Journal of Machine Learning Research 18 (2017) 1-53 Submitted 12/14; Revised 10/16; Published 4/17

# Breaking the Curse of Dimensionality with Convex Neural Networks

**Francis Bach** FRANCIS.BACH@ENS.FR
*INRIA - Sierra Project-team*
*Département d'Informatique de l'Ecole Normale Supérieure (UMR CNRS/ENS/INRIA)*
*2, rue Simone Iff*
*75012 Paris, France*

**Editor:** Yoshua Bengio

## Abstract

We consider neural networks with a single hidden layer and non-decreasing positively homogeneous activation functions like the rectified linear units. By letting the number of hidden units grow unbounded and using classical non-Euclidean regularization tools on the output weights, they lead to a convex optimization problem and we provide a detailed theoretical analysis of their generalization performance, with a study of both the approximation and the estimation errors. We show in particular that they are adaptive to unknown underlying linear structures, such as the dependence on the projection of the input variables onto a low-dimensional subspace. Moreover, when using sparsity-inducing norms on the input weights, we show that high-dimensional non-linear variable selection may be achieved, without any strong assumption regarding the data and with a total number of variables potentially exponential in the number of observations. However, solving this convex optimization problem in infinite dimensions is only possible if the non-convex subproblem of addition of a new unit can be solved efficiently. We provide a simple geometric interpretation for our choice of activation functions and describe simple conditions for convex relaxations of the finite-dimensional non-convex subproblem to achieve the same generalization error bounds, even when constant-factor approximations cannot be found. We were not able to find strong enough convex relaxations to obtain provably polynomial-time algorithms and leave open the existence or non-existence of such tractable algorithms with non-exponential sample complexities.

**Keywords:** Neural networks, non-parametric estimation, convex optimization, convex relaxation.

## 1. Introduction

Supervised learning methods come in a variety of ways. They are typically based on local averaging methods, such as $k$-nearest neighbors, decision trees, or random forests, or on optimization of the empirical risk over a certain function class, such as least-squares regression, logistic regression or support vector machine, with positive definite kernels, with model selection, structured sparsity-inducing regularization, or boosting (see, e.g., Györfi and Krzyzak, 2002; Hastie et al., 2009; Shalev-Shwartz and Ben-David, 2014, and references therein).

©2017 Francis Bach.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v18/14-546.html.

Most methods assume either explicitly or implicitly a certain class of models to learn from. In the non-parametric setting, the learning algorithms may adapt the complexity of the models as the number of observations increases: the sample complexity (i.e., the number of observations) to adapt to any particular problem is typically large. For example, when learning Lipschitz-continuous functions in $\mathbb{R}^d$, at least $n = \Omega(\varepsilon^{-\max\{d,2\}})$ samples are needed to learn a function with excess risk $\varepsilon$ (von Luxburg and Bousquet, 2004, Theorem 15). The exponential dependence on the dimension $d$ is often referred to as the *curse of dimensionality*: without any restrictions, exponentially many observations are needed to obtain optimal generalization performances.

At the other end of the spectrum, parametric methods such as linear supervised learning make strong assumptions regarding the problem and generalization bounds based on estimation errors typically assume that the model is well-specified, and the sample complexity to attain an excess risk of $\varepsilon$ grows as $n = \Omega(d/\varepsilon^2)$, for linear functions in $d$ dimensions and Lipschitz-continuous loss functions (Shalev-Shwartz and Ben-David, 2014, Chapter 9). While the sample complexity is much lower, when the assumptions are not met, the methods underfit and more complex models would provide better generalization performances.

Between these two extremes, there are a variety of models with structural assumptions that are often used in practice. For input data in $x \in \mathbb{R}^d$, prediction functions $f : \mathbb{R}^d \to \mathbb{R}$ may for example be parameterized as:

(a) *Affine functions*: $f(x) = w^\top x + b$, leading to potential severe underfitting, but easy optimization and good (i.e., non exponential) sample complexity.

(b) *Generalized additive models*: $f(x) = \sum_{j=1}^d f_j(x_j)$, which are generalizations of the above by summing functions $f_j : \mathbb{R} \to \mathbb{R}$ which may not be affine (Hastie and Tibshirani, 1990; Ravikumar et al., 2008; Bach, 2008a). This leads to less strong underfitting but cannot model interactions between variables, while the estimation may be done with similar tools than for affine functions (e.g., convex optimization for convex losses).

(c) *Nonparametric ANOVA models*: $f(x) = \sum_{A \in \mathcal{A}} f_A(x_A)$ for a set $\mathcal{A}$ of subsets of $\{1, \dots, d\}$, and non-linear functions $f_A : \mathbb{R}^A \to \mathbb{R}$. The set $\mathcal{A}$ may be either given (Gu, 2013) or learned from data (Lin and Zhang, 2006; Bach, 2008b). Multi-way interactions are explicitly included but a key algorithmic problem is to explore the $2^d - 1$ non-trivial potential subsets.

(d) *Single hidden-layer neural networks*: $f(x) = \sum_{j=1}^k \sigma(w_j^\top x + b_j)$, where $k$ is the number of units in the hidden layer (see, e.g., Rumelhart et al., 1986; Haykin, 1994). The activation function $\sigma$ is here assumed to be fixed. While the learning problem may be cast as a (sub)differentiable optimization problem, techniques based on gradient descent may not find the global optimum. If the number of hidden units is fixed, this is a parametric problem.

(e) *Projection pursuit* (Friedman and Stuetzle, 1981): $f(x) = \sum_{j=1}^k f_j(w_j^\top x)$ where $k$ is the number of projections. This model combines both (b) and (d); the only difference with neural networks is that the non-linear functions $f_j : \mathbb{R} \to \mathbb{R}$ are learned from data. The optimization is often done sequentially and is harder than for neural networks.

(e) *Dependence on a unknown $k$-dimensional subspace*: $f(x) = g(W^\top x)$ with $W \in \mathbb{R}^{d\times k}$, where $g$ is a non-linear function. A variety of algorithms exist for this problem (Li, 1991; Fukumizu et al., 2004; Dalalyan et al., 2008). Note that when the columns of $W$ are assumed to be composed of a single non-zero element, this corresponds to *variable selection* (with at most $k$ selected variables).

In this paper, our main aim is to answer the following question: **Is there a *single* learning method that can deal *efficiently* with all situations above with *provable adaptivity*?** We consider single-hidden-layer neural networks, with non-decreasing homogeneous activation functions such as

$$\sigma(u) = \max\{u, 0\}^\alpha = (u)_+^\alpha,$$

for $\alpha \in \{0, 1, \dots\}$, with a particular focus on $\alpha = 0$ (with the convention that $0^0 = 0$), that is $\sigma(u) = 1_{u>0}$ (a threshold at zero), and $\alpha = 1$, that is, $\sigma(u) = \max\{u, 0\} = (u)_+$, the so-called *rectified linear unit* (Nair and Hinton, 2010; Krizhevsky et al., 2012). We follow the convexification approach of Bengio et al. (2006); Rosset et al. (2007), who consider potentially infinitely many units and let a sparsity-inducing norm choose the number of units automatically. This leads naturally to incremental algorithms such as forward greedy selection approaches, which have a long history for single-hidden-layer neural networks (see, e.g., Breiman, 1993; Lee et al., 1996).

We make the following contributions:

– We provide in Section 2 a review of functional analysis tools used for learning from continuously infinitely many basis functions, by studying carefully the similarities and differences between $L_1$- and $L_2$-penalties on the output weights. For $L_2$-penalties, this corresponds to a positive definite kernel and may be interpreted through random sampling of hidden weights. We also review incremental algorithms (i.e., forward greedy approaches) to learn from these infinite sets of basis functions when using $L_1$-penalties.

– The results are specialized in Section 3 to neural networks with a single hidden layer and activation functions which are positively homogeneous (such as the rectified linear unit). In particular, in Sections 3.2, 3.3 and 3.4, we provide simple geometric interpretations to the non-convex problems of additions of new units, in terms of separating hyperplanes or Hausdorff distance between convex sets. They constitute the core potentially hard computational tasks in our framework of learning from continuously many basis functions.

– In Section 4, we provide a detailed theoretical analysis of the approximation properties of (single hidden layer) convex neural networks with monotonic homogeneous activation functions, with explicit bounds. We relate these new results to the extensive literature on approximation properties of neural networks (see, e.g., Pinkus, 1999, and references therein) in Section 4.7, and show that these neural networks are indeed adaptive to linear structures, by replacing the exponential dependence in dimension by an exponential dependence in the dimension of the subspace of the data can be projected to for good predictions.

| | Functional form | Generalization bound |
|---|---|---|
| No assumption | | $n^{-1/(d+3)} \log n$ |
| Affine function | $w^\top x + b$ | $d^{1/2} \cdot n^{-1/2}$ |
| Generalized additive model | $\sum_{j=1}^k f_j(w_j^\top x)$, $w_j \in \mathbb{R}^d$ | $kd^{1/2} \cdot n^{-1/4} \log n$ |
| Single-layer neural network | $\sum_{j=1}^k \eta_j (w_j^\top x + b_j)_+$ | $kd^{1/2} \cdot n^{-1/2}$ |
| Projection pursuit | $\sum_{j=1}^k f_j(w_j^\top x)$, $w_j \in \mathbb{R}^d$ | $kd^{1/2} \cdot n^{-1/4} \log n$ |
| Dependence on subspace | $f(W^\top x)$ , $W \in \mathbb{R}^{d \times s}$ | $d^{1/2} \cdot n^{-1/(s+3)} \log n$ |

Table 1: Summary of generalization bounds for various models. The bound represents the expected excess risk over the best predictor in the given class. When no assumption is made, the dependence in $n$ goes to zero with an exponent proportional to $1/d$ (which leads to sample complexity exponential in $d$), while making assumptions removes the dependence of $d$ in the exponent.

– In Section 5, we study the generalization properties under a standard supervised learning set-up, and show that these convex neural networks are adaptive to all situations mentioned earlier. These are summarized in Table 1 and constitute the main statistical results of this paper. When using an $\ell_1$-norm on the input weights, we show in Section 5.3 that high-dimensional non-linear variable selection may be achieved, that is, the number of input variables may be much larger than the number of observations, without any strong assumption regarding the data (note that we do not present a polynomial-time algorithm to achieve this).

– We provide in Section 5.5 simple conditions for convex relaxations to achieve the same generalization error bounds, even when constant-factor approximation cannot be found (e.g., because it is NP-hard such as for the threshold activation function and the rectified linear unit). We present in Section 6 convex relaxations based on semi-definite programming, but we were not able to find strong enough convex relaxations (they provide only a provable sample complexity with a polynomial time algorithm which is exponential in the dimension $d$) and leave open the existence or non-existence of polynomial-time algorithms that preserve the non-exponential sample complexity.

## 2. Learning from Continuously Infinitely Many Basis Functions

In this section we present the functional analysis framework underpinning the methods presented in this paper, which learn for a potential continuum of features. While the formulation from Sections 2.1 and 2.2 originates from the early work on the approximation properties of neural networks (Barron, 1993; Kurkova and Sanguineti, 2001; Mhaskar, 2004), the algorithmic parts that we present in Section 2.5 have been studied in a variety of contexts, such as "convex neural networks" (Bengio et al., 2006), or $\ell_1$-norm with infinite dimensional feature spaces (Rosset et al., 2007), with links with conditional gradient algorithms (Dunn and Harshbarger, 1978; Jaggi, 2013) and boosting (Rosset et al., 2004).

In the following sections, note that there will be two different notions of *infinity*: infinitely many inputs $x$ and infinitely many basis functions $x \mapsto \varphi_v(x)$. Moreover, two orthogonal notions of *Lipschitz-continuity* will be tackled in this paper: the one of the prediction functions $f$, and the one of the loss $\ell$ used to measure the fit of these prediction functions.

### 2.1 Variation Norm

We consider an arbitrary measurable input space $\mathcal{X}$ (this will a sphere in $\mathbb{R}^{d+1}$ starting from Section 3), with a set of *basis functions* (a.k.a. *neurons* or *units*) $\varphi_v : \mathcal{X} \to \mathbb{R}$, which are parameterized by $v \in \mathcal{V}$, where $\mathcal{V}$ is a compact topological space (typically a sphere for a certain norm on $\mathbb{R}^d$ starting from Section 3). We assume that for any given $x \in \mathcal{X}$, the functions $v \mapsto \varphi_v(x)$ are continuous. These functions will be the hidden neurons in a single-hidden-layer neural network, and thus $\mathcal{V}$ will be $(d+1)$-dimensional for inputs of dimension $d$ (to represent any affine function). Throughout Section 2, these features will be left unspecified as most of the tools apply more generally.

In order to define our space of functions from $\mathcal{X} \to \mathbb{R}$, we need real-valued Radon measures, which are continuous linear forms on the space of continuous functions from $\mathcal{V}$ to $\mathbb{R}$, equipped with the uniform norm (Rudin, 1987; Evans and Gariepy, 1991). For a continuous function $g : \mathcal{V} \to \mathbb{R}$ and a Radon measure $\mu$, we will use the standard notation $\int_{\mathcal{V}} g(v) d\mu(v)$ to denote the action of the measure $\mu$ on the continuous function $g$. The norm of $\mu$ is usually referred to as its *total variation* (such finite total variation corresponds to having a continuous linear form on the space of continuous functions), and we denote it as $|\mu|(\mathcal{V})$, and is equal to the supremum of $\int_{\mathcal{V}} g(v) d\mu(v)$ over all continuous functions with values in $[-1, 1]$. As seen below, when $\mu$ has a density with respect to a probability measure, this is the $L_1$-norm of the density.

We consider the space $\mathcal{F}_1$ of functions $f$ that can be written as

$$f(x) = \int_{\mathcal{V}} \varphi_v(x) d\mu(v),$$

where $\mu$ is a signed Radon measure on $\mathcal{V}$ with finite total variation $|\mu|(\mathcal{V})$.

When $\mathcal{V}$ is finite, this corresponds to

$$f(x) = \sum_{v \in \mathcal{V}} \mu_v \varphi_v(x),$$

with total variation $\sum_{v \in \mathcal{V}} |\mu_v|$, where the proper formalization for infinite sets $\mathcal{V}$ is done through measure theory.

The infimum of $|\mu|(\mathcal{V})$ over all decompositions of $f$ as $f = \int_{\mathcal{V}} \varphi_v d\mu(v)$, turns out to be a norm $\gamma_1$ on $\mathcal{F}_1$, often called the *variation* norm of $f$ with respect to the set of basis functions (see, e.g., Kurkova and Sanguineti, 2001; Mhaskar, 2004).

Given our assumptions regarding the compactness of $\mathcal{V}$, for any $f \in \mathcal{F}_1$, the infimum defining $\gamma_1(f)$ is in fact attained by a signed measure $\mu$, as a consequence of the compactness of measures for the weak topology (see Evans and Gariepy, 1991, Section 1.9).

In the definition above, if we assume that the signed measure $\mu$ has a density with respect to a fixed *probability* measure $\tau$ with full support on $\mathcal{V}$, that is, $d\mu(v) = p(v)d\tau(v)$, then, the variation norm $\gamma_1(f)$ is also equal to the infimal value of

$$|\mu|(\mathcal{V}) = \int_{\mathcal{V}} |p(v)| d\tau(v),$$

over all integrable functions $p$ such that $f(x) = \int_{\mathcal{V}} p(v)\varphi_v(x)d\tau(v)$. Note however that not all measures have densities, and that the two infimums are the same as all Radon measures are limits of measures with densities. Moreover, the infimum in the definition above is not attained in general (for example when the optimal measure is singular with respect to $d\tau$); however, it often provides a more intuitive definition of the variation norm, and leads to easier comparisons with Hilbert spaces in Section 2.3.

**Finite number of neurons.** If $f : \mathcal{X} \to \mathbb{R}$ is decomposable into $k$ basis functions, that is, $f(x) = \sum_{j=1}^k \eta_j \varphi_{v_j}(x)$, then this corresponds to $\mu = \sum_{j=1}^k \eta_j \delta(v = v_j)$, and the total variation of $\mu$ is equal to the $\ell_1$-norm $\|\eta\|_1$ of $\eta$. Thus the function $f$ has variation norm less than $\|\eta\|_1$ or equal. This is to be contrasted with the number of basis functions, which is the $\ell_0$-pseudo-norm of $\eta$.

### 2.2 Representation from Finitely Many Functions

When minimizing any functional $J$ that depends only on the function values taken at a subset $\hat{\mathcal{X}}$ of values in $\mathcal{X}$, over the ball $\{f \in \mathcal{F}_1,\ \gamma_1(f) \leqslant \delta\}$, then we have a "representer theorem" similar to the reproducing kernel Hilbert space situation, but also with significant differences, which we now present.

The problem is indeed simply equivalent to minimizing a functional on functions restricted to $\mathcal{X}$, that is, to minimizing $J(f_{|\hat{\mathcal{X}}})$ over $f_{|\hat{\mathcal{X}}} \in \mathbb{R}^{\hat{\mathcal{X}}}$, such that $\gamma_{1|\hat{\mathcal{X}}}(f_{|\hat{\mathcal{X}}}) \leqslant \delta$, where

$$\gamma_{1|\hat{\mathcal{X}}}(f_{|\hat{\mathcal{X}}}) = \inf_{\mu} |\mu|(\mathcal{V}) \text{ such that } \forall x \in \hat{\mathcal{X}},\ f_{|\hat{\mathcal{X}}}(x) = \int_{\mathcal{V}} \varphi_v(x) d\mu(v);$$

we can then build a function defined over all $\mathcal{X}$, through the optimal measure $\mu$ above.

Moreover, by Carathéodory's theorem for cones (Rockafellar, 1997), if $\hat{\mathcal{X}}$ is composed of only $n$ elements (e.g., $n$ is the number of observations in machine learning), the optimal function $f_{|\hat{\mathcal{X}}}$ above (and hence $f$) may be decomposed into at most $n$ functions $\varphi_v$, that is, $\mu$ is supported by at most $n$ points in $\mathcal{V}$, among a potential continuum of possibilities.

Note however that the identity of these $n$ functions is not known in advance, and thus there is a significant difference with the representer theorem for positive definite kernels and Hilbert spaces (see, e.g., Shawe-Taylor and Cristianini, 2004), where the set of $n$ functions are known from the knowledge of the points $x \in \hat{\mathcal{X}}$ (i.e., kernel functions evaluated at $x$).

### 2.3 Corresponding Reproducing Kernel Hilbert Space (RKHS)

We have seen above that if the real-valued measures $\mu$ are restricted to have density $p$ with respect to a fixed probability measure $\tau$ with full support on $\mathcal{V}$, that is, $d\mu(v) = p(v)d\tau(v)$, then, the norm $\gamma_1(f)$ is the infimum of the total variation $|\mu|(\mathcal{V}) = \int_{\mathcal{V}} |p(v)| d\tau(v)$, over all decompositions $f(x) = \int_{\mathcal{V}} p(v)\varphi_v(x) d\tau(v)$.

We may also define the infimum of $\int_{\mathcal{V}} |p(v)|^2 d\tau(v)$ over the same decompositions (squared $L_2$-norm instead of $L_1$-norm). It turns out that it defines a squared norm $\gamma_2^2$ and that the function space $\mathcal{F}_2$ of functions with finite norm happens to be a reproducing kernel Hilbert space (RKHS). When $\mathcal{V}$ is finite, then it is well-known (see, e.g., Berlinet and Thomas-Agnan, 2004, Section 4.1) that the infimum of $\sum_{v\in\mathcal{V}} \mu_v^2$ over all vectors $\mu$ such that $f = \sum_{v\in V} \mu_v \varphi_v$ defines a squared RKHS norm with positive definite kernel $k(x,y) = \sum_{v\in V} \varphi_v(x)\varphi_v(y)$.

We show in Appendix A that for any compact set $\mathcal{V}$, we have defined a squared RKHS norm $\gamma_2^2$ with positive definite kernel $k(x,y) = \displaystyle\int_{\mathcal{V}} \varphi_v(x)\varphi_v(y) d\tau(v)$.

**Random sampling.** Note that such kernels are well-adapted to approximations by sampling several basis functions $\varphi_v$ sampled from the probability measure $\tau$ (Neal, 1995; Rahimi and Recht, 2007). Indeed, if we consider $m$ i.i.d. samples $v_1, \dots, v_m$, we may define the approximation $\hat{k}(x,y) = \frac{1}{m}\sum_{i=1}^m \varphi_{v_i}(x)\varphi_{v_i}(y)$, which corresponds to an explicit feature representation. In other words, this corresponds to sampling units $v_i$, using prediction functions of the form $\frac{1}{m}\sum_{i=1}^m \eta_i \varphi_{v_i}(x)$ and then penalizing by the $\ell_2$-norm of $\eta$.

When $m$ tends to infinity, then $\hat{k}(x,y)$ tends to $k(x,y)$ and random sampling provides a way to work efficiently with explicit $m$-dimensional feature spaces. See Rahimi and Recht (2007) for a analysis of the number of units needed for an approximation with error $\varepsilon$, typically of order $1/\varepsilon^2$. See also Bach (2017) for improved results with a better dependence on $\varepsilon$ when making extra assumptions on the eigenvalues of the associated covariance operator.

**Relationship between $\mathcal{F}_1$ and $\mathcal{F}_2$.** The corresponding RKHS norm is always greater than the variation norm (because of Jensen's inequality), and thus the RKHS $\mathcal{F}_2$ is included in $\mathcal{F}_1$. However, as shown in this paper, the two spaces $\mathcal{F}_1$ and $\mathcal{F}_2$ have very different properties; e.g., $\gamma_2$ may be computed easily in several cases, while $\gamma_1$ does not; also, learning with $\mathcal{F}_2$ may either be done by random sampling of sufficiently many weights or using kernel methods, while $\mathcal{F}_1$ requires dedicated convex optimization algorithms with potentially non-polynomial-time steps (see Section 2.5).

Moreover, for any $v \in \mathcal{V}$, $\varphi_v \in \mathcal{F}_1$ with a norm $\gamma_1(\varphi_v) \leqslant 1$, while in general $\varphi_v \notin \mathcal{F}_2$. This is a simple illustration of the fact that $\mathcal{F}_2$ is too small and thus will lead to a lack of adaptivity that will be further studied in Section 5.4 for neural networks with certain activation functions.

### 2.4 Supervised Machine Learning

Given some distribution over the pairs $(x,y) \in \mathcal{X}\times\mathcal{Y}$, a loss function $\ell : \mathcal{Y}\times\mathbb{R} \to \mathbb{R}$, our aim is to find a function $f : \mathcal{X} \to \mathbb{R}$ such that the functional $J(f) = \mathbb{E}[\ell(y, f(x))]$ is small, given some i.i.d. observations $(x_i, y_i)$, $i = 1, \dots, n$. We consider the empirical risk minimization framework over a space of functions $\mathcal{F}$, equipped with a norm $\gamma$ (in our situation, $\mathcal{F}_1$ and $\mathcal{F}_2$, equipped with $\gamma_1$ or $\gamma_2$). The empirical risk $\hat{J}(f) = \frac{1}{n}\sum_{i=1}^n \ell(y_i, f(x_i))$, is minimized either (a) by constraining $f$ to be in the ball $\mathcal{F}^\delta = \{f \in \mathcal{F},\ \gamma(f) \leqslant \delta\}$ or (b) regularizing the empirical risk by $\lambda\gamma(f)$. Since this paper has a more theoretical nature, we focus on constraining, noting that in practice, penalizing is often more robust (see, e.g., Harchaoui

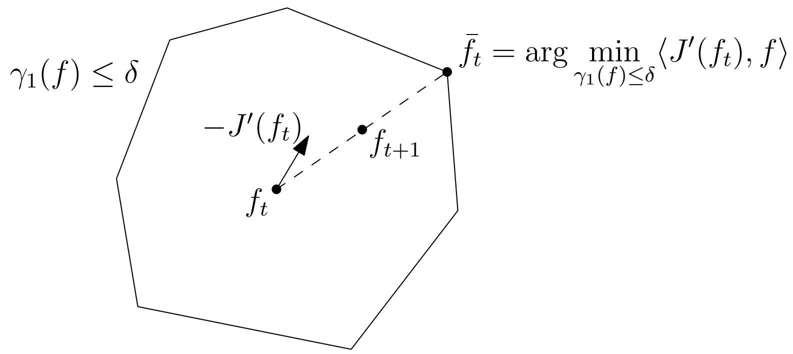
$\gamma_1(f) \leq \delta$
$\bar{f}_t = \arg\min_{\gamma_1(f)\leq\delta} \langle J'(f_t), f\rangle$
$-J'(f_t)$
$f_{t+1}$
$f_t$

Figure 1: Conditional gradient algorithm for minimizing a smooth functional $J$ on $\mathcal{F}_1^\delta = \{f \in \mathcal{F}_1,\ \gamma_1(f) \leqslant \delta\}$: going from $f_t$ to $f_{t+1}$; see text for details.

et al., 2013) and leaving its analysis in terms of learning rates for future work. Since the functional $\hat{J}$ depends only on function values taken at finitely many points, the results from Section 2.2 apply and we expect the solution $f$ to be spanned by only $n$ functions $\varphi_{v_1}, \dots, \varphi_{v_n}$ (but we ignore in advance which ones among all $\varphi_v$, $v \in \mathcal{V}$, and the algorithms in Section 2.5 will provide approximate such representations with potentially less or more than $n$ functions).

