Structure discovery in PAC-Learning by Random Projections

Kaban, Ata; Reeve, Henry

DOI:
10.1007/s10994-024-06531-0

Machine Learning
https://doi.org/10.1007/s10994-024-06531-0

Structure discovery in PAC-learning by random projections

Ata Kabán[1] · **Henry Reeve**[2]

Received: 5 March 2023 / Revised: 30 October 2023 / Accepted: 14 February 2024

Abstract
High dimensional learning is data-hungry in general; however, many natural data sources and real-world learning problems posses some hidden low-complexity structure that permit effective learning from relatively small sample sizes. We are interested in the general question of how to discover and exploit such hidden benign traits when problem-specific prior knowledge is insufficient. In this work, we address this question through random projection's ability to expose structure. We study both compressive learning and high dimensional learning from this angle by introducing the notions of compressive distortion and compressive complexity. We give user-friendly PAC bounds in the agnostic setting that are formulated in terms of these quantities, and we show that our bounds can be tight when these quantities are small. We then instantiate these quantities in several examples of particular learning problems, demonstrating their ability to discover interpretable structural characteristics that make high dimensional instances of these problems solvable to good approximation in a random linear subspace. In the examples considered, these turn out to resemble some familiar benign traits such as the margin, the margin distribution, the intrinsic dimension, the spectral decay of the data covariance, or the norms of parameters—while our general notions of compressive distortion and compressive complexity serve to unify these, and may be used to discover benign structural traits for other PAC-learnable problems.

Keywords Random projection · Generalization · Structure discovery

Editors: Dino Ienco, Roberto Interdonato, Pascal Poncelet.

✉ Ata Kabán
A.Kaban@bham.ac.uk

Henry Reeve
Henry.Reeve@bristol.ac.uk

[1] School of Computer Science, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK

[2] Department of Mathematics, University of Bristol, Woodland Road, Bristol BS8 1UG, UK

Published online: 26 March 2024

1 Introduction

Many high dimensional learning problems require sample sizes that grow with the dimension of the data representation in an essential way in general. Examples include learning with scale-insensitive loss functions such as the 0–1 loss, learning on unbounded input or parameter domains (Mohri et al., 2012; Shalev-Shwartz & Ben-David, 2014), learning Lipschitz classifiers (Gottlieb & Kontorovich, 2014), metric learning (Verma & Branson, 2015), and others. A common approach to deal with these problems is to employ some form of regularisation constraints that reflect prior knowledge about the problem, when available. Indeed, natural data sources and real-world learning problems tend to possess some hidden low complexity structure, and these can permit effective learning from relatively small sample sizes in principle. However, knowing these structures in advance to devise appropriate learning algorithms can be a challenge.

In this work, we are interested in the general question of how to discover and exploit such hidden benign traits when problem-specific prior knowledge is insufficient, based on just a general-purpose low complexity conjecture.

We address this question through random projection's ability to expose structure—an ability previously studied in contexts as distinct as high dimensional phenomena (Bartl & Mendelson, 2021), geometric functional analysis (Liaw et al., 2017), and brain research (Papadimitriou & Vempala, 2019). Random projection (RP) is a simple, computationally efficient linear dimensionality reduction technique that preserves Euclidean structure with high probability. In machine learning, this can speed up computations at the price of a controlled loss of accuracy—this is generally referred to as compressive learning, in analogy with compressive sensing. Moreover, RP has a regularisation effect, and it has also been used as an analytic tool to better understand high dimensional learning in an early conference version of this work (Kabán, 2019).

The remainder of this section sets up the problem and gives a motivating example. In Sect. 2 we give simple PAC-bounds in the agnostic setting, both for compressive learning and for high dimensional learning. Our goal here is to work under minimal assumptions and isolate interpretable structural quantities that help gain intuitive insights into generalisation in high dimensional small sample situations. We term these as compressive distortion and compressive complexity in the compressed and uncompressed settings respectively, and we show that our bounds can be tight when these quantities are small.

In Sect. 3 we instantiate the above by bounding the problem-specific quantities that appear in these bounds for several widely-used model classes. These worked examples demonstrate how these quantities unearth structural characteristics that make these specific problems solvable to good approximation in a random linear subspace. In the examples considered, these turn out to take the form of some familiar benign traits such as the margin, the margin distribution, the intrinsic dimension, the spectral decay of the data covariance, or the norms of parameters—all of which remove dimensionality-dependence from error-guarantees in settings where such dependence is known to be essential in general. At the same time, our general notions of compressive distortion and compressive complexity serve to unify these characteristics, and may be used beyond the examples pursued here. We also show how one can use unlabelled data to estimate these general quantities when analytic bounds are infeasible, and this procedure recovers a form of consistency regularisation (Laine & Aila, 2017), which is a semi-supervised technique widely used in practice.

1.1 Problem setting

1.1.1 High dimensional learning

Let $\mathcal{X}_d \subset \mathbb{R}^d$ be an input domain, and $\mathcal{Y}$ the target domain—e.g. $\mathcal{Y} = \{-1, 1\}$ is classification, $\mathcal{Y} \subseteq \mathbb{R}$ in regression. We are interested in high dimensional problems, so d can be arbitrarily large.

Let $\mathcal{H}_d$ be a function class (hypothesis class) with elements $h : \mathcal{X}_d \to \mathcal{Y}$. The loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, \bar{\ell}]$ quantifies the mismatch between predictions and targets. Throughout this work we assume that the loss is bounded i.e. $\bar{\ell} < \infty$. This simplifying assumption is often made in algorithm-independent theoretical analyses, either by clipping the loss, or by working with bounded functions $h \in \mathcal{H}_d$ e.g. by constraining both the parameter and input spaces to bounded sets. Several examples may be found in (Rosasco et al., 2004). Boundedness is often natural too, since classification losses in use are typically surrogates for the 0–1 loss, which is bounded by $\bar{\ell} = 1$.

We are given a set of labelled examples $\mathcal{T}_N = \{(X_1, Y_1), \ldots, (X_N, Y_N)\}$ drawn i.i.d. from some unknown distribution $\mathbb{P}$ over $\mathcal{X}_d \times \mathcal{Y}$. The learning problem is to select a function from $\mathcal{H}_d$ with smallest generalisation error $E_{(X,Y)\sim\mathbb{P}}[\ell(h(X), Y)]$, using the sample $\mathcal{T}_N$.

Let $\mathcal{G}_d = \ell \circ \mathcal{H}_d = \{(x, y) \to g(x, y) = \ell(h(x), y) : h \in \mathcal{H}_d\}$ denote the loss class under study. Expectations with respect to (w.r.t.) the unknown data distribution $\mathbb{P}$, will be denoted by the shorthand $E[g] := E_{(X,Y)\sim\mathbb{P}}[g(X, Y)] = \int_{\mathcal{X}\times\mathcal{Y}} g d\mathbb{P}$. Sample averages, i.e. expectations w.r.t. the empirical measure $\hat{\mathbb{P}}_N$ defined by a sample $\mathcal{T}_N$ will be denoted as $\hat{E}_{\mathcal{T}_N}[g] := \hat{E}_{\mathcal{T}_N}[g(X, Y)] = \frac{1}{N}\sum_{n=1}^{N} g(X_n, Y_n) = \int_{\mathcal{X}\times\mathcal{Y}} g d\hat{\mathbb{P}}_N$, where $\hat{\mathbb{P}}_N = \frac{1}{N}\sum_{n=1}^{N} \delta_{X_n}$, and δ_X is the probability distribution concentrated at X. A best element of $\mathcal{H}$ is denoted by $h^* \in \arg\inf_{h\in\mathcal{H}_d} E[\ell \circ h]$, $g^* := \ell \circ h^*$; a sample error minimiser is $\hat{h} \in \arg\min_{h\in\mathcal{H}_d} \hat{E}_{\mathcal{T}}[\ell \circ h]$, and $\hat{g} := \ell \circ \hat{h}$.

1.1.2 Compressive learning

Let $k \leq d$ be integers, and $R \in \mathbb{R}^{k\times d}$ a random matrix with independent and identically distributed (i.i.d.) entries from a 0-mean $1/k$-variance distribution, chosen to satisfy the Johnson–Lindenstrauss (JL) property (Property 5.1). This is referred to as a random projection (RP) (Arriaga & Vempala, 1999; Matoušek, 2008). For instance, a random matrix with i.i.d. Gaussian entries is known to satisfy JL. For simplicity, throughout of this paper we will work with Gaussian RP, which serves as a simple dimensionality reduction method. While RP is not a projection in a strict linear-algebraic sense, the rows of R have approximately identical lengths and are approximately orthogonal to each other with high probability—hence the established nomenclature of "random projection".

We denote the compressed input domain by $\mathcal{X}_R \equiv R(\mathcal{X}) \subseteq \mathbb{R}^k$, and have analogous definitions, indexed by R, as follows. The compressed function class $\mathcal{H}_R$ contains functions of the form $h_R : \mathcal{X}_R \to \mathcal{Y}$. The learning algorithm receives the compressed training set, denoted $\mathcal{T}_R^N = \{(RX_n, Y_n)\}_{n=1}^N$, and selects a function from $\mathcal{H}_R$.

We denote a sample error minimiser in this reduced class by $\hat{h}_R \in \arg\inf_{h_R\in\mathcal{H}_R} \hat{E}_{\mathcal{T}_R^N}[\ell \circ h_R]$, where $\hat{E}_{\mathcal{T}_R^N}[\ell \circ h_R] = \frac{1}{N}\sum_{n=1}^{N} \ell(h_R(RX_n), Y_n)$ is the empirical error of the compressed

learning problem, and denote $\hat{g}_R := \ell \circ \hat{h}_R$. Likewise, $h_R^* \in \arg\inf_{h_R \in \mathcal{H}_R} E[\ell \circ h_R]$ denotes a best function in $\mathcal{H}_R$, $g_R^* := \ell \circ h_R^*$.

We are interested in the generalisation error of the compressed sample minimiser $\hat{h}_R$, that is $E_{(X,Y)\sim\mathbb{P}}[\ell(\hat{h}_R(RX), Y)]$, relative to the best $h^* \in \mathcal{H}_d$.

Let us end this introduction with an example that showcases the regularisation effect of RP, and demonstrates a failure of empirical risk minimisation (ERM) without regularisation. This will motivate our approach of introducing novel quantities in Sect. 2, and the instantiations of these quantities later in Sect. 3 may be regarded as a strategy to derive model-specific regularisers from the structure-preserving ability of RP. In our bounds, these quantities will be responsible for dimension-independence.

1.2 A motivating example

Random projection based dimensionality reduction is most commonly motivated by computational speed-up and storage savings, and these benefits may come at the expense of a slight deterioration of accuracy performance. But this is just part of the story. In this section we make the picture more complete by demonstrating a simple example to highlight that RP has a regularisation effect without of which ERM can actually fail.

Theorem 1 (ERM can be arbitrarily bad) *Let e_i be the i-th canonical basis vector, suppose the data distribution is uniform on the finite set $\mathcal{X} \times \mathcal{Y} := S \equiv \{(e_1 + e_i, 1), (-e_1 - e_i, -1) : i = 2, \ldots, d\}$, and let $\mathcal{T}_N$ be an i.i.d. sample of size N. Then,*

1. *There exists a classifier h_{bad} such that $\hat{E}_{(X,Y)\sim\mathcal{T}_N}[\mathbf{1}(h_{\text{bad}}^T XY \leq 0)] = 0$, but*

$$E_{X,Y}[\mathbf{1}(h_{\text{bad}}^T XY \leq 0)] \geq 1 - \frac{N}{d-1}.$$

2. *Let R be a $k \times d$ random projection matrix with i.i.d. sub-gaussian entries independent of $\mathcal{T}_N$, and $d \geq k \geq \lceil 16 \log \frac{4N}{\delta} \rceil$, where $\gamma > 0$ is the normalised margin of h^* in S. Given any $\delta \in (0,1)$, w.p. at least $1 - \delta$ the generalisation error of any compressive ERM, $\hat{h}_R \in \mathbb{R}^k$, is upper bounded as the following*

$$E_{X,Y}\left\{\mathbf{1}\left(\hat{h}_R^T RXY \leq 0\right)\right\} \leq \frac{2}{N}\left(k \log \frac{2eN}{k} + \log \frac{4}{\delta}\right)$$

The proof is given in Appendix Sect. 1. The construction exploits the fact that some ERM classifiers perform badly in small sample problems with large margin; in contrast, RP narrows the margin while keeping separability with high probability, so in this construction compressive ERM enjoys a dimension-free generalisation guarantee.

2 Error bounds for compressible problems

2.1 Learning with compressive ERM

We introduce the following definition, which later we use to bound the error of compressive ERM.

Definition 1 (Compressive distortion of a function) Given a function $g \in \mathcal{G}_d$, we define its compressive distortion as the following:

$$D_R(g) \equiv \inf_{g_R \in \mathcal{G}_R} E_{X,Y}|(g_R \circ R - g)(X,Y)|; \quad D_k(g) \equiv E_R[D_R(g)(X,Y)] \tag{1}$$

Property 2.1 *The following properties are immediate*:

1. *For all $g \in \mathcal{G}_d$ and all $k \in \mathbb{N}$, $D_k(g) \geq 0$.*
2. *There exists $k \leq d$ s.t. $D_k(g) = 0$.*
3. *For any k, if $g(x,y) \in [0, \bar{\ell}]$ for all $(x,y) \in \mathcal{X} \times \mathcal{Y}$, then $D_k(g) \in [0, \bar{\ell}]$.*
4. *If ℓ is L-Lipschitz in its first argument, then $\forall h \in \mathcal{H}_d, D_k(g) \leq L \cdot D_k(h)$, where $g = \ell \circ h$.*

Moreover, these properties also hold for D_R.

Due to the first two properties above, as $k \to d$, the generalisation bounds for compressive ERM will recover those for the original ERM. The last property implies that for many loss functions of interest, the compressive distortion can be bounded independently of label information.

It is natural to conjecture that learning problems whose target function has small compressive distortion are easier for compressive learning. This is indeed the case, as we shall see shortly. Recall the empirical Rademacher complexity of a function class $\mathcal{G}$ is defined as $\hat{\mathcal{R}}_N(\mathcal{G}) = \frac{1}{N} E_\sigma \sup_{g \in \mathcal{G}} \sum_{n=1}^N \sigma_n g(X_n)$, where $\sigma = (\sigma_1, \ldots, \sigma_N) \overset{\text{i.i.d}}{\sim} \text{Uniform}(\pm 1)$. Let us denote by $\hat{g}_R = \ell \circ \hat{h}_R$ the loss of the compressive ERM predictor. We have the following generalisation bound.

Theorem 2 (Generalisation of compressive ERM) *Let $\mathcal{G}_R$ be the loss class associated with the compressive class of functions $\mathcal{H}_R$, and assume that ℓ is uniformly bounded above by $\bar{\ell}$. For any $k \in \mathbb{N}$ and $\delta > 0$, w.p. $1 - 2\delta$,*

$$E[\hat{g}_R] \leq E[g^*] + D_k(g^*) + 2\hat{\mathcal{R}}_N(\mathcal{G}_R) + \bar{\ell} \cdot \xi(k, g^*, \delta) + 4\bar{\ell}\sqrt{\frac{\log(3/\delta)}{2N}} \tag{2}$$

where $\xi(k, g^*, \delta) \equiv \min\left\{ \frac{1-\delta}{\delta} D_k(g^*), \sqrt{\frac{1}{2}\log\frac{1}{\delta}} \right\}$. In particular, if $D_k(g^*) \leq \theta$ for some $\theta \in [0, \bar{\ell}]$, then the compressive ERM satisfies

$$E[\hat{g}_R] \leq E[g^*] + \theta + 2\hat{\mathcal{R}}_N(\mathcal{G}_R) + \bar{\ell} \cdot \xi(k, g^*, \delta) + 4\bar{\ell}\sqrt{\frac{\log(3/\delta)}{2N}}. \tag{3}$$

Proof Fixing R we have an ERM over the compressive class. Hence, we can bound the generalisation error of the function learned, $\hat{g}_R \in \mathcal{G}_R$, using classic uniform bounds such as

(Mohri et al, 2012, Lemma 3.3) (Theorem 29 in Appendix 5) combined with the Hoeffding bound. This gives w.p. $1-\delta$ that

$$E[\hat{g}_R] \leq E[g_R^*] + 2\hat{\mathcal{R}}_N(\mathcal{G}_R) + 4\bar{\ell}\sqrt{\frac{\log(3/\delta)}{2N}} \tag{4}$$

This bound is relative to $g_R^* \in \mathcal{G}_R$, that is the best achievable in the reduced class, while we want a bound relative to the best achievable in the original class, i.e. $g^* \in \mathcal{G}_d$. To this end, we write

$$E[g_R^*] = E[g^*] + E[g_R^* - g^*] \leq E[g^*] + \inf_{g_R \in \mathcal{G}_R} E|g_R - g^*| = E[g^*] + D_R(g^*), \tag{5}$$

where we used Jensen's inequality to draw the infimum out of the expectation, since the infimum is a concave function.

Now, since the loss is bounded, and recalling that $D_k(g^*) = E_R[D_R(g^*)]$, we can bound the last term on the r.h.s. as $D_R(g^*) \leq D_k(g^*) + \sqrt{\frac{1}{2}\log(1/\delta)}$ w.p. $1-\delta$ using Hoeffding's inequality (Lemma 27), or alternatively as $D_R(g^*) \leq \frac{1}{\delta}D_k(g^*) = D_k(g^*) + \frac{1-\delta}{\delta}D_k(g^*)$ w.p. $1-\delta$ using Markov's inequality (Lemma 26). Each of these two bounds can be tighter than the other depending on the magnitude of $D_k(g^*)$. By taking the minimum, we have

$$D_R(g^*) \leq D_k(g^*) + \xi(k, g^*, \delta). \tag{6}$$

Finally, by the union bound, both (4) and (6) hold simultaneously w.p. $1-2\delta$, hence we conclude the statement (2). Equation (3) follows from (2) by substituting the upper bound θ for $D_k(g^*)$. □

The error of the uncompressed ERM is recovered when $D_k(g^*) = 0$, which in the worst case will happen for $k = d$. Moreover, depending on the structure of the problem, $D_k(g^*)$ can become negligible even for $k < d$. Theorem 2 implies that compressive learning will work better on problems where the target function g^* has small compressive distortion.

The benefit of this simple result is to unify the analysis of compressive learning of various models into one framework, which further depends on problem-specific quantities. In particular, the compressive distortion appears in the bound, which depends on the particular model class, and analysing this quantity further will give us a handle on discovering problem-specific characteristics that contribute to the ease of learning from compressed data.

Here we assumed that the distortion threshold θ and the compression dimension k are fixed in advance. The latter may be set to a fraction of the available sample size N, so that the function class complexity remains small. Later in Sect. 3 we develop some intuition about the geometric meaning of compressive distortion in some concrete function classes, and demonstrate how it can be used to learn about benign problem characteristics.

2.2 Learning compressible problems in the dataspace

The main quantity in our analysis of compressive learning in the previous section was the compressive distortion of the target function, $D_k(g^*)$. In this section we return to the original high dimensional problem, and define a notion of distortion for the entire function class, which we refer to as the compresive complexity of the class. We shall then focus on

function classes that have low compressive complexity. The intuition behind this approach is that such classes are in fact a smaller in some sense, which should allow easier learning—albeit this will have to be a non-ERM algorithm that avoids the pitfalls of ERM that we exemplified earlier in Sect. 1.2, and this will indeed follow from our analysis. To this end, in this section we give a uniform bound in terms of compressive complexity.

We introduce an auxiliary construction that involves a random projection for analytic purposes, while the learning problem stays in the original data space without any dimensionality reduction. As before, $R \in \mathbb{R}^{k\times d}, k \leq d$ is a RP matrix, but this time it will serve a purely analytic role. We define an auxiliary function class, $\mathcal{G}_R = \ell \circ \mathcal{H}_R$ with elements $g_R = \ell \circ h_R$—again for analytic purposes. This class may be chosen freely. A natural choice is to have the same functional form as the elements of $\mathcal{G}_d$, but operating on k (rather than d) dimensional inputs, as then from a compressive learning guarantee one can readily infer a dataspace guarantee, as we shall see shortly. However, other choices can be more convenient to work with when the dataspace bound is sought. Next, we define compressive complexity with the aid of an unspecified auxiliary class $\mathcal{G}_R$, as follows.

Definition 2 (Compressive complexity of a function class) Given a function class $\mathcal{G}_d$ and a function $g \in \mathcal{G}_d$, we let $\hat{D}_{R,N}(g) \equiv \inf_{g_R \in \mathcal{G}_R} \hat{E}_{\mathcal{T}_N} |g_R(RX, Y) - g(X, Y)|$, and $\hat{D}_{k,N}(g) \equiv E_R[\hat{D}_{R,N}(g)]$. We define the compressive distortion of $\mathcal{G}_d$ as the following.

$$\hat{\mathcal{C}}_{k,N}(\mathcal{G}_d) \equiv \sup_{g\in\mathcal{G}_d} \hat{D}_{k,N}(g); \qquad \mathcal{C}_{k,N}(\mathcal{G}_d) \equiv E_{\mathcal{T}_N\sim\mathbb{P}^N}[\hat{\mathcal{C}}_{k,N}(\mathcal{G}_d)] \tag{7}$$

We may think of the compressive complexity as the largest (w.r.t. $g \in \mathcal{G}_d$) 'mimicking error' (on average over training sets) of compressive learners that each receive a randomly compressed version of the inputs and learn to behave like g. With the use of Definition 2, we can decompose the Rademacher complexity of the original class as the following.

