Super-fast rates of convergence for Neural Networks Classifiers under the Hard Margin Condition

Anonymous authors
Paper under double-blind review

Abstract

We study the classical binary classification problem for hypothesis spaces of Deep Neural Networks (DNNs) with ReLU activation under Tsybakov's low-noise condition with exponent $q > 0$, as well as its limit case $q = \infty$, which we refer to as the *hard margin condition*. We demonstrate that DNN solutions to the empirical risk minimization (ERM) problem with square loss surrogate and ℓ_p penalty on the weights ($0 < p < \infty$) can achieve excess risk bounds of order $\mathcal{O}\left(n^{-\alpha}\right)$ for arbitrarily large $\alpha > 1$ under the hard-margin condition, provided that the Bayes regression function η satisfies a *distribution-adapted* smoothness condition relative to the marginal data distribution ρ_X. Additionally, we establish minimax lower bounds, showing that these rates cannot be improved upon. Our proof relies on a novel decomposition of the excess risk for general ERM-based classifiers, which may be of independent interest.

1 Introduction

In this article, we study the problem of classifying high-dimensional data points with binary labels. It is well-known that, in the absence of structural assumptions about the data or the underlying model, convergence rates for classification tasks typically decay as $\mathcal{O}(n^{-c/d})$ for some constant $c > 0$, which becomes arbitrarily slow as the dimensionality d increases. This phenomenon is often referred to as the curse of dimensionality (CoD). However, it has been observed that many models used in practice — particularly Deep Neural Networks in recent years — are capable of efficiently solving extremely high-dimensional classification tasks, achieving convergence rates that appear to defy the CoD (Goodfellow et al., 2016; Krizhevsky et al., 2012).

This gap between theoretical results and practical observations can often be bridged by introducing suitable regularity assumptions on the problem. In the context of supervised binary classification, such assumptions frequently take the form of *margin conditions*. First introduced in the seminal work of (Mammen & Tsybakov, 1999), margin conditions characterize the behavior of the data distribution near the decision boundary — the region where classification is most challenging. Over the years, these conditions have enabled the derivation of CoD-free rates of convergence for classifiers based on various hypothesis spaces (Tsybakov, 2004; Audibert & Tsybakov, 2007). Remarkably, margin conditions can not only eliminate the curse of dimensionality, but can also lead to "fast" rates of convergence — faster than the standard $\mathcal{O}(n^{-1/2})$ — and, under their strongest form, even "super-fast" rates, exceeding $\mathcal{O}(n^{-1})$.

Notable examples of hypothesis spaces for which these super-fast (sometimes even exponential) rates of convergence have been observed include local polynomial estimators (Audibert & Tsybakov, 2007), support vector machines (Steinwart & Scovel, 2005; Steinwart & Christmann, 2008; Cabannes & Vigogna, 2022) or Reproducing Kernel Hilbert Spaces (RKHS) (Koltchinskii & Beznosova, 2005; Smale & Zhou, 2007; Vigogna et al., 2022). More recently, it has even been shown that for data coming from an infinite-dimensional Hilbert space, the Delaigle-Hall condition (Delaigle & Hall, 2012), which can be thought of as an infinite-dimensional analogue of the classical margin conditions, can lead to super-fast rates of convergence for RKHS classifiers (Wakayama & Imaizumi, 2021).

Perhaps surprisingly however, for hypothesis spaces of Deep Neural Networks (DNNs), no such "super-fast" rates of convergence have been shown to hold, even under the strongest margin and regularity assumptions.

This fact seemingly contradicts the observation that DNNs outperform all other traditional methods by far when it comes to high-dimensional classification. This naturally raises the question: are Neural Networks truly inferior to traditional classification methods in the hard-margin regime? In this work, we answer negatively to this question by showing that "super-fast" rates of convergence for DNN classifiers can also hold under the hard-margin condition. Before presenting our setup and results in greater detail, we briefly review related literature in the following section.

1.1 Related works

When considering a binary classification problem on $[0,1]^d$ with labels $\{1,-1\}$, there are different possible objects which can be used to characterize the regularity of the problem:

- the Bayes regression function $\eta : x \in [0,1]^d \mapsto \mathbb{E}[Y \mid X = x]$ which, up to an affine transformation, represents the conditional probability of $\{Y = 1\}$ given $\{X = x\}$,
- the Bayes classifier c induced by the Bayes regression function: $c : x \mapsto \text{sign}(\eta(x))$. It is the optimal classifier in the sense that it minimizes the expected 0-1 loss over all admissible classifiers, and is therefore what we are implicitly trying to learn.
- the decision region $\Omega := c^{-1}(\{1\})$ and the induced decision boundary $\partial\Omega$.

The margin condition which we refer to in this work, and originally introduced in (Mammen & Tsybakov, 1999), assumes that for all $t > 0$, $\mathbb{P}(|\eta(X)| \geqslant t) \lesssim t^q$, where $q > 0$ is a constant called the *margin exponent* (note that depending on the source, this is also referred to as a *low-noise* condition). In (Kim et al., 2018), it has been shown that such a margin condition coupled with additional assumptions on respectively the regression function η, the decision boundary $\partial\Omega$, or the probability for data points to be near the decision boundary $\partial\Omega$, leads to minimax optimal fast rates of convergence for sparse DNN classifiers obtained by hinge-loss empirical risk minimization. For instance, if η is assumed to be Hölder continuous, they prove the excess risk bound:

$$\mathcal{E}(\hat{f}_{DNN}) \lesssim \left(\frac{\log^3 n}{n}\right)^{\frac{\beta(q+1)}{\beta(q+2)+d}},$$

where β is the Hölder exponent of η. As we can see, when the margin exponent $q \to \infty$, their result leads to the "fast rate" of $\mathcal{O}\left(n^{-1}\right)$.

In a similar vein, by assuming different kinds of regularity on these objects, and leveraging recent advances on the approximation rates and complexity measures of DNNs hypothesis spaces, various minimax optimal rates of convergence of this kind have been obtained for DNNs under different settings. A non-exhaustive list of such works includes (Feng et al., 2021; Meyer, 2022; Petersen & Voigtlaender, 2021; Bos & Schmidt-Hieber, 2022; Hu et al., 2022; Ko et al., 2023). As it has been mentioned earlier, while these results for DNNs clearly highlight their ability to generalize with CoD-free rates of convergence, none of them obtain a rate faster than $\mathcal{O}\left(n^{-1}\right)$, even under the most idealized regularity assumptions, unlike the more traditional methods.

To the best of the authors' knowledge, it has only been shown in (Hu et al., 2021) that the *hard-margin condition* (which can informally be seen as the limit $q = \infty$ of Tsybakov's low-noise condition), can lead to exponential rates of convergence for the excess risk for Neural Networks: they prove the result for shallow networks in the Neural Tangent Kernel (NTK) regime (Jacot et al., 2018), which are trained to minimize the Empirical Risk with square loss surrogate. (Nitanda & Suzuki, 2020) similarly show how, in the NTK regime, the hard-margin condition leads to an exponential convergence of the averaged stochastic gradient descent (SGD) algorithm with respect to the number of epochs. However, these results are not fully satisfactory, as it is known that the NTK regime does not accurately represent the expressive power of deeper Networks (Bietti & Bach, 2021). This work is thus, to the best of our knowledge, the first to prove super-fast rates of convergence for conventional DNNs hypothesis spaces under the hard-margin condition.

1.2 Our Contributions

We study the binary classification problem over a hypothesis space of parametric functions. The classifiers are learned in a standard supervised learning fashion, by minimizing an Empirical Risk with the square loss as a surrogate and an ℓ_p penalty on the network's weight, where $0 < p < \infty$. For a real-valued, measurable function f, denote the *excess risk* $\mathcal{E}(f)$ of the classifier induced by f as

$$\mathcal{E}(f) := \mathbb{P}_{(X,Y)\sim\rho}\left(\operatorname{sign} f(X) \neq Y\right) - \mathbb{P}_{(X,Y)\sim\rho}\left(c^*(X) \neq Y\right),$$

where c^* is the Bayes classifier, induced by the Bayes regression function η. Our main contributions can be stated as follows:

- In Theorem 2, we provide a novel error decomposition for the excess risk of classifiers induced by general classes of parametric functions under both "weak" ($q > 0$) and "hard" ($q = \infty$) margin conditions. The proof is elementary and relies on an inequality we learned from (Vigogna et al., 2022).
- As a direct application of Theorem 2, we show in Theorem 4 that when the hypothesis space consists of DNNs with ReLU activation, and the regression function η satisfies a *distribution-adapted smoothness* condition relative to the marginal data distribution ρ_X, the excess risk $\mathcal{E}(\hat{f}_{NN})$ converges at a rate of $\mathcal{O}\left(n^{-\alpha}\right)$. Specifically, $\alpha \to 1$ as $s \to \infty$ under the weak-margin condition, and $\alpha \to \infty$ as $s \to \infty$ under the hard-margin condition. Here, $s > 0$ quantifies how efficiently η can be approximated by the DNN space and can be informally interpreted as a smoothness parameter.
- Lastly, we apply Theorem 2 again to a simplified version of the teacher-student setting, which we recast as a binary classification problem in which the training labels are given by fuzzy predictions of a "teacher" neural network: we show that if the teacher network is realizable by the student network, then the excess risk $\mathcal{E}(\hat{f}_{DNN})$ converges at a rate $\mathcal{O}\left(e^{-\beta n}\right)$ for some constant $\beta > 0$.

In all of our results, the excess risk bounds hold in the almost-sure sense and are non-asymptotic: they hold for any integer $n \geqslant n_0$ for some constant n_0 whose expression we give explicitly in terms of the problem's parameters.

1.3 Notations

Function Spaces:

Let $d \geqslant 1$ be an integer. For a closed subset $\mathcal{X} \subseteq \mathbb{R}^d$, a Borel measurable $\mathcal{Y} \subseteq \mathbb{R}$, and an integer $k \geqslant 0$ we will denote by

- $\mathcal{M}(\mathcal{X}, \mathcal{Y})$ the space of Borel measurable functions from $\mathcal{X}$ to $\mathcal{Y}$,
- $\mathcal{C}^k(\mathcal{X}, \mathcal{Y})$ the space of $\mathcal{Y}$-valued, k times continuously differentiable functions on $\mathcal{X}$,
- $L^p(\mathcal{X}, \mathcal{Y}, \mu)$ the space of Borel measurable $\mathcal{Y}$-valued functions on $\mathcal{X}$ whose absolute p-th power is μ-integrable, where μ is a measure on $\mathcal{X}$ and $p \in [1, \infty]$. Whenever μ is the Lebesgue measure, we will omit it from notation and simply write $L^p(\mathcal{X}, \mathcal{Y})$.

For any of these function spaces, we might drop the domain $\mathcal{X}$ and/or the co-domain $\mathcal{Y}$ from notation if context already makes it clear.

Norms:

For any positive integers d, u, v, real $0 < p < \infty$, $x = (x_1, \ldots, x_d)^T \in \mathbb{R}^d$, $A = (a_{i,j}) \in \mathbb{R}^{u\times v}$ and $f \in \mathcal{M}(\mathcal{X}, \mathbb{R})$, we will denote by

- respectively $|x|_p := (|x_1|^p + \ldots + |x_d|^p)^{1/p}$, $|x|_0 := |x_1|^0 + \ldots |x_d|^0$ (with the convention $0^0 := 0$) and $|x|_\infty := \max_{1 \leqslant i \leqslant d} |x_i|$, the ℓ_p, ℓ_0 and ℓ_∞ (quasi-)norm of x.
- $|A|_p := \left(\sum_{i,j} |a_{i,j}|^p \right)^{1/p}$ the $\ell_{p,p}$ norm of A,
- respectively $\|f\|_{\mathcal{C}^k(\mathcal{X})}$ and $\|f\|_{L^p(\mu)}$ the $\mathcal{C}^k(\mathcal{X}, \mathbb{R})$ norm and $L^p(\mathcal{X}, \mathbb{R}, \mu)$ norm of f, which are defined in the standard way.

Other Symbols:

We will also denote by

- $\mathbb{N} := \{1, 2, \ldots\}$ the set of all natural numbers, and $\mathbb{N}_0 := \{0\} \cup \mathbb{N}$,
- $\mathbb{1}_A$ the indicator function of a set A, which equals 1 on A and 0 everywhere else,
- $\operatorname{sign}(x) := \mathbb{1}_{(0,\infty)}(x) - \mathbb{1}_{(-\infty,0)}(x)$ the sign of a real number x. We will also denote by $\operatorname{sign} f := \operatorname{sign} \circ f$ the composition of a real-valued function f with sign,
- $\mathbb{E}[Z]$ the expectation of a random variable Z. If $Z = f(X, Y)$, we may write $\mathbb{E}_X[Z]$ or $\mathbb{E}_Y[Z]$ to indicate with respect to which variables the expectation is taken, or equivalently $\mathbb{E}_\mu[Z]$ to indicate with respect to which distribution the expectation is taken,
- For two sequences of real numbers $(A_n)_{n \geqslant 1}$ and $(B_n)_{n \geqslant 1}$, we will write $A_n \lesssim B_n$ if $A_n \leqslant C B_n$ for some absolute constant $C > 0$, and $A_n = \mathcal{O}(B_n)$ if there exists $n_0 \in \mathbb{N}$ such that $|A_n| \lesssim |B_n|$ for all $n \geqslant n_0$.

2 Problem Setting

Let $d \geqslant 2$ be an integer. We are given a sample of n observations $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ where $\mathcal{X} \equiv [0,1]^d$ is the d-dimensional unit cube and $\mathcal{Y} \equiv \{-1, 1\}$ is the set of possible labels. Each sample is assumed to be i.i.d. data points generated from a distribution ρ on the probability space $(\Omega, \mathscr{A}, \mathbb{P})$. We will call any measurable map $c : \mathcal{X} \to \mathcal{Y}$ a *classifier*, and for any such function c we define its *misclassification risk* by

$$\mathcal{R}(c) := \mathbb{P}_{(X,Y) \sim \rho}(c(X) \neq Y) \tag{1}$$

For any function $f : \mathcal{X} \to \mathbb{R}$, we thus see that $\operatorname{sign} f$ is always a classifier, and we will call $\operatorname{sign} f$ the classifier *induced* by f. It is well known that the misclassification risk is minimized by the Bayes classifier $c^* := \operatorname{sign} \eta$ (Devroye et al., 2013), where

$$\eta(x) := \mathbb{E}_{(X,Y) \sim \rho}[Y \mid X = x]$$

is the so-called Bayes regression function.

We will denote by $\mathcal{R}^* := \mathcal{R}(c^*)$ the optimal risk. As c^* depends on the unknown distribution ρ, it is a priori not possible to achieve the optimal risk $\mathcal{R}_*$, hence we instead aim to learn a classifier $\widehat{c}_n$ from the observations $(x_1, y_1), \ldots, (x_n, y_n)$, such that the *excess risk* $\mathcal{R}(\widehat{c}_n) - \mathcal{R}_*$ converges to zero as fast as possible when n goes to infinity.

2.1 Empirical Risk Minimization

The misclassification risk (1) being a function of ρ, it can't be explicitly computed and hence minimized. We instead minimize the following *Empirical Risk* with square surrogate loss:

$$\widehat{\mathcal{R}}_\ell(f) := \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 \tag{2}$$

Our choice of the square loss $\ell(f(x), y) := (f(x) - y)^2$ as a surrogate is motivated by at least three reasons :

- Empirical evidence suggests that square loss may perform just as well if not better than cross-entropy for classification tasks (Hui & Belkin, 2020). Our result thus provides some theoretical backing for this observation.
- (Hu et al., 2021) prove rates of convergence under the hard-margin condition for Neural Networks classifiers in the NTK regime learned with square loss. Our work shows that their results extend outside of the NTK regime, as they correctly conjectured.
- Most convergence rate results for kernel-based classifiers under margin conditions also consider the square loss as a surrogate (Steinwart & Scovel, 2005; Steinwart & Christmann, 2008). We thus have an analogous setting for DNNs and can meaningfully compare the two approaches.

To match what is often done in practice, we also introduce a penalty function $\mathcal{P} : \mathcal{H} \to \mathbb{R}_{\geqslant 0}$ and a regularization parameter $\lambda \geqslant 0$. This leads to the following *λ-Regularized Empirical Risk Minimization* (λ-ERM) problem :

$$\widehat{f}_\lambda := \operatorname*{argmin}_{f \in \mathcal{H}} \left\{ \widehat{\mathcal{R}}_\ell(f) + \lambda \mathcal{P}(f) \right\}. \tag{3}$$

As stated earlier, we will set the hypothesis space $\mathcal{H}$ as a parametric family of functions, and the penalty $\mathcal{P}$ as the ℓ_p norm. We aim to give fast rates of convergence for the excess risk of sign $\widehat{f}_\lambda$, the classifier induced by $\widehat{f}_\lambda$.

2.2 Hypothesis Spaces of Parametric Functions

2.2.1 Parametric Function Families

We start by defining the parametric families of functions we will be considering in this paper, and the associated terminology. Given integers $L, a_0, a_1, \ldots, a_L \in \mathbb{N}$, we call *parameter vector* and denote by

$$\boldsymbol{\theta} := ((W_1, B_1), \ldots, (W_L, B_L))$$

a tuple of matrix-vector pairs, where $W_l \in \mathbb{R}^{a_l \times a_{l-1}}$ and $B_l \in \mathbb{R}^{a_l}$ are respectively referred to as *weight matrices* and *bias vectors*.

We call $\mathbf{a} = (a_0, a_1, \ldots, a_L) \in \mathbb{N}^{L+1}$ an *architecture vector*, and given any such $\mathbf{a}$, we define the sets of all respectively bounded and unbounded *parametrizations* as:

$$\mathcal{P}_{\mathbf{a},R} := \bigtimes_{l=1}^{L} \left([-R, R]^{a_l \times a_{l-1}} \times [-R, R]^{a_l} \right), \quad \mathcal{P}_{\mathbf{a},\infty} := \bigtimes_{l=1}^{L} \left(\mathbb{R}^{a_l \times a_{l-1}} \times \mathbb{R}^{a_l} \right) \tag{4}$$

where $R > 0$ is a fixed *parameter bound*. We then call a *realization mapping* any fixed map

$$\mathcal{F} : \mathcal{P}_{\mathbf{a},\infty} \to \mathcal{C}(\mathcal{X}, \mathbb{R}) \tag{5}$$

and from then we define

$$\mathcal{H}_{\mathcal{F},\mathbf{a},R} := \{ \mathcal{F}(\boldsymbol{\theta}) \mid \boldsymbol{\theta} \in \mathcal{P}_{\mathbf{a},R} \}$$

as the hypothesis space of functions *induced by* $\mathcal{F}$, parametrized by $\mathbf{a}$ and with parameters bounded by R.

A quantity of interest, after having defined $\mathcal{H}_{\mathcal{F},\mathbf{a},R}$ as we did, is the *sparsity* of $\mathcal{H}_{\mathcal{F},\mathbf{a},R}$. That is, the number of non-zero parameters needed to describe an arbitrary element of $\mathcal{H}_{\mathcal{F},\mathbf{a},R}$. We will denote that quantity — which implicitly depends on $\mathcal{F}$ — $P(\mathbf{a})$, and remark that we always have $P(\mathbf{a}) \leqslant \sum_{l=1}^{L} a_l(a_{l-1} + 1)$. We also remark that any tuple $\boldsymbol{\theta} \in \mathcal{P}_{\mathbf{a},\infty}$ can naturally be identified, up to permutation, with a vector $\widetilde{\boldsymbol{\theta}} \in \mathbb{R}^{P(\mathbf{a})}$, hence the name *parameter vector*.

This rather general and seemingly arbitrary representation for parametric families of function is motivated by hypothesis spaces of Deep Neural Networks, to which it is particularly adapted. However, this representation can be used to represent essentially any kind of parametric family of real-valued functions one might use for binary classification in practice, such as:

- Linear classifiers, induced by maps of the form $x \mapsto W^T x + B$. In this case, the architecture vector is simply given by $\mathbf{a} = (d, 1)$ with corresponding parametrization $\mathcal{P}_{\mathbf{a},\infty} = \mathbb{R}^d \times \mathbb{R}$. The associated realization mapping is defined for all $\boldsymbol{\theta} \equiv (W, B) \in \mathcal{P}_{\mathbf{a},\infty}$ by
$$\mathcal{F}(\boldsymbol{\theta}) = \left(f : x \mapsto W^T x + B\right).$$
- Logistic regression, induced by maps of the form $x \mapsto 2(1 + \exp(W^T x + B))^{-1} - 1$. For this, the architecture vector and parametrization are again given by $\mathbf{a} = (d, 1)$ and $\mathcal{P}_{\mathbf{a},\infty} = \mathbb{R}^d \times \mathbb{R}$. The associated realization mapping is then defined for all $\boldsymbol{\theta} \equiv (W, B) \in \mathcal{P}_{\mathbf{a},\infty}$ by
$$\mathcal{F}(\boldsymbol{\theta}) = \left(f : x \mapsto 2(1 + \exp(W^T x + B))^{-1} - 1\right).$$
- Kernel classifiers, induced by maps of the form $x \mapsto \sum_{i=1}^n \alpha_i K(x_i, \cdot)$ where K is a Mercer kernel defined on $\mathcal{X} \times \mathcal{X}$ and $x_1, \ldots, x_n \in \mathcal{X}$ are the training data points. The architecture and associated parametrization in this case are respectively $\mathbf{a} = (n, 1)$ and $\mathcal{P}_{\mathbf{a},\infty} = \mathbb{R}^n \times \mathbb{R}$. The associated realization mapping is then defined for all $\boldsymbol{\theta} \equiv (\alpha, B) \in \mathcal{P}_{\mathbf{a},\infty}$ by
$$\mathcal{F}(\boldsymbol{\theta}) = \left(f : x \mapsto \sum_{i=1}^n \alpha_i K(x_i, \cdot)\right).$$
Note that for this hypothesis space, the bias parameter is not used. Hence we have an effective number $P(\mathbf{a}) = n$ of parameters.