**Approximation error vs. estimation error.** We consider an $\varepsilon$-approximate minimizer of $\hat{J}(f) = \frac{1}{n}\sum_{i=1}^n \ell(y_i, f(x_i))$ on the convex set $\mathcal{F}^\delta$, that is a certain $\hat{f} \in \mathcal{F}^\delta$ such that $\hat{J}(\hat{f}) \leqslant \varepsilon + \inf_{f\in\mathcal{F}^\delta} \hat{J}(f)$. We thus have, using standard arguments (see, e.g., Shalev-Shwartz and Ben-David, 2014):

$$J(\hat{f}) - \inf_{f\in\mathcal{F}} J(f) \leqslant \Big[ \inf_{f\in\mathcal{F}^\delta} J(f) - \inf_{f\in\mathcal{F}} J(f) \Big] + 2 \sup_{f\in\mathcal{F}^\delta} |\hat{J}(f) - J(f)| + \varepsilon,$$

that is, the excess risk $J(\hat{f}) - \inf_{f\in\mathcal{F}} J(f)$ is upper-bounded by a sum of an *approximation error* $\inf_{f\in\mathcal{F}^\delta} J(f) - \inf_{f\in\mathcal{F}} J(f)$, an *estimation error* $2\sup_{f\in\mathcal{F}^\delta} |\hat{J}(f) - J(f)|$ and an *optimization error* $\varepsilon$ (see also Bottou and Bousquet, 2008). In this paper, we will deal with all three errors, starting from the optimization error which we now consider for the space $\mathcal{F}_1$ and its variation norm.

### 2.5 Incremental Conditional Gradient Algorithms

In this section, we review algorithms to minimize a smooth functional $J : L_2(d\rho) \to \mathbb{R}$, where $\rho$ is a probability measure on $\mathcal{X}$. This may typically be the expected risk or the empirical risk above. When minimizing $J(f)$ with respect to $f \in \mathcal{F}_1$ such that $\gamma_1(f) \leqslant \delta$, we need algorithms that can efficiently optimize a convex function over an infinite-dimensional space of functions. Conditional gradient algorithms allow to incrementally build a set of elements of $\mathcal{F}_1^\delta = \{f \in \mathcal{F}_1,\ \gamma_1(f) \leqslant \delta\}$; see, e.g., Frank and Wolfe (1956); Dem'yanov and Rubinov (1967); Dudik et al. (2012); Harchaoui et al. (2013); Jaggi (2013); Bach (2015).

**Conditional gradient algorithm.** We assume the functional $J$ is convex and $L$-smooth, that is for all $h \in L_2(d\rho)$, there exists a gradient $J'(h) \in L_2(d\rho)$ such that for all $f \in L_2(d\rho)$,

$$0 \leqslant J(f) - J(h) - \langle f - h, J'(h) \rangle_{L_2(d\rho)} \leqslant \frac{L}{2} \| f - h \|^2_{L_2(d\rho)}.$$

When $\mathcal{X}$ is finite, this corresponds to the regular notion of smoothness from convex optimization (Nesterov, 2004).

The conditional gradient algorithm (a.k.a. Frank-Wolfe algorithm) is an iterative algorithm, starting from any function $f_0 \in \mathcal{F}_1^\delta$ and with the following recursion, for $t \geqslant 0$:

$$\begin{aligned} \bar{f}_t &\in \arg\min_{f \in \mathcal{F}_1^\delta} \langle f, J'(f_t) \rangle_{L_2(d\rho)} \\ f_{t+1} &= (1-\rho_t) f_t + \rho_t \bar{f}_t. \end{aligned}$$

See an illustration in Figure 1. We may choose either $\rho_t = \frac{2}{t+1}$ or perform a line search for $\rho_t \in [0,1]$. For all of these strategies, the $t$-th iterate is a convex combination of the functions $\bar{f}_0, \dots, \bar{f}_{t-1}$, and is thus an element of $\mathcal{F}_1^\delta$. It is known that for these two strategies for $\rho_t$, we have the following convergence rate (see, e.g. Jaggi, 2013):

$$J(f_t) - \inf_{f \in \mathcal{F}_1^\delta} J(f) \leqslant \frac{2L}{t+1} \sup_{f,g \in \mathcal{F}_1^\delta} \| f - g \|^2_{L_2(d\rho)}.$$

When, $r^2 = \sup_{v \in \mathcal{V}} \|\varphi_v\|^2_{L_2(d\rho)}$ is finite, we have $\|f\|^2_{L_2(d\rho)} \leqslant r^2 \gamma_1(f)^2$ and thus we get a convergence rate of $\frac{2Lr^2\delta^2}{t+1}$.

Moreover, the basic Frank-Wolfe (FW) algorithm may be extended to handle the regularized problem as well (Harchaoui et al., 2013; Bach, 2013; Zhang et al., 2012), with similar convergence rates in $O(1/t)$. Also, the second step in the algorithm, where the function $f_{t+1}$ is built in the segment between $f_t$ and the newly found extreme function, may be replaced by the optimization of $J$ over the convex hull of all functions $\bar{f}_0, \dots, \bar{f}_t$, a variant which is often referred to as *fully corrective*. Moreover, in our context where $\mathcal{V}$ is a space where local search techniques may be considered, there is also the possibility of "fine-tuning" the vectors $v$ as well (Bengio et al., 2006), that is, we may optimize the function $(v_1, \dots, v_t, \alpha_1, \dots, \alpha_t) \mapsto J(\sum_{i=1}^t \alpha_i \varphi_{v_i})$, through local search techniques, starting from the weights $(\alpha_i)$ and points $(v_i)$ obtained from the conditional gradient algorithm.

**Adding a new basis function.** The conditional gradient algorithm presented above relies on solving at each iteration the "Frank-Wolfe step":

$$\max_{\gamma(f) \leqslant \delta} \langle f, g \rangle_{L_2(d\rho)}.$$

for $g = -J'(f_t) \in L_2(d\rho)$. For the norm $\gamma_1$ defined through an $L_1$-norm, we have for $f = \int_{\mathcal{V}} \varphi_v d\mu(v)$ such that $\gamma_1(f) = |\mu|(\mathcal{V})$:

$$\begin{aligned} \langle f, g \rangle_{L_2(d\rho)} &= \int_{\mathcal{X}} f(x) g(x) d\rho(x) = \int_{\mathcal{X}} \bigg( \int_{\mathcal{V}} \varphi_v(x) d\mu(v) \bigg) g(x) d\rho(x) \\ &= \int_{\mathcal{V}} \bigg( \int_{\mathcal{X}} \varphi_v(x) g(x) d\rho(x) \bigg) d\mu(v) \\ &\leqslant \gamma_1(f) \cdot \max_{v \in \mathcal{V}} \bigg| \int_{\mathcal{X}} \varphi_v(x) g(x) d\rho(x) \bigg|, \end{aligned}$$

with equality if and only if $\mu = \mu_+ - \mu_-$ with $\mu_+$ and $\mu_-$ two non-negative measures, with $\mu_+$ (resp. $\mu_-$) supported in the set of maximizers $v$ of $|\int_{\mathcal{X}} \varphi_v(x) g(x) d\rho(x)|$ where the value is positive (resp. negative).

This implies that:

$$\max_{\gamma_1(f) \leqslant \delta} \langle f, g \rangle_{L_2(d\rho)} = \delta \max_{v \in \mathcal{V}} \bigg| \int_{\mathcal{X}} \varphi_v(x) g(x) d\rho(x) \bigg|, \tag{1}$$

with the maximizers $f$ of the first optimization problem above (left-hand side) obtained as $\delta$ times convex combinations of $\varphi_v$ and $-\varphi_v$ for maximizers $v$ of the second problem (right-hand side).

A common difficulty in practice is the hardness of the Frank-Wolfe step, that is, the optimization problem above over $\mathcal{V}$ may be difficult to solve. See Section 3.2, 3.3 and 3.4 for neural networks, where this optimization is usually difficult.

**Finitely many observations.** When $\mathcal{X}$ is finite (or when using the result from Section 2.2), the Frank-Wolfe step in Eq. (1) becomes equivalent to, for some vector $g \in \mathbb{R}^n$:

$$\sup_{\gamma_1(f) \leqslant \delta} \frac{1}{n} \sum_{i=1}^n g_i f(x_i) \ = \ \delta \max_{v \in \mathcal{V}} \bigg| \frac{1}{n} \sum_{i=1}^n g_i \varphi_v(x_i) \bigg|, \tag{2}$$

where the set of solutions of the first problem is in the convex hull of the solutions of the second problem.

**Non-smooth loss functions.** In this paper, in our theoretical results, we consider non-smooth loss functions for which conditional gradient algorithms do not converge in general. One possibility is to smooth the loss function, as done by Nesterov (2005): an approximation error of $\varepsilon$ may be obtained with a smoothness constant proportional to $1/\varepsilon$. By choosing $\varepsilon$ as $1/\sqrt{t}$, we obtain a convergence rate of $O(1/\sqrt{t})$ after $t$ iterations. See also Lan (2013).

**Approximate oracles.** The conditional gradient algorithm may deal with approximate oracles; however, what we need in this paper is not the additive errors situations considered by Jaggi (2013), but multiplicative ones on the computation of the dual norm (similar to ones derived by Bach (2013) for the regularized problem).

Indeed, in our context, we minimize a function $J(f)$ on $f \in L_2(d\rho)$ over a norm ball $\{\gamma_1(f) \leqslant \delta\}$. A multiplicative approximate oracle outputs for any $g \in L_2(d\rho)$, a vector $\hat{f} \in L_2(d\rho)$ such that $\gamma_1(\hat{f}) = 1$, and

$$\langle \hat{f}, g \rangle \leqslant \max_{\gamma_1(f) \leqslant 1} \langle f, g \rangle \leqslant \kappa \, \langle \hat{f}, g \rangle,$$

for a fixed $\kappa \geqslant 1$. In Appendix B, we propose a modification of the conditional gradient algorithm that converges to a certain $h \in L_2(d\rho)$ such that $\gamma_1(h) \leqslant \delta$ and for which $\inf_{\gamma_1(f) \leqslant \delta} J(f) \leqslant J(h) \leqslant \inf_{\gamma_1(f) \leqslant \delta/\kappa} J(f)$.

Such approximate oracles are not available in general, because they require uniform bounds over all possible values of $g \in L_2(d\rho)$. In Section 5.5, we show that a weaker form of oracle is sufficient to preserve our generalization bounds from Section 5.

**Approximation of any function by a finite number of basis functions.** The Frank-Wolfe algorithm may be applied in the function space $\mathcal{F}_1$ with $J(f) = \frac{1}{2}\mathbb{E}[(f(x)-g(x))^2]$, we get a function $f_t$, supported by $t$ basis functions such that $\mathbb{E}[(f_t(x) - g(x))^2] = O(\gamma(g)^2/t)$. Hence, any function in $\mathcal{F}_1$ may be approximated with averaged error $\varepsilon$ with $t = O([\gamma(g)/\varepsilon]^2)$ units. Note that the conditional gradient algorithm is one among many ways to obtain such approximation with $\varepsilon^{-2}$ units (Barron, 1993; Kurkova and Sanguineti, 2001; Mhaskar, 2004). See Section 4.1 for a (slightly) better dependence on $\varepsilon$ for convex neural networks.

## 3. Neural Networks with Non-decreasing Positively Homogeneous Activation Functions

In this paper, we focus on a specific family of basis functions, that is, of the form

$$x \mapsto \sigma(w^\top x + b),$$

for specific activation functions $\sigma$. We assume that $\sigma$ is non-decreasing and positively homogeneous of some integer degree, i.e., it is equal to $\sigma(u) = (u)_+^\alpha$, for some $\alpha \in \{0, 1, \ldots\}$. We focus on these functions for several reasons:

- Since they are not polynomials, linear combinations of these functions can approximate any measurable function (Leshno et al., 1993).
- By homogeneity, they are invariant by a change of scale of the data; indeed, if all observations $x$ are multiplied by a constant, we may simply change the measure $\mu$ defining the expansion of $f$ by the appropriate constant to obtain exactly the same function. This allows us to study functions defined on the unit-sphere.
- The special case $\alpha = 1$, often referred to as the *rectified linear unit*, has seen considerable recent empirical success (Nair and Hinton, 2010; Krizhevsky et al., 2012), while the case $\alpha = 0$ (hard thresholds) has some historical importance (Rosenblatt, 1958).

The goal of this section is to specialize the results from Section 2 to this particular case and show that the "Frank-Wolfe" steps have simple geometric interpretations.

We first show that the positive homogeneity of the activation functions allows to transfer the problem to a unit sphere.

**Boundedness assumptions.** For the theoretical analysis, we assume that our data inputs $x \in \mathbb{R}^d$ are almost surely bounded by $R$ in $\ell_q$-norm, for some $q \in [2, \infty]$ (typically $q = 2$ and $q = \infty$). We then build the augmented variable $z \in \mathbb{R}^{d+1}$ as $z = (x^\top, R)^\top \in \mathbb{R}^{d+1}$ by appending the constant $R$ to $x \in \mathbb{R}^d$. We therefore have $\|z\|_q \leqslant \sqrt{2}R$. By defining the vector $v = (w^\top, b/R)^\top \in \mathbb{R}^{d+1}$, we have:

$$\varphi_v(x) = \sigma(w^\top x + b) = \sigma(v^\top z) = (v^\top z)_+^\alpha,$$

which now becomes a function of $z \in \mathbb{R}^{d+1}$.

Without loss of generality (and by homogeneity of $\sigma$), we may assume that the $\ell_p$-norm of each vector $v$ is equal to $1/R$, that is $\mathcal{V}$ will be the $(1/R)$-sphere for the $\ell_p$-norm, where $1/p + 1/q = 1$ (and thus $p \in [1, 2]$, with corresponding typical values $p = 2$ and $p = 1$).

This implies by Hölder's inequality that $\varphi_v(x)^2 \leqslant 2^\alpha$. Moreover this leads to functions in $\mathcal{F}_1$ that are bounded everywhere, that is, $\forall f \in \mathcal{F}_1,\ f(x)^2 \leqslant 2^\alpha \gamma_1(f)^2$. Note that the functions in $\mathcal{F}_1$ are also Lipschitz-continuous for $\alpha \geqslant 1$.

Since all $\ell_p$-norms (for $p \in [1,2]$) are equivalent to each other with constants of at most $\sqrt{d}$ with respect to the $\ell_2$-norm, all the spaces $\mathcal{F}_1$ defined above are equal, but the norms $\gamma_1$ are of course different and they differ by a constant of at most $d^{\alpha/2}$—this can be seen by computing the dual norms like in Eq. (2) or Eq. (1).

**Homogeneous reformulation.** In our study of approximation properties, it will be useful to consider the the space of function $\mathcal{G}_1$ defined for $z$ in the unit sphere $\mathbb{S}^d \subset \mathbb{R}^{d+1}$ of the Euclidean norm, such that $g(z) = \int_{\mathbb{S}^d} \sigma(v^\top z) d\mu(v)$, with the norm $\gamma_1(g)$ defined as the infimum of $|\mu|(\mathbb{S}^d)$ over all decompositions of $g$. Note the slight overloading of notations for $\gamma_1$ (for norms in $\mathcal{G}_1$ and $\mathcal{F}_1$) which should not cause any confusion.

In order to prove the approximation properties (with unspecified constants depending only on $d$), we may assume that $p = 2$, since the norms $\|\cdot\|_p$ for $p \in [1,\infty]$ are equivalent to $\|\cdot\|_2$ with a constant that grows at most as $d^{\alpha/2}$ with respect to the $\ell_2$-norm. We thus focus on the $\ell_2$-norm in all proofs in Section 4.

We may go from $\mathcal{G}_1$ (a space of real-valued functions defined on the unit $\ell_2$-sphere in $d+1$ dimensions) to the space $\mathcal{F}_1$ (a space of real-valued functions defined on the ball of radius $R$ for the $\ell_2$-norm) as follows (this corresponds to sending a ball in $\mathbb{R}^d$ into a spherical cap in dimension $d+1$, as illustrated in Figure 2).

– Given $g \in \mathcal{G}_1$, we define $f \in \mathcal{F}_1$, with $f(x) = \Big(\frac{\|x\|_2^2}{R^2} + 1\Big)^{\alpha/2} g\bigg(\frac{1}{\sqrt{\|x\|_2^2 + R^2}} \binom{x}{R}\bigg)$. If $g$ may be represented as $\int_{\mathbb{S}^d} \sigma(v^\top z) d\mu(v)$, then the function $f$ that we have defined may be represented as

$$\begin{aligned} f(x) &= \Big(\frac{\|x\|_2^2}{R^2} + 1\Big)^{\alpha/2} \int_{\mathbb{S}^d} \bigg(v^\top \frac{1}{\sqrt{\|x\|_2^2 + R^2}} \binom{x}{R}\bigg)_+^\alpha d\mu(v) \\ &= \int_{\mathbb{S}^d} \bigg(v^\top \binom{x/R}{1}\bigg)_+^\alpha d\mu(v) = \int_{\mathbb{S}^d} \sigma(w^\top x + b) d\mu(Rw, b), \end{aligned}$$

that is $\gamma_1(f) \leqslant \gamma_1(g)$, because we have assumed that $(w^\top, b/R)^\top$ is on the $(1/R)$-sphere.

– Conversely, given $f \in \mathcal{F}_1$, for $z = (t^\top, a)^\top \in \mathbb{S}^d$, we define $g(z) = g(t,a) = f(\frac{Rt}{a}) a^\alpha$, which we define as such on the set of $z = (t^\top, a)^\top \in \mathbb{R}^d \times \mathbb{R}$ (of unit norm) such that $a \geqslant \frac{1}{\sqrt{2}}$. Since we always assume $\|x\|_2 \leqslant R$, we have $\sqrt{\|x\|_2^2 + R^2} \leqslant \sqrt{2}R$, and the value of $g(z,a)$ for $a \geqslant \frac{1}{\sqrt{2}}$ is enough to recover $f$ from the formula above.

On that portion $\{a \geqslant 1/\sqrt{2}\}$ of the sphere $\mathbb{S}^d$, this function exactly inherits the differentiability properties of $f$. That is, (a) if $f$ is bounded by 1 and $f$ is $(1/R)$-Lipschitz-continuous, then $g$ is Lipschitz-continuous with a constant that only depends on $d$ and $\alpha$ and (b), if all derivatives of order less than $k$ are bounded by $R^{-k}$, then all derivatives of the same order of $g$ are bounded by a constant that only depends on $d$ and $\alpha$. Precise notions of differentiability may be defined on the sphere, using

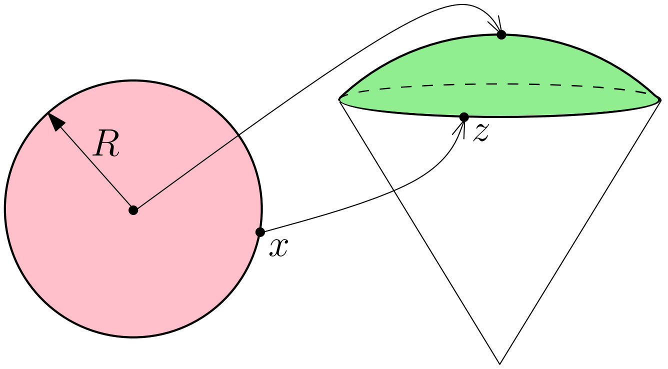

Figure 2: Sending a ball to a spherical cap.

the manifold structure (see, e.g., Absil et al., 2009) or through polar coordinates (see, e.g., Atkinson and Han, 2012, Chapter 3). See these references for more details.

The only remaining important aspect is to define $g$ on the entire sphere, so that (a) its regularity constants are controlled by a constant times the ones on the portion of the sphere where it is already defined, (b) $g$ is either even or odd (this will be important in Section 4). Ensuring that the regularity conditions can be met is classical when extending to the full sphere (see, e.g., Whitney, 1934). Ensuring that the function may be chosen as odd or even may be obtained by multiplying the function $g$ by an infinitely differentiable function which is equal to one for $a \geqslant 1/\sqrt{2}$ and zero for $a \leqslant 0$, and extending by $-g$ or $g$ on the hemi-sphere $a < 0$.

In summary, we may consider in Section 4 functions defined on the sphere, which are much easier to analyze. In the rest of the section, we specialize some of the general concepts reviewed in Section 2 to our neural network setting with specific activation functions, namely, in terms of corresponding kernel functions and geometric reformulations of the Frank-Wolfe steps.

### 3.1 Corresponding Positive-definite Kernels

In this section, we consider the $\ell_2$-norm on the input weight vectors $w$ (that is $p = 2$). We may compute for $x, x' \in \mathbb{R}^d$ the kernels defined in Section 2.3:

$$k_\alpha(x, x') = \mathbb{E}[(w^\top x + b)_+^\alpha (w^\top x' + b)_+^\alpha],$$

for $(Rw, b)$ distributed uniformly on the unit $\ell_2$-sphere $\mathbb{S}^d$, and $x, x' \in \mathbb{R}^{d+1}$. Given the angle $\varphi \in [0, \pi]$ defined through $\dfrac{x^\top x'}{R^2} + 1 = (\cos\varphi)\sqrt{\dfrac{\|x\|_2^2}{R^2} + 1}\sqrt{\dfrac{\|x'\|_2^2}{R^2} + 1}$, we have explicit expressions (Le Roux and Bengio, 2007; Cho and Saul, 2009):

$$\begin{aligned} k_0(z, z') &= \frac{1}{2\pi}(\pi - \varphi) \\ k_1(z, z') &= \frac{\sqrt{\frac{\|x\|_2^2}{R^2} + 1}\sqrt{\frac{\|x'\|_2^2}{R^2} + 1}}{2(d+1)\pi}((\pi - \varphi)\cos\varphi + \sin\varphi) \end{aligned}$$

$$k_2(z,z') \;=\; \frac{\Big(\frac{\|x\|_2^2}{R^2}+1\Big)\Big(\frac{\|x'\|_2^2}{R^2}+1\Big)}{2\pi[(d+1)^2+2(d+1)]}(3\sin\varphi\cos\varphi+(\pi-\varphi)(1+2\cos^2\varphi)).$$

There are key differences and similarities between the RKHS $\mathcal{F}_2$ and our space of functions $\mathcal{F}_1$. The RKHS is smaller than $\mathcal{F}_1$ (i.e., the norm in the RKHS is larger than the norm in $\mathcal{F}_1$); this implies that approximation properties of the RKHS are transferred to $\mathcal{F}_1$. In fact, our proofs rely on this fact.

However, the RKHS norm does not lead to any adaptivity, while the function space $\mathcal{F}_1$ does (see more details in Section 5). This may come as a paradox: both the RKHS $\mathcal{F}_2$ and $\mathcal{F}_1$ have similar properties, but one is adaptive while the other one is not. A key intuitive difference is as follows: given a function $f$ expressed as $f(x) = \int_{\mathcal{V}} \varphi_v(x)p(v)d\tau(v)$, then $\gamma_1(f) = \int_{\mathcal{V}} |p(v)|d\tau(v)$, while the squared RKHS norm is $\gamma_2(f)^2 = \int_{\mathcal{V}} |p(v)|^2 d\tau(v)$. For the $L_1$-norm, the measure $p(v)d\tau(v)$ may tend to a singular distribution with a bounded norm, while this is not true for the $L_2$-norm. For example, the function $(w^\top x+b)_+^\alpha$ is in $\mathcal{F}_1$, while it is not in $\mathcal{F}_2$ in general.

### 3.2 Incremental Optimization Problem for $\alpha = 0$

We consider the problem in Eq. (2) for the special case $\alpha = 0$. For $z_1,\dots,z_n \in \mathbb{R}^{d+1}$ and a vector $y \in \mathbb{R}^n$, the goal is to solve (as well as the corresponding problem with $y$ replaced by $-y$):

$$\max_{v\in\mathbb{R}^{d+1}} \sum_{i=1}^n y_i 1_{v^\top z_i>0} = \max_{v\in\mathbb{R}^{d+1}} \sum_{i\in I_+} |y_i| 1_{v^\top z_i>0} - \sum_{i\in I_-} |y_i| 1_{v^\top z_i>0},$$

where $I_+ = \{i, y_i \geqslant 0\}$ and $I_- = \{i, y_i < 0\}$. As outlined by Bengio et al. (2006), this is equivalent to finding an hyperplane parameterized by $v$ that minimizes a weighted misclassification rate (when doing linear classification). Note that the norm of $v$ has no effect.

