Journal of Machine Learning Research 3 (2002) 175–212 Submitted 12/01; Published 9/02

Algorithmic Luckiness

Ralf Herbrich RHERB@MICROSOFT.COM
Microsoft Research
7 J J Thomson Avenue
CB3 0FB Cambridge
United Kingdom

Robert C. Williamson BOB.WILLIAMSON@NICTA.EDU.AU
National ICT Australia
Research School of Information Sciences and Engineering
Australian National University
Canberra, ACT 0200, Australia

Editor: David A. Cohn

Abstract

Classical statistical learning theory studies the generalisation performance of machine learning algorithms rather indirectly. One of the main detours is that algorithms are studied in terms of the hypothesis class that they draw their hypotheses from. In this paper, motivated by the luckiness framework of Shawe-Taylor et al. (1998), we study learning algorithms more directly and in a way that allows us to exploit the serendipity of the training sample. The main difference to previous approaches lies in the complexity measure; rather than covering all hypotheses in a given hypothesis space it is only necessary to cover the functions which could have been learned using the fixed learning algorithm. We show how the resulting framework relates to the VC, luckiness and compression frameworks. Finally, we present an application of this framework to the maximum margin algorithm for linear classifiers which results in a bound that exploits the margin, the sparsity of the resultant weight vector, and the degree of clustering of the training data in feature space.

1. Introduction

Statistical learning theory has been mainly concerned with the study of *uniform* bounds on the expected error of hypotheses from a given hypothesis space (Vapnik, 1998; Anthony and Bartlett, 1999). Such bounds have the appealing feature that they provide performance guarantees for classifiers found by *any* learning algorithm. However, it has been observed that these bounds tend to be overly pessimistic. One explanation is that *only* in the case of learning algorithms which minimise the training error it has been proven that uniformity of the bounds is equivalent to studying the learning algorithm's generalisation performance directly, and this equivalence only holds in an asymptotic sense. Thus it is not surprising that such analysis tools, which analyse algorithms mainly in terms of the class of functions from which they draw their hypotheses, are somewhat loose.

In this paper we present a theoretical framework for *directly* studying the generalisation error of a learning algorithm rather than taking the detour via the uniform convergence

of training errors to expected errors in a given hypothesis space. In addition, our new model of learning allows the exploitation of the situation that we serendipitously observe a training sample which is easy to learn by a given learning algorithm. Thus our framework is consanguineous to the luckiness framework of Shawe-Taylor et al. (1998). In this paper, the luckiness is a function of a given learning algorithm and a given training sample and characterises the diversity of the algorithm's solutions. This notion of luckiness enables the study of given learning algorithms from many different perspectives. For example, the maximum margin algorithm (Vapnik, 1998) can be studied via the number of dimensions in feature space, the margin of the classifier learned or the sparsity of the resulting classifier. Our main results are two generalisation error bounds for learning algorithms: one for when the training error is zero and one agnostic bound (Section 5). We shall demonstrate the utility of our new framework by studying its relation to the VC framework, the original luckiness framework, the compression framework of Littlestone and Warmuth (1986), and algorithmic stability (Section 6). Finally, we present an application of the new framework to the maximum margin algorithm for linear classifiers (Section 7).

1.1 Notation and Background

Let $\mathcal{X}$ and $\mathcal{Y}$ be sets and let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. We denote vectors using bold face, e.g. $\boldsymbol{x} = (x_1, \ldots, x_m)$ and the length of this vector by $|\boldsymbol{x}|$, i.e. $|\boldsymbol{x}| = m$. The shorthand notation $x \in \boldsymbol{x}$ means $\exists i \in \{1, \ldots, |\boldsymbol{x}|\} : x = x_i$. For natural numbers i and j, $i \leq j$, $[i : j] := (i, i+1, \ldots, j)$. Given a vector $\boldsymbol{z}$ and an index vector $\boldsymbol{i} = (i_1, \ldots, i_k) \in \{1, \ldots, |\boldsymbol{z}|\}^k$, define $\boldsymbol{z}_{\boldsymbol{i}} := (z_{i_1}, \ldots, z_{i_k})$. The symbols $\mathbf{P}_{\mathsf{X}}$, $\mathbf{E}_{\mathsf{X}}[f(\mathsf{X})]$ and $\mathbb{I}$ denote a probability measure over X, the expectation of $f(\cdot)$ over the random draw of its argument x and the indicator function, respectively. For a vector valued function $\mathbf{f}$, we define $\mathbf{var}_{\mathsf{X}}(\mathbf{f}(\mathsf{X})) := \mathbf{E}_{\mathsf{X}}[\|\mathbf{f}(\mathsf{X}) - \boldsymbol{\mu}\|^2]$ where $\boldsymbol{\mu} = \mathbf{E}_{\mathsf{X}}[\mathbf{f}(\mathsf{X})]$. The notation $\mathbf{P}_{\mathsf{X}^m}$ denotes the product measure $\mathbf{P}_{\mathsf{X}_1} \cdots \mathbf{P}_{\mathsf{X}_m}$. The symbols $\mathbb{R}$ and $\mathbb{N}$ denote the real and natural numbers, respectively. The shorthand notation $\mathcal{Z}^{(\infty)} := \cup_{m=1}^{\infty} \mathcal{Z}^m$ denotes the union of all m–fold Cartesian products of the set $\mathcal{Z}$. If $f : \mathcal{X} \to \mathcal{Y}$ is a function, $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ and $\boldsymbol{x} \in \mathcal{X}^m$ then $f_{|\boldsymbol{x}} := (f(x_1), \ldots, f(x_m))$ and $\mathcal{F}_{|\boldsymbol{x}} := \left\{ f_{|\boldsymbol{x}} \mid f \in \mathcal{F} \right\}$ denotes the evaluation of f at $x_1, \ldots, x_m$ and its extension to the set $\mathcal{F}$ respectively. Moreover, if $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ and $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$ then

$$\mathcal{L}_l(\mathcal{H}) := \{(x, y) \mapsto l(h(x), y) \mid h \in \mathcal{H}\} .$$

Since we often consider permutations we introduce the following additional notation: For any $m \in \mathbb{N}$ we define $I_m \subset \{1, \ldots, m\}^m$ as the set of all permutations of the numbers $1, \ldots, m$,

$$I_m := \{(i_1, \ldots, i_m) \in \{1, \ldots, m\}^m \mid \forall j \neq k : \; i_j \neq i_k\} .$$

Note that $|I_m| = m!$. Given a $2m$–vector $\boldsymbol{i} \in I_{2m}$ and a m-vector $\boldsymbol{s} \in \{0, 1\}^m$ we define $\pi_{\boldsymbol{i}} : \{1, \ldots, 2m\} \to \{1, \ldots, 2m\}$ and the swapping permutation $\sigma_{\boldsymbol{s}} : \{1, \ldots, 2m\} \to \{1, \ldots, 2m\}$ by

$$\forall j \in \{1, \ldots, 2m\} : \qquad \pi_{\boldsymbol{i}}(j) := i_j , \quad \sigma_{\boldsymbol{s}}(j) := j + m \cdot (s_j \mathbb{I}_{j \leq m} - s_{j-m} \mathbb{I}_{j > m}) ,$$

that is $\sigma_{\boldsymbol{s}}$ swaps i and $i + m$ if and only if $s_i = 1$. Given a sample $\boldsymbol{z} \in \mathcal{Z}^{2m}$ we denote the action of $\pi_{\boldsymbol{i}}$ and $\sigma_{\boldsymbol{s}}$ on the indices of $\boldsymbol{z}$ by $\Pi_{\boldsymbol{i}}(\boldsymbol{z})$ and $\Sigma_{\boldsymbol{s}}(\boldsymbol{z})$, i.e. $\Pi_{\boldsymbol{i}}(\boldsymbol{z}) := \left(z_{\pi_{\boldsymbol{i}}(1)}, \ldots, z_{\pi_{\boldsymbol{i}}(2m)}\right)$ and $\Sigma_{\boldsymbol{s}}(\boldsymbol{z}) := \left(z_{\sigma_{\boldsymbol{s}}(1)}, \ldots, z_{\sigma_{\boldsymbol{s}}(2m)}\right)$.

Given a set $B \subseteq A$ and a metric $\rho : A \times A \to \mathbb{R}^+$ the *covering number* $\mathcal{N}(\varepsilon, B, \rho)$ is defined as the size of the smallest subset $C \subseteq A$ such that for all $b \in B$ there exists a $c \in C$ such that $\rho(b,c) < \varepsilon$. Such a set C is called a *cover* of the set B at the scale ε. If $A = \mathbb{R}^n$ then we define the ℓ_1^n–metric ρ_1^n by

$$\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n : \qquad \rho_1^n(\boldsymbol{x}, \boldsymbol{y}) := \frac{1}{n} \sum_{i=1}^{n} |x_i - y_i| .$$

If $\boldsymbol{x} \in \mathcal{X}^n$ and f and g are functions from $\mathcal{X}$ to $\mathbb{R}$ then $\rho_1^{\boldsymbol{x}} : \mathbb{R}^{\mathcal{X}} \times \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^+$ is defined by

$$\rho_1^{\boldsymbol{x}}(f,g) := \rho_1^n\left(f_{|\boldsymbol{x}}, g_{|\boldsymbol{x}}\right) .$$

Finally, the ℓ_1^n–norm $\|\cdot\|_1$ is defined by $\|\boldsymbol{x}\|_1 := \sum_{i=1}^n |x_i|$. Note that $\|\boldsymbol{x} - \boldsymbol{y}\|_1 = n \cdot \rho_1^n(\boldsymbol{x}, \boldsymbol{y})$.

2. The Learning Model

Suppose we are given a training sample $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{y}) \in (\mathcal{X} \times \mathcal{Y})^m$ of size $m \in \mathbb{N}$ drawn iid from some unknown distribution $\mathbf{P}_{\mathsf{XY}} = \mathbf{P}_{\mathsf{Z}}$. Suppose furthermore we are given a learning algorithm $\mathcal{A} : \mathcal{Z}^{(\infty)} \to \mathcal{Y}^{\mathcal{X}}$ that maps a given training sample $\boldsymbol{z} \in \mathcal{Z}^{(\infty)}$ to a function from $\mathcal{X}$ to $\mathcal{Y}$, often called a hypothesis. Then we would like to investigate the *generalisation error* of the algorithm.

Definition 1 (Generalisation, prediction and training error) *Given a loss $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$ and a hypothesis $h \in \mathcal{Y}^{\mathcal{X}}$ the* prediction error $R_l[h]$ *of h is defined by*

$$R_l[h] := \mathbf{E}_{\mathsf{XY}}\left[l\left(h(\mathsf{X}), \mathsf{Y}\right)\right] .$$

Given a training sample $\boldsymbol{z} \in \mathcal{Z}^{(\infty)}$, the training error $\widehat{R}_l[h, \boldsymbol{z}]$ *is defined as the empirical counterpart of the prediction error, i.e.*

$$\widehat{R}_l[h, \boldsymbol{z}] := \frac{1}{|\boldsymbol{z}|} \sum_{(x_i, y_i) \in \boldsymbol{z}} l\left(h(x_i), y_i\right) .$$

For any learning algorithm $\mathcal{A} : \mathcal{Z}^{(\infty)} \to \mathcal{Y}^{\mathcal{X}}$, the generalisation error $G_l[\mathcal{A}, \boldsymbol{z}]$ *of $\mathcal{A}$ is defined as the prediction error of $\mathcal{A}(\boldsymbol{z})$,*

$$G_l[\mathcal{A}, \boldsymbol{z}] := R_l[\mathcal{A}(\boldsymbol{z})] .$$

Note that $G_l[\mathcal{A}\, \boldsymbol{z}]$ can never be smaller than the quantity $\inf_{h \in \mathcal{Y}^{\mathcal{X}}} R_l[h]$ which is also known as the *Bayes error.*

Definition 2 (Generalisation error bound) *A function $\varepsilon : \mathcal{Z}^{(\infty)} \times (0,1] \to \mathbb{R}^+$ is called a* generalisation error bound[1] *for $\mathcal{A}$ if and only if*

$$\forall \mathbf{P}_{\mathsf{Z}} : \forall \delta \in (0,1] : \qquad \mathbf{P}_{\mathsf{Z}^m}\left(G_l(\mathcal{A}, \mathbf{Z}) \leq \varepsilon(\mathbf{Z}, \delta)\right) \geq 1 - \delta .$$

1. In classical statistics, such a function is closely related to a confidence interval for the estimator $\mathcal{A}$. In fact, the interval $[0, \varepsilon(\boldsymbol{z}, \delta)]$ is a confidence interval for the difference between the prediction error of $\mathcal{A}(\boldsymbol{z})$ and the smallest prediction error possible at level at least $1 - \delta$.

It is worth noticing that any probabilistic bound on $R_l\left[\mathcal{A}\left(\boldsymbol{z}\right)\right]$ can readily be transformed into a generalisation error bound because the Bayes error does not depend on the training sample $\boldsymbol{z}$. Hence, our ultimate interest is in the predictive performance $R_l\left[\mathcal{A}\left(\boldsymbol{z}\right)\right]$ of $\mathcal{A}\left(\boldsymbol{z}\right) \in \mathcal{Y}^{\mathcal{X}}$. The classical approach taken by many researchers (see, e.g. Vapnik, 1982, 1995; Blumer et al., 1989; Shawe-Taylor et al., 1998) is to derive uniform bounds over some a-priori restricted subset $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ of hypotheses often called *hypothesis space*, i.e. proving an upper bound on the prediction error which holds uniformly for all hypotheses $h \in \mathcal{H}$ (or all consistent hypotheses) we automatically obtain a bound for $\mathcal{A}\left(\boldsymbol{z}\right)$ because $\mathcal{A}\left(\boldsymbol{z}\right) \in \mathcal{H}$ by definition. Clearly, this is much too strong a requirement and leads to bounds which are independent of the algorithm used.

3. A General Recipe for Generalisation Error Bounds

Before presenting our new framework let us consider the general steps typically used to obtain bounds on the generalisation error. The five steps classically used in statistical learning theory are as follows:

1. First, we relate the prediction error $R_l\left[h\right]$ of a given hypothesis $h \in \mathcal{Y}^{\mathcal{X}}$ to some empirical quantity such as the training error $\widehat{R}_l\left[h, \boldsymbol{z}\right]$. The essential requirement on the empirical quantity is that one can show an exponentially fast convergence towards $R_l\left[h\right]$ over the random draw of the training sample $\boldsymbol{z} \in \mathcal{Z}^m$ for a fixed hypothesis h. More formally, we require that[2]

$$\forall h \in \mathcal{Y}^{\mathcal{X}} : \qquad \mathbf{P}_{\mathbf{Z}^m}\left(\left|R_l\left[h\right] - \widehat{R}_l\left[h, \mathbf{Z}\right]\right| > \varepsilon\right) < \exp\left(-c\varepsilon^{\beta}m\right) ,$$

   for some constant $c \in \mathbb{R}^+$ and $\beta \in [1, 2]$. If the loss function l is bounded then an application of Hoeffding's inequality (Hoeffding, 1963) establishes such a convergence (see Feller, 1966; Devroye and Lugosi, 2001, for further results).

2. Exploiting the exponential convergence of the empirical quantity towards the prediction error, it is possible to bound the probability of training samples $\boldsymbol{z} \in \mathcal{Z}^m$ such that the prediction error deviates from the empirical quantity by more than ε, by twice the probability that the empirical quantity deviates by more than $\varepsilon/2$ when evaluated on two training samples[3] $\boldsymbol{z}, \tilde{\boldsymbol{z}} \in \mathcal{Z}^m$ drawn iid. This step is known as *symmetrisation by a ghost sample* and can either be shown to be valid uniformly over some hypothesis space $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ (see Lemma 20 and 21) or for the hypothesis learned from the first training sample (see Lemma 22 and 23). In fact, a closer look at the proof shows that in the uniform case we are effectively studying the algorithm

$$\mathcal{A}_{\text{worst}}\left(\boldsymbol{z}\right) := \operatorname*{argmax}_{\left\{h \in \mathcal{H} \,\middle|\, \widehat{R}_l[h, \boldsymbol{z}] = 0\right\}} R_l\left[h\right] , \qquad \mathcal{A}_{\text{worst}}\left(\boldsymbol{z}\right) := \operatorname*{argmax}_{h \in \mathcal{H}} \left|R_l\left[h\right] - \widehat{R}_l\left[h, \boldsymbol{z}\right]\right| ,$$

2. Depending on the type of result we are looking for, it is also possible to consider a multiplicative form, that is,

$$\forall h \in \mathcal{Y}^{\mathcal{X}} : \qquad \mathbf{P}_{\mathbf{Z}^m}\left(\frac{R_l\left[h\right]}{\widehat{R}_l\left[h, \mathbf{Z}\right]} > \varepsilon\right) < \exp\left(-c\varepsilon^{\beta}m\right) ,$$

   for some $c \in \mathbb{R}^+$ and $\beta \in [1, 2]$ (see, e.g. Anthony and Shawe-Taylor (1993)).

3. Such samples are referred to as *double samples* and the second sample is often called a *ghost sample*.

depending on whether we consider consistent or agnostic learning.

3. In order to bound the probability of the above-mentioned event over the random draw of double samples $\boldsymbol{z} \in \mathcal{Z}^{2m}$ we resort to a technique known as *symmetrisation by permutation* (Kahane, 1968). This technique exploits the assumption that the double sample is drawn iid since $\mathbf{P}_{\mathbf{Z}^{2m}}(\Upsilon(\mathbf{Z})) = \mathbf{P}_{\mathbf{Z}^{2m}}(\Upsilon(\Pi_{\boldsymbol{i}}(\mathbf{Z})))$ for any permutation $\pi_{\boldsymbol{i}}$. Thus,

$$\mathbf{P}_{\mathbf{Z}^{2m}}(\Upsilon(\mathbf{Z})) = \mathbf{E}_{\mathbf{I}}\left[\mathbf{P}_{\mathbf{Z}^{2m}|\mathbf{I}=\boldsymbol{i}}(\Upsilon(\Pi_{\boldsymbol{i}}(\mathbf{Z})))\right] = \mathbf{E}_{\mathbf{Z}^{2m}}\left[\mathbf{P}_{\mathbf{I}|\mathbf{Z}^{2m}=\boldsymbol{z}}(\Upsilon(\Pi_{\mathbf{I}}(\boldsymbol{z})))\right] \quad (1)$$

for any measure $\mathbf{P}_{\mathbf{I}}$ over permutations. The advantage of this conditioning step on permutations is that we can fix the double sample $\boldsymbol{z} \in \mathcal{Z}^{2m}$ and only need to determine the probability of permutations such that $\Upsilon(\Pi_{\boldsymbol{i}}(\boldsymbol{z}))$ holds. If $\mathbf{P}_{\mathbf{I}}$ is uniform then this reduces to simple counting. This technique also works if we only assume the training (and ghost) sample to be exchangeable which is slightly weaker than the iid assumption.

4. For any fixed double sample $\boldsymbol{z} \in \mathcal{Z}^{2m}$ we consider a cover (w.r.t. the defined empirical quantity) of all hypotheses (uniform case) or all hypotheses that can be generated by permuting the double sample and training on the first m training examples (algorithmic case). Given such a cover, we can now apply the *union bound* which naturally brings covering numbers of the hypothesis space (uniform case) or covering numbers of the "reachable" hypotheses (algorithmic case) into the generalisation error bound.

5. Finally, for a fixed hypothesis h in the cover we need to bound the probability that the empirical quantity measured on the training and ghost sample deviates by more than $\varepsilon/2$ over the random draw of permutations. It turns out that this bounding step can easily be reduced to an application of Hoeffding's inequality (Hoeffding, 1963) once we reduced general permutations to swappings (see Theorem 24).

Carrying out this analysis for a given algorithm leads to following VC type generalisation error bounds. Note that for practical purposes we need to bound the complexities[4] $\mathbf{E}_{\mathbf{Z}^{2m}}\left[\mathcal{N}\left(\frac{1}{2m}, \mathcal{L}_l(\mathcal{H}), \rho_1^{\mathbf{Z}}\right)\right]$ and $\mathbf{E}_{\mathbf{Z}^{2m}}[\mathcal{N}(\frac{1}{2m}, \mathcal{L}_l(\{\mathcal{A}(\Pi_{\boldsymbol{i}}(\mathbf{Z})_{[1:m]}) \mid \boldsymbol{i} \in I_{2m}\}), \rho_1^{\mathbf{Z}})]$ by their worst-case counterparts (replacing $\mathbf{E}_{\mathbf{Z}^{2m}}$ by $\sup_{\boldsymbol{z} \in \mathcal{Z}^{2m}}$) because if we know the distribution $\mathbf{P}_{\mathbf{Z}}$ we have solved the learning problem already[5].

Theorem 3 (VC bound) *For any probability measure $\mathbf{P}_{\mathbf{Z}}$, for any hypothesis space $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, for any learning algorithm $\mathcal{A} : \mathcal{Z}^{(\infty)} \to \mathcal{Y}^{\mathcal{X}}$, for any $\delta \in (0, 1]$, for any loss $l : \mathcal{Y} \times \mathcal{Y} \to$*

4. If the range of the loss function l is $\{0, 1\}$ then for any double sample $\boldsymbol{z} \in \mathcal{Z}^{2m}$,

$$\mathcal{N}\left(\frac{1}{2m}, \mathcal{L}_l(\mathcal{H}), \rho_1^{\boldsymbol{z}}\right) = \left|\mathcal{L}_l(\mathcal{H})_{|\boldsymbol{z}}\right|,$$

that is, the covering number at scale $\frac{1}{2m}$ equals the number of dichotomies of $\mathcal{L}_l(\mathcal{H})$ on $\boldsymbol{z}$.

5. An interesting question arising from this analysis is whether or not we can bound $\mathbf{E}_{\mathbf{Z}^{2m}}\left[\mathcal{N}\left(\frac{1}{2m}, \mathcal{L}_l(\mathcal{H}), \rho_1^{\mathbf{Z}}\right)\right]$ by its empirical counterpart $\mathcal{N}\left(\frac{1}{2m}, \mathcal{L}_l(\mathcal{H}), \rho_1^{\boldsymbol{z}}\right)$ using large deviation bounds for functions of random variables (see (Devroye and Lugosi, 2001; Boucheron et al., 2000) for first results).