In all of the remaining text, we will slightly abuse notation and identify any $f \in \mathcal{H}_{\mathcal{F},\mathbf{a},R}$, which is defined on all of $\mathbb{R}^d$, with its restriction to the unit cube $\mathcal{X}$.

2.2.2 Clipping the function outputs

To study the generalization error of our hypothesis space $\mathcal{H}_{\mathcal{F},\mathbf{a},R}$, it is necessary to ensure that the functions within have uniformly bounded supremum norm, as the complexity may grow unboundedly otherwise. A simple way to guarantee this is the following: given a *clipping constant* $D > 0$, we compose all the functions in $\mathcal{H}_{\mathcal{F},\mathbf{a},R}$ with $\mathrm{clip}_D : \mathbb{R} \to \mathbb{R}$ defined by
$$\mathrm{clip}_D(x) = \begin{cases} D & \text{if } x \geqslant D \\ x & \text{if } -D \leqslant x \leqslant D \\ -D & \text{if } x \leqslant -D \end{cases} \tag{6}$$

Although the clipping operator ensures boundedness of outputs, one may worry about it negatively affecting the approximation power of the hypothesis space. The following lemma guarantees that as long as the clipping constant D is chosen larger than $\|\eta\|_{L^\infty(\mathcal{X})}$, the approximation error does not increase.

Lemma 1. *Let $f^* \in L^\infty(\mathcal{X}, \mathbb{R})$ and $D \geqslant \|f^*\|_{L^\infty(\mathcal{X},\mathbb{R})}$. For any $f \in L^\infty(\mathcal{X}, \mathbb{R})$, we have*
$$\|\mathrm{clip}_D \circ f - f^*\|_{L^\infty(\mathcal{X},\mathbb{R})} \leqslant \|f - f^*\|_{L^\infty(\mathcal{X},\mathbb{R})}$$
where clip_D is as defined in (6).

Proof. By the assumption on D, we have $f^*(x) = \mathrm{clip}_D \circ f^*(x)$ for almost all $x \in \mathcal{X}$. Hence, by 1-Lipschitz continuity of clip_D,
$$|\mathrm{clip}_D \circ f(x) - f^*(x)| = |\mathrm{clip}_D \circ f(x) - \mathrm{clip}_D \circ f^*(x)| \leqslant |f(x) - f^*(x)|$$
holds for almost all x, and the conclusion follows by definition of the essential supremum. □

Likewise, it is immediate to see that for any $D > 0$, f and $\mathrm{clip}_D \circ f$ induce the same classifier. Since the composition with clip_D does not affect the number of free parameters, and $\|\eta\|_{L^\infty(\mathcal{X})} \leqslant 1$, we will fix $D = 1$ and assume in the following that all functions we consider have been composed with clip_D, without making it explicit in the notation, which is justified thanks to Lemma 1 above.

2.2.3 ℓ_p Regularization

Lastly, we fix $0 < p < \infty$ and regularize the objective (2) with an ℓ_p penalty term.

We thus define the regularized empirical risk as

$$\widehat{\mathcal{R}}_{\ell,\lambda}(\boldsymbol{\theta}) := \frac{1}{n}\sum_{i=1}^{n}(f(x_i;\boldsymbol{\theta}) - y_i)^2 + \lambda|\boldsymbol{\theta}|_p^p \tag{7}$$

where, for a parameter vector $\boldsymbol{\theta} = ((W_l, B_l))_{l=1}^L \in \mathcal{P}_{\mathbf{a},\infty}$,

$$|\boldsymbol{\theta}|_p^p := \sum_{l=1}^{L} |W_l|_p^p + |B_l|_p^p.$$

This penalty is very popular in practical applications. For $p = 2$, in which case it is often referred to as weight decay, it is known to help training and improve generalization (Krogh & Hertz, 1991). Similarly, $p = 1$ is a popular choice as it tends to promote sparse solutions, which are less expensive to store and more efficient to compute with (Candes et al., 2008). Although not as common, taking $0 < p < 1$ also has its merits, as it can be used as a differentiable approximation of the ℓ_0 penalty, which induces very sparse models but is not compatible with standard gradient-based optimization algorithms (Louizos et al., 2017).

For fixed $R > 0$ and $\lambda \geqslant 0$, the λ-ERM problem (3) thus consists in finding $\widehat{\boldsymbol{\theta}}_\lambda$ satisfying

$$\widehat{\boldsymbol{\theta}}_\lambda \in \operatorname*{argmin}_{\boldsymbol{\theta}\in\mathcal{P}_{\mathbf{a},R}} \widehat{\mathcal{R}}_{\ell,\lambda}(\boldsymbol{\theta}). \tag{8}$$

Note that the objective (8) is, for most hypothesis spaces, highly non-convex. This implies that the set of minimizers is generally not reduced to a singleton. Therefore, we will only consider the minimum norm solutions throughout this paper, i.e. we only consider

$$\widehat{\boldsymbol{\theta}}_\lambda \in \operatorname{argmin}\left\{ |\boldsymbol{\theta}|_\infty, \text{ for } \boldsymbol{\theta} \in \operatorname*{argmin}_{\boldsymbol{\theta}\in\mathcal{P}_{\mathbf{a},R}} \widehat{\mathcal{R}}_{\ell,\lambda}(\boldsymbol{\theta}) \right\}. \tag{9}$$

2.3 Technical Assumptions

In this section, we present the technical assumptions under which we establish our main results.

(A1) The Bayes regression function $\eta : x \mapsto \mathbb{E}[Y \mid X = x]$ satisfies Tsybakov's *low-noise condition*: there exists a *noise exponent* $q > 0$ and a positive constant $C > 0$ such that

$$\mathbb{P}\left(|\eta(X)| \leqslant \delta\right) \leqslant C\delta^q \quad \text{for all } \delta > 0.$$

At times, we will also refer to Assumption **(A1)** as the "weak-margin" condition, using the two terms interchangeably and without particular preference. The so-called *hard-margin condition*, can be thought of as a "limit" of the low-noise condition when $q = \infty$:

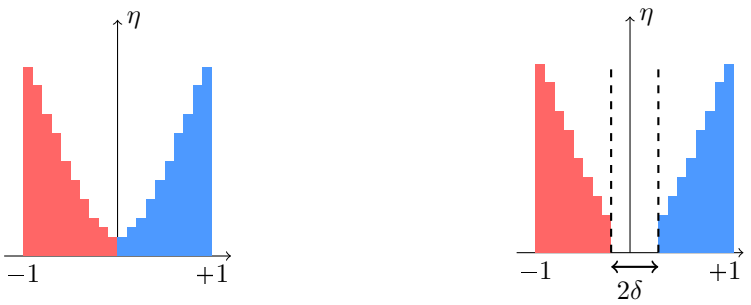
(A2) The Bayes regression function $\eta : x \mapsto \mathbb{E}[Y \mid X = x]$ satisfies the *hard-margin condition*: there exists $\delta > 0$ such that

$$\mathbb{P}\left(|\eta(X)| > \delta\right) = 1.$$

Assumption **(A2)** was originally introduced in (Mammen & Tsybakov, 1999) as a characterization of classification problems for which the two classes are in some sense "separable", and has been repeatedly shown in the literature to lead to faster rates of convergence for various hypothesis classes.

Consider the regularized population risk $\mathcal{R}_{\ell,\lambda}$, which is given for all $\lambda \geqslant 0$ and $\boldsymbol{\theta} \in \mathcal{P}_{\mathbf{a},R}$ by

$$\mathcal{R}_{\ell,\lambda}(\boldsymbol{\theta}) := \mathbb{E}_{(x,y)\sim\rho}\left[(f(x;\boldsymbol{\theta}) - y)^2\right] + \lambda|\boldsymbol{\theta}|_p^p, \tag{10}$$



(a) Weak Margin Condition (b) Hard Margin Condition

Figure 1: Visualization of margin conditions through histograms of values taken by η.

which for convenience, we also denote $\mathcal{R}_\ell$ whenever $\lambda = 0$. We will assume that $\mathcal{R}_{\ell,\lambda}$ satisfies the two following assumptions:

(A3) For any $R, \lambda \geqslant 0$, the set $\operatorname{argmin}_{\boldsymbol{\theta} \in \mathcal{P}_{\mathbf{a},R}} \mathcal{R}_{\ell,\lambda}$ has finitely many elements.

Assumption **(A3)** ensures that all the global minimizers of $\mathcal{R}_{\ell,\lambda}$ are isolated and can not "cluster" arbitrarily close to each other. Note that $\mathcal{R}_{\ell,\lambda}$ being a continuous function of $\boldsymbol{\theta}$ defined over a compact set, its argmin is necessarily not empty.

We now briefly introduce some necessary terminology.

Definition 1 (Analytic functions)**.** *A real-valued function f defined on an open set $U \subseteq \mathbb{R}^k$ is said to be real analytic on U if for all $u \in U$, there exists an open ball B containing u and real coefficients $(c_\alpha)_{\alpha \in \mathbb{N}_0^k}$ such that $f(x) = \sum_{\alpha \in \mathbb{N}_0^k} c_\alpha (x-u)^\alpha$ for all $x \in B$. An $\mathbb{R}^\ell$-valued function $f = (f_1, \ldots, f_\ell)^T$ is said to be analytic if each of its components is analytic.*

Definition 2 (Semialgebraic sets and functions)**.** *A set $S \subseteq \mathbb{R}^k$ is said to be semialgebraic (Bochnak et al., 1998) if it can be written in the form*

$$S = \bigcup_{i=1}^{s} \bigcap_{j=1}^{t} \left\{ x \in \mathbb{R}^k : p_{ij}(x) = 0, q_{ij}(x) > 0 \right\},$$

where $s, t \in \mathbb{N}$ and p_{ij}, q_{ij} are polynomial functions for all $1 \leqslant i \leqslant s, 1 \leqslant j \leqslant t$.
We call a function $f : \mathbb{R}^k \to \mathbb{R}$ semialgebraic if its graph $\{(x, f(x)) : x \in \mathbb{R}^k\}$ is semialgebraic.

Definition 3 (Subanalytic sets and functions)**.** *A set $S \subseteq \mathbb{R}^k$ is said to be subanalytic (Shiota, 1997) if each point of $\mathbb{R}^k$ has a neighborhood U such that $S \cap U$ is a finite union of sets of the form $Imf_1 \backslash Imf_2$, where f_1 and f_2 are analytic $\mathbb{R}^k$ valued proper maps defined on analytic manifolds.*
We call a function $f : \mathbb{R}^k \to \mathbb{R}$ subanalytic if its graph $\{(x, f(x)) : x \in \mathbb{R}^k\}$ is subanalytic.

The concept of semialgebraic function is relatively straightforward: it refers to functions whose graph can be described using finite unions and intersections of polynomial equalities and inequalities. Subanalyticity extends this concept in some sense by allowing not just polynomial, but analytic equalities and inequalities, allowing for a much richer range of function graphs (Shiota, 1997; Bochnak et al., 1998).

As it turns out, most choices of loss functions, hypothesis spaces, and regularizations will lead to subanalytic, if not semialgebraic, population and empirical risk. For DNN hypothesis spaces, (Zeng et al., 2019) give sufficient conditions on the choice of loss, activation function and regularization to ensure subanalyticity of the resulting risk. Those conditions are verified for most practically used models, including ReLU, sigmoid, tanh networks with ℓ_p regularization and square, logistic or cross-entropy loss (Zeng et al., 2019, Appendix B).

The last notion we introduce is that of a *limiting subdifferential*:

Definition 4 (limiting subdifferential)**.** *Given a function $f : \mathbb{R}^k \to \mathbb{R}$ and $x \in \mathbb{R}^k$, we call* Fréchet subdifferential *of f at x and denote $\hat{\partial} f(x)$ the set of all vectors $v \in \mathbb{R}^k$ which satisfy the following:*

$$\lim_{y \neq x} \inf_{y \to x} \frac{f(y) - f(x) - v^T(y-x)}{|y-x|_2} \geqslant 0.$$

The limiting subdifferential of f at $x \in \mathbb{R}^k$, which we denote $\partial f(x)$ is then defined as (Rockafellar & Wets, 1998):

$$\partial f(x) := \left\{ v \in \mathbb{R}^k : \exists (x_i)_{i \in \mathbb{N}} \to x, f(x_i) \to f(x), v_i \in \hat{\partial} f(x_i) \to v \right\}.$$

The limiting subdifferential ∂f generalizes the concept of subdifferential to functions with even less regularity. Just like the former, it necessarily contains zero at any extremum of f, and coincides with its gradient ∇f whenever f is differentiable (Rockafellar & Wets, 1998).

We can now introduce the following assumption:

(A4) For any $\lambda \geqslant 0$, the regularized population risk $\mathcal{R}_{\ell,\lambda} : \mathbb{R}^{P(\mathbf{a})} \to \mathbb{R}$ is subanalytic, and for any $\boldsymbol{\theta} \in \mathbb{R}^{P(\mathbf{a})}$, $\partial \mathcal{R}_{\ell,\lambda}(\boldsymbol{\theta})$ is not empty.
Furthermore, for any $R > 0$, the restriction $\mathcal{R}_{\ell,\lambda} : [-R, R]^{P(\mathbf{a})} \to \mathbb{R}$ is Lipschitz continuous, and we will denote by $\mathrm{Lip}_R(\mathcal{R}_{\ell,\lambda}) \equiv \mathrm{Lip}(\mathcal{R}_{\ell,\lambda})$ its Lipschitz constant.

In light of the preceding discussion, we observe that Assumption **(A4)** is very mild. Both subanalyticity and (limiting) subdifferentiability are satisfied in most practical setups, and local Lipschitz-ness is also guaranteed as long as, e.g., the parametric hypothesis space consists of compositions of locally Lipschitz mappings, which is often the case as well.

The need for Assumption **(A4)** is motivated by the following theorem, due to (Bolte et al., 2007):

Theorem 1 (Adapted from Corollary 16 in (Bolte et al., 2007))**.** *If $f : \mathbb{R}^k \to \mathbb{R}$ is a lower semi-continuous, globally subanalytic function, and $f(x_0) = 0$, then there exist $c, \rho > 0$ and $0 < \kappa < 1$ such that for all $x \in \{x \in \mathbb{R}^k : 0 < |f(x)| < \rho\}$, we have*

$$c|f(x)|^\kappa \leqslant |x^*|_2, \quad \text{for all } x^* \in \partial f(x). \tag{11}$$

Functions satisfying the conclusion of Theorem 1 are said to satisfy the Kurdyka-Lojasiewicz (KL) property, named after pioneering works of Lojasiewicz (Lojasiewicz, 1963) and Kurdyka (Kurdyka, 1998). This property is fundamental in non-convex optimization, as it is a key ingredient when proving convergence guarantees for first-order methods in general settings (Li & Pong, 2018). This property has also been used in recent works to obtain convergence rates for first-order optimization in Deep Learning (Xu et al., 2024; Zeng et al., 2019).

As a consequence of assumptions **(A3)** and **(A4)**, we have the following proposition, whose proof we provide in the last section:

Proposition 1. *Assume that **(A3)** and **(A4)** hold. Fix $R > 0$. There exist constants $K, \Lambda, \rho > 0$, and $r > 1$, which do not depend on R, such that for all $0 \leqslant \lambda \leqslant \Lambda$, $0 < t < \rho/(2\,\mathrm{Lip}_R(\mathcal{R}_{\ell,\lambda}))$, and $\boldsymbol{\theta}_\lambda \in \mathrm{argmin}_{\boldsymbol{\theta} \in \mathcal{P}_{\mathbf{a},R}} \mathcal{R}_{\ell,\lambda}$*

$$\inf_{\boldsymbol{\theta} \in \mathcal{P}_{\mathbf{a},R} : \mathrm{dist}(\boldsymbol{\theta}, \mathrm{argmin}\, \mathcal{R}_{\ell,\lambda}) \geqslant t} \mathcal{R}_{\ell,\lambda}(\boldsymbol{\theta}) - \mathcal{R}_{\ell,\lambda}(\boldsymbol{\theta}_\lambda) \geqslant K t^r, \tag{12}$$

where $\mathrm{dist}(a, A)$ denotes the ℓ_∞ distance between a vector $a \in \mathbb{R}^k$ and a set $A \subseteq \mathbb{R}^k$.

A condition similar to the conclusion of Lemma 1 is often assumed to hold in the empirical process theory literature, where it is referred to as a *well-separation assumption* and is used to prove consistency of M-estimators (Van der Vaart, 2000; Sen, 2018). In that context, the Kt^r lower bound in equation (12) is replaced by $\psi(t)$, for an unknown function ψ which is merely assumed to be positive for small $t > 0$. Our Proposition 1 shows that this assumption does in fact hold for the majority of cases, and explicitly quantifies this lower bound, which in turn will allow us to carry out our error analysis.

Denote respectively by $\widehat{\mathcal{R}}_{\ell,n}$ and $\mathcal{R}_\ell$ the unregularized ($\lambda = 0$) versions of $\widehat{\mathcal{R}}_{\ell,\lambda}$ and $\mathcal{R}_{\ell,\lambda}$ defined in (7) and (10), where the dependence on n is made explicit in the notation. Denote by

$$\operatorname{argmin}^* \widehat{\mathcal{R}}_{\ell,n} := \left\{ \operatorname{argmin} |\boldsymbol{\theta}|_\infty : \boldsymbol{\theta} \in \operatorname*{argmin}_{\boldsymbol{\theta} \in \mathcal{P}_{\mathbf{a},\infty}} \widehat{\mathcal{R}}_{\ell,n} \right\} \tag{13}$$

the minimum-norm minimizers of $\widehat{\mathcal{R}}_{\ell,n}$, taken as a function defined on the unrestricted parameter space $\mathcal{P}_{\mathbf{a},\infty}$ (4). We will assume the following:

(A5) The sets $\operatorname{argmin}_{\boldsymbol{\theta} \in \mathcal{P}_{\mathbf{a},\infty}} \mathcal{R}_\ell$ and $\operatorname{argmin}_{\boldsymbol{\theta} \in \mathcal{P}_{\mathbf{a},\infty}} \widehat{\mathcal{R}}_{\ell,n}$ are (almost surely) not empty, and there exists a constant $R_0 > 0$ such that almost surely over all possible i.i.d. draws $(x_i, y_i)_{i \geqslant 1}$ with distribution ρ, we have

$$\sup_{n \geqslant 1} \left\{ |\boldsymbol{\theta}|_\infty : \boldsymbol{\theta} \in \operatorname{argmin}^* \widehat{\mathcal{R}}_{\ell,n} \right\} \leqslant R_0. \tag{14}$$

Besides the requirement that minimizers exist, which is standard and often implicitly assumed when studying empirical risk minimization, Assumption **(A5)** states that the minimum-norm solutions of the unregularized ERM problem (9) almost surely do not run off to infinity as the sample size n increases. Although it is intuitively expected that minimum-norm solutions do not diverge, in practice such an event could have a small, positive probability. Assumption **(A5)** thus requires ρ to give zero measure to "pathological" datasets where such a thing happens.

Under Assumption **(A5)** we can uniformly bound the norms of λ-ERM solutions, as well as that of approximation error minimizers:

Lemma 2. *Assume that **(A5)** holds with upper bound $R_0 > 0$. The following is true:*

i. Almost surely for all $n \geqslant 1, \lambda > 0$ the sets $\operatorname{argmin}_{\boldsymbol{\theta} \in \mathcal{P}_{\mathbf{a},\infty}} \widehat{\mathcal{R}}_{\ell,\lambda}$ are not empty, and we have the inequality

$$\sup_{\lambda \geqslant 0, n \geqslant 1} \left\{ |\boldsymbol{\theta}|_\infty : \boldsymbol{\theta} \in \operatorname{argmin}^* \widehat{\mathcal{R}}_{\ell,\lambda} \right\} \leqslant R_0 P(\mathbf{a})^{1/p},$$

where $P(\mathbf{a})$ denotes the number of parameters in the architecture $\mathcal{P}_{\mathbf{a},\infty}$ and $\operatorname{argmin}^* \widehat{\mathcal{R}}_{\ell,\lambda}$ is defined as in (13).

ii. There exists $\boldsymbol{\theta}^* \in \mathcal{P}_{\mathbf{a},\infty}$ such that

$$|\boldsymbol{\theta}^*|_\infty \leqslant R_0 \text{ and } \boldsymbol{\theta}^* \in \operatorname*{argmin}_{\boldsymbol{\theta} \in \mathcal{P}_{\mathbf{a},\infty}} \|\mathcal{F}(\boldsymbol{\theta}) - \eta\|_{L^2(\rho_X)}.$$

Lemma 2, whose proof we defer to the last section, guarantees that if the parameter bound R is chosen equal to $R_0 P(\mathbf{a})$ or greater, the hypothesis space $\mathcal{H}_{\mathcal{F},\mathbf{a},R}$ will both contain global solutions for the λ-ERM problem (9), and global minimizers of the $L^2(\rho_X)$ approximation error.

3 Main results

3.1 A general upper bound for the excess risk of Parametric Classifiers

Before stating our main results, we introduce a few useful definitions. The first being that of an *ε-covering* :

Definition 5 (ε-cover)**.** *Let $\varepsilon > 0$ and $\mathcal{G} \subseteq L^\infty(\mathcal{X}, \mathbb{R})$ be a family of functions. Any finite collection of functions $g_1, \ldots, g_N \in L^\infty(\mathcal{X}, \mathbb{R})$ with the property that for any g in $\mathcal{G}$ there is an index $j \equiv j(g)$ such that*

$$\|g - g_j\|_{L^\infty} \leqslant \varepsilon$$

is called an ε-covering (or cover) of $\mathcal{G}$ with respect to $\|\cdot\|_{L^\infty}$.

