

# Global Minima of DNNs: The Plenty Pantry

Nicole Mücke and Ingo Steinwart  
Institute for Stochastics and Applications  
University of Stuttgart

{nicole.muecke;ingo.steinwart}@mathematik.uni-stuttgart.de

May 28, 2019

## Abstract

A common strategy to train deep neural networks (DNNs) is to use very large architectures and to train them until they (almost) achieve zero training error. Empirically observed good generalization performance on test data, even in the presence of lots of label noise, corroborate such a procedure. On the other hand, in statistical learning theory it is known that over-fitting models may lead to poor generalization properties, occurring in e.g. empirical risk minimization (ERM) over too large hypotheses classes. Inspired by this contradictory behavior, so-called interpolation methods have recently received much attention, leading to consistent and optimally learning methods for some local averaging schemes with zero training error. However, there is no theoretical analysis of interpolating ERM-like methods so far. We take a step in this direction by showing that for certain, large hypotheses classes, some interpolating ERMs enjoy very good statistical guarantees while others fail in the worst sense. Moreover, we show that the same phenomenon occurs for DNNs with zero training error and sufficiently large architectures.

## 1 Introduction

During the last few decades statistical learning theory (SLT) has developed powerful techniques to analyze many variants of (regularized) empirical risk minimizers (ERMs), see e.g. [4, 15, 14, 6, 12, 13, 11]. The resulting learning guarantees, which include finite sample bounds, oracle inequalities, learning rates, adaptivity, and consistency, assume in most cases that the effective hypotheses space of the considered method is sufficiently small in terms of some notion of capacity such as VC-dimension, fat-shattering dimension, Rademacher complexities, covering numbers, or eigenvalues. Most training algorithms for DNNs also optimize an (regularized) empirical error term over a hypotheses space, namely the class of functions that can be represented by the architecture of the considered DNN, see [5, Part II]. However, unlike for many classical ERMs, the hypotheses space is parametrized in a rather complicated manner. Consequently, the optimization problem is, in general, harder to solve. A common way to address this is in practice is to use very large DNNs, since despite their size, training them is often easier, see e.g. [10, 8] and the references therein. Now, for sufficiently large DNNs it has been recently observed that common training algorithms can achieve zero training error on randomly, or arbitrarily labeled training sets, see [16]. Because of this ability, their effective hypotheses space can no longer have a sufficiently small capacity in the sense of classical SLT, so that the usual techniques for analyzing learning algorithms are no longer suitable, see e.g. the discussion on this in [16, 2]. In fact, SLT provides examples of large hypotheses spaces for which zero training error is possible but a simple ERM fails to learn. This phenomenon is known as over-fitting, and common wisdom suggests that successful learning algorithms need to avoid over-fitting, see e.g. [6, pp. 21-22]. Yet, recent empirical evidence suggests that learning in the sense of a small test error is still possible for DNNs achieving zero training error, even if the labels of data contain mis-information, see e.g. [16].

This somewhat paradoxical behavior has recently sparked some interests, leading to so-called *interpolating* learning methods, that is, learning methods that achieve zero training error. For

example, [3] establishes optimal least squares rates for the Nadaraya-Watson estimator with a particular kernel. Similar results are established for kernel ridge regression without regularization in [7]. In summary, optimal learning rates are possible for certain interpolating learning methods, so far, however, none of the considered interpolating methods has been ERM-like or even a DNN.

In this paper we consider a simple interpolating ERM as well as interpolating ReLU-DNNs of at least two hidden layers with widths growing linearly in both input dimension and sample size. For both, we show in Theorems 2.2 and 2.3 rigorous versions of the following informal statement:

**Achieving zero training error does not guarantee anything about generalization performance.**

To be more precise, we show

**Theorem 1.1.** *We can find hypotheses spaces or DNNs with exactly described minimal architecture, as well as predictors  $f_D^+$  and  $f_D^-$  from these hypotheses spaces or architectures such that:*

- i) For every training data set  $D$  both  $f_D^+$  and  $f_D^-$  are interpolating (zero training error).*
- ii) The predictor  $f_D^+$  is consistent, i.e. it learns for essentially arbitrary data generating distributions.*
- iii) The predictor  $f_D^-$  fails to learn in the worst possible sense.*
- iv) There are versions of  $f_D^+$  that achieve minmax optimal learning rates under some standard assumptions, while there are other versions of  $f_D^+$  that learn very slowly.*

*Moreover, both  $f_D^+$  and  $f_D^-$  can be found constructively using a simple and efficient training algorithm.*

The rest of the paper is organized as follows: In Section 2 we present our main results and we discuss their consequences. Section 3 is devoted to constructing statistically good and bad interpolating predictors. In Section 4, a similar construction is derived for DNNs. All proofs are deferred to the appendices.

## 2 Results

In this section we present our main results in Theorem 2.2 and Theorem 2.3 and discuss their consequences. To this end, let us begin by introducing the necessary notations and notions. Throughout this work, we consider  $X := [-1, 1]^d$  if not specified otherwise. Moreover,  $L : Y \times \mathbb{R} \rightarrow [0, \infty)$  denotes either the least squares loss  $L_{\text{ls}}(y, t) = (y - t)^2$ , the hinge loss  $L_{\text{hinge}}(y, t) = \max\{0, 1 - ty\}$ , or the binary classification loss  $L_{\text{class}}(y, t) = \mathbf{1}_{(-\infty, 0]}(y \text{ sign } t)$ , where for the latter two we consider  $Y = \{-1, 1\}$ , while for the least squares loss we consider  $Y = [-1, 1]$ . In any case, given a dataset  $D := ((x_1, y_1), \dots, (x_n, y_n)) \in (X \times Y)^n$  drawn i.i.d. from an unknown distribution  $P$  on  $X \times Y$ , the aim of supervised learning is to build a function  $f_D : X \rightarrow \mathbb{R}$  based on  $D$  such that its *risk*

$$\mathcal{R}_{L,P}(f_D) := \int_{X \times Y} L(y, f_D(x)) dP(x, y) ,$$

is close to the smallest possible risk

$$\mathcal{R}_{L,P}^* = \inf_{f: X \rightarrow \mathbb{R}} \mathcal{R}_{L,P}(f) . \tag{1}$$

In the following,  $\mathcal{R}_{L,P}^*$  is called the *Bayes risk* and an  $f_{L,P}^* : X \rightarrow \mathbb{R}$  satisfying  $\mathcal{R}_{L,P}(f_P^*) = \mathcal{R}_{L,P}^*$  is called *Bayes decision function*. Recall, that for the least squares loss,  $f_{L,P}^*$  equals the conditional mean function, i.e.  $f_{L,P}^*(x) = \mathbb{E}_P(Y|x)$  for  $P_X$ -almost all  $x \in X$ , where  $P_X$  denotes the marginal distribution of  $P$  on  $X$ . Moreover,  $\text{sign } f_{L_{\text{ls}},P}^*$  is a Bayes decision function for both  $L_{\text{hinge}}$  and  $L_{\text{class}}$ . Besides the Bayes risk we also need

$$\mathcal{R}_{L,P}^\dagger := \mathcal{R}_{L,P}(-f_{L,P}^*) .$$

Obviously, we have  $\mathcal{R}_{L,P}^\dagger > \mathcal{R}_{L,P}^*$  if  $-f_{L,P}^*$  is not another Bayes decision function. In fact, for the least squares loss a simple calculation shows

$$\mathcal{R}_{L,P}^\dagger = \mathcal{R}_{L,P}^* + 4\|f_{L,P}^*\|_2^2,$$

while for the hinge and the classification loss  $\mathcal{R}_{L,P}^\dagger$  describes the worst possible risk

$$\mathcal{R}_{L,P}^\dagger = \sup_{f: X \rightarrow Y} \mathcal{R}_{L,P}(f)$$

for all  $Y$ -valued predictors. Now, to describe the class of learning algorithms we are interested in, we need the *empirical risk* of an  $f : X \rightarrow \mathbb{R}$ , i.e.

$$\mathcal{R}_{L,D}(f) := \frac{1}{n} \sum_{i=1}^n L(y_i, f(x_i)).$$

In the following we say that an  $f : X \rightarrow Y$  *interpolates*  $D$ , if

$$\mathcal{R}_{L,D}(f) = \mathcal{R}_{L,D}^* := \inf_{\tilde{f}: X \rightarrow \mathbb{R}} \mathcal{R}_{L,D}(\tilde{f}),$$

where we emphasize that  $f$  is required to be  $Y$ -valued, while the infimum is taken over all  $\mathbb{R}$ -valued functions. It is easy to check that for all three losses  $L$  mentioned above and all data sets  $D$  there exists an  $f_D^*$  interpolating  $D$ . Moreover, for these  $L$  we have  $\mathcal{R}_{L,D}^* > 0$  if and only if  $D$  contains *contradicting samples*, i.e.  $x_i = x_k$  but  $y_i \neq y_k$ . Finally, if  $\mathcal{R}_{L,D}^* = 0$ , then any interpolating  $f_D^*$  needs to satisfy  $f_D^*(x_i) = y_i$  for all  $i = 1, \dots, n$ .

There are various ways to define nonparametric regression or classification estimates, see e.g. [6, 4]. In this paper we focus on *ERMs* and *DNNs*. Recall, that an ERM over some set  $\mathcal{F}$  of functions  $f : X \rightarrow \mathbb{R}$  chooses, for every data set  $D$ , an  $f_D \in \mathcal{F}$  that satisfies

$$\mathcal{R}_{L,D}(f_D) = \inf_{f \in \mathcal{F}} \mathcal{R}_{L,D}(f).$$

Note that the definition of ERMs implicitly requires that the infimum on the right hand side is attained, namely by  $f_D$ . In general, however,  $f_D$  does not need to be unique. It is well-known that if we have a suitably increasing sequence of hypotheses classes  $\mathcal{F}_n$  with controlled capacity, then *every* ERM  $D \mapsto f_D$  that ensures  $f_D \in \mathcal{F}_n$  for all data sets  $D$  of length  $n$  learns in the sense of e.g. universal consistency, and under additional assumptions it may also enjoy minmax optimal learning rates, see e.g. [4, 14, 6, 12]. However, the controlled capacity basically rules out interpolating ERMs. One may thus ask what happens if we consider larger hypotheses classes for which there do exist interpolating ERMs for all data sets. Our first main result now shows that in this case there may exist both *a)* well-learning interpolating ERMs and *b)* interpolating ERMs that have the worst possible learning behavior. Before stating our main results we make an assumption on the class of distributions we are considering:

**Assumption 2.1.** Define  $B_\infty := [-1, 1]^d$ . There exists a constant  $c \in (0, \infty)$  such that for any  $t \geq 0$  and  $x \in X$  one has  $P_X(x + tB_\infty) \leq ct$ .

This assumption is satisfied for instance if  $P_X$  has a bounded Lebesgue-density and can be relaxed.

**Theorem 2.2.** Let  $L$  be the least squares, the hinge, or the classification loss and suppose  $P$  is a distribution on  $X \times Y$  satisfying Assumption 2.1. For  $n \geq 1$  and  $s \in (0, 1]$  there exist a hypotheses space  $\mathcal{F}_{s,n}$  of functions  $X \rightarrow Y$  and two interpolating ERMs  $D \mapsto f_{D,s}^+$  and  $D \mapsto f_{D,s}^-$  with  $f_{D,s}^\pm \in \mathcal{F}_{s,n}$  for all data sets  $D$  of length  $|D|$  such that the following statements hold true:

*i)* For all  $(s_n) \subset (0, 1]$  with  $s_n \rightarrow 0$  and  $\frac{\log(n)}{ns_n^d} \rightarrow 0$  we have in probability for  $|D| \rightarrow \infty$ :

$$\mathcal{R}_{L,P}(f_{D,s_n}^+) \rightarrow \mathcal{R}_{L,P}^*, \quad (2)$$

$$\mathcal{R}_{L,P}(f_{D,s_n}^-) \rightarrow \mathcal{R}_{L,P}^\dagger. \quad (3)$$

ii) Let  $L$  be the least squares loss,  $f_{L,P}^*$  be  $\alpha$ -Hölder continuous and  $\gamma \in [0, \frac{2\alpha}{2\alpha+d}]$ . Then the choice

$$s_n = \left( \frac{\log(n)}{n} \right)^{\frac{1-\gamma}{d}}$$

leads to

$$\|f_{D,s_n}^+ - f_{L,P}^*\|_2^2 \leq c \left( \frac{\log(n)}{n} \right)^\gamma \quad (4)$$

$$\|f_{D,s_n}^- - (-f_{L,P}^*)\|_2^2 \leq c \left( \frac{\log(n)}{n} \right)^\gamma \quad (5)$$

with probability  $P^n$  at least  $1 - n^d e^{-n^{(1-\gamma)}}$  and for some constant  $c > 0$ . In particular, the rate in (4) is minimax optimal up to the logarithmic factor only if  $\gamma = \frac{2\alpha}{2\alpha+d}$ . Note that the choice  $s_n = 1/\log(n)$  is also possible, being independent of  $\alpha$ .

SLT shows that for small hypotheses classes, *all versions* of ERM enjoy good statistical guarantees. Theorem 2.2 demonstrates that this is no longer true for large hypotheses classes. In fact, we can find learning ERMs, see (2), (4) and ERMs whose risk converges to the worst possible one, see (3), (5) (recall that  $-f_{L,P}^*$  is **not** the Bayes decision function!). We may even have the whole spectrum between these two, with arbitrarily slow convergence as in (4), (5). For such hypotheses spaces, the description “ERM” is thus not sufficient to identify well-behaving learning algorithms. Instead, additional, or even orthogonal assumptions need to be found for learning in such hypotheses spaces.

Our next result says that the same phenomenon occurs for DNNs. To this end, we denote the class of all ReLU-DNNs with 2 hidden layer by  $\mathcal{A}_p = \mathcal{A}_{p_1,p_2}$ , with  $p = (p_1, p_2) \in \mathbb{N}^2$  and where  $p_j$  is the number of neurons in layer  $j$ , see Section D in the appendix.

**Theorem 2.3.** *Let  $L$  be the least squares loss or the hinge loss and suppose  $P$  is a distribution on  $X \times Y$  satisfying Assumption 2.1. We further let  $p_1 \geq 4dn$  and  $p_2 \geq 2n$  for all  $n \geq 1$ . Then the following statements hold true:*

- i) *For all  $n \geq 1$  there exist two interpolating DNN predictors  $D \mapsto f_D^+$  and  $D \mapsto f_D^-$  with  $f_D^+, f_D^- \in \mathcal{A}_{p_1,p_2}$  for all  $D$  of length  $n$  such that  $f_D^+$  satisfies (2) and  $f_D^-$  satisfies (3).*
- ii) *If  $L$  is the least squares loss,  $f_{L,P}^*$  is  $\alpha$ -Hölder continuous, and  $\gamma \in [0, \frac{2\alpha}{2\alpha+d}]$ , then there exist two interpolating DNN predictors  $D \mapsto f_D^+$  and  $D \mapsto f_D^-$  with  $f_D^+, f_D^- \in \mathcal{A}_{p_1,p_2}$  for all data sets  $D$  of length  $n$  such that  $f_D^+$  and  $f_D^-$  satisfy (4) and (5).*

Finally, all these predictors can be found by explicit algorithms that have an  $\mathcal{O}(d^2 n^2)$  complexity.