$\{0,1\}$, *with probability at least* $1-\delta$ *over the random draw of the training sample* $\boldsymbol{z} \in \mathcal{Z}^m$ *of size* m, *for all hypotheses* $h \in \mathcal{H}$ *such that* $\widehat{R}_l[h, \boldsymbol{z}] = 0$,

$$R_l[h] \leq \frac{4}{m}\left(d_{\mathcal{H}}(2m) + \ln\left(\frac{2}{\delta}\right)\right). \tag{2}$$

Furthermore, if $\widehat{R}_l[\mathcal{A}(\boldsymbol{z}), \boldsymbol{z}] = 0$,

$$R_l[\mathcal{A}(\boldsymbol{z})] \leq \frac{4}{m}\left(d_{\mathcal{A}}(2m) + \ln\left(\frac{2}{\delta}\right)\right)$$

where $d_{\mathcal{H}}$ *and* $d_{\mathcal{A}}$ *are defined as follows:*

$$\begin{aligned} d_{\mathcal{H}}(m) &:= \ln\left(\sup_{\boldsymbol{z}\in\mathcal{Z}^m} \mathcal{N}\left(\frac{1}{m}, \mathcal{L}_l(\mathcal{H}), \rho_1^{\boldsymbol{z}}\right)\right), \\ d_{\mathcal{A}}(m) &:= \ln\left(\sup_{\boldsymbol{z}\in\mathcal{Z}^m} \mathcal{N}\left(\frac{1}{m}, \mathcal{L}_l(\mathcal{H}_{\mathcal{A}}(\boldsymbol{z})), \rho_1^{\boldsymbol{z}}\right)\right), \\ \mathcal{H}_{\mathcal{A}}(\boldsymbol{z}) &:= \left\{ \mathcal{A}\left(\Pi_{\boldsymbol{i}}(\boldsymbol{z})_{\left[1:\frac{|\boldsymbol{z}|}{2}\right]}\right) \;\middle|\; \boldsymbol{i} \in I_{|\boldsymbol{z}|} \right\}. \end{aligned} \tag{3}$$

For any algorithm $\mathcal{A}$ which chooses its hypotheses h from some pre-defined hypothesis space $\mathcal{H}$ we know by definition that $d_{\mathcal{A}}(m) \leq d_{\mathcal{H}}(m)$ because $\mathcal{H}_{\mathcal{A}}(\boldsymbol{z}) \subseteq \mathcal{H}$. Hence, exploiting the specific learning algorithm used we can (perhaps) reduce the complexity compared to a uniform study of hypothesis space. Note, however, that $d_{\mathcal{A}}(m)$ is driven by the worst training sample $\boldsymbol{z}_{\text{worst}} \in \mathcal{Z}^m$ although we might never observe $\boldsymbol{z}_{\text{worst}}$. In the case of uniform guarantees over some hypothesis space $\mathcal{H}$, a refinement of step 4 has been suggested in the "luckiness framework" to make the complexity $d_{\mathcal{H}}$ dependent on the observed training sample.

4. The Classical Luckiness Framework

The classical luckiness framework was introduced by Shawe-Taylor et al. (1998). In order to refine the covering number analysis in step 4 we introduce an *ordering* between all the hypotheses to be covered for a given double sample $\boldsymbol{z} \in \mathcal{Z}^m$. Such an ordering should be thought of as an a-priori preference for hypothesis h for a given sample; it expresses how lucky we are to observe the sample when considering the hypothesis h. Formally, this function, which is also called a *luckiness* function, is defined by $L : \mathcal{Y}^{\mathcal{X}} \times \mathcal{Z}^{(\infty)} \to \mathbb{R}$. Given a luckiness function L and a sample $\boldsymbol{z} \in \mathcal{Z}^m$, we order all hypotheses in descending order of their luckiness, that is, $L(h_i, \boldsymbol{z}) \geq L(h_{i+1}, \boldsymbol{z})$ for all $i \in \mathbb{N}$. Hence, each particular hypothesis $h \in \mathcal{H}$ can be used to index the subset

$$H(h, \boldsymbol{z}) := \{ g \in \mathcal{H} \mid L(g, \boldsymbol{z}) \geq L(h, \boldsymbol{z}) \}$$

of hypotheses which are higher up in the order defined by L (i.e. those more "lucky" than h). This allows the control of the covering number $\mathcal{N}\left(\frac{1}{m}, \mathcal{L}_l(H(h, \boldsymbol{z})), \rho_1^{\boldsymbol{z}}\right)$ by investigating only a subset of hypotheses. This is exactly what we need: imposing some empirically

measurable condition (such as $L(h, \boldsymbol{z}) \geq L_0$) we are able to reduce the hypotheses space size as measured by the covering numbers.

The biggest technical difficulty arises because we already introduced a ghost sample $\tilde{\boldsymbol{z}}$. Since we only observe a training sample $\boldsymbol{z}$ of size m we need to be able to determine the covering number $\mathcal{N}\left(\frac{1}{2m}, \mathcal{L}_l\left(H\left(h, \boldsymbol{z}\tilde{\boldsymbol{z}}\right)\right), \rho_1^{\boldsymbol{z}\tilde{\boldsymbol{z}}}\right)$ on *any* double sample $(\boldsymbol{z}\tilde{\boldsymbol{z}}) \in \mathcal{Z}^{2m}$ only using the *value* of the luckiness function on the first m examples, $L(h, \boldsymbol{z})$. If we can find a function $\omega : \mathbb{R} \times \mathbb{N} \to \mathbb{N}$ such that for all double samples $\boldsymbol{z} \in \mathcal{Z}^{2m}$

$$\forall h \in \mathcal{H} : \qquad \mathcal{N}\left(\frac{1}{2m}, \mathcal{L}_l\left(H\left(h, \boldsymbol{z}\right)\right), \rho_1^{\boldsymbol{z}}\right) \leq \omega\left(L\left(h, \boldsymbol{z}_{[1:m]}\right), m\right) \tag{4}$$

then we can use $\omega\left(L\left(h, \boldsymbol{z}_{[1:m]}\right), m\right)$ in place of the worst-case covering number $d_{\mathcal{H}}(2m)$ and thus incorporate a dependence on the observed training sample into the bound. Unfortunately, (4) is difficult to establish and the requirement needs to be relaxed by allowing the function ω to "use" parts of the training sample $\boldsymbol{z}_{[1:m]}$ to estimate $\mathcal{N}\left(\frac{1}{2m}, \mathcal{L}_l\left(H\left(h, \boldsymbol{z}\right)\right), \rho_1^{\boldsymbol{z}}\right)$. More formally, this reads as follows[6].

Definition 4 (ω-smallness of the luckiness function) *A luckiness function $L : \mathcal{Y}^{\mathcal{X}} \times \mathcal{Z}^{(\infty)} \to \mathbb{R}$ is* ω–small, *$\omega : \mathbb{R} \times \mathbb{N} \times (0, 1] \to \mathbb{N}$, if for all $m \in \mathbb{N}$, all $\delta \in (0, 1]$ and all measures $\mathbf{P}_{\mathbf{Z}}$,*

$$\mathbf{P}_{\mathbf{Z}^{2m}}\left(\exists h \in \mathcal{H} : \mathcal{N}\left(\frac{1}{2m}, \mathcal{L}_l\left(H\left(h, \mathbf{Z}\right)\right), \rho_1^{\mathbf{Z}}\right) > \omega\left(L\left(h, \mathbf{Z}_{[1:m]}\right), m, \delta\right)\right) < \delta\,.$$

It is worth noticing that the training sample is used because by virtue of (1), for any fixed double sample $\boldsymbol{z} \in \mathcal{Z}^{2m}$ we effectively need to count the number of permutations (swappings) such that[7]

$$\mathcal{N}\left(\frac{1}{2m}, \mathcal{L}_l\left(H\left(h, \boldsymbol{z}\right)\right), \rho_1^{\boldsymbol{z}}\right) > \omega\left(L\left(h, \Pi_{\boldsymbol{i}}\left(\boldsymbol{z}\right)_{[1:m]}\right), m, \delta\right)$$

and ensure that there will never be more than a fraction of δ satisfying the above event. In summary, if we carry out the above mentioned analysis in step 4 we arrive at the following main theorem of the luckiness framework.

Theorem 5 (Luckiness bound) *Suppose $L : \mathcal{Y}^{\mathcal{X}} \times \mathcal{Z}^{(\infty)} \to \mathbb{R}$ is a ω–small luckiness function. For any probability measure $\mathbf{P}_{\mathbf{Z}}$, for any binary loss $l : \mathcal{Y} \times \mathcal{Y} \to \{0, 1\}$, for any $d \in \mathbb{N}$ and for any $\delta \in (0, 1]$, with probability at least $1 - \delta$ over the random draw of the training sample $\boldsymbol{z} \in \mathcal{Z}^m$ of size m, if $\widehat{R}_l[h, \boldsymbol{z}] = 0$ and $\omega\left(L\left(h, \boldsymbol{z}\right), m, \frac{\delta}{4}\right) \leq 2^d$ then*

$$R_l[h] \leq \frac{2}{m}\left(d + \log_2\left(\frac{4}{\delta}\right)\right)\,. \tag{5}$$

In order to make the bound independent of the choice of d we "stratify" over the values[8] of $d \in \left\{1, \ldots, \frac{m}{2}\right\}$ by applying Lemma 19 with $p_d = \frac{2}{m}$. Hence, we obtain that with probability

6. Note that the notion of ω-smallness of the luckiness function is referred to as *probable smoothness* in the original paper Shawe-Taylor et al. (1998).
7. Note that Shawe-Taylor et al. (1998) assumed that the luckiness function L is permutation invariant.
8. Although by definition $\omega \leq 2^{2m}$ we can stop the application of Lemma 19 at $d = \frac{m}{2}$ because by the boundedness of the loss function $R_l[h] \leq 1$.

at least $1-\delta$ over the random draw of the training sample $\boldsymbol{z} \in \mathcal{Z}^m$, for all hypotheses h with zero training error, $\widehat{R}_l[h, \boldsymbol{z}] = 0$, the prediction error satisfies

$$R_l[h] \leq \underbrace{\frac{2}{m}\left(\left\lceil \log_2\left(\omega\left(L(h,\boldsymbol{z}), m, \frac{\delta}{2m}\right)\right)\right\rceil + \log_2\left(\frac{2m}{\delta}\right)\right)}_{b(h,\boldsymbol{z})} .$$

In contrast to the standard VC bound we observe that the complexity strongly depends on the training sample via the luckiness function L to be chosen beforehand. The luckiness framework is sometimes also called the data dependent structural risk minimisation framework because the training data $\boldsymbol{z} \in \mathcal{Z}^m$ is used to structure the hypothesis space $\mathcal{H}$ into subsets of increasing complexity $\mathcal{H}_1(\boldsymbol{z}, \delta) \subseteq \mathcal{H}_2(\boldsymbol{z}, \delta) \subseteq \cdots \subseteq \mathcal{H}$ where

$$\mathcal{H}_i(\boldsymbol{z}, \delta) := \left\{ h \in \mathcal{H} \;\middle|\; \omega\left(L(h,\boldsymbol{z}), |\boldsymbol{z}|, \frac{\delta}{2 \cdot |\boldsymbol{z}|}\right) \leq 2^i \right\} .$$

5. The Algorithmic Luckiness Framework

The classical luckiness framework solves the problem of sample dependence of the complexity measure while still suffering from the independence of the specific learning algorithm used. As a consequence, even if we consider the minimiser $h_{\boldsymbol{z}} := \operatorname{argmin}_{\left\{h \in \mathcal{H} \,\middle|\, \widehat{R}_l[h,\boldsymbol{z}]=0\right\}} b(h, \boldsymbol{z})$ of the bound given in Theorem 5, this theorem is also valid for the following algorithm

$$\mathcal{A}_{\text{worst,lucky}}(\boldsymbol{z}) := \operatorname*{argmax}_{\left\{h \in \mathcal{H} \,\middle|\, \left(\widehat{R}_l[h,\boldsymbol{z}]=0\right) \wedge \left(b(h,\boldsymbol{z}) = b(h_{\boldsymbol{z}},\boldsymbol{z})\right)\right\}} R_l[h] .$$

Though this algorithm cannot be implemented, we see that we need to take into account the learning algorithm $\mathcal{A}$ to eventually attain better generalisation error bounds. In contrast to the classical luckiness framework, where a luckiness $L(h, \boldsymbol{z})$ measures to what extent *any* $h \in \mathcal{Y}^{\mathcal{X}}$ "fits" to the training sample $\boldsymbol{z} \in \mathcal{Z}^{(\infty)}$ observed, we now only concentrate on the one function $\mathcal{A}(\boldsymbol{z})$ learned by the given learning algorithm $\mathcal{A} : \mathcal{Z}^{(\infty)} \to \mathcal{Y}^{\mathcal{X}}$. Similarly to the classical luckiness framework we need to introduce an ordering among all the hypotheses $\mathcal{H}_{\mathcal{A}}(\boldsymbol{z})$ which can be learned by the given learning algorithm (see (3)). The set $\mathcal{H}_{\mathcal{A}}(\boldsymbol{z})$ of hypotheses naturally occurs when introducing a ghost sample $\tilde{\boldsymbol{z}} \in \mathcal{Z}^m$ and applying (1).

Definition 6 (Algorithmic luckiness and lucky sets) *Any function L that maps an algorithm $\mathcal{A} : \mathcal{Z}^{(\infty)} \to \mathcal{Y}^{\mathcal{X}}$ and a training sample $\boldsymbol{z} \in \mathcal{Z}^{(\infty)}$ to a real value is called an* algorithmic luckiness. *For all even $m \in \mathbb{N}$, the* lucky set $\mathcal{H}_{\mathcal{A}}(L, \boldsymbol{z}) \subseteq \mathcal{H}_{\mathcal{A}}(\boldsymbol{z}) \subseteq \mathcal{Y}^{\mathcal{X}}$ *is the set of all hypotheses that are learned from the first $\frac{m}{2}$ examples $\left(z_{\pi_{\boldsymbol{i}}(1)}, \ldots, z_{\pi_{\boldsymbol{i}}\left(\frac{m}{2}\right)}\right)$ when permuting the whole sample $\boldsymbol{z}$, i.e.*

$$\mathcal{H}_{\mathcal{A}}(L, \boldsymbol{z}) := \left\{ \mathcal{A}\left(\Pi_{\boldsymbol{i}}(\boldsymbol{z})_{\left[1:\frac{|\boldsymbol{z}|}{2}\right]}\right) \mid \boldsymbol{i} \in \mathcal{I}_{\mathcal{A}}(L, \boldsymbol{z}) \right\} \subseteq \mathcal{H}_{\mathcal{A}}(\boldsymbol{z}) ,$$

where

$$\mathcal{I}_{\mathcal{A}}(L, \boldsymbol{z}) := \left\{ \boldsymbol{i} \in I_{|\boldsymbol{z}|} \;\middle|\; L\left(\mathcal{A}, \Pi_{\boldsymbol{i}}(\boldsymbol{z})_{\left[1:\frac{|\boldsymbol{z}|}{2}\right]}\right) \geq L\left(\mathcal{A}, \boldsymbol{z}_{\left[1:\frac{|\boldsymbol{z}|}{2}\right]}\right) \right\} .$$

We know by definition that for any double sample $\boldsymbol{z} \in \mathcal{Z}^{2m}$,

$$\left|\mathcal{L}_l\left(\mathcal{H}_{\mathcal{A}}\left(L,\boldsymbol{z}\right)\right)_{|\boldsymbol{z}}\right| \leq \left|\mathcal{L}_l\left(\mathcal{H}_{\mathcal{A}}\left(L,\boldsymbol{z}\right)\right)\right| \leq \left|\mathcal{H}_{\mathcal{A}}\left(L,\boldsymbol{z}\right)\right| \leq \left|\mathcal{I}_{\mathcal{A}}\left(L,\boldsymbol{z}\right)\right| \leq \frac{(2m)!}{m!}, \tag{6}$$

because among the $(2m)! = |I_{2m}|$ many different permutations there are equivalence classes of size $m!$ where only examples in the second half $\Pi_{\boldsymbol{i}}\left(\boldsymbol{z}\right)_{[(m+1):2m]}$ (ghost sample) are permuted which does not change $\mathcal{A}(\Pi_{\boldsymbol{i}}(\boldsymbol{z})_{[1:m]})$. The general idea we shall pursue closely follows the argument in the classical luckiness framework. In fact, as we are only given the training sample $(z_1,\ldots,z_m)$ we need to be able to bound the covering number $\mathcal{N}\left(\tau,\mathcal{L}_l\left(\mathcal{H}_{\mathcal{A}}\left(L,\boldsymbol{z}\right)\right),\rho_1^{\boldsymbol{z}}\right)$ only using the luckiness of $\mathcal{A}$ on the first half of the double sample $\boldsymbol{z} \in \mathcal{Z}^{2m}$. Again, if we are able to show that with high probability over the random draw of the training *and* ghost sample this covering number can be bounded by a function of the luckiness on the training sample only then we can exploit the value of the luckiness to devise training sample based bounds on the prediction error.

Definition 7 (ω–smallness of the algorithmic luckiness function) *Given an algorithm $\mathcal{A} : \mathcal{Z}^{(\infty)} \to \mathcal{Y}^{\mathcal{X}}$ and a loss $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$ the algorithmic luckiness function L is* ω–small at scale $\tau \in \mathbb{R}^+$ *if for all $m \in \mathbb{N}$, all $\delta \in (0,1]$ and all $\mathbf{P}_{\mathsf{Z}}$,*

$$\mathbf{P}_{\mathsf{Z}^{2m}}\left(\mathcal{N}\left(\tau,\mathcal{L}_l\left(\mathcal{H}_{\mathcal{A}}\left(L,\mathbf{Z}\right)\right),\rho_1^{\mathbf{Z}}\right) > \omega\left(L\left(\mathcal{A},\mathbf{Z}_{[1:m]}\right),l,m,\delta,\tau\right)\right) < \delta\,. \tag{7}$$

The purpose of the function ω is to exploit the value of the luckiness on the first m examples, $L\left(\mathcal{A},(z_1,\ldots,z_m)\right)$, in order to upper bound the covering number of "reachable" hypotheses (or their induced loss functions) exceeding this value. A hypothesis $h \in \mathcal{Y}^{\mathcal{X}}$ can be reached if the (fixed) learning algorithm returns this function for a certain permutation of the double sample. Using the ω–smallness of L we have our two main theorems.

Theorem 8 (Algorithmic luckiness bound for binary losses) *Suppose we have a learning algorithm $\mathcal{A} : \mathcal{Z}^{(\infty)} \to \mathcal{Y}^{\mathcal{X}}$ and an algorithmic luckiness L that is ω–small at scale $\frac{1}{2m}$ for a binary loss function $l : \mathcal{Y} \times \mathcal{Y} \to \{0,1\}$. For any probability measure $\mathbf{P}_{\mathsf{Z}}$, any $d \in \mathbb{N}$ and any $\delta \in (0,1]$, with probability at least $1-\delta$ over the random draw of the training sample $\boldsymbol{z} \in \mathcal{Z}^m$ of size m according to $\mathbf{P}_{\mathsf{Z}^m}$, if $\widehat{R}_l\left[\mathcal{A}\left(\boldsymbol{z}\right),\boldsymbol{z}\right] = 0$ and $\omega\left(L\left(\mathcal{A},\boldsymbol{z}\right),l,m,\frac{\delta}{4},\frac{1}{2m}\right) \leq 2^d$ then*

$$R_l\left[\mathcal{A}\left(\boldsymbol{z}\right)\right] \leq \frac{2}{m}\left(d + \log_2\left(\frac{4}{\delta}\right)\right).$$

Theorem 9 (Algorithmic luckiness bound for bounded losses) *Suppose we have a learning algorithm $\mathcal{A} : \mathcal{Z}^{(\infty)} \to \mathcal{Y}^{\mathcal{X}}$ and an algorithmic luckiness L that is ω–small at scale τ for a bounded loss function $l : \mathcal{Y} \times \mathcal{Y} \to [0,1]$. For any probability measure $\mathbf{P}_{\mathsf{Z}}$, any $d \in \mathbb{N}$ and any $\delta \in (0,1]$, with probability at least $1-\delta$ over the random draw of the training sample $\boldsymbol{z} \in \mathcal{Z}^m$ of size m according to $\mathbf{P}_{\mathsf{Z}^m}$, if $\omega\left(L\left(\mathcal{A},\boldsymbol{z}\right),l,m,\frac{\delta}{4},\tau\right) \leq 2^d$ then*

$$R_l\left[\mathcal{A}\left(\boldsymbol{z}\right)\right] \leq \widehat{R}_l\left[\mathcal{A}\left(\boldsymbol{z}\right),\boldsymbol{z}\right] + \sqrt{\frac{8}{m}\left(d + \log_2\left(\frac{4}{\delta}\right)\right)} + 4\tau\,.$$

The proofs can be found in Appendix A.6 and A.7; they closely follow the idea outlined in Section 3. These two bounds constitute the main results of the algorithmic luckiness framework. Note that a straightforward application of the multiple testing lemma (see Lemma 19) allows us to remove the assumption on the computable numbers $\log_2\left(\omega\left(L\left(\mathcal{A},\boldsymbol{z}\right),l,m,\delta/4,\tau\right)\right)$ because if this number exceeds $m/2$ both bounds become trivially true. As a consequence, for any learning algorithm $\mathcal{A}$ and any ω–small luckiness function L, with probability at least $1-\delta$ over the random draw of the training sample $\boldsymbol{z}$ of size m, for a binary loss function $l:\mathcal{Y}\times\mathcal{Y}\to\{0,1\}$ we know that if $\widehat{R}_l\left[\mathcal{A}\left(\boldsymbol{z}\right),\boldsymbol{z}\right]=0$,

$$R_l\left[\mathcal{A}\left(\boldsymbol{z}\right)\right]\leq\frac{2}{m}\left(\left\lceil\log_2\left(\omega\left(L\left(\mathcal{A},\boldsymbol{z}\right),l,m,\frac{\delta}{2m},\frac{1}{2m}\right)\right)\right\rceil+\log_2\left(\frac{2m}{\delta}\right)\right),$$

and for any bounded loss $l:\mathcal{Y}\times\mathcal{Y}\to[0,1]$,

$$R_l\left[\mathcal{A}\left(\boldsymbol{z}\right)\right]\leq\widehat{R}_l\left[\mathcal{A}\left(\boldsymbol{z}\right),\boldsymbol{z}\right]+\sqrt{\frac{8}{m}\left(\left\lceil\log_2\left(\omega\left(L\left(\mathcal{A},\boldsymbol{z}\right),l,m,\frac{\delta}{2m},\tau\right)\right)\right\rceil+\log_2\left(\frac{2m}{\delta}\right)\right)}+4\tau\,.$$

The difference to the main results in the classical luckiness framework (see (5)) and the VC framework (see (2)) is only within the definition of the complexity:

- In the VC framework, the complexity $d=d_{\mathcal{H}}\left(2m\right)$ is only dependent on the hypothesis space $\mathcal{H}$; it is always assumed that $\mathcal{A}:\mathcal{Z}^{(\infty)}\to\mathcal{H}\subseteq\mathcal{Y}^{\mathcal{X}}$. The main motivation for doing so stems from the so-called *key theorem of learning theory* which says that the (distribution independent) consistency of the empirical risk minimisation (ERM) algorithm $\mathcal{A}_{\mathrm{ERM}}$ is equivalent to the uniform convergence of training errors $\widehat{R}_l\left[h,\boldsymbol{z}\right]$ to prediction errors $R_l\left[h\right]$ over the whole hypothesis space $\mathcal{H}$. The ERM algorithm is formally defined by
$$\mathcal{A}_{\mathrm{ERM}}\left(\boldsymbol{z}\right):=\operatorname*{argmin}_{h\in\mathcal{H}}\ \widehat{R}_l\left[h,\boldsymbol{z}\right]\,. \tag{8}$$
- In the classical luckiness framework, the complexity $d=\left\lceil\log_2\left(\omega\left(L\left(h,\boldsymbol{z}\right),m,\frac{\delta}{2m}\right)\right)\right\rceil$ is dependent on the training sample $\boldsymbol{z}\in\mathcal{Z}^m$ via the luckiness function. This allows to study learning algorithms which map to a much richer space such as $\mathcal{Y}^{\mathcal{X}}$, given that the luckiness function L is using the training sample. However, the requirement on the luckiness function L is very strict since it has to provide a bound on the covering number which holds uniformly over the hypothesis space regardless of the fact that the given learning algorithm might never have learned certain hypotheses.
- In the algorithmic luckiness framework, the complexity $d=\left\lceil\log_2\left(\omega\left(L\left(\mathcal{A},\boldsymbol{z}\right),l,m,\frac{\delta}{2m},\tau\right)\right)\right\rceil$ is dependent on both the training sample $\boldsymbol{z}\in\mathcal{Z}^m$ and the learning algorithm $\mathcal{A}$ via the algorithmic luckiness function L. The extra parameters are necessary because the algorithmic luckiness framework is also applicable to the case of non-zero training error $\widehat{R}_l\left[\mathcal{A}\left(\boldsymbol{z}\right),\boldsymbol{z}\right]$.