Lemma 3 (Decomposition of Rademacher complexities) *Let $\mathcal{G}_d$ be a class of uniformly bounded real valued functions on $\mathcal{X}$. We have*

$$\hat{\mathcal{R}}_N(\mathcal{G}_d) \leq \hat{\mathcal{C}}_{k,N}(\mathcal{G}_d) + E_R[\hat{\mathcal{R}}_N(\mathcal{G}_R)] \tag{8}$$

$$\mathcal{R}_N(\mathcal{G}_d) \leq \mathcal{C}_{k,N}(\mathcal{G}_d) + E_R[\mathcal{R}_N(\mathcal{G}_R)] \tag{9}$$

$$\mathcal{R}_N(\mathcal{G}_d) \leq \hat{\mathcal{C}}_{k,N}(\mathcal{G}_d) + E_R[\hat{\mathcal{R}}_N(\mathcal{G}_R)] + \bar{\ell}\sqrt{\frac{\log(1/\delta)}{2N}} \text{ w.p. } 1-\delta \tag{10}$$

$$\mathcal{R}_N(\mathcal{G}_d) \leq \hat{\mathcal{C}}_{k,N}(\mathcal{G}_d) + E_R[\mathcal{R}_N(\mathcal{G}_R)] + \bar{\ell}\sqrt{\frac{\log(1/\delta)}{2N}} \text{ w.p. } 1-\delta \tag{11}$$

$$\mathcal{R}_N(\mathcal{G}_d) \le \mathcal{C}_{k,N}(\mathcal{G}_d) + E_R[\hat{\mathcal{R}}_N(\mathcal{G}_R)] + \bar{\ell}\sqrt{\frac{\log(1/\delta)}{2N}} \text{ w.p. } 1-\delta \tag{12}$$

Proof of Lemma 3 By the definition,

$$\hat{\mathcal{R}}_N(\mathcal{G}_d) = E_\sigma \sup_{g\in\mathcal{G}_d} \frac{1}{N}\sum_{n=1}^{N} \sigma_n g(X_n, Y_n).$$

We add and subtract $E_\sigma \sup_{g\in\mathcal{G}_d} E_R \inf_{g_R\in\mathcal{G}_R} \left\{ \frac{1}{N}\sum_{n=1}^{N} \sigma_n g_R(RX_n, Y_n) \right\}$, so

$$\begin{aligned}\hat{\mathcal{R}}_N(\mathcal{G}_d) &\le E_\sigma \sup_{g\in\mathcal{G}_d} E_R \inf_{g_R\in\mathcal{G}_R} \left\{ \frac{1}{N}\sum_{n=1}^{N} \sigma_n (g(X_n, Y_n) - g_R(RX_n, Y_n)) \right\} \\ &\qquad + E_\sigma E_R \sup_{g_R\in\mathcal{G}_R} \left\{ \frac{1}{N}\sum_{n=1}^{N} \sigma_n g_R(RX_n, Y_n) \right\} \\ &\le \hat{\mathcal{C}}_{k,N}(\mathcal{G}_d) + E_R[\hat{\mathcal{R}}_N(\mathcal{G}_R)].\end{aligned}$$

This completes the proof of (8). Taking expectation w.r.t. the distribution of $\mathcal{T}_N$ we obtain (9). Using these, we obtain inequalities (10)–(12) by employing McDiarmid's inequality (Lemma 28), as follows.

Since the loss function is bounded by $\bar{\ell}$, changing one point of $\mathcal{T}_N$ can only change $\hat{\mathcal{R}}_N(\mathcal{G}_d)$ (or $\hat{\mathcal{C}}_{k,N}(\mathcal{G}_d)$), as a functions of a set of N points, by at most $c = \bar{\ell}/N$. Hence, applying one side of McDiarmid's inequality gives each of the following

$$\mathcal{R}_N(\mathcal{G}_d) \le \hat{\mathcal{R}}_N(\mathcal{G}_d) + \bar{\ell}\sqrt{\frac{\log(1/\delta)}{2N}} \text{ w.p. } 1-\delta; \tag{13}$$

$$\mathcal{C}_{k,N}(\mathcal{G}_d) \le \hat{\mathcal{C}}_{k,N}(\mathcal{G}_d) + \bar{\ell}\sqrt{\frac{\log(1/\delta)}{2N}} \text{ w.p. } 1-\delta. \tag{14}$$

Now, combining (13) with (8) gives (10). Combining (9) with (14) gives (11). Finally, using (9) and then applying (13) with the class $\mathcal{G}_R$ gives (12). □

The reason the above decompositions will be useful for our purposes is that, whenever $\mathcal{C}_{k,N}(\mathcal{G}_d)$ is sufficiently small, then the Rademacher complexity of the original function class becomes essentially the complexity of a k rather than a d dimensional function class—therefore, inspecting $\mathcal{C}_{k,N}(\mathcal{G}_d)$ for the class $\mathcal{G}_d$ at hand will help us gain intuitive insight about the structures that make some high dimensional problems actually be less high dimensional than they appear to be. As such, our focus is on problems where $\mathcal{R}_N(\mathcal{G}_d)$ grows with d, and $\mathcal{C}_{k,N}(\mathcal{G}_d)$ is small, and examples will follow in the next section. In such problems, when prior knowledge does not justify any further assumptions, the smallness of compressive distortion represents a *general-purpose simplicity conjecture* that may be used to derive conditions for a high dimensional problem to be solvable in low dimensions. The particular form of these will depend on the particular function class associated with the learning problem, but for now we keep the formalism general and simple.

Theorem 4 (Uniform bounds for problems with small compressive complexity) *Fix some* $\theta \in [0, \bar{\ell}]$. *Suppose that* $\tilde{\mathcal{G}}_d \subseteq \mathcal{G}_d$ *is a function class that satisfies* $\mathcal{C}_{k,N}(\tilde{\mathcal{G}}_d) \leq \theta$. *Then, for any* $\delta > 0$, *w.p.* $1 - \delta$ *the following holds uniformly for all* $g \in \tilde{\mathcal{G}}_d$:

$$E[g] \leq \hat{E}_{\mathcal{T}_N}[g] + 2\theta + 2E_R[\hat{\mathcal{R}}_N(\mathcal{G}_R)] + 3\bar{\ell}\sqrt{\frac{\log(2/\delta)}{2N}} \tag{15}$$

Furthermore, w.p. $1 - \delta$, $\hat{g} := \arg\min_{g \in \tilde{\mathcal{G}}_d} \hat{E}[g]$, *satisfies*

$$E[\hat{g}] \leq E[g^*] + 2\theta + 2E_R[\hat{\mathcal{R}}_N(\mathcal{G}_R)] + 4\bar{\ell}\sqrt{\frac{\log(3/\delta)}{2N}}. \tag{16}$$

Proof By the classic Rademacher bound (Theorem 29) applied to $\tilde{\mathcal{G}}_d$, we have w.p. $1 - \delta/2$ for all $g \in \tilde{\mathcal{G}}_d$ that

$$E[g] \leq \hat{E}_{\mathcal{T}_N}[g] + 2E_R[\mathcal{R}_N(\tilde{\mathcal{G}}_d)] + \bar{\ell}\sqrt{\frac{\log(2/\delta)}{2N}}. \tag{17}$$

Applying (12) from Lemma 3 to $\tilde{\mathcal{G}}_d$, we further have $\mathcal{R}_N(\tilde{\mathcal{G}}_d) \leq \hat{\mathcal{R}}_N(\tilde{\mathcal{G}}_R) + \mathcal{C}_{k,N}(\tilde{\mathcal{G}}_d) + \bar{\ell}\sqrt{\frac{\log(2/\delta)}{2N}}$ w.p. $1 - \delta/2$, where $\tilde{\mathcal{G}}_R \subseteq \mathcal{G}_R$. This combined with (17) using the union bound gives w.p. $1 - \delta$

$$E[g] \leq \hat{E}_{\mathcal{T}_N}[g] + 2E_R[\hat{\mathcal{R}}_N(\tilde{\mathcal{G}}_R)] + 2\mathcal{C}_{k,N}(\tilde{\mathcal{G}}_d) + \bar{\ell}\sqrt{\frac{\log(2/\delta)}{2N}}. \tag{18}$$

Finally, $\tilde{\mathcal{G}}_R \subseteq \mathcal{G}_R$ implies $\hat{\mathcal{R}}_N(\tilde{\mathcal{G}}_R) \leq \hat{\mathcal{R}}_N(\mathcal{G}_R)$, and using that $\mathcal{C}_{k,N}(\tilde{\mathcal{G}}_d) \leq \theta$ completes the proof of (15).

Equation (16) follows from (15). Indeed, as (15) holds uniformly for all $g \in \tilde{\mathcal{G}}_d$, it also holds with $\hat{g}$ in the place of g, and we apply this w.p. $1 - 2\delta/3$ yielding

$$E[\hat{g}] \leq \hat{E}_{\mathcal{T}_N}[\hat{g}] + 2\theta + 2E_R[\hat{\mathcal{R}}_N(\mathcal{G}_R)] + 3\bar{\ell}\sqrt{\frac{\log(2/(3\delta))}{2N}}. \tag{19}$$

By definition of $\hat{g}$, we also have $\hat{E}_{\mathcal{T}_N}[\hat{g}] \leq \hat{E}_{\mathcal{T}_N}[g^*]$, and by Hoeffding's inequality we further have $\hat{E}_{\mathcal{T}_N}[g^*] \leq E[g^*] + \hat{\ell}\sqrt{\frac{\log(3/\delta)}{2N}}$ w.p. $1 - \delta/3$. Finally, we combine this with (19) via the union bound to complete the proof. □

Theorem 4 implies that, if the compressive complexity of the function class is sufficiently small, then the d-dimensional problem is solvable with a guarantee that is almost as good as a $k \ll d$-dimensional version of the problem. This is of interest in problems where the available sample size N is too small relative to d to permit a meaningful guarantee. Observe that k manages a tradeoff, as θ decreases with k while the Rademacher complexity in general may increase with k. As before, k and θ are considered to be fixed before seeing the data. A sensible choice is to set k proportional to N—which is typically known—in other words, in small sample settings we are prepared to take a bias θ and in return gain control over the affordable complexity of the class. The classic bounds are recovered when $k = d$. However, the intuition is that often the geometry of the problem may be favourable for θ to be sufficiently small while $k \ll d$. Our bounds express this intuition, and Sect. 3 will make it more concrete.

Note that the restriction of the function class to obey $\mathcal{C}_{k,N}(\tilde{\mathcal{G}}_d) \leq \theta$ is necessary for the above guarantee. This is important, as in practice it is often easier to specify a large class $\mathcal{G}_d$, and we have seen earlier in Theorem 1 that an unconstrained ERM can be arbitrarily bad. Hence, in order to exploit the guarantee provided by Theorem 4, the learning algorithm must ensure this constraint.

The compressive complexity has similar properties to those of compressive distortion.

Property 2.2 *The following properties hold.*

1. *For all $g \in \mathcal{G}_d$ and all $k \in \mathbb{N}$, $\mathcal{C}_{k,N}(\mathcal{G}_d) \geq 0$.*
2. *There exists $k \leq d$ s.t. $\mathcal{C}_{k,N}(\mathcal{G}_d) = 0$.*
3. *For any k, if $g(x, y) \in [0, \bar{\ell}]$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, then $\mathcal{C}_{k,N}(\mathcal{G}_d) \in [0, \bar{\ell}]$.*
4. *If ℓ is L-Lipschitz in its first argument, then $\mathcal{C}_{k,N}(\mathcal{G}_d) \leq L \cdot \mathcal{C}_{k,N}(\mathcal{H}_d)$.*

Moreover, these properties also hold for $\hat{D}_{R,N}(\cdot), \hat{D}_{k,N}(\cdot)$, and $\hat{\mathcal{C}}_{k,N}(\cdot)$.

Furthermore, we can link compressive distortion with compressive complexity, and this facilitates insights about high dimensional dataspace learning from guarantees obtained on compressive learning.

Property 2.3 (From compressive distortion to compressive complexity) *Let $\hat{\mathbb{P}}$ denote the counting probability measure over the training sample. Suppose we have a bound $D_R(h) \leq \psi_R(h, \mathbb{P})$ for all $h \in \mathcal{H}_d$, where ψ_R is some expression that depends on R. Then, we also have $\mathcal{C}_{k,N}(\mathcal{H}_d) \leq E_{\mathcal{T}_N \sim \mathbb{P}^N}[\sup_{h \in \mathcal{H}_d} E_R[\psi_R(h, \hat{\mathbb{P}})]]$. In particular, if $D_R(h) \leq \phi(\mathbb{P}) \cdot \varphi_R(h)$ for all $h \in \mathcal{H}_d$ with some expressions ϕ and φ_R, then $\mathcal{C}_{k,N}(\mathcal{H}_d) \leq E_{\mathcal{T}_N \sim \mathbb{P}^N}[\phi(\hat{\mathbb{P}})] \cdot \sup_{h \in \mathcal{H}_d} E_R[\varphi_R(h)]$.*

Proof of Property 2.3 Since $D_R(h) \leq \psi_R(h, \mathbb{P})$ for all $h \in \mathcal{H}_d$, we also have $\hat{D}_{R,N}(h) \leq \psi_R(h, \hat{\mathbb{P}})$ for all $h \in \mathcal{H}_d$. Hence,

$$\mathcal{C}_{k,N}(\mathcal{H}_d) = E_{\mathcal{T}_N \sim \mathbb{P}^N} \sup_{h \in \mathcal{H}_d} E_R[\hat{D}_{R,N}(h)] \leq E_{\mathcal{T}_N \sim \mathbb{P}^N} \sup_{h \in \mathcal{H}_d} E_R[\psi_R(h, \hat{\mathbb{P}})]. \tag{20}$$

Applying this to the special case when $\psi(h, \mathbb{P}) = \phi(\mathbb{P}) \cdot \varphi_R(h)$ for all $h \in \mathcal{H}_d$, the second statement follows. □

Below in Lemma 5 we give a simple example of a compressible problem, i.e. a distribution and function class pair where we have both a low compressive distortion and a low compressive complexity.

Definition 3 (Almost low-rank distributions) Given $\theta \in [0, 1]$ and $k \leq d$ we say that a probability measure μ is θ-almost k-rank on $\mathbb{R}^d$, if there exists a k-dimensional linear subspace $V_k \subseteq \mathbb{R}^d$ such that $\mu(V_k) > 1 - \theta$.

Lemma 5 (Compressive distortion and compressive complexity in almost low-rank distributions) *Let $\mathcal{G}_d$ be the linear function class with an $\bar{\ell}$-bounded loss function. Suppose that the marginal $\mathbb{P}_X$ is a θ-almost k-rank distribution on $\mathbb{R}^d$, and R is a $k \times d$ RP matrix having full row-rank a.s. For any $N \in \mathbb{N}$, we have*

$$D_k(g^*) \leq \bar{\ell}\theta \tag{21}$$

$$\mathcal{C}_{k,N}(\mathcal{G}_d) \leq \bar{\ell}\theta. \tag{22}$$

Lemma 5 will be useful in the construction of a lower bound in Sect. 2.3. The idea of the proof is that, knowing that the marginal distribution is almost k-rank, we can choose the auxiliary class $\mathcal{G}_R$ such that $R \in \mathbb{R}^{k\times d}$ leaves the linear subspace V_k unchanged a.s.

The proof of Lemma 5 is given in Appendix Sect. 2.

2.3 Tightness of the bounds

The upper bounds of Theorems 2 and 4 are attractive when θ is small, i.e. for compressible problems. Our goal in this section is to show the tightness of these bounds under the same conditions as those upper bounds. More precisely, we will show that there exists a function class for which the dependence of the bound on the parameters θ, k and N cannot be improved without imposing extra conditions.

First, we need to make explicit the dependence of the relevant quantities on the unknown data distribution $\mathbb{P}_d$. To this end, we shall use the notations $D_k(g^*, \mathbb{P}_d)$ and $\mathcal{C}_{k,N}(\mathcal{G}_d, \mathbb{P}_d)$ for the compressive distortion and the compressive complexity respectively. We drop the index d as it stays the same throughout this section, so $\mathcal{G}$ will stand for $\mathcal{G}_d$, and $\mathcal{H}$ will stand for $\mathcal{H}_d$. As in the previous sections, we assume $\bar{\ell}$-bounded loss functions.

Next, we define the class of distributions for which these quantities are below a specified threshold.

Definition 4 (Compressible distributions) Let $k \leq d$ be an integer, and $\theta \in [0, 1]$.

1. Given a learning problem with target function $g^*(\cdot, \cdot) = \ell(h^*(\cdot), \cdot)$, we say that a distribution $\mathbb{P}$ is D-compressible with parameters (θ, k), if the compressive distortion of g^* satisfies $D_k(g^*, \mathbb{P}) \leq \bar{\ell}\theta$. We denote by $\mathcal{P}_{g^*}(\theta, k) := \{\mathbb{P} : D_k(g^*, \mathbb{P}) \leq \bar{\ell}\theta\}$ the set of all D-compressible distributions with parameters (θ, k).
2. Given a function class $\mathcal{G}$, we say that a distribution $\mathbb{P}$ is C-compressible with parameters (θ, k), if the compressive complexity of $\mathcal{G}$ satisfies $\mathcal{C}_{k,N}(\mathcal{G}, \mathbb{P}) \leq \bar{\ell}\theta$. We denote by $\mathcal{P}_{\mathcal{G}}(\theta, k) := \{\mathbb{P} : C_{k,N}(\mathcal{G}, \mathbb{P}) \leq \bar{\ell}\theta\}$ the set of all C-compressible distributions with parameters (θ, k).

For a distribution $\mathbb{P}$, we denote by $h^*_{\mathbb{P}} \in \arg\inf_{h\in\mathcal{H}} E[\ell(h(X), Y)]$ a best classifier of the class $\mathcal{H}$ in the underlying distribution $\mathbb{P}$. In the construction of the proof of the forthcoming Theorem 6, $h^*_{\mathbb{P}}$ will coincide with the Bayes-optimal classifier. A learning algorithm $\mathcal{A} : (\mathcal{X}\times\mathcal{Y})^N \to \mathcal{H}$ takes a training set of size N and returns a classifier. The loss of this classifier is denoted by $g_{\mathcal{A}(\mathcal{T}_N)}(X, Y) := \ell((\mathcal{A}(\mathcal{T}_N))(X), Y)$.

We have the following lower bound in the high-dimensional small sample setting.

Theorem 6 (Lower bound) *Consider the* 0–1 *loss. For any* $\theta \in [0, 1]$, *any integers* $k \leq N \leq d$, *and any algorithm* $\mathcal{A} : (\mathcal{X}\times\mathcal{Y})^N \times \mathcal{X} \to \mathcal{H}$ *there exists a D-compressible and C-compressible distribution* $\mathbb{P} \in \mathcal{P}_{g^*}(k, \theta) \cap \mathcal{P}_{\mathcal{G}}(k, \theta)$ (*which depends on* θ, k, d, N *and* $\mathcal{A}$) *such that*:

$$E_{\mathcal{T}_N \sim \mathbb{P}^N}[E[g_{\mathcal{A}(\mathcal{T}_N)}]] - E[g^*_{\mathbb{P}}] \geq \frac{1}{32}\left(\theta + \sqrt{\frac{k}{N}}\right). \tag{23}$$

The proof is deferred to Appendix 4. Theorem 6 says that, in the high dimensional setting ($k \leq N \leq d$), for any choice of algorithm there is a bad distribution which, despite it being compressible (i.e. it satisfies the same condition as our upper bounds), the error of the classifier returned by the algorithm on an i.i.d. sample of size N from that distribution is large.

We note that the bad distribution is allowed to depend on the sample size. Therefore Theorem 6 does not imply that, for some distribution, the excess risk converges at a rate no faster than that of the upper bound. However, studying faster rates is beyond the scope of this paper, as require additional assumptions and is pursued elsewhere (Reeve & Kabán, 2021).