**NP-hardness.** This problem is NP-hard in general. Indeed, if we assume that all $y_i$ are equal to $-1$ or 1 and with $\sum_{i=1}^n y_i = 0$, then we have a balanced binary classification problem (we need to assume $n$ even). The quantity $\sum_{i=1}^n y_i 1_{v^\top z_i>0}$ is then $\frac{n}{2}(1-2e)$ where $e$ is the corresponding classification error for a problem of classifying at positive (resp. negative) the examples in $I_+$ (resp. $I_-$) by thresholding the linear classifier $v^\top z$. Guruswami and Raghavendra (2009) showed that for all $(\varepsilon,\delta)$, it is NP-hard to distinguish between instances (i.e., configurations of points $x_i$), where a halfspace with classification error at most $\varepsilon$ exists, and instances where all half-spaces have an error of at least $1/2-\delta$. Thus, it is NP-hard to distinguish between instances where there exists $v \in \mathbb{R}^{d+1}$ such that $\sum_{i=1}^n y_i 1_{v^\top z_i>0} \geqslant \frac{n}{2}(1-2\varepsilon)$ and instances where for all $v \in \mathbb{R}^{d+1}$, $\sum_{i=1}^n y_i 1_{v^\top z_i>0} \leqslant n\delta$. Thus, it is NP-hard to distinguish instances where $\max_{v\in\mathbb{R}^{d+1}} \sum_{i=1}^n y_i 1_{v^\top z_i>0} \geqslant \frac{n}{2}(1-2\varepsilon)$ and ones where it is less than $\frac{n}{2}\delta$. Since this is valid for all $\delta$ and $\varepsilon$, this rules out a constant-factor approximation.

**Convex relaxation.** Given linear binary classification problems, there are several algorithms to approximately find a good half-space. These are based on using convex surrogates (such as the hinge loss or the logistic loss). Although some theoretical results do exist regarding the classification performance of estimators obtained from convex surrogates (Bartlett et al., 2006), they do not apply in the context of linear classification.

### 3.3 Incremental Optimization Problem for $\alpha = 1$

We consider the problem in Eq. (2) for the special case $\alpha = 1$. For $z_1, \dots, z_n \in \mathbb{R}^{d+1}$ and a vector $y \in \mathbb{R}^n$, the goal is to solve (as well as the corresponding problem with $y$ replaced by $-y$):

$$\max_{\|v\|_p \leqslant 1} \sum_{i=1}^n y_i (v^\top z_i)_+ = \max_{\|v\|_p \leqslant 1} \ \sum_{i \in I_+} (v^\top |y_i| z_i)_+ - \sum_{i \in I_-} (v^\top |y_i| z_i)_+,$$

where $I_+ = \{i, y_i \geqslant 0\}$ and $I_- = \{i, y_i < 0\}$. We have, with $t_i = |y_i| z_i \in \mathbb{R}^{d+1}$, using convex duality:

$$\begin{aligned}
\max_{\|v\|_p \leqslant 1} \sum_{i=1}^n y_i (v^\top z_i)_+ &= \max_{\|v\|_p \leqslant 1} \sum_{i \in I_+} (v^\top t_i)_+ - \sum_{i \in I_-} (v^\top t_i)_+ \\
&= \max_{\|v\|_p \leqslant 1} \sum_{i \in I_+} \max_{b_i \in [0,1]} b_i v^\top t_i - \sum_{i \in I_-} \max_{b_i \in [0,1]} b_i v^\top t_i \\
&= \max_{b_+ \in [0,1]^{I_+}} \max_{\|v\|_p \leqslant 1} \min_{b_- \in [0,1]^{I_-}} v^\top [T_+^\top b_+ - T_-^\top b_-] \\
&= \max_{b_+ \in [0,1]^{I_+}} \min_{b_- \in [0,1]^{I_-}} \max_{\|v\|_p \leqslant 1} v^\top [T_+^\top b_+ - T_-^\top b_-] \text{ by Fenchel duality,} \\
&= \max_{b_+ \in [0,1]^{I_+}} \min_{b_- \in [0,1]^{I_-}} \|T_+^\top b_+ - T_-^\top b_-\|_q,
\end{aligned}$$

where $T_+ \in \mathbb{R}^{n_+ \times d}$ has rows $t_i$, $i \in I_+$ and $T_- \in \mathbb{R}^{n_- \times d}$ has rows $t_i$, $i \in I_-$, with $v \in \arg\max_{\|v\|_p \leqslant 1} v^\top (T_+^\top b_+ - T_-^\top b_-)$. The problem thus becomes

$$\max_{b_+ \in [0,1]^{n_+}} \min_{b_- \in [0,1]^{n_-}} \|T_+^\top b_+ - T_-^\top b_-\|_q.$$

For the problem of maximizing $|\sum_{i=1}^n y_i (v^\top z_i)_+|$, then this corresponds to

$$\max \Big\{ \max_{b_+ \in [0,1]^{n_+}} \min_{b_- \in [0,1]^{n_-}} \|T_+^\top b_+ - T_-^\top b_-\|_q, \max_{b_- \in [0,1]^{n_-}} \min_{b_+ \in [0,1]^{n_+}} \|T_+^\top b_+ - T_-^\top b_-\|_q \Big\}.$$

This is exactly the Hausdorff distance between the two convex sets $\{T_+^\top b_+, \ b_+ \in [0,1]^{n_+}\}$ and $\{T_-^\top b_-, \ b_- \in [0,1]^{n_-}\}$ (referred to as zonotopes, see below).

Given the pair $(b_+, b_-)$ achieving the Hausdorff distance, then we may compute the optimal $v$ as $v = \arg\max_{\|v\|_p \leqslant 1} v^\top (T_+^\top b_+ - T_-^\top b_-)$. Note this has not changed the problem at all, since it is equivalent. It is still NP-hard in general (König, 2014). But we now have a geometric interpretation with potential approximation algorithms. See below and Section 6.

**Zonotopes.** A *zonotope* $A$ is the Minkowski sum of a finite number of segments from the origin, that is, of the form

$$A = [0, t_1] + \cdots + [0, t_r] = \Big\{ \sum_{i=1}^r b_i t_i, \ b \in [0,1]^r \Big\},$$

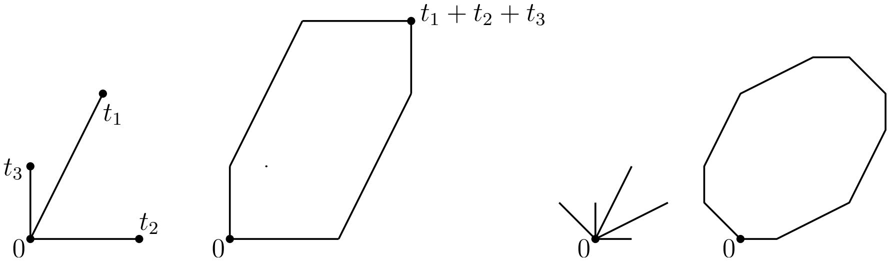

Figure 3: Two zonotopes in two dimensions: (left) vectors, and (right) their Minkowski sum (represented as a polygone).

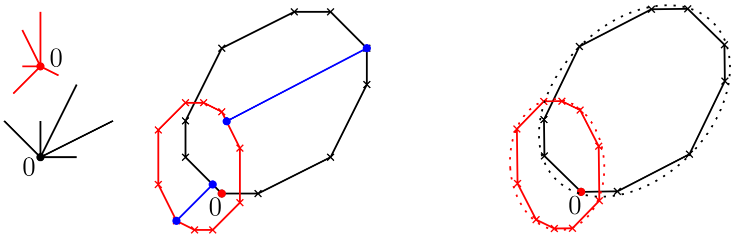

Figure 4: Left: two zonotopes (with their generating segments) and the segments achieving the two sides of the Haussdorf distance. Right: approximation by ellipsoids.

for some vectors $t_i$, $i = 1, \dots, r$ (Bolker, 1969). See an illustration in Figure 3. They appear in several areas of computer science (Edelsbrunner, 1987; Guibas et al., 2003) and mathematics (Bolker, 1969; Bourgain et al., 1989). In machine learning, they appear naturally as the affine projection of a hypercube; in particular, when using a higher-dimensional distributed representation of points in $\mathbb{R}^d$ with elements in $[0, 1]^r$, where $r$ is larger than $d$ (see, e.g., Hinton and Ghahramani, 1997), the underlying polytope that is modelled in $\mathbb{R}^d$ happens to be a zonotope.

In our context, the two convex sets $\{T_+^\top b_+, \ b_+ \in [0,1]^{n_+}\}$ and $\{T_-^\top b_-, \ b_- \in [0,1]^{n_-}\}$ defined above are thus zonotopes. See an illustration of the Hausdorff distance computation in Figure 4 (middle plot), which is the core computational problem for $\alpha = 1$.

**Approximation by ellipsoids.** Centrally symmetric convex polytopes (w.l.o.g. centered around zero) may be approximated by ellipsoids. In our set-up, we could use the minimum volume enclosing ellipsoid (see, e.g. Barvinok, 2002), which can be computed exactly when the polytope is given through its vertices, or up to a constant factor when the polytope is such that quadratic functions may be optimized with a constant factor approximation. For zonotopes, the standard semi-definite relaxation of Nesterov (1998) leads to such constant-factor approximations, and thus the minimum volume inscribed ellipsoid may be computed up to a constant. Given standard results (see, e.g. Barvinok, 2002), a $(1/\sqrt{d})$-scaled version of the ellipsoid is inscribed in this polytope, and thus the ellipsoid is a provably good approximation of the zonotope with a factor scaling as $\sqrt{d}$. However, the approximation ratio is not good enough to get any relevant bound for our purpose (see Section 5.5), as for computing the Haussdorff distance, we care about potentially vanishing differences that are swamped by constant factor approximations.

Nevertheless, the ellipsoid approximation may prove useful in practice, in particular because the $\ell_2$-Haussdorff distance between two ellipsoids may be computed in polynomial time (see Appendix E).

**NP-hardness.** Given the reduction of the case $\alpha = 1$ (rectified linear units) to $\alpha = 0$ (exact thresholds) (Livni et al., 2014), the incremental problem is also NP-hard, so as obtaining a constant-factor approximation. However, this does not rule out convex relaxations with non-constant approximation ratios (see Section 6 for more details).

### 3.4 Incremental Optimization Problem for $\alpha \geqslant 2$

We consider the problem in Eq. (2) for the remaining cases $\alpha \geqslant 2$. For $z_1, \dots, z_n \in \mathbb{R}^{d+1}$ and a vector $y \in \mathbb{R}^n$, the goal is to solve (as well as the corresponding problem with $y$ replaced by $-y$):

$$\max_{\|v\|_p \leqslant 1} \frac{1}{\alpha} \sum_{i=1}^n y_i (v^\top z_i)_+^\alpha = \max_{\|v\|_p \leqslant 1} \sum_{i \in I_+} \frac{1}{\alpha} (v^\top |y_i|^{1/\alpha} z_i)_+^\alpha - \sum_{i \in I_-} \frac{1}{\alpha} (v^\top |y_i|^{1/\alpha} z_i)_+^\alpha,$$

where $I_+ = \{i, y_i \geqslant 0\}$ and $I_- = \{i, y_i < 0\}$. We have, with $t_i = |y_i|^{1/\alpha} z_i \in \mathbb{R}^{d+1}$, and $\beta \in (1, 2]$ defined by $1/\beta + 1/\alpha = 1$ (we use the fact that the function $u \mapsto u^\alpha/\alpha$ and $v \mapsto v^\beta/\beta$ are Fenchel-dual to each other):

$$\begin{aligned}
\max_{\|v\|_p \leqslant 1} \frac{1}{\alpha} \sum_{i=1}^n y_i (v^\top z_i)_+^\alpha &= \max_{\|v\|_p \leqslant 1} \sum_{i \in I_+} \frac{1}{\alpha} (v^\top t_i)_+^\alpha - \sum_{i \in I_-} \frac{1}{\alpha} (v^\top t_i)_+^\alpha \\
&= \max_{\|v\|_p \leqslant 1} \sum_{i \in I_+} \max_{b_i \geqslant 0} \Big\{ b_i v_i^\top t_i - \frac{1}{\beta} b_i^\beta \Big\} - \sum_{i \in I_-} \max_{b_i \geqslant 0} \Big\{ b_i v^\top t_i - \frac{1}{\beta} b_i^\beta \Big\} \\
&= \max_{b_+ \in \mathbb{R}_+^{I_+}} \min_{b_- \in \mathbb{R}_+^{I_-}} \max_{\|v\|_p \leqslant 1} v^\top [T_+^\top b_+ - T_-^\top b_-] - \frac{1}{\beta} \|b_+\|_\beta^\beta + \frac{1}{\beta} \|b_-\|_\beta^\beta \\
&\qquad \text{by Fenchel duality,} \\
&= \max_{b_+ \in [0,1]^{I_+}} \min_{b_- \in [0,1]^{I_-}} \|T_+^\top b_+ - T_-^\top b_-\|_q - \frac{1}{\beta} \|b_+\|_\beta^\beta + \frac{1}{\beta} \|b_-\|_\beta^\beta, \quad (3)
\end{aligned}$$

where $T_+ \in \mathbb{R}^{n_+ \times d}$ has rows $t_i$, $i \in I_+$ and $T_- \in \mathbb{R}^{n_- \times d}$ has rows $t_i$, $i \in I_-$, with $v \in \arg\max_{\|v\|_p \leqslant 1} (T_+^\top b_+ - T_-^\top b_-)^\top v$. Contrary to the case $\alpha = 1$, we do not obtain exactly a formulation as a Hausdorff distance. However, if we consider the convex sets $K_\lambda^+ = \{T_+^\top b_+,\ b_+ \geqslant 0,\ \|b_+\|_\beta \leqslant \lambda\}$ and $K_\mu^- = \{T_-^\top b_-,\ b_- \geqslant 0,\ \|b_-\|_\beta \leqslant \mu\}$, then, a solution of Eq. (3) may be obtained from Hausdorff distance computations between $K_\lambda^+$ and $K_\mu^-$, for certain $\lambda$ and $\mu$.

Note that, while for $\alpha = 1$ we can use the identity $2u_+ = u + |u|$ to replace the rectified linear unit by the absolute value and obtain the same function space, this is not possible for $\alpha = 2$, as $(u_+)^2$ and $u^2$ do not differ by a linear function. This implies that the results from Livni et al. (2014), which state that for the quadratic activation function, the incremental problems is equivalent to an eigendecomposition (and hence solvable in polynomial time), do not apply.

## 4. Approximation Properties

In this section, we consider the approximation properties of the set $\mathcal{F}_1$ of functions defined on $\mathbb{R}^d$. As mentioned earlier, the norm used to penalize input weights $w$ or $v$ is irrelevant for approximation properties as all norms are equivalent. Therefore, we focus on the case $q = p = 2$ and $\ell_2$-norm constraints.

Because we consider homogeneous activation functions, we start by studying the set $\mathcal{G}_1$ of functions defined on the unit $\ell_2$-sphere $\mathbb{S}^d \subset \mathbb{R}^{d+1}$. We denote by $\tau_d$ the uniform probability measure on $\mathbb{S}^d$. The set $\mathcal{G}_1$ is defined as the set of functions on the sphere such that $g(z) = \int_{\mathbb{S}^d} \sigma(v^\top z) p(z) d\tau_d(z)$, with the norm $\gamma_1(g)$ equal to the smallest possible value of $\int_{\mathbb{S}^d} |p(z)| d\tau_d(z)$. We may also define the corresponding squared RKHS norm by the smallest possible value of $\int_{\mathbb{S}^d} |p(z)|^2 d\tau_d(z)$, with the corresponding RKHS $\mathcal{G}_2$.

In this section, we first consider approximation properties of functions in $\mathcal{G}_1$ by a finite number of neurons (only for $\alpha = 1$). We then study approximation properties of functions on the sphere by functions in $\mathcal{G}_1$. It turns out that all our results are based on the approximation properties of the corresponding RKHS $\mathcal{G}_2$: we give sufficient conditions for being in $\mathcal{G}_2$, and then approximation bounds for functions which are not in $\mathcal{G}_2$. Finally we transfer these to the spaces $\mathcal{F}_1$ and $\mathcal{F}_2$, and consider in particular functions which only depend on projections on a low-dimensional subspace, for which the properties of $\mathcal{G}_1$ and $\mathcal{G}_2$ (and of $\mathcal{F}_1$ and $\mathcal{F}_2$) differ. This property is key to obtaining generalization bounds that show adaptivity to linear structures in the prediction functions (as done in Section 5).

Approximation properties of neural networks with finitely many neurons have been studied extensively (see, e.g., Petrushev, 1998; Pinkus, 1999; Makovoz, 1998; Burger and Neubauer, 2001). In Section 4.7, we relate our new results to existing work from the literature on approximation theory, by showing that our results provide an explicit control of the various weight vectors which are needed for bounding the estimation error in Section 5.

### 4.1 Approximation by a Finite Number of Basis Functions

A key quantity that drives the approximability by a finite number of neurons is the variation norm $\gamma_1(g)$. As shown in Section 2.5, any function $g$ such that $\gamma_1(g)$ is finite, may be approximated in $L_2(\mathbb{S}^d)$-norm with error $\varepsilon$ with $n = O(\gamma_1(g)^2 \varepsilon^{-2})$ units. For $\alpha = 1$ (rectified linear units), we may improve the dependence in $\varepsilon$, through the link with zonoids and zonotopes, as we now present.

If we decompose the signed measure $\mu$ as $\mu = \mu_+ - \mu_-$ where $\mu_+$ and $\mu_-$ are positive measures, then, for $g \in \mathcal{G}_1$, we have $g(z) = \int_{\mathbb{S}^d} (v^\top z)_+ d\mu_+(v) - \int_{\mathbb{S}^d} (v^\top z)_+ d\mu_-(v) = g_+(z) - g_-(z)$, which is a decomposition of $g$ as a difference of positively homogenous convex functions.

Positively homogenous convex functions $h$ may be written as the *support function* of a compact convex set $K$ (Rockafellar, 1997), that is, $h(z) = \max_{y \in K} y^\top z$, and the set $K$ characterizes the function $h$. The functions $g_+$ and $g_-$ defined above are not *any* convex positively homogeneous functions, as we now describe.

If the measure $\mu_+$ is supported by finitely many points, that is, $\mu_+(v) = \sum_{i=1}^r \eta_i \delta(v - v_i)$ with $\eta \geqslant 0$, then $g_+(z) = \sum_{i=1}^t \eta_i (v_i^\top z)_+ = \sum_{i=1}^t (\eta_i v_i^\top z)_+ = \sum_{i=1}^t (t_i^\top z)_+$ for $t_i = \eta_i v_i$. Thus the corresponding set $K_+$ is the *zonotope* $[0, t_1] + \cdots + [0, t_r] = \{ \sum_{i=1}^r b_i t_i,\ b \in [0,1]^r \}$ already defined in Section 3.3. Thus the functions $g_+ \in \mathcal{G}_1$ and $g_- \in \mathcal{G}_1$ for finitely supported measures $\mu$ are support functions of zonotopes.

When the measure $\mu$ is not constrained to have finite support, then the sets $K_+$ and $K_-$ are limits of zonotopes, and thus, by definition, *zonoids* (Bolker, 1969), and thus functions in $\mathcal{G}_1$ are differences of support functions of zonoids. Zonoids are a well-studied set of convex bodies. They are centrally symmetric, and in two dimensions, all centrally symmetric compact convexs sets are (up to translation) zonoids, which is not true in higher dimensions (Bolker, 1969). Moreover, the problem of approximating a zonoid by a zonotope with a small number of segments (Bourgain et al., 1989; Matoušek, 1996) is essentially equivalent to the approximation of a function $g$ by finitely many neurons. The number of neurons directly depends on the norm $\gamma_1$, as we now show.

**Proposition 1 (Number of units - $\alpha = 1$)** *Let $\varepsilon \in (0, 1/2)$. For any function $g$ in $\mathcal{G}_1$, there exists a measure $\mu$ supported on at most $r$ points in $\mathcal{V}$, so that for all $z \in \mathbb{S}^d$. $|g(z) - \int_{\mathbb{S}^d} (v^\top z)_+ d\mu(v)| \leqslant \varepsilon \gamma_1(g)$, with $r \leqslant C(d) \varepsilon^{-2d/(d+3)}$, for some constant $C(d)$ that depends only on $d$.*

**Proof** Without loss of generality, we assume $\gamma(g) = 1$. It is shown by Matoušek (1996) that for any probability measure $\mu$ (positive and with finite mass) on the sphere $\mathbb{S}^d$, there exists a set of $r$ points $v_1, \dots, v_r$, so that for all $z \in \mathbb{S}^d$,

$$\left| \int_{\mathbb{S}^d} |v^\top z| d\mu(v) - \frac{1}{r} \sum_{i=1}^r |v_i^\top z| \right| \leqslant \varepsilon, \tag{4}$$

with $r \leqslant C(d) \varepsilon^{-2+6/(d+3)} = C(d) \varepsilon^{-2d/(d+3)}$, for some constant $C(d)$ that depends only on $d$. We may then simply write

$$g(z) = \int_{\mathbb{S}^d} (v^\top z)_+ d\mu(v) = \frac{1}{2} \int_{\mathbb{S}^d} (v^\top z) d\mu(v) + \frac{\mu_+(\mathbb{S}^d)}{2} \int_{\mathbb{S}^d} |v^\top z| \frac{d\mu_+(v)}{\mu_+(\mathbb{S}^d)} - \frac{\mu_-(\mathbb{S}^d)}{2} \int_{\mathbb{S}^d} |v^\top z| \frac{d\mu_-(v)}{\mu_-(\mathbb{S}^d)},$$

and approximate the last two terms with error $\varepsilon \mu_\pm(\mathbb{S}^d)$ with $r$ terms, leading to an approximation of $\varepsilon \mu_+(\mathbb{S}^d) + \varepsilon \mu_-(\mathbb{S}^d) = \varepsilon \gamma_1(g) = \varepsilon$, with a remainder that is a linear function $q^\top z$ of $z$, with $\|q\|_2 \leqslant 1$. We may then simply add two extra units with vectors $q / \|q\|_2$ and weights $-\|q\|_2$ and $\|q\|_2$. We thus obtain, with $2r + 2$ units, the desired approximation result.

Note that Bourgain et al. (1989, Theorem 6.5) showed that the scaling in $\varepsilon$ in Eq. (4) is not improvable, if the measure is allowed to have non equal weights on all points and the proof relies on the non-approximability of the Euclidean ball by centered zonotopes. This results does not apply here, because we may have different weights $\mu_-(\mathbb{S}^d)$ and $\mu_+(\mathbb{S}^d)$. ■

Note that the proposition above is slightly improved in terms of the scaling of the number of neurons with respect to the approximation error $\varepsilon$ (improved exponent), compared to conditional gradient bounds (Barron, 1993; Kurkova and Sanguineti, 2001). Indeed, the simple use of conditional gradient leads to $r \leqslant \varepsilon^{-2}\gamma_1(g)^2$, with a better constant (independent of $d$) but a worse scaling in $\varepsilon$—also with a result in $L_2(\mathbb{S}^d)$-norm and not uniformly on the ball $\{\|x\|_q \leqslant R\}$. Note also that the conditional gradient algorithm gives a constructive way of building the measure. Moreover, the proposition above is related to the result from Makovoz (1998, Theorem 2), which applies for $\alpha = 0$ but with a number of neurons growing as $\varepsilon^{-2d/(d+1)}$, or to the one of Burger and Neubauer (2001, Example 3.1), which applies to a piecewise affine sigmoidal function but with a number of neurons growing as $\varepsilon^{-2(d+1)/(d+3)}$ (both slightly worse than ours).

Finally, the number of neurons needed to express a function with a bound on the $\gamma_2$-norm can be estimated from general results on approximating reproducing kernel Hilbert space described in Section 2.3, whose kernel can be expressed as an expectation. Indeed, Bach (2017) shows that with $k$ neurons, one can approximate a function in $\mathcal{F}_2$ with unit $\gamma_2$-norm with an error measured in $L_2$ of $\varepsilon = k^{-(d+3)/(2d)}$. When inverting the relationship between $k$ and $\varepsilon$, we get a number of neurons scaling as $\varepsilon^{-2d/(d+3)}$, which is the same as in Prop. 1 but with an error in $L_2$-norm instead of $L^\infty$-norm.

### 4.2 Sufficient Conditions for Finite Variation

In this section and the next one, we study more precisely the RKHS $\mathcal{G}_2$ (and thus obtain similar results for $\mathcal{G}_1 \supset \mathcal{G}_2$). The kernel $k(x,y) = \int_{\mathbb{S}^d} (v^\top x)_+ (v^\top y)_+ d\tau_d(v)$ defined on the sphere $\mathbb{S}^d$ belongs to the family of dot-product kernels (Smola et al., 2001) that only depends on the dot-product $x^\top y$, although in our situation, the function is not particularly simple (see formulas in Section 3.1). The analysis of these kernels is similar to one of translation-invariant kernels; for $d = 1$, i.e., on the 2-dimensional sphere, it is done through Fourier series; while for $d > 1$, *spherical harmonics* have to be used as the expansion of functions in series of spherical harmonics make the computation of the RKHS norm explicit (see a review of spherical harmonics in Appendix D.1 with several references therein). Since the calculus is tedious, all proofs are put in appendices, and we only present here the main results. In this section, we provide simple sufficient conditions for belonging to $\mathcal{G}_2$ (and hence $\mathcal{G}_1$) based on the existence and boundedness of derivatives, while in the next section, we show how any Lipschitz-function may be approximated by functions in $\mathcal{G}_2$ (and hence $\mathcal{G}_1$) with precise control of the norm of the approximating functions.