To fully appreciate the Theorem 2.3 let us discuss its good and bad consequences: First, the good interpolating DNN predictors  $f_D^+$  show that it is possible to train sufficiently large DNNs such that they become consistent and enjoy optimal learning rates. In addition, this training can be done in  $\mathcal{O}(d^2 \cdot n^2)$ -time if the DNNs are implemented as fully connected networks. Moreover, the constructed DNNs have a particularly sparse structure and exploiting this can actually reduce the training time to  $\mathcal{O}(d \cdot n \cdot \log n)$ . While we believe that this is one of the very first statistically sound end-to-end<sup>1</sup> proofs of consistency and optimal rates for DNNs, we also need to admit that our training algorithm is mostly interesting from a theoretical point of view, but useless for practical purposes. Nonetheless, Theorem 2.3 also has its consequences for DNNs trained by variants of stochastic gradient descent (SGD) if the resulting predictor is interpolating. Indeed, Theorem 2.3 shows that ending in a global minimum can have all sorts of consequences ranging from very good to very bad learning behavior. So far, however, there is no statistically sound way to distinguish between good and bad interpolating DNNs on the basis of the training set alone, and hence the only

<sup>1</sup>By “end-to-end” we mean the explicit construction of an efficient, feasible, and implementable training algorithm and the rigorous statistical analysis of this very particular algorithm under minimal assumptions.

way to identify good interpolating DNNs obtained by SGD is to use a validation set. Now, for the good interpolating DNNs of Theorem 2.3 it is actually possible to construct a finite set of candidates such that the one with the best validation error achieves the optimal learning rates without knowing  $\alpha$ . For DNNs trained by SGD, however, we do not have this luxury anymore. Indeed, while we can still identify the best predicting DNN from a finite set of SGD-learned interpolating DNNs we no longer have any theoretical understanding of whether there is any useful candidate among them, or whether they all behave like an  $f_D^-$ .

### 3 The Histogram Rule Revisited

In this section we construct the good and bad interpolating ERM of Theorem 2.2. In a nutshell, the basic idea is to first consider classical histogram rules (HR), and then to inflate their hypotheses space so that we can find interpolating ERM in these inflated hypotheses spaces that coincide with either the corresponding HR or its opposite predictor.

Let us begin by saying that  $L$  is *interpolatable* for  $D$  if there exists an  $f : X \rightarrow Y$  that *interpolates*  $D$ , i.e.  $\mathcal{R}_{L,D}(f) = \mathcal{R}_{L,D}^*$ . Clearly, an  $f : X \rightarrow Y$  interpolates  $D$  if and only if

$$\sum_{k:x_k=x_i^*} L(x_i, y_i, f(x_i^*)) = \inf_{c \in \mathbb{R}} \sum_{k:x_k=x_i^*} L(x_i, y_i, c), \quad i = 1, \dots, m, \quad (6)$$

where  $x_1^*, \dots, x_m^*$  are the elements of  $D_X := \{x_i : i = 1, \dots, n\}$ . Note that (6) in particular ensures that the infimum over  $\mathbb{R}$  on the right is attained at some  $c_i^* \in Y$ . Many common losses including the least squares, the hinge, and the classification loss interpolate all  $D$ , and for the latter three losses we have  $\mathcal{R}_{L,D}^* > 0$  if and only if  $D$  contains *contradicting samples*, i.e.  $x_i = x_k$  but  $y_i \neq y_k$ . Moreover, for the least squares loss,  $c_i^*$  can be easily computed by averaging over all labels  $y_k$  that belong to some sample  $x_k$  with  $x_k = x_i$ . For the hinge and classification loss we then have to take  $f(x_i) = \text{sign } c_i^*$ , where  $c_i^*$  is the solution obtained for the least squares loss, and  $\text{sign } 0 := 0$ .

A particular simple ERM are HRs. To recall the latter, we fix a finite partition  $\mathcal{A} = (A_1, \dots, A_m)$  of  $X$  and for  $x \in X$  we write  $A(x)$  for the unique cell  $A_j$  with  $x \in A_j$ . Moreover, we define

$$\mathcal{H}_{\mathcal{A}} := \left\{ \sum_{j=1}^m c_j \mathbf{1}_{A_j} : c_j \in Y \right\}, \quad (7)$$

where  $\mathbf{1}_{A_j}$  denotes the *indicator function* of the cell  $A_j$ . Now, given a data set  $D$  and a loss  $L$  an  $\mathcal{A}$ -*histogram* is an  $h_{D,\mathcal{A}} = \sum_{j=1}^m c_j^* \mathbf{1}_{A_j} \in \mathcal{H}_{\mathcal{A}}$  that satisfies

$$\sum_{i:x_i \in A_j} L(x_i, y_i, c_j^*) = \inf_{c \in Y} \sum_{i:x_i \in A_j} L(x_i, y_i, c) \quad (8)$$

for all *non-empty cells*  $A_j$ , that is  $\{i : x_i \in A_j\} \neq \emptyset$ . Clearly,  $D \mapsto h_{D,\mathcal{A}}$  is an ERM. Moreover, note that in general  $h_{D,\mathcal{A}}$  is *not uniquely determined*, since  $c_j^* \in Y$  can take arbitrary values on empty cells  $A_j$ . In particular, *there are more than one ERM over  $\mathcal{H}_{\mathcal{A}}$  as soon as  $m, n \geq 2$ .*

Before we proceed, let us consider a few examples. First, for the least squares loss, a simple calculation shows that for all non-empty cells  $A_j$ , the coefficient  $c_j^*$  in (8) is uniquely determined by

$$c_j^* = \frac{1}{|\{i : x_i \in A_j\}|} \sum_{i:x_i \in A_j} y_i. \quad (9)$$

In the following, we call every resulting  $D \mapsto h_{D,\mathcal{A}}$  with  $h_{D,\mathcal{A}} = \sum_{j=1}^m c_j^* \mathbf{1}_{A_j} \in \mathcal{H}_{\mathcal{A}}$  an empirical HR for regression (HRR). Moreover, if  $L$  is either the hinge loss  $L(y, t) := \max\{0, 1 - yt\}$  or the classification loss  $L(y, t) := \mathbf{1}_{(-\infty, 0]}(y \text{ sign } t)$  with  $\text{sign } 0 := 1$ , then it is well-known that  $\text{sign } c_j^*$  is a solution of (8), where  $c_j^*$  is given by (9) for all non-empty cells  $A_j$ . Note that this simply means that for the hinge and classification loss the coefficient in (9) is determined by a majority vote over the labels  $y_i$  occurring in the cell  $A_j$ , where a tie is broken by voting for  $y = 1$ . Consequently, the

plug-in estimator  $\text{sign } h_{D,\mathcal{A}}$  is an  $\mathcal{A}$ -histogram for both losses, if  $h_{D,\mathcal{A}}$  is an HRR. In the following, we call every such  $D \mapsto \text{sign } h_{D,\mathcal{A}}$  an empirical HR for classification (HRC).

We are mostly interested in HRs on  $X = [-1, 1]^d$  whose underlying partition essentially consists of cubes with a fixed width. To rigorously deal with boundary effects, we first say that a partition  $(B_j)_{j \geq 1}$  of  $\mathbb{R}^d$  is cubic partition (CP) of width  $s > 0$ , if each cell  $B_j$  is a translated version of  $[0, s)^d$ , i.e. there is an  $x^\dagger \in \mathbb{R}^d$  called *offset* such that for all  $j \geq 1$  there exist  $k := (k_1, \dots, k_d) \in \mathbb{Z}^d$  with

$$B_j = x^\dagger + sk + [0, s)^d. \quad (10)$$

Now, a partition  $\mathcal{A} = (A_j)_{j \in J}$  of  $X = [-1, 1]^d$  is a CP of width  $s > 0$ , if there is a CP  $(B_j)_{j \geq 1}$  of  $\mathbb{R}^d$  with width  $s > 0$  such that  $J = \{j \geq 1 : B_j \cap X \neq \emptyset\}$  and  $A_j = B_j \cap X$  for all  $j \in J$ . If  $s \in (0, 1]$ , then, up to reordering, this  $(B_j)_{j \geq 1}$  is uniquely determined by  $\mathcal{A}$ .

If the hypotheses space (7) is based on a cubic partition of  $X = [-1, 1]^d$  with width  $s > 0$ , then the resulting HRRs and HRCs are well understood. For example, universal consistency and learning rates have been established for both the least squares and the classification loss. In general, these results only require a suitable choice for the widths  $s = s_n$  for  $n \rightarrow \infty$  but no specific choice of the cubic partition of width  $s$ . For this reason we write  $\mathcal{H}_s := \bigcup \mathcal{H}_{\mathcal{A}}$ , where the union runs over all CPs  $\mathcal{A}$  of  $X$  with fixed width  $s \in (0, 1]$ . Our next goal is to consider *inflated versions* of  $\mathcal{H}_s$ . Namely, for  $r, s > 0$  and  $m \geq 0$  we define

$$\mathcal{F}_{s,r,m} := \left\{ h + \sum_{i=1}^m b_i \mathbf{1}_{x_i^* + tB_\infty} : h \in \mathcal{H}_s, b_i \in 2Y \cup \{0\}, x_i^* \in X, t \in [0, r] \right\},$$

where  $B_\infty := [-1, 1]^d$ . In other words, we have  $\mathcal{F}_{s,r,0} = \mathcal{H}_s$  and for  $m \geq 1$ , an  $f \in \mathcal{F}_{s,r,m}$  changes an  $h \in \mathcal{H}_s$  on at most  $m$  small neighborhoods of some arbitrary  $x_1^*, \dots, x_m^*$ . In general, these small neighborhoods  $x_i^* + tB_\infty$  may intersect and may be contained in more than one cell  $A_j$  of the considered  $\mathcal{A}$ . Since this may cause undesired boundary effects we say that an  $f \in \mathcal{F}_{s,r,m}$  is *properly aligned* if it has a representation

$$f = \sum_{j \in J} c_j \mathbf{1}_{A_j} + \sum_{i=1}^m b_i \mathbf{1}_{x_i^* + tB_\infty} \quad (11)$$

as in the definition of  $\mathcal{F}_{s,r,m}$  and for all  $i, k = 1, \dots, m$  we have

$$x_i^* + tB_\infty \subset B(x_i^*), \quad (12)$$

$$x_i^* + tB_\infty \cap x_k^* + tB_\infty = \emptyset \quad \text{whenever } i \neq k, \quad (13)$$

where  $B(x_i^*)$  is the unique cell  $x_i^* \in B(x_i^*)$  of the partition  $(B_j)_{j \geq 1}$  that defines  $\mathcal{A}$ . Note that this gives  $A(x_i^*) = B(x_i^*) \cap X$ . In the following,  $\mathcal{F}_{s,r,m}^*$  denotes the set of all properly aligned  $f \in \mathcal{F}_{s,r,m}$ .

Our next goal is to show that  $\mathcal{F}_{s,r,m}^*$  contains interpolating predictors if  $r$  is sufficiently small and  $m \geq n$ . To this end, note that (13) holds for all  $t > 0$  with  $t < \frac{1}{2} \min_{i,k:i \neq k} \|x_i^* - x_k^*\|_\infty$ . Clearly, a brute-force algorithm finds such a  $t$  in  $\mathcal{O}(dm^2)$ -time. However, a smarter approach is to first sort the first coordinates  $x_{1,1}^*, \dots, x_{m,1}^*$  and to determine the smallest positive distance  $t_1$  of two consecutive, non-identical ordered coordinates. This approach is then repeated for the remaining  $d-1$ -coordinates, so at the end we have  $t_1, \dots, t_d$ . Then  $t := \min\{t_1, \dots, t_d\}/3$  satisfies (13) and the used algorithm is  $\mathcal{O}(d \cdot m \log m)$  in time. Our next result shows that we can also ensure (12) by jiggling the CPs.

**Theorem 3.1.** *For all  $d \geq 1$ ,  $s \in (0, 1]$ , and  $m \geq 1$  there exist  $x_1^\dagger, \dots, x_K^\dagger \in \mathbb{R}^d$  with  $K := (m+1)^d$  such that for all  $x_1^*, \dots, x_m^* \in [-1, 1]^d$  we find an  $\ell \in \{1, \dots, K\}$  such that the CP given by (10) with offset  $x_\ell^\dagger$  satisfies (12) for all  $t > 0$  with  $t \leq \frac{s}{3m+3}$ .*

While at first glance the number  $K$  of candidate offsets seems to be prohibitively large for an efficient search, it turns out that the proof of Theorem (3.1) actually provides a simple  $\mathcal{O}(d \cdot m)$ -algorithm for identifying  $x_\ell^\dagger$  coordinate-wise. This algorithm was used to find the aligned partition in Figure 1.

The next result provides a sufficient condition for interpolating predictors in  $\mathcal{F}_{s,r,m}^*$ .

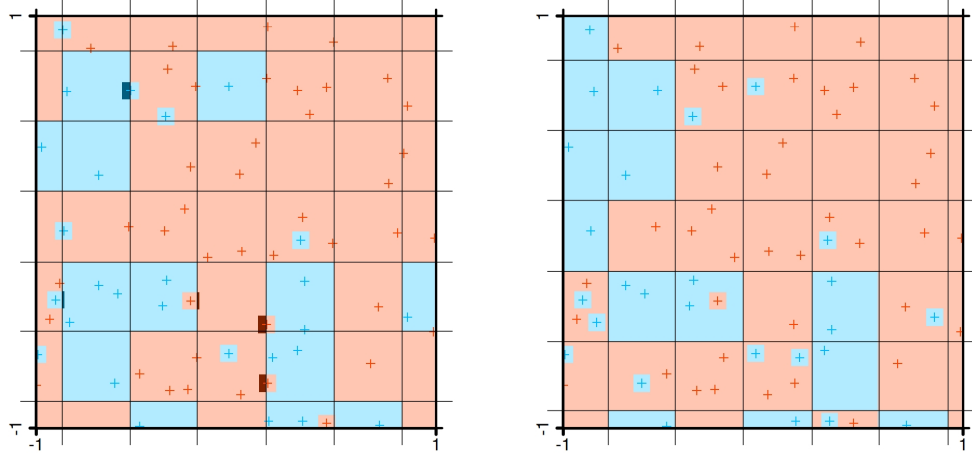


Figure 1: *Left.* An inflated histogram on  $X = [-1, 1]^2$  for binary classification with positively (red) and negatively (blue) labeled samples (crosses). The predictions  $c_j^*$  on the 49 cells of the cubic partition are determined by (8), i.e. by a majority vote if possible, and  $c_j^* = 1$ , otherwise. Misclassified samples are corrected according to (14) on a  $tB_\infty$ -neighborhood for some small  $t > 0$ . The lighter red and blue background colors display the predictions of the inflated HR. Note that a few samples are too close to the cell boundaries, i.e. (12) is violated. If the neighboring cell of such a sample has an opposite prediction, the predictions of the inflated HR are no longer in  $\{-1, 1\}$ . The regions where this happens are colored in dark blue and dark red, respectively. *Right.* An inflated HR on  $X = [-1, 1]^2$  that is properly aligned to the same data set. Note that (12) ensures that boundary effects as for the left HR do not take place. As a result, all predictions are in  $\{-1, 1\}$ . For inflated HRs these effects seem to be a negligible technical nuisance. For their DNN counterparts considered in Section 4, however, such effects may significantly complicate the constructions of interpolating predictors, see Figure 2.