6. Examples of ω–Small Luckiness Functions

In this section we present several algorithmic luckiness functions which illuminate the relationship of the new framework with already existing mathematical models of learning.

As with the classical luckiness framework, the whole difficulty in the algorithmic luckiness framework has been shifted into finding ω–small luckiness functions L. Since the ω–smallness condition (7) has to hold regardless of the measure $\mathbf{P}_{\mathbf{Z}}$, essentially only two techniques can be used:

1. If we can show that for a fixed learning algorithm $\mathcal{A}$ and a given luckiness L, there exists a function ω of $L\left(\mathcal{A}, (z_1, \ldots, z_m)\right)$ that *for all double samples* $\boldsymbol{z} = (z_1, \ldots, z_{2m})$ is a strict upper bound on the covering number $\mathcal{N}\left(\tau, \mathcal{L}_l\left(\mathcal{H}_{\mathcal{A}}\left(L, \boldsymbol{z}\right)\right), \rho_1^{\boldsymbol{z}}\right)$, the function ω will be independent of δ and satisfy the requirements of Definition 7.

2. Since $\mathbf{P}_{\mathbf{Z}^m}$ is a product measure, it is (in general) the only measure which is invariant under any permutation. In other words, if

$$J\left(\boldsymbol{z}\right) \equiv \mathcal{N}\left(\tau, \mathcal{L}_l\left(\mathcal{H}_{\mathcal{A}}\left(L, \boldsymbol{z}\right)\right), \rho_1^{\boldsymbol{z}}\right) > \omega\left(L\left(\mathcal{A}, \boldsymbol{z}_{[1:m]}\right), l, m, \delta, \tau\right)$$

   then it has to hold that

$$\mathbf{P}_{\mathbf{Z}^{2m}}\left(J\left(\mathbf{Z}\right)\right) = \mathbf{E}_{\mathbf{I}}\left[\mathbf{P}_{\mathbf{Z}^{2m}|\mathbf{I}=\boldsymbol{i}}\left(J\left(\Pi_{\boldsymbol{i}}\left(\mathbf{Z}\right)\right)\right)\right] = \mathbf{E}_{\mathbf{Z}^{2m}}\left[\mathbf{P}_{\mathbf{I}|\mathbf{Z}^{2m}=\boldsymbol{z}}\left(J\left(\Pi_{\mathbf{I}}\left(\boldsymbol{z}\right)\right)\right)\right]$$

   for any measure $\mathbf{P}_{\mathbf{I}}$ over permutations $\pi_{\boldsymbol{i}}$. The advantage is that in the last statement it suffices to show that the fraction of permutations $\pi_{\boldsymbol{i}}$ which satisfy

$$\mathcal{N}\left(\tau, \mathcal{L}_l\left(\mathcal{H}_{\mathcal{A}}\left(L, \Pi_{\boldsymbol{i}}\left(\boldsymbol{z}\right)\right)\right), \rho_1^{\boldsymbol{z}}\right) > \omega\left(L\left(\mathcal{A}, \Pi_{\boldsymbol{i}}\left(\boldsymbol{z}\right)_{[1:m]}\right), l, m, \delta, \tau\right)$$

   is less than δ for any $\boldsymbol{z} \in \mathcal{Z}^{2m}$. Thus, the original problem has been reduced to a problem of counting permutations. Note that we exploited the fact that, by definition, $\mathcal{N}\left(\cdot, \cdot, \rho_1^{\boldsymbol{z}}\right)$ is a permutation invariant function of $\boldsymbol{z}$.

These two tricks are the only tools we will need to relate the algorithmic luckiness framework to previous studies. It is sometimes possible to bound the covering number $\mathcal{N}\left(\tau, \mathcal{L}_l\left(\mathcal{H}_{\mathcal{A}}\left(L, \boldsymbol{z}\right)\right), \rho_1^{\boldsymbol{z}}\right)$ using (6).

6.1 VC Dimension of Hypothesis Spaces

In the case of learning algorithms $\mathcal{A}$ that choose their hypothesis from a subset $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ and binary loss functions $l : \mathcal{Y} \times \mathcal{Y} \to \{0, 1\}$ we obtain that, regardless of L, $|\mathcal{L}_l(\mathcal{H}_{\mathcal{A}}\left(L, \boldsymbol{z}\right))_{|\boldsymbol{z}}|$ is no greater than the value of the growth function $\mathcal{G}_{\mathcal{H}}\left(2m\right)$ because the latter is defined as follows

$$\mathcal{G}_{\mathcal{H}}\left(m\right) := \sup_{\boldsymbol{z} \in \mathcal{Z}^m} \left|\mathcal{L}_l\left(\mathcal{H}\right)_{|\boldsymbol{z}}\right|, \tag{9}$$

and by definition $\mathcal{H}_{\mathcal{A}}\left(L, \boldsymbol{z}\right) \subseteq \mathcal{H}$ for any L and $\boldsymbol{z} \in \mathcal{Z}^{2m}$. This quantity can be bounded from above in terms of the VC-dimension $\vartheta_{\mathcal{H}}$ of $\mathcal{H}$ (for details see Vapnik, 1982; Kearns and Vazirani, 1994; Herbrich, 2002). In particular, the result reads as follows (Vapnik and Chervonenkis, 1971; Sauer, 1972).

Theorem 10 (Bound on the growth function) *For any hypothesis space $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ and any loss function $l : \mathcal{Y} \times \mathcal{Y} \to \{0, 1\}$ the growth function (9) either*

1. *satisfies the equality*

$$\forall m \in \mathbb{N}: \qquad \mathcal{G}_{\mathcal{H}}(m) = 2^m,$$

2. *or there exists a natural number* $\vartheta_{\mathcal{H}} \in \mathbb{N}$ *such that*

$$\mathcal{G}_{\mathcal{H}}(m) \begin{cases} = 2^m & \text{if } m < \vartheta_{\mathcal{H}} \\ \leq \sum_{i=0}^{\vartheta_{\mathcal{H}}} \binom{m}{i} & \text{if } m \geq \vartheta_{\mathcal{H}} \end{cases}.$$

Note, however, that this bound is very coarse because we neither exploit any prior knowledge about the algorithm and the unknown probability measure using $L_{\mathrm{VC}} := -\vartheta_{\mathcal{H}}$ nor will, in general, the lucky set $\mathcal{H}_{\mathcal{A}}(L_{\mathrm{VC}}, \boldsymbol{z})$ be $\mathcal{H}$. The sole justification for the usage of the growth function as a complexity measure is due to a theorem which became known as the "key theorem of learning theory" (Vapnik and Chervonenkis, 1991; Vapnik, 1995). The theorem states that the consistency of the empirical risk minimisation algorithm (see (8)) is equivalent to the uniform convergence[9] of training errors to prediction errors over $\mathcal{H}$. Without any assumptions on $\mathbf{P}_{\mathsf{Z}}$, the uniform convergence is equivalent to a sub-exponential scaling of $\mathcal{G}_{\mathcal{H}}$, i.e. finiteness of the VC dimension $\vartheta_{\mathcal{H}}$. This can be seen by applying Lemma 18 to the expression in Theorem 10 yielding

$$\forall 2m > \vartheta_{\mathcal{H}}: \qquad \log_2\left(\omega\left(L_0, l, m, \delta, \frac{1}{2m}\right)\right) \leq \log_2\left(\mathcal{G}_{\mathcal{H}}(2m)\right) \leq \vartheta_{\mathcal{H}} \cdot \log_2\left(\frac{2em}{\vartheta_{\mathcal{H}}}\right). \quad (10)$$

6.2 Sparsity of Compression Schemes

In Littlestone and Warmuth (1986) and Floyd and Warmuth (1995) a set of learning algorithms was studied which have the appealing property that $\mathcal{A}(\boldsymbol{z})$ can be reproduced from a smaller subsample $\tilde{\boldsymbol{z}} \subset \boldsymbol{z}$ of the training sample $\boldsymbol{z}$. More formally this reads as follows.

Definition 11 (Compression schemes) *The algorithm* $\mathcal{A}: \mathcal{Z}^{(\infty)} \to \mathcal{Y}^{\mathcal{X}}$ *is a* compression scheme *if and only if the algorithm can be written as*

$$\mathcal{A}(\boldsymbol{z}) = \mathcal{R}\left(\boldsymbol{z}_{\mathcal{C}(\boldsymbol{z})}\right)$$

where $\mathcal{R}: \mathcal{Z}^{(\infty)} \to \mathcal{Y}^{\mathcal{X}}$ *is called the reconstruction function and is assumed to be permutation invariant. The compression function* $\mathcal{C}: \mathcal{Z}^{(\infty)} \to \mathbb{N}^{(\infty)}$ *maps training samples to index vectors* $(i_1, \ldots, i_d)$, $1 \leq i_1 < \cdots < i_d$.

An example of a compression scheme is given by the maximum margin algorithm also known as a support vector machine (Boser et al., 1992; Cortes and Vapnik, 1995) (see also Section 7):

9. More formally, this reads as follows: *For any measure* $\mathbf{P}_{\mathsf{Z}}$, *for any loss functions* $l: \mathcal{Y} \times \mathcal{Y} \to \{0, 1\}$, *for any* $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ *it holds that*

$$\forall \varepsilon > 0: \lim_{m \to \infty} \mathbf{P}_{\mathsf{Z}^m}\left(\left(\sup_{h \in \mathcal{H}} R_l[h] - \widehat{R}_l[h, \mathbf{Z}]\right) > \varepsilon\right) = 0 \Leftrightarrow$$
$$\forall \varepsilon > 0: \lim_{m \to \infty} \mathbf{P}_{\mathsf{Z}^m}\left(\left(R_l[\mathcal{A}_{\mathrm{ERM}}(\mathbf{Z})] - \inf_{h \in \mathcal{H}} R_l[h]\right) > \varepsilon\right) = 0.$$

The compression function is given by a run of the maximum margin algorithm returning only the indices $\mathcal{C}(\boldsymbol{z})$ of the so-called *support vectors*. By the Karush-Kuhn-Tucker conditions we know that running the maximum margin algorithm only on the support vectors, $\boldsymbol{z}_{\mathcal{C}(\boldsymbol{z})}$, we will obtain the same hypothesis as running the algorithm on the full sample $\boldsymbol{z}$. Intuitively, the smaller the value $|\mathcal{C}(\boldsymbol{z})| \in \{1, \ldots, |\boldsymbol{z}|\}$ the less choices the learning algorithm had in the construction of the hypothesis which should result in a tighter bound on the prediction error $R_l[\mathcal{A}(\boldsymbol{z})]$ of $\mathcal{A}(\boldsymbol{z})$. In order to cast this notion into the algorithmic luckiness framework we introduce the sparsity luckiness.

Theorem 12 (Sparsity luckiness) *Given an algorithm $\mathcal{A} : \mathcal{Z}^{(\infty)} \to \mathcal{Y}^{\mathcal{X}}$ which is a compression scheme the* sparsity luckiness L_{sparse} *is defined by*

$$L_{\text{sparse}}(\mathcal{A}, \boldsymbol{z}) := -|\mathcal{C}(\boldsymbol{z})| .$$

The sparsity luckiness is ω–small at any scale $\tau \in \mathbb{R}^+$ where

$$\omega(L_0, l, m, \delta, \tau) = \sum_{i=0}^{-L_0} \binom{2m}{i} . \tag{11}$$

Proof Let us consider any double sample $\boldsymbol{z} \in \mathcal{Z}^{2m}$. According to Definition 7 and (6) it suffices to bound the size of $|\mathcal{I}_{\mathcal{A}}(L_{\text{sparse}}, \boldsymbol{z})|$ because this is an upper bound on $|\mathcal{L}_l(\mathcal{H}_{\mathcal{A}}(L_{\text{sparse}}, \boldsymbol{z}))|$. Let k be the sparsity of $\mathcal{A}$ on $(z_1, \ldots, z_m)$, i.e. $k = \left|\mathcal{C}\left(\boldsymbol{z}_{[1:m]}\right)\right|$. In order to generate a new hypothesis using $\mathcal{R}$ we have to find another subsample $\tilde{\boldsymbol{z}} \subset \boldsymbol{z}$ of size not larger than k because only permutations where the luckiness does not decrease are considered in $\mathcal{I}_{\mathcal{A}}(L_{\text{sparse}}, \boldsymbol{z})$. Note that the number of i distinct indices from $\{1, \ldots, 2m\}$ is exactly $\binom{2m}{i}$. Noting that the order of the examples is irrelevant for the reconstruction function (see Definition 11), there are no more than $\sum_{i=0}^{k} \binom{2m}{i}$ distinct subsets of size not greater than k of the double sample $\boldsymbol{z} \in \mathcal{Z}^{2m}$ which could (potentially) be used by the learning algorithm to reconstruct hypothesis. The theorem is proven. ■

If we apply Lemma 18 to (11) we obtain that the effective complexity of the function learned by a compression scheme is

$$\log_2(\omega(L_0, l, m, \delta, \tau)) \leq -L_0 \cdot \log_2\left(\frac{2em}{-L_0}\right) ,$$

which should be compared with the corresponding result (10) of VC theory. In contrast to the results by Floyd and Warmuth (1995) we do not need to reduce the effective training sample size to $(m - |\mathcal{C}(\boldsymbol{z})|)$ at the price of $2m$ rather than m in the binomial term and an extra factor of 2 resulting from the basic lemma. For large training sample sizes, the positive effect is that the training error is not discounted by the factor $\frac{m-|\mathcal{C}(\boldsymbol{z})|}{m}$ (see Graepel et al., 2000) but only the complexity term is influenced. Finally we see that the sparsity luckiness makes use of both the learning algorithm (via $\mathcal{C}$) and the training sample (via $|\mathcal{C}(\boldsymbol{z})|$) which somehow explains why compression bounds are very tight in practical applications.

6.3 Uniform Algorithmic Stability

The idea behind uniform algorithmic stability is as follows (Bousquet and Elisseeff, 2001): If the influence of a single training example $(x_i, y_i) \in \boldsymbol{z}$ on the learned function $\mathcal{A}(\boldsymbol{z})$ is decreasing with increasing training sample size m, then it should be possible to exploit this stability for bounding the generalisation error. As we are only interested in the predictions of the function learned on new test point, the influence is usually measured by the maximum change in the functions output. More formally, this reads as follows.

Definition 13 (Uniform stability) *Let $(\beta_i)_{i\in\mathbb{N}}$ be a decreasing sequence of positive real numbers. A learning algorithm $\mathcal{A} : \mathcal{Z}^{(\infty)} \to \mathcal{Y}^{\mathcal{X}}$ is said to be β_m–stable w.r.t. the loss function $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ if for all $i \in \{1, \ldots, m\}$,*

$$\forall \boldsymbol{z} \in \mathcal{Z}^m : \forall (x, y) \in \mathcal{Z} : \quad \left| l\left(\mathcal{A}(\boldsymbol{z})(x), y\right) - l\left(\mathcal{A}\left((z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_m)\right)(x), y\right)\right| \leq \beta_m .$$

Then the following generalisation error bound was proved by Bousquet and Elisseeff (2001).

Theorem 14 (Algorithmic stability bound) *Suppose the learning algorithm $\mathcal{A} : \mathcal{Z}^{(\infty)} \to \mathcal{Y}^{\mathcal{X}}$ is β_m–stable w.r.t. a loss function $l : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$, with probability at least $1 - \delta$ over the random draw of training samples $\boldsymbol{z} \in \mathcal{Z}^m$,*

$$R_l\left[\mathcal{A}(\boldsymbol{z})\right] \leq \widehat{R}_l\left[\mathcal{A}(\boldsymbol{z}), \boldsymbol{z}\right] + 2\beta_m + \sqrt{2m\left(4\beta_m + \frac{1}{m}\right)^2 \ln\left(\frac{1}{\delta}\right)} . \tag{12}$$

Note that the last term dominates the bound. Thus, in order to have a convergence of $\widehat{R}_l\left[\mathcal{A}(\boldsymbol{z}), \boldsymbol{z}\right]$ to $R_l\left[\mathcal{A}(\boldsymbol{z})\right]$, β_m has to decrease at a rate faster than $m^{-\frac{1}{2}}$, i.e. we require

$$\lim_{m\to\infty} \beta_m \cdot \sqrt{m} = 0 . \tag{13}$$

In order to relate algorithmic luckiness to uniform algorithmic stability we consider a more refined version of uniform stability. Broadly speaking, since we consider a ghost sample of size m it is only necessary to know by how much the loss function is changing on the double sample when swapping from the ghost sample and training sample.

Definition 15 (ν–stability) *A permutation invariant learning algorithm $\mathcal{A} : \mathcal{Z}^{(\infty)} \to \mathcal{Y}^{\mathcal{X}}$ is said to be ν–stable w.r.t. the loss function $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ if for all double samples $\boldsymbol{z} \in \mathcal{Z}^{2m}$ and all $\boldsymbol{i}, \boldsymbol{j} \in I_{2m}$,*

$$\frac{1}{2m}\sum_{k=1}^{2m} \left| l\left(\mathcal{A}\left(\Pi_{\boldsymbol{i}}(\boldsymbol{z})_{[1:m]}\right)(x_k), y_k\right) - l\left(\mathcal{A}\left(\Pi_{\boldsymbol{j}}(\boldsymbol{z})_{[1:m]}\right)(x_k), y_k\right)\right| \leq \nu\left(m, D(\boldsymbol{i}, \boldsymbol{j})\right) ,$$

where $D(\boldsymbol{i}, \boldsymbol{j}) := m - |\{i_1, \ldots, i_m\} \cap \{j_1, \ldots, j_m\}|$.

For permutation invariant learning algorithms uniform stability is a stronger notion of stability because any permutation invariant and uniformly stable algorithm is ν–stable.