The important point here is that, there are function classes for which the lower bound of Theorem 6 matches the upper bound up to a constant factor—for instance in k-dimensional linear classification it is well-known that $\hat{\mathcal{R}}_N(\mathcal{G}_R) \in \Theta\left(\sqrt{k/N}\right)$ (Bartlett & Mendelson, 2002). Hence, the lower bound of Theorem 6 implies that Theorem 4 cannot be improved in general by more than a constant factor. To see this more clearly, we rearrange the upper bound from Theorem 4 to have the same left-hand side as (23). Setting $\epsilon := 4\bar{\ell}\sqrt{\frac{\log(3/\delta)}{2N}}$ gives $2\delta = 6\exp\left(-\frac{N\epsilon^2}{8\bar{\ell}^2}\right)$, and we have

$$\mathbb{P}_{\mathcal{T}_N}\left\{E[\hat{g}] > E[g^*] + 2\theta + 2E_R[\hat{\mathcal{R}}_N(\mathcal{G}_R)] + \epsilon\right\} \leq 6\exp\left(-\frac{N\epsilon^2}{8\bar{\ell}^2}\right).$$

This implies that

$$\begin{aligned} &E_{\mathcal{T}_N}[E[\hat{g}]] - E[g^*] - 2\theta + 2E_R[\hat{\mathcal{R}}_N(\mathcal{G}_R)] \\ &\qquad \leq \int_0^\infty \mathbb{P}_{\mathcal{T}_N}\left\{E[\hat{g}] - E[g^*] + 2\theta - 2E_R[\hat{\mathcal{R}}_N(\mathcal{G}_R)] > \epsilon\right\} d\epsilon \\ &\qquad \leq 6\int_0^\infty \exp\left(-\frac{N\epsilon^2}{8\bar{\ell}^2}\right) d\epsilon = 3\sqrt{\frac{8\pi\bar{\ell}^2}{N}} < \frac{16\bar{\ell}}{\sqrt{N}}. \end{aligned}$$

Hence, noting that $\bar{\ell}$ is a constant independent of k, N, d and θ, we have for the linear class that

$$\begin{aligned} E_{\mathcal{T}_N}[E[\hat{g}]] - E[g^*] &\leq 2\theta + 2E_R[\hat{\mathcal{R}}_N(\mathcal{G}_R)] + \frac{16\bar{\ell}}{\sqrt{N}} \\ &\leq 2\theta + 62\sqrt{\frac{k}{N}} + \frac{16\bar{\ell}}{\sqrt{N}} = \mathcal{O}\left(\theta + \sqrt{\frac{k}{N}}\right). \end{aligned} \tag{24}$$

This matches the lower bound up to a constant factor.

In the compressed ERM bound of Theorem 2, the term $\xi(k, g^*, \delta)$ reflects the variability of error due to working in a lower dimensional random subspace of $\mathcal{X}$. This term is irreducible with N, instead it decays with k through $D_k(g^*)$, which is model-specific. The next section will analyse this quantity for several learning problems. Moreover, cf. the second statement of Property 2.1, there is always some integer $k^* \leq d$ such that whenever $k \geq k^*$

we will have $\xi(k, g^*, \delta) = 0$, making the upper bound again match the lower bound up to a constant factor.

3 Discovering problem-specific benign traits

The previous section focused on bounds of a general form, and we argued that these are tight when the problem is compressible. In this section we study the question of what makes learning problems compressible. The answers will depend on the particular learning problem, and we demonstrate how the novel quantities we introduced (the compressive distortion and the compressive complexity) can exploit the structure-exposing ability of random projections to reveal more answers to this question.

The forthcoming subsections are devoted to instantiating these quantities in several models associated to learning tasks, in order to demonstrate their use in revealing structural insights. The proofs of the forthcoming propositions are relegated to Appendix 3, where we also give details on how to use the obtained expressions in the general form of our bounds from the previous sections.

3.1 Thresholded linear models

We start with the classical example of binary classification with linear functions $\mathcal{H}_d = \{x \to h^T x : h, x \in \mathbb{R}^d\}$, and where the loss function of interest is the 0–1 loss, that is $\ell_{01} : \mathcal{Y} \times \mathcal{Y} \to \{0, 1\}, \ell_{01}(\hat{y}, y) = \mathbf{1}(\hat{y}y \leq 0)$. By a slight abuse of notation, we identify the linear classifiers with their weight vectors. As before, we let $\mathcal{G}_d = \ell_{01} \circ \mathcal{H}_d$, and $\mathcal{G}_R = \ell_{01} \circ \mathcal{H}_R$ its compressive counterpart. In this setting, we have the following, proved in Appendix Section "Thresholded linear models".

Proposition 7 *Consider the linear function class with the* 0–1 *loss, as above. We have*

$$D_k(g^*) \leq E_X \left[\exp\left(\frac{-k \cos^2(\measuredangle_X^h)}{8} \right) \right] \cdot \mathbf{1}(k < d) \tag{25}$$

$$\mathcal{C}_{k,N}(\mathcal{G}_d) \leq E_X \left[\sup_{h \in \mathcal{H}_d} \exp\left(\frac{-k \cos^2(\measuredangle_X^h)}{8} \right) \right] \cdot \mathbf{1}(k < d). \tag{26}$$

In the above, $\measuredangle_X^h)$ is the angle, in radians, between the vectors X and h, so $\cos(\measuredangle_X^h)$ is the normalised margin of a point X in terms of its distance to the hyperplane with normal vector h. Consequently, we see that in the case of halfspace learning, the compressive distortion is bounded by the moment generating function of the square of margin distribution. This example recovers as a special case, the main findings of Kabán and Durrant (2020) in a nutshell.

3.2 Linear model with Lipschitz loss

Next we consider the linear function class $\mathcal{H}_d = \{x \to h^T x : h, x \in \mathbb{R}^d\}$, with $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, \bar{\ell}]$ a bounded loss function that is L_ℓ-Lipschitz in its first argument. Common examples of bounded Lipschitz loss functions may be found e.g. in (Rosasco et al., 2004), several of which are surrogates for the 0–1 loss. As before, we let $\mathcal{G}_d = \ell \circ \mathcal{H}_d$, and $\mathcal{G}_R = \ell \circ \mathcal{H}_R$ its compressive version. Let $\Sigma := E_X[XX^T]$, and we require that $\mathrm{Tr}(\Sigma) < \infty$. In this setting, we have the following.

Proposition 8 *Consider the linear model class described above. For any $p \in \mathbb{N}$ s.t. $2 \leq p \leq k-2$, and $k \leq \mathrm{rank}(\Sigma)$, we have*

$$D_k(g^*) \leq L_\ell \|h^*\|_2 \cdot \Xi(k, p, \{\lambda_j(\Sigma)\}_j) \tag{27}$$

$$\mathcal{C}_{k,N}(\mathcal{G}_d) \leq L_\ell E_{\mathcal{T}_N \sim \mathbb{P}^N}[\Xi(k, p, \{\lambda_j(\hat{\Sigma})\}_j)] \cdot \sup_{h \in \mathcal{H}_d} \|h\|_2 \tag{28}$$

where

$$\Xi(k, p, \{\lambda_j(\Sigma)\}_j) := \left(1 + \sqrt{\frac{k-p}{p-1}}\right)\sqrt{\lambda_{k-p+1}(\Sigma)} + \frac{e\sqrt{k}}{p}\sqrt{\sum_{j>k-p} \lambda_j(\Sigma)} \tag{29}$$

From the form of (27)–(28) we infer that both the compressive distortion and the compressive complexity decrease with the inverse margin and the rate of decay of the eigen-spectrum of the data covariance. In conjuction with Theorem 2 this means that the larger the margin of h^*, or/and the faster the eigen-decay of the data covariance, the better the chance that compressive classification with the considered linear model class succeeds. Likewise, in the light of Theorem 4, learning the model in high dimensional settings is eased in situations where compressive complexity is small—i.e. when the margin is large, and the eigen-spectrum has a fast decay.

The proof is deferred to Appendix Section "Linear models with bounded Lipschitz loss". Essentially, we relate the problem to a weighted OLS problem, which was previously analysed (Kabán, 2013; Slawski, 2018), and then manipulate the expressions to apply a seminal result by Halko et al. (Halko et al., 2011).

It may be interesting to note that a coarser alternative that nevertheless retains the main characteristics can be obtained with less sophisticated tools, as the following.

Proposition 9 *Consider the linear model class described above. We have*

$$D_k(g^*) \leq L_\ell \sqrt{\frac{2}{k}}\sqrt{\mathrm{Tr}(\Sigma)}\|h^*\|_2 \tag{30}$$

$$\mathcal{C}_{k,N}(\mathcal{G}_d) \leq L_\ell \sqrt{\frac{2}{k}}\sqrt{\mathrm{Tr}(\Sigma)} \sup_{h \in \mathcal{H}_d} \|h^*\|_2. \tag{31}$$

Proof Using the Lipschitz property of the loss, and relaxing the infimum in the definition of $D_k(g^*)$, we have $D_k(g^*) \leq L_\ell E_X E_R\left[|h^T R^T RX - h^T X|\right] \leq L_\ell \left\{E_X E_R\left[(h^T R^T RX - h^T X)^2\right]\right\}^{1/2} \leq \sqrt{\frac{2}{k}}\mathrm{Tr}(\Sigma)\|h^*\|_2$. Here we used Lemma 2 of Kabán (2014) to compute the matrix expectation, which in our

case of R with i.i.d. entries from $\mathcal{N}(0, 1/k)$ says that $E_R[R^T R\Sigma R R^T] = \frac{1}{k}((k+1)\Sigma + \mathrm{Tr}(\Sigma) I_d)$. We also have $E_R[R^T R] = I_d$, and the final expression (30) then follows by rearranging, and using the Cauchy-Schwartz and Jensen's inequalities.

By the factorised form of (30), Property 2.3 immediately gives (31). □

We see the expressions in Propositions 8 and 9 are driven by the eigen-decay of the unknown true covariance, the margin of the classifier, and k with a decay of order $1/\sqrt{k}$.

3.3 Two-layer perceptron

The purpose of this section is to examine the effect of adding a hidden layer by considering the class of classic fully-connected two-layer perceptrons. It turns of that the distortion bounds can still be expressed in terms of structures that we encountered in the simpler model of the previous section.

Let $\mathcal{H}_d = \{x \to \sum_{i=1}^m v_i \phi(w_i^T x) : x \in \mathcal{X}_d, \|v\|_1 \leq 1\}$ be the class of classic two-layer perceptrons, where $\phi(\cdot)$ is a L_ϕ-Lipschits activation function. We do not regularise the first layer weights, as the RP has a regularisation effect on these. Let $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, \bar{\ell}]$ be an L_ℓ-Lipschitz and $\bar{\ell}$-bounded loss function as before, and let $\mathcal{G}_d = \ell \circ \mathcal{H}_d$, and $\mathcal{G}_R = \ell \circ \mathcal{H}_R$ its compressive version. The ℓ_1-regularisation on the higher-layer weights has the practical benefit of pruning unnecessary components. Again we will assume $\mathrm{Tr}(E_X[XX^T]) < \infty$. In this setting we obtain the following, proved in Appendix Section "Two-layer perceptron".

Proposition 10 *Consider the feed-forward neural network class above. For any $p \in \mathbb{N}$ s.t. $2 \leq p \leq k-2$, and any $k \leq \mathrm{rank}(\Sigma)$, we have*

$$D_k(g^*) \leq L_\ell L_\phi \|v^*\|_2 \|W^*\|_F \cdot \Xi(k, p, \{\lambda_j(\Sigma)\}_j) \cdot \mathbf{1}(k < d) \tag{32}$$

$$\mathcal{C}_{k,N}(\mathcal{G}_d) \leq L_\ell L_\phi E_{\mathcal{T}_N \sim \mathbb{P}^N}[\Xi(k, p, \{\lambda_j(\hat{\Sigma})\}_j)] \cdot \sup_{v,W} \|v\|_2 \|W\|_F \cdot \mathbf{1}(k < d) \tag{33}$$

where $\Xi(k, p, \{\lambda_j(\hat{\Sigma})\}_j)$ is defined in Eq. (29).

We have not considered adding further hidden layers, as the RP only affects the input layer, so deeper networks are unlikely to present further insights on the effect of compressing the data. We have also not attempted to extend our analysis to other types of neural nets in this fast developing field, as analytic bounds of the specific quantities we are interested in would quickly become difficult to obtain and interpret. However, we will return with a generally applicable approach later in Sect. 3.7, where we show how one can use additional unlabelled data to estimate the compressive complexity instead of analytically bounding it. Finally, in the light of multiple equivalent formulations of bounds for layered networks (Munteanu et al., 2022) (under certain conditions), one can argue that the question of what exactly the bounds depend on becomes less interesting for the study of neural nets. Indeed, our only purpose in this section was to demonstrate the intuition that, structures that help learning the linear model also help learning the two-layered model—hence, learning has at least as many (and probably more) benign structures to exploit in the richer class.

3.4 Quadratic model learning

Another interesting non-linear learning problem where we can showcase the ability of RP to discover meaningful structure and eliminate dimension-dependence is learning quadratic models, including Mahalanobis metric learning. Let $\mathcal{M}_d$ be the set of $d \times d$ symmetric matrices, and consider the quadratic function class $\mathcal{H}_d = \{x \to x^T A x : A \in \mathcal{M}_d, x \in \mathbb{R}^d\}$, with ℓ an $\bar{\ell}$-bounded L_ℓ-Lipschitz loss, and we denote be $\mathcal{G}_d = \ell \circ \mathcal{H}_d$ the loss class of $\mathcal{H}_d$. It is known from related analysis of Verma and Branson (2015) that the error of learning a Mahalanobis metric tensor $A \in \mathcal{M}_d$ necessarily grows with $\sqrt{d}$ if no structural assumptions are imposed on the metric tensor. We will use our RP-based analysis to discover a benign structural condition that eliminates the dependence of the error on d.

Let $\mathcal{H}_R$ be the compressive version of $\mathcal{H}_d$, with R having i.i.d. Gaussian entries with 0-mean and variance $1/k$, as before, and $\mathcal{G}_R = \ell \circ \mathcal{H}_R$. In Appendex section "Quadratic models" we show the following.

Proposition 11 *In the quadratic function class above, for any $k \leq d$, we have*

$$D_k(g^*) \leq \sqrt{\frac{4}{k^2} + \frac{3}{k}} L_\ell \mathrm{Tr}(\Sigma) \|A^*\|_* \cdot \mathbf{1}(k < d) \tag{34}$$

$$\mathcal{C}_{k,N}(\mathcal{G}_d) \leq \sqrt{\frac{4}{k^2} + \frac{5}{k}} L_\ell \mathrm{Tr}(\Sigma) \sup_{A \in \mathcal{M}_d} \|A\|_* \cdot \mathbf{1}(k < d) \tag{35}$$

where $\|\cdot\|_*$ is the nuclear norm of the matrix in its argument.

Equation 34 in conjunction with Theorem 2 highlights that, the smaller the nuclear norm of the true parameter matrix A^*, the better the generalisation guarantee for compressively learning the quadratic model. Equation (35) further suggests that learning a quadratic model in high dimensions becomes easier when the nuclear norm of the parameter matrix is small. In addition, both bounds of Proposition 11 scale with the trace of the true covariance of the data distribution, suggesting that spectral decay of the data source is a benign trait.

We find it interesting to relate our findings to recent results by Latorre et al. (2021) which have shown for the quadratic class of classifiers that nuclear norm regularisation in the original data space (no dimensionality reduction considered) has the ability to take advantage of low intrinsic dimensionality of the data to achieve better accuracy, which other regularisers studied therein do not. The fact that the nuclear norm appears in our distortion bounds further validates the ability of our RP-based approach to find meaningful structural traits for the learning problem at hand. In fact, Theorem 4 essentially turns the expression (35) into a regulariser, which is realised by the nuclear norm regulariser in this case, since all the other factors are independent of the model's parameters. Therefore the RP-based analysis following the same recipe as we did in the former sections for other function classes, again succeeded in revealing a meaningful benign trait for the function class under study.

3.5 Nearest neighbour classification

The previous sections concerned various parametric classes. Here we take a representative of a nonparametric class, namely a simplified version of the nearest neighbour classifier proposed by Kontorovich and Weiss (2015).

The nearest neighbour rule can be expressed as the following (Crammer et al., 2002; Kontorovich & Weiss, 2015; von Luxburg & Bousquet, 2004). Denote by $\mathcal{T}_N^+, \mathcal{T}_N^- \subset \mathcal{T}_N, \mathcal{T}_N^+ \cup \mathcal{T}_N^- = \mathcal{T}_N$ the positively and negatively labelled training points respectively. Define the distance of a point $x \in \mathcal{X}$ to a set S as $d(x,S) = \inf_{z\in S}\{\|x-z\|\}$. Then, $N^+(x) \equiv d(x,\mathcal{T}_N^+)$ and $N^-(x) \equiv d(x,\mathcal{T}_N^-)$ are the nearest positive / nearest negative neighbours of x, and the label prediction for $x \in \mathcal{X}$ is given by the sign of the following:

$$h(x : \mathcal{T}_N^+, \mathcal{T}_N^-) = \frac{1}{2}(d(x,\mathcal{T}_N^-) - d(x,\mathcal{T}_N^+)) = \frac{1}{2}\big(\|x-N^-(x)\| - \|x-N^+(x)\|\big) \tag{36}$$

Throughout this section we use Euclidean norms. Like Kontorovich and Weiss (2015), we assume a bounded input domain, $\mathcal{X}_d \subseteq \mathcal{B}(0,B)$. (This can be relaxed, as we will do in the next subsection for a more general case.) We consider the class of classifiers $\mathcal{H}_d = \{x \to h(x : \mathcal{T}_N^+, \mathcal{T}_N^-) = \frac{1}{2}\big(\|x-N^-(x)\| - \|x-N^+(x)\|\big)$, and $\mathcal{G}_d = \ell \circ \mathcal{H}_d$ where we take $\ell(\cdot)$ to be the ramp-loss defined as $\ell(h(x),y) = \min\{\max\{0, 1-h(x)y/\gamma\}, 1\}$, which is $1/\gamma$-Lipschitz.

In the RP-ed domain, we use subscripts: $N_R^+(x)$ and $N_R^-(x)$ denote the points whose images under the random projection R is the nearest positive or nearest negative to Rx. So the compressive class $\mathcal{H}_R$ contains functions of the form:

$$h_R(Rx : R\mathcal{T}_N^+, R\mathcal{T}_N^-) := \frac{1}{2}\big(\|Rx - N_R^-(Rx)\| - \|Rx - N_R^+(Rx)\|\big) \tag{37}$$

Composed with the $1/\gamma$-Lipschitz loss, we have by construction that $\mathcal{G}_d \subseteq \{x \to g(x): x \in \mathcal{X}_d, g \text{ is } 1/\gamma\text{-Lipschitz}\}$, and $\mathcal{G}_R \subseteq \{(Rx) \to g_R(Rx) : x \in \mathcal{X}_d, g_R \text{ is } 1/\gamma\text{-Lipschitz}\}$. That is, the function classes of interest are subsets of the d and k-dimensional class of $1/\gamma$-Lipschitz functions respectively. By the Lipschitz extension theorem (von Luxburg & Bousquet, 2004), for any γ-separated labelled sample there exists a 1-Lipschitz function has the same predictions as the 1-NN induced by that sample, for all points of the input domain $\mathcal{X}$.

For a given value of γ, the ERM classifier in the class of $1/\gamma$-Lipschitz functions of the form defined above is obtained by choosing a sub-sample from the training points such that this sub-sample is γ-separated, and the 1-NN induced by it makes the fewest errors on the full training set (including the points left out). This procedure was proposed by Kontorovich and Weiss (2015) along with an efficient algorithmic implementation.

Let g^* be the best d-dimensional $1/\gamma$-Lipschitz function of the form (36), and g_R the best k-dimensional $1/\gamma$-Lipschitz function of the form (37). We have the following, proved in Appendix Section "Nearest neighbours classification".

Proposition 12 *Let* $T = \left\{ \frac{x-x'}{\|x-x'\|} : x, x' \in \mathcal{X}_d, x \neq x' \right\}$. *For the class of nearest neighbour classifiers described above, we have*

$$D_k(g^*) \leq \frac{2B \cdot w(T)}{\gamma\sqrt{k}}; \qquad \mathcal{C}_{k,N}(\mathcal{G}_d) \leq \frac{2B \cdot w(T)}{\gamma\sqrt{k}}. \tag{38}$$

where $w(T) = E_{r\sim\mathcal{N}(0,1)} \sup_{t\in T}\{\langle r,t\rangle\}$ *is the Gaussian width of the set T.*

In this example, we have the same upper bound on both the compressive distortion and the compressive complexity, featuring the Gaussian width of the normalised distances on the support set. The Gaussian width (see e.g. Vershynin, 2018, sec. 7.5 and references

therein) is a measure of complexity for sets (justifying the name 'compressive complexity'), and it is sensitive not just to the algebraic intrinsic dimension of the support set but also takes fractional values reflecting weakly represented directions in the set, and it is sensitive to structure embedded in Euclidean spaces, such as the existence of a sparse representation, smooth manifold structure, spectral decay, and so on.

The bound we obtain by instantiating Theorems 2 and 4 with the expressions from Proposition 12 and the Rademacher complexity of $\mathcal{G}_R$ holds true with any integer value of k chosen before seeing the data. An interesting connection is obtained if we set k to the value that ensures that the compressive complexity term is below some specified $\eta \in (0, 1)$, i.e. $k \gtrsim \frac{w^2(T)}{\eta^2 \gamma^2}$. With this choice, the associated generalisation bound (Eq. 123) recovers a bound of the form obtained previously for this classifier in doubling metric spaces (Gottlieb et al., 2016; Kontorovich & Weiss, 2015), with the squared Gaussian width taking the place of the doubling dimension. Indeed, there is a known link between the doubling dimension and the squared Gaussian width (Indyk, 2007). In an Euclidean metric space with algebraic dimension d they are both of order $\Theta(d)$, but are otherwise more general and can take fractional values. However, if $w(T)$ is unknown or the sample size N is too small relative to $w(T)^2$, then one may opt to set k proportional to N instead, which is typically known while the Gaussian width or the doubling dimension may be unknown in practice.