The derivatives of functions defined on $\mathbb{S}^d$ may be defined in several ways, using the manifold structure (see, e.g., Absil et al., 2009) or through polar coordinates (see, e.g., Atkinson and Han, 2012, Chapter 3). For $d = 1$, the one-dimensional sphere $\mathbb{S}^1 \subset \mathbb{R}^2$ may be parameterized by a single angle and thus the notion of derivatives and the proof of the following result is simpler and based on Fourier series (see Appendix C.2). For the general proof based on spherical harmonics, see Appendix D.2.

**Proposition 2 (Finite variation on the sphere)** *Assume that $g : \mathbb{S}^d \to \mathbb{R}$ is such that all $i$-th order derivatives exist and are upper-bounded in absolute value by $\eta$ for $i \in \{0, \dots, s\}$, where $s$ is an integer such that $s \geqslant (d-1)/2 + \alpha + 1$. Assume $g$ is even if $\alpha$ is odd (and vice-versa); then $g \in \mathcal{G}_2$ and $\gamma_2(g) \leqslant C(d,\alpha)\eta$, for a constant $C(d,\alpha)$ that depends only on $d$ and $\alpha$.*

We can make the following observations:

- *Tightness of conditions*: as shown in Appendix D.5, there are functions $g$, which have bounded first $s$ derivatives and do not belong to $\mathcal{G}_2$ while $s \leqslant \frac{d}{2} + \alpha$ (at least when $s - \alpha$ is even). Therefore, when $s - \alpha$ is even, the scaling in $(d-1)/2 + \alpha$ is optimal.

- *Dependence on $\alpha$*: for any $d$, the higher the $\alpha$, the stricter the sufficient condition. Given that the estimation error grows slowly with $\alpha$ (see Section 5.1), low values of $\alpha$ would be preferred in practice.

- *Dependence on $d$*: a key feature of the sufficient condition is the dependence on $d$, that is, as $d$ increases the number of derivatives has to increase in $d/2$—like for Sobolev spaces in dimension $d$ (Adams and Fournier, 2003). This is another instantiation of the curse of dimensionality: only very smooth functions in high dimensions are allowed.

- *Special case $d = 1$, $\alpha = 0$*: differentiable functions on the sphere in $\mathbb{R}^2$, with bounded derivatives, belong to $\mathcal{G}_2$, and thus all Lipschitz-continuous functions, because Lipschitz-continuous functions are almost everywhere differentiable with bounded derivative (Adams and Fournier, 2003).

### 4.3 Approximation of Lipschitz-continuous Functions

In order to derive generalization bounds for target functions which are not sufficiently differentiable (and may not be in $\mathcal{G}_2$ or $\mathcal{G}_1$), we need to approximate any Lipschitz-continuous function, with a function $g \in \mathcal{G}_2$ with a norm $\gamma_2(g)$ that will grow as the approximation gets tighter. We give precise rates in the proposition below. Note the requirement for parity of the function $g$. The result below notably shows the density of $\mathcal{G}_1$ in uniform norm in the space of Lipschitz-continuous functions of the given parity, which is already known since our activation functions are not polynomials (Leshno et al., 1993).

**Proposition 3 (Approximation of Lipschitz-continuous functions on the sphere)** *For $\delta$ greater than a constant depending only on $d$ and $\alpha$, for any function $g : \mathbb{S}^d \to \mathbb{R}$ such that for all $x, y \in \mathbb{S}^d$, $g(x) \leqslant \eta$ and $|g(x) - g(y)| \leqslant \eta \|x - y\|_2$, and $g$ is even if $\alpha$ is odd (and vice-versa), there exists $h \in \mathcal{G}_2$, such that $\gamma_2(h) \leqslant \delta$ and*

$$\sup_{x \in \mathbb{S}^d} |h(x) - g(x)| \leqslant C(d, \alpha)\eta \Big(\frac{\delta}{\eta}\Big)^{-1/(\alpha + (d-1)/2)} \log\Big(\frac{\delta}{\eta}\Big).$$

This proposition is shown in Appendix C.3 for $d = 1$ (using Fourier series) and in Appendix D.4 for all $d \geqslant 1$ (using spherical harmonics). We can make the following observations:

- *Dependence in $\delta$ and $\eta$*: as expected, the main term in the error bound $(\delta/\eta)^{-1/(\alpha+(d-1)/2)}$ is a decreasing function of $\delta/\eta$, that is when the norm $\gamma_2(h)$ is allowed to grow, the approximation gets tighter, and when the Lipschitz constant of $g$ increases, the approximation is less tight.

- *Dependence on $d$ and $\alpha$*: the rate of approximation is increasing in $d$ and $\alpha$. In particular the approximation properties are better for low $\alpha$.

- *Special case $d = 1$ and $\alpha = 0$*: up to the logarithmic term we recover the result of Prop. 2, that is, the function $g$ is in $\mathcal{G}_2$.

- *Tightness*: in Appendix D.5, we provide a function which is not in the RKHS and for which the tightest possible approximation scales as $\delta^{-2/(d/2+\alpha-2)}$. Thus the linear scaling of the rate as $d/2 + \alpha$ is not improvable (but constants are).

### 4.4 Linear Functions

In this section, we consider a linear function on $\mathbb{S}^d$, that is $g(x) = v^\top x$ for a certain $v \in \mathbb{S}^d$, and compute its norm (or upper-bound thereof) both for $\mathcal{G}_1$ and $\mathcal{G}_2$, which is independent of $v$ and finite. In the following propositions, the notation $\approx$ means asymptotic equivalents when $d \to \infty$.

**Proposition 4 (Norms of linear functions on the sphere)** *Assume that $g : \mathbb{S}^d \to \mathbb{R}$ is such $g(x) = v^\top x$ for a certain $v \in \mathbb{S}^d$. If $\alpha = 0$, then $\gamma_1(g) \leqslant \gamma_2(g) = \frac{2d\pi}{d-1} \approx 2\pi$. If $\alpha = 1$, then $\gamma_1(g) \leqslant 2$, and for all $\alpha \geqslant 1$, $\gamma_1(g) \leqslant \gamma_2(g) = \frac{d}{d-1}\frac{4\pi}{\alpha}\frac{\Gamma(\alpha/2+d/2+1)}{\Gamma(\alpha/2)\Gamma(d/2+1)} \approx Cd^{\alpha/2}$.*

We see that for $\alpha = 1$, the $\gamma_1$-norm is less than a constant, and is much smaller than the $\gamma_2$-norm (which scales as $\sqrt{d}$). For $\alpha \geqslant 2$, we were not able to derive better bounds for $\gamma_1$ (other than the value of $\gamma_2$).

### 4.5 Functions of Projections

If $g(x) = \varphi(w^\top x)$ for some unit-norm $w \in \mathbb{R}^{d+1}$ and $\varphi$ a function defined on the real-line, then the value of the norms $\gamma_2$ and $\gamma_1$ differ significantly. Indeed, for $\gamma_1$, we may consider a new variable $\tilde{x} \in \mathbb{S}^1 \subset \mathbb{R}^2$, with its first component $\tilde{x}_1 = w^\top x$, and the function $\tilde{g}(x) = \varphi(\tilde{x}_1)$. We may then apply Prop. 2 to $\tilde{g}$ with $d = 1$. That is, if $\varphi$ is $(\alpha+1)$-times differentiable with bounded derivatives, there exists a decomposition $\tilde{g}(\tilde{x}) = \int_{\mathbb{S}^1} \tilde{\mu}(\tilde{v})\sigma(\tilde{v}^\top \tilde{x})d\tilde{\mu}$, with $\gamma_1(\tilde{g}) = |\tilde{\mu}|(\mathbb{S}^1)$, which is *not* increasing in $d$. If we consider any vector $t \in \mathbb{R}^{d+1}$ which is orthogonal to $w$ in $\mathbb{R}^{d+1}$, then, we may define a measure $\mu$ supported in the circle defined by the two vectors $w$ and $t$ and which is equal to $\tilde{\mu}$ on that circle. The total variation of $\mu$ is the one of $\tilde{\mu}$ while $g$ can be decomposed using $\mu$ and thus $\gamma_1(g) \leqslant \gamma_1(\tilde{g})$. Similarly, Prop. 3 could also be applied (and will for obtaining generalization bounds), also our reasoning works for any low-dimensional projections: the dependence on a lower-dimensional projection allows to reduce smoothness requirements.

However, for the RKHS norm $\gamma_2$, this reasoning does not apply. For example, a certain function $\varphi$ exists, which is $s$-times differentiable, as shown in Appendix D.5, for $s \leqslant \frac{d}{2} + \alpha$ (when $s - \alpha$ is even), and is not in $\mathcal{G}_2$. Thus, given Prop. 2, the dependence on a uni-dimensional projection does not make a difference regarding the level of smoothness which is required to belong to $\mathcal{G}_2$.

### 4.6 From the Unit-sphere $\mathbb{S}^d$ to $\mathbb{R}^{d+1}$

We now extend the results above to functions defined on $\mathbb{R}^d$, to be approximated by functions in $\mathcal{F}_1$ and $\mathcal{F}_2$. More precisely, we first extend Prop. 2 and Prop. 3, and then consider norms of linear functions and functions of projections.

**Proposition 5 (Finite variation)** *Assume that $f : \mathbb{R}^d \to \mathbb{R}$ is such that all $i$-th order derivatives exist and are upper-bounded on the ball $\{\|x\|_q \leqslant R\}$ by $\eta/R^i$ for $i \in \{0, \dots, k\}$, where $s$ is the smallest integer such that $s \geqslant (d-1)/2 + \alpha + 1$; then $f \in \mathcal{F}_2$ and $\gamma_2(f) \leqslant C(d,\alpha)\eta$, for a constant $C(d,\alpha)$ that depends only on $d$ and $\alpha$.*

**Proof** By assumption, the function $x \mapsto f(Rx)$ has all its derivatives bounded by a constant times $\eta$. Moreover, we have defined $g(t,a) = f(\frac{Rt}{a})a^\alpha$ so that all derivatives are bounded by $\eta$. The result then follows immediately from Prop. 2. ■

**Proposition 6 (Approximation of Lipschitz-continuous functions)** *For $\delta$ larger than a constant that depends only on $d$ and $\alpha$, for any function $f : \mathbb{R}^d \to \mathbb{R}$ such that for all $x, y$ such that $\|x\|_q \leqslant R$ and $\|y\|_q \leqslant R$, $|f(x)| \leqslant \eta$ and $|f(x) - f(y)| \leqslant \eta R^{-1}\|x-y\|_q$, there exists $g \in \mathcal{F}_2$ such that $\gamma_2(g) \leqslant \delta$ and*

$$\sup_{\|x\|_q \leqslant R} |f(x) - g(x)| \leqslant C(d,\alpha)\eta\Big(\frac{\delta}{\eta}\Big)^{-1/(\alpha+(d-1)/2)} \log\Big(\frac{\delta}{\eta}\Big).$$

**Proof** With the same reasoning as above, we obtain that $g$ is Lipschitz-continuous with constant $\eta$, we thus get the desired approximation error from Prop. 3. ■

**Linear functions.** If $f(x) = w^\top x + b$, with $\|w\|_2 \leqslant \eta$ and $b \leqslant \eta R$, then for $\alpha = 1$, it is straightforward that $\gamma_1(f) \leqslant 2R\eta$. Moreover, we have $\gamma_2(f) \sim CR\eta$. For other values of $\alpha$, we also have $\gamma_1$-norms less than a constant (depending *only* of $\alpha$) times $R\eta$. The RKHS norms are bit harder to compute since linear functions for $f$ leads to linear functions for $g$ only for $\alpha = 1$.

**Functions of projections.** If $f(x) = \varphi(w^\top x)$ where $\|w\|_2 \leqslant \eta$ and $\varphi : \mathbb{R} \to \mathbb{R}$ is a function, then the norm of $f$ is the same as the norm of the function $\varphi$ on the interval $[-R\eta, R\eta]$, and it thus does not depend on $d$. This is a consequence of the fact that the total mass of a Radon measure remains bounded even when the support has measure zero (which might not be the case for the RKHS defined in Section 2.3). For the RKHS, there is no such results and it is in general not adaptive.

More generally, if $f(x) = \Phi(W^\top x)$ for $W \in \mathbb{R}^{d \times s}$ with the largest singular value of $W$ less than $\eta$, and $\Phi$ a function from $\mathbb{R}^s$ to $\mathbb{R}$, then for $\|x\|_2 \leqslant R$, we have $\|W^\top x\|_2 \leqslant R\eta$, and thus we may apply our results for $d = s$.

**$\ell_1$-penalty on input weights ($p$=1).** When using an $\ell_1$-penalty on input weights instead of an $\ell_2$-penalty, the results in Prop. 5 and 6 are unchanged (only the constants that depend on $d$ are changed). Moreover, when $\|x\|_\infty \leqslant 1$ almost surely, functions of the form $f(x) = \varphi(w^\top x)$ where $\|w\|_1 \leqslant \eta$ and $\varphi : \mathbb{R} \to \mathbb{R}$ is a function, will also inherit from properties of $\varphi$ (without any dependence on dimension). Similarly, for functions of the form $f(x) = \Phi(W^\top x)$ for $W \in \mathbb{R}^{d \times s}$ with all columns of $\ell_1$-norm less than $\eta$, we have $\|W^\top x\|_\infty \leqslant R\eta$ and we can apply the $s$-dimensional result.

### 4.7 Related Work

In this section, we show how our results from the previous sections relate to existing work on neural network approximation theory.

**Approximation of Lipschitz-continuous functions with finitely many neurons.** In this section, we only consider the case $\alpha = 1$, for which we have two approximation bounds: Prop. 6 which approximates any $\eta$-Lipschitz-continuous function by a function with finite $\gamma_1$-norm less than $\delta$ and uniform error less than $\eta(\delta/\eta)^{-2/(d+1)} \log(\delta/\eta)$, and Prop. 1 which shows that a function with $\gamma_1$-norm less than $\delta$, may be approximated with $r$ neurons with uniform error $\delta r^{-(d+3)/(2d)}$.

Thus, given $r$ neurons, we get an approximation of the original function with uniform error

$$\eta(\delta/\eta)^{-2/(d+1)} \log(\delta/\eta) + \delta r^{-(d+3)/(2d)}.$$

We can optimize over $\delta$, and use $\delta = \eta n^{(d+1)/(2d)}$, to obtain a uniform approximation bound proportional to $\eta(\log n) n^{-1/d}$, for approximating an $\eta$-Lipschitz-continuous function with $n$ neurons.

**Approximation by ridge functions.** The approximation properties of single hidden layer neural networks have been studied extensively, where they are often referred to as "ridge function" approximations. As shown by Pinkus (1999, Corollary 6.10)—based on a result from Petrushev (1998), the approximation order of $n^{-1/d}$ for the rectified linear unit was already known, but only in $L_2$-norm (and without the factor $\log n$), and without any constraints on the input and output weights. In this paper, we provide an explicit control of the various weights, which is needed for computing estimation errors. Moreover, while the two proof techniques use spherical harmonics, the proof of Petrushev (1998) relies on quadrature formulas for the associated Legendre polynomials, while ours relies on the relationship with the associated positive definite kernels, is significantly simpler, and offers additional insights into the problem (relationship with convex neural networks and zonoids). Maiorov (2006, Theorem 2.3) also derives a similar result, but in $L_2$-norm (rather than uniform norm), and for sigmoidal activation functions (which are bounded). Note finally, that the order $O(n^{-1/d})$ cannot be improved (DeVore et al., 1989, Theorem 4.2). Also, Maiorov and Meir (2000, Theorem 5) derive similar upper and lower bounds based on a random sampling argument which is close to using random features in the RKHS setting described in Section 2.3.

**Relationship to hardness results for Boolean-valued functions.** In this paper, we consider a particular view of the curse of dimensionality and ways of circumventing it, that is, our distribution over inputs is arbitrary, but our aim is to approximate a real-valued function. Thus, all hardness results depending on functions with values in $\{0,1\}$ do not apply there directly—see, e.g., Shalev-Shwartz and Ben-David (2014, Chapter 20), for the need of exponentially many hidden units for approximating most of the functions from $\{0,1\}^d$ to $\{0,1\}$.

Our approximation bounds show that, without any assumption beyond Lipschitz-continuity of the target function, it sufficient to have a number of hidden units which is still exponential in dimension (hence we also suffer from the curse of dimensionality), but a soon as the target function depends on linear low-dimensional structure, then we lose this exponential dependence. It would be interesting to study an extension to $\{0,1\}$-valued functions, and also to relate our results to the number of linear regions delimited by neural networks with rectified linear units (Montufar et al., 2014).

## 5. Generalization Bounds

Our goal is to derive the generalization bounds outlined in Section 2.4 for neural networks with a single hidden layer. The main results that we obtain are summarized in Table 2 and show adaptivity to assumptions that avoid the curse of dimensionality.

More precisely, given some distribution over the pairs $(x,y) \in \mathcal{X} \times \mathcal{Y}$, a loss function $\ell : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}$, our aim is to find a function $f : \mathbb{R}^d \to \mathbb{R}$ such that $J(f) = \mathbb{E}[\ell(y, f(x))]$ is small, given some i.i.d. observations $(x_i, y_i)$, $i = 1, \dots, n$. We consider the empirical risk minimization framework over a space of functions $\mathcal{F}$, equipped with a norm $\gamma$ (in our situations, $\mathcal{F}_1$ and $\mathcal{F}_2$, equipped with $\gamma_1$ or $\gamma_2$). The empirical risk $\hat{J}(f) = \frac{1}{n}\sum_{i=1}^n \ell(y_i, f(x_i))$, is minimized by constraining $f$ to be in the ball $\mathcal{F}^\delta = \{f \in \mathcal{F},\ \gamma(f) \leqslant \delta\}$.

We assume that almost surely, $\|x\|_q \leqslant R$, that for all $y$ the function $u \mapsto \ell(y,u)$ is $G$-Lipschitz-continuous on $\{|u| \leqslant \sqrt{2}\delta\}$, and that almost surely, $\ell(y,0) \leqslant G\delta$. As before $z$ denotes $z = (x^\top, R)^\top$ so that $\|z\|_q \leqslant \sqrt{2}R$. This corresponds to the following examples:

- Logistic regression and support vector machines: we have $G = 1$.
- Least-squares regression: we take $G = \max\{\sqrt{2}\delta + \|y\|_\infty, \frac{\|y\|_\infty^2}{\sqrt{2}\delta}\}$.

Approximation errors $\inf_{f \in \mathcal{F}^\delta} J(f) - \inf_{f \in \mathcal{F}} J(f)$ will be obtained from the approximation results from Section 4 by assuming that the optimal target function $f_*$ has a specific form. Indeed, we have:

$$\inf_{f \in \mathcal{F}^\delta} J(f) - J(f_*) \leqslant G \inf_{f \in \mathcal{F}^\delta} \Big\{ \sup_{\|x\|_q \leqslant R} |f(x) - f_*(x)| \Big\}.$$

We now deal with estimation errors $\sup_{f \in \mathcal{F}^\delta} |\hat{J}(f) - J(f)|$ using Rademacher complexities.

### 5.1 Estimation Errors and Rademacher Complexity

The following proposition bounds the uniform deviation between $J$ and its empirical counterpart $\hat{J}$. This result is standard (see, e.g., Koltchinskii, 2001; Bartlett and Mendelson, 2003) and may be extended in bounds that hold with high-probability.

**Proposition 7 (Uniform deviations)** *We have the following bound on the expected uniform deviation:*

$$\mathbb{E}\Big[\sup_{\gamma_1(f) \leqslant \delta} |J(f) - \hat{J}(f)|\Big] \leqslant 4\frac{G\delta}{\sqrt{n}} C(p,d,\alpha),$$

*with the following constants:*

- *for $\alpha \geqslant 1$, $C(p,d,\alpha) \leqslant \alpha\sqrt{2\log(d+1)}$ for $p = 1$ and $C(p,d,\alpha) \leqslant \frac{\alpha}{\sqrt{p-1}}$ for $p \in (1,2]$*
- *for $\alpha = 0$, $C(p,d,\alpha) \leqslant C\sqrt{d+1}$, where $C$ is a universal constant.*

**Proof** We use the standard framework of Rademacher complexities and get:

$$\begin{aligned}
& \mathbb{E} \sup_{\gamma_1(f) \leqslant \delta} |J(f) - \hat{J}(f)| \\
\leqslant\ & 2\mathbb{E} \sup_{\gamma_1(f) \leqslant \delta} \bigg| \frac{1}{n} \sum_{i=1}^n \tau_i \ell(y_i, f(x_i)) \bigg| \text{ using Rademacher random variables } \tau_i, \\
\leqslant\ & 2\mathbb{E} \sup_{\gamma_1(f) \leqslant \delta} \bigg| \frac{1}{n} \sum_{i=1}^n \tau_i \ell(y_i, 0) \bigg| + 2\mathbb{E} \sup_{\gamma_1(f) \leqslant \delta} \bigg| \frac{1}{n} \sum_{i=1}^n \tau_i [\ell(y_i, f(x_i)) - \ell(y_i, 0)] \bigg| \\
\leqslant\ & 2\frac{G\delta}{\sqrt{n}} + 2G\mathbb{E} \sup_{\gamma(f) \leqslant \delta} \bigg| \frac{1}{n} \sum_{i=1}^n \tau_i f(x_i) \bigg| \text{ using the Lipschitz-continuity of the loss,} \\
\leqslant\ & 2\frac{G\delta}{\sqrt{n}} + 2G\delta\mathbb{E} \sup_{\|v\|_p \leqslant 1/R} \bigg| \frac{1}{n} \sum_{i=1}^n \tau_i (v^\top z_i)_+^\alpha \bigg| \text{ using Eq. (2).}
\end{aligned}$$

We then take different routes for $\alpha \geqslant 1$ and $\alpha = 0$.

For $\alpha \geqslant 1$, we have the upper-bound

$$\begin{aligned}
\mathbb{E} \sup_{\gamma_1(f) \leqslant \delta} |J(f) - \hat{J}(f)| & \leqslant 2\frac{G\delta}{\sqrt{n}} + 2G\delta\alpha\mathbb{E} \sup_{\|v\|_p \leqslant 1/R} \bigg| \frac{1}{n} \sum_{i=1}^n \tau_i v^\top z_i \bigg| \\
& \qquad \text{using the } \alpha\text{-Lipschitz-cont. of } (\cdot)_+^\alpha \text{ on } [-1, 1], \\
& \leqslant 2\frac{G\delta}{\sqrt{n}} + 2\frac{G\alpha\delta}{Rn}\mathbb{E} \bigg\| \sum_{i=1}^n \tau_i z_i \bigg\|_q.
\end{aligned}$$

From Kakade et al. (2009), we get the following bounds on Rademacher complexities:

– If $p \in (1, 2]$, then $q \in [2, \infty)$, and $\mathbb{E}\| \sum_{i=1}^n \tau_i z_i \|_q \leqslant \sqrt{q-1} R \sqrt{n} = \frac{1}{\sqrt{p-1}} R \sqrt{n}$

– If $p = 1$, then $q = \infty$, and $\mathbb{E}\| \sum_{i=1}^n \tau_i z_i \|_q \leqslant R\sqrt{n}\sqrt{2\log(d+1)}$.

Overall, we have $\mathbb{E}\| \sum_{i=1}^n \tau_i z_i \|_q \leqslant \sqrt{n} R C(p, d)$ with $C(p, d)$ defined above, and thus

$$\mathbb{E} \sup_{\gamma(f) \leqslant \delta} |J(f) - \hat{J}(f)| \leqslant 2\frac{G\delta}{\sqrt{n}}(1 + \alpha C(p, d)) \leqslant 4\frac{G\delta\alpha}{\sqrt{n}} C(p, d).$$

For $\alpha = 0$, we can simply go through the VC-dimension of half-hyperplanes, which is equal to $d$, and Theorem 6 from Bartlett and Mendelson (2003), that shows that $\mathbb{E} \sup_{v \in \mathbb{R}^{d+1}} \Big| \frac{1}{n} \sum_{i=1}^n \tau_i 1_{v^\top z_i} \Big| \leqslant C\frac{\sqrt{d+1}}{\sqrt{n}}$, where $C$ is a universal constant.