For a given ε, we can think of the cardinality of an ε-cover as a measure of complexity for the family $\mathcal{G}$. This motivates the definition of a *covering number* :

Definition 6 (ε-covering number)**.** *Let $\varepsilon > 0$ and $\mathcal{G} \subseteq L^\infty(\mathcal{X}, \mathbb{R})$. We denote by $\mathbf{Cov}(\mathcal{G}, \|\cdot\|_{L^\infty}, \varepsilon)$ the size of the smallest ε-cover of $\mathcal{G}$ with respect to $\|\cdot\|_{L^\infty}$, with the convention $\mathbf{Cov}(\mathcal{G}, \|\cdot\|_{L^\infty}, \varepsilon) := \infty$ when no finite cover exists. $\mathbf{Cov}(\mathcal{G}, \|\cdot\|_{L^\infty}, \varepsilon)$ will be called an ε-covering number of $\mathcal{G}$ with respect to $\|\cdot\|_{L^\infty}$.*

For $\mathcal{G} = \mathcal{H}_{\mathcal{F},\mathbf{a},R}$, we will abbreviate and denote

$$\mathbf{Cov}(\mathcal{H}_{\mathcal{F},\mathbf{a},R}, \|\cdot\|_{L^\infty}, \varepsilon) =: \mathbf{Cov}_\infty(\mathcal{H}, \varepsilon).$$

The above two quantities, whose definitions are adapted from (Györfi et al., 2002), are ubiquitous in the learning theory literature, as they give a lot of information on the statistical properties of our estimators.

Another related measure of complexity for parametric hypothesis spaces is the Lipschitz constant of the realization map (5):

Definition 7. *Given a parametrization $\mathcal{P}_{\mathbf{a},R}$, recall the definition of the realization mapping $\mathcal{F} : \mathcal{P}_{\mathbf{a},R} \to \mathcal{C}(\mathcal{X}, \mathbb{R})$ (5). We will denote by*

$$\mathrm{Lip}_R(\mathcal{F}) := \sup_{\substack{\boldsymbol{\theta},\boldsymbol{\theta}' \in \mathcal{P}_{\mathbf{a},R} \\ \boldsymbol{\theta} \neq \boldsymbol{\theta}'}} \frac{\|\mathcal{F}(\boldsymbol{\theta}) - \mathcal{F}(\boldsymbol{\theta}')\|_{L^\infty(\mathcal{X})}}{|\boldsymbol{\theta} - \boldsymbol{\theta}'|_\infty}$$

its Lipschitz constant.

Intuitively, the Lipschitz constant of the realization map estimates the complexity of the induced hypothesis space in the sense that it controls how different two realizations can be given that their parametrizations are close. For this reason, the problem of estimating a Neural Network's Lipschitz constant has garnered a lot of interest over recent years (Fazlyab et al., 2019; Virmaux & Scaman, 2018).

We are now ready to state our main result, which gives an upper bound on the excess risk of parametric classifiers under our setting.

Theorem 2. *Assume that Assumptions **(A3)**, **(A4)** and **(A5)** hold. Fix an architecture $\mathbf{a}$ with parameter bound $R \equiv R_0 P(\mathbf{a})^{1/p}$, where $R_0 > 0$ satisfies (14), and denote*

$$\varepsilon_{approx} := \inf_{f \in \mathcal{H}_{\mathcal{F},\mathbf{a},R}} \|f - \eta\|_{L^2(\rho_X)}$$

the approximation error of $\mathcal{H}_{\mathcal{F},\mathbf{a},R}$. We have the following excess risk bounds:

- *If the low-noise condition **(A1)** holds, then for all $\delta > \varepsilon_{approx}$ and $0 < \nu < \delta$, any minimum-norm solution $\widehat{\boldsymbol{\theta}}_\lambda$ of the λ-ERM problem (9) with $0 \leqslant \lambda < (\delta - \varepsilon_{approx})^2 (P(\mathbf{a}) 2^p R^p)^{-1}$ satisfies for all $n \geqslant 1$:*

$$\begin{aligned} \mathcal{R}(\operatorname{sign} f(\cdot; \widehat{\boldsymbol{\theta}}_\lambda)) - \mathcal{R}^* \leqslant{}& \varepsilon_{approx} + \sqrt{\lambda P(\mathbf{a}) 2^p R^p} + C\delta^q \\ &+ (\delta - \nu)^{-2} \left(\varepsilon_{approx} + \sqrt{\lambda P(\mathbf{a}) 2^p R^p} \right)^2 \\ &+ 4\, \mathbf{Cov}_\infty \left(\mathcal{H}, \frac{K\nu^r}{24\, \mathrm{Lip}_{2R}(\mathcal{F})^{1+r}} \right) \exp\left(\frac{-nK^2\nu^{2r}}{288\, \mathrm{Lip}_{2R}(\mathcal{F})^{2r}} \right) \end{aligned} \tag{15}$$

- *If the hard-margin condition **(A2)** holds with margin $\delta > 0$, and $\varepsilon_{approx} < \delta$, then for all $0 < \nu < \delta$, any minimum-norm solution $\widehat{\boldsymbol{\theta}}_\lambda$ of the λ-ERM problem (9) with $0 \leqslant \lambda < (\delta - \varepsilon_{approx})^2 (P(\mathbf{a}) 2^p R^p)^{-1}$*

satisfies for all $n \geqslant 1$*:*

$$\begin{aligned}\mathcal{R}(\operatorname{sign} f(\cdot;\widehat{\boldsymbol{\theta}}_\lambda)) - \mathcal{R}^* \leqslant{}& \varepsilon_{approx} + \sqrt{\lambda P(\mathbf{a}) 2^p R^p} \\ &+ (\delta-\nu)^{-2}\left(\varepsilon_{approx} + \sqrt{\lambda P(\mathbf{a}) 2^p R^p}\right)^2 \\ &+ 4\,\mathbf{Cov}_\infty\left(\mathcal{H}, \frac{K\nu^r}{24\operatorname{Lip}_{2R}(\mathcal{F})^{1+r}}\right)\exp\left(\frac{-nK^2\nu^{2r}}{288\operatorname{Lip}_{2R}(\mathcal{F})^{2r}}\right)\end{aligned} \tag{16}$$

The bounds in Theorem 2 differ significantly from those typically encountered in the literature for similar problems. While the appearance of the approximation error term is standard, the remaining terms are not. Specifically, the bounds include three terms that capture the interplay between the noise condition, the approximation error, and the regularization parameter, as well as a final exponential term that can be interpreted as the *statistical error* in classical learning theory.

At first glance, it is not immediately evident that Theorem 2 offers improved rates of convergence for ERM-based binary classification. Although the non-exponential terms in (15) and (16) can often be effectively controlled when the regression function η belongs to an appropriate function space, the exponential term presents a challenge. In some cases, this term may fail to vanish as the architecture size $\mathbf{a}$ grows with the sample size n. In the next section, however, we demonstrate that for DNN hypothesis spaces, Theorem 2 can indeed yield fast rates of convergence.

3.2 Super Fast Rates of Convergence for Deep ReLU Networks with Distribution-Adapted Smoothness

We now focus our attention on the case where the hypothesis space $\mathcal{H}_{\mathcal{F},\mathbf{a},R}$ consists of Fully Connected Neural Networks (FCNNs) with ReLU activation. That is, we let $\sigma : x \mapsto \max\{0, x\}$ be the ReLU activation function, and for a given architecture $\mathbf{a} = (a_0, \ldots, a_L) \in \mathbb{N}^{L+1}$, and parameter vector $\boldsymbol{\theta} = ((W_1, B_1), \ldots, (W_L, B_L)) \in \mathcal{P}_{\mathbf{a},\infty}$, we define the affine maps

$$T_\ell : \mathbb{R}^{a_{\ell-1}} \to \mathbb{R}^{a_\ell}, \quad x \mapsto W_\ell x + B_\ell, \quad (1 \leqslant \ell \leqslant L).$$

We then define for all $\boldsymbol{\theta} \in \mathcal{P}_{\mathbf{a},\infty}$ the realization mapping as

$$\mathcal{F}_{NN}(\boldsymbol{\theta}) := \Big[x \mapsto T_L \circ \sigma \circ T_{L-1} \circ \cdots \circ \sigma \circ T_1\Big], \tag{17}$$

where σ is understood component-wise when applied to vectors.

For notational convenience, we will denote by $\mathcal{NN}(\mathbf{a}, W, L, R)$ the hypothesis space $\mathcal{H}_{\mathcal{F}_{NN},\mathbf{a},R}$ induced by $\mathcal{F}_{NN}$. As the width and depth of a Neural Network architecture play a crucial role in bounding its complexity, it is advisable to make them clearly visible in the notation.

3.2.1 Upper bounds

We begin by stating our estimates on FCNN complexity measures:

Lemma 3 (Theorem 2.6 in (Berner et al., 2020))**.** *For any architecture vector* $\mathbf{a}$ *with depth* $L \in \mathbb{N}$ *and width* $W \in \mathbb{N}$*, and any parameter bound* $R > 0$*, we have the upper bound on* $\operatorname{Lip}_R(\mathcal{F}_{NN})$*:*

$$\sup_{\substack{\boldsymbol{\theta},\boldsymbol{\theta}' \in \mathcal{P}_{\mathbf{a},R} \\ \boldsymbol{\theta} \neq \boldsymbol{\theta}'}} \frac{\|\mathcal{F}_{NN}(\boldsymbol{\theta}) - \mathcal{F}_{NN}(\boldsymbol{\theta}')\|_{L^\infty(\mathcal{X})}}{|\boldsymbol{\theta} - \boldsymbol{\theta}'|_\infty} \leqslant 2L^2 R^{L-1} W^L.$$

The above inequality, which is tight, shows that the Lipschitz constant of $\mathcal{F}_{NN}$ grows *exponentially* with depth. As one could expect, the covering number behaves similarly:

Lemma 4. *Let $\varepsilon > 0$. For any DNN architecture $\mathbf{a}$ with depth $L \in \mathbb{N}$ and width $W \in \mathbb{N}$, and any parameter bound $R > 0$, we have the upper bound*

$$\mathbf{Cov}_\infty(\mathcal{NN}, \varepsilon) \leqslant \left(1 + \frac{2R \operatorname{Lip}_R(\mathcal{F}_{NN})}{\varepsilon}\right)^{P(\mathbf{a})}$$

Proof of Lemma 4. Let $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathcal{P}_{\mathbf{a},R}$. Because of the inequality

$$\|f(\cdot;\boldsymbol{\theta}) - f(\cdot;\boldsymbol{\theta}')\|_{L^\infty(\mathcal{X})} \leqslant \operatorname{Lip}_R(\mathcal{F}_{NN})|\boldsymbol{\theta} - \boldsymbol{\theta}'|_\infty,$$

we get that $\mathbf{Cov}_\infty(\mathcal{NN}, \varepsilon)$ is bounded by the number of ℓ_∞ balls of radius $\varepsilon/\operatorname{Lip}_R(\mathcal{F}_{NN})$ needed to cover the hypercube $[-R, R]^{P(\mathbf{a})}$. It is straightforward to check that the collection of such balls centered at the points

$$-R\vec{\mathbf{1}} + \varepsilon\vec{\mathbf{k}}, \quad \text{where } \vec{\mathbf{k}} = [k_1, k_2, \dots k_{P(\mathbf{a})}]^T, \text{ and } k_i \in \left\{0, 1, \dots, \left\lceil \frac{2R \operatorname{Lip}_R(\mathcal{F}_{NN})}{\varepsilon} \right\rceil\right\},$$

where $\vec{\mathbf{1}}$ is the vector whose entries are all ones, covers $[-R, R]^{P(\mathbf{a})}$ and has $\lceil 2R \operatorname{Lip}_R(\mathcal{F}_{NN})/\varepsilon \rceil^{P(\mathbf{a})}$ elements. The proof is thus complete. □

In light of the above estimates, we see that the only way for Theorem 2 to yield any useful results is for the approximation error to decay faster than its complexity increases with respect to the network's size. A rich literature on DNN approximation theory has shown that for target functions in suitable smoothness spaces, such as Sobolev, Hölder, Besov or Korobov spaces, fast rates of approximation were possible for ReLU DNNs (Yarotsky, 2017; Suzuki, 2018; Petersen & Voigtlaender, 2018; Mao & Zhou, 2022). In particular, we have the following result due to (Lu et al., 2021), which provides exact approximation bounds of s times continuously differentiable functions by Deep ReLU FCNNs:

Theorem 3 (Theorem 1.1 from (Lu et al., 2021))**.** *Let $s \in \mathbb{N}$ and $h \in \mathcal{C}^s(\mathcal{X})$. For any $W_0, L_0 \in \mathbb{N}$ there exists a neural network $f(\cdot;\boldsymbol{\theta}) \in \mathcal{NN}(\mathbf{a}, W, L, R)$ with width $W(\mathbf{a}) = C_1(W_0 + 2)\log_2(8W_0)$ and depth $L(\mathbf{a}) = C_2(L_0 + 2)\log_2(4L_0) + 2d$ such that*

$$\|f(\cdot;\boldsymbol{\theta}) - h\|_{L^\infty(\mathcal{X})} \leqslant C_3\|h\|_{\mathcal{C}^s(\mathcal{X})} W_0^{-2s/d} L_0^{-2s/d},$$

where $C_1 = 17s^{d+1}3^d d$, $C_2 = 18s^2$ and $C_3 = 85(s+1)^d 8^s$.

We now introduce the following Assumption **(A6)**, which generalizes Theorem 3 to the $L^2(\rho_X)$ norm and incorporates the interaction between the regression function η and the marginal data distribution:

(A6) Let η be the Bayes regression function. There exist a natural number $L_0 \in \mathbb{N}$, and a smoothness parameter $s > 0$, such that for any $W_0 \in \mathbb{N}_{\geqslant 2}$, one can find a neural network $f(\cdot;\boldsymbol{\theta}) \in \mathcal{NN}(\mathbf{a}, W, L, \infty)$ with width $W(\mathbf{a}) = C_1 W_0 \log_2(W_0)$ and depth $L(\mathbf{a}) = C_2 L_0 \log_2(L_0) + 2d$ such that

$$\|f(\cdot;\boldsymbol{\theta}) - \eta\|_{L^2(\rho_X)} \leqslant C_3 W_0^{-2s/d},$$

where $C_1 = ds^d$, $C_3 = (3s)^d 8^s d C_\eta$ for some $C_\eta > 0$ depending on η only, and the positive quantity $C_2 \equiv C_2(s)$ is such that $C_2(s)/s \to 0$ as $s \to \infty$.

Assumption **(A6)**, which we refer to as *distribution-adapted smoothness*, extends Theorem 3 by shifting the error metric to the $L^2(\rho_X)$ norm. While it shares some similarities with Theorem 3, the assumption places additional constraints on the depth parameter L, requiring it to grow at a slightly slower rate with respect to the smoothness parameter s, specifically at $o(s)$. This slower growth requirement is motivated by prior results showing that, under a uniform measure, achieving L^2 approximation error rates of $(WL)^{-2s/d}$ for non-linear C^2 functions requires a depth proportional to s/d (Safran & Shamir, 2017; Voigtlaender & Petersen, 2019). Assumption **(A6)** accounts for this by considering not only the regularity of η but also the interaction between η and the marginal data distribution ρ_X. This interaction allows for slightly faster

approximation rates when ρ_X is "adapted" to η. In Section 6.4, we give examples of discrete marginal distributions ρ_X for which Assumption **(A6)** holds with $C_2(s) = 0$ for all $s > 0$.

This approximation error bound directly quantifies the width and depth required to achieve a given error level, leading to the following excess risk bounds for deep FCNN classifiers:

Theorem 4. *Assume that Assumptions **(A3)**, **(A4)**, **(A5)** and **(A6)** hold, and let $\alpha > 0$ be a desired order of convergence. For any $n \in \mathbb{N}$, there exists a FCNN architecture $\mathbf{a}_n$ with width W_n, depth L_n and parameter bound R_n satisfying*

$$\begin{cases} W_n &= \tilde{\mathcal{O}}(n^{\alpha d/2rs}) \\ L_n &= C_2 L_0 \log_2 L_0 + 2d \\ R_n &= R_0 \left[L_n(W_n^2 + W_n)\right]^{1/p} = \tilde{\mathcal{O}}(n^{\alpha d/ps}), \end{cases}$$

*where C_2 and L_0 are given in **(A6)**, R_0 is given in **(A5)**, and $\tilde{\mathcal{O}}$ hides polylogarithmic factors, such that the following holds:*

- *If the low-noise condition **(A1)** holds with noise exponent $q > 0$, and $\alpha < \left(1 + \frac{C_2 B_1 + B_2}{s}\right)^{-1}$, then any minimum-norm solution $\hat{\boldsymbol{\theta}}_\lambda$ of the λ-ERM problem (9) with $\lambda = \tilde{\mathcal{O}}\left(n^{-\frac{2\alpha(s+d)}{rs}}\right)$, satisfies the excess risk bound:*

$$\mathcal{R}\left(\operatorname{sign} f(\cdot; \hat{\boldsymbol{\theta}}_\lambda)\right) - \mathcal{R}^* \lesssim \max\left\{n^{-\frac{\alpha}{r}}, n^{-\frac{\alpha q}{2r}}\right\}, \tag{18}$$

  *where $q > 0$ is the noise exponent in Assumption **(A1)**, $r > 1$ is the "separation exponent" in (12), $B_1 = rdL_0 \log_2(L_0)(2+p)/p$, and $B_2 = 2rd\big[d(2+p) - 1\big]/p$.*

- *If the hard-margin condition **(A2)** holds with margin $\delta > 0$, $\alpha < \frac{s}{C_2 B_1 + B_2}$, and $n > (\delta/2)^{-\alpha/r}$, then any minimum-norm solution $\hat{\boldsymbol{\theta}}_\lambda$ of the λ-ERM problem (9) with $\lambda = \tilde{\mathcal{O}}\left(n^{-\frac{2\alpha(s+d)}{rs}}\right)$ satisfies the excess risk bound:*

$$\mathcal{R}\left(\operatorname{sign} f(\cdot; \hat{\boldsymbol{\theta}}_\lambda)\right) - \mathcal{R}^* \lesssim n^{-\frac{\alpha}{r}}, \tag{19}$$

  where $r > 1$ is the "separation exponent" in (12), $B_1 = rdL_0 \log_2(L_0)(2+p)/p$, and $B_2 = 2rd\big[d(2+p) - 1\big]/p$.

As the smoothness parameter s increases to infinity, we get, in the limit $\alpha \to \left(1 + \frac{C_2 B_1 + B_2}{s}\right)^{-1}$, a convergence rate of $\mathcal{O}\left(n^{-\frac{\min\{1, q/2\}}{r}}\right)$ under the low-noise condition **(A1)**. Since $r > 1$, the bound we get is thus slightly worse than the minimax optimal "fast-rate" of $\mathcal{O}\left(n^{-1}\right)$ which typically holds under Assumption **(A1)** when $q \to \infty$ (Kim et al., 2018; Audibert & Tsybakov, 2007).

On the other hand, under the hard-margin Assumption **(A2)**, the exponent α/r grows unbounded as the approximation speed s goes to ∞. Theorem 4 thus shows how deep FCNNs can leverage the hard-margin condition **(A2)** together with higher regularity to achieve potentially arbitrarily fast rates of convergence for the excess risk. A result which, to the best of our knowledge, is the first of its kind for this hypothesis space.

3.2.2 Lower bounds

A natural question which arises from Theorem 4, is whether the obtained rates of convergence can be improved further, in the minimax sense. Although we've seen that the rates obtained under the low-noise condition **(A1)** are slightly suboptimal, one may still wonder whether those obtained under Assumption **(A2)** can be improved further. The following Theorem gives a negative answer to this question.

Theorem 5. *Let $s > 0$ be fixed, and $C_2 \equiv C_2(s), B_1, B_2 > 0$ and $r > 1$ all be as in Theorem 4. For $q, \delta > 0$, denote respectively by $\mathcal{P}_{s,q}^{(1)}$ and $\mathcal{P}_{s,\delta}^{(2)}$ the sets of probability distributions on $\mathcal{X} \times \{-1, 1\}$ such that both Assumption **(A6)** and Assumption **(A1)** with exponent q (respectively Assumption **(A2)** with margin δ) hold. We have the following lower bounds:*

- *There exists a constant* $\kappa_1 > 0$ *such that for any* $n \in \mathbb{N}$*, and any classifier* $\hat{c}_n : (\mathcal{X} \times \{-1, 1\})^n \to \mathcal{M}(\mathcal{X}, \{-1, 1\})$*, we have*

$$\sup_{\mathbb{P} \in \mathcal{P}_{s,q}^{(1)}} \{\mathbb{E}_{\mathbb{P}^{\otimes n}} [\mathcal{R}_{\mathbb{P}}(\hat{c}_n) - \mathcal{R}^*]\} \geqslant \begin{cases} \kappa_1 n^{-\frac{2q+2}{r(d+q-2)}} & \text{if } 0 < q \leqslant 1, \\ \kappa_1 n^{-\frac{s}{r(s+C_2 B_1 + B_2)}} & \text{if } q > 1. \end{cases} \tag{20}$$

- *There exists a constant* $\kappa_2 > 0$ *such that for any* $n \in \mathbb{N}$*, and any classifier* $\hat{c}_n : (\mathcal{X} \times \{-1, 1\})^n \to \mathcal{M}(\mathcal{X}, \{-1, 1\})$*, we have*

$$\sup_{\mathbb{P} \in \mathcal{P}_{s,\delta}^{(2)}} \{\mathbb{E}_{\mathbb{P}^{\otimes n}} [\mathcal{R}_{\mathbb{P}}(\hat{c}_n) - \mathcal{R}^*]\} \geqslant \kappa_2 n^{-\frac{s}{r(C_2 B_1 + B_2)}}. \tag{21}$$

We make some remarks regarding Theorem 5. First, we can see that for Assumption **(A1)** in the regime $0 < q \leqslant 1$, the upper bound of $\mathcal{O}\left(n^{-\frac{qs}{2r(s+C_2 B_1 + B_2)}}\right)$ provided by Theorem 4 is quite loose, as $C_2 B_1 + B_2 \geqslant 4d^2$. However, for both the $q > 1$ regime and the case where Assumption **(A2)** holds, we recover exactly the same rates as in Theorem 4. We still need to highlight however that, formally, the best rates in Theorem 4 can not be exactly attained, but can be approached arbitrarily close, hence the minimum norm DNN classifiers can not be said to be minimax optimal in a strict sense, but we can call them *essentially* minimax optimal, in the sense that they can achieve rates arbitrarily close to the optimal one.