3.6 General Lipschitz classifiers

The nearest neighbour example from the previous section generalises to the class of all Lipschitz classifiers (Gottlieb & Kontorovich, 2014; von Luxburg & Bousquet, 2004), examples of which, besides nearest neighbours, also include the support vector machine and others (von Luxburg & Bousquet, 2004). Let $\mathcal{H}_d$ and $\mathcal{H}_R$ be the sets of L_h-Lipschitz functions on $\mathcal{X}_d$ and $\mathcal{X}_R$ respectively. We take the exact same setting as previous margin-based analyses (Gottlieb et al., 2016), including an L_ℓ-Lipschitz loss functions bounded by $\bar{\ell}$. For instance $\bar{\ell}$ can be 1, since classification losses (e.g. the hinge loss), are surrogates to the 0–1 loss, so clipping at 1 makes sense, as it was done by Gottlieb et al. (2016). We restrict ourselves to the Euclidean space to leverage the computational advantages of random projections. In addition, we relax the requirement for the input space $\mathcal{X}_d$ to be bounded, and instead only require that most of the probability lies in a bounded subset. This relaxation is also applicable to our previous section.

Let $\mathbb{P}_X$ denote the marginal probability, and for each $\epsilon \geq 0$ we define

$$w_\epsilon(\mathcal{X}_d, \mathbb{P}_X) := \inf_{A \subseteq \mathcal{X}_d : \mathbb{P}_X(X \in A) \geq 1-\epsilon} w(A) \tag{39}$$

This lets us relax the boundedness assumption of the domain $\mathcal{X}_d$, instead we only need it to have a bounded subset A of $1 - \epsilon$ probability mass for $w_\epsilon(\mathbb{P}_X)$ to be finite. The familiar Gaussian width is recovered when $\epsilon = 0$, i.e. $w_0(\mathcal{X}_d, \mathbb{P}) = w(\mathcal{X}_d)$. In the sequel, we use the shorthand

$$\mathcal{X}_d^\epsilon := \{A \subseteq \mathcal{X}_d : \mathbb{P}(X \in A) \geq 1 - \epsilon, w_\epsilon(\mathcal{X}_d, \mathbb{P}_X) = w(A)\}.$$

In this setting, we have the following, proved in Appendix Section "General Lipschitz classifiers".

Proposition 13 *Consider the class of Lipschitz classifiers described above. We have*

$$D_k(g^*) \leq L_\ell L_h \mathrm{diam}(\mathcal{X}_d^\epsilon) \frac{w(\mathcal{X}_d^\epsilon)}{\sqrt{k}} + \epsilon \cdot \bar{\ell} \tag{40}$$

$$\mathcal{C}_{k,N}(\mathcal{G}_d) \leq L_\ell L_h \mathrm{diam}(\mathcal{X}_d^\epsilon) \frac{w(\mathcal{X}_d^\epsilon)}{\sqrt{k}} + \epsilon \cdot \bar{\ell} \tag{41}$$

Originally, the Lipschitz classifier (Gottlieb & Kontorovich, 2014) was proposed as a classification approach in doubling metric spaces. The analysis of Gottlieb and Kontorovich (2014) highlighted that the generalisation error can be expressed in terms of the doubling dimension of the metric space. As we commented in the Nearest Neighbour section a particular choice of k proportional to the square of the Gaussian width makes this connection explicit, while in contrast we are also free to choose other values of k. Another difference is in the methodological focus: In (Gottlieb & Kontorovich, 2014; Kontorovich & Weiss, 2015; Gottlieb et al., 2016), bounding the error in terms of a notion of intrinsic dimension was made possible due to a specific property of the Lipschitz class, by which the covering numbers of the function class are upper bounded in terms of the covering numbers of the input space. By contrast, in our strategy the starting point was to exploit random projection to obtain an auxiliary class of a lower complexity, and as such, the Lipschitz property of the classifier functions is not in generally required in our framework. Indeed, we have seen throughout the various examples in this section that the same starting point has drawn together some widely used regularisation schemes in the case of parametric models, as well as the Gaussian width in the nearest neighbour and Lipschitz classifier examples.

3.7 Turning compressive complexity into a regulariser

In several examples of the previous section, the upper bound on $\mathcal{C}_{k,N}$ has taken the form $\sup_{g \in \mathcal{G}_d} \mathcal{C}_k(g)$, where $\mathcal{C}_k$ is some function that only depends on the data through g. Structural Risk Minimisation (SRM) (Vapnik, 1998) is a classic approach that can be applied to turn the expression of $\mathcal{C}_k$ into a regulariser—this would ensure that ERM is confined to an appropriate subset of $\mathcal{G}_d$ that satisfy the compressibility constraint in our theorems.

For more complicated models, however, bounding the compressive complexity in a useful way may be difficult or out of reach. In the absence of a suitable analytic upper bound, in this section we show that one can instead estimate it from unlabelled data, whenever the loss function is Lipschitz, yielding semi-supervised regularisation algorithms that learn the regularisation term from an independent unlabelled data set. This recovers a form of consistency regularisation (Laine & Aila, 2017)—a semi-supervised technique widely used in practice—giving it a theoretical justification. We describe this in the sequel.

Exploiting the uniform nature of the bound of Theorem 4, we use structural risk minimisation (SRM). This will give us a regulariser whose general form comes from the compressive distortion of the function class, and which takes care of the required low-distortion constraint so the resulting predictor enjoys the guarantee stated in Theorem 4. The reason this works is that, by construction, a uniform bound is equivalent to the objective of a learning algorithm (as it can be iterated as many times as needed, so this algorithm enjoys the generalisation guarantee indicated by the bound).

Suppose we have an independent unlabelled data set drawn i.i.d. from the marginal distribution of the data. For each $\theta \in [0, \bar{\ell}]$, we define the class

$$\mathcal{G}_d^\theta := \left\{ g \in \mathcal{G}_d \,:\, \hat{D}_k(g) \le \theta \right\} \subseteq \mathcal{G}_d. \tag{42}$$

Note, these classes depend on the independent unlabelled data set, but not on the labelled data. Fix an increasing sequence $(\theta_i)_{i\in\mathbb{N}}$. This defines a nested sequence of subsets of the function class $\mathcal{G}_d$, as we have $\mathcal{G}_d^{\theta_1} \subseteq \mathcal{G}_d^{\theta_2} \subseteq ... \subseteq \mathcal{G}_d$. Let $(\mu_i)_{i\in\mathbb{N}}$ be an associated sequence of probability weights s.t. $\sum_{i\in\mathbb{N}} \mu_i \le 1$. By Theorem 4 applied to $\mathcal{G}_d^\theta$, for any fixed value of θ, we have uniformly for all $g \in \mathcal{G}_d^\theta$, w.p. $1-\delta$ that

$$E[g] \le \hat{E}_{\mathcal{T}_N}[g] + 2\theta + 2E_R[\hat{\mathcal{R}}_N(\mathcal{G}_R^\theta)] + 3\sqrt{\frac{\log(2/\delta)}{2N}} \tag{43}$$

where $\mathcal{G}_R^\theta$ is the RP-ed version of $\mathcal{G}_d^\theta$, and note that $\hat{\mathcal{R}}_N(\mathcal{G}_R^\theta) \le \hat{\mathcal{R}}_N(\mathcal{G}_R)$. We now use this bound for each $i \in \mathbb{N}$ with failure probabilities $\delta\mu_i$. By the union bound, w.p. $1-\delta$ uniformly for all $i \in \mathbb{N}$ and all $g \in \mathcal{G}_d^{\theta_i}$,

$$E[g] \le \hat{E}_{\mathcal{T}_N}[g] + 2\theta_i + 2E_R[\hat{\mathcal{R}}_N(\mathcal{G}_R)] + 3\sqrt{\frac{\log(2/\delta\mu_i)}{2N}}. \tag{44}$$

This suggests the following algorithm. For each $g \in \mathcal{G}_d$, let $i(g)$ denote the smallest integer such that $g \in \mathcal{G}_d^{\theta_{i(g)}}$; more precisely, $i(g) := \min\{i \in \mathbb{N} \,:\, \hat{D}_k(g) < \theta_i\}$. Define the following minimisation objective as a learning algorithm:

$$g^{reg} := \arg\min_{g\in\mathcal{G}_d} \left\{ \hat{E}_{\mathcal{T}_N}[g] + 2\theta_{i(g)} + 3\sqrt{\frac{\log(1/\mu_{i(g)})}{2N}} \right\}. \tag{45}$$

In practice, one can set $(\mu_i)_{i\in\mathbb{N}}$ as a uniform distribution on a finite sequence, so the last term becomes constant and omitted. Regarded as a guiding principle, the above suggests a practical algorithm using $\hat{D}_k(g)$ directly in place of its discretised version $\theta_{i(g)}$. We have the following guarantee about g^{reg}.

Theorem 14 *With probability at least* $1-\delta$,

$$E[g^{reg}] \le E[g^*] + 2\theta_{i(g^*)} + 2E_R[\hat{\mathcal{R}}_N(\mathcal{G}_R)] + 4\sqrt{\frac{\log(4/(\delta\mu_{i(g^*)}))}{2N}}. \tag{46}$$

Proof of Theorem 14 We apply the uniform bound of Eq. (44) with the choice $\theta := \theta_{i(g^{reg})}$, so

$$E[g^{reg}] \le_{1-\delta/2} \hat{E}_{\mathcal{T}_N}[g^{reg}] + 2\theta_{i(g^{reg})} + 2E_R[\hat{\mathcal{R}}_N(\mathcal{G}_R)] + 3\sqrt{\frac{\log(4/(\delta\mu_{i(g^{reg})}))}{2N}} \tag{47}$$

By the definition of g^{reg}, for any $g \neq g^{reg}, g \in \mathcal{G}_d$, the right hand side is further upper bounded as

$$\le \hat{E}_{\mathcal{T}_N}[g^*] + 2\theta_{i(g^*)} + 2E_R[\hat{\mathcal{R}}_N(\mathcal{G}_R)] + 3\sqrt{\frac{\log(4/(\delta\mu_{i(\theta_{g^*})}))}{2N}} \tag{48}$$

We subtract $E[g^*]$ from both sides, and use Hoeffding's inequality to bound $\hat{E}_{\mathcal{T}_N}[g^*] - E_{\mathcal{T}_N}[g^*]$, yielding

$$E[g^{reg}] - E[g^*] \leq \hat{E}_{\mathcal{T}_N}[g^*] + 2\theta_{i(g^*)} + 2E_R[\hat{\mathcal{R}}_N(\mathcal{G}_R)] + 3\sqrt{\frac{\log(4/(\delta\mu_{i(g^*)}))}{2N}} - E[g^*] \tag{49}$$

$$\leq_{1-\delta/2} 2\theta_{i(g*)} + 2E_R[\hat{\mathcal{R}}_N(\mathcal{G}_R)] + 4\sqrt{\frac{\log(4/(\delta\mu_{i(g^*)}))}{2N}} \tag{50}$$

Combining (47) and (50) by the union bound completes the proof. □

Comments. The bound contains $\theta_{i(g^*)}$, which, is an upper estimate of $\hat{D}_k(g^*)$. This might not be a quantity of particular interest in itself, but we can relate it to $D_k(g^*)$, as follows. Provided sufficient unlabelled data to ensure, for a given $\eta \in (0, 1)$, that $\sup_{g\in\mathcal{G}_d} |\hat{D}_k(g) - D_k(g)| \leq \eta$ w.p $1 - \delta$, then whenever we have $\hat{D}_k(g^*) \leq \theta_{i(g^*)}$ this also implies $D_k(g^*) \leq \theta_{i(g^*)} + \eta$ w.p. $1 - \delta$; consequently, with the overall probability of $1 - 2\delta$, we have

$$E[g^{reg}] \leq E[g^*] + 2\theta^* + 2E_R[\hat{\mathcal{R}}_N(\mathcal{G}_R)] + 4\sqrt{\frac{\log(4/(\delta\mu_{i(g^*)}))}{2N}}. \tag{51}$$

where $\theta^* = \theta_{i(g^*)} + \eta$ is our high probability upper estimate on $D_k(g^*)$. Thus, for the chosen $k < d$, if a learning problem exhibits small $D_k(g^*)$, and provided we have a large enough unlabelled set, then the algorithm (45) adapts to take advantage of this structure.

We have not elaborated here on how much unlabelled data would be needed. One can leverage and adapt the findings of Turner and Kabán (2023), where it was found (albeit in a deterministic model-compression setting) that the problem of ensuring a that η is as small as we like is in general statistically as difficult as the original learning problem, but it becomes surprisingly easy in many natural problem settings, namely when the compression only affects the predictions for a small number of sample points.

As a final comment, we assumed throughout that the choice of k is made before seeing the data, e.g. based on the available sample size N. Instead, if desired, one can pursue a hierarchical SRM to allow the value of k to be also determined from the training sample. The parameter k needs to be large enough to ensure that θ_{g^*} is sufficiently small, and it needs to be small enough to match the available sample size N in order to keep the Rademacher complexity term small.

4 Conclusions

We presented a framework to study the general question of how to discover and exploit such hidden benign traits when problem-specific prior knowledge is insufficient, using random projection's ability to expose structure. We considered both compressive learning and high dimensional learning, and give simple and general PAC bounds in the agnostic setting, in terms of some general notions of compressive distortion and compressive complexity that we introduced. We have also shown the tightness of our bounds when these quantities are small. The novel quantities of compressive distortion and compressive complexity take different forms in different learning tasks, and we instantiate these in several of these. This demonstrated their ability to capture and discover interpretable structural characteristics that make

high dimensional instances of these problems solvable to good approximation in a random linear subspace. In the examples considered, these turned out to resemble the margin, the margin distribution, the intrinsic dimension, the spectral decay of the data covariance, or the norms of parameters. In future work it will be interesting to use this strategy to discover benign structural traits in further PAC-learnable problems, and to develop regularised algorithms suggested by the bounds.

Appendix 1 Proof of Theorem 1

Proof By construction, all data live on the set S. Let $J \equiv \{j_1, \dots j_N\} \subseteq \{2, \dots d\}$ denote the set of indices of basis vectors that appear in the training set. The training set must have the form $\mathcal{T}_N = \{(X_n, Y_n) : n = 1, \dots, N\}$ with $(X_n, Y_n) = ((e_1 + e_{j_n})Y_n, Y_n)$ where $j_n \in J, n \in [N]$.

We define $h_{\text{bad}} \in \mathbb{R}^d$ with components $(h_{\text{bad}})_j, j = 1, \dots, d$ as the following

$$(h_{\text{bad}})_j \equiv \begin{cases} 1 & \text{if } j = 1 \\ 0 & \text{if } j \in J \\ -2 & \text{otherwise.} \end{cases} \tag{52}$$

Observe this is an ERM, since for all $n = 1, \dots, N$ we have $h_{\text{bad}}^T X_n = h_{\text{bad}}^T (e_1 + e_{j_n}) Y_n = (1 + 0) Y_n = Y_n$, so the training error of h_{bad} is zero.

Now, take a new input point $X = (e_1 + e_j)Y$; its correct target is Y. There are two cases: If $j \in J$ then we have $(h_{\text{bad}})^T X = (1 + 0)Y = Y$, so the prediction is correct. But if $j \notin J \cup \{1\}$ then $(h_{\text{bad}})^T X = (1 - 2)Y = -Y$, so the prediction is wrong. Thus, the generalisation error is the probability that, out of $d - 1$ basis vectors, a uniform sampling returns an element outside of J. The cardinality of J is at most N, hence we have

$$\mathbb{P}_{X,Y}\left\{h_{\text{bad}}^T XY \le 0\right\} = \frac{E[|\{2, \dots, d\} \setminus J|]}{d - 1} \ge \frac{d - 1 - N}{d - 1} = 1 - \frac{N}{d - 1}. \tag{53}$$

This completes the proof of the first part.

We now turn to the second part, considering the compressive ERM. The classes are separable by construction, so in $\mathbb{R}^d$ we are in the realisable case. Let us fix $\mathcal{T}_N$, and choose the smallest k for random projection to preserve realisability with high probability.

Let $\hat{h}_R \in \mathcal{H}_R$ be a compressive ERM, and $h^* \in \mathcal{H}_d$ the unknown best high dimensional classifier. Note that $Rh^* \in \mathcal{H}_R$, so we have $\hat{E}_{\mathcal{T}_N}[\mathbf{1}((\hat{h}_R)^T RXY \le 0)] \le \hat{E}_{\mathcal{T}_N}[\mathbf{1}(h^{*T} R^T RXY \le 0)]$, and we evaluate this further.

Fix $X \in S$. By the Johnson-Lindenstrauss lemma for dot products (Kaban, 2015), for any $\gamma \in (0, 1)$ it holds w.p. $1 - 2\exp(-k\gamma^2/8)$ that

$$\left| \frac{h^{*T}}{\|h^*\|_2} R^T R \frac{X}{\|X\|_2} Y - \frac{h^{*T}}{\|h^*\|_2} \frac{X}{\|X\|_2} Y \right| < \gamma. \tag{54}$$

Choose $\gamma := \left| \frac{h^{*T} X}{\|h^*\|_2 \|X\|_2} \right| = \frac{\sqrt{2}}{\sqrt{2} \cdot \sqrt{2}} = \frac{1}{\sqrt{2}}$, i.e. the normalised margin of h^* in the data support. By realisability with margin γ in the original space, we have $\frac{(h^*)^T X}{\|h^*\|_2 \|X\|_2} Y \ge \gamma$. This combined with Eq. (54) gives

$$h^{*T}R^TRXY > 0. \tag{55}$$

Taking union bound over the training examples, w.p. at least $1 - 2N\exp(-k\gamma^2/8)$, we have that (55) holds for all $(X_n, Y_n), n = 1, \ldots, N$ simultaneously. Hence, with the same probability, the training error of $\hat{h}_R$ is $\hat{E}_{\mathcal{T}_N}[\mathbf{1}(h^{*T}R^TRXY \leq 0)] = 0$.

By setting $2N\exp(-k\gamma^2/8) \leq \delta/2$, we have $k \geq k^* = \lceil \frac{8}{\gamma^2} \log \frac{4N}{\delta} \rceil = \lceil 16 \log \frac{4N}{\delta} \rceil$. Hence, for such values of k the problem remains realisable in the compressed space w.p. $\delta/2$. Therefore all compressive ERMs will have zero training error w.p. $1 - \delta/2$.

Now, to evaluate the genralisation error, we apply the fundamental theorem of statistical learning theory in the realisable case (Kearns & Vazirani, 1994; Vapnik, 1998), and use the fact that the VC dimension of $\mathcal{H}_R$ is k in this example. We have

$$\begin{aligned} \mathbb{P}_{\mathcal{T}_N}\left\{\exists h_R \in \mathcal{H}_R : \hat{E}_{\mathcal{T}_N}[\mathbf{1}(h_R^TRXY \leq 0)] = 0,\ \mathbb{P}_{X,Y}[h_R^TRXY \leq 0] \right. \\ \left. \leq 2\left(\frac{2eN}{k}\right)^k \exp\left(-\frac{\epsilon N}{2}\right)\right. \end{aligned} \tag{56}$$

Setting the r.h.s. to $\delta/2$ and combining with (55), w.p. $1 - \delta$ the following holds for any compressive ERM $\hat{h}_R \in \mathbb{R}^k$, $\mathbb{P}_{X,Y}[\hat{h}_R^TRXY \leq 0] \leq \frac{2}{N}\left(k \log \frac{2eN}{k} + \log \frac{4}{\delta}\right)$. □

Appendix 2 Proof of Lemma 5

Proof Choose $\mathcal{G}_R$ as the linear class of functions constructed from $\mathcal{G}_d$ such that $g_R \in \mathcal{G}_R$ has parameter $w_R \in \mathbb{R}^k$ equal to the least square solution of the system of equations $w_R^TRA = w^TA$, where $A \in \mathbb{R}^{d\times k}$ contains in its columns an orthonormal basis of the subspace V_k, and $w \in \mathbb{R}^d$ is the parameter of some $g \in \mathcal{G}_d$. Since R is full row-rank a.s., $RA \in \mathbb{R}^{k\times k}$ is invertible a.s., so $w_R = (RA)^{-T}A^Tw$.

Hence, for any point of the subspace, $X \in V_k$, we have $w_R^TRX = w^TX$; therefore $|\ell(w_R^TRX, Y) - \ell(w^TX, Y)| = 0$ for all $X \in V_k$, all $w \in \mathbb{R}^d$ and all $Y \in \{-1, 1\}$.