Note that using standard results from Rademacher complexities, we have, with probability greater than $1 - u$, $\sup_{\gamma_1(f) \leqslant \delta} |J(f) - \hat{J}(f)| \leqslant \mathbb{E} \sup_{\gamma_1(f) \leqslant \delta} |J(f) - \hat{J}(f)| + \frac{2G\delta}{\sqrt{n}} \sqrt{\log \frac{2}{u}}$. ■

### 5.2 Generalization Bounds for $\ell_2$-norm Constraints on Input Weights ($p = 2$)

We now provide generalization bounds for the minimizer of the empirical risk given the contraint that $\gamma_1(f) \leqslant \delta$ for a well chosen $\delta$, that will depend on the assumptions regarding the target function $f_*$, listed in Section 1. In this section, we consider an $\ell_2$-norm on input weights $w$, while in the next section, we consider the $\ell_1$-norm. The two situations are summarized and compared in Table 2, where we consider that $\|x\|_\infty \leqslant r$ almost surely, which implies that our bound $R$ will depend on dimension as $R \leqslant r\sqrt{d}$.

Our generalization bounds are expected values of the excess expected risk for a our estimator (where the expectation is taken over the data).

**Affine functions.** We assume $f_*(x) = w^\top x + b$, with $\|w\|_2 \leqslant \eta$ and $|b| \leqslant R\eta$. Then, as seen in Section 4.6, $f_* \in \mathcal{F}_1$ with $\gamma_1(f_*) \leqslant C(\alpha)\eta R$ (the constant is independent of $d$ because we approximate an affine function). From Prop. 7, we thus get a generalization bound proportional to $\frac{GR\eta}{\sqrt{n}}$ times a constant (that may depend on $\alpha$), which is the same as assuming directly that we optimize over linear predictors only. The chosen $\delta$ is then a constant times $R\eta$, and does not grow with $n$, like in parametric estimation (although we do use a non-parametric estimation procedure).

**Projection pursuit.** We assume $f_*(x) = \sum_{j=1}^k f_j(w_j^\top x)$, with $\|w_j\|_2 \leqslant \eta$ and each $f_j$ bounded by $\eta R$ and 1-Lipschitz continuous. From Prop. 6, we may approach each $x \mapsto f_j(w_j^\top x)$ by a function with $\gamma_1$-norm less than $\delta\eta R$ and uniform approximation $C(\alpha)\eta R\delta^{-1/\alpha}\log\delta$. This leads to a total approximation error of $kC(\alpha)G\eta R\delta^{-1/\alpha}\log\delta$ for a norm less than $k\delta\eta R$ (the constant is independent of $d$ because we approximate a function of one-dimensional projection).

For $\alpha \geqslant 1$, from Prop. 7, the estimation error is $\frac{kGR\eta\delta}{\sqrt{n}}$, with an overall bound of $C(\alpha)kGR\eta(\frac{\delta}{\sqrt{n}}+\delta^{-1/\alpha}\log\delta)$. With $\delta = n^{\alpha/2(\alpha+1)}$ (which grows with $n$), we get an optimized generalization bound of $C(\alpha)kGR\eta\frac{\log n}{n^{1/(2\alpha+2)}}$, with a scaling independent of the dimension $d$ (note however that $R$ typically grow with $\sqrt{d}$, i.e., $r\sqrt{d}$, if we have a bound in $\ell_\infty$-norm for all our inputs $x$).

For $\alpha = 0$, from Prop. 5, the target function belongs to $\mathcal{F}_1$ with a norm less than $kGR\eta$, leading to an overall generalization bound of $\frac{kGR\eta\sqrt{d}}{\sqrt{n}}$.

Note that when the functions $f_j$ are exactly the activation functions, the bound is better, as these functions directly belong to the space $\mathcal{F}_1$.

**Multi-dimensional projection pursuit.** We extend the situation above, by assuming $f_*(x) = \sum_{j=1}^k F_j(W_j^\top x)$ with each $W_j \in \mathbb{R}^{d\times s}$ having all singular values less than $\eta$ and each $F_j$ bounded by $\eta R$ and 1-Lipschitz continuous. From Prop. 6, we may approach each $x \mapsto F_j(W_j^\top x)$ by a function with $\gamma_1$-norm less than $\delta\eta R$ and uniform approximation $C(\alpha,s)\eta R\delta^{-1/(\alpha+(s-1)/2)}\log\delta$. This leads to a total approximation error of $kC(\alpha,s)G\eta R\delta^{-1/(\alpha+(s-1)/2)}\log\delta$.

For $\alpha \geqslant 1$, the estimation error is $kGR\eta\delta/\sqrt{n}$, with an overall bound of $C(\alpha,s)kGR\eta(\delta/\sqrt{n}+\delta^{-1/(\alpha+(s-1)/2)}\log\delta)$. With $\delta = n^{(\alpha+(s-1)/2)/(2\alpha+s-1)}$, we get an optimized bound of $\frac{C(\alpha,s)kGR\eta}{n^{1/(2\alpha+s+1)}}\log n$.

For $\alpha = 0$, we have an overall bound of $C(s)kGR\eta(\delta^{-2/(s-1)}\log\delta + \frac{\delta\sqrt{d}}{\sqrt{n}})$, and with $\delta = (n/d)^{(s-1)/(s+1)}$, we get a generalization bound scaling as $\frac{C(s)kGR\eta}{(n/d)^{1/(s+1)}}\log(n/d)$.

| function space | $\Vert \cdot \Vert_2,\ \alpha \geqslant 1$ | $\Vert \cdot \Vert_1,\ \alpha \geqslant 1$ | $\alpha = 0$ |
|---|---|---|---|
| $w^\top x + b$ | $\frac{d^{1/2}}{n^{1/2}}$ | $\sqrt{q}(\frac{\log d}{n})^{1/2}$ | $\frac{(dq)^{1/2}}{n^{1/2}}$ |
| $\sum_{j=1}^{k} f_j(w_j^\top x),\ w_j \in \mathbb{R}^d$ | $\frac{kd^{1/2}}{n^{1/(2\alpha+2)}} \log n$ | $\frac{kq^{1/2}(\log d)^{1/(\alpha+1)}}{n^{1/(2\alpha+2)}} \log n$ | $\frac{k(dq)^{1/2}}{n^{1/2}}$ |
| $\sum_{j=1}^{k} f_j(W_j^\top x),\ W_j \in \mathbb{R}^{d\times s}$ | $\frac{kd^{1/2}}{n^{1/(2\alpha+s+1)}} \log n$ | $\frac{kq^{1/2}(\log d)^{1/(\alpha+(s+1)/2)}}{n^{1/(2\alpha+s+1)}} \log n$ | $\frac{(dq)^{1/2}d^{1/(s+1)}}{n^{1/(s+1)}} \log n$ |

Table 2: Summary of generalization bounds with different settings. See text for details.

Note that for $s = d$ and $k = 1$, we recover the usual Lipschitz-continuous assumption, with a rate of $\frac{C(\alpha,d)kGR\eta}{n^{1/(2\alpha+d+1)}} \log n$.

We can make the following observations:

– *Summary table*: when we know a bound $r$ on all dimensions of $x$, then we may take $R = r\sqrt{d}$; this is helpful in comparisons in Table 2, where $R$ is replaced by $r\sqrt{d}$ and the dependence in $r$ is removed as it is the same for all models.

– *Dependence on $d$*: when making only a global Lipschitz-continuity assumption, the generalization bound has a bad scaling in $n$, i.e., as $n^{-1/(2\alpha+d+1)}$, which goes down to zero slowly when $d$ increases. However, when making structural assumptions regarding the dependence on unknown lower-dimensional subspaces, the scaling in $d$ disappears.

– *Comparing different values of $\alpha$*: the value $\alpha = 0$ always has the best scaling in $n$, but constants are better for $\alpha \geqslant 1$ (among which $\alpha = 1$ has the better scaling in $n$).

– *Bounds for $\mathcal{F}_2$*: The simplest upper bound for the penalization by the space $\mathcal{F}_2$ depends on the approximation properties of $\mathcal{F}_2$. For linear functions and $\alpha = 1$, it is less than $\sqrt{d}\eta R$, with a bound $\frac{GR\eta\sqrt{d}}{\sqrt{n}}$. For the other values of $\alpha$, there is a constant $C(d)$. Otherwise, there is no adaptivity and all other situations only lead to upper-bounds of $O(n^{-1/(2\alpha+d+1)})$. See more details in Section 5.4.

– *Sample complexity*: Note that the generalization bounds above may be used to obtain sample complexity results such as $d\varepsilon^{-2}$ for affine functions, $(\varepsilon k^{-1}d^{-1/2})^{-2\alpha-2}$ for projection pursuit, and $(\varepsilon k^{-1}d^{-1/2})^{-s-1-2\alpha}$ for the generalized version (up to logarithmic terms).

– *Relationship to existing work*: Maiorov (2006, Theorem 1.1) derives similar results for neural networks with sigmoidal activation functions (that tend to one at infinity) and the square loss only, and for a level of smoothness of the target function which grows with dimension (in this case, once can get easily rates of $n^{-1/2}$). Our result holds for problems where only bounded first-order derivatives are assumed, but by using Prop. 2, we would get similar rate by ensuring the target function belongs to $\mathcal{F}_2$ and hence to $\mathcal{F}_1$.

**Lower bounds.** In the sections above, we have only provided generalization bounds. Although interesting, deriving lower-bounds for the generalization performance when the target function belongs to certain function classes is out of the scope of this paper. Note however, that results from Sridharan (2012) suggest that the Rademacher complexities of the associated function classes provide such lower-bounds. For general Lipschitz-functions, these Rademacher complexities decreases as $n^{-\max\{d,2\}}$ (von Luxburg and Bousquet, 2004).

### 5.3 Generalization Bounds for $\ell_1$-norm Constraints on Input Weights ($p = 1$)

We consider the same three situations, assuming that linear predictors have at most $q$ non-zero elements. We assume that each component of $x$ is almost surely bounded by $r$ (i.e., a bound in $\ell_\infty$-norm).

**Affine functions.** We assume $f_*(x) = w^\top x + b$, with $\|w\|_2 \leqslant \eta$ and $|b| \leqslant R\eta$. Given that we have assumed that $w$ has at most $q$ non-zeros, we have $\|w\|_1 \leqslant \sqrt{q}\eta$.

Then, $f_* \in \mathcal{F}_1$ with $\gamma_1(f) \leqslant C(\alpha)\eta r\sqrt{q}$, with a constant that is independent of $d$ because we have an affine function.

From Prop. 7, we thus get a rate of $\frac{Gr\eta\sqrt{q\log(d)}}{\sqrt{n}}$ times a constant (that may depend on $\alpha$), which is the same as assuming directly that we optimize over linear predictors only (see, for example, Bühlmann and Van De Geer, 2011). We recover a high-dimensional phenomenon (although with a slow rate in $1/\sqrt{n}$), where $d$ may be much larger than $n$, as long as $\log d$ is small compared to $n$. The chosen $\delta$ is then a constant times $r\eta\sqrt{q}$ (and does not grow with $n$).

**Projection pursuit.** We assume $f_*(x) = \sum_{j=1}^k f_j(w_j^\top x)$, with $\|w_j\|_2 \leqslant \eta$ (which implies $\|w_j\|_1 \leqslant \sqrt{q}\eta$ given our sparsity assumption) and each $f_j$ bounded by $\eta r\sqrt{q}$ and 1-Lipschitz continuous. We may approach each $x \mapsto f_j(w_j^\top x)$ by a function with $\gamma_1$-norm less than $\delta\eta r\sqrt{q}$ and uniform approximation $C(\alpha)\eta r\sqrt{q}\delta^{-1/\alpha}\log\delta$, with a constant that is independent of $d$ because we have a function of one-dimensional projection. This leads to a total approximation error of $kC(\alpha)G\eta r\sqrt{q}\delta^{-1/\alpha}\log\delta$ for a norm less than $k\delta\eta r\sqrt{q}$.

For $\alpha \geqslant 1$, the estimation error is $\frac{kGr\eta\delta\sqrt{q\log d}}{\sqrt{n}}$, with an overall bound of $C(\alpha)kGr\sqrt{q}\eta(\delta^{-1/\alpha}\log\delta + \frac{\delta\sqrt{\log d}}{\sqrt{n}})$. With $\delta = (n/\log d)^{\alpha/2(\alpha+1)}$, we get an optimized bound of $C(\alpha)kGr\sqrt{q}\eta\frac{\log n(\log d)^{1/(2\alpha+2)}}{n^{1/(2\alpha+2)}}$, with a scaling only dependent in $d$ with a logarithmic factor.

For $\alpha = 0$, the target function belongs to $\mathcal{F}_1$ with a norm less than $kGr\sqrt{q}\eta$, leading to an overal bound of $\frac{kGr\eta\sqrt{q\log d}}{\sqrt{n}}$ (the sparsity is not helpful in this case).

**Multi-dimensional projection pursuit.** We assume $f_*(x) = \sum_{j=1}^k F_j(W_j^\top x)$ with each $W_j \in \mathbb{R}^{d\times s}$, having all columns with $\ell_2$-norm less than $\eta$ (note that this is a weaker requirement than having all singular values that are less than $\eta$). If we assume that each of these columns has at most $q$ non-zeros, then the $\ell_1$-norms are less than $r\sqrt{q}$ and we may use the approximation properties described at the end of Section 4.6. We also assume that each $F_j$ is bounded by $\eta r\sqrt{q}$ and 1-Lipschitz continuous (with respect to the $\ell_2$-norm).

We may approach each $x \mapsto F_j(W_j^\top x)$ by a function with $\gamma_1$-norm less than $\delta\eta r\sqrt{q}$ and uniform approximation $C(\alpha,s)\eta r\sqrt{q}\delta^{-1/(\alpha+(s-1)/2)}\log\delta$. This leads to a total approximation error of $kC(\alpha,s)G\eta r\sqrt{q}\delta^{-1/(\alpha+(s-1)/2)}\log\delta$.

For $\alpha \geqslant 1$, the estimation error is $kGr\sqrt{q}\eta\delta\sqrt{\log d}/\sqrt{n}$, with an overall bound which is equal to $C(\alpha, s)kGr\sqrt{q}\eta(\delta^{-1/(\alpha+(s-1)/2)}\log\delta+\frac{\delta\sqrt{\log d}}{\sqrt{n}})$. With $\delta = (n/\log d)^{(\alpha+(s-1)/2)/(2\alpha+s-1)}$, we get an optimized bound of $\dfrac{C(\alpha, s)kGr\sqrt{q}\eta(\log d)^{1/(2\alpha+s+1)}}{n^{1/(2\alpha+s+1)}}\log n$.

For $\alpha = 0$, we have the bound $\frac{C(s)kGr\sqrt{q}\eta}{(n/d)^{1/(s+1)}}\log(n/d)$, that is we cannot use the sparsity as the problem is invariant to the chosen norm on hidden weights.

We can make the following observations:

– *High-dimensional variable selection*: when $k = 1$, $s = q$ and $W_1$ is a projection onto $q$ variables, then we obtain a bound proportional to $\frac{\sqrt{q}\eta(\log d)^{1/(2\alpha+s+1)}}{n^{1/(2\alpha+s+1)}}\log n$, which exhibits a high-dimensional scaling in a non-linear setting. Note that beyond sparsity, no assumption is made (in particular regarding correlations between input variables), and we obtain a high-dimensional phenomenon where $d$ may be much larger than $n$.

– *Group penalties*: in this paper, we only consider $\ell_1$-norm on input weights; when doing joint variable selection for all basis functions, it may be worth using a group penalty (Yuan and Lin, 2006; Bach, 2008a).

### 5.4 Relationship to Kernel Methods and Random Sampling

The results presented in the two sections above were using the space $\mathcal{F}_1$, with an $L_1$-norm on the outputs weights (and either an $\ell_1$- or $\ell_2$-norm on input weights). As seen in Sections 2.3 and 3.1, when using an $L_2$-norm on output weights, we obtain a reproducing kernel Hilbert space $\mathcal{F}_2$.

As shown in Section 6, the space $\mathcal{F}_2$ is significantly smaller than $\mathcal{F}_1$, and in particular is not adaptive to low-dimensional linear structures, which is the main advantage of the space $\mathcal{F}_1$. However, algorithms for $\mathcal{F}_2$ are significantly more efficient, and there is no need for the conditional gradient algorithms presented in Section 2.5. The first possibility is to use the usual RKHS representer theorem with the kernel functions computed in Section 3.1, leading to a computation complexity of $O(n^2)$. Alternatively, as shown by Rahimi and Recht (2007), one may instead sample $m$ basis functions that is $m$ different hidden units, keep the input weights fixed and optimize only the output layer with a squared $\ell_2$-penalty. This will quickly (i.e., the error goes down as $1/\sqrt{m}$) approach the non-parametric estimator based on penalizing by the RKHS norm $\gamma_2$. Note that this argument of random sampling has been used to study approximation bounds for neural networks with finitely many units (Maiorov and Meir, 2000).

Given the usage of random sampling with $L_2$-penalties, it is thus tempting to sample weights, but now optimize an $\ell_1$-penalty, in order to get the non-parametric estimator obtained from penalizing by $\gamma_1$. When the number of samples $m$ tends to infinity, we indeed obtain an approximation that converges to $\gamma_1$ (this is simply a uniform version of the law of large numbers). However, the rate of convergence does depend on the dimension $d$, and in general exponentially many samples would be needed for a good approximation—see Bach (2013, Section 6) for a more precise statement in the related context of convex matrix factorizations.

### 5.5 Sufficient Condition for Polynomial-time Algorithms

In order to preserve the generalization bounds presented above, it is sufficient to be able to solve the following problem, for any $y \in \mathbb{R}^n$ and $z_1, \dots, z_n \in \mathbb{R}^{d+1}$:

$$\sup_{\|v\|_p=1} \left| \frac{1}{n} \sum_{i=1}^n y_i (v^\top z_i)_+^\alpha \right|, \tag{5}$$

*up to a constant factor*. That is, there exists $\kappa \geqslant 1$, such that for all $y$ and $z$, we may compute $\hat{v}$ such that $\|\hat{v}\|_p = 1$ and

$$\left| \frac{1}{n} \sum_{i=1}^n y_i (\hat{v}^\top z_i)_+^\alpha \right| \geqslant \frac{1}{\kappa} \sup_{\|v\|_p=1} \left| \frac{1}{n} \sum_{i=1}^n y_i (v^\top z_i)_+^\alpha \right|.$$

This is provably NP-hard for $\alpha = 0$ (see Section 3.2), and for $\alpha = 1$ (see Section 3.3). If such an algorithm is available, the approximate conditional gradient presented in Section 2.5 leads to an estimator with the same generalization bound. Moreover, given the strong hardness results for improper learning in the situation $\alpha = 0$ (Klivans and Sherstov, 2006; Livni et al., 2014), a convex relaxation that would consider a larger set of predictors (e.g., by relaxing $vv^\top$ into a symmetric positive-definite matrix), and obtained a constant approximation guarantee, is also ruled out.

However, this is only a sufficient condition, and a simpler sufficient condition may be obtained. In the following, we consider $\mathcal{V} = \{v \in \mathbb{R}^{d+1},\ \|v\|_2 = 1\}$ and basis functions $\varphi_v(z) = (v^\top z)_+^\alpha$ (that is we specialize to the $\ell_2$-norm penalty on weight vectors). We consider a new variation norm $\hat{\gamma}_1$ which has to satisfy the following assumptions:

- *Lower-bound on* $\gamma_1$: It is defined from functions $\hat{\varphi}_{\hat{v}}$, for $\hat{v} \in \hat{\mathcal{V}}$, where for any $v \in \mathcal{V}$, there exists $\hat{v} \in \hat{\mathcal{V}}$ such that $\varphi_v = \hat{\varphi}_{\hat{v}}$. This implies that the corresponding space $\hat{\mathcal{F}}_1$ is larger than $\mathcal{F}_1$ and that if $f \in \mathcal{F}_1$, then $\hat{\gamma}_1(f) \leqslant \gamma_1(f)$.

- *Polynomial-time algorithm for dual norm*: The dual norm $\sup_{\hat{v} \in \hat{\mathcal{V}}} \left| \frac{1}{n} \sum_{i=1}^n y_i \hat{\varphi}_{\hat{v}}(z_i) \right|$ may be computed in polynomial time.

- *Performance guarantees for random direction*: There exists $\kappa > 0$, such that for any vectors $z_1, \dots, z_n \in \mathbb{R}^{d+1}$ with $\ell_2$-norm less than $R$, and random standard Gaussian vector $y \in \mathbb{R}^n$,

  $$\sup_{\hat{v} \in \hat{\mathcal{V}}} \left| \frac{1}{n} \sum_{i=1}^n y_i \hat{\varphi}_{\hat{v}}(x_i) \right| \leqslant \kappa \frac{R}{\sqrt{n}}. \tag{6}$$

  We may also replace the standard Gaussian vectors by Rademacher random variables.

We can then penalize by $\hat{\gamma}$ instead of $\gamma$. Since $\hat{\gamma}_1 \leqslant \gamma_1$, approximation properties are transferred, and because of the result above, the Rademacher complexity for $\hat{\gamma}_1$-balls scales as well as for $\gamma_1$-balls. In the next section, we show convex relaxations which cannot achieve these and leave the existence or non-existence of such norm $\hat{\gamma}_1$ as an open problem.

## 6. Convex Relaxations of the Frank-Wolfe Step

In this section, we provide approximation algorithms for the following problem of maximizing, for a given $y \in \mathbb{R}^n$ and vectors $z_1, \dots, z_n$:

$$\sup_{\|v\|_p=1} \frac{1}{n} \sum_{i=1}^n y_i (v^\top z_i)_+^\alpha$$

These approximation algorithms may be divided in three families, as they may be based on (a) geometric interpretations as linear binary classification or computing Haussdorff distances (see Section 3.2 and Section 3.3), (b) on direct relaxations, on (c) relaxations of sign vectors. For simplicity, we only focus on the case $p = 2$ (that is $\ell_2$-constraint on weights) and on $\alpha = 1$ (rectified linear units). As described in Section 5.5, constant-factor approximation ratios are not possible, while approximation ratios that increases with $n$ are possible (but as of now, we only obtain scalings in $n$ that provide a provable sample complexity with a polynomial time algorithm which is exponential in the dimension $d$.

### 6.1 Semi-definite Programming Relaxations

We present two relaxations, which are of the form described in Section 5.5 (leading to potential generalization bounds) but do not attain the proper approximation scaling (as was checked empirically).

Note that all relaxations that end up being Lipschitz-continuous functions of $z$, will have at least the same scaling than the set of these functions. The Rademacher complexity of such functions is well-known, that is $1/\sqrt{n}$ for $d = 1$, $\sqrt{\frac{\log n}{n}}$ for $d = 2$ and $n^{-1/d}$ for larger $d$ (von Luxburg and Bousquet, 2004). Unfortunately, the decay in $n$ is too slow to preserve generalization bounds (which would require a scaling in $1/\sqrt{n}$).

**$d$-dimensional relaxation.** We denote $u_i = (v^\top z_i)_+ = \frac{1}{2} v^\top z_i + \frac{1}{2}|v^\top z_i|$. We may then use $2u_i - v^\top z_i = |v^\top z_i|$ and, for $\|v\|_2 = 1$, $\|vv^\top z_i\|_2 = |v^\top z_i| = \sqrt{z_i^\top vv^\top z_i}$. By denoting $V = vv^\top$, the constraint that $u_i = (v^\top z_i)_+ = \frac{1}{2} v^\top z_i + \frac{1}{2}|v^\top z_i|$ is equivalent to

$$\|Vz_i\|_2 \leqslant 2u_i - v^\top z_i \leqslant \sqrt{z_i^\top V z_i} \;\text{ and }\; V \succcurlyeq 0,\; \operatorname{tr} V = 1,\; \operatorname{rank}(V) = 1.$$

We obtain a convex relaxation when removing the rank constraint, that is

$$\sup_{V \succcurlyeq 0,\; \operatorname{tr} V = 1,\; u \in \mathbb{R}^n} u^\top y \;\text{ such that }\; \forall i \in \{1, \dots, n\},\; \|Vz_i\|_2 \leqslant 2u_i - v^\top z_i \leqslant \sqrt{z_i^\top V z_i}.$$

**$(n+d)$-dimensional relaxation.** We may go further by also considering quadratic forms in $u \in \mathbb{R}^n$ defined above. Indeed, we have:

$$(2u_i - v^\top z_i)(2u_j - v^\top z_j) = |v^\top z_i| \cdot |v^\top z_j| = |v^\top z_i z_j^\top v| = |\operatorname{tr} V z_i z_j^\top|,$$

which leads to a convex program in $U = uu^\top$, $V = vv^\top$ and $J = uv^\top$, that is a semidefinite program with $d + n$ dimensions, with the constraints

$$4U_{ij} + x_j^\top V z_i - 2\delta_i^\top J z_j - 2\delta_j^\top J z_i \geqslant |\operatorname{tr} V z_i z_j^\top|,$$

and the usual semi-definite contraints $\begin{pmatrix} U & J \\ J^\top & V \end{pmatrix} \succcurlyeq \begin{pmatrix} u \\ v \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix}^\top$, with the additional constraint that $4U_{ii} + z_i^\top V z_i - 4\delta_i^\top J z_i = \operatorname{tr} V z_i z_i^\top$.