3.3 A case of exponential convergence rate: well-specified teacher-student learning

The takeaway message from Theorem 4 is that whenever the regression function η lies in a suitable space, such that it can be approximated by FCNNs whose size grows slowly, the margin conditions **(A1)** and **(A2)** will lead to fast rates for the excess risk. Taking this idea a step further, we look in this subsection at what happens when the regression function η is *exactly representable* by our hypothesis space of FCNNs. Our starting point is the following Lemma:

Lemma 5. *Let* $R^* > 0, L^* \in \mathbb{N}$ *be fixed, and* $\mathbf{a}^* \in \mathbb{N}^{L+1}$ *be an arbitrary FCNN architecture. For any parametrization* $\boldsymbol{\theta}^* \in \mathcal{P}_{\mathbf{a}^*, R^*}$*, there exists a distribution* $\rho_{\boldsymbol{\theta}^*}$ *on* $\mathcal{X} \times \mathcal{Y}$ *such that*

$$\mathbb{E}_{(X,Y) \sim \rho_{\boldsymbol{\theta}^*}} [Y \mid X = x] = f(x; \boldsymbol{\theta}^*), \quad \text{for } \rho_X\text{-a.e. } x \in \mathcal{X},$$

where $f(\cdot; \boldsymbol{\theta}^*) : \mathcal{X} \to [-1, 1]$ *is the function realized by* $\boldsymbol{\theta}^*$*.*

Proof. Let $X \sim \rho_X$ and $U \sim \text{Uniform}([-1, 1])$ be two independent random variables on the same probability space, and define:

$$Y := \mathbb{1}[U \leqslant f(X; \boldsymbol{\theta}^*)] - \mathbb{1}[U > f(X; \boldsymbol{\theta}^*)] = \begin{cases} 1, & \text{if } U \leqslant f(X; \boldsymbol{\theta}^*), \\ -1, & \text{if } U > f(X; \boldsymbol{\theta}^*). \end{cases}$$

Now let $\rho_{\boldsymbol{\theta}^*}$ be the joint distribution of (X, Y): we then have that for ρ_X-almost every $x \in \mathcal{X}$,

$$\begin{aligned} \mathbb{E}[Y \mid X = x] &= \mathbb{E}[\mathbb{1}[U \leqslant f(x; \boldsymbol{\theta}^*)] - \mathbb{1}[U > f(x; \boldsymbol{\theta}^*)]] \\ &= \mathbb{P}[U \leqslant f(x; \boldsymbol{\theta}^*)] - \mathbb{P}[U > f(x; \boldsymbol{\theta}^*)] \\ &= \frac{1}{2}(1 + f(x; \boldsymbol{\theta}^*)) - \frac{1}{2}(1 - f(x; \boldsymbol{\theta}^*)) \\ &= f(x; \boldsymbol{\theta}^*). \end{aligned}$$

□

From Lemma 5, it follows that any function realized by a DNN can serve as the Bayes regression function $\eta(x) = \mathbb{E}[Y \mid X = x]$ of a carefully constructed classification problem. Specifically, given a target DNN $f(\cdot; \boldsymbol{\theta}^*)$ with fixed architecture and parameters, we can construct a data distribution $\rho_{\boldsymbol{\theta}^*}$ on $\mathcal{X} \times \mathcal{Y}$ such that

the associated Bayes regression function η is exactly equal to $f(\cdot;\boldsymbol{\theta}^*)$. This observation can be thought of as a formalization of the *knowledge distillation* framework, which consists in training Neural Networks of small size to solve problems at which bigger Neural Networks are very successful with comparable performance. This approach, also known as the *teacher-student setting*, is typically implemented by training a smaller ("student") network to predict a smoothed-out version of the outputs of a larger ("teacher") network, and has shown to be very successful in practice (Hinton et al., 2015; Xu et al., 2023).

Previous work has characterized the expressivity of deep ReLU fully-connected neural networks (FCNNs). Given an architecture $\mathbf{a}$ with input dimension d, width W, and depth L, the number of linear regions it can induce ranges from $\mathcal{O}(1)$ to $\mathcal{O}\big(W^{dL}\big)$ (Montufar et al., 2014; Serra et al., 2018). This exponential gap suggests that a large network with width W and depth L can, in many cases, be represented by a much smaller network with width $W' \ll W$ and depth $L' \ll L$. Such observations help explain the practical success of knowledge distillation and lend support to the *lottery ticket hypothesis*, which posits that large networks contain small subnetworks capable of comparable generalization performance (Frankle & Carbin, 2019).

Building on this understanding, we now examine the learning rates achievable in our idealized "teacher-student" setting, where the "teacher" network is realizable by the "student" network. In this scenario, the excess risk decays at an exponential rate, as formalized in the following theorem:

Theorem 6. *Let $f(\cdot;\boldsymbol{\theta}^*) \in \mathcal{NN}(\mathbf{a}^*, W^*, L^*, R^*)$ be a "teacher" neural network, and let $\rho_{\boldsymbol{\theta}^*}$ be a distribution on $\mathcal{X}\times\mathcal{Y}$ such that the conclusion of Lemma 5 holds. Suppose that the following conditions are satisfied for $\rho_{\boldsymbol{\theta}^*}$:*

1. *Assumptions **(A3)**, **(A4)**, and **(A5)** hold.*
2. *For some $W_0 < W^*$, $L_0 < L^*$, $\mathbf{a}_0 \in \mathbb{N}^{L_0+1}$, and $R_0 > 0$, we have $f(\cdot;\boldsymbol{\theta}^*) \in \mathcal{NN}(\mathbf{a}_0, W_0, L_0, R_0)$.*
3. *The hard-margin condition **(A2)** holds with margin $\delta > 0$.*

Then, for any minimum-norm solution $\widehat{\boldsymbol{\theta}}_\lambda$ of the λ-regularized ERM problem (9) over the hypothesis space $\mathcal{NN}(\mathbf{a}_0, W_0, L_0, R_0)$, with i.i.d. data $\big((x_i, y_i)\big)_{1\leqslant i\leqslant n}$ sampled from $\rho_{\boldsymbol{\theta}^*}$, and with $0 \leqslant \lambda \leqslant \frac{\exp(-2n\beta_1)}{P(\mathbf{a}_0)R_0^p}$, the excess risk satisfies, for all $n \geqslant 1$:

$$\mathcal{R}\big(\operatorname{sign} f(\cdot;\widehat{\boldsymbol{\theta}}_\lambda)\big) - \mathcal{R}^* \leqslant \beta_2 \exp(-n\beta_1) + \frac{4}{\delta^2}\exp(-2n\beta_1),$$

where

$$\beta_1 = \frac{K^2(2^{-L_0}\delta)^{2r}}{288\operatorname{Lip}_{2R_0}(\mathcal{F}_{NN})^{2r}}, \quad \beta_2 = 1 + 4\,\mathbf{Cov}_\infty\left(\mathcal{NN}, \frac{K(2^{-L_0}\delta)^r}{24\operatorname{Lip}_{2R_0}(\mathcal{F}_{NN})^{1+r}}\right),$$

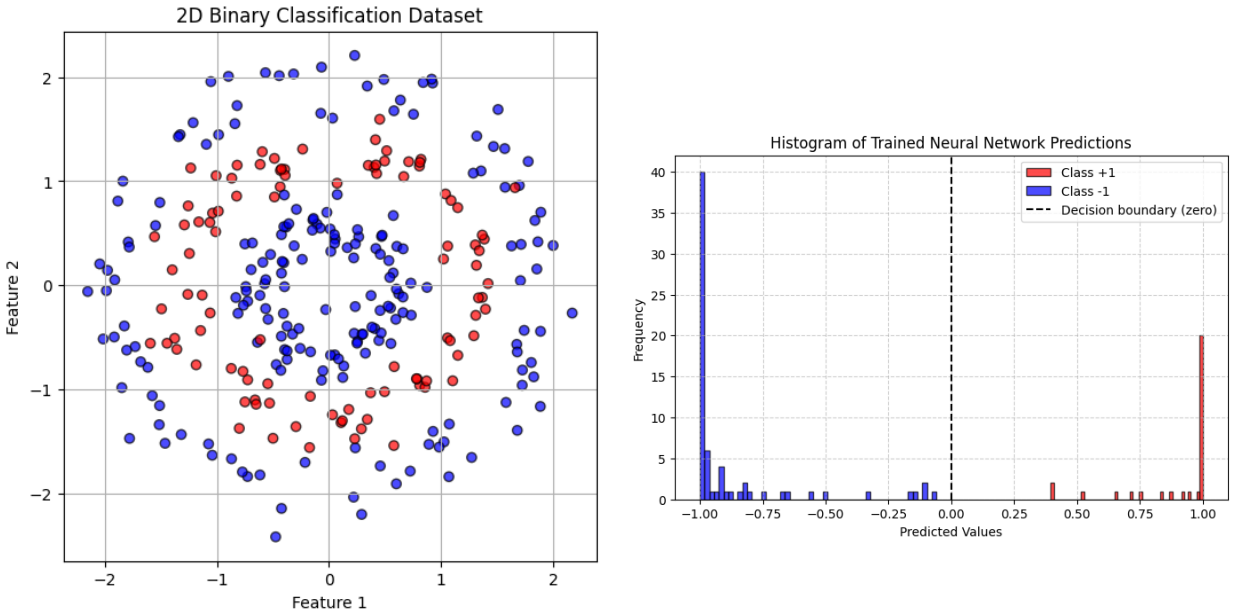
are constants which do not depend on n.

Proof. We have that the Bayes regression function η is given by $f(\cdot;\boldsymbol{\theta}^*) \in \mathcal{NN}(\mathbf{a}_0, W_0, L_0, R_0)$. By applying Theorem 2 for the hypothesis space $\mathcal{NN}(\mathbf{a}_0, W_0, L_0, R_0)$, we thus get that $\varepsilon_{\text{approx}} = 0$, and $R_0, W_0, L_0, P(\mathbf{a}_0), \operatorname{Lip}_{2R_0}(\mathcal{F}_{NN})$ are all independent of n. Applying Theorem 2 with $\nu \equiv \delta/2$ thus immediately yields the desired result. □

4 Numerical Experiments

In this section, we illustrate Theorem 4 through simple numerical experiments, considering both the weak margin and strong margin cases. To achieve this, we generate data points from two toy distributions and empirically verify that the margin assumptions **(A1)** and **(A2)** are satisfied. This verification is performed by plotting the histogram of the outputs of a DNN trained on each distribution using the square loss.
For each case, we train a DNN with the square loss for 5000 epochs, using an increasing number of training samples n, and plot the evolution of the test error as n grows. Consistent with Theorem 4, we observe a convergence rate slower than $O(n^{-1})$ under the weak margin condition **(A1)**, and faster than $O(n^{-1})$ under the hard margin condition **(A2)**.
In all the numerical experiments presented below, the DNN under consideration is a fully connected ReLU network with architecture $\mathbf{a} = (2, 64, 32, 1)$.

4.1 Convergence rate under Weak Margin condition



(a) A toy dataset where the weak margin condition holds. (b) Histogram of predicted labels for a DNN trained on this dataset.

Figure 2: A toy dataset which satisfies the weak margin condition. By plotting the histogram of a DNN trained to predict each class label, we confirm empirically that the condition is satisfied.

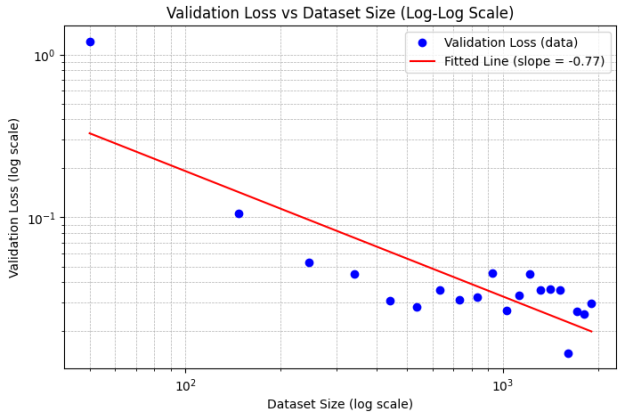
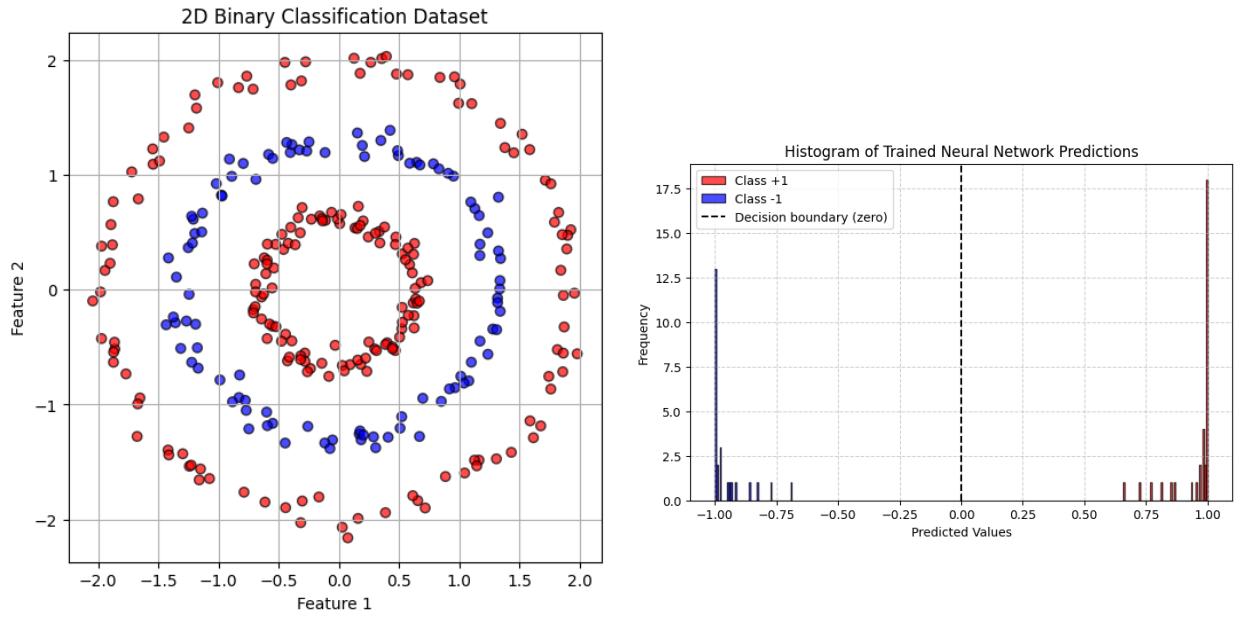


Figure 3: Evolution of the testing error for a DNN trained on n datapoints. The linear fit suggest an error decay rate of $O(n^{-0.77})$.

4.2 Convergence rate under Strong Margin condition



(a) A toy dataset where the weak margin condition holds. (b) Histogram of predicted labels for a DNN trained on this dataset.

Figure 4: A toy dataset which satisfies the hard margin condition. By plotting the histogram of a DNN trained to predict each class label, we confirm empirically that the condition is satisfied.

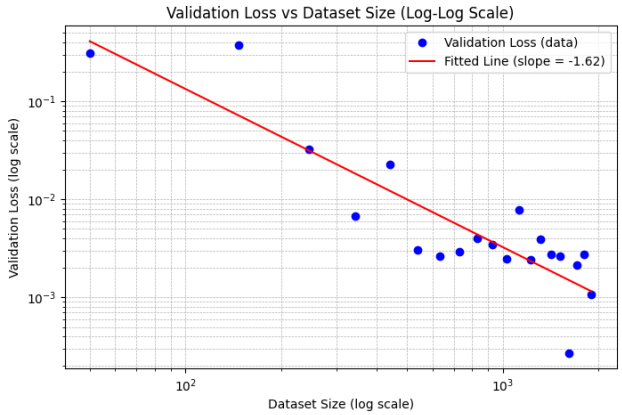


Figure 5: Evolution of the testing error for a DNN trained on n datapoints. The linear fit suggest an error decay rate of $O(n^{-1.62})$.

5 Conclusion and discussion

We have established in this work a general upper bound on the excess risk of ERM classifiers induced by parametric function classes under Mammen-Tsybakov margin conditions. As a consequence, we have deduced “super-fast" rates of convergence for ERM classifiers induced by deep ReLU networks under some suitable regularity conditions on the regression function η and the marginal data distribution ρ_X. We briefly discuss in this section some possible extensions of our results, together with related open questions.

Possible extensions:

We believe that Theorem 2 and its consequence Theorem 4 could be generalized to the following setups by a direct adaptation of our arguments:

- **General Lipschitz activation functions.** While this work focuses on ReLU activations for simplicity, our arguments rely only on two key properties of ReLU: its Lipschitz continuity and its approximation capabilities. Since many other popular activation functions achieve similar approximation rates (Ohn & Kim, 2019; Zhang et al., 2023), we expect that the results in Theorem 4 can be extended to any such Lipschitz activation function.
- **Multi-class classification.** Similarly, we believe that if we define an appropriate notion of Bayes regression function and margin conditions in the multiclass setting, such as what was done in (Vigogna et al., 2022), the results should extend naturally.

Open questions and future work:

We now highlight some questions that we believe are worth investigating further:

- **More efficient approximation.** One of the main limitations of our work is the reliance on Assumption **(A6)**, which is more restrictive than usual smoothness assumptions on η. Investigating the existence of hypothesis spaces whose approximation power and complexity scale such that standard regularity assumptions on η alone are enough to derive these "super-fast" excess risk bounds would be insightful. We believe that potential candidates could be DNNs using super-expressive activation functions (Yarotsky, 2021; Zhang et al., 2022), or other expressive and parameter-efficient architectures, such as Kolmogorov-Arnold networks (Liu et al., 2025).
- **Sparse architectures.** Another aspect not thoroughly explored in this work is the role of sparsity. For sparse architectures, both the number of parameters and the overall complexity may considerably decrease. It would thus be interesting to determine whether better excess risk bounds can be achieved for sparse hypothesis spaces. A natural candidate would be families of deep Convolutional Neural Networks, which, despite their sparsity, possess approximation capabilities comparable to FCNNs (Petersen & Voigtlaender, 2020; Zhou, 2020).
- **Other loss functions.** Our theoretical analysis crucially relies on the properties of the square loss surrogate, and so our results do not extend to popular loss functions used in practice, such as hinge loss and cross-entropy. Establishing similar results for more general losses — perhaps under different types of margin conditions — would be an interesting avenue of future research.

6 Proofs

6.1 Proof of Proposition 1

To prove Proposition 1, we will need the following growth estimate for functions satisfying the KL property, due to (van Ngai & Théra, 2009):

Theorem 7 (Adapted from Corollary 2.(ii) in (van Ngai & Théra, 2009))**.** *Let $f : \mathbb{R}^k \to [0, +\infty)$ be a lower semi-continuous function, and $f(x_0) = 0$. If there exist $c, \gamma, \varepsilon > 0$ such that $\gamma |x^*|_2 [f(x)]^{\gamma-1} \geqslant c$ for all $x \in \{x : |x - x_0|_2 < \varepsilon\} \backslash \{x : f(x) = 0\}$ and $x^* \in \hat{\partial} f(x)$, then*

$$\mathrm{dist}_2(x, \{x : f(x) = 0\}) \leqslant \frac{1}{c}[f(x)]^{\gamma}, \quad \text{for all } x \in \{x : |x - x_0|_2 < \varepsilon/2\},$$

where $\mathrm{dist}_2(a, A)$ denotes the ℓ_2 distance between a vector $a \in \mathbb{R}^k$ and a set $A \subseteq \mathbb{R}^k$.