Using the above, given a pair of functions $g \in \mathcal{G}_d$ and $g_R \in \mathcal{G}_R$ we have

$$E_{X,Y}|(g\circ R - g_R)(X, Y)| = E_{X,Y}[\mathbf{1}(X \notin V_k)|(g\circ R - g_R)(X, Y)|] \tag{57}$$

$$\leq \bar{\ell} \cdot P_{X\sim\mathbb{P}_X}[X \notin V_k] = \bar{\ell} \cdot \theta. \tag{58}$$

Hence, the compressive distortion of the target g^* is bounded as

$$D_k(g^*) = E_R\left[\inf_{g_R\in\mathcal{G}_R} E_{X,Y}|(g\circ R - g_R)(X, Y)|\right] \leq \bar{\ell}\theta. \tag{59}$$

To prove (22), we have for the compressive complexity that

$$\frac{1}{N}\sum_{n=1}^{N}|(g\circ R - g_R)(X_n, Y_n)| = \frac{1}{N}\sum_{n=1}^{N}\mathbf{1}(X_n \notin V_k) \cdot |(g\circ R - g_R)(X_n, Y_n)| \tag{60}$$

$$\leq \bar{\ell} \cdot \frac{1}{N} \sum_{n=1}^{N} \mathbf{1}(X_n \notin V_k). \tag{61}$$

Consequently,

$$\begin{aligned} \mathcal{C}_{k,N}(\mathcal{G}_d) &= E_{\mathcal{T}_N \sim \mathbb{P}^N} \left[\sup_{g \in \mathcal{G}_d} \left\{ E_R \left[\inf_{g_R \in \mathcal{G}_R} \left\{ \frac{1}{N} \sum_{n=1}^{N} |(g \circ R - g_R)(X_n, Y_n)| \right\} \right] \right\} \right] \\ &\leq \bar{\ell} \cdot E_{\mathcal{T}_N \sim \mathbb{P}^N} \left[\frac{1}{N} \sum_{n=1}^{N} \mathbf{1}(X_n \notin V_k) \right] \end{aligned} \tag{62}$$

$$= \bar{\ell} \cdot \frac{1}{N} \sum_{n=1}^{N} P[X_n \notin V_k] = \bar{\ell}\theta \tag{63}$$

as required. □

Appendix 3 Proofs of Propositions for Section 3, and additional Corollaries

Thresholded linear models

$$\begin{aligned} D_k(g^*) &= E_R \left[\inf_{g_R \in \mathcal{G}_R} E_{(X,Y)\sim\mathbb{P}} \left[|g_R(RX, Y) - g^*(X, Y)| \right] \right] \\ &= E_R \inf_{h_R \in \mathcal{H}_R} E_{(X,Y)\sim\mathbb{P}} \left[|\mathbf{1}\{h_R^T RXY < 0\} - \mathbf{1}\{h^{*T} XY < 0\}| \right] \cdot \mathbf{1}\{k < d\} \\ &\leq E_R E_{(X,Y)\sim\mathbb{P}} \left[|\mathbf{1}\{h^{*T} R^T RXY < 0\} - \mathbf{1}\{h^{*T} XY < 0\}| \right] \cdot \mathbf{1}\{k < d\} \end{aligned} \tag{64}$$

Proof of Proposition 7

$$\leq E_X E_R \left[\mathbf{1}\{\text{sign}(h^{*T} R^T RX) \neq \text{sign}(h^{*T} X)\} \right] \cdot \mathbf{1}\{k < d\}. \tag{65}$$

Eq. (64) holds because both h_R and Rh^* belong to $\mathcal{H}_R$. Equation (65) tells us that the compressive distortion is related to the average effect that the input perturbation has on the decision boundary. In conjuction with Theorem 2, this means that the smaller this effect, the better for the compressive classifier.

The expectation w.r.t. R that appears in (65) was extensively studied by Durrant and Kabán (2013),Kabán and Durrant (2020) when R has i.i.d. Gaussian or sub-gaussian entries, and is known to be bounded as

$$E_X E_R \left[\mathbf{1}\{\text{sign}(h^T R^T RX)\} \neq \text{sign}(h^T X) \right] \leq E_X \left[\exp\left(\frac{-k \cos^2(\measuredangle_X^h)}{8} \right) \right]. \tag{66}$$

Moreover, by property 2.3, Eq. (65) also implies a bound for the compressive complexity, which again turns out to be a function of the margin distribution.

$$\mathcal{C}_{k,N}(\mathcal{G}_d) \le E_{\mathcal{T}_N\sim\mathbb{P}^N}\left[\sup_{h\in\mathcal{H}_d} \hat{E}_{X\sim\mathcal{T}_N} E_R\left[\mathbf{1}\left\{\text{sign}(h^{*T}R^TRX) \neq \text{sign}(h^{*T}X)\right\}\right]\right] \cdot \mathbf{1}(k<d)$$

$$\le E_{\mathcal{T}_N\sim\mathbb{P}^N}\left[\sup_{h\in\mathcal{H}_d} \hat{E}_{X\sim\mathcal{T}_N} \exp\left(\frac{-k\cos^2(\measuredangle_X^h)}{8}\right)\right] \cdot \mathbf{1}(k<d) \tag{67}$$

$$\le E_X\left[\sup_{h\in\mathcal{H}_d} \exp\left(\frac{-k\cos^2(\measuredangle_X^h)}{8}\right)\right] \cdot \mathbf{1}(k<d), \tag{68}$$

where (67) follows from (66), and (68) from Jensen's inequality. □

Corollary 15 (Binary linear classification) *Consider the linear function class as above, and let* $\mathcal{G}_d = \ell_{01}\circ\mathcal{H}_d$. *Take any* $k \le d, \delta \in (0,1)$.

a) Suppose that the best classifier in the class, h^*, *satisfies* $E_X\left[\exp\left(\frac{-k\cos^2(\measuredangle_X^{h^*})}{8}\right)\right] \cdot \delta(k<d) \le \theta = \theta(k)$. *Then, w.p.* $1-2\delta$, *the compressive ERM satisfies*

$$E[\hat{g}_R] \le E[g^*] + \theta + \xi(k, g^*, \delta) + 62\sqrt{\frac{k}{N}} + 4\sqrt{\frac{\log(3/\delta)}{2N}}. \tag{69}$$

where ξ *is defined in Theorem 2.*

b) If $E_X\left[\sup_{h\in\mathcal{H}_d} \exp\left(\frac{-k\cos^2(\measuredangle_X^h)}{8}\right)\right] \cdot \mathbf{1}(k<d) \le \theta$, *then, for any* $\delta > 0$, *w.p.* $1-\delta$ *the following holds uniformly for all* $g \in \ell_{01}\circ\mathcal{H}_d$

$$E[g] \le \hat{E}[g] + 2\theta + 62\sqrt{\frac{k}{N}} + 3\sqrt{\frac{\log(2/\delta)}{2N}}. \tag{70}$$

Proof We plug the expressions from Proposition 7 into the bounds of Theorems 2 and 4 respectively, and bound the Rademacher complexity of the compressive function class with its VC dimension (with explicit constant given by (Wolf, 2020, Corollary 1.25)) as $\hat{\mathcal{R}}_N(\mathcal{G}_R) \le 31\sqrt{\frac{k}{N}}$. Putting everything together completes the proof. □

Preliminary Lemmas for proving the results of Sects. 3.2-3.3

The following lemma is inspired by (Slawski, 2018), with a concise proof tailored to Gaussian RP so we can deploy a bound by Halko, Martinsson, Tropp (Halko et al., 2011).

Lemma 16 *Given a matrix* $W^* \in \mathbb{R}^{d\times m}$, *a random vector* $X \in \mathbb{R}^d$ *with* $\Sigma := E[XX^T]$, *and a random matrix* $R \in \mathbb{R}^{k\times d}, k \le d$ *with i.i.d. 0-mean Gaussian entries. For any* $p \in \mathbb{N}$ *s.t.* $2 \le p \le k-2$ *and* $k \le \text{rank}(\Sigma)$, *we have*:

$$E_R\left[\inf_{\tilde{W}\in\mathbb{R}^{m\times k}} E_X\|\tilde{W}^T RX - W^{*T}X\|\right]$$
$$\leq \|W^*\|_F \min\left\{\left(1+\sqrt{\frac{k-p}{p-1}}\right)\sqrt{\lambda_{k-p+1}(\Sigma)} + \frac{e\sqrt{k}}{p}\sqrt{\sum_{j>k-p}\lambda_j(\Sigma)}\right\} \tag{71}$$

As commented by Halko et al. (2011), the parameter p is an oversampling factor, for a target dimension $k-p$. If we increase p the second term declines quicker than the first. If p is chosen proportional to k then the first term is proportional to $\sqrt{\lambda_{k-p+1}}$ (which is the minimum $(k-p)$-rank approximation error of $\Sigma^{1/2}$ in the spectral norm, by the Eckart-Young-Mirsky Theorem), and the second term decreases at the rate $k^{-1/2}$. The spectral tail $\sqrt{\sum_{j>k-p}\lambda_j(\Sigma)}$ in the second term is the minimum $(k-p)$-rank approximation error in the Frobenius norm.

Proof of Lemma 16 By Jensen's inequality,

$$E_R\left[\inf_{W_R\in\mathbb{R}^{m\times k}} E_X\|W_R^T RX - W^{*T}X\|\right] \leq E_R \inf_{W_R\in\mathbb{R}^{m\times k}} \left\{E_X\|W_R^T RX - W^{*T}X\|^2\right\}^{1/2}. \tag{72}$$

The infimum is at $W_R^T = W^{*T}\Sigma R^T(R\Sigma R^T)^{-1}$, so

$$\text{Eq. (72)} = E_R\left[E_X(W^{*T}\Sigma R^T(R\Sigma R^T)^{-1}RX - W^{*T}X)^2\right]^{1/2} \tag{73}$$

$$= E_R\left[W^{*T}\Sigma W^* - W^{*T}\Sigma R^T(R\Sigma R^T)^{-1}R\Sigma W^*\right]^{1/2} \tag{74}$$

$$\leq \|W^*\|_F E_R\left[\lambda_{\max}(\Sigma^{1/2}(I-\Sigma^{1/2}R^T(R\Sigma^T R^T)^{-1}R\Sigma^{1/2})\Sigma^{1/2})\right]^{1/2} \tag{75}$$

$$\leq \|W^*\|_F E_R\left[\lambda_{\max}((I-\Sigma^{1/2}R^T(R\Sigma^T R^T)^{-1}R\Sigma^{1/2})\Sigma)\right]^{1/2} \tag{76}$$

$$\leq \|W^*\|_F E_R\left[\lambda_{\max}((I-\Sigma^{1/2}R^T(R\Sigma^T R^T)^{-1}R\Sigma^{1/2})\Sigma^{1/2})\right] \tag{77}$$

by the idempotent property of projection matrices. In the above, $\lambda_{\max}(\cdot)$ denotes the largest eigenvalue of the matrix in its argument, and we will use $\lambda_j(\cdot)$ to denote the j-th largest eigenvalue.

Now, using (Halko et al, 2011, Theorem 10.6), for any $p\in\mathbb{N}$ s.t. $2\leq p\leq k-2$ and $k\leq \text{rank}(\Sigma)$, this is bounded by:

$$\text{Eq. (77)} \leq \|W^*\|_F\left\{\left(1+\sqrt{\frac{k-p}{p-1}}\right)\sqrt{\lambda_{k-p+1}(\Sigma)} + \frac{e\sqrt{k}}{p}\sqrt{\sum_{j>k-p}\lambda_j(\Sigma)}\right\}$$

□

The following lemma gives a dimension-dependent bound on the empirical Rademacher complexity of bounded Lipschitz functions of a linear class when the parameter domain is unconstrained.

Lemma 17 *Let* $\mathcal{F}_k = \{x \to f(w^T x) \in [0,1] : x \in \mathbb{R}^k\}$, *where* f *is* 1*-Lipschitz and bounded by* 1. *Then,* $\hat{\mathcal{R}}_N(\mathcal{F}_k) \leq c\sqrt{\frac{k}{N}}$, *where* $c \leq 92$.

Proof of Lemma 17 By Dudley's entropy integral inequality (Dudley, 1999), the Rademacher complexity of any [0, 1]-valued function class can be bounded in terms of covering numbers,

$$\hat{\mathcal{R}}_N(\mathcal{F}_k) \leq 12 \int_0^1 \sqrt{\frac{\log \mathcal{N}(\alpha, \mathcal{F}_k, \|\cdot\|_2)}{N}} d\alpha \tag{78}$$

where $\|\cdot\|_2$ is the $\mathcal{L}_2$-norm with respect to the empirical measure i.e. for an $f \in \mathcal{F}_k$, $\|f\|_2 = \sqrt{\frac{1}{N}\sum_{n=1}^N f^2(X_n)}$.

The covering number can be further bounded in terms of the fat shattering dimension[1]. We use a result of (Alon et al., 1997) (see also Theorem 2.18 of Mendelson (2003)), which yields for every sample and any scale $\alpha \in (0,1)$:

$$\mathcal{N}(\alpha, \mathcal{F}_k, \|\cdot\|_2) \leq \left(\frac{13}{2\alpha}\right)^{20 \cdot \mathrm{fat}_{96\alpha}(\mathcal{F}_k)} \tag{79}$$

where $\mathrm{fat}_\gamma(\cdot)$ is the fat shattering dimension of the function class, and the constants have been computed by (Guermeur, 2017, Lemma 3) (see also (Lauer, 2019, Lemma 6)).

It is known that linear function classes have fat shattering dimension upper bounded by their input dimension (Gurvits & Koiran, 1995) for any γ, and composition with a Lipschitz function does not change the fat shattering dimension by more than a constant (Gurvits & Koiran, 1995).

Plugging this back, Eq. (78) is bounded as:

$$\hat{\mathcal{R}}_N(\mathcal{F}_k) \leq 12 \int_0^1 \sqrt{\frac{20\, k \log \frac{13}{2\alpha}}{N}} d\alpha = c\sqrt{\frac{k}{N}} \tag{80}$$

where $c = 12\sqrt{20(\log(13/2)+1)} \leq 92$. □

Linear models with bounded Lipschitz loss

Proof of Proposition 8 Recall that R has i.i.d. 0-mean 1/k-variance Gaussian entries. So for any $p \in \mathbb{N}$ s.t. $2 \leq p \leq k-2$ and any $k \leq \mathrm{rank}(\Sigma)$ we have

[1] The fat shattering dimension is a measure of the complexity of a real valued function class. Definition. Let $\gamma > 0$ be fixed, and let $\mathcal{F}$ be a function class. We say that F γ-shatters a set $A \subset X$ if $\exists s : A \to \mathbb{R}$ s.t. $\forall E \subseteq A, \exists f_E \in \mathcal{F}$ satisfying that $\forall x \in A \backslash E, f_E(x) \leq s(x) - \gamma$ and $\forall x \in E, f_E(x) \geq s(x) + \gamma$. The maximum cardinality of $A \subseteq X$ that is γ-shattered by $\mathcal{F}$ is defined as the fat-shattering dimension of $\mathcal{F}$, denoted $\mathrm{fat}_\gamma(\mathcal{F})$.

$$D_k(g^*) = E_R\left[\inf_{h_R \in \mathcal{H}_R} E_{(X,Y)}|\ell(h_R^T RX, Y) - \ell(h^{*T}X, Y)|\right] \tag{81}$$

$$\leq L_\ell E_R\left[\inf_{h_R \in \mathcal{H}_R} E_X|h_R^T RX - h^{*T}X|\right] \cdot \mathbf{1}(k < d) \tag{82}$$

$$\leq L_\ell \|h^*\|_2 \cdot \Xi(k, p, \{\lambda_j(\Sigma)\}_j) \cdot \mathbf{1}(k < d). \tag{83}$$

where

$$\Xi(k, p, \{\lambda_j(\Sigma)\}_j) := \left(1 + \sqrt{\frac{k-p}{p-1}}\right)\sqrt{\lambda_{k-p+1}(\Sigma)} + \frac{e\sqrt{k}}{p}\sqrt{\sum_{j>k-p}\lambda_j(\Sigma)} \tag{84}$$

This follows by Lemma 16, which made use of (Halko et al., 2011, Theorem 10.6). As noted by Halko et al. (2011), with the choice of p of order k, the second term on the right hand side of (84) decreases as $1/\sqrt{k}$.

Moreover, by property 2.3 applied to (83), we also have the following upper bound on the compressive complexity

$$\mathcal{C}_{k,N}(\mathcal{G}_d) \leq L_\ell E_{\mathcal{T}_N \sim \mathbb{P}^N}[\Xi(k, p, \{\lambda_j(\hat{\Sigma})\}_j)] \cdot \sup_{h \in \mathcal{H}_d} \|h\|_2 \cdot \mathbf{1}(k < d) \tag{85}$$

where the function Ξ is defined in (84). □

Corollary 18 (Linear models with bounded Lipschitz loss) *Let $\mathcal{G}_d$ be the class of generalised linear models of the form $\mathcal{G}_d = \ell \circ \mathcal{H}_d$, where $\mathcal{H}_d = \{x \to h^T x : h, x \in \mathbb{R}^d\}$, and the loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, \bar{\ell}]$ is L_ℓ-Lipschitz in its first argument. Let $\mathcal{T}_N = \{(X_n, Y_n)_{n=1}^N\} \sim \mathbb{P}_d^N$ be a training set in $\mathcal{X}_d \times \mathcal{Y}$, where $\mathbb{P}_d$ satisfies $\mathrm{Tr}(E_{X \sim \mathbb{P}_d}[XX^T]) < \infty$. Take ant $k \leq d, \delta \in (0, 1)$.*

a) Suppose that $\|h^*\|_2 \leq \tau = \tau(k)$. Then, with probability $1 - 2\delta$, the compresive ERM satisfies

$$\begin{aligned} E[\hat{g}_R] \leq E[g^*] + L_\ell \cdot \tau \cdot \Xi(k, p, \{\lambda_j(\Sigma))\}_j \cdot \mathbf{1}(k < d) + \bar{\ell} \cdot \xi(k, g^*, \delta) \\ +184\bar{\ell}\sqrt{\frac{k}{N}} + 4\bar{\ell}\sqrt{\frac{\log(3/\delta)}{2N}} \end{aligned} \tag{86}$$

b) If $\sup_{h \in \mathbb{R}^d} \|h\|_2 \leq \tau = \tau(k)$, then, w.p. $1 - \delta$ we have uniformly for all $g \in \mathcal{G}_d$ that

$$\begin{aligned} E[g] \leq \hat{E}_{\mathcal{T}_N}[g] + 2L_\ell \cdot \tau \cdot E_{\mathcal{T}_N}[\Xi(k, p, \{\lambda_j(\hat{\Sigma})\}_j)] \cdot \mathbf{1}(k < d) + 184\bar{\ell}\sqrt{\frac{k}{N}} \\ +3\bar{\ell}\sqrt{\frac{\log(2/\delta)}{2N}}. \end{aligned} \tag{87}$$

Proof We plug the expressions from Proposition 8 in the bounds of Theorems 2-4, and bound the Rademacher complexity of the reduced class. There is no constraint on the parameters or the input domain, but we exploit that the loss function is bounded, and by

Lemma 17 given in the Appendix we have $\hat{\mathcal{R}}_N(\mathcal{G}_R) \leq \bar{\ell}\hat{\mathcal{R}}_N(\mathcal{G}_R/\bar{\ell}) \leq 92\bar{\ell}\sqrt{\frac{k}{N}}$. Putting everything together, completes the proof. □

Two-layer perceptron

Proof of Proposition 10 For R having i.i.d. Gaussian entries, the compressive distortion can be bounded similarly as before, using the Lipschitz property of ℓ and ϕ, along with Hölder's inequality, as follows.

$$D_k(g^*) = E_R\left[\inf_{W_R\in\mathbb{R}^{m\times k}, v_R\in\mathbb{R}^m} E_{(X,Y)}|\ell\left(\sum_{i=1}^m (v_R)_i\phi((w_R)_i^T RX), Y\right) - \ell\left(\sum_{i=1}^m v_i^*\phi(w_i^{*T}X), Y\right)|\right]$$
$$= L_\ell E_R\left[\inf_{W_R\in\mathbb{R}^{m\times k}} E_X|\sum_{i=1}^m v_i^*\phi((w_R)_i^T RX) - \sum_{i=1}^m v_i^*\phi(w_i^{*T}X)|\right]\cdot\mathbf{1}(k<d) \tag{88}$$

$$\leq L_\ell\|v^*\|_s E_R\left[\inf_{W_R\in\mathbb{R}^{m\times k}} E_X\left(\sum_{i=1}^m|\phi((w_R)_i^T RX) - \phi(w_i^{*T}X)|^q\right)^{1/q}\right]\cdot\mathbf{1}(k<d) \tag{89}$$

$$\leq L_\ell L_\phi\|v^*\|_s E_R\left[\inf_{W_R\in\mathbb{R}^{m\times k}} E_X\|W_R^T RX - W^{*T}X\|_q\right]\cdot\mathbf{1}(k<d) \tag{90}$$

where $s, q \geq 1, 1/s + 1/q = 1$, and the matrices W^* and W_R have the parameter vectors w_i^* and $(w_R)_i$ in their i-th columns. For simplicity, let us choose $s = q = 2$, so by Lemma 16 we have the following upper bound on (90), for any $p \in \mathbb{N}$ s.t. $2 \leq p \leq k-2$

$$D_k(g^*) \leq L_\ell L_\phi\|v^*\|_2\|W^*\|_F\cdot\Xi(k,p,\{\lambda_j(\Sigma)\}_j)\cdot\mathbf{1}(k<d) \tag{91}$$

where $\Xi(k,p,\{\lambda_j(\Sigma)\}_j)$ is the expression defined in Eq. (84).