If we add these constraints on top of the ones above, we obtain a tighter relaxation. Note that for this relaxation, we must have $[(2u_i - v^\top z_i) - (2u_j - v^\top z_j)]$ less than a constant times $\|z_i - z_j\|_2$. Hence, the result mentioned above regarding Lipschitz-continuous functions and the scaling of the upper-bound for random $y$ holds (with the dependence on $n$ which is not good enough to preserve the generalization bounds with a polynomial-time algorithm).

### 6.2 Relaxation of Sign Vectors

By introducing a sign vector $s \in \mathbb{R}^n$ such that $s_i \in \{-1, 1\}$ and $s_i v^\top x_i = |v^\top x_i|$, we have the following relaxation with $S = ss^\top$, $V = vv^\top$ and $J = sv^\top$:

- Usual semi-definite constraint: $\begin{pmatrix} S & J \\ J^\top & V \end{pmatrix} \succcurlyeq \begin{pmatrix} s \\ v \end{pmatrix}\begin{pmatrix} s \\ v \end{pmatrix}^\top$,
- Unit/trace constraints: $\operatorname{diag}(S) = 1$ and $\operatorname{tr} V = 1$,
- Sign constraint: $\delta_i^\top J x_i \geqslant \max_{j \neq i} |\delta_j^\top J x_i|$.
- Additional constraint: $(x_i^\top V x_i)^{1/2} \leqslant \delta_i^\top J x_i$.

We then need to maximize $\frac{1}{2n}\sum_{i=1}^n y_i \delta_i^\top J x_i + \frac{1}{2n}\sum_{i=1}^n y_i v^\top x_i$, which leads to a semidefinte program. Again empirically, it did not lead to the correct scaling as a function of $n$ for random Gaussian vectors $y \in \mathbb{R}^n$.

## 7. Conclusion

In this paper, we have provided a detailed analysis of the generalization properties of convex neural networks with positively homogenous non-decreasing activation functions. Our main new result is the adaptivity of the method to underlying linear structures such as the dependence on a low-dimensional subspace, a setting which includes non-linear variable selection in presence of potentially many input variables.

All our current results apply to estimators for which no polynomial-time algorithm is known to exist and we have proposed sufficient conditions under which convex relaxations could lead to the same bounds, leaving open the existence or non-existence of such algorithms. Interestingly, these problems have simple geometric interpretations, either as binary linear classification, or computing the Haussdorff distance between two zonotopes.

In this work, we have considered a single real-valued output; the functional analysis framework readily extends to outputs in a finite-dimensional vector-space where vector-valued measures could be used, and then apply to multi-task or multi-class problems. However, the extension to multiple hidden layers does not appear straightforward as the units of the last hidden layers share the weights of the first hidden layers, which should require a new functional analysis framework.

## Acknowledgements

The author was partially supported by the European Research Council (SIERRA Project), and thanks Nicolas Le Roux for helpful discussions. The author also thanks Varun Kanade for pointing the NP-hardness linear classification result. The main part of this work was carried through while visiting the Centre de Recerca Matemàtica (CRM) in Barcelona.

## Appendix A. Reproducing Kernel Hilbert Spaces for $\ell_2$-norm Penalization

In this section, we consider a Borel probability measure $\tau$ on the compact space $\mathcal{V}$, and functions $\varphi_v : \mathcal{X} \to \mathbb{R}$ such that the functions $v \mapsto \varphi_v(x)$ are measurable for all $x \in \mathcal{X}$. We study the set $\mathcal{F}_2$ of functions $f$ such that there exists a squared-integrable function $p : \mathcal{X} \to \mathbb{R}$ with $f(x) = \int_{\mathcal{V}} p(v)\varphi_v(x)d\tau(v)$ for all $x \in \mathcal{X}$. For $f \in \mathcal{F}_2$, we define $\gamma_2^2(f)$ as the infimum of $\int_{\mathcal{V}} p(v)^2 d\tau(v)$ over all decompositions of $f$. We now show that $\mathcal{F}_2$ is an RKHS with kernel $k(x,y) = \int_{\mathcal{V}} \varphi_v(x)\varphi_v(y)d\tau(v)$.

We follow the proof of Berlinet and Thomas-Agnan (2004, Section 4.1) and extend it to integrals rather than finite sums. We consider the linear mapping $T : L_2(d\tau) \to \mathcal{F}_2$ defined by $(Tp)(x) = \int_{\mathcal{V}} p(v)\varphi_v(x)d\tau(v)$, with null space $\mathcal{K}$. When restricted to the orthogonal complement $\mathcal{K}^\perp$, we obtain a bijection $U$ from $\mathcal{K}^\perp$ to $\mathcal{F}_2$. We then define a dot-product on $\mathcal{F}_2$ as $\langle f, g\rangle = \int_{\mathcal{V}} (U^{-1}f)(v)(U^{-1}g)(v)d\tau(v)$.

We first show that this defines an RKHS with kernel $k$ defined above. For this, we trivially have $k(\cdot, y) \in \mathcal{F}_2$ for all $y \in \mathcal{X}$. Moreover, for any $y \in \mathcal{X}$, we have with $p = U^{-1}k(\cdot, y) \in \mathcal{K}^\perp$ and $q : v \mapsto \varphi_v(y)$, $p - q \in \mathcal{K}$, which implies that $\langle f, k(\cdot, y)\rangle = \int_{\mathcal{V}} (U^{-1}f)(v)p(v)d\tau(v) = \int_{\mathcal{V}} (U^{-1}f)(v)q(v)d\tau(v) = \int_{\mathcal{V}} (U^{-1}f)(v)\varphi_v(y)d\tau(v) = T(U^{-1}f)(y) = f(y)$, hence the reproducing property is satisfied. Thus, $\mathcal{F}_2$ is an RKHS.

We now need to show that the RKHS norm which we have defined is actually $\gamma_2$. For any $f \in \mathcal{F}_2$ such that $f = Tp$, for $p \in L_2(d\tau)$, we have $p = U^{-1}f + q$, where $q \in \mathcal{K}$. Thus, $\int_{\mathcal{V}} p(v)^2 d\tau(v) = \|p\|^2_{L_2(d\tau)} = \|U^{-1}f\|^2_{L_2(d\tau)} + \|q\|^2_{L_2(d\tau)} = \|f\|^2 + \|q\|^2_{L_2(d\tau)}$. This implies that $\int_{\mathcal{V}} p(v)^2 d\tau(v) \geqslant \|f\|^2$ with equality if and only if $q = 0$. This shows that $\gamma_2(f) = \|f\|$.

## Appendix B. Approximate Conditional Gradient with Multiplicative Oracle

In this section, we wish to minimize a smooth convex functional $J(h)$ on for $h$ in a Hilbert-space over a norm ball $\{\gamma(h) \leqslant \delta\}$. A multiplicative approximate oracle outputs for any $g \in \mathbb{R}^n$, $\hat{h}$ such that $\gamma(\hat{h}) = 1$, and

$$\langle \hat{h}, g\rangle \leqslant \max_{\gamma(h)\leqslant 1} \langle h, g\rangle \leqslant \kappa \, \langle \hat{h}, g\rangle,$$

for a fixed $\kappa \geqslant 1$. We now propose a modification of the conditional gradient algorithm that converges to a certain $h$ such that $\gamma(h) \leqslant \delta$ and for which $\inf_{\gamma(h)\leqslant\delta} J(h) \leqslant J(\hat{h}) \leqslant \inf_{\gamma(h)\leqslant\delta/\kappa} J(h)$.

We assume the smoothness of the function $J$ with respect to the norm $\gamma$, that is, for a certain $L > 0$, for all $h, h'$ such that $\gamma(h) \leqslant \delta$, then

$$J(h') \leqslant J(h) + \langle J'(h), h' - h \rangle + \frac{L}{2}\gamma(h - h')^2. \tag{7}$$

We consider the following recursion

$$\begin{aligned} \hat{h}_t &= -\delta \times \text{ output of approximate oracle at } -J'(h_t) \\ h_{t+1} &\in \arg\min_{\rho\in[0,1]} J((1-\rho)h_t + \rho\hat{h}_t). \end{aligned}$$

In the previous recursion, one may replace the minimization of $J$ on the segment $[h_t, \hat{h}_t]$ with the minimization of its upper-bound of Eq. (7) taken at $h = h_t$. From the recursion, all iterates are in the $\gamma$-ball of radius $\delta$. Following the traditional convergence proof for the conditional gradient method (Dunn and Harshbarger, 1978; Jaggi, 2013), we have, for any $\rho$ in $[0, 1]$:

$$\begin{aligned} J(h_{t+1}) &\leqslant J(h_t) - \rho\langle J'(h_t), h_t - \hat{h}_t\rangle + 2L\rho^2\delta^2 \\ &= J(h_t) - \rho J'(h_t)^\top h_t + \kappa\langle J'(h_t), \frac{\hat{h}_t}{\kappa}\rangle + 2L\rho^2\delta^2 \\ &\leqslant J(h_t) - \rho J'(h_t)^\top h_t - \max_{\gamma(h)\leqslant\delta/\kappa} \{ -\langle J'(h_t), h\rangle\} + 2L\rho^2\delta^2. \end{aligned}$$

If we take $h_*$ the minimizer of $J$ on $\{\gamma(h) \leqslant \delta/\kappa\}$, we get:

$$J(h_{t+1}) \leqslant J(h_t) - \rho\langle J'(h_t), h_t - h_*\rangle + 2L\rho^2\delta^2.$$

Then, by using $J(h_t) \geqslant J(h_*) + \langle J'(h_t), h_* - h_t\rangle$, we get:

$$J(h_{t+1}) - J(h_*) \leqslant (1-\rho)[J(h_t) - J(h_*)] + 2L\rho^2\delta^2.$$

This is valid for any $\rho \in [0, 1]$. If $J(h_t) - J(h_*) \leqslant 0$ for some $t$, then by taking $\rho = 0$ it remains the same of all greater $t$. Therefore, up to (the potentially never happening) point where $J(h_t) - J(h_*) \leqslant 0$, we can apply the regular proof of the conditional gradien to obtain: $J(h_t) \leqslant \inf_{\gamma(h)\leqslant\delta/\kappa} J(h) + \frac{4L\rho^2\delta^2}{t}$, which leads to the desired result. Note that a similar reasoning may be used for $\rho = 2/(t + 1)$.

## Appendix C. Proofs for the 2-dimensional Sphere ($d = 1$)

In this section, we consider only the case $d = 1$, where the sphere $\mathbb{S}^d$ is isomorphic to $[0, 2\pi]$ (with periodic boundary conditions). We may then compute the norm $\gamma_2$ in closed form. Indeed, if we can decompose $g$ as $g(\theta) = \frac{1}{2\pi}\int_0^{2\pi} p(\varphi)\sigma(\cos(\varphi-\theta))d\varphi$, then the decomposition of $g$ into the $k$-th frequency elements (the combination of the two $k$-th elements of the Fourier series) is equal to, for $\sigma(u) = (u)_+^\alpha$, and for $k > 0$:

$$g_k(\theta) = \frac{1}{\pi}\int_0^{2\pi} g(\eta)\cos k(\theta - \eta)d\eta$$

$$
\begin{aligned}
&= \frac{1}{\pi}\int_0^{2\pi} \frac{1}{2\pi}\bigg(\int_0^{2\pi} p(\varphi)\sigma(\cos(\eta-\varphi))d\varphi\bigg)\cos k(\theta-\eta)d\eta \text{ through the decomposition of } g,\\
&= \frac{1}{2\pi^2}\int_0^{2\pi} p(\varphi)\bigg(\int_0^{2\pi} \sigma(\cos(\eta-\varphi))\cos k(\theta-\eta)d\eta\bigg)d\varphi\\
&= \frac{1}{2\pi^2}\int_0^{2\pi} p(\varphi)\bigg(\int_0^{2\pi} \sigma(\cos\eta)\cos k(\theta-\varphi-\eta)d\eta\bigg)d\varphi \text{ by a change of variable,}\\
&= \frac{1}{2\pi^2}\int_0^{2\pi} p(\varphi)\bigg(\cos k(\theta-\varphi)\int_0^{2\pi} \sigma(\cos\eta)\cos k\eta\, d\eta\\
&\qquad\qquad + \sin k(\theta-\varphi)\int_0^{2\pi}\sigma(\cos\eta)\sin k\eta\,d\eta\bigg)d\varphi \text{ by expanding the cosine,}\\
&= \bigg(\frac{1}{2\pi}\int_0^{2\pi}\sigma(\cos\eta)\cos k\eta\,d\eta\bigg)\bigg(\frac{1}{\pi}\int_0^{2\pi}p(\varphi)\cos k(\theta-\varphi)\bigg) + 0 \text{ by a parity argument,}\\
&= \lambda_k p_k(\theta) \text{ with } \lambda_k = \frac{1}{2\pi}\int_0^{2\pi}\sigma(\cos\eta)\cos k\eta\,d\eta.
\end{aligned}
$$

For $k=0$, the same equality holds (except that the two coefficients $g_0$ and $p_0$ are divided by $2\pi$ except of $\pi$).

Thus we may express $\|p\|^2_{L_2(\mathbb{S}^d)}$ as

$$
\begin{aligned}
\|p\|^2_{L_2(\mathbb{S}^d)} &= \sum_{k\geqslant 0}\|p_k\|^2_{L_2(\mathbb{S}^d)} = \sum_{\lambda_k\neq 0}\|p_k\|^2_{L_2(\mathbb{S}^d)} + \sum_{\lambda_k=0}\|p_k\|^2_{L_2(\mathbb{S}^d)}\\
&= \sum_{\lambda_k\neq 0}\frac{1}{\lambda_k^2}\|g_k\|^2_{L_2(\mathbb{S}^d)} + \sum_{\lambda_k=0}\|p_k\|^2_{L_2(\mathbb{S}^d)}.
\end{aligned}
$$

If we minimize over $p$, we thus need to have $\|p_k\|^2_{L_2(\mathbb{S}^d)}=0$ for $\lambda_k=0$, and we get

$$
\gamma_2(g)^2 = \sum_{\lambda_k\neq 0}\frac{1}{\lambda_k^2}\|g_k\|^2_{L_2(\mathbb{S}^d)}. \tag{8}
$$

We thus simply need to compute $\lambda_k$ and its decay for all values of $\alpha$, and then relate them to the smoothness properties of $g$, which is standard for Fourier series.

### C.1 Computing $\lambda_k$

We now detail the computation of $\lambda_k = \frac{1}{2\pi}\int_0^{2\pi}\sigma(\cos\eta)\cos k\eta\,d\eta$ for the different functions $\sigma=(\cdot)^\alpha_+$. We have for $\alpha=0$:

$$
\frac{1}{2\pi}\int_0^{2\pi}1_{\cos\eta\geqslant 0}\cos k\eta\,d\eta = \frac{1}{2\pi}\int_{-\pi/2}^{\pi/2}\cos k\eta\,d\eta = \frac{1}{\pi k}\sin\frac{k\pi}{2} \text{ if } k\neq 0.
$$

For $k=0$ it is equal to $\frac{1}{2}$. It is equal to zero for all other even $k$, and different from zero for all odd $k$, with $\lambda_k$ going to zero as $1/k$.

We have for $\alpha=1$:

$$
\frac{1}{2\pi}\int_0^{2\pi}(\cos\eta)_+\cos k\eta\,d\eta = \frac{1}{2\pi}\int_{-\pi/2}^{\pi/2}\cos\eta\cos k\eta\,d\eta
$$

$$\begin{aligned} &= \frac{1}{2\pi}\int_{-\pi/2}^{\pi/2} [\frac{1}{2}\cos(k+1)\eta + \frac{1}{2}\cos(k-1)\eta]\, d\eta \\ &= \frac{1}{4\pi}\bigg(\frac{2}{k+1}\sin(k+1)\frac{\pi}{2} + \frac{2}{k-1}\sin(k-1)\frac{\pi}{2}\bigg) \\ &= \frac{\cos\frac{k\pi}{2}}{2\pi}\bigg(\frac{1}{k+1} - \frac{1}{k-1}\bigg) = \frac{-\cos\frac{k\pi}{2}}{\pi(k^2-1)} \text{ for } k \neq 1. \end{aligned}$$

For $k = 1$, it is equal to $1/4$. It is equal to zero for all other odd $k$, and different from zero for all even $k$, with $\lambda_k$ going to zero as $1/k^2$.

For $\alpha = 2$, we have:

$$\begin{aligned} \frac{1}{2\pi}\int_0^{2\pi}(\cos\eta)_+^2\cos k\eta\, d\eta &= \frac{1}{2\pi}\int_{-\pi/2}^{\pi/2}(\cos\eta)^2\cos k\eta\, d\eta = \frac{1}{2\pi}\int_{-\pi/2}^{\pi/2}\frac{1+\cos 2\eta}{2}\cos k\eta\, d\eta \\ &= \frac{1}{2\pi}\int_{-\pi/2}^{\pi/2}[\frac{1}{2}\cos k\eta + \frac{1}{4}\cos(k+2)\eta + \frac{1}{4}\cos(k-2)\eta]\, d\eta \\ &= \frac{1}{4\pi}\bigg(\frac{2}{k}\sin k\frac{\pi}{2} + \frac{1}{k+2}\sin(k+2)\frac{\pi}{2} + \frac{1}{k-2}\sin(k-2)\frac{\pi}{2}\bigg) \\ &= \frac{\sin(k\frac{\pi}{2})}{4\pi}\bigg(\frac{2}{k} - \frac{1}{k+2} - \frac{1}{k-2}\bigg) \\ &= \frac{\sin(k\frac{\pi}{2})}{4\pi}\bigg(\frac{2k^2-8-k^2+2k-k^2-2k}{k(k^2-4)}\bigg) \\ &= \frac{-8\sin(k\frac{\pi}{2})}{4\pi k(k^2-4)} \text{ for } k \notin \{0,2\}. \end{aligned}$$

For $k = 0$, it is equal to $1/4$, and for $k = 2$, it is equal to $1/8$. It is equal to zero for all other even $k$, and different from zero for all odd $k$, with $\lambda_k$ going to zero as $1/k^3$.

The general case for $\alpha \geqslant 2$ will be shown for for all $d$ in Appendix D.2: for all $\alpha \in \mathbb{N}$, $\lambda_k$ is different from zero for $k$ having the opposite parity of $\alpha$, with a decay as $1/k^{\alpha+1}$. All values from $k = 0$ to $\alpha$ are also different from zero. All larger values with the same parity as $\alpha$ are equal to zero.

### C.2 Proof of Prop. 2 for $d = 1$

We only consider the proof for $d = 1$. For the proof for general $d$, see Appendix D.3.

Given the zero values of $\lambda_k$ given above, if $g$ has the opposite parity than $\alpha$ (that is, is even when $\alpha$ is odd, and vice-versa), then we may define $p$ through its Fourier series, which is obtained by multiplying the one of $g$ by a strictly positive sequence growing as $k^{\alpha+1}$.

Thus, if $g$ is such that its $(\alpha+1)$-th order derivative is squared-integrable, then $p$ defined above is squared-integrable, that is, $g \in \mathcal{G}_2$. Moreover, if all derivatives of order less than $(\alpha + 1)$ are bounded by $\eta$, $p$ is squared-integrable and $\|p\|_{L_2(\mathbb{S}^d)}^2$ is upper-bounded by a constant times $\eta^2$, i.e., $\gamma_2(g)^2 \leqslant C(\alpha)^2\eta^2$.

Note that we could relax the assumption that $g$ is even (resp. odd) by adding all trigonometric polynomials of order less than $\alpha$.

### C.3 Proof of Prop. 3 for $d = 1$

Again, we only consider the proof for $d = 1$. For the proof for general $d$, see Appendix D.4.

Without loss of generality, we assume that $\eta = 1$. For $d = 1$, we essentially want to approximate a Lipschitz-continuous function by a function which is $(\alpha + 1)$-times differentiable.

For $\alpha = 0$, then the function $g$ is already in $\mathcal{G}_2$ with a norm less than one, because Lipschitz-continuous functions are almost everywhere differentiable with bounded derivative (Adams and Fournier, 2003). We thus now consider $\alpha > 0$.

Given $\lambda_k$ defined above and $r \in (0, 1)$, we define $\hat{p}$ through

$$\hat{p}_k(\theta) = \sum_{k, \lambda_k \neq 0} \lambda_k^{-1} g_k(\theta) r^k.$$

Our goal is to show that for $r$ chosen close enough to 1, then the function $\hat{g}$ defined from $\hat{p}$ has small enough norm $\gamma_2(\hat{g}) \leqslant \|\hat{p}\|_{L_2(\mathbb{S}^d)}$, and is close to $g$.

**Computation of norm.** We have

$$\|\hat{p}\|^2_{L_2(\mathbb{S}^d)} \quad = \sum_{k, \lambda_k \neq 0} \lambda_k^{-2} r^{2k} \|g_k\|^2_{L_2(\mathbb{S}^d)}.$$

Since $g$ is 1-Lipschitz-continuous with constant 1, then it has a squared-integrable derivative $f = g'$ with norm less than 1 (Adams and Fournier, 2003), so that

$$\|f\|^2_{L_2(\mathbb{S}^d)} = \sum_{k \geqslant 0} \|f_k\|^2_{L_2(\mathbb{S}^d)} \leqslant 1.$$

This implies that using $\lambda_k^{-1} = O(k^{\alpha+1})$:

$$\|\hat{p}\|^2_{L_2(\mathbb{S}^d)} \leqslant \lambda_0^{-2} \|g_0\|^2_{L_2(\mathbb{S}^d)} + \|g'\|^2_{L_2(\mathbb{S}^d)} \max_{k \geqslant 1, \lambda_k \neq 0} r^{2k} \lambda_k^{-2} k^{-2} \leqslant C + C \|g'\|^2_{L_2(\mathbb{S}^d)} \max_{k \geqslant 1} r^{2k} k^{2\alpha},$$

because $\|g_0\|^2_{L_2(\mathbb{S}^d)}$ and $\|f\|^2_{L_2(\mathbb{S}^d)}$ are bounded by 1.

We may now compute the derivative of $k \mapsto r^{2k} k^{2\alpha}$ with respect to $k$ (now considered a real number), that is $2\alpha k^{2\alpha-1} r^{2k} + k^{2\alpha} r^{2k} 2 \log r$, which is equal to zero for $\frac{\alpha}{k} = \log \frac{1}{r}$, that is $k = \frac{\alpha}{\log \frac{1}{r}}$, the maximum being then $e^{-2\alpha} (\frac{\alpha}{\log \frac{1}{r}})^{2\alpha} = O((1-r)^{-2\alpha})$, by using the concavity of the logarithm. Thus $\|\hat{p}\|_{L_2(\mathbb{S}^d)} \leqslant C(1-r)^{-\alpha}$. This defines $\hat{g}$ with $\gamma(\hat{g}) \leqslant C(1-r)^{-\alpha}$.