We now proceed with the proof:

Proof of Proposition 1. Fix $R > 0$, $\lambda \geqslant 0$, and let $\boldsymbol{\theta}_\lambda \in \operatorname{argmin}_{\mathcal{P}_{\mathbf{a},R}} \mathcal{R}_{\ell,\lambda}$, which by Assumption **(A3)** is isolated. By Assumption **(A4)**, we can apply Theorem 1 to the continuous function $\psi_\lambda : \boldsymbol{\theta} \mapsto \mathcal{R}_{\ell,\lambda}(\boldsymbol{\theta}) - \mathcal{R}_{\ell,\lambda}(\boldsymbol{\theta}_\lambda)$ to deduce the existence of $c, \rho > 0$ and $0 < \kappa < 1$ such that

$$|\boldsymbol{\theta}^*|_2 \psi_\lambda(\boldsymbol{\theta})^{-\kappa} \geqslant c, \quad \text{for all } \boldsymbol{\theta} \in \mathcal{P}_{\mathbf{a},R} \backslash \{\boldsymbol{\theta}_\lambda\} \text{ s.t. } |\boldsymbol{\theta} - \boldsymbol{\theta}_\lambda|_\infty \leqslant \frac{\rho}{\operatorname{Lip}_R(\mathcal{R}_{\ell,\lambda})}. \tag{22}$$

By applying Theorem 7, we can thus deduce

$$\psi_\lambda(\boldsymbol{\theta})^{1-\kappa} \geqslant (1-\kappa) c \operatorname{dist}\left(\boldsymbol{\theta}, \operatorname*{argmin}_{\boldsymbol{\theta} \in \mathcal{P}_{\mathbf{a},R}} \psi_\lambda\right), \quad \forall \boldsymbol{\theta} \in \mathcal{P}_{\mathbf{a},R} \text{ s.t. } |\boldsymbol{\theta} - \boldsymbol{\theta}_\lambda|_\infty \leqslant \frac{\rho}{2 \operatorname{Lip}_R(\mathcal{R}_{\ell,\lambda})},$$

where $\operatorname{dist}(a, A)$ denotes the ℓ_∞ distance between a vector $a \in \mathbb{R}^k$ and a set $A \subseteq \mathbb{R}^k$.

$\boldsymbol{\theta}_\lambda$ being isolated, we have, by shrinking ρ if necessary, that $\operatorname{dist}\left(\boldsymbol{\theta}, \operatorname{argmin}_{\boldsymbol{\theta} \in \mathcal{P}_{\mathbf{a},R}} \psi_\lambda\right) = |\boldsymbol{\theta} - \boldsymbol{\theta}_\lambda|_\infty$ for all $\boldsymbol{\theta}$ in a $\rho/(2 \operatorname{Lip}_R(\mathcal{R}_{\ell,\lambda}))$ ball centered on $\boldsymbol{\theta}_\lambda$. Taking $r' := 1/(1-\kappa)$ and $K' = ((1-\kappa)c)^{1/(1-\kappa)}$, we have thus shown that for all such $\boldsymbol{\theta}$:

$$\mathcal{R}_{\ell,\lambda}(\boldsymbol{\theta}) - \mathcal{R}_{\ell,\lambda}(\boldsymbol{\theta}_\lambda) \geqslant K' t^{r'} \tag{23}$$

Notice that the constants c and κ given by Theorem 1 do not depend on R, hence K' and r' do not depend on R either. However, these constants might depend on λ, hence we let $\Lambda > 0$ be an arbitrary universal constant, and define

$$\kappa^* := \inf_{0 \leqslant \lambda \leqslant \Lambda} \{\kappa \in (0,1) : \kappa \text{ satisfies (22) for some } c > 0\},$$

$$c^* := \sup\{c > 0 : c \text{ satisfies (22) for } \kappa = \kappa^*\}.$$

Note that $(\kappa^*, c^*) \in (0,1) \times (0, \infty)$, and do not depend on λ. We thus define $r = 1/(1-\kappa^*)$, $K = ((1-\kappa^*)c^*)^{1/(1-\kappa^*)}$, and find that equation (23) holds for these values of K and r, which completes the proof. □

6.2 Proof of Lemma 2

Proof of Lemma 2. Proof of (i.): Denote by $\mathbf{0} \in \mathcal{P}_{\mathbf{a},\infty}$ the parametrization whose entries are all zeros, and $b := \widehat{\mathcal{R}}_{\ell,\lambda}(\mathbf{0})$. Clearly, $\widehat{\mathcal{R}}_{\ell,\lambda}(f(\cdot;\boldsymbol{\theta})) > b$ for all $|\boldsymbol{\theta}|_p^p > b/\lambda$, hence $\widehat{\mathcal{R}}_{\ell,\lambda}$ is minimized somewhere in $\{\boldsymbol{\theta} : |\boldsymbol{\theta}|_p^p \leqslant b/\lambda\}$ and the minimum is attained by compactness and continuity.

Now let $0 \leqslant \lambda \leqslant \lambda'$ and $\boldsymbol{\theta}, \boldsymbol{\theta}'$ respectively in $\operatorname{argmin}^* \widehat{\mathcal{R}}_{\ell,\lambda}$ and $\operatorname{argmin}^* \widehat{\mathcal{R}}_{\ell,\lambda'}$. By optimality we have

$$\begin{aligned}
\widehat{\mathcal{R}}_\ell(\boldsymbol{\theta}) + \lambda |\boldsymbol{\theta}|_p^p &\leqslant \widehat{\mathcal{R}}_\ell(\boldsymbol{\theta}') + \lambda |\boldsymbol{\theta}'|_p^p \\
&= \widehat{\mathcal{R}}_\ell(\boldsymbol{\theta}') + \lambda' |\boldsymbol{\theta}'|_p^p + (\lambda - \lambda') |\boldsymbol{\theta}'|_p^p \\
&\leqslant \widehat{\mathcal{R}}_\ell(\boldsymbol{\theta}) + \lambda' |\boldsymbol{\theta}|_p^p + (\lambda - \lambda') |\boldsymbol{\theta}'|_p^p
\end{aligned}$$

Hence we have shown

$$0 \leqslant (\lambda - \lambda') \left(|\boldsymbol{\theta}'|_p^p - |\boldsymbol{\theta}|_p^p \right) \tag{24}$$

which implies that $|\boldsymbol{\theta}'|_p \leqslant |\boldsymbol{\theta}|_p$ whenever $\lambda \leqslant \lambda'$. Finally, by basic properties of ℓ_p norms, we have

$$\begin{aligned}
\sup_{\lambda \geqslant 0, n \geqslant 1} \left\{ |\boldsymbol{\theta}|_\infty : \boldsymbol{\theta} \in \operatorname{argmin}^* \widehat{\mathcal{R}}_{\ell,\lambda} \right\} &\leqslant \sup_{\lambda \geqslant 0, n \geqslant 1} \left\{ |\boldsymbol{\theta}|_p : \boldsymbol{\theta} \in \operatorname{argmin}^* \widehat{\mathcal{R}}_{\ell,\lambda} \right\} \\
&\leqslant \sup_{n \geqslant 1} \left\{ |\boldsymbol{\theta}|_p : \boldsymbol{\theta} \in \operatorname{argmin}^* \widehat{\mathcal{R}}_{\ell,n} \right\} \\
&\leqslant \sup_{n \geqslant 1} \left\{ P(\mathbf{a})^{1/p} |\boldsymbol{\theta}|_\infty : \boldsymbol{\theta} \in \operatorname{argmin}^* \widehat{\mathcal{R}}_{\ell,n} \right\} \\
&= R^* P(\mathbf{a})^{1/p}
\end{aligned}$$

Proof of (ii.): First note that for any $f \in L^2(\rho_X)$, we have

$$\begin{aligned}
\mathcal{R}_\ell(f) &:= \mathbb{E}_{(x,y)\sim\rho}\left[(f(x)-y)^2\right] \\
&= \mathbb{E}_{(x,y)\sim\rho}\left[(f(x)-\eta(x))^2\right] + \mathbb{E}_{(x,y)\sim\rho}\left[(\eta(x)-y)^2\right] + 2\mathbb{E}_{(x,y)\sim\rho}\left[(f(x)-y)(\eta(x)-y)\right] \\
&= \|f-\eta\|^2_{L^2(\rho_X)} + C + 2\mathbb{E}_{(x,y)\sim\rho}\mathbb{E}\left[(f(x)-\eta(x))(\eta(x)-y) \mid x\right] \\
&= \|f-\eta\|^2_{L^2(\rho_X)} + C + 2\mathbb{E}_{(x,y)\sim\rho}\left[(f(x)-\eta(x))(\eta(x)-\mathbb{E}[y \mid x])\right] \\
&= \|f-\eta\|^2_{L^2(\rho_X)} + C + 0
\end{aligned}$$

Where $C \equiv \mathbb{E}_{(x,y)\sim\rho}\left[(\eta(x)-y)^2\right] \geqslant 0$ is a constant which does not depend on f. This shows that minimizing $\mathcal{R}_\ell$ is equivalent to minimizing the $L^2(\rho_X)$ distance to η. It is thus enough to show that there exists $\boldsymbol{\theta} \in \operatorname{argmin}_{\boldsymbol{\theta}\in\mathcal{P}_{\mathbf{a},\infty}} \mathcal{R}_\ell$ with $|\boldsymbol{\theta}|_\infty \leqslant R^*$ to conclude.

To that end, observe that by the strong law of large numbers, we have for any $\boldsymbol{\theta} \in \left[-R^*, R^*\right]^{P(\mathbf{a})}$ that $\widehat{\mathcal{R}}_{\ell,n}(\boldsymbol{\theta}) \to \mathcal{R}_\ell(\boldsymbol{\theta})$ almost surely as $n \to \infty$. Furthermore, because the realization mapping $\mathcal{F}$ is Lipschitz, its composition with the mapping $\ell : (f,x,y) \mapsto (f(x)-y)^2$ is uniformly Lipschitz over $(x,y) \in \mathcal{X} \times \{-1,+1\}$, and we can denote by $L_R > 0$ its Lipschitz constant. Finally, for any convergent subsequence $(\boldsymbol{\theta}_n)_{n\geqslant 1} \subseteq \left[-R^*, R^*\right]^{P(\mathbf{a})}$ satisfying $\boldsymbol{\theta}_n \in \operatorname{argmin}_{\boldsymbol{\theta}\in\mathcal{P}_{\mathbf{a},\infty}} \widehat{\mathcal{R}}_{\ell,n}$ for all $n \geqslant 1$ (which is guaranteed to exist by Assumption **(A5)**) and with limit $\boldsymbol{\theta}^*$, we have

$$|\widehat{\mathcal{R}}_{\ell,n}(\boldsymbol{\theta}_n) - \mathcal{R}_\ell(\boldsymbol{\theta}^*)| \leqslant L_R|\boldsymbol{\theta}_n - \boldsymbol{\theta}^*|_\infty + |\mathcal{R}_\ell(\boldsymbol{\theta}_n) - \mathcal{R}_\ell(\boldsymbol{\theta}^*)|.$$

Thus for any $\boldsymbol{\theta} \in \mathcal{P}_{\mathbf{a},\infty}$, we can take the limit $n \to \infty$ to find

$$\widehat{\mathcal{R}}_{\ell,n}(\boldsymbol{\theta}_n) \leqslant \widehat{\mathcal{R}}_{\ell,n}(\boldsymbol{\theta}) \implies \mathcal{R}_\ell(\boldsymbol{\theta}^*) \leqslant \mathcal{R}_\ell(\boldsymbol{\theta}),$$

which implies that $\boldsymbol{\theta}^* \in \operatorname{argmin}_{\boldsymbol{\theta}\in\mathcal{P}_{\mathbf{a},\infty}} \mathcal{R}_\ell$, as desired. □

6.3 Upper bounds

6.3.1 Preliminary Results

We start by collecting a number of useful lemmas which will be needed to prove the main results. Throughout the following, recall the definition of the *misclassification risk* $\mathcal{R}(\operatorname{sign} f)$ (1) for a real-valued function f :

$$\mathcal{R}(\operatorname{sign} f) := \mathbb{P}_{(X,Y)\sim\rho}(\operatorname{sign} f(X) \neq Y)$$

Our first lemma is a bound on the difference of the misclassification risks of classifiers induced by measurable functions $f, g \in L^\infty(\mathcal{X})$:

Lemma 6. *For any two $f, g \in L^\infty(\mathcal{X})$, we have*

$$|\mathcal{R}(\operatorname{sign} f) - \mathcal{R}(\operatorname{sign} g)| \leqslant \mathbb{P}_{x\sim\rho_X}\left(\|f-g\|_{L^\infty(\mathcal{X})} \geqslant |f(x)|\right)$$

Proof of Lemma 6. We have

$$\begin{aligned}
|\mathcal{R}(\operatorname{sign} f) - \mathcal{R}(\operatorname{sign} g)| &= \left|\mathbb{E}\left[\mathbb{1}\left\{\operatorname{sign} f(X) \neq Y\right\} - \mathbb{1}\left\{\operatorname{sign} g(X) \neq Y\right\}\right]\right| \\
&\leqslant \mathbb{E}\left[\left|\mathbb{1}\left\{\operatorname{sign} f(X) \neq Y\right\} - \mathbb{1}\left\{\operatorname{sign} g(X) \neq Y\right\}\right|\right] \\
&\leqslant \mathbb{E}\left[\mathbb{1}\left\{\operatorname{sign} f(X) \neq \operatorname{sign} g(X)\right\}\right] = \mathbb{P}\left(\operatorname{sign} f(X) \neq \operatorname{sign} g(X)\right)
\end{aligned}$$

But now observe that for any $x \in \mathcal{X}$, $\operatorname{sign} f(x) \neq \operatorname{sign} g(x) \implies |f(x)-g(x)| \geqslant |f(x)|$. Hence the inclusion of events

$$\left\{\operatorname{sign} f(X) \neq \operatorname{sign} g(X)\right\} \subseteq \left\{\|f-g\|_{L^\infty(\rho_X)} \geqslant |f(X)|\right\},$$

which implies the claimed inequality. □

We next have an upper bound on the excess misclassification risk of a classifier $\operatorname{sign} f$ in terms of the $L^2(\rho_X)$ distance between f and the regression function η.

Lemma 7. *For any $f \in L^2(\rho_X)$, we have the inequality*

$$\mathcal{R}(\operatorname{sign} f) - \mathcal{R}(\operatorname{sign} \eta) \leqslant \|f - \eta\|_{L^2(\rho_X)}$$

Proof. Note that we have

$$\eta(X) = \mathbb{E}[Y \mid X] = \mathbb{P}(Y = 1 \mid X) - \mathbb{P}(Y = -1 \mid X),$$

hence by the law of total expectation :

$$\begin{aligned}
\mathcal{R}(\operatorname{sign} f) - \mathcal{R}(\operatorname{sign} \eta) &= \mathbb{E}_X \left[\mathbb{E}_Y[\mathbb{1}\{\operatorname{sign} f(X) \neq Y\} - \mathbb{1}\{\operatorname{sign} \eta(X) \neq Y\} \mid X]\right] \\
&= \mathbb{E}_X \left[(\mathbb{1}\{\operatorname{sign} f(X) \neq 1\} - \mathbb{1}\{\operatorname{sign} \eta(X) \neq 1\}) \cdot \mathbb{P}(Y = 1 \mid X)\right. \\
&\quad \left. + (\mathbb{1}\{\operatorname{sign} f(X) \neq -1\} - \mathbb{1}\{\operatorname{sign} \eta(X) \neq -1\}) \cdot \mathbb{P}(Y = -1 \mid X)\right] \\
&\leqslant \mathbb{E}_X[|\eta(X)| \mathbb{1}\{\operatorname{sign} f(X) \neq \operatorname{sign} \eta(X)\}] \\
&\leqslant \mathbb{E}_X[|\eta(X) - f(X)| \mathbb{1}\{\operatorname{sign} f(X) \neq \operatorname{sign} \eta(X)\}] \\
&\leqslant \|f - \eta\|_{L^2(\rho_X)}
\end{aligned}$$

□

The following result states that, whenever η satisfies either the low-noise Assumption **(A1)** or the hard margin condition **(A2)**, any sufficiently good $L^2(\rho_X)$ approximation of η will satisfy the same assumption with high probability.

Lemma 8. *Let $f \in L^2(\rho_X)$ be such that $\|f - \eta\|_{L^2(\rho_X)} \leqslant \varepsilon$ for some $\varepsilon > 0$. The following is true :*

- *If η satisfies the low-noise Assumption **(A1)**, we have for all $\delta > \varepsilon$ and $0 < \nu < \delta$:*

$$\mathbb{P}(|f(X)| \leqslant \nu) \leqslant \frac{\varepsilon^2}{(\delta - \nu)^2} + C\delta^q$$

- *If η satisfies the hard-margin Assumption **(A2)** with margin $\delta > 0$ and $\varepsilon < \delta$, we have for all $\nu < \delta$:*

$$\mathbb{P}(|f(X)| \leqslant \nu) \leqslant \frac{\varepsilon^2}{(\delta - \nu)^2}$$

Proof.

- Assume that Assumption **(A1)** holds. Observe that for any $\delta > 0$

$$\begin{aligned}
\mathbb{P}(|f(X)| \leqslant \nu) &= \mathbb{P}(|f(X)| \leqslant \nu; |\eta(X)| > \delta) + \mathbb{P}(|f(X)| \leqslant \nu; |\eta(X)| \leqslant \delta) \\
&\leqslant \mathbb{P}(|f(X)| \leqslant \nu; |\eta(X)| > \delta) + C\delta^q
\end{aligned}$$

Now note that on the event $|\eta(X)| > \delta$, we have by triangle inequality

$$|f(X) - \eta(X)| + |f(X)| \geqslant |\eta(X)| > \delta \implies |f(X) - \eta(X)| \geqslant \delta - |f(X)|$$

Finally, Chebyshev's inequality yields

$$\begin{aligned}
\mathbb{P}(|f(X)| \leqslant \nu; |\eta(X)| > \delta) &\leqslant \mathbb{P}(|f(X) - \eta(X)| \geqslant \delta - \nu) \\
&\leqslant \frac{\|f - \eta\|^2_{L^2(\rho_X)}}{(\delta - \nu)^2} \\
&\leqslant \frac{\varepsilon^2}{(\delta - \nu)^2},
\end{aligned}$$

this yields the claimed inequality.

- If we now assume that η satisfies the hard-margin condition **(A2)**, we proceed similarly as in the previous case, with the only difference being that the term $\mathbb{P}(|f(X)| \leqslant \nu; |\eta(X)| \leqslant \delta)$ is now equal to zero. The rest of the argument carries through.

□

The following lemma quantifies the approximation error of minimizers $\boldsymbol{\theta}_\lambda$ of the regularized population risk $\mathcal{R}_{\ell,\lambda}$ over $\mathcal{H}_{\mathcal{F},\mathbf{a},R}$ in terms of the approximation error of the parametric function class $\mathcal{H}_{\mathcal{F},\mathbf{a},R}$.

Lemma 9. *Let $\mathbf{a} \in \mathbb{N}^{L+1}$ be an architecture, and $R > 0$ a parameter bound such that*

$$\inf_{f \in \mathcal{H}_{\mathcal{F},\mathbf{a},R}} \|f - \eta\|_{L^2(\rho_X)} \leqslant \varepsilon$$

for some constant $\varepsilon \geqslant 0$. Then, for any $\lambda \geqslant 0$, we have that any minimizer $\boldsymbol{\theta}_\lambda$ of the regularized population risk $\mathcal{R}_{\ell,\lambda}$ over $\mathcal{H}_{\mathcal{F},\mathbf{a},R}$ satisfies

$$\|f(\cdot;\boldsymbol{\theta}_\lambda) - \eta\|_{L^2(\rho_X)} \leqslant \varepsilon + \sqrt{\lambda P(\mathbf{a}) R^p},$$

where $P(\mathbf{a})$ denotes the number of parameters in the architecture $\mathbf{a}$.

Proof. First note that for any $g \in L^2(\rho_X)$, we have

$$\begin{aligned}
\mathcal{R}_\ell(g) &:= \mathbb{E}_{(x,y)\sim\rho}\left[(g(x)-y)^2\right] \\
&= \mathbb{E}_{(x,y)\sim\rho}\left[(g(x)-\eta(x))^2\right] + \mathbb{E}_{(x,y)\sim\rho}\left[(\eta(x)-y)^2\right] + 2\mathbb{E}_{(x,y)\sim\rho}\left[(g(x)-y)(\eta(x)-y)\right] \\
&= \|g-\eta\|^2_{L^2(\rho_X)} + C + 2\mathbb{E}_{(x,y)\sim\rho}\mathbb{E}\left[(g(x)-\eta(x))(\eta(x)-y) \mid x\right] \\
&= \|g-\eta\|^2_{L^2(\rho_X)} + C + 2\mathbb{E}_{(x,y)\sim\rho}\left[(g(x)-\eta(x))(\eta(x)-\mathbb{E}[y \mid x])\right] \\
&= \|g-\eta\|^2_{L^2(\rho_X)} + C + 0
\end{aligned}$$

Where $C \equiv \mathbb{E}_{(x,y)\sim\rho}\left[(\eta(x)-y)^2\right] \geqslant 0$ is a constant which does not depend on g. This shows that minimizing $\mathcal{R}_\ell$ is equivalent to minimizing the $L^2(\rho_X)$ distance to η, and in particular for two square-integrable functions $f, g \in L^2(\rho_X)$, we have the identity

$$\mathcal{R}_\ell(f) - \mathcal{R}_\ell(g) = \|f-\eta\|^2_{L^2(\rho_X)} - \|g-\eta\|^2_{L^2(\rho_X)}. \tag{25}$$

Now denote by $\boldsymbol{\theta}^*$ any minimizer of $\|f(\cdot;\boldsymbol{\theta}) - \eta\|^2_{L^2(\rho_X)}$ over $\mathcal{P}_{\mathbf{a},R}$. For any positive λ, we have

$$\begin{aligned}
\mathcal{R}_\ell(f(\cdot;\boldsymbol{\theta}_\lambda)) &= \mathcal{R}_{\ell,\lambda}(f(\cdot;\boldsymbol{\theta}_\lambda)) - \lambda|\boldsymbol{\theta}_\lambda|_p^p \\
&\leqslant \mathcal{R}_{\ell,\lambda}(f(\cdot;\boldsymbol{\theta}_\lambda)) \\
&\leqslant \mathcal{R}_{\ell,\lambda}(f(\cdot;\boldsymbol{\theta}^*)) \\
&= \mathcal{R}_\ell(f(\cdot;\boldsymbol{\theta}^*)) + \lambda|\boldsymbol{\theta}^*|_p^p \\
&\leqslant \mathcal{R}_\ell(f(\cdot;\boldsymbol{\theta}^*)) + \lambda P(\mathbf{a}) R^p
\end{aligned}$$

Where $P(\mathbf{a})$ is the number of parameters in the architecture $\mathbf{a}$. From the identity (25) above, we deduce that $\|f(\cdot;\boldsymbol{\theta}_\lambda) - \eta\|^2_{L^2(\rho_X)}$ differs from $\|f(\cdot;\boldsymbol{\theta}^*) - \eta\|^2_{L^2(\rho_X)}$ by at most $\lambda P(\mathbf{a}) R^p$. We thus find that

$$\begin{aligned}
\|f(\cdot;\boldsymbol{\theta}_\lambda) - \eta\|^2_{L^2(\rho_X)} &\leqslant \|f(\cdot;\boldsymbol{\theta}^*) - \eta\|^2_{L^2(\rho_X)} + \lambda P(\mathbf{a}) R^p \\
&\leqslant \|f(\cdot;\boldsymbol{\theta}^*) - \eta\|^2_{L^\infty(\rho_X)} + \lambda P(\mathbf{a}) R^p \\
&\leqslant \varepsilon^2 + \lambda P(\mathbf{a}) R^p,
\end{aligned}$$

and we conclude the proof by using the subadditivity of $x \mapsto \sqrt{x}$. □

The last result we will need is a large deviation type estimate on the probability that a minimizer $\widehat{\boldsymbol{\theta}}_\lambda$ of the empirical risk $\widehat{\mathcal{R}}_{\ell,\lambda}$ is far away from the argmin of $\mathcal{R}_{\ell,\lambda}$. Such estimate can be readily obtained by applying covering number based concentration bounds, which are a standard tool in Learning Theory literature (Györfi et al., 2002).