Moreover, noting that in (91) the effect of the predictor factorises from that of the data distribution, by Property 2.3 we also have an upper bound on the compressive complexity of the original class,

$$\mathcal{C}_{k,N}(\mathcal{G}_d) \leq L_\ell L_\phi E_{\mathcal{T}_N\sim\mathbb{P}^N}[\Xi(k,p,\{\lambda_j(\hat{\Sigma})\}_j)]\cdot\sup_{v,W}\|v\|_2\|W\|_F\cdot\mathbf{1}(k<d) \tag{92}$$

where $\Xi(k,p,\{\lambda_j(\hat{\Sigma})\}_j)$ is defined in Eq. (84). □

Recall $\mathcal{H}_d = \{x \to \sum_{i=1}^m v_i\phi(w_i^T x) : x \in \mathcal{X}_d, \|v\|_1 \leq 1\}$ is the class of classic two-layer perceptrons, and take $\phi : \mathbb{R} \to [-b, b]$ to be an L_ϕ-Lipschitz anti-symmetric activation function (i.e. $\phi(-u) = -\phi(u), \forall u \in \mathbb{R}$; for instance *tanh*). A bounded activation function is chosen here for convenience, to allow us to easily work with un-regularised input layer weights—since the RP itself exerts a regularisation effect. Then we have the following.

Corollary 19 (Two-layer perceptron) *Let $\mathcal{H}_d$ be the class of 2-layer networks as above, and $\mathcal{G}_d = \ell\circ\mathcal{H}_d$, and $\ell : \mathcal{Y}\times\mathcal{Y} \to [0,\bar{\ell}]$ an L_ℓ-Lipschitz loss*

function. Let $\mathcal{T}_N = \{(X_n, Y_n)_{n=1}^N \sim \mathbb{P}_d^N\}$ *be a training set in* $\mathcal{X}_d \times \mathcal{Y}$*, where* $\mathbb{P}_d$ *satisfies* $\mathrm{Tr}(E_{X\sim\mathbb{P}_d}[XX^T]) \leq \infty$*. Take any* $k \leq d, \delta \in (0,1)$*.*

a) Suppose that $\|v^*\|_2 \|W^*\|_F \leq \tau = \tau(k)$*. Then, with probability* $1-2\delta$*, the compressive ERM satisfies*

$$E[\hat{g}_R] \leq E[g^*] + L_\ell L_\phi \|v^*\|_2 \|W^*\|_F \cdot \Xi(k, p, \{\lambda_j(\Sigma)\}_j) \cdot \mathbf{1}(k < d) + \bar{\ell} \cdot \xi(k, g^*, \delta) + 286 L_\ell b \sqrt{\frac{k}{N}} + 4\bar{\ell}\sqrt{\frac{\log(3/\delta)}{2N}} \tag{93}$$

b) Suppose that $\sup_{v,W} \|v\|_2 \|W\|_F \leq \tau = \tau(k)$*. Then, w.p.* $1-\delta$ *we have uniformly for all* $g \in \mathcal{G}_d$*,*

$$E[g] \leq \hat{E}_{\mathcal{T}_N}[g] + 2L_\ell \cdot \tau \cdot E_{\mathcal{T}_N}[\Xi(k, p, \{\lambda_j(\hat{\Sigma})\}_j)] \cdot \mathbf{1}(k < d) + 268 L_\ell b \sqrt{\frac{k}{N}} + 3\bar{\ell}\sqrt{\frac{\log(2/\delta)}{2N}} \tag{94}$$

Proof We plug the expressions from Proposition 10 into Theorems 2-4, and bound the Rademacher complexity of the class of compressive networks. Since the first layer weights are unconstrained, we use the boundedness of $\phi(\cdot)$ to do this. Assume $\|v\|_1 \leq 1$, so we can use the property of empirical Rademacher complexities by which for any class H it hold that $\hat{\mathcal{R}}_N(\mathrm{conv}(H)) = \hat{\mathcal{R}}_N(H)$ (Bartlett & Mendelson, 2002). Using this combined with Talagrand's contraction lemma,

$$\hat{\mathcal{R}}_N(\mathcal{G}_R) = \hat{\mathcal{R}}_N(\ell \circ \mathcal{H}_R) \leq L_\ell \hat{\mathcal{R}}_N(\mathcal{H}_R) = L_\ell 2b \hat{\mathcal{R}}_N(\mathcal{F}_R), \tag{95}$$

and $\mathcal{F}_R = \{x \mapsto \phi(w^T x)/(2b) + 1/2 : \mathbb{R}^k \to [0,1] \text{ s.t. } w \in \mathbb{R}^k, x = Rx, x \in \mathcal{X}_d\}$. We bound the empirical Rademacher complexity of $\mathcal{F}_R$ using the fact that this class has a bounded range of values. Using Lemma 17 we have $\hat{\mathcal{R}}_N(\mathcal{F}_R) \leq 92\sqrt{\frac{k}{N}}$, and plugging back we have $\hat{\mathcal{R}}_N(\mathcal{G}_R) \leq L_\ell 184 b \sqrt{\frac{k}{N}}$. Putting everything together completes the proof. □

Quadratic models

Proof of Proposition C.5 We will first consider the case where A^* is positive semi-definite, so all of its eigenvalues are non-negative.

By the Lipschitz property of ℓ, and using Jensen's inequality, we have

$$D_k(g^*) = E_R\left[\inf_{\tilde{A}\in\mathcal{M}_k} E_{X,Y}[\ell(X^T R^T \tilde{A} R X, Y) - \ell(X^T A^* X, Y)]\right] \tag{96}$$

$$\leqslant L_\ell E_X E_R\left[|X^T R^T R A^* R^T R X - X^T A^* X|\right] \tag{97}$$

Let a_i be the i-th column of $A^{*1/2}$. Then,

$$\text{Eq. (97)} = L_\ell E_{X,R}[|\sum_{i\geq 1}\{(a_i^T R^T R X)^2 - (a_i^T X)^2\}|] \tag{98}$$

$$= L_\ell E_{X,R}[|\sum_{i\geq 1}(a_i^T R^T RX - a_i^T X)(a_i^T R^T RX + a_i^T X)|] \tag{99}$$

$$\leq L_\ell E_{X,R}[\sum_{i\geq 1}|a_i^T R^T RX - a_i^T X| \cdot |a_i^T R^T RX + a_i^T X|] \tag{100}$$

$$\leq L_\ell \sum_{i\geq 1} E_{X,R}[|a_i^T R^T RX - a_i^T X| \cdot |a_i^T R^T RX + a_i^T X|] \tag{101}$$

$$\leq L_\ell \sum_{i\geq 1} \left\{E_{X,R}[(a_i^T R^T RX - a_i^T X)^2] \cdot E_{X,R}[(a_i^T R^T RX + a_i^T X)^2]\right\}^{1/2}. \tag{102}$$

where the last line used the Cauchy-Schwartz inequality.

Now, the first expectation is of the form we encountered before, and the second expectation can be treated similarly. We can use Lemma 2 of Kabán (2014) to compute matrix expectations, as

$$a_i^T E_{X,R}[R^T RXX^T R^T R]a_i = \frac{1}{k}a_i^T((k+1)\Sigma + \mathrm{Tr}(\Sigma)I_d)a_i. \tag{103}$$

where we denoted $\Sigma = E[XX^T]$. Note also that $E[R^T R] = I_d$. So after some algebra we have

$$\begin{aligned}\text{Eq. } (C51) = L_\ell \sum_{i\geq 1} \Big\{ \Big(\frac{1}{k}a_i^T\Sigma a_i + \frac{1}{k}\mathrm{Tr}(\Sigma)\|a_i\|^2\Big) \\ \cdot\Big(\frac{1}{k}a_i^T\Sigma a_i + \frac{1}{k}\mathrm{Tr}(\Sigma)\|a_i\|^2 + 4a_i^T\Sigma a_i\Big)\Big\}^{1/2}\end{aligned} \tag{104}$$

$$\leq L_\ell \sum_{i\geq 1}\sqrt{\frac{1}{k^2}+\frac{5}{k}}\|a_i\|^2\mathrm{Tr}(\Sigma) \tag{105}$$

$$= \sqrt{\frac{4}{k^2}+\frac{5}{k}}L_\ell \sum_{i\geq 1} A^*_{ii}\mathrm{Tr}(\Sigma) \tag{106}$$

$$\leq \sqrt{\frac{4}{k^2}+\frac{5}{k}}L_\ell \mathrm{Tr}(\Sigma)\mathrm{Tr}(A^*) \tag{107}$$

Finally, if A^* is not positive definite, recall that it is symmetric and any symmetric matrix can be written as $A^* = A^*_+ - A^*_-$, where A^*_+, A^*_- are positive semi-definite. Indeed, writing $A^* = U\Lambda U^T$ for the SVD of A^*, and decomposing $\Lambda = \Lambda_+ + \Lambda_-$ where Λ_+ and Λ_- contain the positive and the absolutes of the negative eigenvalues of A^* respectively and their remaining eigenvalues area zero, we have $A^*_+ = U\Lambda_+ U^T$ and $A^*_- = U\Lambda_- U^T$. By the triangle inequality,

$$|X^T R^T RA^* R^T RX - X^T A^* X| \leqslant |X^T R^T RA^*_+ R^T RX - X^T A^*_+ X| \tag{108}$$

$$+ |X^T R^T RA^*_- R^T RX - X^T A^*_- X| \tag{109}$$

We invoke (107) twice, i.e. for A^*_+ and A^*_- respectively, and note that $\mathrm{Tr}(A^*_+) + \mathrm{Tr}(A^*_-) = \|A^*\|_*$ is the nuclear norm of A^*. This yields

$$D_k(g^*) \leq \sqrt{\frac{4}{k^2} + \frac{5}{k}} L_\ell \mathrm{Tr}(\Sigma) \|A^*\|_*. \tag{110}$$

□

Moving on to the compressive complexity, and noting the factorised form of $D_k(g^*)$, by Property 2.3 we also have

$$\mathcal{C}_{k,N}(\mathcal{G}_d) \leq \sqrt{\frac{4}{k^2} + \frac{5}{k}} L_\ell \mathrm{Tr}(\Sigma) \sup_{A \in \mathcal{M}_d} \|A\|_* \tag{111}$$

Corollary 20 (Quadratic classifier learning) *Let $\mathcal{G}_d$ be the class $\mathcal{G}_d = \ell \circ \mathcal{H}_d$, where $\mathcal{H}_d = \{x \to x^T A x : A \in \mathcal{M}_d, x \in \mathbb{R}^d\}$, $\mathcal{M}_d$ is the set of $d \times d$ symmetric matrices, and $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, \bar{\ell}]$ is L_ℓ-Lipschitz in its first argument. Let $\mathcal{T}_N = \{(X_n, Y_n)_{n=1}^N\} \sim \mathbb{P}_d^N$ be a training set in $\mathcal{X}_d \times \mathcal{Y}$, where $\mathbb{P}_d$ satisfies $\mathrm{Tr}(E_{X \sim \mathbb{P}_d}[XX^T]) < \infty$. Take any $k \leq d, \delta \in (0, 1)$.*

a) Suppose that $\|A^*\|_* \leq \tau = \tau(k)$. Then, with probability $1 - 2\delta$, the compresive ERM satisfies

$$\begin{aligned} E[\hat{g}_R] \leq E[g^*] + L_\ell \cdot \tau \cdot \sqrt{\frac{4}{k^2} + \frac{5}{k}} L_\ell \mathrm{Tr}(\Sigma) \cdot \mathbf{1}(k < d) + \bar{\ell} \cdot \xi(k, g^*, \delta) \\ + 184 \bar{\ell} \sqrt{\frac{k(k+1)}{2N}} + 4 \bar{\ell} \sqrt{\frac{\log(3/\delta)}{2N}} \end{aligned} \tag{112}$$

b) If $\sup_{A \in \mathcal{M}_d} \|A\|_* \leq \tau = \tau(k)$, then, w.p. $1 - \delta$ we have uniformly for all $g \in \mathcal{G}_d$ that

$$E[g] \leq \hat{E}_{\mathcal{T}_N}[g] + 2L_\ell \cdot \tau \cdot \sqrt{\frac{4}{k^2} + \frac{5}{k}} \mathrm{Tr}(\Sigma) \cdot \mathbf{1}(k < d) + 184 \bar{\ell} \sqrt{\frac{k(k+1)}{2N}} + 3 \bar{\ell} \sqrt{\frac{\log(2/\delta)}{2N}}. \tag{113}$$

Proof Note that any $h \in \mathcal{H}_k$ has the form $h(X) = X^T A X = \sum_{i=1}^k \sum_{j=1}^k A_{ij} X_i X_j$, where X_i and X_j are the i-th and j-th feature components of the point X. Hence H_d is equivalent to a linear model over a $k(k+1)/2$-dimensional instance space, so we can apply the Rademacher complexity bound from the previous section, yielding $\hat{\mathcal{R}}_N(\mathcal{G}_R) \leq 92 \ell \sqrt{\frac{k(k+1)}{2}}$. Plugging this, along with the upper bounds obtained on $D_k(g^*)$ and $\mathcal{C}_{k,N}(\mathcal{G}_d)$ into the general Theorems 2 and 4 respectively completes the proof. □

Nearest neighbours classification

Proof of Proposition 12 We will need the following result by Gordon (1985).

Lemma 21 (Gordon) *Let $T \subseteq \mathbb{S}^{d-1}$, and R with entries $(R_{ij})_{i=1,\ldots k, j=1,\ldots d} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, 1/k)$. Then,*

$$E_R\left[\sup_{x\in T}|\|Rx\|_2-1|\right]\leq\frac{w(T)}{\sqrt{k}} \tag{114}$$

where $w(T)=E_{r\sim\mathcal{N}(0,1)}\sup_{t\in T}\{\langle r,t\rangle\}$ *denotes the Gaussian width of the set in its argument.*

We proceed to bound compressive distortion,

$$D_k(g^*)=E_R\inf_{g_R\in\mathcal{G}_R}E_{(x,y)}|g_R(Rx,y)-g^*(x,y)| \tag{115}$$

$$\leq E_RE_{(x,y)}|g_R^*(Rx,y)-g^*(x,y)|. \tag{116}$$

On a given sample, we have $|g_R(RX,Y)-g(X,Y)|$

$$\begin{aligned}&\leq\frac{1}{2\gamma}|\|RX-RN_R^-(X)\|-\|RX-RN_R^+(X)\|-\|X-N^-(X)\|+\|X-N^+(X)\||\\&\leq\frac{1}{2\gamma}\big(|\|RX-RN_R^-(X)\|-\|X-N^-(X)\||+|\|RX-RN_R^+(X)\|-\|X-N^+(X)\||\big).\end{aligned} \tag{117}$$

Note that $\|RX-RN_R^{\pm}(X)\|\leq\|RX-RN^{\pm}(X)\|$, and $\|X-N^{\pm}(X)\|\leq\|X-N_R^{\pm}(X)\|$, hence (117) is further bounded as

$$\begin{aligned}&\leq\frac{1}{2\gamma}(\max\left\{|\|RX-RN^-(X)\|-\|X-N^-(X)\||,|\|RX-RN_R^-(X)\|-\|X-N_R^-(X)\||\right\}\\&+\max\left\{|\|RX-RN^+(X)\|-\|X-N^+(X)\||,|\|RX-RN_R^+(X)\|-\|X-N_R^+(X)\||\right\})\end{aligned} \tag{118}$$

To make this independent on the given sample, we take the supremum over the neighbouring points involved, and plugging this back yields:

$$D_k(g^*)\leq\frac{1}{\gamma}E_XE_R\sup_{x'\in\mathcal{X}_d}|\|RX-Rx'\|-\|X-x'\||\tag{119}$$

$$\leq\frac{1}{\gamma}E_XE_R\sup_{x'\in\mathcal{X}_d}|\frac{\|RX-Rx'\|}{\|X-x'\|}-1|\cdot\operatorname{diam}(\mathcal{X}_d) \tag{120}$$

$$\leq\frac{2B\cdot w(T)}{\gamma\sqrt{k}} \tag{121}$$

where $T=\left\{\frac{x-x'}{\|x-x'\|}:x,x'\in\mathcal{X}_d\right\}$, and the last step used a result by Gordon (1985) (Lemma 21) (see also (Vershynin, 2018), sec. 7.5 and references therein).

Moreover, by applying Property 2.3 to (121), we also have

$$\mathcal{C}_{k,N}(\mathcal{G}_d)\leq\frac{2B\cdot w(T)}{\gamma\sqrt{k}}. \tag{122}$$

□

Corollary 22 (Nearest Neighbour) *Let $\mathcal{G}_d$ be the class of nearest neighbour classifiers of the form (36) with the $1/\gamma$-Lipschitz ramp-loss. Let $\mathcal{X}_d \subseteq \mathcal{B}(0,B), \mathcal{Y} = \{-1,1\}$, and $T \equiv \left\{ \frac{x-x'}{\|x-x'\|} : x, x' \in \mathcal{X}_d \right\}$. Take any $k \leq d, \gamma > 0, \delta \in (0,1)$.*

a) With probability $1-2\delta$,

$$E[\hat{g}_R] \leq E[g^*] + \frac{2B}{\gamma} \cdot \frac{w(T)}{\sqrt{k}} \cdot \mathbf{1}(k<d) + \bar{\ell} \cdot \xi(k, g^*, \delta) + \frac{56B}{\gamma}\left(1 + \frac{w(T)}{\sqrt{k}}\right) \cdot N^{-\frac{1}{k+1}} + 4\bar{\ell}\sqrt{\frac{\log(1/\delta)}{2N}} \tag{123}$$

where $w(\cdot)$ is the Gaussian width of the set in its argument, and g^* is the best 1-Lipschitz classifier.

b) With probability $1-\delta$, uniformly for all $g \in \mathcal{G}_d$ we have

$$E[g] \leq \hat{E}_{T_N}[g] + \frac{4B}{\gamma} \cdot \frac{w(T)}{\sqrt{k}} \cdot \mathbf{1}(k<d) + \frac{112B}{\gamma}\left(1 + \frac{w(T)}{\sqrt{k}}\right) \cdot N^{-\frac{1}{k+1}} + 3\bar{\ell}\sqrt{\frac{\log(2/\delta)}{2N}} \tag{124}$$

Proof Before we plug the expressions from Proposition 12 into Theorems 2-4, we need a bound on the complexity of the compressive class. We make use of the existing estimate for the class of Lipschitz functions with a fixed Lipschitz constant given by Gottlieb et al. (2016), which in our case takes the following form:

$$E_R[\hat{\mathcal{R}}_N(\mathcal{G}_R)] \leq E_R\left[\left[\frac{34(4\frac{1}{\gamma}\mathrm{diam}(R\mathcal{X}_d))^{k/2}}{\sqrt{N}}\left(\frac{k-1}{2}\right)\right]^{\frac{2}{k+1}}\right] \tag{125}$$

$$\leq 28\frac{1}{\gamma}E_R[\mathrm{diam}(R\mathcal{X}_d)]N^{-\frac{1}{k+1}} \tag{126}$$

$$\leq \frac{112B}{\gamma}\left(1 + \frac{2w(T)}{\sqrt{k}}\right) \cdot N^{-\frac{1}{k+1}} \tag{127}$$

Here we used that

$$E_R[\mathrm{diam}(R\mathcal{X}_d)] = E_R[\sup_{x,x' \in \mathcal{X}_d} \|R(x-x')\|] \leq E_R\left[\sup_{x,x' \in \mathcal{X}_d} \frac{\|R(x-x')\|}{\|x-x'\|}\right]\mathrm{diam}(\mathcal{X}_d) \tag{128}$$

$$\leq \left(1 + \frac{w(T)}{\sqrt{k}}\right)\mathrm{diam}(\mathcal{X}_d) \leq \left(1 + \frac{w(T)}{\sqrt{k}}\right)2B, \tag{129}$$

and noted that $34^{2/(k+1)}\left(\frac{k-1}{2}\right)^{2/(k+1)}$ has maximum at $k=2$ taking value ≤ 6.62, and $u^{k/(k+1)} \leq u$. Putting everything together completes the proof. □

General Lipschitz classifiers

Proof of Proposition 13 We will need the following lemma, proved later in this section.