**Proposition 3.2.** *Let  $L$  be a loss that is interpolatable for  $D = ((x_1, y_1), \dots, (x_n, y_n))$  and let  $x_1^*, \dots, x_m^*$  be as in (6). Moreover, for  $s \in (0, 1]$  and  $r > 0$  we fix an  $f^* \in \mathcal{F}_{s,r,m}^*$  with representation (11) satisfying (12) and (13). For  $i = 1, \dots, m$  let  $j_i$  be the index such that  $x_i^* \in A_{j_i}$ . Then  $f^*$  interpolates  $D$  if for all  $i = 1, \dots, m$  we have*

$$b_i = -c_{j_i} + \arg \min_{c \in Y} \sum_{k: x_k = x_i^*} L(x_k, y_k, c). \quad (14)$$

Note that for all  $c_{j_i} \in Y$  the value  $b_i$  given by (14) satisfies  $b_i \in 2Y \cup \{0\}$  and we have  $b_i = 0$  if  $c_{j_i}$  is contained in the  $\arg \min$  in (14). Moreover, (14) shows that an interpolating  $f^* \in \mathcal{F}_{s,r,m}^*$  can be an arbitrary  $\mathcal{A}$ -step function  $h \in \mathcal{H}_{\mathcal{A}}$  outside the small  $tB_\infty$ -neighborhoods around the samples of  $D$ . In other words, as soon as we have found at least one such  $f^* \in \mathcal{F}_{s,r,m}^*$ , we can arbitrarily change it outside these small neighborhoods by changing its  $\mathcal{H}_{\mathcal{A}}$ -part and recomputing the  $b_i$ 's by (14). Based on this observation, we can now construct different, interpolating  $f_D^* \in \mathcal{F}_{s,r,m}^*$  that have particularly good and bad learning behaviors.

**Example 3.3 (Good and bad interpolating ERM's).** Let  $L$  be the least squares loss and  $D = ((x_1, y_1), \dots, (x_n, y_n))$  be a data set. For  $s \in (0, 1]$ ,  $r := 2^{-n}$ , and  $m := n$  we fix an  $f_{D,s}^+ \in \mathcal{F}_{s,r,m}^*$  with representation (11) satisfying (12) and (13) such that (14) holds and such that its  $\mathcal{H}_{\mathcal{A}}$ -part

$$h_{D,\mathcal{A}}^+ := \sum_{j \in J} c_j^+ \mathbf{1}_{A_j}$$

is an HRR. Analogously, for the hinge and classification loss  $L$  we demand  $h_{D,\mathcal{A}}^+$  to be an HRC and (14) needs to hold for  $L$ , instead. In all three cases  $f_{D,s}^+$  is an interpolating predictor. In the

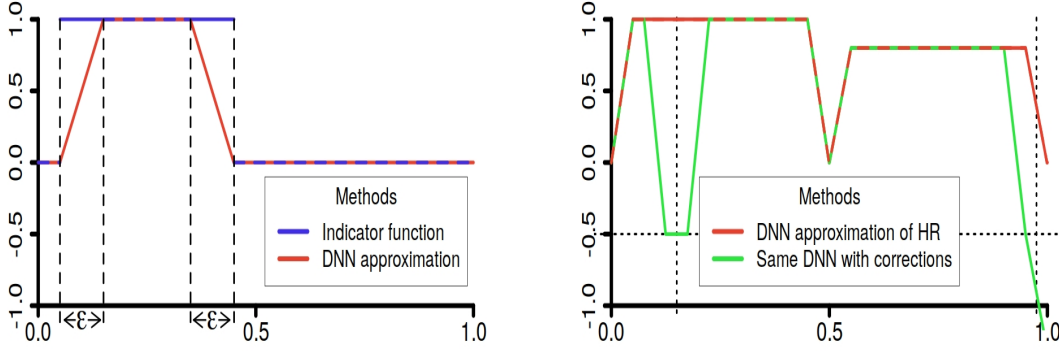


Figure 2: Left. Approximation  $g_A^{(\varepsilon)}$  (orange) of the indicator function  $\mathbf{1}_A$  for  $A = [0.05, 0.45]$  (blue) according to Lemma 4.1 for  $\varepsilon = 0.1$  on  $X = [0, 1]$ . The construction of  $g_A^{(\varepsilon)}$  ensures that  $g_A^{(\varepsilon)}$  coincides with  $\mathbf{1}_A$  modulo a small set that is controlled by  $\varepsilon > 0$ . Right. A DNN (orange) for regression that approximates the histogram  $\mathbf{1}_{[0,0.5]} + 0.8 \cdot \mathbf{1}_{[0.5,1]}$  and a DNN (green) that additionally tries to interpolate two samples  $x_1 = 0.15$  and  $x_2 = 0.975$  (located at the two vertical dotted lines) with  $y_i = -0.5$ . The label  $y_1$  is correctly interpolated since the alignment condition (12) is satisfied for  $x_1$  with  $t = 0.15$  and  $\varepsilon = \delta = t/3 = 0.05$  as in Example 4.2. In contrast,  $y_2$  is not correctly interpolated since condition (12) is violated for this  $t$  and hence  $\varepsilon$  and  $\delta$  are too large.

following, a  $D \mapsto f_{D,s}^+$  that chooses the CP  $\mathcal{A}$  from the candidates described in Theorem 3.1 is called a *good interpolating histogram rule*.

To find a bad interpolating predictor for  $D$  we consider the good  $f_{D,s}^+ \in \mathcal{F}_{s,r,m}^*$  just constructed. Then  $f_{D,s}^- \in \mathcal{F}_{s,r,m}^*$  denotes a predictor whose  $\mathcal{H}_A$ -part is  $h_{D,A}^- := -h_{D,A}^+$  and whose  $b_i$ 's are obtained by (14). Clearly,  $f_{D,s}^-$  is an interpolating predictor with representation (11) satisfying (12), (13) and (14). In the following,  $D \mapsto f_{D,s}^-$  is called a *bad interpolating histogram rule*.

## 4 Approximation of Histograms with ReLU Networks

The goal of this section is to construct the DNNs of Theorem 2.3. To this end, we mimic inflated histogram rules with DNNs of suitable depth and width.

Motivated by the representation (11) for histograms, the first step of our construction approximates the indicator function of an multi-dimensional interval by a small part of a possibly large DNN. This will be our main building block. Note that the ReLU activation function is particularly suited for this approximation and it thus plays a key role in our entire construction.

For the formulation of the corresponding lemma we fix some notation. For  $z_1, z_2 \in \mathbb{R}^d$  we write  $z_1 \leq z_2$  if each coordinate satisfies  $z_{1,i} \leq z_{2,i}$ ,  $i = 1, \dots, d$ . We define  $z_1 < z_2$  analogously. In addition, if  $z_1 \leq z_2$ , then the multi-dimensional interval is  $[z_1, z_2] := \{z \in \mathbb{R}^d : z_1 \leq z \leq z_2\}$ , and we similarly define  $(z_1, z_2)$  if  $z_1 < z_2$ . Finally, for  $s \in \mathbb{R}$ , we let  $z_1 + s := (z_{1,1} + s, \dots, z_{1,d} + s)$ .

**Lemma 4.1.** *Let  $z_1, z_2 \in \mathbb{R}^d$  with  $z_1 < z_2$  and  $\varepsilon > 0$  with  $\varepsilon < \min\{z_{2,i} - z_{1,i} : i = 1, \dots, d\}$ . Then for all  $A \subset X$  with  $[z_1 + \varepsilon, z_2 - \varepsilon] \subset A \subset [z_1, z_2]$  there exists a DNN  $g_A^{(\varepsilon)} \in \mathcal{A}_{2d,1}$  with*

$$\{g_A^{(\varepsilon)} = \mathbf{1}_A\} = [z_1 + \varepsilon, z_2 - \varepsilon] \cup (X \setminus (z_1, z_2)) .$$

Figure 2 illustrates  $g_A^{(\varepsilon)}$  for  $d = 1$ . Moreover, the proof of Lemma 4.1 shows that out of the  $2d^2$  weight parameters of the first layer, only  $2d$  are non-zero. In addition, the  $2d$  weight parameters of the neuron in the second layer are all identical. In order to approximate inflated histograms we need to know how to combine several functions of the form provided by Lemma 4.1 into a single neural



network. A particularly appealing feature of our DNNs is that the concatenation of layer structures is very easy. To be more precise, if  $c \in \mathbb{R}$ ,  $(p_1, p_2) \in \mathbb{N}^2$ , and  $g \in \mathcal{A}_p$ ,  $g' \in \mathcal{A}_{p'}$ , then  $cg \in \mathcal{A}_p$  and  $g + g' \in \mathcal{A}_{p+p'}$ , see Lemma D.1. In particular, our constructed DNNs have a particularly sparse structure and the number of required neurons behaves in a very controlled and natural fashion.

With these insights, we are now able to find a representation similar to (11) and to define good and bad interpolating DNNs, similarly to good and bad interpolating ERMs presented in Example 3.3. To this end, we choose a CP  $\mathcal{A} = (A_j)_{j \in J}$  of  $X$  with width  $s > 0$  and define

$$\mathcal{H}_{\mathcal{A}}^{(\varepsilon)} := \left\{ \sum_{j \in J} c_j g_{A_j}^{(\varepsilon)} : c_j \in Y \right\}, \quad 0 < \varepsilon \leq \frac{s}{3},$$

where  $g_{A_j}^{(\varepsilon)} := (g_{B_j}^{(\varepsilon)})|_{A_j}$  is the restriction of  $g_{B_j}^{(\varepsilon)}$  to  $A_j$  and  $g_{B_j}^{(\varepsilon)}$  is the  $\varepsilon$ -approximation of  $\mathbf{1}_{B_j}$  of Lemma 4.1, where  $B_j$  is cell with  $A_j = \cap B_j \cap X$ , see the text around (10). Moreover, we write  $\mathcal{H}_s^{(\varepsilon)} := \bigcup \mathcal{H}_{\mathcal{A}}^{(\varepsilon)}$ , where the union runs over all CPs  $\mathcal{A}$  of  $X$  with width  $s > 0$ . Our considerations above show that we have  $\mathcal{H}_s^{(\varepsilon)} \subset \mathcal{A}_{p_1, p_2}$  with  $p_1 = 2d|J|$  and  $p_2 = |J|$ . Any  $\varepsilon$ -approximate histogram, i.e., any function in  $\mathcal{H}_s^{(\varepsilon)}$ , can therefore be represented by a DNN with 2 hidden layers. Inflated versions are now straightforward. Namely, for  $r, s, \varepsilon > 0$  and  $m \geq 0$  we define

$$\mathcal{F}_{s,r,m}^{(\varepsilon)} := \left\{ h^{(\varepsilon)} + \sum_{i=1}^m b_i g_{x_i^* + tB_\infty}^{(\delta)} : h^{(\varepsilon)} \in \mathcal{H}_s^{(\varepsilon)}, b_i \in 2Y \cup \{0\}, x_i^* \in X, t \in (0, r], \delta \in (0, t/3] \right\},$$

where  $g_{x_i^* + tB_\infty}^{(\delta)}$  is a  $\delta$ -approximation of  $\mathbf{1}_{x_i^* + tB_\infty}$ . A short calculation shows that  $\mathcal{F}_{s,r,m}^{(\varepsilon)} \subset \mathcal{A}_{p_1, p_2}$  with  $p_1 = 2d(m + |J|)$ ,  $p_2 = m + |J|$  and  $|J| \leq (2/s)^d$ . With these preparations, we can now introduce good and bad interpolating DNNs.

**Example 4.2 (Good and bad interpolating DNN).** Let  $L$  be the least squares or the hinge loss, and  $D = ((x_1, y_1), \dots, (x_n, y_n))$  be a data set. For  $s \in (0, 1]$ ,  $r := 2^{-n}$ , and  $m := n$  let

$$f_{D,s}^+ = \sum_{j \in J} c_j^+ \mathbf{1}_{A_j} + \sum_{i=1}^m b_i \mathbf{1}_{x_i^* + tB_\infty} \in \mathcal{F}_{s,r,m}^*$$

be the good interpolating HR according to Example 3.3. In particular,  $t > 0$  satisfies (12) and (13). For  $\varepsilon := \delta := t/3$  we then define the *good interpolating DNN* by

$$g_{D,s}^+ = \sum_{j \in J} c_j^+ g_{A_j}^{(\varepsilon)} + \sum_{i=1}^m b_i g_{x_i^* + tB_\infty}^{(\delta)}.$$

Clearly, we have  $\varepsilon = \delta \leq t/3 \leq \min\{2^{-n}, s/6\}$  and  $g_{D,s}^+ \in \mathcal{F}_{s,r,n}^{(\varepsilon)}$ . Moreover, for  $s \geq 2n^{-1/d}$  we find  $|J| \leq n$ , and hence  $g_{D,s}^+ \in \mathcal{A}_{4dn, 2n}$ . Note that every wider and/or deeper architecture includes  $\mathcal{A}_{4dn, 2n}$ . Finally, the *bad interpolating DNN*  $g_{D,s}^-$  is defined analogously using the bad interpolating HR from Example 3.3, instead.

## Acknowledgments

NM is supported by the German Research Foundation under DFG Grant STE 1074/4-1.

## References

- [1] H. Bauer. *Measure and Integration Theory*. De Gruyter, Berlin, 2001.
- [2] M. Belkin, D. J. Hsu, and P. Mitra. Overfitting or perfect fitting? Risk bounds for classification and regression rules that interpolate. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 2300–2311. Curran Associates, Inc., 2018.