Lemma 10. *For any $\lambda \geqslant 0$, let $\widehat{\boldsymbol{\theta}}_\lambda \in \mathcal{P}_{\mathbf{a},R}$ be a minimum-norm solution of the λ-ERM problem (9), and denote by $\mathcal{R}_{\ell,\lambda}$ the regularized population risk (10). If Assumptions **(A3)** and **(A4)** hold, then for all $t > 0$, we have the estimate*

$$\mathbb{P}(\operatorname{dist}(\widehat{\boldsymbol{\theta}}_\lambda, \operatorname{argmin} \mathcal{R}_{\ell,\lambda}) \geqslant t) \leqslant 4\, \mathbf{Cov}_\infty\left(\mathcal{H}, \frac{Kt^r}{24\operatorname{Lip}_R(\mathcal{F})}\right) \exp\left(\frac{-nK^2t^{2r}}{288}\right)$$

Proof. Observe the inclusion of events

$$\begin{aligned}
\operatorname{dist}(\widehat{\boldsymbol{\theta}}_\lambda, \operatorname{argmin} \mathcal{R}_{\ell,\lambda}) \geqslant t \implies & \mathcal{R}_{\ell,\lambda}(f(\cdot,\widehat{\boldsymbol{\theta}}_\lambda)) \geqslant \inf_{\boldsymbol{\theta}\in\mathcal{P}_{\mathbf{a},R}:\operatorname{dist}(\boldsymbol{\theta},\operatorname{argmin}\mathcal{R}_{\ell,\lambda})\geqslant t} \mathcal{R}_{\ell,\lambda}(f(\cdot,\boldsymbol{\theta})) \\
\implies & \mathcal{R}_{\ell,\lambda}(f(\cdot,\widehat{\boldsymbol{\theta}}_\lambda)) - \mathcal{R}_{\ell,\lambda}(f(\cdot,\boldsymbol{\theta}_\lambda)) \geqslant Kt^r \\
\implies & \mathcal{R}_{\ell,\lambda}(f(\cdot,\widehat{\boldsymbol{\theta}}_\lambda)) - \widehat{\mathcal{R}}_{\ell,\lambda}(f(\cdot,\widehat{\boldsymbol{\theta}}_\lambda)) \\
& + \widehat{\mathcal{R}}_{\ell,\lambda}(f(\cdot,\boldsymbol{\theta}_\lambda)) - \mathcal{R}_{\ell,\lambda}(f(\cdot,\boldsymbol{\theta}_\lambda)) \geqslant Kt^r \\
\implies & \widehat{\mathcal{R}}_{\ell,\lambda}(f(\cdot,\boldsymbol{\theta}_\lambda)) - \mathcal{R}_{\ell,\lambda}(f(\cdot,\boldsymbol{\theta}_\lambda)) \geqslant Kt^r/2 \\
& \text{OR} \quad \mathcal{R}_{\ell,\lambda}(f(\cdot,\widehat{\boldsymbol{\theta}}_\lambda)) - \widehat{\mathcal{R}}_{\ell,\lambda}(f(\cdot,\widehat{\boldsymbol{\theta}}_\lambda)) \geqslant Kt^r/2
\end{aligned}$$

Where we used Proposition 1 in the second line. Now set $\varepsilon := Kt^r/2$ and let

$$\{f(\cdot;\boldsymbol{\theta}_\varepsilon) : \boldsymbol{\theta}_\varepsilon \in \boldsymbol{\Theta}_\varepsilon\}$$

be a minimal size $\varepsilon/(12\operatorname{Lip}(\mathcal{F}))$-cover of $\mathcal{H}_{\mathcal{F},\mathbf{a},R}$. By observing that the map

$$\varphi : \mathcal{P}_{\mathbf{a},R} \to \mathbb{R}, \quad \boldsymbol{\theta} \mapsto (f(x;\boldsymbol{\theta}) - y)^2$$

is $4\operatorname{Lip}_R(\mathcal{F})$-Lipschitz continuous uniformly over $(x,y) \in \mathcal{X} \times \{-1,1\}$, we get that for any $\boldsymbol{\theta} \in \mathcal{P}_{\mathbf{a},R}$, and $\boldsymbol{\theta}_\varepsilon \in \boldsymbol{\Theta}_\varepsilon$ such that $|\boldsymbol{\theta} - \boldsymbol{\theta}_\varepsilon|_\infty \leqslant \varepsilon/(12\operatorname{Lip}_R(\mathcal{F}))$:

$$\begin{aligned}
|\widehat{\mathcal{R}}_{\ell,\lambda}(f(\cdot,\boldsymbol{\theta})) - \mathcal{R}_{\ell,\lambda}(f(\cdot,\boldsymbol{\theta}))| &= \left|\frac{1}{n}\sum_{i=1}^n (f(x_i;\boldsymbol{\theta}) - y_i)^2 - \mathbb{E}\left[(f(x;\boldsymbol{\theta}) - y)^2\right]\right| \\
&\leqslant \left|\frac{1}{n}\sum_{i=1}^n (f(x_i;\boldsymbol{\theta}_\varepsilon) - y_i)^2 - \frac{1}{n}\sum_{i=1}^n (f(x_i;\boldsymbol{\theta}) - y_i)^2\right| \\
&+ \left|\mathbb{E}\left[(f(x;\boldsymbol{\theta}_\varepsilon) - y)^2\right] - \mathbb{E}\left[(f(x;\boldsymbol{\theta}) - y)^2\right]\right| \\
&+ \left|\frac{1}{n}\sum_{i=1}^n (f(x_i;\boldsymbol{\theta}_\varepsilon) - y_i)^2 - \mathbb{E}\left[(f(x;\boldsymbol{\theta}_\varepsilon) - y)^2\right]\right| \\
&\leqslant 4\operatorname{Lip}_R(\mathcal{F})|\boldsymbol{\theta} - \boldsymbol{\theta}_\varepsilon|_\infty + \left|\frac{1}{n}\sum_{i=1}^n (f(x_i;\boldsymbol{\theta}_\varepsilon) - y_i)^2 - \mathbb{E}\left[(f(x;\boldsymbol{\theta}_\varepsilon) - y)^2\right]\right| \\
&\leqslant \frac{2\varepsilon}{3} + \left|\frac{1}{n}\sum_{i=1}^n (f(x_i;\boldsymbol{\theta}_\varepsilon) - y_i)^2 - \mathbb{E}\left[(f(x;\boldsymbol{\theta}_\varepsilon) - y)^2\right]\right|
\end{aligned}$$

After taking the supremum over $\boldsymbol{\theta} \in \mathcal{P}_{\mathbf{a},R}$ in the above inequality, and observing that the $Z_i := (f(x_i;\boldsymbol{\theta}_\varepsilon) - y_i)^2$ are i.i.d. and taking value in $[0,4]$ almost surely, we apply the union bound together with Hoeffding's

inequality to find:

$$\begin{aligned}
\mathbb{P}\left(\operatorname{dist}(\widehat{\boldsymbol{\theta}}_\lambda, \operatorname{argmin}\mathcal{R}_{\ell,\lambda}) \geqslant t\right) &\leqslant \mathbb{P}\left(\widehat{\mathcal{R}}_{\ell,\lambda}(f(\cdot,\boldsymbol{\theta}_\lambda)) - \mathcal{R}_{\ell,\lambda}(f(\cdot,\boldsymbol{\theta}_\lambda)) \geqslant Kt^r/2\right) \\
&\quad + \mathbb{P}\left(\mathcal{R}_{\ell,\lambda}(f(\cdot,\widehat{\boldsymbol{\theta}}_\lambda)) - \widehat{\mathcal{R}}_{\ell,\lambda}(f(\cdot,\widehat{\boldsymbol{\theta}}_\lambda)) \geqslant Kt^r/2\right) \\
&\leqslant 2\mathbb{P}\left(\sup_{\boldsymbol{\theta}\in\mathcal{P}_{\mathbf{a},R}} |\mathcal{R}_{\ell,\lambda}(f(\cdot,\boldsymbol{\theta})) - \widehat{\mathcal{R}}_{\ell,\lambda}(f(\cdot,\boldsymbol{\theta}))| \geqslant Kt^r/2\right) \\
&= 2\mathbb{P}\left(\sup_{\boldsymbol{\theta}\in\mathcal{P}_{\mathbf{a},R}} |\mathcal{R}_{\ell,\lambda}(f(\cdot,\boldsymbol{\theta})) - \widehat{\mathcal{R}}_{\ell,\lambda}(f(\cdot,\boldsymbol{\theta}))| \geqslant \varepsilon\right) \\
&\leqslant 2\mathbb{P}\left(\sup_{\boldsymbol{\theta}_\varepsilon\in\boldsymbol{\Theta}_\varepsilon} |\mathcal{R}_{\ell,\lambda}(f(\cdot,\boldsymbol{\theta}_\varepsilon)) - \widehat{\mathcal{R}}_{\ell,\lambda}(f(\cdot,\boldsymbol{\theta}_\varepsilon))| \geqslant \varepsilon/3\right) \\
&\leqslant 4\,\mathbf{Cov}_\infty\left(\mathcal{H}, \frac{\varepsilon}{12\operatorname{Lip}_R(\mathcal{F})}\right)\exp\left(\frac{-n\varepsilon^2}{72}\right).
\end{aligned}$$

Finally, after substituting $\varepsilon \equiv Kt^r/2$, we find

$$\mathbb{P}\left(\operatorname{dist}(\widehat{\boldsymbol{\theta}}_\lambda, \operatorname{argmin}\mathcal{R}_{\ell,\lambda}) \geqslant t\right) \leqslant 4\,\mathbf{Cov}_\infty\left(\mathcal{H}, \frac{Kt^r}{24\operatorname{Lip}_R(\mathcal{F})}\right)\exp\left(\frac{-nK^2t^{2r}}{288}\right),$$

as desired. □

6.3.2 Proof of Theorem 2

We prove Theorem 2 under the low-noise Assumption **(A1)** only, as the case **(A2)** can be shown using the exact same argument.

To begin, we decompose the excess risk in two parts :

$$\begin{aligned}
\mathcal{R}(\operatorname{sign} f(\cdot;\widehat{\boldsymbol{\theta}}_\lambda)) - \mathcal{R}^* &:= \mathcal{R}(\operatorname{sign} f(\cdot;\widehat{\boldsymbol{\theta}}_\lambda)) - \mathcal{R}(\operatorname{sign}\eta) \\
&= \mathcal{R}(\operatorname{sign} f(\cdot;\widehat{\boldsymbol{\theta}}_\lambda)) - \mathcal{R}(\operatorname{sign} f(\cdot;\boldsymbol{\theta}_\lambda)) \\
&\quad + \mathcal{R}(\operatorname{sign} f(\cdot;\boldsymbol{\theta}_\lambda)) - \mathcal{R}(\operatorname{sign}\eta)
\end{aligned}$$

Where $\widehat{\boldsymbol{\theta}}_\lambda \in \mathcal{P}_{\mathbf{a},R}$ and $\boldsymbol{\theta}_\lambda \in \mathcal{P}_{\mathbf{a},2R}$ are respectively minimum-norm minimizers of the empirical and population risk (9), such that

$$|\widehat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}_\lambda|_\infty = \operatorname{dist}(\widehat{\boldsymbol{\theta}}_\lambda, \operatorname*{argmin}_{\mathcal{P}_{\mathbf{a},R}}\mathcal{R}_{\ell,\lambda}).$$

Note that by Assumption **(A5)** and closedness of $\operatorname{argmin}_{\mathcal{P}_{\mathbf{a},R}}\mathcal{R}_{\ell,\lambda}$, the above is always possible as long as the parameter bound R has been chosen larger than $R_0\cdot P(\mathbf{a})^{1/p}$, but the ℓ_∞ norm of $\boldsymbol{\theta}_\lambda$ can only be bounded by $2R$ instead of R.

Combining Lemma 7 and Lemma 9, we immediately get the bound on the first summand :

$$\mathcal{R}(\operatorname{sign} f(\cdot;\boldsymbol{\theta}_\lambda)) - \mathcal{R}(\operatorname{sign}\eta) \leqslant \|f(\cdot;\boldsymbol{\theta}_\lambda) - \eta\|_{L^2(\rho_X)} \leqslant \varepsilon_{\text{approx}} + \sqrt{\lambda P(\mathbf{a})2^pR^p}. \tag{26}$$

It only remains to bound the second summand. To that end, we apply Lemma 6, which yields :

$$\mathcal{R}(\operatorname{sign} f(\cdot;\widehat{\boldsymbol{\theta}}_\lambda)) - \mathcal{R}(\operatorname{sign} f(\cdot;\boldsymbol{\theta}_\lambda)) \leqslant \mathbb{P}\left\{\|f(\cdot;\widehat{\boldsymbol{\theta}}_\lambda) - f(\cdot;\boldsymbol{\theta}_\lambda)\|_{L^\infty(\rho_X)} \geqslant |f(X;\boldsymbol{\theta}_\lambda)|\right\}.$$

Now note that thanks to inequality (26), we can apply the "high-probability margin" property from Lemma 8 to get for all $\delta > \varepsilon_{\text{approx}}$, $\lambda < (\delta - \varepsilon_{\text{approx}})^2 (P(\mathbf{a}) 2^p R^p)^{-1}$, and $0 < \nu < \delta$:

$$\begin{aligned}
\mathbb{P}\left\{\|f(\cdot;\widehat{\boldsymbol{\theta}}_\lambda) - f(\cdot;\boldsymbol{\theta}_\lambda)\|_{L^\infty(\rho_X)} \geqslant |f(X;\boldsymbol{\theta}_\lambda)|\right\} &= \mathbb{P}\left\{\|f(\cdot;\widehat{\boldsymbol{\theta}}_\lambda) - f(\cdot;\boldsymbol{\theta}_\lambda)\|_{L^\infty(\rho_X)} \geqslant |f(X;\boldsymbol{\theta}_\lambda)|; |f(X;\boldsymbol{\theta}_\lambda)| > \nu\right\} \\
&\quad + \mathbb{P}\left\{\|f(\cdot;\widehat{\boldsymbol{\theta}}_\lambda) - f(\cdot;\boldsymbol{\theta}_\lambda)\|_{L^\infty(\rho_X)} \geqslant |f(X;\boldsymbol{\theta}_\lambda)|; |f(X;\boldsymbol{\theta}_\lambda)| \leqslant \nu\right\} \\
&\leqslant \mathbb{P}\left\{\|f(\cdot;\widehat{\boldsymbol{\theta}}_\lambda) - f(\cdot;\boldsymbol{\theta}_\lambda)\|_{L^\infty(\rho_X)} \geqslant \nu\right\} \\
&\quad + (\delta - \nu)^{-2}\left(\varepsilon_{\text{approx}} + \sqrt{\lambda P(\mathbf{a}) 2^p R^p}\right)^2 + C\delta^q
\end{aligned}$$

We are now left with estimating the probability that $\|f(\cdot;\widehat{\boldsymbol{\theta}}_\lambda) - f(\cdot;\boldsymbol{\theta}_\lambda)\|_{L^\infty} \geqslant \nu$. By Lipschitzness of $\mathcal{F}$, we have

$$\begin{aligned}
\mathbb{P}\left\{\|f(\cdot;\widehat{\boldsymbol{\theta}}_\lambda) - f(\cdot;\boldsymbol{\theta}_\lambda)\|_{L^\infty(\rho_X)} \geqslant \nu\right\} &\leqslant \mathbb{P}\left(|\widehat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}_\lambda|_\infty \geqslant \nu / \operatorname{Lip}_{2R}(\mathcal{F})\right) \\
&= \mathbb{P}\left(\operatorname{dist}(\widehat{\boldsymbol{\theta}}_\lambda, \operatorname{argmin} \mathcal{R}_{\ell,\lambda}) \geqslant \nu / \operatorname{Lip}_{2R}(\mathcal{F})\right) \\
&\leqslant 4\,\mathbf{Cov}_\infty\left(\mathcal{H}, \frac{K\nu^r}{24\operatorname{Lip}_{2R}(\mathcal{F})^{1+r}}\right) \exp\left(\frac{-nK^2\nu^{2r}}{288\operatorname{Lip}_{2R}(\mathcal{F})^{2r}}\right),
\end{aligned}$$

where the exponential inequality follows from Lemma 10. Combining all of these inequalities, we have thus shown that for all $\delta > \varepsilon_{\text{approx}}$, $\lambda < (\delta - \varepsilon_{\text{approx}})^2 (P(\mathbf{a}) 2^p R^p)^{-1}$, and $0 < \nu < \delta$:

$$\begin{aligned}
\mathcal{R}(\operatorname{sign} f(\cdot;\widehat{\boldsymbol{\theta}}_\lambda)) - \mathcal{R}^* &\leqslant \varepsilon_{\text{approx}} + \sqrt{\lambda P(\mathbf{a}) 2^p R^p} + C\delta^q \\
&\quad + (\delta - \nu)^{-2}\left(\varepsilon_{\text{approx}} + \sqrt{\lambda P(\mathbf{a}) 2^p R^p}\right)^2 \\
&\quad + 4\,\mathbf{Cov}_\infty\left(\mathcal{NN}, \frac{K\nu^r}{24\operatorname{Lip}_{2R}(\mathcal{F})^{1+r}}\right) \exp\left(\frac{-nK^2\nu^{2r}}{288\operatorname{Lip}_{2R}(\mathcal{F})^{2r}}\right)
\end{aligned}$$

which concludes the proof of Theorem 2 under Assumption **(A1)**. As was mentioned in the beginning, the proof under **(A2)** can be done with the exact same argument, the only difference being that the $C\delta^q$ term will disappear when applying Lemma 8.

6.3.3 Proof of Theorem 4

Start by fixing $\alpha > 0$, and recall the approximation error bound given by Assumption **(A6)**, according to which

$$\inf_{f \in \mathcal{NN}(\mathbf{a},W,L,R)} \|f - \eta\|_{L^2(\rho_X)} \leqslant C_3 W_0^{-2s/d},$$

for some architecture $\mathbf{a}$ such that $W(\mathbf{a}) = C_1 W_0 \log_2(W_0)$, $L(\mathbf{a}) = C_2 L_0 \log_2(L_0) + 2d$, where $W_0 \in \mathbb{N}_{\geqslant 2}$ is arbitrary, $C_1 = d(3s)^d d$, $C_3 = C_\eta s^d 8^s$, and L_0 and $C_2 \equiv C_2(s)$ are fixed.

