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

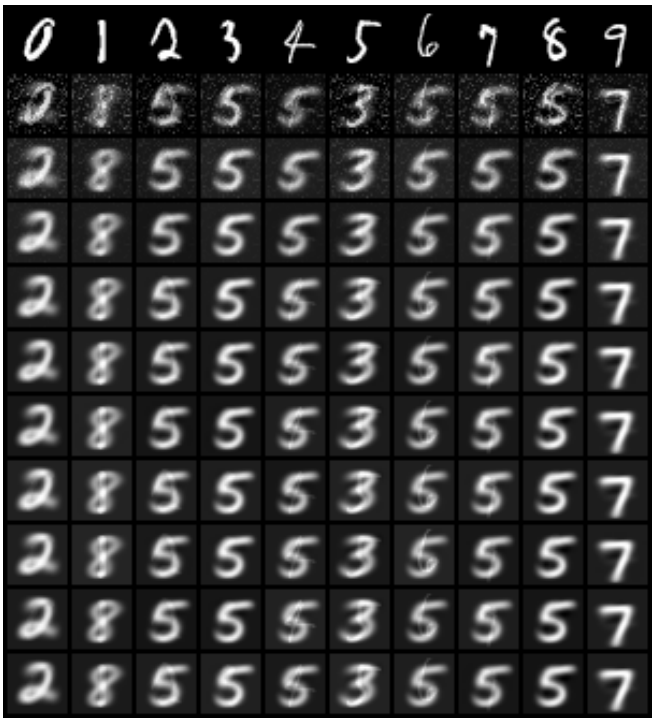


Figure 12: Perturbed images at the end of 100-step PGD attack using the (**untargeted**) C&W approximation. The top row shows 10 random images, one sampled from each class. The 10 rows below show the images after 10, 20, ..., 100 steps of PGD.