Lemma 16 (Uniform stability implies ν–stability) *If a permutation invariant learning algorithm $\mathcal{A} : \mathcal{Z}^{(\infty)} \to \mathcal{Y}^{\mathcal{X}}$ is β_m–stable w.r.t. the loss function $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ then it is ν–stable with*

$$\nu(m, d) = 2 \cdot d \cdot \beta_m .$$

Proof Consider an arbitrary $\boldsymbol{z} \in \mathcal{Z}^{2m}$. Given two permutations $\pi_{\boldsymbol{i}} \in I_{2m}$ and $\pi_{\boldsymbol{j}} \in I_{2m}$ such that $D(\boldsymbol{i}, \boldsymbol{j}) = d$, without loss of generality we can assume that $i_1 \neq j_1, \ldots, i_d \neq j_d$ and $i_{d+1} = j_{d+1}, \ldots, i_m = j_m$ because $\mathcal{A}$ is permutation invariant. Then we can always find d permutations $\pi_{\boldsymbol{i}_1}, \ldots, \pi_{\boldsymbol{i}_d}$ such $(\boldsymbol{i}_k)_1 \neq j_1, \ldots, (\boldsymbol{i}_k)_{d-k} \neq j_{d-k}$ and $(\boldsymbol{i}_k)_{d-k+1} = j_{d-k+1}, \ldots, (\boldsymbol{i}_k)_m = j_m$. For a given $(x, y) \in \mathcal{Z}$, let $l_k(x, y) := l(\mathcal{A}(\Pi_{\boldsymbol{i}_k}(\boldsymbol{z})_{[1:m]})(x), y)$ and $l_{k\setminus n}(x, y) := l(\mathcal{A}(\Pi_{\boldsymbol{i}_k}(\boldsymbol{z})_{(1,\ldots,n-1,n+1,\ldots,m)})(x), y)$. By Definition 13 we know that for all $(x, y) \in \mathcal{Z}$

$$\begin{aligned} |l_k(x, y) - l_{k+1}(x, y)| &= \left|l_k(x, y) - l_{k\setminus k+1}(x, y) + l_{k\setminus k+1}(x, y) - l_{k+1}(x, y)\right| \\ &\leq \left|l_k(x, y) - l_{k\setminus k+1}(x, y)\right| + \left|l_{k\setminus k+1}(x, y) - l_{k+1}(x, y)\right| \\ &\leq 2 \cdot \beta_m . \end{aligned}$$

Noticing that $\mathcal{A}\left(\Pi_{\boldsymbol{i}}(\boldsymbol{z})_{[1:m]}\right) = h_0$ and $\mathcal{A}\left(\Pi_{\boldsymbol{j}}(\boldsymbol{z})_{[1:m]}\right) = h_d$ we thus obtain

$$\begin{aligned} |l_0(x_j, y_j) - l_d(x_j, y_j)| &= \left|\sum_{k=0}^{d-1} l_k(x_j, y_j) - l_{k+1}(x_j, y_j)\right| \\ &\leq \sum_{k=0}^{d-1} |l_k(x_j, y_j) - l_{k+1}(x_j, y_j)| \\ &\leq d \cdot 2 \cdot \beta_m . \end{aligned}$$

The lemma is proven. ∎

Since $D(\boldsymbol{i}, \boldsymbol{j}) \leq m$ by definition we know that the hypothesis $\mathcal{A}\left(\boldsymbol{z}_{[1:m]}\right)$ covers all possible loss functions $\mathcal{L}_l(\mathcal{H}_{\mathcal{A}}(\boldsymbol{z}))$ at scale $\nu(m, m)$ w.r.t. $\rho_1^{\boldsymbol{z}}$. As a consequence, for any double sample $\boldsymbol{z} \in \mathcal{Z}^{2m}$ and any luckiness function, $\omega\left(L\left(\mathcal{A}, \boldsymbol{z}_{[1:m]}\right), l, m, \delta, \nu(m, m)\right) = 1$. Using this result together with Theorem 9 we have shown that for any ν–stable (β_m–stable) learning algorithm $\mathcal{A} : \mathcal{Z}^{(\infty)} \to \mathcal{Y}^{\mathcal{X}}$ with probability at least $1 - \delta$ over the random draw of training samples $\boldsymbol{z}$,

$$\begin{aligned} R_l[\mathcal{A}(\boldsymbol{z})] &\leq \widehat{R}_l[\mathcal{A}(\boldsymbol{z}), \boldsymbol{z}] + \sqrt{\frac{8}{m}\log_2\left(\frac{4}{\delta}\right)} + 4\nu(m, m) \\ &\leq \widehat{R}_l[\mathcal{A}(\boldsymbol{z}), \boldsymbol{z}] + \sqrt{\frac{8}{m}\log_2\left(\frac{4}{\delta}\right)} + 8m\beta_m , \end{aligned}$$

where the last line follows from Lemma 16. In order to have a convergence of $\widehat{R}_l[\mathcal{A}(\boldsymbol{z}), \boldsymbol{z}]$ to $R_l[\mathcal{A}(\boldsymbol{z})]$, β_m has to decrease at a rate faster than m^{-1}, i.e.

$$\lim_{m\to\infty} \beta_m \cdot m = 0 ,$$

which should be compared with (13).

Our aim was not to improve the result in Theorem 14 but to reveal the relation of algorithmic stability results — which are usually proved using concentration inequalities — to algorithmic luckiness. The result obtained is weaker insofar as it makes a stronger requirement on the behaviour of β_m as a function of m. On the other hand, the way we obtained the result involved some very crude bounding steps. In order to improve the current result two ways are conceivable:

1. Rather than covering the induced loss function set by *one* single swapping at a very large scale of $\nu(m,m)$, one could envisage a larger cover at a smaller scale with the concomitant reduction of $\nu(m,\cdot)$. One approach to cover this set is to construct a cover of I_{2m} using the $D(\boldsymbol{i},\boldsymbol{j})$–metric. This can be related to the Hamming distance of binary strings and thus one can use results on the maximal size of a constant weight code to bound the covering number we seek McEliece et al. (see 1977).

2. Instead of using Lemma 16 one should try to bound $\nu(\cdot,\cdot)$ directly for some particular algorithms. In contrast to uniform stability, it would suffice to bound ν with high probability which can be readily incorporated into the notion of ω–smallness.

Finally, we remark that results such as Theorem 14 should be interpreted very carefully in relation to particular algorithms. For example, although Bousquet and Elisseeff have shown for algorithms which minimise a regularised functional of the form

$$R_{\text{reg}}[h,\boldsymbol{z}] := \widehat{R}_l[h,\boldsymbol{z}] + \lambda \|h\|$$

that $\beta_m \leq \frac{c}{\lambda m}$ for some constants $c \in \mathbb{R}^+$, one should not simply substitute this bound into (12) and consider the scaling behaviour in m since in practice the optimal λ (often chosen by cross-validation) will itself be a function of m. Furthermore, it is not possible to determine the scaling behaviour of λ as a function of m a-priori because it depends on the unknown target function. It would seem that some form of luckiness argumentation is necessary to allow λ to be dependent on the training sample. In contrast to concentration inequalities (which so far seems to require a uniform notion of stability) the algorithmic luckiness framework offers the advantage of exploiting knowledge of the target function directly using the luckiness function (which, so far, was constant and independent of the sample).

7. An Application of Algorithmic Luckiness to Linear Classifiers

In this section we study the maximum margin algorithm for linear classifiers, $\mathcal{A}_{\text{MM}} : \mathcal{Z}^{(\infty)} \to \mathcal{H}_{\boldsymbol{\phi}}$, where $\mathcal{H}_{\boldsymbol{\phi}} := \{x \mapsto \langle \boldsymbol{\phi}(x), \mathbf{w}\rangle \mid \mathbf{w} \in \mathcal{K}\}$ and $\boldsymbol{\phi} : \mathcal{X} \to \mathcal{K} \subseteq \ell_2^N$ is known as the feature mapping (see Boser et al., 1992; Cortes and Vapnik, 1995, for details). As our hypotheses $h \in \mathcal{H}_{\boldsymbol{\phi}}$ map to real-valued functions we define the zero-one loss w.r.t. $\text{sign}(h)$ as $l(h(x), y) = l_{0-1}(h(x), y) := \mathbb{I}_{yh(x)\leq 0}$. Classical VC generalisation error bounds exploit the fact that $\vartheta_{\mathcal{H}_{\boldsymbol{\phi}}} = N$ and thus use (10) in Theorem 8. According to this result it seems crucial that we have more training examples than dimensions of feature space, $m \gg N$, because otherwise the resulting bound on the prediction error becomes trivial. The impossibility to obtain

any practically useful results (bounds that are *independent* of N) in the VC framework is believed to be closely related to the intuitive notion of the *curse of dimensionality*. However, Shawe-Taylor et al. (1998) have shown that we can use $\mathrm{fat}_{\mathcal{H}_{\boldsymbol{\phi}}}(\gamma_{\boldsymbol{z}}(\mathbf{w})) \leq (\gamma_{\boldsymbol{z}}(\mathbf{w}))^{-2}$ (at the price of an extra $\log_2(32m)$ factor) in place of $\vartheta_{\mathcal{H}_{\boldsymbol{\phi}}}$ where

$$\gamma_{\boldsymbol{z}}(\mathbf{w}) := \min_{(x_i,y_i)\in\boldsymbol{z}} \frac{y_i \langle \boldsymbol{\phi}(x_i), \mathbf{w}\rangle}{\|\mathbf{w}\|}$$

is known as the margin. This result was proven essentially using a luckiness-based reasoning[10]. The most important difference to the VC-type result lies in the independence of the bound on the number of dimensions of feature space, N. If the training sample $\boldsymbol{z}$ could be correctly classified with a hypothesis $h_{\mathbf{w}}$ that has a margin $\gamma_{\boldsymbol{z}}(\mathbf{w})$ which is substantially larger than $\sqrt{m}$ then, regardless of the dimensionality N of the feature space $\mathcal{K}$, $h_{\mathbf{w}}$ has a small prediction error. This result forms the theoretical basis for the *maximum margin algorithm* which finds the weight vector $\mathbf{w}_{\mathrm{MM}}$ that maximises $\gamma_{\boldsymbol{z}}(\mathbf{w})$. It is known (Schölkopf et al., 2001) that $\mathbf{w}_{\mathrm{MM}}$ can be written as a linear combination of the $\boldsymbol{\phi}(x_i)$, i.e.

$$\mathbf{w}_{\mathrm{MM}} = \sum_{i=1}^{m} \alpha_i \boldsymbol{\phi}(x_i) \ .$$

Interestingly, however, the bound by Shawe-Taylor et al. (1998) does not only hold for the large margin classifier $\mathbf{w}_{\mathrm{MM}}$ but for every classifier which has zero training error.

In the following we shall present an algorithmic luckiness function which is only valid for the maximum margin algorithm. For notational convenience, we shall assume that $\mathcal{A}_{\mathrm{MM}} : \mathcal{Z}^{(\infty)} \to \mathbb{R}^{(\infty)}$ maps to the expansion coefficients $\boldsymbol{\alpha}$ such that $\|\mathbf{w}_{\boldsymbol{\alpha}}\| = 1$ where $\mathbf{w}_{\boldsymbol{\alpha}} := \sum_{i=1}^{|\boldsymbol{\alpha}|} \alpha_i \boldsymbol{\phi}(x_i)$. Thus, $\|\mathcal{A}_{\mathrm{MM}}(\boldsymbol{z})\|_1$ means the 1–norm of the expansion coefficients $\boldsymbol{\alpha}$. Then, our new margin bound follows from the following theorem together with Theorem 8.

Theorem 17 *Let $\epsilon_i(\boldsymbol{x})$ be the smallest $\epsilon > 0$ such that $\{\boldsymbol{\phi}(x_1), \ldots, \boldsymbol{\phi}(x_m)\}$ can be covered by at most i balls of radius less than or equal ϵ. Let $\Gamma_{\boldsymbol{z}}(\mathbf{w})$ be defined by*

$$\Gamma_{\boldsymbol{z}}(\mathbf{w}) := \min_{(x_i,y_i)\in\boldsymbol{z}} \frac{y_i \langle \boldsymbol{\phi}(x_i), \mathbf{w}\rangle}{\|\boldsymbol{\phi}(x_i)\| \cdot \|\mathbf{w}\|} . \tag{14}$$

For the zero-one loss l_{0-1} and the maximum margin algorithm $\mathcal{A}_{\mathrm{MM}}$, the luckiness function

$$L_{\mathrm{MM}}(\mathcal{A}_{\mathrm{MM}}, \boldsymbol{z}) := -\min \left\{ i = 4j,\ j \in \mathbb{N} \;\middle|\; i \geq 4 \cdot \left(\frac{\epsilon_{\frac{i}{4}}(\boldsymbol{x}) \cdot \|\mathcal{A}_{\mathrm{MM}}(\boldsymbol{z})\|_1}{\Gamma_{\boldsymbol{z}}\left(\mathbf{w}_{\mathcal{A}_{\mathrm{MM}}(\boldsymbol{z})}\right)} \right)^2 \right\} , \tag{15}$$

is ω-small at scale $1/2m$ w.r.t. the function

$$\omega\left(L_0, l, m, \delta, \frac{1}{2m}\right) = \left(\frac{2em}{-L_0}\right)^{-2L_0} . \tag{16}$$

10. The derivation of this result needs a slightly more complicated version of ω–smallness of the luckiness function $L_{\mathrm{margin}}(\mathbf{w}, \boldsymbol{z}) := -\gamma_{\boldsymbol{z}}(\mathbf{w})$ (see Shawe-Taylor et al. (1998) for details).

Proof First we note that by a slight refinement of a theorem of Makovoz (1996) (see Corollary 29 in Appendix A.8) we know that for any $\boldsymbol{z} \in \mathcal{Z}^m$ there exists a weight vector $\widehat{\mathbf{w}} = \sum_{i=1}^{m} \widehat{\alpha}_i \boldsymbol{\phi}(x_i)$ such that

$$\left\| \widehat{\mathbf{w}} - \mathbf{w}_{\mathcal{A}_{\mathrm{MM}}(\boldsymbol{z})} \right\|^2 \leq \Gamma_{\boldsymbol{z}}^2 \left(\mathbf{w}_{\mathcal{A}_{\mathrm{MM}}(\boldsymbol{z})} \right) \tag{17}$$

and $\widehat{\boldsymbol{\alpha}} \in \mathbb{R}^m$ has no more than $-L(\mathcal{A}_{\mathrm{MM}}, \boldsymbol{z})$ non-zero components. Although only $\mathbf{w}_{\mathcal{A}_{\mathrm{MM}}(\boldsymbol{z})}$ is of unit length, we show in Theorem 30 in Appendix A.8 that (17) implies

$$\left\langle \mathbf{w}_{\mathcal{A}_{\mathrm{MM}}(\boldsymbol{z})}, \frac{\widehat{\mathbf{w}}}{\|\widehat{\mathbf{w}}\|} \right\rangle \geq \sqrt{1 - \Gamma_{\boldsymbol{z}}^2 \left(\mathbf{w}_{\mathcal{A}_{\mathrm{MM}}(\boldsymbol{z})} \right)}\,.$$

Using equation (10) of Herbrich and Graepel (2001) this implies that $\Gamma_{\boldsymbol{z}}(\widehat{\mathbf{w}}) > 0$, that is, $\widehat{\mathbf{w}}$ correctly classifies $\boldsymbol{z} \in \mathcal{Z}^m$. Consider a fixed double sample $\boldsymbol{z} \in \mathcal{Z}^{2m}$ and let $L_0 := L(\mathcal{A}_{\mathrm{MM}}, (z_1, \ldots, z_m))$. By virtue of Definition 6 and the above argument we only need to consider permutations $\pi_{\boldsymbol{i}}$ such that there exists a weight vector $\widehat{\mathbf{w}} = \sum_{j=1}^{m} \widehat{\alpha}_j \boldsymbol{\phi}(x_j)$ with no more than L_0 non-zero $\widehat{\alpha}_i$. As there are exactly $\binom{2m}{i}$ distinct choices of $i \in \{1, \ldots, L_0\}$ training examples from the $2m$ examples $\boldsymbol{z}$ there are no more than $(2em/L_0)^{L_0}$ different subsamples to be used in $\widehat{\mathbf{w}}$ (see Lemma 18). For each particular subsample $\overline{\boldsymbol{z}} \subseteq \boldsymbol{z}$ the weight vector $\widehat{\mathbf{w}}$ is a member of the class of linear classifiers in a L_0 (or less) dimensional space. Thus, from (10) it follows that for the given subsample $\overline{\boldsymbol{z}}$ there are no more $(2em/L_0)^{L_0}$ different dichotomies induced on the double sample $\boldsymbol{z} \in \mathcal{Z}^{2m}$. As this holds for any double sample, the theorem is proven. ■

There are several interesting features about this result. Firstly, observe that $\|\mathcal{A}_{\mathrm{MM}}(\boldsymbol{z})\|_1$ is a measure of sparsity of the solution found by the maximum margin algorithm which, in the present case, is combined with margin. Note that for normalised data, i.e. $\|\boldsymbol{\phi}(\cdot)\| = \text{constant}$, the two notion of margins coincide, i.e. $\Gamma_{\boldsymbol{z}}(\mathbf{w}) = \gamma_{\boldsymbol{z}}(\mathbf{w})$. Secondly, the quantity $\epsilon_i(\boldsymbol{x})$ can be considered as a measure of the distribution of the mapped data points in feature space. From the definition, for all $i \in \mathbb{N}$, $\epsilon_i(\boldsymbol{x}) \leq \epsilon_1(\boldsymbol{x}) \leq \max_{j \in \{1,\ldots,m\}} \|\boldsymbol{\phi}(x_j)\|$. It is worthwhile mentioning that bounding $\epsilon_i(\boldsymbol{x})$ from above by $\max_{j \in \{1,\ldots,m\}} \|\boldsymbol{\phi}(x_j)\|$ is too crude a step and will lead to a bound on the effective complexity $\lceil \log_2(\omega(L_0, l_{0-1}, m, \cdot, 1/2m)) \rceil$ which scales as $\Gamma_{\boldsymbol{z}}^{-4}\left(\mathbf{w}_{\mathcal{A}_{\mathrm{MM}}(\boldsymbol{z})}\right)$ as opposed to $\Gamma_{\boldsymbol{z}}^{-2}\left(\mathbf{w}_{\mathcal{A}_{\mathrm{MM}}(\boldsymbol{z})}\right)$ because $\|\mathcal{A}_{\mathrm{MM}}(\boldsymbol{z})\|_1 = \Gamma_{\boldsymbol{z}}^{-1}\left(\mathbf{w}_{\mathcal{A}_{\mathrm{MM}}(\boldsymbol{z})}\right)$ (Vapnik, 1995). Supposing that the two class-conditional probabilities $\mathbf{P}_{\mathsf{X}|\mathsf{Y}=y}$ are highly clustered, $\epsilon_2(\boldsymbol{x})$ will be very small. An extension of this reasoning is useful in the multi-class case; binary maximum margin classifiers are often used to solve multi-class problems (Vapnik, 1998; Weston and Watkins, 1999). There appears to be also a close relationship of $\epsilon_i(\boldsymbol{x})$ with the notion of kernel alignment recently introduced in (Cristianini et al., 2002). The computation of $\epsilon_i(\boldsymbol{x})$ seems closely related to some classical questions in computational geometry and there are some fast approximate algorithms (Feder and Greene, 1988; Har-Peled, 2001). Finally, one can use standard entropy number techniques to bound $\epsilon_i(\boldsymbol{x})$ in terms of eigenvalues of the inner product matrix or its centred variants. It is worth mentioning that although our aim was to study the maximum margin algorithm the above theorem actually holds for any algorithm whose solution can be represented as a linear combination of the data points. This includes for example the perceptron learning algorithm (Rosenblatt, 1958), Bayes point machines (Herbrich et al., 2001; Ruján and Marchand, 2000), relevance vector machines (Tipping, 2001) and the Fisher discriminant (Mika et al., 1999).

8. Conclusions

In this paper we have introduced a new theoretical framework to study the generalisation error of learning algorithms. In contrast to previous approaches, we considered specific learning algorithms rather than specific hypothesis spaces. We introduced the notion of algorithmic luckiness which allowed us to devise data dependent generalisation error bounds. Thus we were able to relate the compression framework of Littlestone and Warmuth with the VC framework. Furthermore, we presented a new bound for the maximum margin algorithm which not only exploits the margin but also the distribution of the *actual* training data in feature space. Perhaps the most appealing feature of our margin based bound is that it naturally combines the three factors considered important for generalisation with linear classifiers: margin, sparsity and the distribution of the data. Further research is concentrated on studying Bayesian algorithms and whether one can avoid the union bound argument using other techniques.

Acknowledgements

Most of this work was done during a visit of Bob Williamson to Microsoft Research in Cambridge. We would like to thank Chris Bishop for providing this opportunity. This work was also supported by the Australian Research Council. Furthermore, we would like to thank Olivier Bousquet and Andre Elisseeff for interesting discussions that motivated the current approach. We are greatly indebted to Thore Graepel, Petra Philips and the anonymous reviewers for useful suggestions.

Appendix A. Proofs

A.1 An Upper Bound on the Sum of Binomials

We will sometimes use the following upper bound on the sum of binomials.

Lemma 18 (Bound on the sum of binomials) *For any $m \in \mathbb{N}$ and $n \in \{1, \ldots, m\}$*

$$\sum_{i=0}^{n} \binom{m}{i} < \left(\frac{em}{n}\right)^n .$$

Proof The proof follows from the binomial theorem together with the inequality $1 + x < \exp(x)$ for $x \neq 0$. Noticing that $\left(\frac{m}{n}\right)^{n-i} \geq 1$ for all $i \in \{0, \ldots, n\}$ we have

$$\begin{aligned}
\sum_{i=0}^{n} \binom{m}{i} &\leq \sum_{i=0}^{n} \binom{m}{i} \left(\frac{m}{n}\right)^{n-i} = \left(\frac{m}{n}\right)^n \sum_{i=0}^{n} \binom{m}{i} \left(\frac{n}{m}\right)^i \\
&\leq \left(\frac{m}{n}\right)^n \sum_{i=0}^{m} \binom{m}{i} \left(\frac{n}{m}\right)^i = \left(\frac{m}{n}\right)^n \left(1 + \frac{n}{m}\right)^m \\
&< \left(\frac{m}{n}\right)^n \exp(n) = \left(\frac{em}{n}\right)^n .
\end{aligned}$$

■

A.2 Multiple Testing Lemma

The following result is essentially a union bound argument which allows us to combine several generalisation error bounds into a uniform statement. This result has its roots in classical statistical test theory and is known as the *Bonferroni lemma.*

Lemma 19 (Multiple testing) *Suppose we are given a set* $\{\Upsilon_1, \ldots, \Upsilon_s\}$ *of* s *measurable logical formulae* $\Upsilon_i : \mathcal{Z}^{(\infty)} \times (0,1] \to \{\text{true}, \text{false}\}$ *together with* s *positive numbers* $p_1, \ldots, p_s$ *which sum up to one. If, for some probability measure* P_{Z}*,*

$$\forall i \in \{1, \ldots, s\} : \forall \delta \in (0,1] : \quad \mathsf{P}_{\mathsf{Z}^m}\left(\Upsilon_i\left(\mathbf{Z}, \delta\right)\right) \geq 1 - \delta\,,$$

then, for the same probability measure P_{Z}*,*

$$\forall \delta \in (0,1] : \quad \mathsf{P}_{\mathsf{Z}^m}\left(\Upsilon_1\left(\mathbf{Z}, \delta p_1\right) \wedge \cdots \wedge \Upsilon_s\left(\mathbf{Z}, \delta p_s\right)\right) \geq 1 - \delta\,.$$

Proof The proof is a simple union bound argument. By definition

$$\begin{aligned}
\mathsf{P}_{\mathsf{Z}^m}\left(\Upsilon_1\left(\mathbf{Z}, \delta p_1\right) \wedge \cdots \wedge \Upsilon_s\left(\mathbf{Z}, \delta p_s\right)\right) &= 1 - \mathsf{P}_{\mathsf{Z}^m}\left(\neg\Upsilon_1\left(\mathbf{Z}, \delta p_1\right) \vee \cdots \vee \neg\Upsilon_s\left(\mathbf{Z}, \delta p_s\right)\right) \\
&\geq 1 - \sum_{i=1}^{s} \mathsf{P}_{\mathsf{Z}^m}\left(\neg\Upsilon_i\left(\mathbf{Z}, \delta p_i\right)\right) \\
&> 1 - \sum_{i=1}^{s} \delta p_i = 1 - \delta \sum_{i=1}^{s} p_i = 1 - \delta\,.
\end{aligned}$$

■

A.3 Basic Lemma for Hypothesis Spaces

In this subsection we present the proof of two basic lemmas originally developed by Vapnik and Chervonenkis (1971). In contrast to their result we do not restrict the ghost sample to be of the same size as the training sample.