By letting $W_0 = n^{\alpha d/2s} \times C_3^{d/2s}$, we deduce that there is a Neural Network architecture $\mathbf{a}_n$ with respective depth and width

$$L_n = C_2 L_0 \log_2(L_0) + 2d, \quad W_n = C_1 W_0 \log_2(W_0) = \tilde{\mathcal{O}}\left(n^{\alpha d/2s}\right),$$

where $\tilde{\mathcal{O}}$ hides logarithmic factors, such that

$$\inf_{f \in \mathcal{NN}(\mathbf{a}_n,W_n,L_n,R)} \|f - \eta\|_{L^2(\rho_X)} \leqslant n^{-\alpha}$$

Furthermore, the number of parameters in $\mathbf{a}_n$ is bounded as

$$P(\mathbf{a}_n) = \sum_{l=1}^{L_n} \mathbf{a}_n^{(l)} \mathbf{a}_n^{(l-1)} + \mathbf{a}_n^{(l)} \leqslant L_n(W_n^2 + W_n) = \tilde{\mathcal{O}}\left(n^{\alpha d/s}\right).$$

Similarly, recall the Lipschitz constant bound given by Lemma 3:

$$\sup_{\substack{\boldsymbol{\theta},\boldsymbol{\theta}'\in\mathcal{P}_{\mathbf{a}_n,R}\\ \boldsymbol{\theta}\neq\boldsymbol{\theta}'}} \frac{\|\mathcal{F}_\sigma(\boldsymbol{\theta})-\mathcal{F}_\sigma(\boldsymbol{\theta}')\|_{\mathcal{C}(\mathcal{X})}}{|\boldsymbol{\theta}-\boldsymbol{\theta}'|_\infty} \leqslant 2L_n^2 R^{L_n-1} W_n^{L_n},$$

and note that with $R \equiv R_n \equiv R_0 P(\mathbf{a_n})^{1/p}$, we have

$$R_n = \tilde{\mathcal{O}}\left(n^{\alpha d/ps}\right)$$

Putting these together we get

$$\begin{aligned}
\sup_{\substack{\boldsymbol{\theta},\boldsymbol{\theta}'\in\mathcal{P}_{\mathbf{a}_n,R_n}\\ \boldsymbol{\theta}\neq\boldsymbol{\theta}'}} \frac{\|\mathcal{F}_\sigma(\boldsymbol{\theta})-\mathcal{F}_\sigma(\boldsymbol{\theta}')\|_{\mathcal{C}(\mathcal{X})}}{|\boldsymbol{\theta}-\boldsymbol{\theta}'|_\infty} &\leqslant 2L_n^2\tilde{\mathcal{O}}\left(\left(n^{\alpha d/ps}\right)^{L_n-1}\left(n^{\alpha d/2s}\right)^{L_n}\right)\\
&\leqslant \tilde{\mathcal{O}}\left(n^{\frac{\alpha d}{s}\cdot\left[\frac{L_n-1}{p}+\frac{L_n}{2}\right]}\right)\\
&= \tilde{\mathcal{O}}\left(n^{\frac{\alpha d}{ps}\cdot[L_n(1+p/2)-1]}\right)\\
&= \tilde{\mathcal{O}}\left(n^{\frac{\alpha d}{ps}\cdot\left[\left(C_2(s)L_0\log_2 L_0+2d\right)(1+p/2)-1\right]}\right),
\end{aligned}$$

where all the logarithmic factors and terms which do not depend on n are hidden in the $\tilde{\mathcal{O}}$.

After noting that $\mathrm{Lip}_{2R}(\mathcal{F}_{NN}) \leqslant 2^{L-1}\,\mathrm{Lip}_R(\mathcal{F}_{NN})$, we are left with bounding the quantity

$$\mathbf{Cov}_\infty\left(\mathcal{NN}, \frac{K(2^{1-L_n}\nu)^r}{24\,\mathrm{Lip}_{R_n}(\mathcal{F}_{NN})^{1+r}}\right),$$

which by Lemma 4, we know is bounded by

$$\left(1+\frac{48R_n\,\mathrm{Lip}_{R_n}(\mathcal{F}_{NN})^{2+r}}{K(2^{1-L_n}\nu)^r}\right)^{P(\mathbf{a}_n)} \leqslant \left(\frac{49R_n\,\mathrm{Lip}_{R_n}(\mathcal{F}_{NN})^{2+r}}{K(2^{1-L_n}\nu)^r}\right)^{P(\mathbf{a}_n)}.$$

Using the bounds on R_n and $\mathrm{Lip}_{R_n}(\mathcal{F}_{NN})$ above, we thus find that

$$\frac{49R_n\,\mathrm{Lip}_{R_n}(\mathcal{F}_{NN})^{2+r}}{K\cdot 2^{r(1-L_n)}} \leqslant \tilde{\mathcal{O}}\left(n^{\frac{\alpha d}{ps}}\cdot n^{\frac{(2+r)\alpha d}{ps}\cdot\left[\left(C_2L_0\log_2 L_0+2d\right)(1+p/2)-1\right]}\right).$$

The above quantity being polynomial in n, we thus find that the covering number grows as the exponential of $P(\mathbf{a}_n)$, up to a multiplicative logarithmic factor:

$$\log\left[\mathbf{Cov}_\infty\left(\mathcal{NN}, \frac{K(2^{1-L}\nu)^r}{24\,\mathrm{Lip}_{R_n}(\mathcal{F}_{NN})^{1+r}}\right)\right] = \mathcal{O}\Big(P(\mathbf{a}_n)\log(n^\beta\cdot\nu^{-r})\Big),$$

where

$$\beta \equiv \frac{\alpha d}{ps}\Big(1+(2+r)\cdot\left[\big(C_2(s)L_0\log_2 L_0+2d\big)(1+p/2)-1\right]\Big)$$

To conclude the proof for the case **(A1)**, we let $\varepsilon_{\text{approx}} \equiv n^{-\frac{\alpha}{r}}$, $\delta \equiv 2n^{-\frac{\alpha}{2r}}$ and $\nu \equiv n^{-\frac{\alpha}{2r}}$: observe that by picking λ such that

$$0 \leqslant \lambda \leqslant \varepsilon_{\text{approx}}^2(P(\mathbf{a}_n)2^pR_n^p)^{-1} = \mathcal{O}\left(n^{-\frac{2\alpha(s+d)}{rs}}\right),$$

we have $\lambda < (\delta-\varepsilon_{\text{approx}})^2(P(\mathbf{a}_n)2^pR_n^p)^{-1}$ and

$$\varepsilon_{\text{approx}} + \sqrt{\lambda P(\mathbf{a}_n)2^pR^p} \leqslant 2\varepsilon_{\text{approx}}.$$

We are thus allowed to apply Theorem 2 with these values of λ, which yields the excess risk bound:

$$\mathcal{R}\left(\operatorname{sign} f(\cdot;\hat{\boldsymbol{\theta}}_\lambda)\right) - \mathcal{R}_* \leqslant 2n^{-\frac{\alpha}{r}} + 2Cn^{-\frac{\alpha q}{2r}} + 4n^{-\frac{\alpha}{r}} + 4\exp\left(-A_1 n^{1-A_2} + n^{\frac{\alpha d}{s}}\log(\gamma n^{(\alpha+2\beta)/2})\right),$$

where

$$A_1 \equiv \frac{K^2 2^{2r(1-L_n)}}{288}, \quad A_2 \equiv \alpha\left(1 + \frac{rd}{ps}\cdot\left(\left[C_2(s)L_0\log_2(L_0) + 2d\right]\cdot(2+p) - 2\right)\right),$$

and $\gamma > 0$ is a quantity which does not depend on n. Hence we see that the exponential term converges to zero as $n \to \infty$ if $1 - A_2 > 0$ and $1 - A_2 > \alpha d/s$, or equivalently if $1 - A_2 > \alpha d/s$, which is equivalent to the following inequality for α:

$$\alpha < \left(1 + \frac{rd}{ps}\cdot\left(\left[C_2(s)L_0\log_2(L_0) + 2d\right]\cdot(2+p) - 2\right)\right)^{-1}.$$

This concludes the proof under Assumption **(A1)**. The proof under Assumption **(A2)** with margin $\delta > 0$ is very similar: we now pick $\varepsilon_{\text{approx}} \equiv n^{-\frac{\alpha}{r}}$, $\nu \equiv \delta/2$, and

$$0 \leqslant \lambda \leqslant \varepsilon_{\text{approx}}^2 (P(\mathbf{a}_n)2^p R_n^p)^{-1} = \mathcal{O}\left(n^{-\frac{2\alpha(s+d)}{rs}}\right),$$

such that Theorem 2 can be applied, to yield for all $n \geqslant \left\lceil (\delta/2)^{-r/\alpha}\right\rceil$:

$$\mathcal{R}\left(\operatorname{sign} f(\cdot;\hat{\boldsymbol{\theta}}_\lambda)\right) - \mathcal{R}_* \leqslant 2n^{-\frac{\alpha}{r}} + 16n^{-\frac{\alpha}{r}} + 4\exp\left(-A_1 n^{1-A_2} + n^{\alpha d/s}\log(\gamma n^{\beta}(\delta/2)^{-r})\right),$$

where

$$A_1 \equiv \frac{K^2(\delta 2^{-L_n})^{2r}}{288}, \quad A_2 \equiv \alpha\frac{rd}{sp}\left(\left[C_2(s)L_0\log_2 L_0 + 2d\right]\cdot(2+p) - 2\right).$$

Hence, we see as before that in this case the term $4\exp\left(-A_1 n^{1-A_2} + n^{\alpha d/s}\log(\gamma n^{\beta}(\delta/2)^{-r})\right)$ vanishes exponentially fast as $n \to \infty$ if $1 - A_2 > \alpha d/s$, which equivalently means that α needs to satisfy the following inequality

$$\alpha < \frac{sp}{rd\left(\left[C_2(s)L_0\log_2 L_0 + 2d\right]\cdot(2+p) - 2\right)}.$$

6.4 Lower bounds

6.4.1 Preliminary Results

We begin by giving a sufficient condition on the marginal distribution ρ_X for Assumption **(A6)** to hold. We first formally define *discrete distributions* on the cube $\mathcal{X} = [0,1]^d$.

Definition 8 (Discrete Distribution)**.** *For any $x \in \mathcal{X}$, let δ_x denote the Dirac probability measure at x, which satisfies $\delta_x(A) = \mathbb{1}_A(x)$ for any measurable $A \subseteq \mathcal{X}$. We say that the marginal data distribution ρ_X is discrete if there exist a sequence $x_1, x_2, \ldots \subseteq \mathcal{X}$ of pairwise distinct points, and a sequence $\lambda_1, \lambda_2, \ldots$ of non-negative real numbers, such that $\sum_{i\geqslant 1}\lambda_i = 1$ and*

$$\rho_X := \sum_{i\geqslant 1}\lambda_i \delta_{x_i}.$$

Lemma 11. *Assume that $\rho_X = \sum_{i\geqslant 1}\lambda_i\delta_{x_i}$ is a discrete distribution, and fix $s > 0$. If the coefficients $(\lambda_i)_{i\geqslant 1}$ are such that*

$$\sum_{i\geqslant n+1}\lambda_i \leqslant C_\rho n^{-4s/d} \quad \text{for all } n \geqslant 1,$$

*where C_ρ is a positive universal constant, then Assumption **(A6)** is satisfied with smoothness parameter s and constant $C_2(s) = 0$.*

Proof of Lemma 11. We will prove the lemma by constructing a sequence $(\Psi_n)_{n\geqslant 1}$ of DNN with 2 hidden layers and width $O(n)$ which, for every $n \in \mathbb{N}$, interpolate the Bayes regression function η at every $x \in \{x_1, \dots, x_n\}$.

To that end, fix $n \in \mathbb{N}$: since $x_1, \dots, x_n$ are pairwise distinct, the set $\mathbb{R}^d \setminus \cup_{i\neq j} (x_i - x_j)^\perp$ is not empty, and we can find a direction $v \in \mathbb{R}^d$ such that $v\cdot x_1, \dots, v\cdot x_n$ are pairwise distinct. Now let $b = 2\max_{1\leqslant i\leqslant n} |v\cdot x_i|$, and note that the map $NN_n : x \mapsto \mathrm{ReLU}(v\cdot x + b)$ maps each x_i to $v\cdot x_i + b$.

Given the n real numbers $v\cdot x_1 + b < \ldots < v\cdot x_n + b$ (reindexing as necessary), one can construct a continuous piecewise linear map $PL_n : \mathbb{R} \to \mathbb{R}$ with breakpoints $(v\cdot x_1 + b, \eta(x_1)), \dots, (v\cdot x_n + b, \eta(x_n))$. Such a map PL_n can easily be realized by a shallow ReLU network of width n (see e.g. (Arora et al., 2018) for an explicit construction). Each $\Psi_n = PL_n \circ NN_n$ is thus realized by a DNN with 2 hidden layers and width equal to $\max(d, n)$. Furthermore we have $|\Psi_n(x)| \leqslant \|\eta(x)\|_{L^\infty(\mathcal{X})}$ for all $x \in [0,1]^d$, and by construction, we have

$$\|\Psi_n - \eta\|^2_{L^2(\rho_X)} = \sum_{i\geqslant n+1} \lambda_i \big(\Psi_n(x_i) - \eta(x_i)\big)^2 \leqslant 4\|\eta\|^2_{L^\infty(\mathcal{X})} \sum_{i\geqslant n+1} \lambda_i.$$

Lastly, note that for any depth $L \geqslant 2$, the identity mapping $Id : x \in \mathbb{R}^d \mapsto x$ can be realized by the following parameter vector

$$\boldsymbol{\theta} = \left(\left(\begin{pmatrix} I_d \\ -I_d \end{pmatrix}, 0 \right), (I_{2d}, 0), \dots, (I_{2d}, 0), ((I_d - I_d), 0) \right),$$

where I_d is the $d \times d$ identity matrix, and the tuple $(I_{2d}, 0)$ is repeated $L-2$ times. Note that it is easy to modify the above architecture so that it yields a representation of the identity mapping with same depth and arbitrary width $W \geqslant 2d$. Hence, by "padding" the width and depth of Ψ_n as necessary, we see that Assumption **(A6)** is indeed satisfied with smoothness parameter s and constant $C_2(s) = 0$ whenever $\sum_{i\geqslant n+1} \lambda_i \leqslant C_\rho n^{-4s/d}$, as claimed. □

The main tool we will need to obtain our minimax lower bounds is Assouad's lemma. Before stating the lemma, we define the *probability hypercube*, following notation from (Audibert, 2004).

Definition 9 (Probability Hypercube)**.** *Let $m \in \mathbb{N}, w \in (0,1], b \in (0,1]$, and $b' \in (0,1]$. A (m, w, b, b')-hypercube of probability distributions is a family*

$$\{\mathbb{P}_{\vec{\sigma}} \mid \vec{\sigma} = (\sigma_1, \dots, \sigma_m) \in \{-1,1\}^m\}$$

of 2^m probability distributions on $\mathcal{X} \times \{-1,1\}$ having the same first marginal:

$$\mathbb{P}_{\vec{\sigma}}(dX) = \mathbb{P}_{(+1,\dots,+1)}(dX) =: \mu \quad \text{for all } \vec{\sigma} \in \{-1,1\}^m,$$

and such that there exists a partition $\mathcal{X}_0, \dots, \mathcal{X}_m$ of the unit cube $\mathcal{X}$ satisfying

- *for any $j \in \{1, \dots, m\}$, we have $\mu(\mathcal{X}_j) = w$*
- *for any $j \in \{0, \dots, m\}$, and any $x \in \mathcal{X}_j$, we have*

$$\mathbb{E}_{\vec{\sigma}}[Y \mid X = x] = \sigma_j\, \xi(x),$$

where $\sigma_0 := 1$ and $\xi : \mathcal{X} \to [0,1]$ is such that for any $j \in \{1, \dots, m\}$,

$$\begin{cases} wb = \sqrt{w^2 - \Big(\mathbb{E}_{X\sim\mu}\big[\sqrt{1-\xi^2(X)}\mathbb{1}_{\mathcal{X}_j}\big]\Big)^2}, \\ wb' = \mathbb{E}_{X\sim\mu}\big[\xi(X)\mathbb{1}_{\mathcal{X}_j}\big]. \end{cases}$$

We are now ready to state Assouad's lemma, in a version adapted to the setting in which the label set is given by $\mathcal{Y} = \{-1, 1\}$:

Lemma 12 (Adapted from Lemma 5.1 in (Audibert, 2004)). *If a set $\mathcal{P}$ of probability distributions contains a (m, w, b, b')-hypercube, then for any measurable estimator*

$$\hat{c} : (\mathcal{X} \times \{-1, 1\})^N \to \mathcal{M}(\mathcal{X}, \{-1, 1\}),$$

we have

$$\sup_{\mathbb{P} \in \mathcal{P}} \left\{ \mathbb{E}_{\mathbb{P}^{\otimes N}} \left[\mathcal{R}_{\mathbb{P}}(\hat{c}) \right] - \mathcal{R}^*_{\mathbb{P}} \right\} \geqslant \frac{1 - b\sqrt{Nw}}{2} \, m \, w \, b',$$

where $\mathcal{R}_{\mathbb{P}}(f) = \mathbb{P}\{f(X) \neq Y\}$ and $\mathcal{R}^*_{\mathbb{P}}$ is the Bayes optimal risk under $\mathbb{P}$.

Although the original result in (Audibert, 2004) only proves an analogous version of Lemma 12 for the case $\mathcal{Y} = \{0, 1\}$, one can readily check that Lemma 12 is a direct consequence of applying the transformation $t \mapsto (t + 1)/2$ to the labels.

6.4.2 Proof of Theorem 5

We now prove the minimax lower bound. Note that our argument mostly follows the approach taken in (Audibert & Tsybakov, 2007), with suitable adjustments made as needed. We first prove the result for the low noise condition **(A1)**. The lower bound for the hard margin condition **(A2)** follows from a straightforward modification of the argument, which we explain at the end.

For an integer $h \geqslant 1$, we partition the unit cube $\mathcal{X} = [0, 1]^d$ as follows: define the regular grid G_h as

$$G_h := \left\{ \left(\frac{2k_1 + 1}{2h}, \ldots, \frac{2k_d + 1}{2h} \right) : k_i \in \{0, \ldots, h - 1\}, i = 1, \ldots, d \right\},$$

and for $x \in \mathcal{X}$ let $n_h(x) \in G_h$ be the minimum-norm element of G_h closest to x in Euclidean norm, so that $n_h : \mathcal{X} \to G_h$ is a well-defined (single-valued) function. We then define $\mathcal{X}'_1, \ldots \mathcal{X}'_{h^d}$ as the canonical partition of $\mathcal{X}$ induced by n_h. That is, $x, y \in \mathcal{X}$ belong to the same subset $\mathcal{X}_i$ if and only if $n_h(x) = n_h(y)$. Now fix an integer $m \geqslant 1$, and for $1 \leqslant i \leqslant m$, define $\mathcal{X}_i := \mathcal{X}'_i$, and let $\mathcal{X}_0 := [0, 1]^d \setminus \cup_{1 \leqslant i \leqslant m} \mathcal{X}_i$, so that $\mathcal{X}_0, \ldots, \mathcal{X}_m$ is a partition of $\mathcal{X}$ as well.

Let $u : \mathbb{R}^+ \to \mathbb{R}^+$ be the continuous piecewise linear function defined by $u(x) = 1$ for $x \in [0, 1/4]$, $u(x) = -4x + 2$ for $x \in [1/4, 1/2]$, and $u(x) = 0$ for $x \in [1/2, \infty)$, and define $\phi : \mathcal{X} \to \mathbb{R}^+$ by $\phi(x) = u(|x|_2)$, where $|\cdot|_2$ denotes the ℓ_2 (Euclidean) norm.

We now define the hypercube $\mathcal{H} = \{\mathbb{P}_{\vec{\sigma}} \mid \vec{\sigma} \in \{-1, 1\}^m\}$ of probability distributions on $\mathcal{X} \times \{-1, 1\}$. For all $\vec{\sigma}$, we set the marginal distribution $\mathbb{P}_{\vec{\sigma}}(dX) =: \mu$ on $\mathcal{X}$ as a discrete distribution (as per Definition 8) defined in the following way: denote $z_1, \ldots, z_{h^d}$ the centers of the grid G_h, and for $1 \leqslant i \leqslant m$, fix $\widetilde{\mathcal{X}}^{(i)} := (x_j^{(i)})_{j \geqslant 1}$ as a dense countable subset of $B(z_i, 1/4h)$, the Euclidean ball with center z_i and radius $1/4h$. Likewise, let $\widetilde{\mathcal{X}}^{(0)} := (x_j^{(0)})_{j \geqslant 1}$ be a dense countable subset of $\mathcal{X}_0$. Finally, let $0 < w \leqslant m^{-1}$, and define the marginal probability μ for all $x \in \mathcal{X}$ by

$$\mu(\{x\}) := \begin{cases} C_s w 2^{-js} & \text{if } x = x_j^{(i)} \text{ for some } (i, j) \in \{1, \ldots, m\} \times \mathbb{N}, \\ C_s (1 - mw) 2^{-js} & \text{if } x = x_j^{(0)} \text{ for some } j \in \mathbb{N}, \\ 0 & \text{otherwise,} \end{cases}$$

where $C_s := \left(\sum_{j \geqslant 1} 2^{-js} \right)^{-1}$ is a normalization constant. Observe that μ does not depend on $\vec{\sigma}$.

We now define for each $\vec{\sigma} \in \{-1, 1\}^m$ the regression function $\eta_{\vec{\sigma}} : x \mapsto \mathbb{E}_{(X,Y) \sim \mathbb{P}_{\vec{\sigma}}}[Y \mid X = x]$, which will fully determine $\mathbb{P}_{\vec{\sigma}}$. For all $x \in \mathcal{X}_j$, $1 \leqslant j \leqslant m$, we set $\eta_{\vec{\sigma}}(x) := \sigma_j \varphi(x)$, where $\varphi(x) := h^{-1} \phi(h[x - n_h(x)])$, and for $x \in \mathcal{X}_0$, we let $\eta_{\vec{\sigma}}(x) = 1$. Note that by Lemma 11, we have that Assumption **(A6)** is satisfied.

We now check the margin assumption **(A1)**: fix $z_1 = (1/2h, \dots, 1/2h) \in \mathcal{X}_1$. For any $\vec{\sigma} \in \{-1,1\}^m$ we have

$$\begin{aligned}
\mathbb{P}_{\vec{\sigma}}(|\eta_{\vec{\sigma}}(X)| \leqslant t) &= m\mathbb{P}_{\vec{\sigma}}(\phi(h\lfloor X - z_1 \rfloor) \leqslant th) \\
&= m \sum_{x \in B(z_1, 1/4h)} \mathbb{1}_{\{\phi(h\lfloor X - z_1 \rfloor) \leqslant th\}}(x)\mu(\{x\}) \\
&= C_s m w \sum_{j \geqslant 1} 2^{-js} \mathbb{1}_{\{\phi(h\lfloor X - z_1 \rfloor) \leqslant th\}}(x_j^{(1)}) \\
&= C_s m w \sum_{j \geqslant 1} 2^{-js} \mathbb{1}_{\{\phi^{-1}([0,th])\}}(x_j^{(1)}/h + z_1) \\
&= mw \mathbb{1}_{\{th \geqslant 1\}}.
\end{aligned}$$

We thus see that the low noise condition **(A1)** holds as long as $mw \leqslant Ch^{-q}$.