- [3] M. Belkin, A. Rakhlin, and A. B. Tsybakov. Does data interpolation contradict statistical optimality? Technical report, 2018. <https://arxiv.org/abs/1806.09471>.
- [4] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, New York, 1996.
- [5] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, Cambridge, MA, 2016.
- [6] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer-Verlag, New York, 2002.
- [7] T. Liang and A. Rakhlin. Just interpolate: Kernel” ridgeless” regression can generalize. Technical report, 2018. <https://arxiv.org/abs/1808.00387>.
- [8] S. Ma, R. Bassily, and M. Belkin. The power of interpolation: Understanding the effectiveness of SGD in modern over-parametrized learning. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3325–3334, 2018.
- [9] M. Meister and I. Steinwart. Optimal learning rates for localized SVMs. *J. Mach. Learn. Res.*, 17:1–44, 2016.
- [10] R. Salakhutdinov. Deep learning tutorial at the Simons Institute, 2017. <https://simons.berkeley.edu/talks/ruslan-salakhutdinov-01-26-2017-1>.
- [11] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, Cambridge, 2014.
- [12] I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, New York, 2008.
- [13] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer-Verlag, New York, 2009.
- [14] S. van de Geer. *Applications of Empirical Process Theory*. Cambridge University Press, Cambridge, 2000.
- [15] V. N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, New York, 1998.
- [16] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. Technical report, 2016. <http://arxiv.org/abs/1611.03530>.

## A Histograms based on Data-Dependent Partitions

Our aim is to show consistency and to derive learning rates for histograms based on a random partition of the input space. We first introduce some notation: We denote by  $\mathcal{P}_s$  the set of all CPs of fixed width  $s \in (0, 1]$  of  $\mathbb{R}^d$  and by  $\mathcal{P}_s^X$  the set of all CPs on  $X$ . Note that cells of  $\mathcal{A} \in \mathcal{P}_s^X$  are obtained by intersecting the cells of  $\mathcal{B} \in \mathcal{P}_s$  with  $X$ . An  $m$ -sample CP rule of fixed width  $s \in (0, 1]$  for  $\mathbb{R}^d$  is a map  $\pi_{s,m} : \mathbb{R}^{dm} \rightarrow \mathcal{P}_s$  to which we associate a non-random family of partitions  $\mathcal{P}_{s,m} := \pi_{s,m}(\mathbb{R}^{dm}) \subset \mathcal{P}_s$ . Thus,  $\mathcal{P}_{s,m}$  is the set of all partitions generated by the rule  $\pi_{s,m}$  for all possible realizations of a training set  $D_X$ . In particular,  $\pi_{s,m}$  applied to the input training sample  $(x_1^*, \dots, x_m^*)$  produces a *data-dependent* CP. Again, by restricting the cells of an element  $\mathcal{B} \in \mathcal{P}_{s,m}$  to  $X$ , we obtain a partition  $\mathcal{A}$  of  $X$  and the set of all such  $\mathcal{A}$  will be denoted by  $\mathcal{P}_{s,m}^X$ . Recall that Theorem 3.1 provides us with a special partitioning rule  $\pi_{s,m} : \mathbb{R}^{dm} \rightarrow \mathcal{P}_s$ , where  $|\mathcal{P}_{s,m}| = (m+1)^d$ .

### A.1 Regression

Let us also introduce an infinite sample version of a classical histogram

$$h_{P,\mathcal{A}} = \sum_{j \in J} c_j \mathbf{1}_{A_j}, \quad c_j = \frac{1}{P_X(A_j)} \int_{A_j} f_{L,P}^*(x) dP_X(x).$$

Similarly to empirical histograms one has

$$\mathcal{R}_{L,P}(h_{P,\mathcal{A}}) = \inf_{h \in \mathcal{H}_{\mathcal{A}}} \mathcal{R}_{L,P}(h).$$

Recall, the histogram rule for regression based on a data-dependent partition  $\mathcal{A}_D \in \mathcal{P}_{s,n}^X$  is a map  $\mathcal{L} : D \mapsto h_{D,\mathcal{A}_D}$  with

$$h_{D,\mathcal{A}_D} = \sum_{j \in J} c_j^* \mathbf{1}_{A_j},$$

where the  $c_j^*$  are defined in (9).

Our first result establishes an oracle inequality for the excess risk of the histogram rule based on a data-dependent partition for regression when using the least square loss.

**Proposition A.1** (Oracle Inequality). *Let  $L$  be the least square loss. Assume that  $|\mathcal{P}_{s,n}^X| = K_{s,n} < \infty$ . For any  $\varepsilon > 0$ ,  $\tau > 0$*

$$\begin{aligned} \sup_{\mathcal{A} \in \mathcal{P}_{s,n}^X} \mathcal{R}_{L,P}(h_{D,\mathcal{A}}) - \mathcal{R}_{L,P}^* &\leq 6 \sup_{\mathcal{A} \in \mathcal{P}_{s,n}^X} (\mathcal{R}_{L,P,\mathcal{H}_{\mathcal{A}}}^* - \mathcal{R}_{L,P}^*) + 8\varepsilon \\ &+ 128 \sup_{\mathcal{A} \in \mathcal{P}_{s,n}^X} \frac{\tau + 1 + \log \mathcal{N}(\mathcal{H}_{\mathcal{A}}, \|\cdot\|_{\infty}, \varepsilon)}{n}, \end{aligned}$$

with  $P^n$ -probability at least  $1 - 2K_{s,n}e^{-\tau}$ .

**Proof of Proposition A.1:** This follows from the result in [12], p. 284 by taking a union bound. In particular, the assumptions required there are satisfied with  $\theta = 1$  and  $B = V = 4$ , resulting from  $L$  being the least square loss (see [12], p. 245).  $\square$

In the following, we establish universal consistency of  $\mathcal{L}$  with respect to the least square loss and derive learning rates. To do so we need an additional assumption for the a-priori smoothness of the regression function.

**Assumption A.2 (Regularity).** For  $\alpha \in (0, 1]$  and  $C > 0$  we let  $\Sigma(\alpha, C)$  denote the class of  $\alpha$ -Hölder continuous functions  $f : X \rightarrow \mathbb{R}$ , i.e.,

$$|f(x) - f(x')| \leq C \|x - x'\|^\alpha,$$

for any  $x, x' \in X$ .

**Lemma A.3** (Approximation Error). *Let  $\mathcal{A}$  be a CP of width  $s \in (0, 1]$ . Then, for any  $\varepsilon > 0$  there exists  $s_\varepsilon > 0$  such that for any CP of width  $s \in (0, s_\varepsilon]$  one has*

$$\mathcal{R}_{L,P}(h_{P,\mathcal{A}}) - \mathcal{R}_{L,P}^* < \varepsilon.$$

Moreover, if the regularity Assumption A.2 holds, then for all  $s \in (0, 1]$  we have

$$\|h_{P,\mathcal{A}} - f_{L,P}^*\|_2 < 2Cs^\alpha.$$

**Proof of Lemma A.3:** For the proof of the first assertion we fix an  $\varepsilon > 0$ . Then recall that there exists a continuous  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  with compact support such that

$$\|f_{L,P}^* - f\|_2 \leq \varepsilon, \quad (15)$$

see e.g. [1, Theorem 29.14 and Lemma 26.2]. Moreover, since  $\|f_{L,P}^*\|_\infty \leq 1$ , we can assume without loss of generality that  $\|f\|_\infty \leq 1$ . Moreover, since  $f$  is continuous and has compact support,  $f$  is uniformly continuous, and hence there exists a  $\delta \in (0, 1]$  such that for all  $x, x' \in X$  with  $\|x - x'\|_\infty \leq \delta$  we have

$$|f(x) - f(x')| \leq \varepsilon. \quad (16)$$

We define  $s_\varepsilon := \delta$ . Now, we fix a CP  $\mathcal{A} = (A_j)_{j \in J}$  of width  $s > 0$  for some  $s \in (0, s_\varepsilon]$ . For  $x \in X$  with  $P_X(A(x)) > 0$  we then have

$$h_{P,\mathcal{A}}(x) = \frac{1}{P_X(A(x))} \int_{A(x)} f_{L,P}^* dP_X.$$

For such  $x$  we then define

$$\bar{f}(x) := \frac{1}{P_X(A(x))} \int_{A(x)} f dP_X.$$

For the remaining  $x \in X$  we simply set  $\bar{f}(x) := 0$ . With these preparations we then have

$$\|h_{P,\mathcal{A}} - f_{L,P}^*\|_2 \leq \|h_{P,\mathcal{A}} - \bar{f}\|_2 + \|\bar{f} - f\|_2 + \|f - f_{L,P}^*\|_2. \quad (17)$$

Clearly, (15) shows that the third term is bounded by  $\varepsilon$ . Let us now consider the second term. Here we note that for an  $x \in X$  with  $P_X(A(x)) > 0$  we have

$$\begin{aligned} |f(x) - \bar{f}(x)| &= \frac{1}{P_X(A(x))} \left| \int_{A(x)} f(x) - f(x') dP_x(x') \right| \\ &\leq \frac{1}{P_X(A(x))} \int_{A(x)} |f(x) - f(x')| dP_x(x') \\ &\leq \varepsilon, \end{aligned}$$

where in the last step we used (16). Consequently, we obtain

$$\|f - \bar{f}\|_2^2 = \sum_{j \in J: P_X(A_j) > 0} \int_{A(x)} |f(x) - \bar{f}(x)|^2 dP_x(x') \leq \sum_{j \in J: P_X(A_j) > 0} \varepsilon^2 \cdot P_X(A_j) \leq \varepsilon^2.$$

In other words, the second term is bounded by  $\varepsilon$ , too. Let us finally consider the first term. Here

we have

$$\begin{aligned}
\|h_{P,\mathcal{A}} - \bar{f}\|_2^2 &= \sum_{j \in J: P_X(A_j) > 0} \int_{A_j} |h_{P,\mathcal{A}} - \bar{f}|^2 dP_X \\
&= \sum_{j \in J: P_X(A_j) > 0} \int_{A_j} \left| \frac{1}{P_X(A(x))} \int_{A_j} f_{L,P}^* dP_X - \frac{1}{P_X(A_j)} \int_{A(x)} f dP_X \right|^2 dP_X \\
&= \sum_{j \in J: P_X(A_j) > 0} \left| \int_{A_j} f_{L,P}^* dP_X - \int_{A_j} f dP_X \right|^2 \\
&\leq \left| \sum_{j \in J: P_X(A_j) > 0} \int_{A_j} f_{L,P}^* dP_X - \int_{A_j} f dP_X \right|^2 \\
&= \|f_{L,P}^* - f\|_1^2 \\
&\leq \|f_{L,P}^* - f\|_2^2 \\
&\leq \varepsilon^2.
\end{aligned}$$

Consequently, the first term is bounded by  $\varepsilon$ , too, and hence we conclude by (17) that

$$\mathcal{R}_{L,P}(h_{P,\mathcal{A}}) - \mathcal{R}_{L,P}^* = \|h_{P,\mathcal{A}} - f_{L,P}^*\|_2^2 \leq 9\varepsilon^2.$$

A simple variable transformation then yields the first assertion.

To show the second assertion we simply consider  $f = f_{L,P}^*$  and note that we can choose  $s_\varepsilon = (\varepsilon/C)^{1/\alpha}$ . Repeating the proof of the first case for *arbitrary*  $\varepsilon > 0$  and  $s \in (0, s_\varepsilon]$  then yields

$$\|h_{P,\mathcal{A}} - f_{L,P}^*\|_2 \leq 2\varepsilon.$$

Now let  $s \in (0, 1]$ . For  $\varepsilon := Cs^\alpha$  we then obtain the assertion.  $\square$

Based on the oracle inequality Proposition A.1 and the Approximation error bound Lemma A.3 we can now establish universal consistency of any histogram rule  $h_{D,\mathcal{A}_D}$  based on a data-dependent partition  $\mathcal{A}_D$  from  $\mathcal{P}_{s,n}^X$ , growing at most polynomially with the sample size.

**Proposition A.4** (Consistency). *Assume  $s_n \rightarrow 0$  and  $\frac{\log(n)}{ns_n^d} \rightarrow 0$  as  $n \rightarrow \infty$ . Further suppose that  $|\mathcal{P}_{s_n,n}^X| = K_{s_n,n} \leq cn^\beta$ , for some  $c < \infty$ ,  $0 < \beta$ . The learning method  $\mathcal{L} : D \mapsto h_{D,\mathcal{A}_D}$  is universally consistent, that is,*

$$\mathcal{R}_{L,P}(h_{D,\mathcal{A}_D}) \rightarrow \mathcal{R}_{L,P}^*$$

in probability for  $|D| \rightarrow \infty$ .

**Proof of Proposition A.4:** Let  $\varepsilon > 0$  and  $\tau > 0$ . Applying the oracle inequality in Proposition A.1 gives with  $P^n$ -probability at least  $1 - 2K_{s,n}e^{-\tau}$

$$\begin{aligned}
\mathcal{R}_{L,P}(h_{D,\mathcal{A}_D}) - \mathcal{R}_{L,P}^* &\leq \sup_{\mathcal{A} \in \mathcal{P}_{s,n}^X} \mathcal{R}_{L,P}(h_{D,\mathcal{A}}) - \mathcal{R}_{L,P}^* \\
&\leq 6 \sup_{\mathcal{A} \in \mathcal{P}_{s,n}^X} (\mathcal{R}_{L,P,\mathcal{H}_\mathcal{A}}^* - \mathcal{R}_{L,P}^*) + 8\varepsilon \\
&\quad + 128 \sup_{\mathcal{A} \in \mathcal{P}_{s,n}^X} \frac{\tau + 1 + \log \mathcal{N}(\mathcal{H}_\mathcal{A}, \|\cdot\|_\infty, \varepsilon)}{n}
\end{aligned} \tag{18}$$

Next, Lemma A.3 gives for any CP  $\mathcal{A}$  of width  $s \in (0, s_\varepsilon]$

$$\mathcal{R}_{L,P,\mathcal{H}_\mathcal{A}}^* - \mathcal{R}_{L,P}^* = \mathcal{R}_{L,P}(h_{P,\mathcal{A}}) - \mathcal{R}_{L,P}(f_{L,P}^*) < \varepsilon. \tag{19}$$

Further, the covering number is bounded by

$$\log \mathcal{N}(\mathcal{H}_\mathcal{A}, \|\cdot\|_\infty, \varepsilon) \leq c_d s^{-d} \log(1/\varepsilon), \tag{20}$$

for some  $c_d < \infty$ . Thus, combining this bound with (19) and with (18) yields with  $P^n$ -probability at least  $1 - 2K_{s,n}e^{-\tau}$

$$\mathcal{R}_{L,P}(h_{D,\mathcal{A}_D}) - \mathcal{R}_{L,P}^* \leq 14\varepsilon + 128 \frac{\tau + 1 + c_d s^{-d} \log(1/\varepsilon)}{n}. \quad (21)$$

Finally, the result follows by choosing  $\tau = \sqrt{n}$  and  $\varepsilon = 1/\log(n)$ .  $\square$

Finally, we derive learning rates.