**Computing distance between $\hat{g}$ and $g$.** We have:

$$\begin{aligned}
\hat{g}(\theta) &= \sum_{k \geqslant 0} g_k(\theta) r^k = \sum_{k > 0} \frac{1}{\pi} \int_0^{2\pi} g(\eta) r^k \cos k(\theta - \eta) d\eta + \frac{1}{2\pi} \int_0^{2\pi} g(\eta) d\eta \\
&= \frac{1}{\pi} \int_0^{2\pi} \bigg( \sum_{k \geqslant 0} r^k \cos k(\theta - \eta) \bigg) g(\eta) d\eta - \frac{1}{2\pi} \int_0^{2\pi} g(\eta) d\eta \\
&= \frac{1}{\pi} \int_0^{2\pi} \text{Real} \bigg( \frac{1}{1 - r e^{i(\theta - \eta)}} \bigg) g(\eta) d\eta - \frac{1}{2\pi} \int_0^{2\pi} g(\eta) d\eta
\end{aligned}$$

$$\begin{aligned}
&= \frac{1}{\pi}\int_0^{2\pi}\left(\frac{1-r\cos(\theta-\eta)}{(1-r\cos(\theta-\eta))^2+r^2(\sin(\theta-\eta))^2}\right)g(\eta)d\eta-\frac{1}{2\pi}\int_0^{2\pi}g(\eta)d\eta\\
&= \frac{1}{\pi}\int_0^{2\pi}\left(\frac{1-r\cos(\theta-\eta)}{1+r^2-2r\cos(\theta-\eta)}\right)g(\eta)d\eta-\frac{1}{2\pi}\int_0^{2\pi}g(\eta)d\eta\\
&= \frac{1}{2\pi}\int_0^{2\pi}\left(\frac{1-r^2+1+r^2-2r\cos(\theta-\eta)}{1+r^2-2r\cos(\theta-\eta)}\right)g(\eta)d\eta-\frac{1}{2\pi}\int_0^{2\pi}g(\eta)d\eta\\
&= \frac{1}{2\pi}\int_0^{2\pi}\left(\frac{1-r^2}{1+r^2-2r\cos(\theta-\eta)}\right)g(\eta)d\eta.
\end{aligned}$$

We have, for any $\theta \in [0, 2\pi]$

$$\begin{aligned}
|\hat{g}(\theta)-g(\theta)| &= \left|\frac{1}{2\pi}\int_0^{2\pi}\left(\frac{1-r^2}{1+r^2-2r\cos(\theta-\eta)}\right)[g(\eta)-g(\theta)]d\eta\right|\\
&\leqslant \frac{1}{2\pi}\int_0^{2\pi}\left(\frac{1-r^2}{1+r^2-2r\cos(\theta-\eta)}\right)|g(\eta)-g(\theta)|d\eta\\
&= \frac{1}{2\pi}\int_{\pi}^{\pi}\left(\frac{1-r^2}{1+r^2-2r\cos\eta}\right)|g(\theta)-g(\theta+\eta)|d\eta \text{ by periodicity,}\\
&= \frac{1}{\pi}\int_{\pi/2}^{\pi/2}\left(\frac{1-r^2}{1+r^2-2r\cos\eta}\right)|g(\theta)-g(\theta+\eta)|d\eta \text{ by parity of } g,\\
&\leqslant \frac{1}{\pi}\int_{\pi/2}^{\pi/2}\left(\frac{1-r^2}{1+r^2-2r\cos\eta}\right)\sqrt{2}|\sin\eta|d\eta\\
&\qquad\qquad \text{because the distance on the sphere is bounded by the sine,}\\
&\leqslant \frac{2}{\pi}\int_0^{\pi}\left(\frac{1-r^2}{1+r^2-2r\cos\eta}\right)\sin\eta\, d\eta\\
&= \frac{1}{\pi}\int_0^{1}\left(\frac{1-r^2}{1+r^2-2rt}\right)dt \text{ by the change of variable } t=\cos\theta,\\
&\leqslant C(1-r)\int_0^{1}\left(\frac{1}{1+r^2-2rt}\right)dt\\
&= C(1-r)\left[\frac{-1}{2r}\log(1+r^2-2rt)\right]_0^1 = C(1-r)\frac{1}{2r}\log\frac{1+r^2}{(1-r)^2}.
\end{aligned}$$

It can be easily checked that for any $r \in (1/2, 1)$, the last function is less than a constant times $\frac{5}{2}C(1-r)\log\frac{1}{1-r}$. We thus get for $\delta$ large enough, by taking $r = 1-(C/\delta)^{1/\alpha} \in (1/2, 1)$, an error of

$$(C/\delta)^{1/\alpha}\log(C/\delta)^{-1/\alpha} = O(\delta^{-1/\alpha}\log\delta).$$

This leads to the desired result.

## Appendix D. Approximations on the $d$-dimensional Sphere

In this appendix, we first review tools from spherical harmonic analysis, before proving the two main propositions regarding the approximation properties of the Hilbert space $\mathcal{G}_2$. Using spherical harmonics in our set-up is natural and is common in the analysis of ridge functions (Petrushev, 1998) and zonotopes (Bourgain and Lindenstrauss, 1988).

### D.1 Review of Spherical Harmonics Theory

In this section, we review relevant concepts from spherical harmonics. See Frye and Efthimiou (2012); Atkinson and Han (2012) for more details. Spherical harmonics may be seen as extension of Fourier series to spheres in dimensions more than 2 (i.e., with our convention $d \geqslant 1$).

For $d \geqslant 1$, we consider the sphere $\mathbb{S}^d = \{x \in \mathbb{R}^{d+1},\ \|x\|_2 = 1\} \subset \mathbb{R}^{d+1}$, as well as its normalized rotation-invariant measure $\tau_d$ (with mass 1). We denote by $\omega_d = \frac{2\pi^{(d+1)/2}}{\Gamma((d+1)/2)}$ the surface area of the sphere $\mathbb{S}^d$.

**Definition and links with Laplace-Beltrami operator.** For any $k \geqslant 1$ (for $k = 0$, the constant function is the corresponding basis element), there is an orthonormal basis of spherical harmonics, $Y_{kj} : \mathbb{S}^d \to \mathbb{R}$, $1 \leqslant j \leqslant N(d,k) = \frac{2k+d-1}{k}\binom{k+d-2}{d-1}$. They are such $\langle Y_{ki}, Y_{si}\rangle_{\mathbb{S}^d} = \int_{\mathbb{S}^d} Y_{ki}(x) Y_{sj} d\tau_d(x) = \delta_{ij}\delta_{sk}$.

Each of these harmonics may be obtained from homogeneous polynomials in $\mathbb{R}^d$ with an Euclidean Laplacian equal to zero, that is, if we define a function $H_k(y) = Y_{ki}(y/\|y\|_2)\|y\|_2^k$ for $y \in \mathbb{R}^{d+1}$, then $H_k$ is a homogeneous polynomial of degree $k$ with zero Laplacian. From the relationship between the Laplacian in $\mathbb{R}^{d+1}$ and the Laplace-Beltrami operator $\Delta$ on $\mathbb{S}^d$, $Y_{ki}$ is an eigenfunction of $\Delta$ with eigenvalue $-k(k+d-1)$. Like in Euclidean spaces, the Laplace-Beltrami operator may be used to characterize differentiability of functions defined on the sphere (Frye and Efthimiou, 2012; Atkinson and Han, 2012).

**Legendre polynomials.** We have the addition formula

$$\sum_{j=1}^{N(d,k)} Y_{kj}(x) Y_{kj}(y) = N(d,k) P_k(x^\top y),$$

where $P_k$ is a Legendre polynomial of degree $k$ and dimension $d+1$, defined as (Rodrigues' formula):

$$P_k(t) = (-1/2)^k \frac{\Gamma(d/2)}{\Gamma(k+d/2)} (1-t^2)^{(2-d)/2} \Big(\frac{d}{dt}\Big)^k (1-t^2)^{k+(d-2)/2}.$$

They are also referred to as Gegenbauer polynomials. For $d = 1$, $P_k$ is the $k$-th Chebyshev polynomial, such that $P_k(\cos\theta) = \cos(k\theta)$ for all $\theta$ (and we thus recover the Fourier series framework of Appendix C). For $d = 2$, $P_k$ is the usual Legendre polynomial.

The polynomial $P_k$ is even (resp. odd) when $k$ is even (resp. odd), and we have

$$\int_{-1}^{1} P_k(t) P_j(k) (1-t^2)^{(d-2)/2} dt = \delta_{jk} \frac{\omega_d}{\omega_{d-1}} \frac{1}{N(d,k)}.$$

For small $k$, we have $P_0(t) = 1$, $P_1(t) = t$, and $P_2(t) = \frac{(d+1)t^2 - 1}{d}$.

The Hecke-Funk formula leads to, for any linear combination $Y_k$ of $Y_{kj}$, $j \in \{1, \dots, N(d,k)\}$:

$$\int_{\mathbb{S}^d} f(x^\top y) Y_k(y) d\tau_d(y) = \frac{\omega_{d-1}}{\omega_d} Y_k(x) \int_{-1}^{1} f(t) P_k(t) (1-t^2)^{(d-2)/2} dt.$$

**Decomposition of functions in $L_2(\mathbb{S}^d)$.** Any function $g : \mathbb{S}^d \to \mathbb{R}$, such that we have $\int_{\mathbb{S}^d} g(x) d\tau_d(x) = 0$ may be decomposed as

$$\begin{aligned} g(x) &= \sum_{k=1}^{\infty} \sum_{j=1}^{N(d,k)} \langle Y_{kj}, g \rangle Y_{kj}(x) = \sum_{k=1}^{\infty} \sum_{j=1}^{N(d,k)} \int_{\mathbb{S}^d} Y_{kj}(y) Y_{kj}(x) g(y) d\tau_d(y) \\ &= \sum_{k=1}^{\infty} g_k(x) \text{ with } g_k(x) = N(d,k) \int_{\mathbb{S}^d} g(y) P_k(x^\top y) d\tau_d(y). \end{aligned}$$

This is the decomposition in harmonics of degree $k$. Note that

$$g_1(x) = x^\top \Big[ d \int_{\mathbb{S}^d} y g(y) d\tau_d(y) \Big]$$

is the linear part of $g$ (i.e., if $g(x) = w^\top x$, $g_1 = g$). Moreover, if $g$ does not have zero mean, we may define $g_0(x) = \int_{\mathbb{S}^d} g(y) d\tau_d(y)$ as the average value of $g$. Since the harmonics of different degrees are orthogonal to each other, we have the Parseval formula:

$$\|g\|_{\mathbb{S}^d}^2 = \sum_{k \geqslant 0} \|g_k\|_{\mathbb{S}^d}^2.$$

**Decomposition of functions of one-dimensional projections.** If $g(x) = \varphi(x^\top v)$ for $v \in \mathbb{S}^d$ and $\varphi : [-1,1] \to \mathbb{R}$, then

$$\begin{aligned} g_k(x) &= N(d,k) \int_{\mathbb{S}^d} \varphi(v^\top y) P_k(x^\top y) d\tau(y) \\ &= N(d,k) \frac{\omega_{d-1}}{\omega_d} P_k(x^\top v) \int_{-1}^{1} \varphi(t) P_k(t) (1-t^2)^{(d-2)/2} dt \\ &= \Big( \frac{\omega_{d-1}}{\omega_d} P_k(x^\top v) \int_{-1}^{1} \varphi(t) P_k(t) (1-t^2)^{(d-2)/2} dt \Big) \sum_{j=1}^{N(d,k)} Y_{kj}(x) Y_{kj}(y), \end{aligned}$$

and

$$\begin{aligned} \|g_k\|_{L_2(\mathbb{S}^d)}^2 &= \Big( \frac{\omega_{d-1}}{\omega_d} P_k(x^\top v) \int_{-1}^{1} \varphi(t) P_k(t) (1-t^2)^{(d-2)/2} dt \Big)^2 \sum_{j=1}^{N(d,k)} Y_{kj}(y)^2 \\ &= \Big( \frac{\omega_{d-1}}{\omega_d} P_k(x^\top v) \int_{-1}^{1} \varphi(t) P_k(t) (1-t^2)^{(d-2)/2} dt \Big)^2 N(d,k) P_k(1) \\ &= \Big( \frac{\omega_{d-1}}{\omega_d} P_k(x^\top v) \int_{-1}^{1} \varphi(t) P_k(t) (1-t^2)^{(d-2)/2} dt \Big)^2 N(d,k). \end{aligned}$$

### D.2 Computing the RKHS Norm $\gamma_2$

Like for the case $d = 1$, we may compute the RKHS norm $\gamma_2$ of a function $g$ in closed form given its decomposition in the basis of spherical harmonics $g = \sum_{k \geqslant 0} g_k$. If we can decompose $g(x) = \int_{\mathbb{S}^d} p(w)\sigma(w^\top x) d\tau_d(w)$ for a certain function $p : \mathbb{S}^d \to \mathbb{R}$, then we have, for $k \geqslant 0$:

$$\begin{aligned}
g_k(x) &= N(d,k) \int_{\mathbb{S}^d} g(y) P_k(x^\top y) d\tau_d(y) \\
&= N(d,k) \int_{\mathbb{S}^d} \int_{\mathbb{S}^d} p(w)\sigma(w^\top y) P_k(x^\top y) d\tau_d(y) d\tau_d(w) \\
&= N(d,k) \int_{\mathbb{S}^d} p(w) \bigg( \int_{\mathbb{S}^d} \sigma(w^\top y) P_k(x^\top y) d\tau_d(y) \bigg) d\tau_d(w) \\
&= \frac{\omega_{d-1}}{\omega_d} N(d,k) \int_{\mathbb{S}^d} p(w) P_k(x^\top w) \bigg( \int_{-1}^1 \sigma(t) P_k(t) (1-t^2)^{(d-2)/2} dt \bigg) d\tau_d(w) \\
& \qquad \qquad \qquad \qquad \qquad \qquad \qquad \qquad \text{using the Hecke-Funk formula,} \\
&= \lambda_k p_k(x) \text{ with } \lambda_k = \frac{\omega_{d-1}}{\omega_d} \int_{-1}^1 \sigma(t) P_k(t) (1-t^2)^{(d-2)/2} dt.
\end{aligned}$$

If $k \equiv \alpha$ mod. 2, then $\lambda_k \propto \frac{1}{2}\int_{-1}^1 t^\alpha P_k(t)(1-t^2)^{(d-2)/2} dt = 0$, for $k > \alpha$ since $P_k$ is orthogonal to all polynomials of degree strictly less than $k$ for that dot-product. Otherwise, $\lambda_k \neq 0$, since $t^\alpha$ may be decomposed as combination with non-zero coefficients of polynomials $P_j$ for $j \equiv \alpha$ mod. 2, $j \leqslant \alpha$.

We now provide an explicit formula extending the proof technique (for $\alpha = 1$) of Schneider (1967) and Bourgain and Lindenstrauss (1988) to all values of $\alpha$. See also Mhaskar (2006).

We have, by $\alpha$ successive integration by parts, for $k \geqslant \alpha + 1$:

$$\begin{aligned}
& \int_0^1 t^\alpha \Big(\frac{d}{dt}\Big)^k (1-t^2)^{k+(d-2)/2} dt \\
= \ & (-1)^\alpha \alpha! \int_0^1 \Big(\frac{d}{dt}\Big)^{k-\alpha} (1-t^2)^{k+(d-2)/2} dt = -(-1)^\alpha \alpha! \Big(\frac{d}{dt}\Big)^{k-\alpha-1} (1-t^2)^{k+(d-2)/2} \bigg|_{t=0} \\
= \ & -(-1)^\alpha \alpha! \Big(\frac{d}{dt}\Big)^{k-\alpha-1} \sum_{j \geqslant 0} \binom{k+(d-2)/2}{j} (-1)^j t^{2j} \bigg|_{t=0} \text{ using the binomial formula,} \\
= \ & -(-1)^\alpha \alpha! \Big(\frac{d}{dt}\Big)^{k-\alpha-1} \binom{k+(d-2)/2}{j} (-1)^j t^{2j} \bigg|_{t=0} \text{ for } 2j = k-\alpha-1, \\
= \ & -(-1)^\alpha \alpha! \binom{k+(d-2)/2}{j} (-1)^j (2j)! \text{ for } 2j = k-\alpha-1.
\end{aligned}$$

Thus

$$\begin{aligned}
\lambda_k &= -\frac{\omega_{d-1}}{\omega_d} (-1/2)^k \frac{\Gamma(d/2)}{\Gamma(k+d/2)} (-1)^\alpha \alpha! \binom{k+(d-2)/2}{j} (-1)^j (2j)! \text{ for } 2j = k-\alpha-1, \\
&= -\frac{\omega_{d-1}}{\omega_d} (-1/2)^k \frac{\Gamma(d/2)}{\Gamma(k+d/2)} (-1)^\alpha \alpha! \frac{\Gamma(k+\frac{d}{2})}{\Gamma(j+1)\Gamma(k+\frac{d}{2}-j)} (-1)^j \Gamma(2j+1) \\
&= -\frac{\omega_{d-1}}{\omega_d} (-1/2)^k \frac{\Gamma(d/2)}{\Gamma(k+d/2)} (-1)^\alpha \alpha! \frac{\Gamma(k+\frac{d}{2})(-1)^{(k-\alpha-1)/2}\Gamma(k-\alpha)}{\Gamma(\frac{k}{2}-\frac{\alpha}{2}+\frac{1}{2})\Gamma(\frac{k}{2}+\frac{d}{2}+\frac{\alpha}{2}+\frac{1}{2})} \\
&= \frac{d-1}{2\pi} \frac{\alpha!(-1)^{(k-\alpha-1)/2}}{2^k} \frac{\Gamma(d/2)\Gamma(k-\alpha)}{\Gamma(\frac{k}{2}-\frac{\alpha}{2}+\frac{1}{2})\Gamma(\frac{k}{2}+\frac{d}{2}+\frac{\alpha}{2}+\frac{1}{2})}.
\end{aligned}$$

By using Stirling formula $\Gamma(x) \approx x^{x-1/2}e^{-x}\sqrt{2\pi}$, we get an equivalent when $k$ or $d$ tends to infinity as a constant (that depends on $\alpha$) times

$$d^{d/2+1/2}k^{k/2-\alpha/2+1/2}(k+d)^{-k/2-d/2-\alpha/2}.$$

Note that all exponential terms cancel out. Moreover, when $k$ tends to infinity and $d$ is considered constant, then we get the equivalent $k^{-d/2-\alpha-1/2}$, which we need for the following sections. Finally, when $d$ tends to infinity and $k$ is considered constant, then we get the equivalent $d^{-\alpha/2-k/2+1/2}$.

We will also need expressions of $\lambda_k$ for $k=0$ and $k=1$. For $k=0$, we have:

$$\begin{aligned}\int_0^1 t^\alpha(1-t^2)^{d/2-1}dt &= \int_0^1 (1-u)^{\alpha/2}u^{d/2-1}\frac{du}{2\sqrt{1-u}} \text{ with } t=\sqrt{1-u},\\ &= \frac{1}{2}\int_0^1(1-u)^{\alpha/2+1/2-1}u^{d/2-1}du = \frac{1}{2}\frac{\Gamma(\alpha/2+1/2)\Gamma(d/2)}{\Gamma(\alpha/2+1/2+d/2)},\end{aligned}$$

using the normalization factor of the Beta distribution. This leads to

$$\lambda_0 \;=\; \frac{\omega_{d-1}}{\omega_d}\frac{1}{2}\frac{\Gamma(\alpha/2+1/2)\Gamma(d/2)}{\Gamma(\alpha/2+1/2+d/2)} = \frac{d-1}{2\pi}\frac{1}{2}\frac{\Gamma(\alpha/2+1/2)\Gamma(d/2)}{\Gamma(\alpha/2+1/2+d/2)},$$

which is equivalent to $d^{1/2-\alpha/2}$ as $d$ tends to infinity.

Moreover, for $k=1$, we have (for $\alpha>0$):

$$\begin{aligned}\int_0^1 t^\alpha\Big(\frac{d}{dt}\Big)(1-t^2)^{d/2}dt &= -\alpha\int_0^1 t^{\alpha-1}(1-t^2)^{d/2}dt = -\alpha\int_0^1(1-u)^{\alpha/2-1/2}u^{d/2}\frac{du}{2\sqrt{1-u}}\\ &= -\alpha/2\int_0^1(1-u)^{\alpha/2-1}u^{d/2+1-1}du = -\alpha/2\frac{\Gamma(\alpha/2)\Gamma(d/2+1)}{\Gamma(\alpha/2+d/2+1)}.\end{aligned}$$

This leads to, for $\alpha>0$:

$$\lambda_1 \;=\; (-1/2)\frac{2}{d}\frac{d-1}{2\pi}(-\alpha/2)\frac{\Gamma(\alpha/2)\Gamma(d/2+1)}{\Gamma(\alpha/2+d/2+1)} = \frac{d-1}{d}\frac{\alpha}{4\pi}\frac{\Gamma(\alpha/2)\Gamma(d/2+1)}{\Gamma(\alpha/2+d/2+1)},$$

which is equivalent to $d^{-\alpha/2}$ as $d$ tends to infinity.

Finally, for $\alpha=0$, $\lambda_1=\frac{d-1}{2d\pi}$. More generally, we have $|\lambda_k|\sim C(d)k^{-(d-1)/2-\alpha-1}$.

**Computing the RKHS norm.** Given $g$ with the correct parity, then we have

$$\gamma_2(g)^2 = \sum_{k\geqslant 0}\lambda_k^{-2}\|g_k\|^2_{L_2(\mathbb{S}^d)}.$$

### D.3 Proof of Prop. 2 for $d>1$

Given the expression of $\lambda_k$ from the section above, the proof is essentially the same than for $d=1$ in Appendix C.3. If $g$ is $s$-times differentiable with all derivatives bounded uniformly by $\eta$, then is equal to $g=\Delta^{s/2}f$ for a certain function $f$ such that $\|f\|_{L_2(\mathbb{S}^d)}\leqslant\eta$ (where $\Delta$ is the Laplacian on the sphere) (Frye and Efthimiou, 2012; Atkinson and Han, 2012).

Moreover, since $g$ has the correct parity,

$$\gamma_2(g)^2 \leqslant \|p\|^2_{L_2(\mathbb{S}^d)} \leqslant \sum_{k \geqslant 1, \lambda_k \neq 0} \lambda_k^{-2} \|g_k\|^2_{L_2(\mathbb{S}^d)}$$

Also, $g_k$ are eigenfunctions of the Laplacian with eigenvalues $k(k+d-1)$. Thus, we have

$$\|g_k\|_2^2 \leqslant \|f_k\|^2_{L_2(\mathbb{S}^d)} \frac{1}{[k(k+d-1)]^s} \leqslant \|f_k\|^2_{L_2(\mathbb{S}^d)} / k^{2s},$$

leading to $\gamma_2(g)^2 \leqslant \max_{k \geqslant 2} \lambda_k^{-2} k^{-2s} \|f\|^2_{L_2(\mathbb{S}^d)} \leqslant \max_{k \geqslant 2} k^{d-1+2\alpha+2} k^{-2s} \|f\|^2_{L_2(\mathbb{S}^d)} \leqslant C(d)\eta^2$, if $s \geqslant (d-1)/2 + \alpha + 1$, which is the desired result.

### D.4 Proof of Prop. 3 for $d > 1$

Without loss of generality we assume that $\eta = 1$, and we follow the same proof as for $d = 1$ in Appendix C.3. We have assumed that for all $x, y \in \mathbb{S}^d$, $|g(x) - g(y)| \leqslant \eta \|x - y\|_2 = \eta\sqrt{2}\sqrt{1 - x^\top y}$. Given the decomposition in the $k$-th harmonics, with

$$g_k(x) \quad = \quad N(d,k) \int_{\mathbb{S}^d} g(y) P_k(x^\top y) d\tau_d(y),$$

we may now define, for $r < 1$:

$$\hat{p}(x) \quad = \quad \sum_{k, \lambda_k \neq 0} \lambda_k^{-1} r^k g_k(x),$$

which is always defined when $r \in (0,1)$ because the series is absolutely convergent. This defines a function $\hat{g}$ that will have a finite $\gamma_2$-norm and be close to $g$.

**Computing the norm.** Given our assumption regarding the Lipschitz-continuity of $g$, we have $g = \Delta^{1/2} f$ with $f \in L_2(\mathbb{S}^d)$ with norm less than 1 (Atkinson and Han, 2012). Moreover $\|g_k\|^2_{L_2(\mathbb{S}^d)} \leqslant C k^2 \|f_k\|^2_{L_2(\mathbb{S}^d)}$. We have

$$\begin{aligned} \|\hat{p}\|^2_{L_2(\mathbb{S}^d)} &= \sum_{k, \lambda_k \neq 0} \lambda_k^{-2} r^{2k} \|g_k\|^2_{L_2(\mathbb{S}^d)} \\ &\leqslant C(d,\alpha) \max_{k \geqslant 0} k^{d-1+2\alpha} r^{2k} \|f\|^2_{L_2(\mathbb{S}^d)} \text{ because } \lambda_k = \Omega(k^{-d/2-\alpha-1/2}), \\ &\leqslant C(d,\alpha)(1-r)^{-d+1-2\alpha} \text{ (see Appendix C.3).} \end{aligned}$$

The function $\hat{p}$ thus defines a function $\hat{g} \in \mathcal{G}_1$ by $\hat{g}_k = \lambda_k p_k$, for which $\gamma_2(g) \leqslant C(d,\alpha)(1-r)^{(-d+1)/2-\alpha}$.

**Approximation properties.** We now show that $g$ and $\hat{g}$ are close to each other. Because of the parity of $g$, we have $\hat{g}_k = r^k g_k$. We have, using Theorem 4.28 from Frye and Efthimiou (2012):

$$\hat{g}(x) \quad = \quad \sum_{k \geqslant 0} r^k = \sum_{k \geqslant 0} r^k N(d,k) \int_{\mathbb{S}^d} g(y) P_k(x^\top y) d\tau_d(y)$$

$$
\begin{aligned}
&= \int_{\mathbb{S}^d} g(y)\bigg(\sum_{k\geqslant 0} r^k N(d,k) P_k(x^\top y)\bigg) d\tau_d(y) \\
&= \int_{\mathbb{S}^d} g(y) \frac{1-r^2}{(1+r^2-2r(x^\top y))^{(d+1)/2}} d\tau_d(y).
\end{aligned}
$$

Moreover, following Bourgain and Lindenstrauss (1988), we have:

$$
\begin{aligned}
g(x)-\hat{g}(x) &= \int_{\mathbb{S}^d} [g(x)-g(y)] \frac{1-r^2}{(1+r^2-2r(x^\top w))^{(d+1)/2}} d\tau_d(y) \\
&= 2\int_{\mathbb{S}^d,\ y^\top x\geqslant 0} [g(x)-g(y)] \frac{1-r^2}{(1+r^2-2r(x^\top w))^{(d+1)/2}} d\tau_d(y) \text{ by parity of } g, \\
|g(x)-\hat{g}(x)| &\leqslant \int_{\mathbb{S}^d,\ y^\top x\geqslant 0} \sqrt{2}\sqrt{1-x^\top y} \frac{1-r^2}{(1+r^2-2r(x^\top y))^{(d+1)/2}} d\tau_d(y).
\end{aligned}
$$

As shown by Bourgain and Lindenstrauss (1988, Eq. (2.13)), this is less than a constant that depends on $d$ times $(1-r)\log\frac{1}{1-r}$. We thus get for $\delta$ large enough, by taking $1-r = (C/\delta)^{1/(\alpha+(d-1)/2)} \in (0,1)$, an error of

$$
(C/\delta)^{1/(\alpha+(d-1)/2)} \log(C/\delta)^{-1/(\alpha+(d-1)/2)}] = O(\delta^{1/(\alpha+(d-1)/2)}\log\delta),
$$

which leads to the desired result.