We are now ready to prove the lower bound. By applying Lemma 12, we have for any classifier $\hat{c}_n$

$$\sup_{\mathbb{P} \in \mathcal{H}} \left\{ \mathbb{E}_{\mathbb{P}^{\otimes n}} \left[\mathcal{R}_{\mathbb{P}}(\hat{c}_n) \right] - \mathcal{R}^*_{\mathbb{P}} \right\} \geqslant \frac{1 - b\sqrt{nw}}{2} \, m \, w \, b', \tag{27}$$

where $b = b' = h^{-1}$. Denote by $\alpha^* := s/(s + C_2 B_1 + B_2)$ the optimal rate, and remember that $q > 0$ denotes the noise exponent in Assumption **(A1)**. By taking

$$h = \begin{cases} \lceil n^{2/(d+q-2)} \rceil & \text{if } q \leqslant 1, \\ \lceil n^{\alpha^*/(q-1)} \rceil & \text{if } q > 1, \end{cases} \quad m = h^d, \quad w = h^{-d-q},$$

and replacing in (27), we obtain the claimed lower bounds in equation (20), which proves the first case of Theorem 5.

We now address the case where we require the probability hypercube $\mathcal{H}$ to satisfy the stronger Assumption **(A2)**. In that case, we can simply modify the previous construction as follows: define the grid G_h, partition $(\mathcal{X}_i)_{1 \leqslant i \leqslant m}$, functions u, ϕ, and marginal probability distribution μ in the exact same way as before, but now let $\varphi(x) = \delta\phi(h\lfloor x - n_h(x) \rfloor)$ for all $x \in \mathcal{X}$. It is straightforward to check that with these choices, all distributions in this probability hypercube satisfy Assumptions **(A2)** and **(A6)**. Furthermore, $\mathcal{H}$ as defined is a (m, w, b, b')-hypercube with $b = b' = \delta$. We then set

$$\begin{cases} h = \lceil m^{1/d} \rceil, \quad m = \left\lceil n^{1-\alpha^*} \right\rceil, \quad w = \frac{1}{\delta^2 n} & \text{if } 0 < \alpha^* \leqslant 1, \\ h = 1, \qquad m = 1, \qquad w = \frac{y_n^2}{n} & \text{if } \alpha^* > 1, \end{cases}$$

where $\alpha^* = s/(C_2 B_1 + B_2)$ is the optimal rate, and $y_n \in (0, 1/\delta)$ is such that $\delta^2 y_n^3 - \delta y_n^2 + 2n^{1-\alpha^*} = 0$. By plugging these values into (27), we find that the lower bound (21) holds. The proof is thus complete.

References

Raman Arora, Amitabh Basu, Poorya Mianjy, and Anirbit Mukherjee. Understanding deep neural networks with rectified linear units. In *International Conference on Learning Representations*, 2018.

Jean-Yves Audibert. Classification under polynomial entropy and margin assumptions and randomized estimators. Preprint, Laboratoire de Probabilités et Modèles Aléatoires, Univ. Paris VI and VII, 2004. URL https://imagine.enpc.fr/publications/papers/04PMA-908.pdf.

Jean-Yves Audibert and Alexandre B Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of statistics*, 35(2):608–633, 2007.

Julius Berner, Philipp Grohs, and Arnulf Jentzen. Analysis of the generalization error: Empirical risk minimization over deep artificial neural networks overcomes the curse of dimensionality in the numerical approximation of black–scholes partial differential equations. *SIAM Journal on Mathematics of Data Science*, 2(3):631–657, 2020.

Alberto Bietti and Francis Bach. Deep equals shallow for relu networks in kernel regimes, 2021.

J. Bochnak, M. Coste, and M.-F. Roy. *Real Algebraic Geometry*, volume 36 of *Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge*. Springer, Berlin, 1998. ISBN 978-3-540-64663-1.

Jérôme Bolte, Aris Daniilidis, Adrian Lewis, and Masahiro Shiota. Clarke subgradients of stratifiable functions. *SIAM Journal on Optimization*, 18(2):556–572, 2007.

Thijs Bos and Johannes Schmidt-Hieber. Convergence rates of deep relu networks for multiclass classification. *Electronic Journal of Statistics*, 16(1):2724–2773, 2022.

Vivien Cabannes and Stefano Vigogna. A case of exponential convergence rates for svm. *arXiv preprint arXiv:2205.10055*, 2022.

Emmanuel J Candes, Michael B Wakin, and Stephen P Boyd. Enhancing sparsity by reweighted ℓ_1 minimization. *Journal of Fourier analysis and applications*, 14:877–905, 2008.

Aurore Delaigle and Peter Hall. Achieving near perfect classification for functional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):267–286, 2012.

Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

Mahyar Fazlyab, Alexander Robey, Hamed Hassani, Manfred Morari, and George Pappas. Efficient and accurate estimation of lipschitz constants for deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Han Feng, Shuo Huang, and Ding-Xuan Zhou. Generalization analysis of cnns for classification on spheres. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *ICLR*. OpenReview.net, 2019. URL http://dblp.uni-trier.de/db/conf/iclr/iclr2019.html#FrankleC19.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

László Györfi, Michael Köhler, Adam Krzyżak, and Harro Walk. *A distribution-free theory of nonparametric regression*, volume 1. Springer, 2002.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015. URL http://dblp.uni-trier.de/db/journals/corr/corr1503.html#HintonVD15.

Tianyang Hu, Jun Wang, Wenjia Wang, and Zhenguo Li. Understanding square loss in training overparametrized neural network classifiers. *arXiv preprint arXiv:2112.03657*, 2021.

Tianyang Hu, Ruiqi Liu, Zuofeng Shang, and Guang Cheng. Minimax optimal deep neural network classifiers under smooth decision boundary. *arXiv preprint arXiv:2207.01602*, 2022.

Like Hui and Mikhail Belkin. Evaluation of neural architectures trained with square loss vs cross-entropy in classification tasks. *arXiv preprint arXiv:2006.07322*, 2020.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Yongdai Kim, Ilsang Ohn, and Dongha Kim. Fast convergence rates of deep neural networks for classification. *arXiv preprint arXiv:1812.03599*, 2018.

Hyunouk Ko, Namjoon Suh, and Xiaoming Huo. On excess risk convergence rates of neural network classifiers. *arXiv preprint arXiv:2309.15075*, 2023.

Vladimir Koltchinskii and Olexandra Beznosova. Exponential convergence rates in classification. In *Learning Theory: 18th Annual Conference on Learning Theory, COLT 2005, Bertinoro, Italy, June 27-30, 2005. Proceedings 18*, pp. 295–307. Springer, 2005.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Anders Krogh and John Hertz. A simple weight decay can improve generalization. *Advances in neural information processing systems*, 4, 1991.

Krzysztof Kurdyka. On gradients of functions definable in o-minimal structures. *Annales de l'Institut Fourier*, 48(3):769–783, 1998. doi: 10.5802/aif.1638. URL https://www.numdam.org/item/10.5802/aif.1638.pdf.

Guoyin Li and Ting Kei Pong. Calculus of the exponent of kurdyka–łojasiewicz inequality and its applications to linear convergence of first-order methods. *Foundations of computational mathematics*, 18(5):1199–1232, 2018.

Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljacic, Thomas Y. Hou, and Max Tegmark. KAN: Kolmogorov–arnold networks. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Ozo7qJ5vZi.

Stanislaw Lojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. *Les équations aux dérivées partielles*, 117(87-89), 1963.

Christos Louizos, Max Welling, and Diederik P Kingma. Learning sparse neural networks through l_0 regularization. *arXiv preprint arXiv:1712.01312*, 2017.

Jianfeng Lu, Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network approximation for smooth functions. *SIAM Journal on Mathematical Analysis*, 53(5):5465–5506, 2021.

Enno Mammen and Alexandre B Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27 (6):1808–1829, 1999.

Tong Mao and Ding-Xuan Zhou. Approximation of functions from korobov spaces by deep convolutional neural networks. *Advances in Computational Mathematics*, 48(6):84, 2022.

Joseph T Meyer. Optimal convergence rates of deep neural networks in a classification setting. *arXiv preprint arXiv:2207.12180*, 2022.

Guido F Montufar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. On the number of linear regions of deep neural networks. *Advances in neural information processing systems*, 27, 2014.

Atsushi Nitanda and Taiji Suzuki. Optimal rates for averaged stochastic gradient descent under neural tangent kernel regime. *arXiv preprint arXiv:2006.12297*, 2020.

Ilsang Ohn and Yongdai Kim. Smooth function approximation by deep neural networks with general activation functions. *Entropy*, 21(7):627, 2019.

Philipp Petersen and Felix Voigtlaender. Optimal approximation of piecewise smooth functions using deep relu neural networks. *Neural Networks*, 108:296–330, 2018.

Philipp Petersen and Felix Voigtlaender. Equivalence of approximation by convolutional neural networks and fully-connected networks. *Proceedings of the American Mathematical Society*, 148(4):1567–1581, 2020.

Philipp Petersen and Felix Voigtlaender. Optimal learning of high-dimensional classification problems using deep neural networks. *arXiv preprint arXiv:2112.12555*, 2021.

R Tyrrell Rockafellar and Roger JB Wets. *Variational analysis*. Springer, 1998.

Itay Safran and Ohad Shamir. Depth-width tradeoffs in approximating natural functions with neural networks. In *International conference on machine learning*, pp. 2979–2987. PMLR, 2017.

Bodhisattva Sen. A gentle introduction to empirical process theory and applications. *Lecture Notes, Columbia University*, 11:28–29, 2018.

Thiago Serra, Christian Tjandraatmadja, and Srikumar Ramalingam. Bounding and counting linear regions of deep neural networks. In *International conference on machine learning*, pp. 4558–4566. PMLR, 2018.

Masahiro Shiota. *Geometry of Subanalytic and Semialgebraic Sets*, volume 150 of *Progress in Mathematics*. Birkhäuser, Boston, 1997. ISBN 978-0-8176-3913-4. doi: 10.1007/978-1-4612-1970-5.

Steve Smale and Ding-Xuan Zhou. Learning theory estimates via integral operators and their approximations. *Constructive approximation*, 26(2):153–172, 2007.

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

Ingo Steinwart and Clint Scovel. Fast rates for support vector machines. In *Learning Theory: 18th Annual Conference on Learning Theory, COLT 2005, Bertinoro, Italy, June 27-30, 2005. Proceedings 18*, pp. 279–294. Springer, 2005.

Taiji Suzuki. Adaptivity of deep relu network for learning in besov and mixed smooth besov spaces: optimal rate and curse of dimensionality. *arXiv preprint arXiv:1810.08033*, 2018.

Alexander B Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Huynh van Ngai and Michel Théra. Error bounds for systems of lower semicontinuous functions in asplund spaces. *Mathematical Programming*, 116(1-2):397–427, 2009.

Stefano Vigogna, Giacomo Meanti, Ernesto De Vito, and Lorenzo Rosasco. Multiclass learning with margin: exponential rates with no bias-variance trade-off. *arXiv preprint arXiv:2202.01773*, 2022.

Aladin Virmaux and Kevin Scaman. Lipschitz regularity of deep neural networks: analysis and efficient estimation. *Advances in Neural Information Processing Systems*, 31, 2018.

Felix Voigtlaender and Philipp Petersen. Approximation in l p (μ) with deep relu neural networks. In *2019 13th International conference on Sampling Theory and Applications (SampTA)*, pp. 1–4. IEEE, 2019.

Tomoya Wakayama and Masaaki Imaizumi. Fast convergence on perfect classification for functional data. *arXiv preprint arXiv:2104.02978*, 2021.

Chuanyun Xu, Wenjian Gao, Tian Li, Nanlan Bai, Gang Li, and Yang Zhang. Teacher-student collaborative knowledge distillation for image classification. *Applied Intelligence*, 53(2):1997–2009, 2023.

Jintao Xu, Chenglong Bao, and Wenxun Xing. Convergence rates of training deep neural networks via alternating minimization methods. *Optimization Letters*, 18(4):909–923, 2024.

Dmitry Yarotsky. Error bounds for approximations with deep relu networks. *Neural Networks*, 94:103–114, 2017.

Dmitry Yarotsky. Elementary superexpressive activations. In *International conference on machine learning*, pp. 11932–11940. PMLR, 2021.

J. Zeng, T. T. K. Lau, S. Lin, and Y. Yao. Global convergence of block coordinate descent in deep learning. In *International Conference on Machine Learning*, pp. 7313–7323. PMLR, May 2019.

Shijun Zhang, Zuowei Shen, and Haizhao Yang. Deep network approximation: Achieving arbitrary accuracy with fixed number of neurons. *Journal of Machine Learning Research*, 23(276):1–60, 2022.

Shijun Zhang, Jianfeng Lu, and Hongkai Zhao. Deep network approximation: Beyond relu to diverse activation functions. *arXiv preprint arXiv:2307.06555*, 2023.

Ding-Xuan Zhou. Universality of deep convolutional neural networks. *Applied and computational harmonic analysis*, 48(2):787–794, 2020.

A Margin conditions for real-life datasets

The goal of this section is to empirically evaluate the validity of margin conditions on two widely-used benchmark classification datasets: Fashion MNIST and CIFAR-10. Since both datasets contain more than two classes, we fall back to the binary classification setting by restricting our analysis to two arbitrarily selected classes from each dataset. Following a similar approach to (Kim et al., 2018), we investigate the margin conditions in two ways: first, we generate interpolated images between the two classes and visually examine whether such interpolated samples, which are inherently harder to classify, could realistically appear in the original datasets. Second, we train a deep Convolutional Neural Network (CNN) with ReLU activation and squared loss until convergence, and analyze the histogram of its outputs for all images in the test set. For both cases, the resulting histograms strongly suggest that the weak margin condition **(A1)** hold.
The CNN used to obtain the numerical results consists of two convolutional layers with 32 and 64 filters of size 3×3, each followed by ReLU activation and 2×2 max pooling. The output of the second convolutional layer goes through a fully connected ReLU network with architecture $\mathbf{a} = (64 \times 8 \times 8, 128, 1)$.

A.1 Fashion MNIST: “T-Shirt" vs “Pullover"

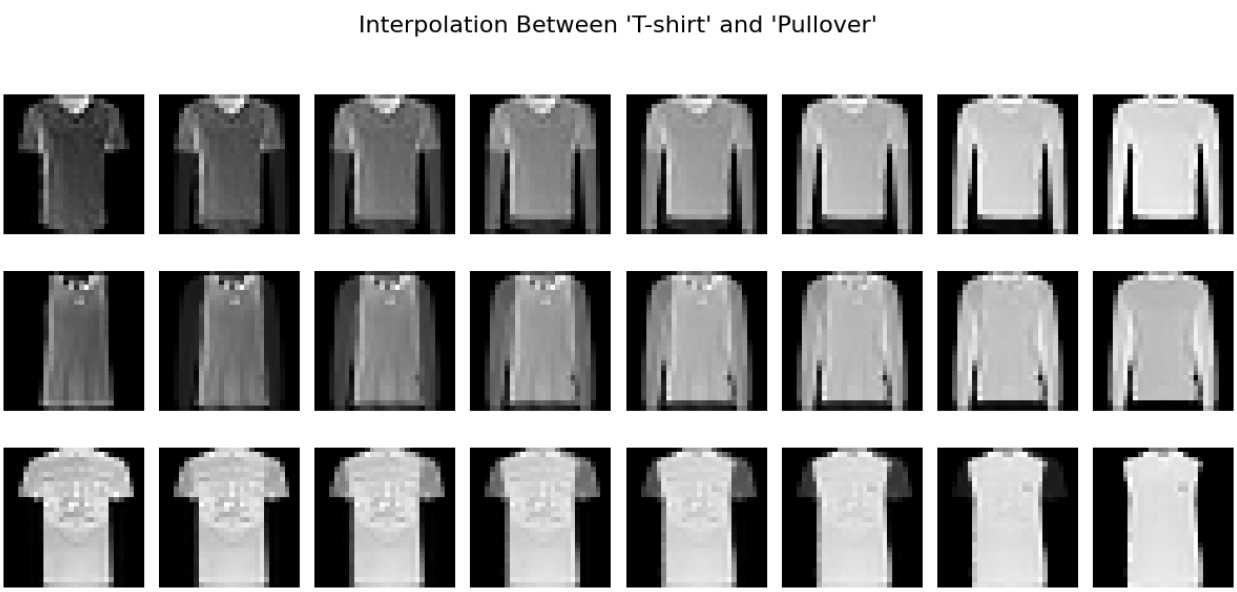


Figure 6: Interpolation of randomly selected images respectively in the “T-shirt" and “Pullover" class.

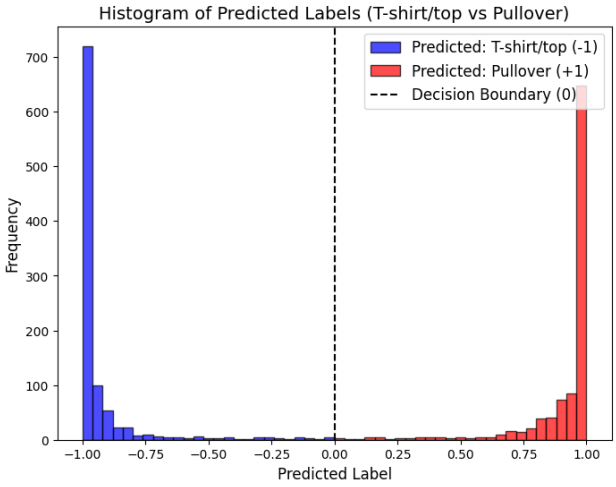


Figure 7: Histogram of predicted labels for images in the testing dataset: the histogram matches the function $t \mapsto C|t|^q$.

Despite the classes “T-Shirt" and “Pullover" exhibiting a relatively high level of visual similarity, the obtained histogram for the CNN approximation $\hat{\eta}$ of the Bayes regression function strongly suggests that the low-noise condition **(A1)** holds with a seemingly large exponent q.

A.2 CIFAR-10: “Automobile" vs “Truck"

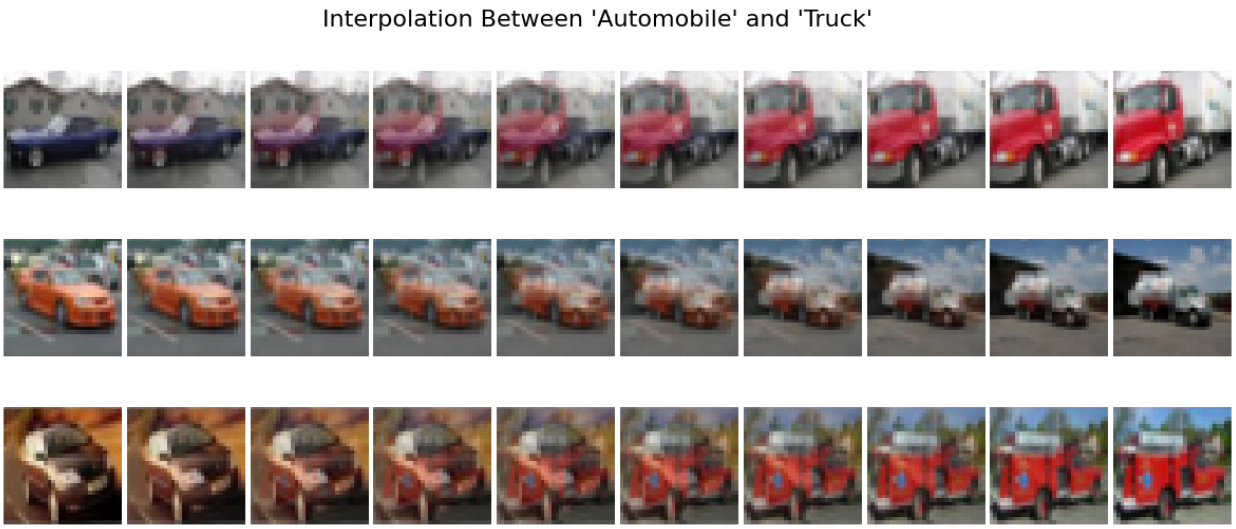


Figure 8: Interpolation of randomly selected images respectively in the “Automobile" and “Truck" class.

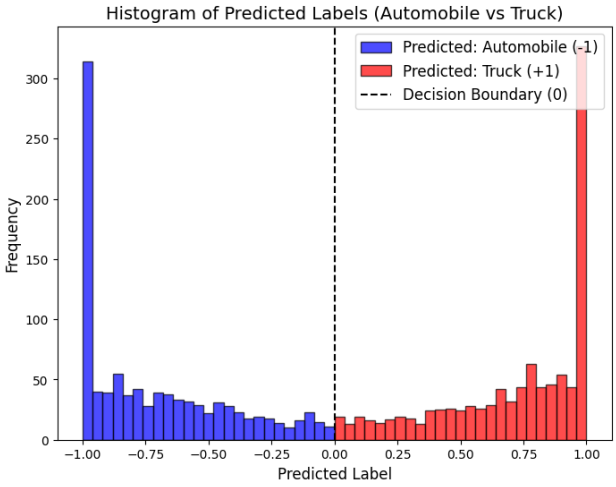


Figure 9: Histogram of predicted labels for images in the testing dataset: the decay as $t \to 0$ is much slower than for the Fashion MNIST dataset.

As for the Fashion MNIST dataset, the two selected classes have a relatively high amount of visual similarity. In this case, the histogram of labels predicted by CNN approximation $\hat{\eta}$ of the Bayes regression function suggests that the low-noise condition **(A1)** holds, but for noticeably lower values of the exponent q and multiplicative constant C.