Lemma 20 (Basic lemma for consistent classifiers) *For all binary losses* $l : \mathcal{Y} \times \mathcal{Y} \to \{0,1\}$*, all probability measures* P_{Z}*, all hypothesis spaces* $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$*, all measurable formulae* $\Upsilon : \mathcal{Z}^{(\infty)} \to \{\text{true}, \text{false}\}$ *and all* $n > m$*, if* $\varepsilon\,(n-m) \geq 2$ *we have*

$$\begin{aligned}
&\mathsf{P}_{\mathsf{Z}^m}\left(\exists h \in \mathcal{H} : \left(R_l\left[h\right] > \varepsilon\right) \wedge \left(\widehat{R}_l\left[h, \mathbf{Z}\right] = 0\right) \wedge \Upsilon\left(\mathbf{Z}\right)\right) < \\
&\quad 2 \cdot \mathsf{P}_{\mathsf{Z}^n}\left(\exists h \in \mathcal{H} : \left(\widehat{R}_l\left[h, \mathbf{Z}_{[1:m]}\right] = 0\right) \wedge \left(\widehat{R}_l\left[h, \mathbf{Z}_{[(m+1):n]}\right] \geq \frac{\varepsilon}{2}\right) \wedge \Upsilon\left(\mathbf{Z}_{[1:m]}\right)\right)\,.
\end{aligned}$$

Proof Given a sample $\boldsymbol{z} \in \mathcal{Z}^m$ let $H\left(\boldsymbol{z}\right) \in \mathcal{H}$ be such that $\left(R_l\left[H\left(\boldsymbol{z}\right)\right] > \varepsilon\right) \wedge \left(\widehat{R}_l\left[H\left(\boldsymbol{z}\right), \boldsymbol{z}\right] = 0\right)$ if such an hypothesis exists or any $h \in \mathcal{H}$ otherwise. Let us introduce

the following shorthand notations where $\boldsymbol{z} \in \mathcal{Z}^m$ and $\tilde{\boldsymbol{z}} \in \mathcal{Z}^{n-m}$

$$\begin{aligned} Q_1(\boldsymbol{z}\tilde{\boldsymbol{z}}) &\equiv \widehat{R}_l[H(\boldsymbol{z}), \tilde{\boldsymbol{z}}] \geq \frac{\varepsilon}{2}, \qquad Q_2(\boldsymbol{z}) \equiv \left(\widehat{R}_l[H(\boldsymbol{z}), \boldsymbol{z}] = 0\right) \wedge \Upsilon(\boldsymbol{z}), \\ Q_3(\boldsymbol{z}) &\equiv R_l[H(\boldsymbol{z})] > \varepsilon. \end{aligned}$$

Then, it holds that

$$\begin{aligned} \mathbf{P}_{\mathsf{Z}^n}\left(Q_1(\mathbf{Z}) \wedge Q_2\left(\mathbf{Z}_{[1:m]}\right)\right) &\geq \mathbf{P}_{\mathsf{Z}^n}\left(Q_1(\mathbf{Z}) \wedge Q_2\left(\mathbf{Z}_{[1:m]}\right) \wedge Q_3\left(\mathbf{Z}_{[1:m]}\right)\right) \\ &= \mathbf{P}_{\mathsf{Z}^n|Q_2(\mathbf{Z}_{[1:m]})\wedge Q_3(\mathbf{Z}_{[1:m]})}(Q_1(\mathbf{Z}))\, \mathbf{P}_{\mathsf{Z}^n}\left(Q_2\left(\mathbf{Z}_{[1:m]}\right) \wedge Q_3\left(\mathbf{Z}_{[1:m]}\right)\right) \\ &= \mathbf{E}_{\mathsf{Z}_1^m}\left[\mathbb{I}_{Q_2(\mathbf{Z}_1)\wedge Q_3(\mathbf{Z}_1)} \mathbf{P}_{\mathsf{Z}_2^{n-m}|\mathsf{Z}_1^m=\boldsymbol{z}_1}(Q_1(\boldsymbol{z}_1\mathbf{Z}_2))\right]. \end{aligned}$$

Observe that by the conditioning we know that $R_l[H(\boldsymbol{z})] > \varepsilon$ whenever we have to evaluate the probability of $Q_1(\boldsymbol{z}_1\boldsymbol{z}_2)$ over the random draw of $\boldsymbol{z}_2 \in \mathcal{Z}^{n-m}$. Now, this probability is the probability that a binomially distributed random variable with an expectation of more than ε is greater than or equal to $\frac{\varepsilon(n-m)}{2} \geq 1$ which is equivalent to $\varepsilon(n-m) \geq 2$. Hence, this quantity satisfies

$$\mathbf{P}_{\mathsf{Z}_2^{n-m}|\mathsf{Z}_1^m=\boldsymbol{z}_1}(Q_1(\boldsymbol{z}_1\mathbf{Z}_2)) \geq 1 - (1-\varepsilon)^{n-m} > 1 - \exp(-\varepsilon(n-m)) > \frac{1}{2},$$

where we have used the assumption that $\varepsilon(n-m) \geq 2$. As a consequence we have shown

$$\begin{aligned} &\mathbf{P}_{\mathsf{Z}^n}\left(\exists h \in \mathcal{H} : \left(\widehat{R}_l\left[h, \mathbf{Z}_{[1:m]}\right] = 0\right) \wedge \left(\widehat{R}_l\left[h, \mathbf{Z}_{[(m+1):n]}\right] \geq \frac{\varepsilon}{2}\right) \wedge \Upsilon\left(\mathbf{Z}_{[1:m]}\right)\right) \\ &\quad = \mathbf{P}_{\mathsf{Z}^n}\left(Q_1(\mathbf{Z}) \wedge Q_2\left(\mathbf{Z}_{[1:m]}\right)\right) > \frac{1}{2}\mathbf{P}_{\mathsf{Z}^m}(Q_2(\mathbf{Z}) \wedge Q_3(\mathbf{Z})) \\ &\quad = \frac{1}{2}\mathbf{P}_{\mathsf{Z}^m}\left(\exists h \in \mathcal{H} : (R_l[h] > \varepsilon) \wedge \left(\widehat{R}_l[h, \mathbf{Z}] = 0\right) \wedge \Upsilon(\mathbf{Z})\right). \end{aligned}$$

■

Lemma 21 (General basic lemma) *For all bounded loss functions $l : \mathcal{Y} \times \mathcal{Y} \to [0,1]$, all probability measures $\mathbf{P}_{\mathsf{Z}}$, all hypothesis spaces $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, all measurable formulae $\Upsilon : \mathcal{Z}^{(\infty)} \to \{\text{true}, \text{false}\}$ and all $n > m$, if $\varepsilon^2(n-m) > 2$ we have*

$$\begin{aligned} &\mathbf{P}_{\mathsf{Z}^m}\left(\left(\sup_{h\in\mathcal{H}} R_l[h] - \widehat{R}_l[h, \mathbf{Z}] > \varepsilon\right) \wedge \Upsilon(\mathbf{Z})\right) < \\ &\quad 2 \cdot \mathbf{P}_{\mathsf{Z}^n}\left(\left(\sup_{h\in\mathcal{H}} \widehat{R}_l\left[h, \mathbf{Z}_{[(m+1):n]}\right] - \widehat{R}_l\left[h, \mathbf{Z}_{[1:m]}\right] > \frac{\varepsilon}{2}\right) \wedge \Upsilon\left(\mathbf{Z}_{[1:m]}\right)\right). \end{aligned}$$

Proof Given a sample $\boldsymbol{z} \in \mathcal{Z}^m$ let $H(\boldsymbol{z}) \in \mathcal{H}$ be given by

$$H(\boldsymbol{z}) := \operatorname*{argmax}_{h\in\mathcal{H}} R_l[h] - \widehat{R}_l[h, \boldsymbol{z}].$$

Let us introduce the following shorthand notations, where $\boldsymbol{z} \in \mathcal{Z}^m$ and $\tilde{\boldsymbol{z}} \in \mathcal{Z}^{n-m}$

$$\begin{aligned} Q_1(\boldsymbol{z}\tilde{\boldsymbol{z}}) &\equiv \left(\widehat{R}_l[H(\boldsymbol{z}), \tilde{\boldsymbol{z}}] - \widehat{R}_l[H(\boldsymbol{z}), \boldsymbol{z}] > \frac{\varepsilon}{2}\right) \wedge \Upsilon(\boldsymbol{z}), \\ Q_2(\boldsymbol{z}) &\equiv \left(R_l[H(\boldsymbol{z})] - \widehat{R}_l[H(\boldsymbol{z}), \boldsymbol{z}] > \varepsilon\right) \wedge \Upsilon(\boldsymbol{z}), \\ Q_3(\boldsymbol{z}\tilde{\boldsymbol{z}}) &\equiv \left(R_l[H(\boldsymbol{z})] - \widehat{R}_l[H(\boldsymbol{z}), \tilde{\boldsymbol{z}}] < \frac{\varepsilon}{2}\right). \end{aligned}$$

Since $Q_2(\boldsymbol{z}) \wedge Q_3(\boldsymbol{z}\tilde{\boldsymbol{z}}) \Rightarrow Q_1(\boldsymbol{z}\tilde{\boldsymbol{z}})$ we know that

$$\begin{aligned} \mathbf{P}_{\mathbf{Z}^n}(Q_1(\mathbf{Z})) &\geq \mathbf{P}_{\mathbf{Z}^n}\left(Q_2\left(\mathbf{Z}_{[1:m]}\right) \wedge Q_3(\mathbf{Z})\right) = \mathbf{P}_{\mathbf{Z}^n|Q_2\left(\mathbf{Z}_{[1:m]}\right)}(Q_3(\mathbf{Z}))\, \mathbf{P}_{\mathbf{Z}^n}\left(Q_2\left(\mathbf{Z}_{[1:m]}\right)\right) \\ &= \mathbf{E}_{\mathbf{Z}_1^m}\left[\mathbb{I}_{Q_2(\mathbf{Z}_1)} \mathbf{P}_{\mathbf{Z}_2^{n-m}|\mathbf{Z}_1^m = \boldsymbol{z}_1}(Q_3(\boldsymbol{z}_1\mathbf{Z}_2))\right]. \end{aligned}$$

Now, the probability $\mathbf{P}_{\mathbf{Z}_2^{n-m}|\mathbf{Z}_1^m=\boldsymbol{z}_1}(Q_3(\boldsymbol{z}_1\mathbf{Z}_2))$ is the probability that the mean of $n-m$ random variables taking values in $[0,1]$ is no more than $\frac{\varepsilon}{2}$ smaller than their common expectation $R_l[H(\boldsymbol{z}_1)]$. According to Hoeffding's inequality (Hoeffding, 1963) this probability is bounded from below by $1 - \exp\left(-\frac{\varepsilon^2(n-m)}{2}\right)$. Thus

$$\mathbf{P}_{\mathbf{Z}_2^{n-m}|\mathbf{Z}_1^m=\boldsymbol{z}_1}(Q_3(\boldsymbol{z}_1\mathbf{Z}_2)) \geq 1 - \exp\left(-\frac{\varepsilon^2(n-m)}{2}\right) > 1 - \exp(-1) > \frac{1}{2},$$

where we used the assumption $\varepsilon^2(n-m) > 2$. In summary, we have

$$\begin{aligned} &\mathbf{P}_{\mathbf{Z}^n}\left(\left(\sup_{h\in\mathcal{H}} \widehat{R}_l\left[h, \mathbf{Z}_{[(m+1):n]}\right] - \widehat{R}_l\left[h, \mathbf{Z}_{[1:m]}\right] > \frac{\varepsilon}{2}\right) \wedge \Upsilon\left(\mathbf{Z}_{[1:m]}\right)\right) \\ &= \mathbf{P}_{\mathbf{Z}^n}(Q_1(\mathbf{Z})) > \frac{1}{2}\mathbf{P}_{\mathbf{Z}^m}(Q_2(\mathbf{Z})) \\ &= \frac{1}{2}\mathbf{P}_{\mathbf{Z}^m}\left(\left(\sup_{h\in\mathcal{H}} R_l[h] - \widehat{R}_l[h, \mathbf{Z}] > \varepsilon\right) \wedge \Upsilon(\mathbf{Z})\right). \end{aligned}$$

Note that for the special case of $n = 2m$ we obtain the basic lemma in its standard form. ■

A.4 Basic Lemma for Learning Algorithms

In this subsection we prove two modified versions of the basic lemmas of the previous section. Our extension makes effective use of the learning algorithm used since we only consider the prediction error of the hypotheses learned using a fixed learning algorithm.

Lemma 22 (Basic lemma for consistent algorithms) *For all binary losses $l : \mathcal{Y}\times\mathcal{Y} \to \{0,1\}$, all probability measures $\mathbf{P}_{\mathbf{Z}}$, all algorithms $\mathcal{A} : \mathcal{Z}^{(\infty)} \to \mathcal{Y}^{\mathcal{X}}$, all measurable formulae $\Upsilon : \mathcal{Z}^{(\infty)} \to \{\text{true}, \text{false}\}$ and all $n > m$, if $\varepsilon(n-m) \geq 2$,*

$$\begin{aligned} &\mathbf{P}_{\mathbf{Z}^m}\left(\left(R_l[\mathcal{A}(\mathbf{Z})] > \varepsilon\right) \wedge \left(\widehat{R}_l[\mathcal{A}(\mathbf{Z}), \mathbf{Z}] = 0\right) \wedge \Upsilon(\mathbf{Z})\right) < \\ &2 \cdot \mathbf{P}_{\mathbf{Z}^n}\left(\left(\widehat{R}_l\left[\mathcal{A}\left(\mathbf{Z}_{[1:m]}\right), \mathbf{Z}_{[(m+1):n]}\right] \geq \frac{\varepsilon}{2}\right) \wedge \left(\widehat{R}_l\left[\mathcal{A}\left(\mathbf{Z}_{[1:m]}\right), \mathbf{Z}_{[1:m]}\right] = 0\right) \wedge \Upsilon\left(\mathbf{Z}_{[1:m]}\right)\right). \end{aligned}$$

Proof Let us introduce the following shorthand notations where $\boldsymbol{z} \in \mathcal{Z}^m$ and $\tilde{\boldsymbol{z}} \in \mathcal{Z}^{n-m}$

$$\begin{aligned} Q_1(\boldsymbol{z}\tilde{\boldsymbol{z}}) &\equiv \widehat{R}_l[\mathcal{A}(\boldsymbol{z}), \tilde{\boldsymbol{z}}] \geq \frac{\varepsilon}{2}, \quad Q_2(\boldsymbol{z}) \equiv \left(\widehat{R}_l[\mathcal{A}(\boldsymbol{z}), \boldsymbol{z}] = 0\right) \wedge \Upsilon(\boldsymbol{z}), \\ Q_3(\boldsymbol{z}) &\equiv R_l[\mathcal{A}(\boldsymbol{z})] > \varepsilon. \end{aligned}$$

By simple probability theory we know that

$$\begin{aligned} \mathbf{P}_{\mathbf{Z}^n}\left(Q_1(\mathbf{Z}) \wedge Q_2\left(\mathbf{Z}_{[1:m]}\right)\right) &\geq \mathbf{P}_{\mathbf{Z}^n}\left(Q_1(\mathbf{Z}) \wedge Q_2\left(\mathbf{Z}_{[1:m]}\right) \wedge Q_3\left(\mathbf{Z}_{[1:m]}\right)\right) \\ &= \mathbf{P}_{\mathbf{Z}^n|Q_2\left(\mathbf{Z}_{[1:m]}\right)\wedge Q_3\left(\mathbf{Z}_{[1:m]}\right)}\left(Q_1(\mathbf{Z})\right) \mathbf{P}_{\mathbf{Z}^n}\left(Q_2\left(\mathbf{Z}_{[1:m]}\right) \wedge Q_3\left(\mathbf{Z}_{[1:m]}\right)\right) \\ &= \mathbf{E}_{\mathbf{Z}_1^m}\left[\mathbb{I}_{Q_2(\mathbf{Z}_1)\wedge Q_3(\mathbf{Z}_1)} \mathbf{P}_{\mathbf{Z}_2^{n-m}|\mathbf{Z}_1^m=\boldsymbol{z}_1}\left(Q_1(\boldsymbol{z}_1\mathbf{Z}_2)\right)\right]. \end{aligned}$$

Observe that by the conditioning we know that $R_l[\mathcal{A}(\boldsymbol{z}_1)] > \varepsilon$ whenever we have to evaluate the probability of $Q_1(\boldsymbol{z}_1\boldsymbol{z}_2)$ over the random draw of $\boldsymbol{z}_2 \in \mathcal{Z}^{n-m}$. Now this probability is the probability that a binomially distributed random variable with an expectation of more than ε is greater than or equal to $\frac{\varepsilon(n-m)}{2} \geq 1$ which is equivalent to $\varepsilon(n-m) \geq 2$. Hence, this quantity is bounded from below by

$$\mathbf{P}_{\mathbf{Z}_2^{n-m}|\mathbf{Z}_1^m=\boldsymbol{z}_1}\left(Q_1(\boldsymbol{z}_1\mathbf{Z}_2)\right) \geq 1 - (1-\varepsilon)^{n-m} > 1 - \exp(-\varepsilon(n-m)) > \frac{1}{2},$$

where we have used the assumption that $\varepsilon(n-m) \geq 2$. As a consequence we have shown

$$\mathbf{P}_{\mathbf{Z}^n}\left(Q_1(\mathbf{Z}) \wedge Q_2\left(\mathbf{Z}_{[1:m]}\right)\right) > \frac{1}{2}\mathbf{P}_{\mathbf{Z}^m}\left(Q_2(\mathbf{Z}) \wedge Q_3(\mathbf{Z})\right).$$

■

Lemma 23 (General basic lemma for learning algorithms) *For all bounded loss functions $l : \mathcal{Y} \times \mathcal{Y} \to [0,1]$, all probability measures $\mathbf{P}_{\mathbf{Z}}$, all algorithms $\mathcal{A} : \mathcal{Z}^{(\infty)} \to \mathcal{Y}^{\mathcal{X}}$, all measurable formulae $\Upsilon : \mathcal{Z}^{(\infty)} \to \{\text{true}, \text{false}\}$ and all $n > m$, if $\varepsilon^2(n-m) > 2$,*

$$\begin{aligned} &\mathbf{P}_{\mathbf{Z}^m}\left(\left(R_l[\mathcal{A}(\mathbf{Z})] - \widehat{R}_l[\mathcal{A}(\mathbf{Z}), \mathbf{Z}] > \varepsilon\right) \wedge \Upsilon(\mathbf{Z})\right) < \\ &\quad 2 \cdot \mathbf{P}_{\mathbf{Z}^n}\left(\left(\widehat{R}_l\left[\mathcal{A}\left(\mathbf{Z}_{[1:m]}\right), \mathbf{Z}_{[(m+1):n]}\right] - \widehat{R}_l\left[\mathcal{A}\left(\mathbf{Z}_{[1:m]}\right), \mathbf{Z}_{[1:m]}\right] > \frac{\varepsilon}{2}\right) \wedge \Upsilon\left(\mathbf{Z}_{[1:m]}\right)\right). \end{aligned}$$

Proof Let us introduce the following shorthand notations where $\boldsymbol{z} \in \mathcal{Z}^m$ and $\tilde{\boldsymbol{z}} \in \mathcal{Z}^{n-m}$

$$\begin{aligned} Q_1(\boldsymbol{z}\tilde{\boldsymbol{z}}) &\equiv \left(\widehat{R}_l[\mathcal{A}(\boldsymbol{z}), \tilde{\boldsymbol{z}}] - \widehat{R}_l[\mathcal{A}(\boldsymbol{z}), \boldsymbol{z}] > \frac{\varepsilon}{2}\right) \wedge \Upsilon(\boldsymbol{z}), \\ Q_2(\boldsymbol{z}) &\equiv \left(R_l[\mathcal{A}(\boldsymbol{z})] - \widehat{R}_l[\mathcal{A}(\boldsymbol{z}), \boldsymbol{z}] > \varepsilon\right) \wedge \Upsilon(\boldsymbol{z}), \\ Q_3(\boldsymbol{z}\tilde{\boldsymbol{z}}) &\equiv \left(R_l[\mathcal{A}(\boldsymbol{z})] - \widehat{R}_l[\mathcal{A}(\boldsymbol{z}), \tilde{\boldsymbol{z}}] < \frac{\varepsilon}{2}\right). \end{aligned}$$

Since $Q_2(\boldsymbol{z}) \wedge Q_3(\boldsymbol{z}\tilde{\boldsymbol{z}}) \Rightarrow Q_1(\boldsymbol{z}\tilde{\boldsymbol{z}})$ we know that

$$\begin{aligned} \mathbf{P}_{\mathbf{Z}^n}(Q_1(\mathbf{Z})) &\geq \mathbf{P}_{\mathbf{Z}^n}\left(Q_2\left(\mathbf{Z}_{[1:m]}\right) \wedge Q_3(\mathbf{Z})\right) = \mathbf{P}_{\mathbf{Z}^n|Q_2\left(\mathbf{Z}_{[1:m]}\right)}(Q_3(\mathbf{Z})) \mathbf{P}_{\mathbf{Z}^n}\left(Q_2\left(\mathbf{Z}_{[1:m]}\right)\right) \\ &= \mathbf{E}_{\mathbf{Z}_1^m}\left[\mathbb{I}_{Q_2(\mathbf{Z}_1)} \mathbf{P}_{\mathbf{Z}_2^{n-m}|\mathbf{Z}_1^m=\boldsymbol{z}_1}(Q_3(\boldsymbol{z}_1\mathbf{Z}_2))\right]. \end{aligned}$$

Now, the probability $\mathbf{P}_{\mathbf{Z}_2^{n-m}|\mathbf{Z}_1^m=\boldsymbol{z}_1}\left(Q_3\left(\boldsymbol{z}_1\mathbf{Z}_2\right)\right)$ is the probability that the mean of $n-m$ random variables taking values in $[0,1]$ is no more than $\frac{\varepsilon}{2}$ smaller than their common expectation $R_l\left[\mathcal{A}\left(\boldsymbol{z}_1\right)\right]$. According to Hoeffding's inequality (see Hoeffding, 1963) this probability is bounded from below by $1-\exp\left(-\frac{\varepsilon^2(n-m)}{2}\right)$. Thus

$$\mathbf{P}_{\mathbf{Z}_2^{n-m}|\mathbf{Z}_1^m=\boldsymbol{z}_1}\left(Q_3\left(\boldsymbol{z}_1\mathbf{Z}_2\right)\right) \geq 1-\exp\left(-\frac{\varepsilon^2\left(n-m\right)}{2}\right) > 1-\exp\left(-1\right) > \frac{1}{2},$$

where we used the assumption $\varepsilon^2\left(n-m\right)>2$. In summary, we have

$$\mathbf{P}_{\mathbf{Z}^n}\left(Q_1\left(\mathbf{Z}\right)\right) > \frac{1}{2}\mathbf{P}_{\mathbf{Z}^m}\left(Q_2\left(\mathbf{Z}\right)\right).$$

■

A.5 Reduction of general permutations to swapping permutations

In this section we prove a simple result on the reduction of all $(2m)!$ permutations to 2^m swapping permutations if an event is independent of the ordering within the first and the second m examples[11]. The theorem reads as follows.