**Proposition A.5** (Learning Rates). *Suppose the regularity Assumption A.2 holds. Let the width  $s_n$  be chosen according to*

$$s_n = \left( \frac{\log(n)}{n} \right)^{\frac{1}{2\alpha+d}}.$$

If  $|\mathcal{P}_{s_n,n}^X| = K_{s_n,n} \leq cn^\beta$ , for some  $c < \infty$ ,  $0 \leq \beta$ , then

$$\mathcal{R}_{L,P}(h_{D,\mathcal{A}_D}) - \mathcal{R}_{L,P}^* \leq c_{d,\alpha} \left( \frac{\log(n)}{n} \right)^{\frac{2\alpha}{2\alpha+d}}$$

with  $P^n$ -probability at least  $1 - cn^\beta e^{-n^\gamma}$ , with  $\gamma = \frac{d}{2\alpha+d}$ .

**Proof of Proposition A.5:** Under the regularity Assumption A.2, Lemma A.3 gives us

$$\begin{aligned} \mathcal{R}_{L,P,\mathcal{H}_\mathcal{A}}^* - \mathcal{R}_{L,P}^* &= \|h_{P,\mathcal{A}} - f_{L,P}^*\|_{L^2}^2 \\ &\leq \|h_{P,\mathcal{A}} - f_{L,P}^*\|_{L^\infty}^2 \\ &< Cs^{2\alpha}. \end{aligned}$$

Plugging this bound into (18) and using (20) once more leads us to

$$\mathcal{R}_{L,P}(h_{D,\mathcal{A}_D}) - \mathcal{R}_{L,P}^* \leq Cs^{2\alpha} + 8\varepsilon + 128 \frac{\tau + 1 + c_d s^{-d} \log(1/\varepsilon)}{n}, \quad (22)$$

with  $P^n$ -probability at least  $1 - 2K_{s,n}e^{-\tau}$ , for any  $\tau > 0$  and  $\varepsilon > 0$ . Finally, choosing

$$s_n = \left( \frac{\log(n)}{n} \right)^{\frac{1}{2\alpha+d}}, \quad \varepsilon_n = n^{-\frac{2\alpha}{2\alpha+d}}, \quad \tau_n = n^{\frac{d}{2\alpha+d}}$$

gives the result.  $\square$

**Remark A.6.** The proof of Proposition A.5 shows actually more: If we let  $\gamma \in [0, \frac{2\alpha}{2\alpha+d}]$  and choose in (22)

$$\tau_n = n^{1-\gamma}, \quad \varepsilon_n = n^{-\gamma}, \quad s_n = \left( \frac{\log(n)}{n} \right)^{\frac{1-\gamma}{d}}$$

we get

$$\mathcal{R}_{L,P}(h_{D,\mathcal{A}_D}) - \mathcal{R}_{L,P}^* \leq c \left( \frac{\log(n)}{n} \right)^\gamma$$

with  $P^n$ -probability at least  $1 - 2K_{s,n}e^{-\tau_n}$ , for some  $c < \infty$ . This rate is optimal in the mini-max sense only if  $\gamma = \frac{2\alpha}{2\alpha+d}$ . In other words: If the width of the random cubic partition is chosen inappropriately, the method  $\mathcal{L} : D \mapsto h_{D,\mathcal{A}_D}$  can learn arbitrarily bad.

## A.2 Classification

**Proposition A.7** (Consistency). *Let  $L$  be the classification loss and  $\text{sign } h_{D, \mathcal{A}_D}$  be a plug-in classification rule, where  $h_{D, \mathcal{A}_D}$  is an HRR based on a random partition. Assume  $s_n \rightarrow 0$  and  $\frac{\log(n)}{ns_n^d} \rightarrow 0$  as  $n \rightarrow \infty$ . Further suppose that  $|\mathcal{P}_{s_n, n}^X| = K_{s_n, n} \leq cn^\beta$ , for some  $c < \infty$ ,  $0 < \beta$ . The learning method  $\mathcal{L} : D \mapsto h_{D, \mathcal{A}_D}$  is universally consistent, that is,*

$$\mathcal{R}_{L, P}(\text{sign } h_{D, \mathcal{A}_D}) \rightarrow \mathcal{R}_{L, P}^*$$

in probability for  $|D| \rightarrow \infty$ .

**Proof of Proposition A.7:** The assertion follows by applying a well-known calibration inequality between the classification and the least squares loss, namely

$$\mathcal{R}_{L_{\text{class}}, P}(f) - \mathcal{R}_{L_{\text{class}}, P}^* \leq \sqrt{\mathcal{R}_{L_{\text{ls}}, P}(f) - \mathcal{R}_{L_{\text{ls}}, P}^*} = \|f - f_{L_{\text{ls}}, P}^*\|_2,$$

see e.g. [12, Table 3.1 and Theorem 3.22] for the inequality and [12, Example 2.6] for the identity. Then apply Proposition A.4.  $\square$

## B General Aspects of Histograms and other Lego Pieces

In this section we collect some useful Lemmas which we shall need for proving our main Theorem 2.2. The first Lemma provides a simple characterization of ERMs.

**Lemma B.1** (Characterization of ERMs). *Let  $A \subseteq X$  be non-empty,  $\mathcal{A} := (A_1, \dots, A_m)$  be a partition of  $A$ , and*

$$\mathcal{F}_{\mathcal{A}} := \left\{ \sum_{j=1}^m \alpha_j \mathbf{1}_{A_j} : \alpha_j \in Y \right\}$$

with  $Y = \mathbb{R}$  or  $Y = \{-1, +1\}$ . Moreover, let  $D = ((x_1, y_1), \dots, (x_n, y_n)) \in (X \times Y)^n$  be a data set and let  $L_A(x, y, t) = \mathbf{1}_A(x)L(y, t)$ , where  $L$  is either the classification loss or least squares loss. Further denote by  $N_j = \sum_{i: x_i \in A_j} \mathbf{1}_{A_j}(x_i)$  the number of covariates falling into  $A_j$ . Then, for every  $f^* \in \mathcal{F}_{\mathcal{A}}$  with representation  $f^* = \sum_{j=1}^m \alpha_j \mathbf{1}_{A_j}$ , the following statements are equivalent:

i) The function  $f^*$  is an empirical risk minimizer, that is

$$\mathcal{R}_{L_A, D}(f^*) = \min_{f \in \mathcal{F}_{\mathcal{A}}} \mathcal{R}_{L_A, D}(f).$$

ii) • Let  $Y = \mathbb{R}$ . For all  $j \in \{1, \dots, m\}$  satisfying  $N_j \neq 0$  and  $\sum_{i: x_i \in A_j} y_i \neq 0$  we have

$$\alpha_j = \frac{1}{N_j} \sum_{i: x_i \in A_j} y_i. \quad (23)$$

If  $\sum_{i: x_i \in A_j} y_i = 0$ , then  $\alpha_j = 0$ . If  $N_j = 0$ , then any  $\alpha_j \in Y$  is a minimizer.

• Let  $Y = \{-1, +1\}$ . For all  $j \in \{1, \dots, m\}$  satisfying  $\sum_{i: x_i \in A_j} y_i \neq 0$  we have

$$\alpha_j = \text{sign}\left(\frac{1}{N_j} \sum_{i: x_i \in A_j} y_i\right) = \text{sign}\left(\sum_{i: x_i \in A_j} y_i\right). \quad (24)$$

If  $\sum_{i: x_i \in A_j} y_i = 0$ , both  $\alpha_j = -1$  and  $\alpha_j = 1$  are minimizers.

**Proof of Lemma B.1:** We first note that for an  $f^* \in \mathcal{F}_A$  with representation  $f^* = \sum_{j=1}^m \alpha_j \mathbf{1}_{A_j}$  we have

$$\mathcal{R}_{L_A, D}(f^*) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_A(x_i) L(y_i, f^*(x_i)) = \frac{1}{n} \sum_{j=1}^m \sum_{i: x_i \in A_j} L(y_i, \alpha_j).$$

Consequently,  $f^*$  is an empirical risk minimizer, if and only if  $\alpha_j$  minimizes  $\sum_{i: x_i \in A_j} L(y_i, \cdot)$  for all  $j \in \{1, \dots, m\}$ .

Now let  $Y = \mathbb{R}$  and  $L_A(x, y, t) = \mathbf{1}_A(x) L(y, t)$ , with  $L$  the least square loss. Then

$$\sum_{i: x_i \in A_j} L(y_i, \alpha_j) = \left( \sum_{i: x_i \in A_j} \mathbf{1}_{A_j}(x_i) \right) \alpha_j^2 - 2\alpha_j \left( \sum_{i: x_i \in A_j} y_i \right) + y_i^2$$

which is minimized for  $\alpha_j = \frac{\sum_{i: x_i \in A_j} y_i}{\sum_{i: x_i \in A_j} \mathbf{1}_{A_j}(x_i)}$ .

If  $Y = \{-1, +1\}$  and  $L_A(x, y, t) = \mathbf{1}_A(x) L(y, t)$  with  $L$  the classification loss, then (24) is the only minimizer, whereas, in the case  $\sum_{i: x_i \in A_j} y_i = 0$ , both  $\alpha_j = -1$  and  $\alpha_j = 1$  minimize the term. This completes the proof.  $\square$

The next Lemma provides a bound on the difference of the risks of two measurable functions with respect to both, the classification loss and the least squares loss.

**Lemma B.2.** *Let  $f_1, f_2 : X \rightarrow Y$  be measurable functions and let  $A \subset X$  be measurable and non-empty.*

i) *If  $Y = \{-1, +1\}$  and  $L_A(x, y, t) = \mathbf{1}_A(x) L(y, t)$  with  $L$  the classification loss, then*

$$|\mathcal{R}_{L_A, P}(f_1) - \mathcal{R}_{L_A, P}(f_2)| \leq P_X(A \cap \{f_1 \neq f_2\}).$$

ii) *If  $Y = [-1, 1]$  and  $L_A(x, y, t) = \mathbf{1}_A(x) L(y, t)$  with  $L$  the least square loss, then*

$$|\mathcal{R}_{L_A, P}(f_1) - \mathcal{R}_{L_A, P}(f_2)| \leq M P_X(A \cap \{f_1 \neq f_2\}),$$

where  $M = M_1 + M_2$  with

$$M_j = \sup_{x \in X} |f_{L, P}^*(x) - f_j(x)|^2, \quad j = 1, 2.$$

**Proof of Lemma B.2:**

i) By definition, we have

$$\begin{aligned} |\mathcal{R}_{L, P}(f_1) - \mathcal{R}_{L, P}(f_2)| &= \left| \int_{A \times Y} \mathbf{1}_{(-\infty, 0]}(yf_1(x)) - \mathbf{1}_{(-\infty, 0]}(yf_2(x)) dP(x, y) \right| \\ &\leq \int_{A \times Y} |\mathbf{1}_{(-\infty, 0]}(yf_1(x)) - \mathbf{1}_{(-\infty, 0]}(yf_2(x))| dP(x, y) \\ &\leq \int_{A \times Y} \mathbf{1}_{\{f_1 \neq f_2\}} dP(x, y), \end{aligned}$$

and this yields the assertion.

ii) Again by definition, we have

$$\mathcal{R}_{L, P}(f_1) - \mathcal{R}_{L, P}(f_2) = \int_A \int_Y (y - f_1(x))^2 - (y - f_2(x))^2 dP(y|x) dP_X(x).$$



Using  $(a-b)^2 - (a-c)^2 = (b-c)((b-a) - (c-a))$  and integrating w.r.t.  $y$  yields for the absolute value of the rhs of the above equation

$$\begin{aligned}
& \left| \int_A (f_1(x) - f_2(x))((f_1(x) - y) + (f_2(x) - y)) dP(y|x) dP_X(x) \right| \\
&= \left| \int_A (f_1(x) - f_2(x))((f_1(x) - f_{L,P}^*(x)) + (f_2(x) - f_{L,P}^*(x))) dP_X(x) \right| \\
&= \left| \int_{A \cap \{f_1 \neq f_2\}} (f_{L,P}^*(x) - f_1(x))^2 - (f_{L,P}^*(x) - f_2(x))^2 dP_X(x) \right| \\
&\leq (M_1 + M_2) P_X(A \cap \{f_1 \neq f_2\}) ,
\end{aligned}$$

with

$$M_j = \sup_{x \in X} |f_{L,P}^*(x) - f_j(x)|^2, \quad j = 1, 2 .$$

□

**Lemma B.3.** *Let  $A_1, A_2 \subset X$  be non-empty, disjoint, and measurable with  $A_1 \cup A_2 = X$ . Let  $L : Y \times \mathbb{R} \rightarrow \mathbb{R}$  be loss and define  $L_{A_j}(x, y, t) = \mathbf{1}_{A_j}(x) L(y, t)$ , for  $j = 1, 2$ . Then we have*

$$\mathcal{R}_{L,P}^* = \mathcal{R}_{L_{A_1},P}^* + \mathcal{R}_{L_{A_2},P}^* .$$

**Proof of Lemma B.3:** See e.g. [9], Lemma 9. □

**Lemma B.4** (Label Flipping). *Let  $Y = \{-1, 1\}$ , and  $P$  be a distribution on  $X \times Y$ . Moreover, let  $P_\varphi$  denote the pushforward measure of the label flipping map  $\varphi : X \times Y \rightarrow X \times Y$  given by*

$$\varphi(x, y) := (x, -y) .$$

*Then for all  $f : X \rightarrow \mathbb{R}$  we have*

$$|\mathcal{R}_{L_{\text{class}},P}(f) - (1 - \mathcal{R}_{L_{\text{class}},P}^*)| \leq \|f - f_{L_{\text{class}},P_\varphi}^*\|_2 .$$