Lemma 23 *Let $A \subset \mathbb{R}^d$ be a bounded set, $f : A \to \mathbb{R}$ a given L-Lipschitz function, and $R : \mathbb{R}^d \to \mathbb{R}^k$ a linear mapping. There exists an L-Lipschitz function $f_R : R(A) \to \mathbb{R}$, such that for all $x \in A$,*

$$|f_R(Rx) - f(x)| \leq L \cdot \sup_{x' \in A} |\|x - x'\| - \|Rx - Rx'\||.$$

We proceed to bound $D_k(g^*)$,

$$\begin{aligned} D_k(g^*) &= E_R[\inf_{g_R \in \mathcal{G}_R} E_{(X,Y)}[|g_R(RX, Y) - g^*(X, Y)|]] \\ &= E_R\left[\inf_{g_R \in \mathcal{G}_R} E_{(X,Y)}\left[\mathbf{1}(X \in \mathcal{X}_d^\epsilon) \cdot |g_R(RX, Y) - g^*(X, Y)| + \mathbf{1}(X \notin \mathcal{X}_d^\epsilon)|g_R(RX, Y) - g^*(X, Y)|\right]\right] \\ &\leq L_\ell E_R\left[\inf_{h_R \in \mathcal{G}_R} E_X[\mathbf{1}(X \in \mathcal{X}_d^\epsilon) \cdot |h_R(RX) - h^*(X)|]\right] + \epsilon \cdot \bar{\ell} \end{aligned}$$

where the last line used that for any $g_R \in \mathcal{G}_R$, $|g_R(RX, Y) - g^*(X, Y)| \leq \bar{\ell}$ by the boundedness of the loss function.

Now, using Lemma 23, we further upper bound the expression in (130) on the set $\mathcal{X}_d^\epsilon$ by choosing $h_R \in \mathcal{H}_R$ to be the L_h-Lipschitz function associated with h^* from Lemma 23. So for all $x \in \mathcal{X}_d^\epsilon$ we have $|h_R(Rx) - h^*(x)| \leq L_h \cdot \sup_{x' \in \mathcal{X}_d^\epsilon} |\|x - x'\| - \|Rx - Rx'\||$. Hence, bounding Eq. (130) gives:

$$D_k(g^*) \leq L_\ell L_h E_x E_R[\sup_{x' \in \mathcal{X}_d^\epsilon} |\|x - x'\| - \|Rx - Rx'\||] + \epsilon \cdot \bar{\ell} \tag{130}$$

$$\leq L_\ell L_h E_x E_R \sup_{x' \in \mathcal{X}_d^\epsilon} |\frac{\|Rx - Rx'\|}{\|x - x'\|} - 1| \cdot \mathrm{diam}(\mathcal{X}_d^\epsilon) + \epsilon \cdot \bar{\ell} \tag{131}$$

$$\leq L_\ell L_h \mathrm{diam}(\mathcal{X}_d^\epsilon) \frac{w(\mathcal{X}_d^\epsilon)}{\sqrt{k}} + \epsilon \cdot \bar{\ell} \tag{132}$$

where (132) follows from Gordon's lemma (Lemma 21).

Moreover, by using Property 2.3, this also gives us the same upper bound for the distortion-complexity,

$$\mathcal{C}_{k,N}(\mathcal{G}_d) \leq L_\ell L_h \mathrm{diam}(\mathcal{X}_d^\epsilon) \frac{w(\mathcal{X}_d^\epsilon)}{\sqrt{k}} + \epsilon \cdot \bar{\ell} \tag{133}$$

□

Proof We use the Rademacher complexity of the $L_\ell L_h$-Lipschitz function class, adapted to the relaxation of bounded domain.

$$E_R[\mathcal{R}_N(\mathcal{G}_R)] = \frac{1}{N} E_R E_{\mathcal{T}_N \sim \mathbb{P}^N} E_\sigma \left[\sup_{g_R \in \mathcal{G}_R} \sum_{n=1}^{N} \sigma_n \cdot \left(g_R(Rx_n)\mathbf{1}(x_n \in \mathcal{X}_d^\epsilon) + g_R(Rx_n)\mathbf{1}(x_n \notin \mathcal{X}_d^\epsilon) \right) \right] \tag{134}$$

$$\leq 28 L_\ell L_h \cdot E_R[\mathrm{diam}(R\mathcal{X}_d^\epsilon)] N^{-\frac{1}{k+1}} + \frac{1}{N} E_{\mathcal{T}_N} E_\sigma \left[\left| \sum_{n=1}^{N} \sigma_n \mathbf{1}(x_i \notin \mathcal{X}_d^\epsilon) \right| \right] \bar{\ell} \tag{135}$$

$$\leq 28 L_\ell L_h \cdot \mathrm{diam}(\mathcal{X}_d^\epsilon) \left(1 + \frac{2w(\mathcal{X}_d^\epsilon)}{\sqrt{k}} \right) \cdot N^{-\frac{1}{k+1}} + \frac{\epsilon \cdot \bar{\ell}}{\sqrt{N}} \tag{136}$$

The last step follows from bounding the expected diameter of the projected set $R\mathcal{X}_d^\epsilon$ in terms of the diameter of $\mathcal{X}_d^\epsilon$ in the first term, as before in Eqs. (128–129), and the Hölder and Jensen inequalities in the second term.

Finally, putting everything together with the expressions from Proposition 13 completes the proof. □

Corollary 24 (Lipschitz classifiers) *Let $\mathcal{G}_d$ be the class of L_h-Lipschitz classifiers with an L_ℓ-Lipschitz loss function. Let $T \equiv \left\{ \frac{x-x'}{\|x-x'\|} : x, x' \in \mathcal{X}_d \right\}$. Take any $k \leq d, \gamma > 0, \delta \in (0,1)$.*

a) With probability $1 - 2\delta$ the compressive Lipschitz classifier satisfies

$$E[\hat{g}_R] \leq E[g^*] + \left\{ L_\ell L_h \cdot \mathrm{diam}(\mathcal{X}_d^\epsilon) \cdot \frac{w(\mathcal{X}_d^\epsilon)}{\sqrt{k}} + \epsilon \cdot \bar{\ell} \right\} \cdot \mathbf{1}(k < d) + \bar{\ell} \cdot \xi(k, g^*, \delta)$$
$$+ 56 L_\ell L_h \cdot \mathrm{diam}(\mathcal{X}_d^\epsilon) \left(1 + \frac{2w(\mathcal{X}_d^\epsilon)}{\sqrt{k}} \right) \cdot N^{-\frac{1}{k+1}} + \frac{2\epsilon \cdot \bar{\ell}}{\sqrt{N}} + 4\bar{\ell} \sqrt{\frac{\log(3/\delta)}{2N}} \tag{137}$$

where $w(\cdot)$ is the Gaussian width of the set in its argument, and g^* is the best L_h-Lipschitz classifier.

b) W.p. $1 - \delta$, all $g \in \mathcal{G}_d$ satisfy

$$E[g] \leq \hat{E}[g] + 2 \left\{ L_\ell L_h \cdot \mathrm{diam}(\mathcal{X}_d^\epsilon) \cdot \frac{w(\mathcal{X}_d^\epsilon)}{\sqrt{k}} + \epsilon \cdot \bar{\ell} \right\} \cdot \mathbf{1}(k < d)$$
$$+ 56 L_\ell L_h \cdot \mathrm{diam}(\mathcal{X}_d^\epsilon) \left(1 + \frac{2w(\mathcal{X}_d^\epsilon)}{\sqrt{k}} \right) \cdot N^{-\frac{1}{k+1}} + \frac{2\epsilon \cdot \bar{\ell}}{\sqrt{N}} + 3\bar{\ell} \sqrt{\frac{\log(2/\delta)}{2N}} \tag{138}$$

Proof of Lemma 23 We define the following function, and show that it satisfies the required properties.

$$f_R : \mathbb{R}^k \to \mathbb{R}, \quad f_R(\tilde{x}) = \sup_{z \in A} \left\{ f(z) - L \cdot \sup_{z \in A} \|Rz - \tilde{x}\| \right\} \tag{139}$$

This function is L-Lipschitz: For all $\tilde{x}_1, \tilde{x}_2 \in \mathbb{R}^k$,

$$|f_R(\tilde{x}_1) - f_R(\tilde{x}_2)| = |\sup_{z \in A} \left\{ f(z) - L \cdot \|Rz - \tilde{x}_1\| \right\} - \sup_{z \in A} \left\{ f(z) - L \cdot \|Rz - \tilde{x}_2\| \right\}| \tag{140}$$

$$\leq L \cdot \sup_{z \in A} |\|Rz - \tilde{x}_2\| - \|Rz - \tilde{x}_1\||\tag{141}$$

$$\leq L \cdot \|\tilde{x}_2 - \tilde{x}_1\| \tag{142}$$

by the reverse triangle inequality.

Using the definition of f_R and the L-Lipschitz property of f, we have:

$$f_R(Rx) - f(x) = \sup_{z \in A} \{ f(z) - L \cdot \|Rz - Rx\| \} - f(x) \tag{143}$$

$$\leq L \sup_{z \in A} \{ \|z - x\| - \|Rz - Rx\| \}. \tag{144}$$

Furthermore, by choosing $z := x$ in the supremum,

$$f(x) - f_R(Rx) = f(x) - \sup_{z \in A} \{ f(z) - L \cdot \|Rz - Rx\| \} \tag{145}$$

$$\leq f(x) - \{ f(x) - L \cdot \|Rx - Rx\| \} \tag{146}$$

$$= 0 \tag{147}$$

$$\leq L \sup_{z \in A} |\|z - x\| - \|Rz - Rx\||\tag{148}$$

Hence, $|f(x) - f_R(Rx)| \leq L \sup_{z \in A} |\|z - x\| - \|Rz - Rx\||$. □

Appendix 4 Proof of lower bound, Theorem 6

Roadmap and tools

The proof uses techniques from (Tsybakov, 2004). The high level idea is to replace the infinite set of distributions $\mathbb{P}_{g^*}(k,\theta)$ or $\mathbb{P}_{\mathcal{G}}(k,\theta)$ with a finite family, which we need to construct to satisfy a balance between two antagonistic goals: Firstly, the distributions must be similar enough to make it difficult to determine which distribution generated a given i.i.d. sample of size N, and secondly, they must be different enough so that failure in doing so incurs a sufficiently high loss.

For the sake of intuition, suppose a finite support set of size q; then there are a total of 2^q possible binary classifiers, each of which can be identified with a binary string that encodes its outputs for on the points in the support. Equivalently, the set of all possible classifiers corresponds to the vertices of a q-dimensional hypercube. Our goal is to construct and associate a distribution to each $\sigma \in \Sigma$ from the set of distributions of interest i.e. from $\mathcal{P}_{g^*}(\theta,k)$ and from $\mathcal{P}_{\mathcal{G}_d}(\theta,k)$. As the two compressibility notions are related, the same construction involving θ-almost rank k distributions will work for both.

The following result from nonparametric statistics, known as the Assouad lemma (Tsybakov, 2004, Chapter 2, pp. 77–136), will guide our construction.

Lemma 25 (Assouad lemma) *Let $\Sigma = \{0,1\}^q$ be the set of binary strings of length q indexing a set $\{P_\sigma : \sigma \in \Sigma\}$ of 2^q probability measures on $\mathcal{Z}$. If $KL(P_\sigma||P_{\sigma'}) \le \zeta < \infty$ for all pairs $\sigma,\sigma' \in \Sigma$ with Hamming distance $H(\sigma,\sigma') = 1$, then*

$$\inf_{\hat{\sigma}} \sup_{\sigma\in\Sigma} E_{P_\sigma}[H(\hat{\sigma},\sigma)] \ge \frac{q}{2} \cdot \max\left\{\frac{1}{2}\exp(-\zeta), (1-\sqrt{\zeta/2})\right\} \tag{149}$$

where the infimum is with respect to all measurable functions $\hat{\sigma} : \mathcal{Z} \to \Sigma$, and $KL(\cdot||\cdot)$ is the Kullback–Leibler divergence between a pair of distributions.

Lemma 25 says that, if we can find a family of 2^q distributions such that the ones having neighbouring indexes on the hypercube are close in the KL sense, then for every estimator $\hat{\sigma}$ (which also corresponds to a vertex of the hypercube) there is another vertex σ whose associated distribution expects the Hamming distance of their hypercube-indexes to be large.

In the context of classification, P_σ will correspond to the distribution of the training set, and for any learning algorithm that returns a classifier from a sample set drawn from P_σ, $\hat{\sigma}$ will be an encoding the outputs of this classifier. We shall see that the excess error of this classifier relative to the best classifier, when the underlying distribution is P_σ, can be lower bounded in terms of the Hamming distance $H(\hat{\sigma},\sigma)$.

We start by specifying the family of distributions in a parameterised form. We will later determine appropriate values for the parameters to ensure both the KL condition of the Assouad lemma, and that all distributions are in the required compressible classes.

Construction of a parameterised set of distributions

Take an integer $q \le d$ and a parameter $\lambda \in [0,1]$, to be determined later. We define the following family of 2^q distributions indexed by binary strings of length q, supported on the

following finite set: $\{e_1, \dots e_q, 0_d\}$, where e_i is the i-th canonical basis vector. The q basis vectors will support a q-dimensional Euclidean space, and the setting of q, along with the parameter λ, and the inclusion of the origin 0_d into the support set will be used to handle the case when a relatively large probability mass lies outside of this subspace.

Our family of distributions will differ only in their class-conditional probability for the q basis vectors, while the marginals on $\mathcal{X}$ and the class conditional probability at 0_d are taken to be identical in all distributions.

With a slight abuse of notation, we will write $\mathbb{P}^{(\sigma)}(x)$ for $\mathbb{P}^{(\sigma)}(\{x\})$. We define the marginals as the following

$$\mathbb{P}^{(\sigma)}(0_d) := 1 - \lambda$$
$$\mathbb{P}^{(\sigma)}(e_i) := \lambda/q,\ i = 1, \dots q$$

and one can easily verify that $\sum_{i=1}^{q} \mathbb{P}^{(\sigma)}(e_i) + \mathbb{P}^{(\sigma)}(0_d) = 1$.

With appropriate choices of the parameters λ and q, a marginal distribution of this form is able to represent compressible distributions that belong to both $\mathcal{P}_{g^*}(\theta, k)$ and $\mathcal{P}_{\mathcal{G}_d}(\theta, k)$. For instance, if $q = d$ and $\lambda = \theta$, we have a θ-almost k-rank distribution; if $q = k, \lambda > 0$ then we have an exactly rank-k distribution.

The class-conditional probabilities at $e_i, i \in [q]$ are defined to fluctuate around 1/2.

$$\mathbb{P}^{(\sigma)}[Y = 1 | X = 0_d] := 1/2;$$
$$\mathbb{P}^{(\sigma)}[Y = 1 | X = e_i] := \frac{1 + \sigma_i \Delta}{2},\ i = 1, \dots q$$

where $\sigma = (\sigma_1, ..., \sigma_q) \in S_q \subset \{-1, +1\}^q$, and $\Delta \in (0, 1/2)$ is another parameter to be determined later in a way to ensure that the distributions $\mathbb{P}^{(\sigma)}$ indexed by neighouring strings are similar enough in the KL sense, as required in Assouad's lemma.

Observe that there is a bijection between the above family of distributions $\mathcal{P} \equiv \{(\mathbb{P}^{(\sigma)})^N\}_{\sigma \in S_q}$ and the set of binary strings Σ (or the hypercube vertices).

Setting the parameter Δ

Take two strings σ, σ' that only differ in one coordinate, $i' \in [q]$. We shall set the parameter Δ with the aim to have $KL(\mathbb{P}^{(\sigma)} || \mathbb{P}^{(\sigma')})$ below a threshold of 1/2—this will make the maximum on the r.h.s. of Assouad's lemma is $1 - \sqrt{1/4} = 1/2$. First, note that, since the sample is i.i.d., we have $KL((\mathbb{P}^{(\sigma)})^N || (\mathbb{P}^{(\sigma')})^N) = N \cdot KL(\mathbb{P}^{(\sigma)} || \mathbb{P}^{(\sigma')})$. We bound $KL(\mathbb{P}^{(\sigma)} || \mathbb{P}^{(\sigma')})$ using the χ^2 distance, and using the definition of the latter, as follows.

$$KL(\mathbb{P}^{(\sigma)} || \mathbb{P}^{(\sigma')}) \leq \chi^2(\mathbb{P}^{(\sigma)}, \mathbb{P}^{(\sigma')}) \tag{150}$$

$$= \sum_{i=1}^{q} \sum_{y \in \{-1,1\}} \frac{\left(\mathbb{P}^{(\sigma)}(e_i, y) - \mathbb{P}^{(\sigma')}(e_i, y)\right)^2}{\mathbb{P}^{(\sigma')}(e_i, y)} + \sum_{y \in \{-1,1\}} \frac{\left(\mathbb{P}^{(\sigma)}(0_d, y) - \mathbb{P}^{(\sigma')}(0_d, y)\right)^2}{\mathbb{P}^{(\sigma')}(0_d, y)} \tag{151}$$

The last term is zero, since the probability at $(0_d, y)$ was defined identically in both $\mathbb{P}^{(\sigma)}$ and $\mathbb{P}^{(\sigma')}$.

Writing $P(X, Y) = P(X)P(Y|X)$, we will condition on X. It is also useful to rewrite the label conditional probability can be written as the following

$$\begin{aligned}\mathbb{P}^{(\sigma)}(Y=y|X=e_i) &= \left[\mathbb{P}^{(\sigma)}(Y=1|X=e_i)\right]^{\mathbf{1}(y=1)}\left[\mathbb{P}^{(\sigma)}(Y=-1|X=e_i)\right]^{\mathbf{1}(y=-1)} \\ &= \left(\frac{1+\sigma_i\Delta}{2}\right)^{\mathbf{1}(y=1)}\left(\frac{1-\sigma_i\Delta}{2}\right)^{\mathbf{1}(y=-1)} \\ &= \frac{1+y\sigma_i\Delta}{2}\end{aligned}$$

Plugging this into (151), and taking into account that only the $i=i'$ term is nonzero in the sum (since σ and σ' only differ in their i'-th coordinate), we have

$$\text{Eq. (151)} = \frac{\lambda}{q}\sum_{y\in\{-1,1\}}\frac{\left(\mathbb{P}^{(\sigma)}(Y=y|X=e_{i'})-\mathbb{P}^{(\sigma')}(Y=y|X=e_{i'})\right)^2}{\mathbb{P}^{(\sigma')}(Y=y|X=e_{i'})} \tag{152}$$

$$= \frac{\lambda}{q}\sum_{y\in\{-1,1\}}\frac{\left(\frac{1+y\sigma_{i'}\Delta}{2}-\frac{1+y\sigma'_{i'}\Delta}{2}\right)^2}{\frac{1+y\sigma'_{i'}\Delta}{2}} \tag{153}$$

$$\leq \frac{\lambda}{q}\sum_{y\in\{-1,1\}}\frac{((\sigma'_{i'}-\sigma_{i'})\Delta y)^2}{1+\sigma'_{i'}\Delta y} \tag{154}$$

$$\leq \frac{4\lambda}{q}\sum_{y\in\{-1,1\}}\frac{\Delta^2}{\min\{1-\Delta,1+\Delta\}} \tag{155}$$

$$\leq \frac{4\Delta^2\lambda}{q}\frac{1}{1-\Delta} \tag{156}$$

$$\leq 8\lambda\Delta^2/q \tag{157}$$

In (155) we used that $\sigma'_{i'},\sigma_{i'}\in\{-1,1\}$, hence $(\sigma'_{i'}-\sigma_{i'})^2\leq 4$. The last inequality used the assumption that $\Delta\in(0,1/2)$.

In sum, for the product distribution we have $KL((\mathbb{P}^{(\sigma)})^N||(\mathbb{P}^{(\sigma')})^N)\leq 8\lambda\Delta^2N/q$. Now we set Δ by putting this quantity below 1/2, and also ensuring that $\Delta\in(0,1/2)$, as follows

$$\Delta := \frac{\min\{1,\sqrt{q/(\lambda N)}\}}{4}. \tag{158}$$

Before we can set the remaining parameters, q and λ, we need to link in the learning algorithm.

Defining $\hat{\sigma}$ and σ

We start by defining $\hat{\sigma}$ and σ in the context of a learning problem as follows. An arbitrary learning algorithm $\mathcal{A}$ receives an i.i.d. sample from $\mathbb{P}^{(\sigma)}$ and returns a classifier, which we map onto $\hat{\sigma}\in\Sigma$. Likewise, we map h^* to $\sigma\in\Sigma$. Using these definitions, we then lower bound

the excess error of the classifier learned by the algorithm in terms of the Hamming distance $H(\hat{\sigma}, \sigma)$.

Given any learning algorithm $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^N \to \mathcal{H}_d$ trained on a training set $\mathcal{T}_N \in (\mathcal{X} \times \mathcal{Y})^N$ drawn from $(\mathbb{P}^{(\sigma)})^N$, we let $\hat{w}_i := (\mathcal{A}(\mathcal{T}_N))(e_i), i = 1, \ldots, q$, and $\hat{w} = (\hat{w}_1, \ldots, \hat{w}_q) \in \mathbb{R}^q$. Furthermore, let $\hat{\sigma}_i = \text{sign}(\hat{w}_i), i = 1, \ldots, q$, and $\hat{\sigma} = (\hat{\sigma}_1, \ldots, \hat{\sigma}_q)$.