Theorem 24 (General permutations to swappings) *For any $m\in\mathbb{N}$, consider a logical formula $\Upsilon:\mathcal{Z}^{2m}\to\{\text{true},\text{false}\}$ with the property*

$$\forall\boldsymbol{z}\in\mathcal{Z}^{2m}:\forall\boldsymbol{i}_1\in I_m:\forall\boldsymbol{i}_2\in I_m:\qquad \Upsilon\left(\boldsymbol{z}\right)=\Upsilon\left(\Pi_{\boldsymbol{i}_1}\left(\boldsymbol{z}_{[1:m]}\right)\Pi_{\boldsymbol{i}_2}\left(\boldsymbol{z}_{[(m+1):2m]}\right)\right). \tag{18}$$

Then there exists a non-zero measure $\mathbf{P}_{\mathbf{I}}$ over I_{2m} given by (20) such that

$$\forall\boldsymbol{z}\in\mathcal{Z}^{2m}:\qquad \mathbf{P}_{\mathbf{I}|\mathbf{Z}^{2m}=\boldsymbol{z}}\left(\Upsilon\left(\Pi_{\mathbf{I}}\left(\boldsymbol{z}\right)\right)\right)=\frac{1}{2^m}\sum_{\boldsymbol{s}\in\{0,1\}^m}\mathbb{I}_{\Upsilon\left(\Sigma_{\boldsymbol{s}}\left(\boldsymbol{z}\right)\right)},$$

that is the probability of a permutation such that $\Upsilon\left(\Pi_{\boldsymbol{i}}\left(\boldsymbol{z}\right)\right)$ is true can be computed by counting the number of swappings $\Sigma_{\boldsymbol{s}}$ such that $\Upsilon\left(\Sigma_{\boldsymbol{s}}\left(\boldsymbol{z}\right)\right)$ is true.

Proof In the course of the proof we shall use the shorthand notation $\#\left(\boldsymbol{i}\right)$ to denote the number of swappings from the first m to the second m examples induced by $\pi_{\boldsymbol{i}}$; i.e.

$$\forall\boldsymbol{i}\in I_{2m}:\qquad \#\left(\boldsymbol{i}\right):=\left|\{j\in\{1,\ldots,m\}\mid\pi_{\boldsymbol{i}}(j)>m\}\right|. \tag{19}$$

In the case of a swapping permutation $\sigma_{\boldsymbol{s}}$, $\#\left(\boldsymbol{s}\right):=\sum_{i=1}^m s_i$. For $k=0,\ldots m$ let $J_k\subset I_{2m}$ and $S_k\subset\{0,1\}^m$ be the set of parameters for general and swapping permutations respectively, which swap exactly k examples from the first m to the second m examples; i.e.

$$J_k:=\{\boldsymbol{i}\in I_{2m}\mid\#\left(\boldsymbol{i}\right)=k\},\qquad S_k:=\{\boldsymbol{s}\in\{0,1\}^m\mid\#\left(\boldsymbol{s}\right)=k\}.$$

11. The importance of this result can best be seen by looking at its application in the proofs of Theorems 8 and 9: by virtue of Theorem 24 we can use all permutations $(2m)!$ for construction of the cover, yet resort to a simple counting argument on swappings for the single tail bounds. This would not be possible if we use the 2^m swapping permutations from the beginning (see Anthony and Shawe-Taylor (1993) for a similar approach).

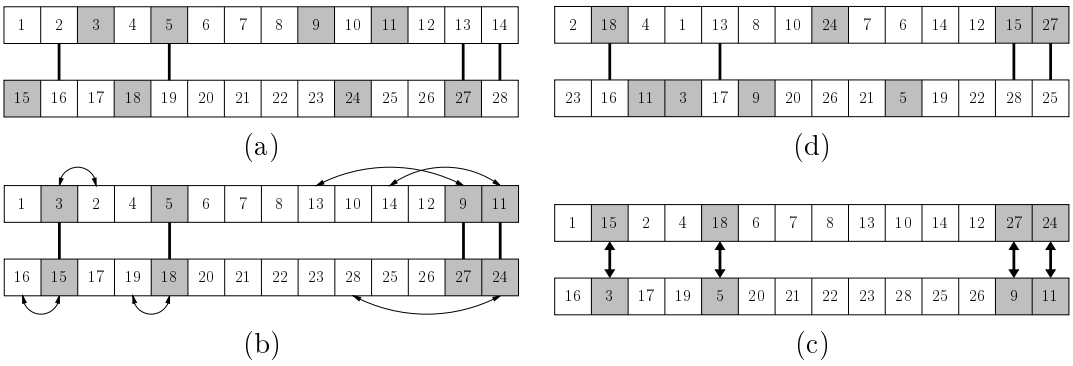

Figure 1: Suppose we want to permute the numbers $(1, \ldots, 28)$ (a) to obtain the permutation $(2, 18, \ldots, 15, 27, 23, 16, \ldots, 28, 25)$ (d), only using the swapping permutation $\sigma_{\mathbf{s}}$ with $\mathbf{s} = (0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1)$ (the vertical black lines). Since we know that any permutation *within* the first and the second $m = 14$ indices does not change the value of the function Υ we proceed as follows: having identified the 4 indices which are swapped from the first to the second half (denoted by shaded boxes), i.e. $(3, 5, 9, 11)$ and $(15, 18, 24, 27)$, we permute $(1, \ldots, 14)$ and $(15, \ldots, 28)$ such that the four numbers are at the 2nd, 5th, 13th and 14th position (see the curved arrows in (b)). Then we apply the swapping permutation $\sigma_{\mathbf{s}}$ (which might change the value of Υ) (see (c)). Finally we permute the first and second half again to obtain the permutation shown in (d).

Observe that $|J_k| = \left(\binom{m}{k}(m!)\right)^2$ because there are $\binom{m}{k}$ choices of k distinct indices from $\{1, \ldots, m\}$ and $\{m+1, \ldots, 2m\}$ (making the $\left(\binom{m}{k}\right)^2$ term) and any of the $m!$ permutations of the two half samples of size m (before swapping) leads to a new permutation. Furthermore we know that $|S_k| = \binom{m}{k}$ because this is the number of ways to choose k distinct indices from the set $\{1, \ldots, m\}$. Using the above notation consider the distribution $\mathbf{P}_{\mathsf{I}}^{\text{swap}}$:

$$\mathbf{P}_{\mathsf{I}}^{\text{swap}}(\boldsymbol{i}) := \frac{\binom{m}{\#(\boldsymbol{i})}}{2^m} \cdot \frac{1}{\left|J_{\#(\boldsymbol{i})}\right|}. \tag{20}$$

In other words, $\mathbf{P}_{\mathsf{I}}^{\text{swap}}$ is a distribution over all permutations $\pi_{\boldsymbol{i}}$ which is uniform for a given number $\#(\boldsymbol{i})$ of swappings but is not uniform overall. To see that this is a valid probability measure observe

$$\sum_{\boldsymbol{i} \in I_{2m}} \mathbf{P}_{\mathsf{I}}^{\text{swap}}(\boldsymbol{i}) = \sum_{k=0}^{m} \sum_{\boldsymbol{i} \in J_k} \mathbf{P}_{\mathsf{I}}^{\text{swap}}(\boldsymbol{i}) = \sum_{k=0}^{m} \sum_{\boldsymbol{i} \in J_k} \frac{\binom{m}{k}}{2^m} \cdot \frac{1}{|J_k|} = \frac{1}{2^m} \sum_{k=0}^{m} \binom{m}{k} \cdot \frac{|J_k|}{|J_k|} = 1\,.$$

Consider an arbitrary but fixed $\boldsymbol{z} \in \mathcal{Z}^{2m}$. For a fixed value k of $\#(\boldsymbol{i})$ we will now show that J_k can be subdivided into $|S_k|$ non-overlapping subsets $J_{k,1}, \ldots, J_{k,|S_k|}$ of size $\frac{|J_k|}{|S_k|}$ using $S_k = \left\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_{|S_k|}\right\}$ such that

$$\forall j \in \{1, \ldots, |S_k|\} : \forall \boldsymbol{i} \in J_{k,j} : \qquad \Upsilon\left(\Pi_{\boldsymbol{i}}(\boldsymbol{z})\right) = \Upsilon\left(\Sigma_{\boldsymbol{s}_j}(\boldsymbol{z})\right)\,. \tag{21}$$

In order to prove this we demonstrate that *any* permutation $\pi_{\boldsymbol{i}}$ for $\boldsymbol{i} \in J_k$ can be constructed by *any* swapping permutation $\Sigma_{\boldsymbol{s}}$ for $\boldsymbol{s} \in S_k$ applied to $(1, \ldots, 2m)$. The result follows because $\frac{|J_k|}{|S_k|} = \binom{m}{k} (m!)^2$ is an integer. Consider a binary vector $\boldsymbol{s} \in S_k$ and let $j_1, \ldots, j_k \in \{1, \ldots, m\}$ be the indices at which $s_{j_1} = \cdots = s_{j_k} = 1$. Now we will separately permute $(1, \ldots, m)$ and $(m+1, \ldots, 2m)$ such that the k indices in $\{i_{m+1}, \ldots, i_{2m}\}$ which are less than or equal to m are at the positions $j_1, \ldots, j_k$ and the k indices in $\{i_1, \ldots, i_m\}$ which are greater than m are at the positions $j_1 + m, \ldots, j_k + m$. Note that these permutations do not change the function Υ due to (18). Now apply the swapping permutation $\sigma_{\boldsymbol{s}}$. Then we can find two permutations of the resulting half samples (which, again, do not change the function Υ by (18)) such that the result of all these steps is the permutation $\pi_{\boldsymbol{i}}$ (see Figure 1).

By the above argument we can therefore decompose $\mathbf{P}^{\text{swap}}_{\mathbf{I}|\mathbf{Z}^{2m}=\boldsymbol{z}} (\Upsilon (\Pi_{\mathbf{I}} (\boldsymbol{z})))$ as follows:

$$
\begin{aligned}
\mathbf{P}^{\text{swap}}_{\mathbf{I}|\mathbf{Z}^{2m}=\boldsymbol{z}} (\Upsilon (\Pi_{\mathbf{I}} (\boldsymbol{z}))) &= \sum_{\boldsymbol{i} \in I_{2m}} \mathbb{I}_{\Upsilon(\Pi_{\boldsymbol{i}}(\boldsymbol{z}))} \mathbf{P}^{\text{swap}}_{\mathbf{I}} (\boldsymbol{i}) \\
&= \sum_{k=0}^{m} \frac{\binom{m}{k}}{2^m} \cdot \frac{1}{|J_k|} \sum_{\boldsymbol{i} \in J_k} \mathbb{I}_{\Upsilon(\Pi_{\boldsymbol{i}}(\boldsymbol{z}))} \\
&= \sum_{k=0}^{m} \frac{\binom{m}{k}}{2^m} \cdot \frac{1}{|J_k|} \sum_{j=1}^{|S_k|} \sum_{\boldsymbol{i} \in J_{k,j}} \mathbb{I}_{\Upsilon(\Pi_{\boldsymbol{i}}(\boldsymbol{z}))} \\
&= \sum_{k=0}^{m} \frac{\binom{m}{k}}{2^m} \cdot \frac{1}{|J_k|} \sum_{\boldsymbol{s} \in S_k} \frac{|J_k|}{|S_k|} \mathbb{I}_{\Upsilon(\Sigma_{\boldsymbol{s}}(\boldsymbol{z}))} \\
&= \frac{1}{2^m} \sum_{k=0}^{m} \sum_{\boldsymbol{s} \in S_k} \mathbb{I}_{\Upsilon(\Sigma_{\boldsymbol{s}}(\boldsymbol{z}))} \\
&= \frac{1}{2^m} \sum_{\boldsymbol{s} \in \{0,1\}^m} \mathbb{I}_{\Upsilon(\Sigma_{\boldsymbol{s}}(\boldsymbol{z}))} ,
\end{aligned}
$$

where we used (20) in the second line, (21) in the fourth line and $|S_k| = \binom{m}{k}$ in the fifth line. ■

A.6 Proof of Algorithmic Luckiness Theorem 8

This subsection contains the proof of our main result for binary loss functions. The ideas of the proof are similar to those used in the original luckiness framework. However, in the present case we make use of general covering numbers which allows us to devise a generalisation error bound for the agnostic case (Section A.7).

Proof In order to prove the theorem we bound the probability of training samples $\boldsymbol{z} \in \mathcal{Z}^m$ such that

1. the prediction error of $\mathcal{A}(\boldsymbol{z})$ is greater than ε, i.e. $J_1(\boldsymbol{z}) \equiv R_l[\mathcal{A}(\boldsymbol{z})] > \varepsilon$
2. the training error of $\mathcal{A}(\boldsymbol{z})$ on $\boldsymbol{z}$ is zero, i.e. $J_2(\boldsymbol{z}) \equiv \widehat{R}_l[\mathcal{A}(\boldsymbol{z}), \boldsymbol{z}] = 0$, and

3. the function $\omega\left(L\left(\mathcal{A},\boldsymbol{z}\right),l,m,\frac{\delta}{4},\frac{1}{2m}\right)$ is smaller than 2^d, i.e. $J_3\left(\boldsymbol{z}\right) \equiv \omega\left(L\left(\mathcal{A},\boldsymbol{z}\right),l,m,\frac{\delta}{4},\frac{1}{2m}\right) \leq 2^d$

by our preset value of δ. Using Lemma 22 and noticing that by assumption $\varepsilon m > 2$ we have that[12]

$$\mathbf{P}_{\mathbf{Z}^m}\left(J_1\left(\mathbf{Z}\right) \wedge J_2\left(\mathbf{Z}\right) \wedge J_3\left(\mathbf{Z}\right)\right) < 2 \cdot \mathbf{P}_{\mathbf{Z}^{2m}} \underbrace{\left(J_4\left(\mathbf{Z}\right) \wedge J_2\left(\mathbf{Z}_{[1:m]}\right) \wedge J_3\left(\mathbf{Z}_{[1:m]}\right)\right)}_{J\left(\mathbf{Z}\right)}, \tag{22}$$

where $J_4\left(\boldsymbol{z}\right) \equiv \widehat{R}_l\left[\mathcal{A}\left(\boldsymbol{z}_{[1:m]}\right),\boldsymbol{z}_{[(m+1):2m]}\right] \geq \frac{\varepsilon}{2}$. We now exploit the ω–smallness of L by considering the following proposition for $\boldsymbol{z} \in \mathcal{Z}^{2m}$

$$S\left(\boldsymbol{z}\right) \equiv \left|\mathcal{L}_l\left(\mathcal{H}_{\mathcal{A}}\left(L,\boldsymbol{z}\right)\right)\right| > \omega\left(L\left(\mathcal{A},\boldsymbol{z}_{[1:m]}\right),l,m,\frac{\delta}{4},\frac{1}{2m}\right).$$

Since for any double sample $\boldsymbol{z} \in \mathcal{Z}^{2m}$, $J\left(\boldsymbol{z}\right) \equiv \left(J\left(\boldsymbol{z}\right) \wedge S\left(\boldsymbol{z}\right)\right) \vee \left(J\left(\boldsymbol{z}\right) \wedge \overline{S}\left(\boldsymbol{z}\right)\right)$, it follows that

$$\begin{aligned} 2 \cdot \mathbf{P}_{\mathbf{Z}^{2m}}\left(J\left(\mathbf{Z}\right)\right) &= 2 \cdot \mathbf{P}_{\mathbf{Z}^{2m}}\left(J\left(\mathbf{Z}\right) \wedge S\left(\mathbf{Z}\right)\right) + 2 \cdot \mathbf{P}_{\mathbf{Z}^{2m}}\left(J\left(\mathbf{Z}\right) \wedge \overline{S}\left(\mathbf{Z}\right)\right) \\ &\leq \frac{\delta}{2} + 2 \cdot \mathbf{P}_{\mathbf{Z}^{2m}}\left(\underbrace{J_4\left(\mathbf{Z}\right) \wedge J_2\left(\mathbf{Z}_{[1:m]}\right)}_{J_{42}\left(\mathbf{Z}\right)} \wedge \underbrace{J_3\left(\mathbf{Z}_{[1:m]}\right) \wedge \overline{S}\left(\mathbf{Z}\right)}_{J_{3\overline{S}}\left(\mathbf{Z}\right)}\right), \end{aligned} \tag{23}$$

where we used Definition 7. We now resort to a technique known as *symmetrisation by permutation* (Kahane, 1968): Since we consider the product measure $\mathbf{P}_{\mathbf{Z}^{2m}}$ we know that any permutation of the double sample does not change the probability. Consequently, for any measure $\mathbf{P}_{\mathbf{I}}$ over $\boldsymbol{i} \in I_{2m}$ we have

$$\begin{aligned} \mathbf{P}_{\mathbf{Z}^{2m}}\left(J_{42}\left(\mathbf{Z}\right) \wedge J_{3\overline{S}}\left(\mathbf{Z}\right)\right) &= \mathbf{E}_{\mathbf{I}}\left[\mathbf{P}_{\mathbf{Z}^{2m}|\mathbf{I}=\boldsymbol{i}}\left(J_{42}\left(\Pi_{\boldsymbol{i}}\left(\mathbf{Z}\right)\right) \wedge J_{3\overline{S}}\left(\Pi_{\boldsymbol{i}}\left(\mathbf{Z}\right)\right)\right)\right] \\ &= \mathbf{E}_{\mathbf{Z}^{2m}}\left[\mathbf{P}_{\mathbf{I}|\mathbf{Z}^{2m}=\boldsymbol{z}}\left(J_{42}\left(\Pi_{\mathbf{I}}\left(\boldsymbol{z}\right)\right) \wedge J_{3\overline{S}}\left(\Pi_{\mathbf{I}}\left(\boldsymbol{z}\right)\right)\right)\right]. \end{aligned} \tag{24}$$

For a fixed double sample $\boldsymbol{z} \in \mathcal{Z}^{2m}$ let us arrange all permutations parameterised by $\boldsymbol{i}_j \in I_{2m}$ such that $L\left(\mathcal{A},\left(\Pi_{\boldsymbol{i}_{j+1}}\left(\boldsymbol{z}\right)\right)_{[1:m]}\right) \leq L\left(\mathcal{A},\left(\Pi_{\boldsymbol{i}_j}\left(\boldsymbol{z}\right)\right)_{[1:m]}\right)$ for all $j \in \{1,\ldots,(2m)!\}$. Then, let

$$H_j\left(\boldsymbol{z}\right) := \left\{\mathcal{A}\left(\Pi_{\boldsymbol{i}_k}\left(\boldsymbol{z}\right)\right)_{[1:m]} \mid k \in \{1,\ldots,j\}\right\} \subseteq \mathcal{Y}^{\mathcal{X}}.$$

Note that $\mathcal{N}\left(\frac{1}{2m},\mathcal{L}_l\left(H_{j+1}\left(\boldsymbol{z}\right)\right),\rho_1^{\boldsymbol{z}}\right) \geq \mathcal{N}\left(\frac{1}{2m},\mathcal{L}_l\left(H_j\left(\boldsymbol{z}\right)\right),\rho_1^{\boldsymbol{z}}\right)$ and let $j^* \in \{1,\ldots,(2m)!\}$ be such that

$$\mathcal{N}\left(\frac{1}{2m},\mathcal{L}_l\left(H_{j^*+1}\left(\boldsymbol{z}\right)\right),\rho_1^{\boldsymbol{z}}\right) > 2^d \quad \text{but} \quad \mathcal{N}\left(\frac{1}{2m},\mathcal{L}_l\left(H_{j^*}\left(\boldsymbol{z}\right)\right),\rho_1^{\boldsymbol{z}}\right) \leq 2^d.$$

Then $J_{3\overline{S}}\left(\Pi_{\mathbf{i}_j}\left(\boldsymbol{z}\right)\right)$ is only true if $j \leq j^*$. As a consequence, let $\widehat{H}\left(\boldsymbol{z}\right)$ be a minimal cover of $H_{j^*}\left(\boldsymbol{z}\right)$ at scale $\frac{1}{2m}$ w.r.t. $\rho_1^{\boldsymbol{z}}$, i.e. it contains a minimal set of hypotheses that incur all

12. Note that we consider double samples $\boldsymbol{z} \in \mathcal{Z}^{2m}$ on the right hand side of this inequality.

different zero-one loss patterns achieved on the double sample $\boldsymbol{z}$. Note that by definition of j^* we know that $\left|\widehat{H}\left(\boldsymbol{z}\right)\right| \leq 2^d$. Hence, whenever $J_{42}\left(\Pi_{\boldsymbol{i}}\left(\boldsymbol{z}\right)\right) \wedge J_{3\overline{S}}\left(\Pi_{\boldsymbol{i}}\left(\boldsymbol{z}\right)\right)$ is true for a permutation $\pi_{\boldsymbol{i}}$ then

$$\exists h \in \widehat{H}\left(\boldsymbol{z}\right) : \left(\widehat{R}_l\left[h, \left(\Pi_{\boldsymbol{i}}\left(\boldsymbol{z}\right)\right)_{[1:m]}\right] = 0\right) \wedge \left(\widehat{R}_l\left[h, \left(\Pi_{\boldsymbol{i}}\left(\boldsymbol{z}\right)\right)_{[(m+1):2m]}\right] \geq \frac{\varepsilon}{2}\right) .$$