**Proof of Lemma B.4:** As usual, we write  $\eta(x) := P(y = 1|x)$ . Obviously, this gives  $\eta_\varphi = 1 - \eta$ , where  $\eta_\varphi := P_\varphi(y = 1|x)$ . By the last equation in the proof of [12, Theorem 2.31] we then find

$$\begin{aligned}
\mathcal{R}_{L_{\text{class}},P_\varphi}(f) - \mathcal{R}_{L_{\text{class}},P_\varphi}^* &= \int_X |2\eta_\varphi - 1| \cdot \mathbf{1}_{(-\infty, 0]}((2\eta_\varphi - 1) \text{sign } f) dP_X \\
&= \int_X |2\eta - 1| \cdot \mathbf{1}_{(-\infty, 0]}((1 - 2\eta) \text{sign } f) dP_X \\
&= \int_X |2\eta - 1| dP_X - \int_X |2\eta - 1| \cdot \mathbf{1}_{(0, \infty)}((1 - 2\eta) \text{sign } f) dP_X \\
&= \int_X |2\eta - 1| dP_X - \int_X |2\eta - 1| \cdot \mathbf{1}_{(-\infty, 0]}((2\eta - 1) \text{sign } f) dP_X \\
&= \int_X |2\eta - 1| dP_X - \int_X |2\eta - 1| \cdot \mathbf{1}_{(-\infty, 0]}((2\eta - 1) \text{sign } f) dP_X \\
&= \int_X |2\eta - 1| dP_X - (\mathcal{R}_{L_{\text{class}},P}(f) - \mathcal{R}_{L_{\text{class}},P}^*) ,
\end{aligned}$$

where in the second to last step we used  $|2\eta - 1| \cdot \mathbf{1}_{\{0\}}((2\eta - 1) \text{sign } f) = 0$ , and the last step is another application of the last equation in the proof of [12, Theorem 2.31]. Now observe that we have

$$|2\eta - 1| + \min\{\eta, 1 - \eta\} = 1 - \min\{\eta, 1 - \eta\}$$

and since  $\mathcal{R}_{L_{\text{class}}, P}^* = \mathbb{E}_{P_X} \min\{\eta, 1 - \eta\}$ , see e.g. [12, Example 2.4] we find

$$(1 - \mathcal{R}_{L_{\text{class}}, P}^*) - \mathcal{R}_{L_{\text{class}}, P}(f) = \mathcal{R}_{L_{\text{class}}, P_\varphi}(f) - \mathcal{R}_{L_{\text{class}}, P_\varphi}^*.$$

Now the assertion follows by applying a well-known calibration inequality between the classification and the least squares loss, namely

$$\mathcal{R}_{L_{\text{class}}, P_\varphi}(f) - \mathcal{R}_{L_{\text{class}}, P_\varphi}^* \leq \sqrt{\mathcal{R}_{L_{\text{ls}}, P_\varphi}(f) - \mathcal{R}_{L_{\text{ls}}, P_\varphi}^*} = \|f - f_{L_{\text{ls}}, P_\varphi}^*\|_2,$$

see e.g. [12, Table 3.1 and Theorem 3.22] for the inequality and [12, Example 2.6] for the identity.  $\square$

## C Proof of Main Theorem 2.2

### C.1 Proof of Main Theorem 2.2

The first Lemma in this section gives a useful bound of the excess risk of inflated histograms in terms of related classical histograms. Handling classical histograms can be done with the results from Section A. This is the main step for proving Theorem 2.2.

**Lemma C.1.** *Let  $L$  be the least squares, the hinge or the classification loss. For  $s \in (0, 1]$ ,  $r > 0$  and  $m \geq 0$  let  $f^* \in \mathcal{F}_{s, r, m}^*$  be an (interpolating) predictor having representation (11), with  $h_{\mathcal{A}}$  being its  $\mathcal{H}_{\mathcal{A}}$ -part. If Assumption 2.1 is satisfied, then the excess risk satisfies*

$$\mathcal{R}_{L, P}(f^*) - \mathcal{R}_{L, P}^* \leq \mathcal{R}_{L, P}(h_{\mathcal{A}}) - \mathcal{R}_{L, P}^* + mMcr, \quad (25)$$

for some  $M \in \mathbb{R}_+$ , depending on the loss function and where  $c \in (0, \infty)$  is from Assumption 2.1.

**Proof of Proposition C.1:** We split the excess risk as

$$\mathcal{R}_{L, P}(f^*) - \mathcal{R}_{L, P}^* \leq (\mathcal{R}_{L, P}(f^*) - \mathcal{R}_{L, P}(h_{\mathcal{A}})) + (\mathcal{R}_{L, P}(h_{\mathcal{A}}) - \mathcal{R}_{L, P}^*). \quad (26)$$

By definition, for  $t \in [0, r]$ , one has

$$\{f^* \neq h_{\mathcal{A}}\} = \bigcup_{i=1}^m x_i^* + tB_\infty. \quad (27)$$

Applying Lemma B.2 and taking Assumption 2.1 into account, we find almost surely

$$\mathcal{R}_{L, P}(f^*) - \mathcal{R}_{L, P}(h_{\mathcal{A}}) \leq M \sum_{i=1}^m P_X(x_i^* + tB_\infty) \leq mMcr, \quad (28)$$

for some  $M \in \mathbb{R}_+$ .  $\square$

We now prove our main result.

**Proof of Theorem 2.2:** Choose a good interpolating histogram rule  $f_{D, s}^+ \in \mathcal{F}_{s, r, n}^*$  as in Example 3.3 and a bad interpolating histogram rule  $f_{D, s}^- \in \mathcal{F}_{s, r, n}^*$  according to Example 3.3.

- i) Let  $L$  be the least squares, the hinge or the classification loss. Recall that Theorem 3.1 defines a partitioning rule  $\pi_{s, n} : \mathbb{R}^{dn} \rightarrow \mathcal{P}_s$ , where  $|\mathcal{P}_{s, n}| = (n+1)^d$ . The claim in Eq. (2) follows from Proposition A.4, Proposition A.7 and by applying Lemma C.1. More precisely, (25) gives us with probability at least  $1 - 2(n+1)^d e^{-\sqrt{n}}$

$$\mathcal{R}_{L, P}(f_{D, s_n}^+) - \mathcal{R}_{L, P}^* \leq \mathcal{R}_{L, P}(h_{D, s_n}^+) - \mathcal{R}_{L, P}^* + Mnr_n \leq \epsilon$$

provided  $s_n$  satisfies the assumptions of Theorem 2.2,  $r_n = 2^{-n}$  and  $n$  is sufficiently large.

Next, consider a bad interpolating histogram rule  $f_{D,s}^- \in \mathcal{F}_{s,r,n}^*$ . We first consider the case of  $L$  being the least squares loss and have owing to Lemma C.1

$$\begin{aligned} \mathcal{R}_{L,P}(f_{D,s}^-) - \mathcal{R}_{L,P}^\dagger &= \mathcal{R}_{L,P}(f_{D,s}^-) - \mathcal{R}_{L,P}(h_{D,s}^-) \\ &\quad + \mathcal{R}_{L,P}(h_{D,s}^-) - \mathcal{R}_{L,P}(-f_{L,P}^*) \\ &\leq nMc r + \|h_{D,s}^- - f_{L,P}^*\|_2^2 \\ &= nMc r + \|h_{D,s}^+ - f_{L,P}^*\|_2^2. \end{aligned} \quad (29)$$

Now, the analysis of  $D \mapsto h_{D,s_n}^+$  in Proposition A.4 ensures  $\|h_{D,s_n}^+ - f_{L,P}^*\|_2 \rightarrow 0$  for  $n \rightarrow \infty$ , and thus (29) immediately shows that

$$\mathcal{R}_{L,P}(f_{D,s_n}^-) \longrightarrow \mathcal{R}_{L,P}^\dagger$$

in probability for  $n \rightarrow \infty$ , if additionally  $\epsilon_n = r_n = 2^{-n}$ .

A similar observation can be made for the classification loss, since by Lemma B.4 and

$$\mathcal{R}_{L_{\text{class}},P}(\text{sign } f) = \mathcal{R}_{L_{\text{class}},P}(f)$$

we have

$$\begin{aligned} |\mathcal{R}_{L_{\text{class}},P}(\text{sign } h_{D,s}^-) - (1 - \mathcal{R}_{L_{\text{class}},P}^*)| &\leq \|h_{D,s}^- - f_{L_{\text{class}},P}^*\|_2 \\ &= \|h_{D,s}^+ - f_{L_{\text{class}},P}^*\|_2, \end{aligned} \quad (30)$$

where  $h_{D,s}^\pm$  denote the  $\mathcal{H}_{\mathcal{A}}$ -part of the good and bad interpolating HRs for regression, and  $\text{sign } h_{D,s}^-$  is the  $\mathcal{H}_{\mathcal{A}}$ -part of the bad interpolating HR for classification. Finally, for the hinge loss,  $f_{D,s}^-$  coincides with the corresponding function for the classification loss.

- ii) Let  $L$  be the least squares loss,  $f_{L,P}^*$  be  $\alpha$ -Hölder continuous and suppose  $P_X$  satisfies Assumption 2.1. Choose  $s_n$  as in Theorem 2.2 and  $r_n = 2^{-n}$ . From Lemma C.1 and Proposition A.5 with  $|\mathcal{P}_{s_n,n}^X| = (n+1)^d$  we obtain for  $n$  sufficiently large

$$\begin{aligned} \mathcal{R}_{L,P}(f_{D,s_n}^+) - \mathcal{R}_{L,P}^* &\leq \mathcal{R}_{L,P}(h_{D,s}^+) - \mathcal{R}_{L,P}^* + Mcnr_n \\ &\leq c'_{M,d,\alpha} \left( \frac{\log(n)}{n} \right)^{\frac{2\alpha}{2\alpha+d}} \end{aligned}$$

for some  $c'_{M,d,\alpha} < \infty$  and with  $P^n$ -probability at least  $1 - n^d e^{-n^\beta}$ , with  $\beta = \frac{d}{2\alpha+d}$ . Finally, the rates for the full range  $\gamma \in [0, \frac{2\alpha}{2\alpha+d}]$  follow by using the arguments as in Remark A.6.

□

## C.2 Auxiliary Technical Proofs

**Proof of Theorem 3.1:** Let us write  $\delta := s/(m+1)$ . For  $j \in \{0, \dots, m\}$  we then define

$$z_j^\dagger := \left(j + \frac{1}{2}\right) \delta.$$

Moreover, our candidate offsets  $x_1^\dagger, \dots, x_K^\dagger \in \mathbb{R}^d$  are exactly those vectors whose coordinates equal  $z_j^\dagger$  for some  $j \in \{0, \dots, m\}$ . Clearly, this gives  $K = (m+1)^d$ . Now let  $x_1^*, \dots, x_m^* \in [-1, 1]^d$ . In the following, we will identify the offset  $x_\ell^\dagger$  coordinate-wise. We begin by determining its first coordinate  $x_{\ell,1}^\dagger$ . To this end, we define

$$I_j := \bigcup_{k \in \mathbb{Z}} [ks + j\delta, ks + (j+1)\delta].$$

Our first goal is to show that  $I_0, \dots, I_m$  are a partition of  $\mathbb{R}$ . To this end, we fix an  $x \in \mathbb{R}$ . Then there exists a unique  $k \in \mathbb{Z}$  with  $ks \leq x < (k+1)s$ . Moreover, for  $y := x - ks \in [0, s)$ , there exists a unique  $j \in \{0, \dots, m\}$  with  $j\delta \leq y < (j+1)\delta$ . Consequently, we have found  $x \in [ks + j\delta, ks + (j+1)\delta)$ . This shows  $\mathbb{R} \subset I_0, \dots, I_m$ , and the converse inclusion is trivial. Let us now fix an  $x \in I_j \cap I_{j'}$ . Then there exist  $k, k' \in \mathbb{Z}$  such that

$$x \in [ks + j\delta, ks + (j+1)\delta) \cap [k's + j'\delta, k's + (j'+1)\delta) \quad (31)$$

Since  $(j+1)\delta \leq s$  and  $(j'+1)\delta \leq s$  we conclude that  $ks \leq x < (k+1)s$  and  $k's \leq x < (k'+1)s$ . As observed above this implies  $k = k'$ . Now consider  $y := x - ks \in [0, s)$ . Then (31) implies

$$y \in [j\delta, (j+1)\delta) \cap [j'\delta, (j'+1)\delta),$$

and again we have seen above that this implies  $j = j'$ . This shows  $I_j \cap I_{j'} = \emptyset$  for all  $j \neq j'$ .

Let us now denote the first coordinate of  $x_i^*$  by  $x_{i,1}^*$ . Then  $D_{X,1}^* := \{x_{i,1}^* : i = 1, \dots, m\}$  satisfies  $|D_{X,1}^*| \leq m$  and since we have  $m+1$  cells  $I_j$ , we conclude that there exists a  $j_1^* \in \{0, \dots, m\}$  with  $D_{X,1}^* \cap I_{j_1^*} = \emptyset$ . We define

$$x_{\ell,1}^\dagger := z_{j_1^*}^\dagger = \left(j_1^* + \frac{1}{2}\right) \delta.$$

Next we repeat this construction for the remaining  $d-1$  coordinates, so that we finally obtain  $x_\ell^\dagger := (z_{j_1^*}^\dagger, \dots, z_{j_d^*}^\dagger) \in \mathbb{R}^d$  for indices  $j_1^*, \dots, j_d^* \in \{0, \dots, m\}$  found by the above reasoning.

It remains to show that (12) holds the CP (10) with offset  $x_\ell^\dagger$  and all  $t > 0$  with  $t \leq \frac{s}{3m+3} = \delta/3$ . To this end, we fix an  $x_i^*$ . Then its cell  $B(x_i^*)$  is described by a unique  $k := (k_1, \dots, k_d) \in \mathbb{Z}^d$ , namely

$$B(x_i^*) = [x_{\ell,1}^\dagger + k_1s, x_{\ell,1}^\dagger + (k_1+1)s) \times \dots \times [x_{\ell,d}^\dagger + k_ds, x_{\ell,d}^\dagger + (k_d+1)s).$$

Let us now consider the first coordinate  $x_{i,1}^*$ . By construction we then know that  $x_{i,1}^* \notin I_{j_1^*}$  and

$$\left(j_1^* + \frac{1}{2}\right) \cdot \delta + k_1s \leq x_{i,1}^* < \left(j_1^* + \frac{1}{2}\right) \cdot \delta + (k_1+1)s. \quad (32)$$

Now  $x_{i,1}^* \notin I_{j_1^*}$  implies

$$x_{i,1}^* \notin [(k_1+1)s + j_1^*\delta, (k_1+1)s + (j_1^*+1)\delta)$$

Since the right hand side of (32) excludes the case  $x_{i,1}^* \geq (k_1+1)s + (j_1^*+1)\delta$ , hence we find

$$x_{i,1}^* < (k_1+1)s + j_1^*\delta = x_{\ell,1}^\dagger + (k_1+1)s - \delta/2$$

This shows  $x_{i,1}^* + r < x_{\ell,1}^\dagger + (k_1+1)s$  for all  $r \in [-t, t]$ . Analogously, we can show  $x_{i,1}^* + r > x_{\ell,1}^\dagger + k_1s$  for all  $r \in [-t, t]$ . By repeating these considerations for the remaining  $d-1$ -coordinates, we conclude that  $x_i^* + tB_\infty \subset B(x_i^*)$ .  $\square$

**Proof of Proposition 3.2:** By our assumptions we have

$$c_i^* := b_i + c_{j_i} \in \arg \min_{c \in Y} \sum_{k: x_k = x_i^*} L(x_k, y_k, c) = \arg \min_{c \in \mathbb{R}} \sum_{k: x_k = x_i^*} L(x_k, y_k, c),$$

where the last equality is a consequence of the fact that there is an  $f : X \rightarrow Y$  satisfying (6). Moreover, since (12) and (13) hold, we find  $f^*(x_i^*) = h(x_i^*) + b_i = c_{j_i} + b_i = c_i^*$ , and therefore  $f^*$  interpolates  $D$  by (6).  $\square$

## D General Aspects of Neural Nets and other Lego Pieces

### D.1 DNN Algebra

This Section is devoted to showing how networks of different sizes can be combined to build new ones.