### D.5 Finding Differentiable Functions which are not in $\mathcal{G}_2$

In this section, we consider functions on the sphere which have the proper parity with respect to $\alpha$, which are $s$-times differentiable with bounded derivatives, but which are not in $\mathcal{G}_2$. We then provide optimal approximation rates for these functions.

We assume that $s-\alpha$ is even, we consider $g(x)=(w^\top x)_+^s$ for a certain arbitrary $w\in\mathbb{S}^d$. As computed at the end of Appendix D.1, we have $\|g_k\|^2_{L_2(\mathbb{S}^d)} = \bigg(\frac{\omega_{d-1}}{\omega_d} P_k(x^\top v)\int_{-1}^1 \varphi(t)P_k(t)(1-t^2)^{(d-2)/2}dt\bigg)^2 N(d,k)$. Given the computations from Appendix D.2, the squared norm equal to $\bigg(\frac{\omega_{d-1}}{\omega_d} P_k(x^\top v)\int_{-1}^1 \varphi(t)P_k(t)(1-t^2)^{(d-2)/2}dt\bigg)^2$ goes down to zero as $k^{-d-2s-1}$, while $N(d,k)$ grows as $k^{d-1}$. In order to use the computation of the RKHS norm derived in Appendix D.2, we need to make sure that $g$ has the proper parity. This can de done by removing all harmonics with $k\leqslant s$ (note that these harmonics are also functions of $w^\top x$, and thus the function that we obtain is also a function of $w^\top x$). That function then has a squared RKHS norm equal to

$$
\sum_{k\geqslant s,\lambda_k\neq 0} \|g_k\|^2_{L_2(\mathbb{S}^d)}\lambda_k^{-2}.
$$

The summand has an asymptotic equivalent proportional to $k^{-d-2s-1}k^{d-1}k^{d+2\alpha+1}$ which is equal to $k^{d+2\alpha-2s-1}$. Thus if $d+2\alpha-2s\geqslant 0$, the series is divergent (the function is not in the RKHS), i.e., if $s\leqslant\alpha+\frac{d}{2}$.

**Best approximation by a function in $\mathcal{G}_2$.** The squared norm of the $k$-th harmonic $\|g_k\|^2_{L_2(\mathbb{S}^d)}$ goes down to zero as $k^{-2s-2}$ and the squared RKHS norm of a $h$ is equivalent to $\displaystyle\sum_{k\geqslant 0}\|h_k\|^2_{L_2(\mathbb{S}^d)}k^{d+2\alpha+1}$. Given $\delta$, we may then find the function $h$ such that $\gamma_2(h)^2 = \sum_{k\geqslant 0}\|h_k\|^2_{L_2(\mathbb{S}^d)}k^{d+2\alpha+1} \leqslant \delta^2$ with smallest $L_2(\mathbb{S}^d)$ norm distance to $g$, that is, $\sum_{k\geqslant 0}\|g_k - h_k\|^2_{L_2(\mathbb{S}^d)}$. The optimal approximation is $h_k = \alpha_k g_k$ for some $\alpha_k \in \mathbb{R}_+$, with error $\sum_{k\geqslant 0}(1-\alpha_k)^2\|g_k\|^2_{L_2(\mathbb{S}^d)} \sim \sum_{k\geqslant 0}(1-\alpha_k)^2 k^{-2s-2}$ and squared $\gamma_2$-norm $\sum_{k\geqslant 0}\alpha_k^2 k^{d+2\alpha+1}k^{-2s-2} = \sum_{k\geqslant 0}\alpha_k^2 k^{d+2\alpha-2s-1}$. The optimal $\alpha_k$ is obtained by considering a Lagrange multiplier $\lambda$ such that $(\alpha_k - 1)k^{-2s-2} + \lambda\alpha_k k^{d+2\alpha-2s-1} = 0$, that is, $\alpha_k = (k^{-2s-2} + \lambda k^{d+2\alpha-2s-1})^{-1}k^{-2s-2} = (1+\lambda k^{d+2\alpha+1})^{-1}$. We then have

$$\begin{aligned}
\sum_{k\geqslant 0}\alpha_k^2 k^{d+2\alpha-2s-1} &= \sum_{k\geqslant 0}(1+\lambda k^{d+2\alpha+1})^{-2}k^{d+2\alpha-2s-1} \\
&\approx \int_0^\infty (1+\lambda t^{d+2\alpha+1})^{-2}t^{d+2\alpha-2s-1}dt \text{ by approximation by an integral,} \\
&\propto \int_0^\infty (1+u)^{-2}d(t^{d+2\alpha-2s}) \text{ with the change of variable } u = \lambda t^{d+2\alpha+1} \\
&\propto \lambda^{-(d+2\alpha-2s)/(d+2\alpha+1)} \text{ up to constants,}
\end{aligned}$$

which should be of order $\delta^2$ (this gives the scaling of $\lambda$ as a function of $\delta$). Then the squared error is

$$\begin{aligned}
\sum_{k\geqslant 0}(1-\alpha_k)^2 k^{-2s-2} &= \sum_{k\geqslant 0}\frac{\lambda^2 t^{2d+4\alpha+2}}{(1+\lambda k^{d+2\alpha+1})^2}k^{-2s-2} \\
&\approx \int_0^\infty \frac{\lambda^2 t^{2d+4\alpha-2s}}{(1+\lambda t^{d+2\alpha+1})^2}dt \\
&\approx \lambda^2\lambda^{-(2d+4\alpha-2s+1)/(d+2\alpha+1)} = \lambda^{-(2d+4\alpha-2s+1-2d-4\alpha-2)/(d+2\alpha+1)} \\
&= \lambda^{(2s+1)/(d+2\alpha+1)} \approx \delta^{-2(2s+1)/(d+2\alpha-2s)},
\end{aligned}$$

and thus the (non-squared) approximation error scales as $\delta^{-(2s+1)/(d+2\alpha-2s)}$. For $s=1$, this leads to a scaling as $\delta^{-3/(d+2\alpha-2)}$.

### D.6 Proof of Prop. 4

For $\alpha = 1$, by writing $v^\top x = (v^\top x)_+ - (-v^\top x)_+$ we obtain the upperbound $\gamma_1(g) \leqslant 2$. For all other situations, we may compute

$$\gamma_2(g)^2 = \sum_{k\geqslant 0}\frac{\|g_k\|^2_{L_2(\mathbb{S}^d)}}{\lambda_k^2}.$$

For $g$ a linear function $g_k = 0$ except for $k=1$, for which, we have $g_1(x) = v^\top x$, and thus $\|g_k\|^2_{L_2(\mathbb{S}^d)} = \int_{\mathbb{S}^d}(v^\top x)^2 d\tau_d(x) = v^\top\big(\int_{\mathbb{S}^d}xx^\top\tau_d(x)\big)v = 1$. This implies that $\gamma_2(g) = \lambda_1^{-1}$. Given the expression (from Appendix D.2) $\lambda_1 = \frac{d-1}{d}\frac{\alpha}{4\pi}\frac{\Gamma(\alpha/2)\Gamma(d/2+1)}{\Gamma(\alpha/2+d/2+1)}$ for $\alpha > 1$ and $\lambda_1 = \frac{d-1}{2d\pi}$.

## Appendix E. Computing $\ell_2$-Haussdorff Distance between Ellipsoids

We assume that we are given two ellipsoids defined as $(x-a)^\top A^{-1}(x-a) \leqslant 1$ and $(x-b)^\top B^{-1}(x-b) \leqslant 1$ and we want to compute their Hausdorff distance. This leads to the two equivalent problems

$$\max_{\|w\|_2 \leqslant 1} w^\top (a-b) - \|B^{1/2}w\|_2 + \|A^{1/2}w\|_2,$$

$$\max_{\|u\|_2 \leqslant 1} \min_{\|v\|_2 \leqslant 1} \|a + A^{1/2}u - b - B^{1/2}v\|_2,$$

which are related by $w = a + A^{1/2}u - b - B^{1/2}v$. We first review classical methods for optimization of quadratic functions over the $\ell_2$-unit ball.

**Minimizing convex quadratic forms over the sphere.** We consider the following convex optimization problem, with $Q \succcurlyeq 0$; we have by Lagrangian duality:

$$\begin{aligned}
&\min_{\|x\|_2 \leqslant 1} \frac{1}{2}x^\top Qx - q^\top x \\
&\max_{\lambda \geqslant 0} \min_{x \in \mathbb{R}^d} \frac{1}{2}x^\top Qx - q^\top x + \frac{\lambda}{2}(\|x\|_2^2 - 1) \\
&\max_{\lambda \geqslant 0} -\frac{1}{2}q^\top (Q+\lambda I)^{-1} q - \frac{\lambda}{2} \text{ with } x = (Q+\lambda I)^{-1}q.
\end{aligned}$$

If $\|Q^{-1}q\|_2 \leqslant 1$, then $\lambda = 0$ and $x = Q^{-1}q$. Otherwise, at the optimum, $\lambda > 0$ and $\|x\|_2^2 = q^\top (Q+\lambda I)^{-2} q = 1$, which implies $1 \leqslant \frac{1}{\lambda + \lambda_{\min}(Q)} q^\top Q^{-1} q$, which leads to $\lambda \leqslant q^\top Q^{-1} q - \lambda_{\min}(Q)$, which is important to reduce the interval of possible $\lambda$. The optimal $\lambda$ may then be obtained by binary search (from a single SVD of $Q$).

**Minimizing concave quadratic forms over the sphere.** We consider the following non-convex optimization problem, with $Q \succcurlyeq 0$, for which strong Lagrangian duality is known to hold (Boyd and Vandenberghe, 2004):

$$\begin{aligned}
&\min_{\|x\|_2 \leqslant 1} -\frac{1}{2}x^\top Qx + q^\top x = \min_{\|x\|_2 = 1} -\frac{1}{2}x^\top Qx + q^\top x \\
&\max_{\lambda \geqslant 0} \min_{x \in \mathbb{R}^d} -\frac{1}{2}x^\top Qx + q^\top x + \frac{\lambda}{2}(\|x\|_2^2 - 1) \\
&\max_{\lambda \geqslant \lambda_{\max}(Q)} -\frac{1}{2}q^\top (\lambda I - Q)^{-1} q - \frac{\lambda}{2} \text{ with } x = (Q - \lambda I)^{-1}q.
\end{aligned}$$

At the optimum, we have $q^\top (\lambda I - Q)^{-2} q = 1$, which implies $1 \leqslant \frac{1}{[\lambda - \lambda_{\max}(Q)]^2}\|q\|_2^2$, which leads to $0 \leqslant \lambda - \lambda_{\max}(Q) \leqslant \|q\|_2$. We may perform binary search on $\lambda$ from a single SVD of $Q$.

**Computing the Haussdorff distance.** We need to compute:

$$\max_{\|u\|_2 \leqslant 1} \min_{\|v\|_2 \leqslant 1} \frac{1}{2}\|a + A^{1/2}u - b - B^{1/2}v\|_2^2$$

$$
\begin{aligned}
&= \max_{\|u\|_2 \leqslant 1} \max_{\lambda \geqslant 0} \min_{\|v\|_2 \leqslant 1} \frac{1}{2}\|a + A^{1/2}u - b - B^{1/2}v\|_2^2 + \frac{\lambda}{2}(\|v\|_2^2 - 1) \\
&= \max_{\|u\|_2 \leqslant 1} \max_{\lambda \geqslant 0} -\frac{\lambda}{2} + \frac{1}{2}\|a - b + A^{1/2}u\|^2 - \frac{1}{2}(a - b + A^{1/2}u)^\top B(B + \lambda I)^{-1}(a - b + A^{1/2}u) \\
&= \max_{\|u\|_2 \leqslant 1} \max_{\lambda \geqslant 0} -\frac{\lambda}{2} + \frac{\lambda}{2}(a - b + A^{1/2}u)^\top (B + \lambda I)^{-1}(a - b + A^{1/2}u)
\end{aligned}
$$

with $v = (B + \lambda I)^{-1}B^{1/2}(a - b + A^{1/2}u)$. The interval in $\lambda$ which is sufficient to explore is

$$\lambda \in [0, -\lambda_{\min}(B) + (\|a - b\|_2^2 + \lambda_{\max}(A^{1/2}))^2],$$

which are bounds that are independent of $u$.

Given $\lambda \geqslant 0$, we have the problem of

$$
\begin{aligned}
& \min_{\mu \geqslant 0} \max_{u \in \mathbb{R}^d} \frac{\lambda}{2}(a - b + A^{1/2}u)^\top (B + \lambda I)^{-1}(a - b + A^{1/2}u) - \frac{\mu}{2}(\|u\|_2^2 - 1) - \frac{\lambda}{2} \\
= \; & \min_{\mu \geqslant 0} \max_{u \in \mathbb{R}^d} \frac{\lambda}{2}(a - b)^\top (B + \lambda I)^{-1}(a - b) + \frac{\mu - \lambda}{2} + \lambda u^\top A^{1/2}(B + \lambda I)^{-1}(a - b) \\
& \qquad\qquad - \frac{1}{2}u^\top(\mu I - \lambda A^{1/2}(B + \lambda I)^{-1}A^{1/2})u \\
= \; & \min_{\mu \geqslant 0} \frac{\lambda}{2}(a - b)^\top (B + \lambda I)^{-1}(a - b) + \frac{\mu - \lambda}{2} \\
& + \lambda^2 (a - b)^\top (B + \lambda I)^{-1}A^{1/2}(\mu I - \lambda A^{1/2}(B + \lambda I)^{-1}A^{1/2})^{-1}A^{1/2}(B + \lambda I)^{-1}(a - b)
\end{aligned}
$$

We have $u = (\frac{\mu}{\lambda}I - A^{1/2}(B + \lambda I)^{-1}A^{1/2})^{-1}A^{1/2}(B + \lambda I)^{-1}(a - b)$, leading to $w \propto (\lambda^{-1}B - \mu^{-1}A + I)(a - b)$. We need $\frac{\mu}{\lambda} \geqslant \lambda_{\max}(A^{1/2}(B + \lambda I)^{-1}A^{1/2})$. Moreover

$$0 \leqslant \frac{\mu}{\lambda} - \lambda_{\max}(A^{1/2}(B + \lambda I)^{-1}A^{1/2}) \leqslant \|A^{1/2}(B + \lambda I)^{-1}(a - b)\|.$$

This means that the $\ell_2$-Haussdorff distance may be computed by solving in $\lambda$ and $\mu$, by exhaustive search with respect to $\lambda$ and by binary search (or Newton's method) for $\mu$. The complexity of each iteration is that of a singular value decomposition, that is $O(d^3)$. For more details on optimization of quadratic functions on the unit-sphere, see Forsythe and Golub (1965).

## References

P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2009.

R. A. Adams and J. F. Fournier. *Sobolev Spaces*, volume 140. Academic Press, 2003.

K. Atkinson and W. Han. *Spherical Harmonics and Approximations on the Unit Sphere: an Introduction*, volume 2044. Springer, 2012.

F. Bach. Consistency of the group Lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008a.

F. Bach. Exploring large feature spaces with hierarchical multiple kernel learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2008b.

F. Bach. Convex relaxations of structured matrix factorizations. Technical Report 00861118, HAL, 2013.

F. Bach. Duality between subgradient and conditional gradient methods. *SIAM Journal on Optimization*, 25(1):115–129, 2015.

F. Bach. On the equivalence between kernel quadrature rules and random feature expansions. *Journal of Machine Learning Research*, 18:1–38, 2017.

A. R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, 1993.

P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2003.

P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

A. Barvinok. *A Course in Convexity*, volume 54. American Mathematical Society, 2002.

Y. Bengio, N. Le Roux, P. Vincent, O. Delalleau, and P. Marcotte. Convex neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2006.

A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*, volume 3. Springer, 2004.

E. D. Bolker. A class of convex bodies. *Transactions of the American Mathematical Society*, 145:323–345, 1969.

L. Bottou and O. Bousquet. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.

J. Bourgain and J. Lindenstrauss. Projection bodies. In *Geometric Aspects of Functional Analysis*, pages 250–270. Springer, 1988.

J. Bourgain, J. Lindenstrauss, and V. Milman. Approximation of zonoids by zonotopes. *Acta Mathematica*, 162(1):73–141, 1989.

S. P. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

L. Breiman. Hinging hyperplanes for regression, classification, and function approximation. *IEEE Transactions on Information Theory*, 39(3):999–1013, 1993.

P. Bühlmann and S. Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer, 2011.

M. Burger and A. Neubauer. Error bounds for approximation with neural networks. *Journal of Approximation Theory*, 112(2):235–250, 2001.

Y. Cho and L. K. Saul. Kernel methods for deep learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2009.

A. S. Dalalyan, A. Juditsky, and V. Spokoiny. A new algorithm for estimating the effective dimension-reduction subspace. *Journal of Machine Learning Research*, 9:1647–1678, 2008.

V. F. Dem'yanov and A. M. Rubinov. The minimization of a smooth convex functional on a convex set. *SIAM Journal on Control*, 5(2):280–294, 1967.

R. A. DeVore, R. Howard, and C. Micchelli. Optimal nonlinear approximation. *Manuscripta Mathematica*, 63(4):469–478, 1989.

M. Dudik, Z. Harchaoui, and J. Malick. Lifted coordinate descent for learning with trace-norm regularization. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012.

J. C. Dunn and S. Harshbarger. Conditional gradient algorithms with open loop step size rules. *Journal of Mathematical Analysis and Applications*, 62(2):432–444, 1978.

H. Edelsbrunner. *Algorithms in Combinatorial Geometry*, volume 10. Springer, 1987.

L. C. Evans and R. F. Gariepy. *Measure Theory and Fine Properties of Functions*, volume 5. CRC Press, 1991.

G. E. Forsythe and G. H. Golub. On the stationary values of a second-degree polynomial on the unit sphere. *Journal of the Society for Industrial & Applied Mathematics*, 13(4): 1050–1068, 1965.

M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1-2):95–110, 1956.

J. H. Friedman and W. Stuetzle. Projection pursuit regression. *Journal of the American Statistical Association*, 76(376):817–823, 1981.

C. Frye and C. J. Efthimiou. Spherical Harmonics in $p$ Dimensions. Technical Report 1205.3548, ArXiv, 2012.

K. Fukumizu, F. Bach, and M. I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. *Journal of Machine Learning Research*, 5:73–99, 2004.

C. Gu. *Smoothing Spline ANOVA Models*, volume 297. Springer, 2013.

L. J. Guibas, A. Nguyen, and L. Zhang. Zonotopes as bounding volumes. In *Proceedings of the ACM-SIAM symposium on Discrete Algorithms*, 2003.

V. Guruswami and P. Raghavendra. Hardness of learning halfspaces with noise. *SIAM Journal on Computing*, 39(2):742–765, 2009.

L. Györfi and A. Krzyzak. *A Distribution-free Theory of Nonparametric Regression.* Springer, 2002.

Z. Harchaoui, A. Juditsky, and A. Nemirovski. Conditional gradient algorithms for norm-regularized smooth convex optimization. *Mathematical Programming*, pages 1–38, 2013.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning.* Springer, 2009. 2nd edition.

T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models.* Chapman & Hall, 1990.

S. Haykin. *Neural Networks: A Comprehensive Foundation.* Prentice Hall, 1994.

G. E. Hinton and Z. Ghahramani. Generative models for discovering sparse distributed representations. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 352(1358):1177–1190, 1997.

M. Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2013.

S. M. Kakade, K. Sridharan, and A. Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Advances in Neural Information Processing Systems (NIPS)*, 2009.

A. R. Klivans and A. A. Sherstov. Cryptographic hardness for learning intersections of halfspaces. In *Annual Symposium on Foundations of Computer Science (FOCS)*, 2006.

V. Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902–1914, 2001.

S. König. Computational aspects of the Hausdorff distance in unbounded dimension. Technical Report 1401.1434, ArXiv, 2014.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012.

V. Kurkova and M. Sanguineti. Bounds on rates of variable-basis and neural-network approximation. *IEEE Transactions on Information Theory*, 47(6):2659–2665, Sep 2001.

G. Lan. The complexity of large-scale convex programming under a linear optimization oracle. Technical Report 1309.5550, arXiv, 2013.

N. Le Roux and Y. Bengio. Continuous neural networks. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2007.

W. S. Lee, P. L. Bartlett, and R. C. Williamson. Efficient agnostic learning of neural networks with bounded fan-in. *IEEE Transactions on Information Theory*, 42(6):2118–2132, 1996.

M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6 (6):861–867, 1993.

K.-C. Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991.

Y. Lin and H. H. Zhang. Component selection and smoothing in multivariate nonparametric regression. *Annals of Statistics*, 34(5):2272–2297, 2006.

Roi Livni, Shai Shalev-Shwartz, and Ohad Shamir. On the computational efficiency of training neural networks. In *Advances in Neural Information Processing Systems*, 2014.

V. Maiorov. Approximation by neural networks and learning theory. *Journal of Complexity*, 22(1):102–117, 2006.

V. E. Maiorov and R. Meir. On the near optimality of the stochastic approximation of smooth functions by neural networks. *Advances in Computational Mathematics*, 13(1): 79–103, 2000.

Y. Makovoz. Uniform approximation by neural networks. *Journal of Approximation Theory*, 95(2):215–228, 1998.

J. Matoušek. Improved upper bounds for approximation by zonotopes. *Acta Mathematica*, 177(1):55–73, 1996.

H. N. Mhaskar. On the tractability of multivariate integration and approximation by neural networks. *Journal of Complexity*, 20(4):561–590, 2004.

H. N. Mhaskar. Weighted quadrature formulas and approximation by zonal function networks on the sphere. *Journal of Complexity*, 22(3):348–370, 2006.

G. F. Montufar, R. Pascanu, K. Cho, and Y. Bengio. On the number of linear regions of deep neural networks. In *Advances in neural information processing systems*, 2014.

V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of International Conference on Machine Learning (ICML)*, 2010.

R. M. Neal. *Bayesian Learning for Neural Networks.* PhD thesis, University of Toronto, 1995.

Y. Nesterov. Semidefinite relaxation and nonconvex quadratic optimization. *Optimization Methods and Software*, 9(1-3):141–160, 1998.

Y. Nesterov. *Introductory lectures on convex optimization: a basic course.* Kluwer Academic Publishers, 2004.

Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, 2005.

P. P. Petrushev. Approximation by ridge functions and neural networks. *SIAM Journal on Mathematical Analysis*, 30(1):155–189, 1998.

A. Pinkus. Approximation theory of the MLP model in neural networks. *Acta Numerica*, 8:143–195, 1999.

A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems (NIPS)*, 2007.

P. Ravikumar, H. Liu, J. Lafferty, and L. Wasserman. SpAM: Sparse additive models. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.

R. T. Rockafellar. *Convex Analysis.* Princeton University Press, 1997.

F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386, 1958.

S. Rosset, J. Zhu, and T. Hastie. Boosting as a regularized path to a maximum margin classifier. *Journal of Machine Learning Research*, 5:941–973, 2004.

S. Rosset, G. Swirszcz, N. Srebro, and J. Zhu. $\ell_1$-regularization in infinite dimensional feature spaces. In *Proceedings of the Conference on Learning Theory (COLT)*, 2007.

W. Rudin. *Real and Complex Analysis.* Tata McGraw-Hill Education, 1987.

D. E. Rumelhart, G. E Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986.

R. Schneider. Zu einem problem von shephard über die projektionen konvexer körper. *Mathematische Zeitschrift*, 101(1):71–82, 1967.

S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms.* Cambridge University Press, 2014.

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis.* Cambridge University Press, 2004.

A. J. Smola, Z. L. Ovari, and R. C. Williamson. Regularization with dot-product kernels. *Advances in Neural Information Processing Systems (NIPS)*, 2001.

K. Sridharan. *Learning from an Optimization Viewpoint.* PhD thesis, Toyota Technological Institute at Chicago, 2012.

U. von Luxburg and O. Bousquet. Distance–based classification with Lipschitz functions. *Journal of Machine Learning Research*, 5:669–695, 2004.

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

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

X. Zhang, D. Schuurmans, and Y. Yu. Accelerated training for matrix-norm regularization: A boosting approach. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.