Consequently, we can use the union bound to obtain

$$\begin{aligned}
&\mathbf{P}_{\mathbf{I}|\mathsf{Z}^{2m}=\boldsymbol{z}}\left(J_{42}\left(\Pi_{\mathbf{I}}\left(\boldsymbol{z}\right)\right) \wedge J_{3\overline{S}}\left(\Pi_{\mathbf{I}}\left(\boldsymbol{z}\right)\right)\right) \\
&\quad \leq \mathbf{P}_{\mathbf{I}|\mathsf{Z}^{2m}=\boldsymbol{z}}\left(\exists h \in \widehat{H}\left(\boldsymbol{z}\right) : \left(\widehat{R}_l\left[h, \left(\Pi_{\mathbf{I}}\left(\boldsymbol{z}\right)\right)_{[1:m]}\right] = 0\right) \wedge \left(\widehat{R}_l\left[h, \left(\Pi_{\mathbf{I}}\left(\boldsymbol{z}\right)\right)_{[(m+1):2m]}\right] \geq \frac{\varepsilon}{2}\right)\right) \\
&\quad \leq \sum_{h \in \widehat{H}(\boldsymbol{z})} \mathbf{P}_{\mathbf{I}|\mathsf{Z}^{2m}=\boldsymbol{z}}\left(\left(\widehat{R}_l\left[h, \left(\Pi_{\mathbf{I}}\left(\boldsymbol{z}\right)\right)_{[1:m]}\right] = 0\right) \wedge \left(\widehat{R}_l\left[h, \left(\Pi_{\mathbf{I}}\left(\boldsymbol{z}\right)\right)_{[(m+1):2m]}\right] \geq \frac{\varepsilon}{2}\right)\right) .
\end{aligned}$$

Now we will choose $\mathbf{P}_{\mathbf{I}} = \mathbf{P}_{\mathbf{I}}^{\text{swap}}$ as given by (20). From Theorem 24 we obtain

$$\begin{aligned}
&\mathbf{P}_{\mathbf{I}|\mathsf{Z}^{2m}=\boldsymbol{z}}^{\text{swap}}\left(J_{42}\left(\Pi_{\mathbf{I}}\left(\boldsymbol{z}\right)\right) \wedge J_{3\overline{S}}\left(\Pi_{\mathbf{I}}\left(\boldsymbol{z}\right)\right)\right) \\
&\qquad \leq \sum_{h \in \widehat{H}(\boldsymbol{z})} \frac{1}{2^m} \sum_{\boldsymbol{s} \in \{0,1\}^m} \mathbb{I}_{\left(\widehat{R}_l\left[h, \left(\Sigma_{\boldsymbol{s}}\left(\boldsymbol{z}\right)\right)_{[1:m]}\right] = 0\right) \wedge \left(\widehat{R}_l\left[h, \left(\Sigma_{\boldsymbol{s}}\left(\boldsymbol{z}\right)\right)_{[(m+1):2m]}\right] \geq \frac{\varepsilon}{2}\right)} .
\end{aligned}$$

For each fixed $h \in \widehat{H}\left(\boldsymbol{z}\right)$ the maximum number of swappings that satisfy the requirement stated (the argument to the indicator function) is given by $2^{m-\frac{m\varepsilon}{2}}$ because whenever we swap one of the at least $\frac{m\varepsilon}{2}$ examples that incur a mistake into the first half of the double sample we violate the condition of zero training error. Hence setting $\varepsilon = \frac{2}{m}\left(d + 2 + \ln\left(\frac{1}{\delta}\right)\right)$, for any $\boldsymbol{z} \in \mathcal{Z}^{2m}$,

$$\mathbf{P}_{\mathbf{I}|\mathsf{Z}^{2m}=\boldsymbol{z}}^{\text{swap}}\left(J_{42}\left(\Pi_{\mathbf{I}}\left(\boldsymbol{z}\right)\right)\right) \leq 2^d \cdot 2^{-\frac{m\varepsilon}{2}} = 2^d \cdot 2^{-\frac{m}{2}\cdot\frac{2}{m}\left(d+2+\log_2\left(\frac{1}{\delta}\right)\right)} = \frac{\delta}{4} . \tag{25}$$

In summary, combining (22), (23), (24) and (25) we have shown that

$$\mathbf{P}_{\mathsf{Z}^m}\left(\left(R_l\left[\mathcal{A}\left(\mathbf{Z}\right)\right] > \varepsilon\right) \wedge \left(\widehat{R}_l\left[\mathcal{A}\left(\mathbf{Z}\right), \mathbf{Z}\right] = 0\right) \wedge \left(\omega\left(L\left(\mathcal{A}, \mathbf{Z}\right), l, m, \frac{\delta}{4}, \frac{1}{2m}\right) \leq 2^d\right)\right) < \delta .$$

■

A.7 Proof of Algorithmic Luckiness Theorem 9

Lemma 25 *Suppose we are given two vectors* $\mathbf{a} \in \mathbb{R}^{2m}$ *and* $\mathbf{b} \in \mathbb{R}^{2m}$ *such that*

$$\frac{1}{m}\sum_{i=1}^{m} a_i - a_{i+m} > \varepsilon + 2\delta \quad \textit{and} \quad \frac{1}{2m}\sum_{i=1}^{2m} |a_i - b_i| \leq \delta ,$$

for some positive numbers $\varepsilon, \delta \in \mathbb{R}^+$. *Then*

$$\frac{1}{m}\sum_{i=1}^{m} b_i - b_{i+m} > \varepsilon .$$

Proof The result follows by using $x \geq -|x|$ in the third line below

$$\begin{aligned}
\frac{1}{m}\sum_{i=1}^{m}(b_i - b_{i+m}) &= \frac{1}{m}\sum_{i=1}^{m}\left(b_i - b_{i+m} + \underbrace{a_i - a_i + a_{i+m} - a_{i+m}}_{=0}\right) \\
&= \frac{1}{m}\left(\sum_{i=1}^{m}(a_i - a_{i+m}) + \left(\sum_{i=1}^{m}(b_i - a_i)\right) + \left(\sum_{i=1}^{m}(a_{i+m} - b_{i+m})\right)\right) \\
&\geq \frac{1}{m}\sum_{i=1}^{m}(a_i - a_{i+m}) - \left(\frac{1}{m}\sum_{i=1}^{m}|b_i - a_i| + \frac{1}{m}\sum_{i=1}^{m}|a_{i+m} - b_{i+m}|\right) \\
&= \underbrace{\frac{1}{m}\sum_{i=1}^{m}(a_i - a_{i+m})}_{>\varepsilon+2\delta} - \underbrace{\frac{1}{m}\sum_{i=1}^{2m}|b_i - a_i|}_{\leq 2\delta} > \varepsilon\,.
\end{aligned}$$

■

We now prove the main theorem for real-valued function classes.

Proof In order to prove the theorem we bound the probability of training samples $\boldsymbol{z} \in \mathcal{Z}^m$ such that

1. the prediction error of $\mathcal{A}(\boldsymbol{z})$ is more than ε greater than $\widehat{R}_l[\mathcal{A}(\boldsymbol{z}), \boldsymbol{z}]$, i.e.
$$J_1(\boldsymbol{z}) \equiv R_l[\mathcal{A}(\boldsymbol{z})] - \widehat{R}_l[\mathcal{A}(\boldsymbol{z}), \boldsymbol{z}] > \varepsilon$$
2. the function $\omega\left(L(\mathcal{A}, \boldsymbol{z}), l, m, \frac{\delta}{4}, \tau\right)$ is smaller than 2^d, i.e.
$$J_2(\boldsymbol{z}) \equiv \omega\left(L(\mathcal{A}, \boldsymbol{z}), l, m, \frac{\delta}{4}, \tau\right) \leq 2^d$$

by our preset value of δ. Using Lemma 23 and noticing that by assumption $\varepsilon^2 m > 2$ we have that

$$\mathbf{P}_{\mathbf{Z}^m}(J_1(\mathbf{Z}) \wedge J_2(\mathbf{Z})) < 2 \cdot \mathbf{P}_{\mathbf{Z}^{2m}} \underbrace{\left(J_3(\mathbf{Z}) \wedge J_2\left(\mathbf{Z}_{[1:m]}\right)\right)}_{J(\mathbf{Z})}, \tag{26}$$

where

$$J_3(\boldsymbol{z}) \equiv \widehat{R}_l\left[\mathcal{A}\left(\boldsymbol{z}_{[1:m]}\right), \boldsymbol{z}_{[(m+1):2m]}\right] - \widehat{R}_l\left[\mathcal{A}\left(\boldsymbol{z}_{[1:m]}\right), \boldsymbol{z}_{[1:m]}\right] > \frac{\varepsilon}{2}\,.$$

We now exploit the ω–smallness of L by considering the proposition for $\boldsymbol{z} \in \mathcal{Z}^{2m}$

$$S(\boldsymbol{z}) \equiv \mathcal{N}\left(\tau, \mathcal{L}_l\left(\mathcal{H}_{\mathcal{A}}(L, \boldsymbol{z})\right), \rho_1^{\boldsymbol{z}}\right) > \omega\left(L\left(\mathcal{A}, \boldsymbol{z}_{[1:m]}\right), l, m, \frac{\delta}{4}, \tau\right)\,.$$

Since for any double sample $\boldsymbol{z} \in \mathcal{Z}^{2m}$ we know $J(\boldsymbol{z}) \equiv (J(\boldsymbol{z}) \wedge S(\boldsymbol{z})) \vee \left(J(\boldsymbol{z}) \wedge \overline{S}(\boldsymbol{z})\right)$ it follows that

$$\begin{aligned}
2 \cdot \mathbf{P}_{\mathbf{Z}^{2m}}(J(\mathbf{Z})) &= 2 \cdot \mathbf{P}_{\mathbf{Z}^{2m}}(J(\mathbf{Z}) \wedge S(\mathbf{Z})) + 2 \cdot \mathbf{P}_{\mathbf{Z}^{2m}}\left(J(\mathbf{Z}) \wedge \overline{S}(\mathbf{Z})\right) \\
&\leq \frac{\delta}{2} + 2 \cdot \mathbf{P}_{\mathbf{Z}^{2m}}\left(J_3(\mathbf{Z}) \wedge \underbrace{J_2\left(\mathbf{Z}_{[1:m]}\right) \wedge \overline{S}(\mathbf{Z})}_{J_{2\overline{S}}(\mathbf{Z})}\right)\,.
\end{aligned} \tag{27}$$

where we used Definition 7. We again make use of *symmetrisation by permutation* (Kahane, 1968): Since we consider the product measure $\mathbf{P}_{\mathbf{Z}^{2m}}$ we know that any permutation of the double sample does not change the probability. Consequently, for any measure $\mathbf{P}_{\mathbf{I}}$ over $\boldsymbol{i} \in I_{2m}$ we have

$$\begin{aligned}\mathbf{P}_{\mathbf{Z}^{2m}}\left(J_3\left(\mathbf{Z}\right) \wedge J_{2\overline{S}}\left(\mathbf{Z}\right)\right) &= \mathbf{E}_{\mathbf{I}}\left[\mathbf{P}_{\mathbf{Z}^{2m}|\mathbf{I}=\boldsymbol{i}}\left(J_3\left(\Pi_{\boldsymbol{i}}\left(\mathbf{Z}\right)\right) \wedge J_{2\overline{S}}\left(\Pi_{\boldsymbol{i}}\left(\mathbf{Z}\right)\right)\right)\right] \\ &= \mathbf{E}_{\mathbf{Z}^{2m}}\left[\mathbf{P}_{\mathbf{I}|\mathbf{Z}^{2m}=\boldsymbol{z}}\left(J_3\left(\Pi_{\mathbf{I}}\left(\boldsymbol{z}\right)\right) \wedge J_{2\overline{S}}\left(\Pi_{\mathbf{I}}\left(\boldsymbol{z}\right)\right)\right)\right]. \qquad (28)\end{aligned}$$

For a fixed double sample $\boldsymbol{z} \in \mathcal{Z}^{2m}$ let us arrange all permutations parameterised by $\boldsymbol{i}_j \in I_{2m}$ such that $L\left(\mathcal{A}, \left(\Pi_{\boldsymbol{i}_{j+1}}\left(\boldsymbol{z}\right)\right)_{[1:m]}\right) \leq L\left(\mathcal{A}, \left(\Pi_{\boldsymbol{i}_j}\left(\boldsymbol{z}\right)\right)_{[1:m]}\right)$ for all $j \in \{1, \ldots, (2m)!\}$. Let

$$H_j\left(\boldsymbol{z}\right) := \left\{\mathcal{A}\left(\Pi_{\boldsymbol{i}_k}\left(\boldsymbol{z}\right)\right)_{[1:m]} \mid k \in \{1, \ldots, j\}\right\} \subseteq \mathcal{Y}^{\mathcal{X}}.$$

Note that $\mathcal{N}\left(\tau, \mathcal{L}_l\left(H_{j+1}\left(\boldsymbol{z}\right)\right), \rho_1^{\boldsymbol{z}}\right) \geq \mathcal{N}\left(\tau, \mathcal{L}_l\left(H_j\left(\boldsymbol{z}\right)\right), \rho_1^{\boldsymbol{z}}\right)$. Let j^* be such that

$$\mathcal{N}\left(\tau, \mathcal{L}_l\left(H_{j^*+1}\left(\boldsymbol{z}\right)\right), \rho_1^{\boldsymbol{z}}\right) > 2^d \quad \text{but} \quad \mathcal{N}\left(\tau, \mathcal{L}_l\left(H_{j^*}\left(\boldsymbol{z}\right)\right), \rho_1^{\boldsymbol{z}}\right) \leq 2^d.$$

Then $J_{2\overline{S}}\left(\Pi_{\boldsymbol{i}_j}\left(\boldsymbol{z}\right)\right)$ is true only if $j \leq j^*$. Let $\widehat{H}\left(\boldsymbol{z}\right)$ be a minimal cover of $H_{j^*}\left(\boldsymbol{z}\right)$ at scale τ w.r.t. the metric

$$\rho^{\boldsymbol{z}}\left(h, \tilde{h}\right) := \frac{1}{2m} \sum_{(x_i, y_i) \in \boldsymbol{z}} \left| l\left(h\left(x_i\right), y_i\right) - l\left(\tilde{h}\left(x_i\right), y_i\right)\right|,$$

that is for every $h \in H_{j^*}\left(\boldsymbol{z}\right)$ there exists a $\widehat{h} \in \widehat{H}\left(\boldsymbol{z}\right)$ such that

$$\frac{1}{2m} \sum_{(x_i, y_i) \in \boldsymbol{z}} \left| l\left(h\left(x_i\right), y_i\right) - l\left(\widehat{h}\left(x_i\right), y_i\right)\right| \leq \tau. \qquad (29)$$

Whenever $J_{2\overline{S}}\left(\boldsymbol{z}\right)$ is true we know that $\left|\widehat{H}\left(\boldsymbol{z}\right)\right| \leq 2^d$. Hence, whenever $J_3\left(\Pi_{\boldsymbol{i}}\left(\boldsymbol{z}\right)\right) \wedge J_{2\overline{S}}\left(\Pi_{\boldsymbol{i}}\left(\boldsymbol{z}\right)\right)$ is true for a permutation $\pi_{\boldsymbol{i}}$ then

$$\exists h \in \widehat{H}\left(\boldsymbol{z}\right) : \widehat{R}_l\left[h, \left(\Pi_{\boldsymbol{i}}\left(\boldsymbol{z}\right)\right)_{[(m+1):2m]}\right] - \widehat{R}_l\left[h, \left(\Pi_{\boldsymbol{i}}\left(\boldsymbol{z}\right)\right)_{[1:m]}\right] > \frac{\varepsilon}{2} - 2\tau,$$

using (29) and Lemma 25. Thus we can use the union bound to obtain

$$\begin{aligned}&\mathbf{P}_{\mathbf{I}|\mathbf{Z}^{2m}=\boldsymbol{z}}\left(J_3\left(\Pi_{\mathbf{I}}\left(\boldsymbol{z}\right)\right)\right) \\ &\quad \leq \mathbf{P}_{\mathbf{I}|\mathbf{Z}^{2m}=\boldsymbol{z}}\left(\exists h \in \widehat{H}\left(\boldsymbol{z}\right) : \left(\widehat{R}_l\left[h, \left(\Pi_{\mathbf{I}}\left(\boldsymbol{z}\right)\right)_{[(m+1):2m]}\right] - \widehat{R}_l\left[h, \left(\Pi_{\mathbf{I}}\left(\boldsymbol{z}\right)\right)_{[1:m]}\right]\right) > \frac{\varepsilon}{2} - 2\tau\right) \\ &\quad \leq \sum_{h \in \widehat{H}(\boldsymbol{z})} \mathbf{P}_{\mathbf{I}|\mathbf{Z}^{2m}=\boldsymbol{z}}\left(\left(\widehat{R}_l\left[h, \left(\Pi_{\mathbf{I}}\left(\boldsymbol{z}\right)\right)_{[(m+1):2m]}\right] - \widehat{R}_l\left[h, \left(\Pi_{\mathbf{I}}\left(\boldsymbol{z}\right)\right)_{[1:m]}\right]\right) > \frac{\varepsilon}{2} - 2\tau\right)\end{aligned}$$

Now we will choose $\mathbf{P}_{\mathbf{I}} = \mathbf{P}_{\mathbf{I}}^{\text{swap}}$ as given by (20). From Theorem 24 we obtain that $\mathbf{P}_{\mathbf{I}|\mathbf{Z}^{2m}=\boldsymbol{z}}^{\text{swap}}\left(J_3\left(\Pi_{\mathbf{I}}\left(\boldsymbol{z}\right)\right)\right)$ is less than or equal to

$$\sum_{h \in \widehat{H}(\boldsymbol{z})} \underbrace{\frac{1}{2^m} \sum_{\boldsymbol{s} \in \{0,1\}^m} \mathbb{I}_{\left(\widehat{R}_l\left[h, \left(z_{\sigma_{\boldsymbol{s}}(m+1)}, \ldots, z_{\sigma_{\boldsymbol{s}}(2m)}\right)\right] - \widehat{R}_l\left[h, \left(z_{\sigma_{\boldsymbol{s}}(1)}, \ldots, z_{\sigma_{\boldsymbol{s}}(m)}\right)\right]\right) > \frac{\varepsilon}{2} - 2\tau}}_{\chi}.$$

For a fixed $h \in \widehat{H}(\boldsymbol{z})$ consider the m random variables

$$\mathsf{W}_i := l\left(h\left(x_{\sigma_{\mathbf{S}}(i+m)}\right), y_{\sigma_{\mathbf{S}}(i+m)}\right) - l\left(h\left(x_{\sigma_{\mathbf{S}}(i)}\right), y_{\sigma_{\mathbf{S}}(i)}\right), \quad i = 1, \ldots, m,$$

which are mutually independent with mean zero. Since $\mathsf{W}_i \in [-1, +1]$ we can use the one-sided Hoeffding's inequality (Hoeffding, 1963) to obtain

$$\chi = \mathbf{P}_{\mathsf{W}^m}\left(\frac{1}{m}\sum_{i=1}^{m}\mathsf{W}_i > \frac{\varepsilon}{2} - 2\tau\right) < \exp\left(-\frac{m\left(\frac{\varepsilon}{2} - 2\tau\right)^2}{2}\right) < 2^{-\frac{m(\varepsilon-4\tau)^2}{8}}.$$

Hence setting $\varepsilon = \sqrt{\frac{8}{m}\left(d + 2 + \log_2\left(\frac{1}{\delta}\right)\right)} + 4\tau$, for any $\boldsymbol{z} \in \mathcal{Z}^{2m}$,

$$\mathbf{P}^{\text{swap}}_{\mathbf{I}|\mathsf{Z}^{2m}=\boldsymbol{z}}\left(J_3\left(\Pi_{\mathbf{I}}(\boldsymbol{z})\right)\right) \leq 2^d \cdot 2^{-\frac{m(\varepsilon-4\tau)^2}{8}} = \frac{\delta}{4}. \tag{30}$$

In summary, combining (26), (27), (28) and (30) we have shown that

$$\mathbf{P}_{\mathsf{Z}^m}\left(\left(R_l\left[\mathcal{A}(\mathbf{Z})\right] - \widehat{R}_l\left[\mathcal{A}(\mathbf{Z}), \mathbf{Z}\right] > \varepsilon\right) \wedge \left(\omega\left(L(\mathcal{A}, \mathbf{Z}), l, m, \frac{\delta}{4}, \tau\right) \leq 2^d\right)\right) < \delta.$$

■

A.8 Auxiliary Results for the Proof of Theorem 17

In this section we present two theorems which will be needed in the proof of Theorem 17. The first theorem together with the first corollary is a refinement of a result proven by Makovoz (1996) which bounds the approximation error of sparse linear combination of functions. We will present the proof in terms of elements of ℓ_2^N, where N may be infinite (simply in order to align the notation with that used elsewhere in the present paper). The most appealing feature of this result is that the approximation error is related to the geometrical configuration of the basis vectors. To this end, we need the notion of entropy numbers.

Definition 26 (Entropy numbers) *Given a subset X of $\mathcal{K} \subseteq \ell_2^N$, the n–th* entropy number $\epsilon_n(X)$ *of X is defined as*

$$\epsilon_n(X) := \inf\{\epsilon > 0 \mid \mathcal{N}(\epsilon, X, \|\cdot - \cdot\|) \leq n\}.$$

In other words, $\epsilon_n(X)$ is the smallest radius such that X can be covered by not more than n balls.

Theorem 27 *Let $X := \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\} \subseteq \mathcal{K}$ be an arbitrary sequence of elements of $\mathcal{K} \subseteq \ell_2^N$. For every $\mathbf{w} \in \mathcal{K}$ of the form*

$$\mathbf{w} = \sum_{i=1}^{m}\alpha_i\mathbf{x}_i, \qquad \boldsymbol{\alpha} \geq \mathbf{0},$$

and for every even $n \in \mathbb{N}$, $n \leq m$, *there is a* $\widehat{\mathbf{w}} = \sum_{i=1}^{m} \widehat{\alpha}_i \mathbf{x}_i$ *with at most* n *non-zero coefficients* $\widehat{\alpha}_i > 0$ *for which*

$$\|\mathbf{w} - \widehat{\mathbf{w}}\| \leq \frac{\sqrt{2}\epsilon_{\frac{n}{2}}(X) \cdot \|\boldsymbol{\alpha}\|_1}{\sqrt{n}}.$$

The proof follows Makovoz' proof closely, but uses a more refined argument in order to improve on the constant.