Given an activation function  $\sigma : \mathbb{R} \rightarrow \mathbb{R}$  and  $\mathbf{b} \in \mathbb{R}^m$  we define the *shifted activation function*  $\sigma_{\mathbf{b}} : \mathbb{R}^m \rightarrow \mathbb{R}^m$  as

$$\sigma_{\mathbf{b}}(y) := (\sigma(y_1 + b_1), \dots, \sigma(y_m + b_m))^T.$$

A *hidden layer* of width  $m$  and with input dimension  $M$  is a function  $H : \mathbb{R}^M \rightarrow \mathbb{R}^m$  of the form

$$H(x) = (\sigma_{\mathbf{b}} \circ A)(x), \quad x \in \mathbb{R}^M,$$

where  $A$  is a  $M \times m$  *weight matrix* and  $\mathbf{b} \in \mathbb{R}^m$  is a *shift vector* or *bias*. Clearly, each pair  $(A, \mathbf{b})$  describes a layer, but in general, a layer, if viewed as a *function*, can be described by more than one such pair. A network architecture is described by a *width vector*  $\mathbf{m} = (m_0, \dots, m_L) \in \mathbb{N}^{L+1}$ , where the positive integer  $L$  is the number of *layers*,  $L - 1$  is the number of *hidden layers* or the *depth*. Here,  $m_0$  is the input dimension and  $m_L$  is the output dimension. A neural network with architecture  $\mathbf{m} \in \mathbb{N}^{L+1}$  is a function  $f : \mathbb{R}^{m_0} \rightarrow \mathbb{R}^{m_L}$ , having a representation of the form

$$\begin{aligned} f(\mathbf{x}) &= H_L \circ H_{L-1} \circ \dots \circ H_1(x) \\ &= A^{(L)} \sigma_{\mathbf{b}^{(L-1)}} A^{(L-1)} \sigma_{\mathbf{b}^{(L-2)}} \dots A^{(2)} \sigma_{\mathbf{b}^{(1)}} A^{(1)}(x) + \mathbf{b}^{(L)}, \end{aligned} \quad (33)$$

where  $A^{(l)}$  is a  $m_l \times m_{l-1}$  *weight matrix* and  $\mathbf{b}^{(l)} \in \mathbb{R}^{m_l}$  is a *shift vector* or *bias*, associated to layer  $l = 1, \dots, L$ . Moreover, if the layers  $H_1, \dots, H_{L-1}, H_L$  are represented by the pairs

$$(A^{(1)}, \mathbf{b}^{(1)}), \dots, (A^{(L)}, \mathbf{b}^{(L)})$$

then we say that the network is represented by  $(\mathfrak{A}, \mathfrak{B})$ , where  $\mathfrak{A} := (A^{(1)}, \dots, A^{(L)})$  and  $\mathfrak{B} := (\mathbf{b}^{(1)}, \dots, \mathbf{b}^{(L)})$ . As for layers, each pair  $(\mathfrak{A}, \mathfrak{B})$  determines a neural network, but in general, a neural network, if viewed as a function, can be described by more than one such pair. We denote the set of all neural networks of architecture  $(m_0, m_1, \dots, m_L)$  by  $\mathcal{A}_{m_0, (m_1, \dots, m_{L-1}), m_L}$ .

In the sequel, we confine ourselves to ReLU-activation functions  $|\cdot|_+ : \mathbb{R} \rightarrow [0, \infty)$  defined by  $|t|_+ := \max\{0, t\}$ . Moreover, we consider networks where the last layer is a single neuron without activation function, that is,

$$H_L(x) = \langle a, x \rangle + b, \quad x \in \mathbb{R}^{m_{L-1}}$$

**Lemma D.1.** *i) Let  $d \geq 1$  and  $m \in \mathbb{N}^{L-1}$ . Assume  $f \in \mathcal{A}_{d, m, 1}$  has representation  $\mathfrak{A} := (A^{(1)}, \dots, A^{(L-1)}, a)$  and  $\mathfrak{B} := (B^{(1)}, \dots, B^{(L-1)}, b)$ . Then for  $\alpha \in \mathbb{R}$  and  $c \in \mathbb{R}$ , the function  $\alpha f + c \in \mathcal{A}_{d, m, 1}$  has representation*

$$(A^{(1)}, \dots, A^{(L-1)}, \alpha a) \quad (B^{(1)}, \dots, B^{(L-1)}, \alpha b + c)$$

*and  $|f|_+ \in \mathcal{A}_{d, (m, 1), 1}$  has representation*

$$(A^{(1)}, \dots, A^{(L-1)}, a, 1) \quad (B^{(1)}, \dots, B^{(L-1)}, b, 0).$$

*ii) Let  $d \geq 1$ ,  $L \geq 2$ , and  $\tilde{m}, \hat{m} \in \mathbb{N}^{L-1}$ . Then for all  $f \in \mathcal{A}_{d, \tilde{m}, 1}$  and  $g \in \mathcal{A}_{d, \hat{m}, 1}$  we have*

$$f + g \in \mathcal{A}_{d, \tilde{m} + \hat{m}, 1}.$$

*More precisely, if  $(\tilde{\mathfrak{A}}, \tilde{\mathfrak{B}})$  ( $\hat{\mathfrak{A}}, \hat{\mathfrak{B}}$ ) are representations of  $f$  and  $g$ , then  $\mathfrak{A} := (A^{(1)}, \dots, A^{(L-1)}, a)$  and  $\mathfrak{B} := (B^{(1)}, \dots, B^{(L-1)}, b)$  defined by*

$$A^{(1)} := \begin{pmatrix} \tilde{A}^{(1)} \\ \hat{A}^{(1)} \end{pmatrix} \in \mathbb{R}^{(\tilde{m}_1 + \hat{m}_1) \times d}, \quad b^{(1)} := \begin{pmatrix} \tilde{b}^{(1)} \\ \hat{b}^{(1)} \end{pmatrix} \in \mathbb{R}^{\tilde{m}_1 + \hat{m}_1}$$

*and for  $l = 2, \dots, L - 1$*

$$A^{(l)} := \begin{pmatrix} \tilde{A}^{(l)} & 0 \\ 0 & \hat{A}^{(l)} \end{pmatrix} \in \mathbb{R}^{(\tilde{m}_l + \hat{m}_l) \times (\tilde{m}_{l-1} + \hat{m}_{l-1})}, \quad b^{(l)} := \begin{pmatrix} \tilde{b}^{(l)} \\ \hat{b}^{(l)} \end{pmatrix} \in \mathbb{R}^{\tilde{m}_l + \hat{m}_l},$$

$$a := \begin{pmatrix} \tilde{a} & \hat{a} \end{pmatrix} \in \mathbb{R}^{\tilde{m}_L + \hat{m}_L}, \quad b := \tilde{b} + \hat{b} \in \mathbb{R},$$

*gives a representation  $(\mathfrak{A}, \mathfrak{B})$  of  $f + g$ .*

**Proof of Lemma D.1:**

i) The first assertion immediately follows from representation (33) while for the second note that building the positive part is nothing else than applying a shifted ReLU activation function with weight equal to 1 and zero bias.

ii) Let

$$f(x) = \tilde{H}_L \circ \dots \circ \tilde{H}_1(x), \quad x \in \mathbb{R}^d$$

and

$$g(x) = \hat{H}_L \circ \dots \circ \hat{H}_1(x), \quad x \in \mathbb{R}^d.$$

For  $K = 1, \dots, L$  we introduce the concatenation of layers

$$\tilde{W}_K(x) := \tilde{H}_K \circ \dots \circ \tilde{H}_1(x), \quad x \in \mathbb{R}^d$$

and similarly  $\hat{W}_K$ . Then, since the last layer is just a single neuron without activation function given by

$$\tilde{H}_L(x) = \langle \tilde{a}, x \rangle + \tilde{b}, \quad x \in \mathbb{R}^{\tilde{m}_{L-1}}$$

and

$$\hat{H}_L(x) = \langle \hat{a}, x \rangle + \hat{b}, \quad x \in \mathbb{R}^{\hat{m}_{L-1}},$$

we immediately obtain

$$\begin{aligned} (f + g)(x) &= \langle \tilde{a}, \tilde{W}_{L-1}(x) \rangle + \tilde{b} + \langle \hat{a}, \hat{W}_{L-1}(x) \rangle + \hat{b} \\ &= \left\langle \begin{pmatrix} \tilde{a} \\ \hat{a} \end{pmatrix}, \begin{pmatrix} \tilde{W}_{L-1}(x) \\ \hat{W}_{L-1}(x) \end{pmatrix} \right\rangle + \tilde{b} + \hat{b} \end{aligned}$$

and thus

$$a = \begin{pmatrix} \tilde{a} & \hat{a} \end{pmatrix} \in \mathbb{R}^{\tilde{m}_L + \hat{m}_L} \quad b = \tilde{b} + \hat{b} \in \mathbb{R}.$$

Moreover, for any  $l = 2, \dots, L - 1$  we have

$$\begin{aligned} \begin{pmatrix} \tilde{W}_l(x) \\ \hat{W}_l(x) \end{pmatrix} &= \begin{pmatrix} \tilde{H}_l \circ \tilde{W}_{l-1}(x) \\ \hat{H}_l \circ \hat{W}_{l-1}(x) \end{pmatrix} \\ &= \left| \begin{pmatrix} \tilde{A}^{(l)} & 0 \\ 0 & \hat{A}^{(l)} \end{pmatrix} \begin{pmatrix} \tilde{W}_{l-1}(x) \\ \hat{W}_{l-1}(x) \end{pmatrix} + \begin{pmatrix} \tilde{b}^{(l)} \\ \hat{b}^{(l)} \end{pmatrix} \right|_+, \end{aligned}$$

with

$$\tilde{A}^{(l)} \in \mathbb{R}^{\tilde{m}_l \times \tilde{m}_{l-1}}, \quad \hat{A}^{(l)} \in \mathbb{R}^{\hat{m}_l \times \hat{m}_{l-1}}$$

and

$$\tilde{b}^{(l)} \in \mathbb{R}^{\tilde{m}_l}, \quad \hat{b}^{(l)} \in \mathbb{R}^{\hat{m}_l}$$

and where we apply  $|\cdot|_+$  on each component. Finally, the representation for  $A^{(1)}$  and  $b^{(1)}$  follows again from simple algebra.  $\square$

## D.2 Neuron Lego

In this section we collect the main pieces to approximate histograms with DNNs. The first Lemma is a longer and more detailed version of Lemma 4.1 and shows how to approximate an indicator function on a multidimensional interval by a ReLU-DNN with 2 hidden layers.

**Lemma D.2.** Let  $z_1 = (z_{1,1}, \dots, z_{1,d}) \in \mathbb{R}^d$  and  $z_2 = (z_{2,1}, \dots, z_{2,d}) \in \mathbb{R}^d$  with  $z_1 < z_2$ , and let  $\varepsilon > 0$  satisfy

$$\varepsilon < \min \left\{ \frac{z_{2,i} - z_{1,i}}{2} : i = 1, \dots, d \right\}.$$

Moreover, for  $i = 1, \dots, d$  and  $j = 1, 2$  let  $h_i^{(j)} : \mathbb{R} \rightarrow [0, \infty)$  be the neurons with weights  $a_i^{(j)} \in \mathbb{R}^d$  and biases  $b_i^{(j)} \in \mathbb{R}$  given by

$$\begin{aligned} a_i^{(1)} &= -\frac{1}{\varepsilon} e_i, & b_i^{(1)} &= \frac{z_{1,i} + \varepsilon}{\varepsilon}, \\ a_i^{(2)} &= \frac{1}{\varepsilon} e_i, & b_i^{(2)} &= -\frac{z_{2,i} - \varepsilon}{\varepsilon}. \end{aligned}$$

Let  $H_\varepsilon : \mathbb{R}^d \rightarrow \mathbb{R}^{2d}$  be the hidden layer defined by

$$H_\varepsilon(x) := \left( h_1^{(1)}(x), \dots, h_d^{(1)}(x), h_1^{(2)}(x), \dots, h_d^{(2)}(x) \right), \quad x \in \mathbb{R}^d,$$

and in addition, let  $h : \mathbb{R}^{2d} \rightarrow \mathbb{R}$  be the neuron with weight  $a \in \mathbb{R}^{2d}$  and bias  $b$  given by

$$a = (-1, -1, \dots, -1), \quad b = 1.$$

Then  $f_\varepsilon : \mathbb{R}^d \rightarrow [0, \infty)$  defined by  $f_\varepsilon := h \circ H_\varepsilon$  is continuous and we have

$$\{f_\varepsilon > 0\} \subset (z_1, z_2), \quad \{f_\varepsilon = 1\} = [z_1 + \varepsilon, z_2 - \varepsilon], \quad \text{and} \quad \{f_\varepsilon > 1\} = \emptyset.$$

In particular,  $f_\varepsilon \in \mathcal{A}_{d,(2d,1),1}$ .

**Proof of Lemma D.2:** The specific form of  $h$  shows that

$$\begin{aligned} f_\varepsilon(x) &= h \circ H = \left| \sum_{i=1}^d a_i h_i^{(1)}(x) + \sum_{i=1}^d a_{d+i} h_i^{(2)}(x) + b \right|_+ \\ &= \left| -\sum_{i=1}^d h_i^{(1)}(x) - \sum_{i=1}^d h_i^{(2)}(x) + 1 \right|_+ \\ &= \left| \sum_{i=1}^d (1 - h_i^{(1)}(x) - h_i^{(2)}(x)) - d + 1 \right|_+, \end{aligned} \tag{34}$$

and hence we first investigate the functions  $1 - h_i^{(1)} - h_i^{(2)}$ . To this end, let us fix an  $i \in \{1, \dots, d\}$  and an  $x = (x_1, \dots, x_d) \in \mathbb{R}^d$ . Then we obviously have

$$h_i^{(1)}(x) = |\langle a_i^{(1)}, x \rangle + b_i^{(1)}|_+ = \left| -\frac{x_i}{\varepsilon} + \frac{z_{1,i} + \varepsilon}{\varepsilon} \right|_+ = \begin{cases} \frac{-x_i + z_{1,i} + \varepsilon}{\varepsilon} & \text{if } x_i \leq z_{1,i} + \varepsilon \\ 0 & \text{else,} \end{cases}$$

and

$$h_i^{(2)}(x) = |\langle a_i^{(2)}, x \rangle + b_i^{(2)}|_+ = \left| \frac{x_i}{\varepsilon} - \frac{z_{2,i} - \varepsilon}{\varepsilon} \right|_+ = \begin{cases} \frac{x_i - z_{2,i} + \varepsilon}{\varepsilon} & \text{if } x_i \geq z_{2,i} - \varepsilon \\ 0 & \text{else.} \end{cases}$$

Since  $z_{1,i} + \varepsilon < z_{2,i} - \varepsilon$ , we consequently find

$$1 - h_i^{(1)}(x) - h_i^{(2)}(x) = \begin{cases} \frac{x_i - z_{1,i}}{\varepsilon} & \text{if } x_i \leq z_{1,i} + \varepsilon \\ 1 & \text{if } x_i \in [z_{1,i} + \varepsilon, z_{2,i} - \varepsilon] \\ -\frac{x_i - z_{2,i}}{\varepsilon} & \text{if } x_i \geq z_{2,i} - \varepsilon. \end{cases}$$

In particular, we have

$$\{1 - h_i^{(1)} - h_i^{(2)} > 0\} = \{(x_1, \dots, x_d) \in \mathbb{R}^d : x_i \in (z_{1,i}, z_{2,i})\}, \quad (35)$$

$$\{1 - h_i^{(1)} - h_i^{(2)} = 1\} = \{(x_1, \dots, x_d) \in \mathbb{R}^d : x_i \in [z_{1,i} + \varepsilon, z_{2,i} - \varepsilon]\}, \quad (36)$$

$$\{1 - h_i^{(1)} - h_i^{(2)} > 1\} = \emptyset. \quad (37)$$

Combining our initial equation (34) with (36) and (37) yields

$$\begin{aligned} \{f_\varepsilon = 1\} &= \left\{ \left| \sum_{i=1}^d (1 - h_i^{(1)} - h_i^{(2)}) - d + 1 \right|_+ = 1 \right\} = \left\{ \sum_{i=1}^d (1 - h_i^{(1)} - h_i^{(2)}) = d \right\} \\ &= [z_1 + \varepsilon, z_2 - \varepsilon]. \end{aligned}$$

and a combination of (34) with (35) gives

$$\begin{aligned} \{f_\varepsilon > 0\} &= \left\{ \left| \sum_{i=1}^d (1 - h_i^{(1)} - h_i^{(2)}) - d + 1 \right|_+ > 0 \right\} = \left\{ \sum_{i=1}^d (1 - h_i^{(1)} - h_i^{(2)}) > d - 1 \right\} \\ &\subset \bigcap_{i=1}^d \{1 - h_i^{(1)} - h_i^{(2)} > 0\} \\ &= (z_1, z_2), \end{aligned}$$

where for the proof of the inclusion assume that it was not true. Then there would be an  $x \in \mathbb{R}^d$  and an  $i_0 \in \{1, \dots, d\}$  with

$$\sum_{i=1}^d (1 - h_i^{(1)}(x) - h_i^{(2)}(x)) > d - 1 \quad \text{and} \quad 1 - h_{i_0}^{(1)}(x) - h_{i_0}^{(2)}(x) \leq 0.$$

without loss of generality we may assume that  $i_0 = d$ . Then combining both inequalities we find

$$d - 1 < \sum_{i=1}^{d-1} (1 - h_i^{(1)}(x) - h_i^{(2)}(x)) + (1 - h_d^{(1)}(x) - h_d^{(2)}(x)) \leq \sum_{i=1}^{d-1} (1 - h_i^{(1)}(x) - h_i^{(2)}(x))$$

and this shows that there would be an  $i \in \{1, \dots, d - 1\}$  with  $1 - h_i^{(1)}(x) - h_i^{(2)}(x) > 1$ . This contradicts (37).

Finally, the equation  $\{f_\varepsilon > 1\} = \emptyset$  immediately follows from combining (34) and (37). The continuity of  $f_\varepsilon$  is obvious.  $\square$

As a second step in our construction presented in Section 4 we use Lemma D.2 and combine that with Lemma D.1 to approximate step-functions (i.e., histograms based on cubic partitions) with ReLU-DNNs with 2 layers.

**Proposition D.3.** *Let  $A_1, \dots, A_k$  be mutually disjoint subsets of  $X := [-1, 1]^d$  such that for each  $i \in \{1, \dots, k\}$  there exist  $z_i^-, z_i^+ \in X$  with  $z_i^- < z_i^+$  and  $(z_i^-, z_i^+) \subset A_i \subset [z_i^-, z_i^+]$ . Moreover, let  $z_{i,j}^\pm$  be the  $j$ -th coordinate of  $z_i^\pm$ . Then for all  $f : X \rightarrow \mathbb{R}$  of the form*

$$f = \sum_{i=1}^k \alpha_i \mathbf{1}_{A_i} \quad (38)$$

where  $\alpha_i \in \mathbb{R}$ , all  $\varepsilon > 0$  satisfying

$$\varepsilon < \min \left\{ \frac{z_{i,j}^+ - z_{i,j}^-}{2} : i = 1, \dots, k \text{ and } j = 1, \dots, d \right\}$$

and all  $m_1 \geq 2dk$  and  $m_2 \geq k$ , there exists a neural network  $g_\varepsilon$  of architecture  $(d, m_1, m_2, 1)$  such that

$$\{f = g_\varepsilon\} = \bigcup_{i=1}^k [z_i^- + \varepsilon, z_i^+ - \varepsilon] \cup (X \setminus (z_i^-, z_i^+)).$$



**Proof of Proposition D.3:** By assumption, for any  $\varepsilon > 0$  we have for any  $i = 1, \dots, k$  the inclusion

$$[z_i^- + \varepsilon, z_i^+ - \varepsilon] \subset A_i \subset [z_i^-, z_i^+].$$

For each pair  $z_i^-, z_i^+ \in X$  Lemma 4.1 gives us a function  $g_i^{(\varepsilon)} = h_i \circ H_i \in \mathcal{A}_{d,(2d,1),1}$  such that

$$\{g_i^{(\varepsilon)} > 0\} \subset (z_i^-, z_i^+), \quad \{g_i^{(\varepsilon)} = 1\} = [z_i^- + \varepsilon, z_i^+ - \varepsilon] \quad \text{and} \quad \{g_i^{(\varepsilon)} > 1\} = \emptyset.$$

In particular

$$\{g_i^{(\varepsilon)} = \mathbf{1}_{A_i}\} = A_i \cap \{g_i^{(\varepsilon)} = 1\} \cap \{g_i^{(\varepsilon)} = 0\} = [z_i^- + \varepsilon, z_i^+ - \varepsilon] \cup X \setminus (z_i^-, z_i^+).$$

Moreover, for any  $\alpha_i \in \mathbb{R}$  Lemma D.1 ensures  $\alpha_i g_i^{(\varepsilon)} \in \mathcal{A}_{d,(2d,1),1}$  with

$$\{\alpha_i g_i^{(\varepsilon)} = \alpha_i \mathbf{1}_{A_i}\} = [z_i^- + \varepsilon, z_i^+ - \varepsilon] \cup X \setminus (z_i^-, z_i^+).$$

Finally, applying Lemma D.1 once more shows that

$$g_\varepsilon := \sum_{i=1}^k \alpha_i g_i^{(\varepsilon)}$$

belongs to  $\mathcal{A}_{d,(2kd,k),1}$  and satisfies

$$\{f = g_\varepsilon\} = \bigcup_{i=1}^k [z_i^- + \varepsilon, z_i^+ - \varepsilon] \cup (X \setminus (z_i^-, z_i^+)).$$

□

## E Proof of Main Theorem 2.3

The first Lemma provides a bound for the excess risk of approximations of inflated histograms in terms of the excess risk of the corresponding classical histograms.

**Lemma E.1.** *Let  $L$  be the least squares, the hinge or the classification loss. Let  $\varepsilon, r, s > 0$  and  $m \geq 0$  and  $f^{(\varepsilon)} \in \mathcal{F}_{s,r,m}^{(\varepsilon)}$  be a DNN having representation*

$$f^{(\varepsilon)} = h_{\mathcal{A}}^{(\varepsilon)} + \sum_{i=1}^m b_i g_{x_i^* + tB_\infty}^{(\delta)}, \quad \delta \leq t/2,$$

*with  $h_{\mathcal{A}}^{(\varepsilon)}$  being its  $\varepsilon$ -approximation  $\mathcal{H}_{\mathcal{A}}$ -part and  $h_{\mathcal{A}}$  its exact  $\mathcal{H}_{\mathcal{A}}$ -part. If Assumption 2.1 is satisfied, then the excess risk satisfies*

$$\mathcal{R}_{L,P}(f^{(\varepsilon)}) - \mathcal{R}_{L,P}^* \leq \mathcal{R}_{L,P}(h_{\mathcal{A}}) - \mathcal{R}_{L,P}^* + 2mM\tilde{c}(\varepsilon + r), \quad (39)$$

*for some  $M \in \mathbb{R}_+$ , depending on the loss function and where  $\tilde{c} \in (0, \infty)$  is from Assumption 2.1.*

**Proof of Proposition E.1:** We split the excess risk as

$$\begin{aligned} \mathcal{R}_{L,P}(f^{(\varepsilon)}) - \mathcal{R}_{L,P}^* &\leq \left( \mathcal{R}_{L,P}(f^{(\varepsilon)}) - \mathcal{R}_{L,P}(h_{\mathcal{A}}^{(\varepsilon)}) \right) + \left( \mathcal{R}_{L,P}(h_{\mathcal{A}}^{(\varepsilon)}) - \mathcal{R}_{L,P}(h_{\mathcal{A}}) \right) \\ &\quad + \left( \mathcal{R}_{L,P}(h_{\mathcal{A}}) - \mathcal{R}_{L,P}^* \right). \end{aligned} \quad (40)$$

By construction, for  $t \in [0, r]$ , one has

$$\{f^{(\varepsilon)} \neq h_{\mathcal{A}}^{(\varepsilon)}\} = \bigcup_{i=1}^m x_i^* + (t + \delta)B_\infty$$

and thus Lemma B.2 together with Assumption 2.1 gives

$$\mathcal{R}_{L,P}(f^{(\varepsilon)}) - \mathcal{R}_{L,P}(h_{\mathcal{A}}^{(\varepsilon)}) \leq M \sum_{i=1}^m P_X(x_i^* + (t + \delta)B_\infty) \leq 2mMc r, \quad (41)$$

for some  $M \in \mathbb{R}_+$  depending on the loss function. Furthermore, by the same token and with Proposition D.3 we find

$$\mathcal{R}_{L,P}(h_{\mathcal{A}}^{(\varepsilon)}) - \mathcal{R}_{L,P}(h_{\mathcal{A}}) \leq Mc' \varepsilon, \quad (42)$$

for some  $c' \in (0, \infty)$ . Collecting (42), (41) and (40) finishes the proof by setting  $\tilde{c} = c + c'$ .  $\square$

**Proof of Theorem 2.3:** The proof follows closely the lines of the proof of Theorem 2.2. Choose a good interpolating DNN  $g_{D,s}^+ \in \mathcal{F}_{s,r,n}^{(\varepsilon)}$  and a bad interpolating DNN  $g_{D,s}^- \in \mathcal{F}_{s,r,n}^{(\varepsilon)}$  as in Example 4.2.

- i) Let  $L$  be the least squares or the hinge loss. Recall that Theorem 3.1 defines a partitioning rule  $\pi_{s,n} : \mathbb{R}^{dn} \rightarrow \mathcal{P}_s$ , where  $|\mathcal{P}_{s,n}| = (n+1)^d$ . The claim in Eq. (2) follows from Proposition A.4, Proposition A.7 and by applying Lemma E.1. More precisely, for any  $\varepsilon' > 0$ , (39) gives us with probability at least  $1 - 2(n+1)^d e^{-\sqrt{n}}$

$$\mathcal{R}_{L,P}(g_{D,s}^+) - \mathcal{R}_{L,P}^* \leq \mathcal{R}_{L,P}(h_{D,s}^+) - \mathcal{R}_{L,P}^* + 2M\tilde{c}n(\varepsilon + r) \leq \varepsilon'$$

provided  $s = s_n$  satisfies the assumptions of Theorem 2.2,  $\varepsilon = \varepsilon_n = 2^{-n}$ ,  $r = r_n = 2^{-n}$  and  $n$  is sufficiently large.

Next, consider a bad interpolating DNN  $g_{D,s}^- \in \mathcal{F}_{s,r,n}^{(\varepsilon)}$ . We first consider the case of  $L$  being the least squares loss and have owing to Lemma E.1

$$\begin{aligned} \mathcal{R}_{L,P}(g_{D,s}^-) - \mathcal{R}_{L,P}^* &\leq 2M\tilde{c}n(\varepsilon + r) + \|h_{D,s}^- - f_{L,P}^*\|_2^2 \\ &= 2M\tilde{c}n(\varepsilon + r) + \|h_{D,s}^+ - f_{L,P}^*\|_2^2. \end{aligned} \quad (43)$$

Now, the analysis of  $D \mapsto h_{D,s_n}^+$  in Proposition A.4 ensures  $\|h_{D,s_n}^+ - f_{L,P}^*\|_2 \rightarrow 0$  for  $n \rightarrow \infty$ , and thus (29) immediately shows that

$$\mathcal{R}_{L,P}(g_{D,s}^-) \rightarrow \mathcal{R}_{L,P}^*$$

in probability for  $n \rightarrow \infty$ , if additionally  $\varepsilon_n = r_n = 2^{-n}$ . The reasoning for the classification loss and hinge loss is the same as in the proof of Theorem 2.2.

- ii) Let  $L$  be the least squares loss,  $f_{L,P}^*$  be  $\alpha$ -Hölder continuous and suppose  $P_X$  satisfies Assumption 2.1. Choose  $s_n, \varepsilon_n$  as in Theorem 2.2 and  $r_n = 2^{-n}$ . From Lemma E.1 and Proposition A.5 with  $|\mathcal{P}_{s_n,n}^X| = (n+1)^d$  we obtain for  $n$  sufficiently large

$$\begin{aligned} \mathcal{R}_{L,P}(g_{D,s_n}^+) - \mathcal{R}_{L,P}^* &\leq \mathcal{R}_{L,P}(h_{D,s_n}^+) - \mathcal{R}_{L,P}^* + 2M\tilde{c}n(\varepsilon_n + r_n) \\ &\leq c'_{M,d,\alpha} \left( \frac{\log(n)}{n} \right)^{\frac{2\alpha}{2\alpha+d}} \end{aligned}$$

for some  $c'_{M,d,\alpha} < \infty$  and with  $P^n$ -probability at least  $1 - n^d e^{-n^\gamma}$ , with  $\gamma = \frac{d}{2\alpha+d}$ . Finally, the rates for the full range  $\gamma \in [0, \frac{2\alpha}{2\alpha+d}]$  follow by using the arguments as in Remark A.6.  $\square$