Proof Lets define $p := \frac{n}{2} \Leftrightarrow n = 2p$. First we notice that we only need to consider the case that $\|\boldsymbol{\alpha}\|_1 = 1$. If it is not then for every even $n \in \mathbb{N}$ we can approximate $\frac{1}{\|\boldsymbol{\alpha}\|_1}\mathbf{w}$ with an n–sparse vector $\widehat{\mathbf{w}}$ to accuracy $\epsilon_p(X)/\sqrt{p}$, i.e.

$$\left\| \frac{1}{\|\boldsymbol{\alpha}\|_1}\mathbf{w} - \widehat{\mathbf{w}} \right\| \leq \frac{\epsilon_p(X)}{\sqrt{p}} \quad \Leftrightarrow \quad \|\mathbf{w} - \|\boldsymbol{\alpha}\|_1 \cdot \widehat{\mathbf{w}}\| \leq \frac{\epsilon_p(X) \cdot \|\boldsymbol{\alpha}\|_1}{\sqrt{p}}.$$

In the construction below we will allocate a fraction of the n terms available to us in the approximation $\widehat{\mathbf{w}}$. By definition, for a given even n and some $\epsilon > \epsilon_p(X)$, we can find p disjoint and nonempty subsets $X_1, \ldots, X_p$ such that X_j is covered w.r.t. $\|\cdot - \cdot\|$ at radius ϵ, that is,

$$\forall j \in \{1, \ldots, p\}: \qquad \exists \mathbf{c}_j \in \mathcal{K}: \forall \widetilde{\mathbf{x}} \in X_j: \ \|\mathbf{c}_j - \widetilde{\mathbf{x}}\| \leq \epsilon \tag{31}$$

and $\bigcup_{j=1}^{p} X_j = X$. Hence, we can decompose the set $\{1, \ldots, m\}$ as $\{1, \ldots, m\} = \bigcup_{j=1}^{p} I_j$ where the $I_j \subset \{1, \ldots, m\}$ are all disjoint and nonempty and the sets X_j are defined by

$$X_j := \{\mathbf{x}_i \mid i \in I_j\}.$$

Let $\mathbf{w}_j := \sum_{i \in I_j} \alpha_i \mathbf{x}_i$, and $C_j := \sum_{i \in I_j} \alpha_i$. We will approximate $\mathbf{w}_j$ by linear combinations $\widehat{\mathbf{w}}_j = \sum_{i \in I_j} \widehat{\alpha}_i \mathbf{x}_i$ with a small number of non-zero $\widehat{\alpha}_i$, respectively. The proof uses the probabilistic method (Alon et al., 1991). To this end, we assume that $\mathbf{w}_j$ is always approximated by $l_j := \lceil pC_j \rceil$ many randomly drawn points from the set I_j. The effect of such an allocation is that index sets I_j which contribute largely (in terms of the coefficients α_i) to the linear combination $\mathbf{w} = \sum_{j=1}^{p} \sum_{i \in I_j} \alpha_i \mathbf{x}_i$ are used more often in the random n–sparse approximation $\widehat{\mathbf{w}}$. Note that

$$l := \sum_{j=1}^{p} l_j \leq \sum_{j=1}^{p} (pC_j + 1) = p\left(\sum_{j=1}^{p} C_j\right) + p = 2p = n.$$

Within each subset I_j we also need to select which of the $|I_j|$ many $\mathbf{x}_i$, $i \in I_j$, are used. Hence, we define p groups of l_j, $j \in \{1, \ldots, p\}$, iid random variables $\mathsf{K}_1, \ldots, \mathsf{K}_p$ taking values in $I_1^{l_1}, \ldots, I_p^{l_p}$ and having the probability distributions

$$\forall j \in \{1, \ldots, p\}: \forall \mu \in \{1, \ldots, l_j\}: \forall i \in I_j: \qquad \mathsf{P}_{\mathsf{K}_{j,\mu}}(\mathsf{K}_{j,\mu} = i) := \frac{\alpha_i}{C_j}.$$

In a manner similar to the allocation policy for subsets I_j we ensure that points $\mathbf{x}_i$ with large coefficients α_i are more likely to appear in the n–sparse approximation $\widehat{\mathbf{w}}$. Thus, for

any given sample $(\boldsymbol{k}_1, \ldots, \boldsymbol{k}_p) \sim \mathsf{P}_{\mathsf{K}_1} \cdots \mathsf{P}_{\mathsf{K}_p}$ we define the n–sparse approximation $\widehat{\mathbf{w}}$ by

$$\widehat{\mathbf{w}}\left(\boldsymbol{k}_1, \ldots, \boldsymbol{k}_p\right) := \sum_{j=1}^{p} \widehat{\mathbf{w}}_j\left(\boldsymbol{k}_j\right) , \qquad \widehat{\mathbf{w}}_j\left(\boldsymbol{k}_j\right) := \frac{C_j}{l_j} \sum_{\mu=1}^{l_j} \mathbf{x}_{k_{j,\mu}} .$$

First we observe that for all $j \in \{1, \ldots, p\}$, $\mathsf{E}_{\mathsf{K}_j}\left[\widehat{\mathbf{w}}_j\left(\mathsf{K}_j\right)\right] = \mathbf{w}_j$ because,

$$\begin{aligned}
\mathsf{E}_{\mathsf{K}_j}\left[\widehat{\mathbf{w}}_j\left(\mathsf{K}_j\right)\right] &= \mathsf{E}_{\mathsf{K}_j}\left[\frac{C_j}{l_j} \sum_{\mu=1}^{l_j} \mathbf{x}_{\mathsf{K}_{j,\mu}}\right] \\
&= \frac{C_j}{l_j} \sum_{\mu=1}^{l_j} \mathsf{E}_{\mathsf{K}_{j,\mu}}\left[\mathbf{x}_{\mathsf{K}_{j,\mu}}\right] \\
&= \frac{C_j}{l_j} \sum_{\mu=1}^{l_j} \sum_{\nu \in I_j} \frac{\alpha_\nu}{C_j} \mathbf{x}_\nu \\
&= \frac{1}{l_j} \sum_{\mu=1}^{l_j} \mathbf{w}_j = \mathbf{w}_j , \qquad (32)
\end{aligned}$$

where the second line is a consequence of the mutual of the independence $\mathsf{K}_{j,\mu}$, $\mu \in \{1, \ldots, l_j\}$. This implies that $\mathsf{E}_{\mathsf{K}_1 \cdots \mathsf{K}_p}\left[\widehat{\mathbf{w}}\left(\mathsf{K}_1, \ldots, \mathsf{K}_p\right)\right] = \mathbf{w}$. Note that for all $j \in \{1, \ldots, p\}$,

$$\begin{aligned}
\mathsf{var}_{\mathsf{K}_j}\left(\widehat{\mathbf{w}}_\mu\left(\mathsf{K}_j\right)\right) &= \mathsf{var}_{\mathsf{K}_j}\left(\frac{C_j}{l_j} \sum_{\mu=1}^{l_j} \mathbf{x}_{\mathsf{K}_{j,\mu}}\right) \\
&= \frac{C_j^2}{l_j^2} \sum_{\mu=1}^{l_j} \underbrace{\mathsf{var}_{\mathsf{K}_{j,\mu}}\left(\mathbf{x}_{\mathsf{K}_{j,\mu}}\right)}_{\leq \epsilon^2} \\
&\leq \frac{C_j^2 \epsilon^2}{l_j} \leq \frac{C_j \epsilon^2}{n} , \qquad (33)
\end{aligned}$$

because by construction $\mathbf{x}_{\mathsf{K}_j}$ only takes values in X_j which has by definition a radius[13] of ϵ. Combining (33) together with and (32), we can now bound the expected approximation

13. Note that by definition of variance and (31),

$$\begin{aligned}
\mathsf{var}_{\mathsf{K}_{j,\mu}}\left(\mathbf{x}_{\mathsf{K}_{j,\mu}}\right) &= \mathsf{var}_{\mathsf{K}_{j,\mu}}\left(\mathbf{x}_{\mathsf{K}_{j,\mu}} - \mathbf{c}_j\right) \\
&= \mathsf{E}_{\mathsf{K}_{j,\mu}}\left[\left\|\mathbf{x}_{\mathsf{K}_{j,\mu}} - \mathbf{c}_j\right\|^2\right] - \left\|\mathsf{E}_{\mathsf{K}_{j,\mu}}\left[\mathbf{x}_{\mathsf{K}_{j,\mu}} - \mathbf{c}_j\right]\right\|^2 \\
&\leq \mathsf{E}_{\mathsf{K}_{j,\mu}}\left[\left\|\mathbf{x}_{\mathsf{K}_{j,\mu}} - \mathbf{c}_j\right\|^2\right] = \mathsf{E}_{\mathsf{K}_{j,\mu}}\left[\epsilon^2\right] \leq \epsilon^2 .
\end{aligned}$$

error from above as follows

$$\begin{aligned}
\mathbf{E}_{\mathbf{K}_1 \cdots \mathbf{K}_p} \left\| \widehat{\mathbf{w}} \left(\mathbf{K}_1, \ldots, \mathbf{K}_p \right) - \mathbf{w} \right\|^2 &= \mathbf{var}_{\mathbf{K}_1 \cdots \mathbf{K}_p} \left(\widehat{\mathbf{w}} \left(\mathbf{K}_1, \ldots, \mathbf{K}_p \right) \right) \\
&= \mathbf{var}_{\mathbf{K}_1 \cdots \mathbf{K}_p} \left(\sum_{j=1}^{p} \widehat{\mathbf{w}}_j \left(\mathbf{K}_j \right) \right) \\
&= \sum_{j=1}^{p} \mathbf{var}_{\mathbf{K}_j} \left(\widehat{\mathbf{w}}_j \left(\mathbf{K}_j \right) \right) \\
&\leq \sum_{j=1}^{p} \frac{C_j \epsilon^2}{p} = \frac{\epsilon^2}{p} .
\end{aligned}$$

Since for any random draw $\boldsymbol{k}_1, \ldots, \boldsymbol{k}_p$, $\widehat{\mathbf{w}} \left(\boldsymbol{k}_1, \ldots, \boldsymbol{k}_p \right)$ is l–sparse, $l \leq n$, there must exists at least one weight vector $\widehat{\mathbf{w}}$ for which the approximation error is less than $\epsilon / \sqrt{p}$. The theorem is proved. ■

Corollary 28 *Let $X := \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\} \subseteq \mathcal{K}$ be an arbitrary sequence of elements of $\mathcal{K} \subseteq \ell_2^N$. For every $\mathbf{w} \in \mathcal{K}$ of the form*

$$\mathbf{w} = \sum_{i=1}^{m} \alpha_i \mathbf{x}_i , \qquad \left\| \boldsymbol{\alpha} \right\|_1 < \infty , \tag{34}$$

and for every $n = 4j$, $j \in \mathbb{N}$, $n \leq m$, there is a $\widehat{\mathbf{w}} = \sum_{i=1}^{m} \widehat{\alpha}_i \mathbf{x}_i$ with at most n non-zero coefficients $\widehat{\alpha}_i$ for which

$$\left\| \mathbf{w} - \widehat{\mathbf{w}} \right\| \leq \frac{2 \epsilon_{\frac{n}{4}} \left(X \right) \cdot \left\| \boldsymbol{\alpha} \right\|_1}{\sqrt{n}} . \tag{35}$$

Proof Define the two subsets $I_+ := \{ i \in \{1, \ldots, m\} \mid \alpha_i > 0 \}$ and $I_- := \{ i \in \{1, \ldots, m\} \mid \alpha_i < 0 \}$, the two vectors $\mathbf{w}_\pm := \sum_{i \in I_\pm} |\alpha_i| \, \mathbf{x}_i$, the two vectors $\boldsymbol{\alpha}_\pm := (a_i)_{i \in I_\pm}$, and the two sets $X_\pm := \{ \mathbf{x}_i \mid i \in I_\pm \}$. For any number $j \in \mathbb{N}$, by virtue of Theorem 27 we know that there exists a $2j$–sparse vector $\widehat{\mathbf{w}}_\pm$ such that

$$\left\| \mathbf{w}_\pm - \widehat{\mathbf{w}}_\pm \right\| \leq \frac{\epsilon_j \left(X_\pm \right) \left\| \boldsymbol{\alpha}_\pm \right\|_1}{\sqrt{j}} .$$

Note that $\epsilon_j \left(X_\pm \right) \leq \epsilon_j \left(X \right)$ because $X_\pm \subseteq X$. If we define $\widehat{\mathbf{w}} := \mathbf{w}_+ - \mathbf{w}_-$ then by application of the the triangle inequality in the third line

$$\begin{aligned}
\left\| \mathbf{w} - \widehat{\mathbf{w}} \right\| &= \left\| \left(\mathbf{w}_+ - \mathbf{w}_- \right) - \left(\widehat{\mathbf{w}}_+ - \widehat{\mathbf{w}}_- \right) \right\| \\
&= \left\| \left(\mathbf{w}_+ - \widehat{\mathbf{w}}_+ \right) - \left(\mathbf{w}_- - \widehat{\mathbf{w}}_- \right) \right\| \\
&\leq \left\| \mathbf{w}_+ - \widehat{\mathbf{w}}_+ \right\| + \left\| \mathbf{w}_- - \widehat{\mathbf{w}}_- \right\| \\
&\leq \frac{\epsilon_j \left(X_+ \right) \left\| \boldsymbol{\alpha}_+ \right\|_1 + \epsilon_j \left(X_- \right) \left\| \boldsymbol{\alpha}_- \right\|_1}{\sqrt{j}} \\
&\leq \frac{\epsilon_j \left(X \right) \left(\left\| \boldsymbol{\alpha}_+ \right\|_1 + \left\| \boldsymbol{\alpha}_- \right\|_1 \right)}{\sqrt{j}} \\
&= \frac{\epsilon_j \left(X \right) \left\| \boldsymbol{\alpha} \right\|_1}{\sqrt{j}} .
\end{aligned}$$

Noticing that $\widehat{\mathbf{w}}$ is a $4j = n$–sparse approximation of $\mathbf{w}$ proves the corollary. ■

Corollary 29 *Suppose we are given a training sample* $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{y}) \in (\mathcal{X} \times \{-1, +1\})^m$*, a feature map* $\boldsymbol{\phi} : \mathcal{X} \to \mathcal{K} \subseteq \ell_2^N$ *and a vector* $\boldsymbol{\alpha} \in \mathbb{R}^m$ *such that* $\mathbf{w} = \sum_{i=1}^m \alpha_i \boldsymbol{\phi}(x_i)$ *has a positive margin,* $\Gamma_{\boldsymbol{z}}(\mathbf{w}) \geq 0$*. If the natural number* $n = 4j$*,* $j \in \mathbb{N}$*, satisfies*

$$n > \frac{4 \cdot \epsilon^2_{\frac{n}{4}}(\{\boldsymbol{\phi}(x_1), \ldots, \boldsymbol{\phi}(x_m)\}) \cdot \|\boldsymbol{\alpha}\|_1^2}{\Gamma^2_{\boldsymbol{z}}(\mathbf{w})},$$

then there exists a weight vector $\widehat{\mathbf{w}}$ *with a representation*

$$\widehat{\mathbf{w}} = \sum_{i=1}^m \widehat{\alpha}_i \boldsymbol{\phi}(x_i)$$

such that at most n *of the coefficients* $\widehat{\alpha}_i$ *are non-zero and* $\|\mathbf{w} - \widehat{\mathbf{w}}\|^2 \leq \Gamma^2_{\boldsymbol{z}}(\mathbf{w})$*.*

Proof Observe that with the assumed choice of n we have

$$\frac{4 \cdot \epsilon^2_{\frac{n}{4}}(\{\boldsymbol{\phi}(x_1), \ldots, \boldsymbol{\phi}(x_m)\}) \cdot \|\boldsymbol{\alpha}\|_1^2}{n} < \Gamma^2_{\boldsymbol{z}}(\mathbf{w})$$

But Corollary 28 implies that for any n there exists $\widehat{\mathbf{w}} = \sum_{i=1}^m \widehat{\alpha}_i \boldsymbol{\phi}(x_i)$ with no more than n non-zero coefficients $\widehat{\alpha}_i$ such that

$$\|\mathbf{w} - \widehat{\mathbf{w}}\|^2 \leq \frac{4 \cdot \epsilon^2_{\frac{n}{4}}(\{\boldsymbol{\phi}(x_1), \ldots, \boldsymbol{\phi}(x_m)\}) \cdot \|\boldsymbol{\alpha}\|_1^2}{n} < \Gamma^2_{\boldsymbol{z}}(\mathbf{w}) \ .$$

■

Our second theorem lower bounds the inner product of two normalised vectors if we only know the distance between the two vectors, one of which is already normalised. More formally, this reads as follows.

Theorem 30 *Suppose* $\|\mathbf{w}\| = 1$ *and* $\|\mathbf{w} - \widehat{\mathbf{w}}\|^2 \leq c^2$*,* $c < 1$*. Then*

$$\left\langle \mathbf{w}, \frac{\widehat{\mathbf{w}}}{\|\widehat{\mathbf{w}}\|} \right\rangle \geq \sqrt{1 - c^2}\,.$$

Proof With no loss of generality consider the subspace spanned by $\mathbf{w}$ and $\widehat{\mathbf{w}}$. Let $\theta := \angle(\mathbf{w}, \widehat{\mathbf{w}})$. The worst $\widehat{\mathbf{w}}$ is such that the line ℓ passing through the origin and $\widehat{\mathbf{w}}$ is tangential (at the point denoted $\mathbf{v}$) to the circle of radius c centred at $\mathbf{w}$. The line $(\mathbf{v}, \mathbf{w}) \perp \ell \ \Rightarrow \ \sin(\theta) \leq c \Rightarrow \langle \mathbf{w}, \widehat{\mathbf{w}} / \|\widehat{\mathbf{w}}\| \rangle = \cos(\theta) \geq \sqrt{1 - c^2}$. ■

References

N. Alon, J. H. Spencer, and P. Erdös. *The Probabilistic Method*. John Wiley and Sons, 1991.

M. Anthony and P. Bartlett. *A Theory of Learning in Artificial Neural Networks*. Cambridge University Press, 1999.

M. Anthony and J. Shawe-Taylor. A result of Vapnik with applications. *Discrete Applied Mathematics*, 47:207–217, 1993.

A. Blumer, A. Ehrenfeucht, D. Haussler, and M. Warmuth. Learnability and the Vapnik-Chervonenkis Dimension. *Journal of the ACM*, 36(4):929–965, 1989.

B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In D. Haussler, editor, *Proceedings of the Annual Conference on Computational Learning Theory*, pages 144–152, Pittsburgh, PA, July 1992. ACM Press.

S. Boucheron, G. Lugosi, and P. Massart. A sharp concentration inequality with applications. *Random Structures and Algorithms*, 16(3):277–292, 2000.

O. Bousquet and A. Elisseeff. Algorithmic stability and generalization performance. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 196–202. MIT Press, 2001.

C. Cortes and V. Vapnik. Support vector networks. *Machine Learning*, 20:273–297, 1995.

N. Cristianini, J. Shawe-Taylor, A. Elisseeff, and J. Kandola. On kernel-target alignment. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 367–373, Cambridge, MA, 2002. MIT Press.

L. Devroye and G. Lugosi. *Combinatorial Methods in Density Estimation*. Springer, 2001.

T. Feder and D. H. Greene. Optimal algorithms for approximate clustering. In *Proceedings of the 20th Symposium on the Theory of Computing*, pages 434–444, 1988.

W. Feller. *An Introduction To Probability Theory and Its Application*, volume 2. John Wiley and Sons, New York, 1966.

S. Floyd and M. Warmuth. Sample compression, learnability, and the Vapnik Chervonenkis dimension. *Machine Learning*, 27:1–36, 1995.

T. Graepel, R. Herbrich, and J. Shawe-Taylor. Generalisation error bounds for sparse linear classifiers. In *Proceedings of the Annual Conference on Computational Learning Theory*, pages 298–303, 2000.

S. Har-Peled. Clustering motion. In *Proceedings of the 42nd Annual IEEE Symposium on Foundations of Computer Science*, pages 84–93, 2001.

R. Herbrich. *Learning Kernel Classifiers: Theory and Algorithms*. MIT Press, 2002.

R. Herbrich and T. Graepel. A PAC-Bayesian margin bound for linear classifiers: Why SVMs work. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 224–230, Cambridge, MA, 2001. MIT Press.

R. Herbrich, T. Graepel, and C. Campbell. Bayes point machines. *Journal of Machine Learning Research*, 1:245–279, 2001.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

J. P. Kahane. *Some Random Series of Functions*. Cambridge University Press, 1968.

M. J. Kearns and U. V. Vazirani. *An Introduction to Computational Learning Theory*. MIT Press, Cambridge, Massachusetts, 1994.

N. Littlestone and M. Warmuth. Relating data compression and learnability. Technical report, University of California Santa Cruz, 1986.

Y. Makovoz. Random approximants and neural networks. *Journal of Approximation Theory*, 85:98–109, 1996.

R. J. McEliece, E. R. Rodemich, H. Rumsey, and L. R. Welch. New upper bounds on the rate of a code via the Delsarte-MacWilliams inequalities. *IEEE Transactions on Information Theory*, 23(2):157–166, 1977.

S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K.-R. Müller. Fisher discriminant analysis with kernels. In Y.-H. Hu, J. Larsen, E. Wilson, and S. Douglas, editors, *Neural Networks for Signal Processing IX*, pages 41–48. IEEE, 1999.

F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958.

P. Ruján and M. Marchand. Computing the Bayes kernel classifier. In A. J. Smola, P. L. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 329–347, Cambridge, MA, 2000. MIT Press.

N. Sauer. On the density of families of sets. *Journal of Combinatorial Theory*, 13:145–147, 1972.

B. Schölkopf, R. Herbrich, and A. Smola. A generalized representer theorem. In *Proceedings of the Annual Conference on Computational Learning Theory*, pages 416–426, 2001.

J. Shawe-Taylor, P. L. Bartlett, R. C. Williamson, and M. Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, 44 (5):1926–1940, 1998.

M. Tipping. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244, 2001.

V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 1995. ISBN 0-387-94559-8.

V. Vapnik. *Statistical Learning Theory*. John Wiley and Sons, New York, 1998.

V. N. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer, Berlin, 1982.

V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–281, 1971.

V. N. Vapnik and A. Y. Chervonenkis. The necessary and sufficient conditions for consistency in the empirical risk minimization method. *Pattern Recognition and Image Analysis*, 1 (3):283–305, 1991.

J. Weston and C. Watkins. Multi-class support vector machines. In M. Verleysen, editor, *Proceedings ESANN*, Brussels, 1999. D Facto.