Likewise, we let $w^*_{\mathbb{P}^{(\sigma)}} = (w^*_1, \ldots w^*_q) \in \mathbb{R}^q$ with $w^*_i := h^*_{\mathbb{P}^{(\sigma)}}(e_i), i = 1, \ldots, q$ the outputs of the best classifier in the class, h^*, and $\sigma = (\sigma_1, \ldots, \sigma_q) \in \Sigma$ with $\sigma_i = \text{sign}((w^*_{\mathbb{P}^{(\sigma)}})_i), i = 1, \ldots, q$ It may be worth observing that, on the constructed family of distributions any learning algorithm is equivalent to a halfspace classifier, since the q canonical basis vectors are the only inputs where the function outputs can differ. For the same reason, h^* (equivalently w^*) is also a Bayes-optimal classifier under the distribution $\mathbb{P}^{(\sigma)}$. Hence, for any x in the support, we can write $(\mathcal{A}(\mathcal{T}_N))(x) = \hat{w}^T x$, and $h^*_{\mathbb{P}^{(\sigma)}}(x) = w^{*T}_{\mathbb{P}^{(\sigma)}} x$.

Lower bounding the excess risk by a Hamming distance

Our next goal is to lower bound the excess risk of the learned classifier in terms of a Hamming distance. In particular, the following holds, where ℓ is the 0–1 loss

$$E_{(X,Y)\sim\mathbb{P}^{(\sigma)}}[\ell((\mathcal{A}(\mathcal{T}_N))(X), Y)] - E_{(X,Y)\sim\mathbb{P}^{(\sigma)}}[\ell(h^*_{\mathbb{P}^{(\sigma)}}(X), Y)] \geq \frac{\lambda}{q}\Delta \cdot H(\hat{\sigma}, \sigma). \tag{159}$$

To see (159), we lower bound the l.h.s. using the law of iterated expectation

$$\sum_{i=1}^{q} \left(\mathbb{P}^{(\sigma)}_{Y|X}[Y \neq \text{sign}(\hat{w}^T X)|X = e_i] - \mathbb{P}^{(\sigma)}_{Y|X}[Y \neq \text{sign}(w^{*T}_{\mathbb{P}^{(\sigma)}} X)|X = e_i] \right) \frac{\lambda}{q}$$
$$+ \left(\mathbb{P}^{(\sigma)}_{Y|X}[Y \neq \text{sign}(\hat{w}^T X)|X = 0_d] - \mathbb{P}^{(\sigma)}_{Y|X}[Y \neq \text{sign}(w^{*T}_{\mathbb{P}^{(\sigma)}} X)|X = 0_d] \right)(1 - \lambda) \tag{160}$$

$$= \sum_{i=1}^{q} \left(\mathbb{P}^{(\sigma)}_{Y|X}[Y \neq \text{sign}(\hat{w}^T X)|X = e_i] - \mathbb{P}^{(\sigma)}_{Y|X}[Y \neq \text{sign}(w^{*T}_{\mathbb{P}^{(\sigma)}} X)|X = e_i] \right) \frac{\lambda}{q} \tag{161}$$

since the multiplier of $1 - \lambda$ in the last term of Eq. (160) evaluates to zero.

Consequently, by using the definitions of $\mathbb{P}^{(\sigma)}_{Y|X}$,

$$\text{Eq. (161)} \geq \sum_{i=1}^{q} \left(\mathbb{P}^{(\sigma)}_{Y|X}[Y \neq \hat{\sigma}_i] - \mathbb{P}^{(\sigma)}_{Y|X}[Y \neq \sigma_i] \right) \frac{\lambda}{q} \tag{162}$$

$$= \sum_{i=1}^{q} \sum_{y \in \{-1,1\}} \frac{\lambda}{q} \cdot \frac{1 + y\sigma_i \Delta}{2} \cdot \left(\mathbf{1}(\hat{\sigma}_i \neq y) - \mathbf{1}(\sigma_i \neq y) \right) \tag{163}$$

$$= \sum_{i=1}^{q} \sum_{y \in \{-1,1\}} \frac{\lambda}{q} \cdot \frac{1 + y\sigma_i \Delta}{2} \cdot \mathbf{1}(\hat{\sigma}_i \neq \sigma_i) \tag{164}$$

If $\sigma_i = y$, then $\frac{1+y\sigma_i\Delta}{2} = \frac{1+\Delta}{2}$; if $\sigma_i \neq y$, then $\frac{1+y\sigma_i\Delta}{2} = \frac{1-\Delta}{2}$. Consequently, (163) equals

$$\sum_{i=1}^{q} \frac{\lambda}{q}\Big[\frac{1+\Delta}{2}\mathbf{1}(\sigma_i \neq \hat{\sigma}_i) - \frac{1-\Delta}{2}\mathbf{1}(\sigma_i \neq \hat{\sigma}_i)\Big] = \sum_{i=1}^{q} \frac{\lambda}{q}\mathbf{1}(\sigma_i \neq \hat{\sigma}_i)\Delta = \frac{\lambda}{q}\Delta \cdot H(\hat{\sigma}, \sigma) \tag{165}$$

which concludes the statement of Eq. (159).

Applying Assouad's lemma

Having constructed the family of distributions in a way that neighbouring ones on the hypercube are similar in the KL sense, we now want to show that a classifier trained on a sample drawn from one of these distributions will have high expected error for some setting of the remaining distributional parameters.

To recall the setting, suppose that one of the members of our family of distributions, $\mathbb{P}^{(\sigma)}, \sigma \in \Sigma$ is the true underlying distribution from which we have a sample $\mathcal{T}_N \sim (\mathbb{P}^{(\sigma)})^N$. An arbitrary learning algorithm trained on $\mathcal{T}_N$ returns the classifier $\mathcal{A}(\mathcal{T}_N)$. Using Assouad's lemma, we want to show that, we can set q and λ such that $\mathcal{A}(\mathcal{T}_N)$ has high expected risk of failing to identify the correct distribution—in other words its expected excess error will be higher than some lower bound.

We use the encoding of the classifier $\mathcal{A}(\mathcal{T}_N)$ into $\hat{\sigma} = \hat{\sigma}(\mathcal{T}_N)$ described earlier—this is an estimator of σ – and use the lower bound on its excess error from (159),

$$\begin{aligned} E[g_{\mathcal{A}(\mathcal{T}_N)}]] - E[g^*_{\mathbb{P}}] &= E_{(X,Y)\sim\mathbb{P}^{(\sigma)}}[\ell((\mathcal{A}(\mathcal{T}_N))(X), Y)] - E_{(X,Y)\sim\mathbb{P}^{(\sigma)}}[\ell(h^*_{\mathbb{P}^{(\sigma)}}(X), Y)] \\ &\geq \frac{\lambda}{q}\Delta \cdot H(\hat{\sigma}(\mathcal{T}_N), \sigma), \end{aligned} \tag{166}$$

where we made explicit the dependence of $\hat{\sigma}$ on $\mathcal{T}_N$.

Taking expectation w.r.t. the distribution of $\mathcal{T}_N$ on both sides, we now apply the Assouad lemma (Lemma 25) with the distribution family $\{(\mathbb{P}^{(\sigma)})^N\}_{\sigma\in\Sigma}$ on $\mathcal{Z} = (\mathcal{X}\times\mathcal{Y})^N$. Hence, the expectation of (166) is lower bounded as

$$E_{\mathcal{T}_N\sim(\mathbb{P}^{(\sigma)})^N}\left[E[g_{\mathcal{A}(\mathcal{T}_N)}]] - E[g^*_{\mathbb{P}}]\right] \geq \frac{\lambda}{q}\Delta \cdot E_{\mathcal{T}_N\sim(\mathbb{P}^{(\sigma)})^N}[H(\hat{\sigma}(\mathcal{T}_N), \sigma)] \tag{167}$$

$$\geq \frac{\lambda}{q} \cdot \frac{\min\{1, \sqrt{q/(\lambda N)}\}}{4} \cdot \frac{q}{4} \tag{168}$$

$$= \frac{\lambda}{16}\min\left\{1, \sqrt{\frac{q}{\lambda N}}\right\} \tag{169}$$

The lower bound (169) still depends on the distributional parameters q, λ. It now remains to set these so as to ensure that $\mathbb{P}^{(\sigma)}$ is both D-compressible and C-compressible.

Final construction of a bad distribution

We are finally ready to set the parameters q and λ in the family of distributions constructed early in the proof; these will be set with the aim to construct the required bad distribution.

We apply the findings of the previous section, Eq. (169). There are 2 cases to consider: small θ and large θ.

1. Case $\theta \geq \sqrt{\frac{k}{N}}$. In this case we choose $q = d$, $\lambda = \theta$, and the marginal on $\mathcal{X}$ becomes

$$\mathbb{P}^{(\sigma)}(0_d) = 1 - \theta; \quad \mathbb{P}^{(\sigma)}(e_i) = \theta/d, \ i = 1, \dots d. \tag{170}$$

Observe, this is a θ-almost k-rank distribution, cf our Definition 3, with the underlying linear subspace V_k – indeed, $0_d \in V_k$ and $\mathbb{P}^{(\sigma)}(0_d) = 1 - \theta$, so we have $\mathbb{P}^{(\sigma)}(V_k) = 1 - \theta + k\theta/d > 1 - \theta$. Hence, this distribution is both D-compressible and C-compressible, with the same parameters (θ, k). Plugging these parameter choices back into (169), we have

$$\begin{aligned} E_{(X,Y)\sim\mathbb{P}^{(\sigma)}}[\ell((\mathcal{A}(\mathcal{T}_N))(X), Y)] - E_{(X,Y)\sim\mathbb{P}^{(\sigma)}}[\ell(h^*_{\mathbb{P}^{(\sigma)}}(X), Y)] \\ \geq \frac{\theta}{16} \min\left\{1, \sqrt{\frac{d}{\theta N}}\right\} \\ \geq \frac{\theta}{16} = \frac{1}{32} 2\theta \end{aligned} \tag{171}$$

$$\geq \frac{1}{32}\left(\theta + \sqrt{\frac{k}{N}}\right). \tag{172}$$

The inequality (171) holds because $N < d$ and $\theta \in [0, 1]$ so the minimum is 1; the inequality (172) follows from $\theta \geq \sqrt{k/N}$.

2. Case $\theta < \sqrt{\frac{k}{N}}$. Now we choose $q = k$, $\lambda = 1$, so the marginal becomes

$$\mathbb{P}^{(\sigma)}(0_d) = 0; \quad \mathbb{P}^{(\sigma)}(e_i) = 1/k, \ i = 1, \dots k. \tag{173}$$

This is again a θ-almost k-rank distribution (with $\theta = 0$—exactly k-rank in fact), therefore it belongs to both D-compressible and C-compressible distributions with the same parameters (θ, k). By Eq. (169), in this case we have:

$$\begin{aligned} E_{(X,Y)\sim\mathbb{P}^{(\sigma)}}[\ell((\mathcal{A}(\mathcal{T}_N))(X), Y)] - E_{(X,Y)\sim\mathbb{P}^{(\sigma)}}[\ell(h^*_{\mathbb{P}^{(\sigma)}}(X), Y)] \\ \geq \frac{1}{16} \min\left\{1, \sqrt{\frac{k}{N}}\right\} \end{aligned} \tag{174}$$

$$\begin{aligned} &\geq \frac{1}{16}\sqrt{\frac{k}{N}} = \frac{1}{32} 2\sqrt{\frac{k}{N}} \\ &\geq \frac{1}{32}\left(\theta + \sqrt{\frac{k}{N}}\right). \end{aligned} \tag{175}$$

The inequality (174) holds because $k < N$ so the minimum is $\sqrt{k/N}$, and inequality (175) follows from $\theta \leq \sqrt{k/N}$.

Therefore, in both cases we found a distribution for which the excess risk of $\mathcal{A}(\mathcal{T}_N)$ is greater than $c(\theta + \sqrt{k/N})$, where $c = \frac{1}{32}$.

Appendix 5 Standard inequalities

For reference, here we list the classic inequalities that we made use of; these can be found in textbooks such as (Shalev-Shwartz & Ben-David, 2014; Mohri et al., 2012).

Property 5.1 (Johnson-Lindenstrauss) *Let $\{x_1, x_2, \ldots, x_N\} \subset \mathbb{R}^d$ be a set of N points, and a random matrix $R \in \mathbb{R}^{k \times d}$. For any $\epsilon \in (0,1)$, $\delta \in (0,1)$, if $k \geq C\epsilon^{-2}\log(N/\delta)$, we have*

$$\forall i, j \in [N], (1-\epsilon)\|x_i - x_j\|^2 \leq \|Rx_i - Rx_j\|^2 \leq (1+\epsilon)\|x_i - x_j\|^2 \tag{176}$$

with probability at least $1-\delta$, where $C > 0$ is a constant.

Lemma 26 (Markov inequality) *Let X be a non-negative random variable. Then, for any $\delta > 0$, with probability at least $1-\delta$, we have*

$$\mathbb{P}(X \geq \epsilon) \leq \frac{E[X]}{\delta}. \tag{177}$$

Lemma 27 (Hoeffding inequality) *Let $X_1, X_2, \ldots, X_n$ be independent random variables such that $X_i \in [a,b]$ a.s. for all $i \in [n]$. Then, for any $\epsilon, \delta > 0$, w.p. at least $1-\delta$, we have*

$$\left| \frac{1}{n}\sum_{i=1}^{n} X_i - E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \right| \leq \sqrt{\frac{(b-a)^2 \log(2/\delta)}{2n}}. \tag{178}$$

Lemma 28 (McDiarmid inequality) *Let $\mathcal{X}$ be a set, and $f : \mathcal{X}^N \to \mathbb{R}$ be a function s.t. for some $c > 0$, for all $i \in [N]$ and for all $x_1, \ldots x_N, x_i' \in \mathcal{X}$ we have*

$$|f(x_1, \ldots, x_N) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_N)| \leq c. \tag{179}$$

Let $X_1, \ldots, \mathcal{X}_N$ be N independent random variables taking values in $\mathcal{X}$. Then, w.p. at least $1-\delta$ we have

$$|f(X_1, \ldots, X_N) - E[f(X_1, \ldots, X_N)]| \leq c\sqrt{N \log(2/\delta)/2}. \tag{180}$$

The following classic generalisation bound is derived using McDiarmid inequality.

Theorem 29 (Rademacher bounds (Shalev-Shwartz & Ben-David, 2014) *Lemma* 3.3.) *Let $\mathcal{G}$ be the loss class of a function class, and suppose the loss is bounded by $\bar{\ell}$. With probability at least $1-\delta$ we have each of the following uniformly for all $g \in \mathcal{G}$:*

$$E[g] \leq \hat{E}[g] + 2\mathcal{R}_N(\mathcal{G}) + \bar{\ell}\sqrt{\frac{\log(1/\delta)}{2N}}, \quad \text{and}$$
$$E[g] \leq \hat{E}[g] + 2\hat{\mathcal{R}}_N(\mathcal{G}) + 3\bar{\ell}\sqrt{\frac{\ln(2/\delta)}{2N}}.$$

Acknowledgements The authors are grateful for the generous support of EPSRC, though the Fellowship grant EP/P004245/1, "Fortuitous Geometries and Compressive Learning". This work was undertaken when HR was with the University of Birmingham.

Author contributions conception and design: AK, HR; supervision: AK; writing and editing: AK, HR.

Funding This work was funded by EPSRC Fellowship EP/P004245/1.

Declarations

Conflict of interest The authors declare that they have no conflicts of interest or competing interests relating to the content of this article.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

Alon, N., Ben-David, S., Cesa-Bianchi, N., & Haussler, D. (1997). Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM, 4*, 615–631.

Arriaga, R. I., & Vempala, S. (1999). An algorithmic theory of learning: Robust concepts and random projection. In *40th Annual Symposium on Foundations of Computer Science (FOCS)* (pp. 616–623).

Bartl, D., & Mendelson, S. (2022). Random embeddings with an almost Gaussian distortion. *Advances in Mathematics*, *400*, 108261.

Bartlett, P. L., & Mendelson, S. (2002). Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research, 3*, 463–482.

Crammer, K., Gilad-Bachrach, R., Navot, A., & Tishby, N. (2002). Margin analysis of the LVQ algorithm. In *Neural information processing systems (NIPS)*.

Dudley, R. M. (1999). *Uniform central limit theorems*. Cambridge, MA: Cambridge University Press.

Durrant, R. J., & Kabán, A. (2013). Sharp generalization error bounds for randomly-projected classifiers. In *Proceedings of 30-th international conference on machine learning (ICML)*. Journal of Machine Learning Research W &CP 28(3) (pp. 693–701).

Gordon, Y. (1985). Some inequalities for Gaussian processes and applications. *Israel Journal of Mathematics, 50*(4), 265–289.

Gottlieb, L. A., & Kontorovich, A. (2014). Efficient classification for metric data. *IEEE Transactions on Information Theory, 60*(9), 5750–5759.

Gottlieb, L. A., Kontorovich, A., & Krauthgamer, R. (2016). Adaptive metric dimensionality reduction. *Theoretical Computer Science, 620*(21), 105–118.

Guermeur, Y. (2017). LP-norm Sauer–Shelah lemma for margin multi-category classifiers. *Journal of Computer and System Sciences, 89*, 450–473.

Gurvits, L., & Koiran, P. (1995). Approximation and learning of convex superpositions. In *Computational learning theory (EUROCOLT)*.
Halko, N., Martinsson, P. G., & Tropp, J. A. (2011). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review, 53*(2), 217–288.
Indyk, P. (2007). Nearest-neighbor-preserving embeddings. *ACM Transactions on Algorithms, 3*, 3.
Kabán A (2014) New bounds on compressed linear least squares regression. In *International conference on artificial intelligence and statistics (AISTATS)*, JMLR W &P (vol. 33, pp. 448–456).
Kabán A (2019) Dimension-free error bounds from random projections. In *The thirty-third AAAI conference on artificial intelligence*. AAAI Press (pp. 4049–4056).
Kabán, A. (2013). A new look at compressed ordinary least squares. In Ding, W., Washio, T., Xiong, H., et al. (eds.) *13th IEEE international conference on data mining workshops, ICDM workshops*, TX, USA, December 7–10, 2013. IEEE Computer Society (pp 482–488).
Kaban, A. (2015). Improved bounds on the dot product under random projection and random sign projection. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*
Kabán, A., & Durrant, R. J. (2020). Structure from randomness in halfspace learning with the zero-one loss. *Journal of Artificial Intelligence Research, 69*, 733–764.
Kearns, M., & Vazirani, U. (1994). *An introduction to computational learning theory*. The MIT Press.
Kontorovich, A. & Weiss, R. (2015). A Bayes consistent 1-NN classifier. In *AISTATS*.
Laine, S. & Aila, T. (2017). Temporal ensembling for semi-supervised learning. In *ICLR*.
Latorre, F., Dadi, L. T., Rolland, P., & Cevher, V. (2021). The effect of the intrinsic dimension on the generalization of quadratic classifiers. *Advances in Neural Information Processing Systems, 34*, 21138–21149.
Lauer, F. (2019). Optimization and statistical learning theory for piecewise smooth and switching regression. Habilitation à diriger des recherches, Université de Lorraine. https://hal.univ-lorraine.fr/tel-02307957
Liaw, C., Mehrabian, A., Plan, Y., & Vershynin, R. (2017). A simple tool for bounding the deviation of random matrices on geometric sets. In *Geometric aspects of functional analysis* (pp. 277–299).
Matoušek, J. (2008). On variants of the Johnson–Lindenstrauss lemma. *Random Structures & Algorithms, 33*(2), 142–156.
Mendelson, S. (2003). A few notes on statistical learning theory. Lecture notes in computer scienceIn S. Mendelson & A. J. Smola (Eds.), *Advanced lectures in machine learning* (Vol. 2600, pp. 1–40). Berlin: Springer-Verlag.
Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). *Foundations of machine learning*. MIT Press.
Munteanu, A., Omlor, S., & Song, Z., Woodruff, D. (2022). Bounding the width of neural networks via coupled initialization A worst case analysis. In *International conference on machine learning (ICML)*, (pp. 16083–16122).
Papadimitriou, C. H. & Vempala, S. S. (2019). Random projection in the brain and computation with assemblies of neurons. In *Information technology convergence and services*.
Reeve, H. W. J. & Kabán, A. (2021). Statistical optimality conditions for compressive ensembles. CoRR abs/2106.01092. arXiv:2106.01092
Rosasco, L., Vito, E. D., Caponnetto, A., et al. (2004). Are loss functions all the same? *Neural Computation, 16*(5), 1063–1076.
Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge University Press.
Slawski, M. (2018). On principal components regression, random projections, and column subsampling. *Electronic Journal of Statistics, 12*(2), 3673–3712.
Tsybakov, A. B. (2004). *Introduction to nonparametric estimation.* Mathématiques & applications (Paris) (Vol. 41). Springer.
Turner, A.J. & Kabán, A. (2023). Pac learning with approximate predictors. Machine Learning
Vapnik, V. N. (1998). *Statistical learning theory*. Wiley-Interscience.
Verma, N., & Branson, K. (2015). Sample complexity of learning Mahalanobis distance metrics. *Advances in Neural Information Processing Systems (NIPS), 28*, 2584–2592.
Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*. Cambridge University Press.
von Luxburg, U., & Bousquet, O. (2004). Distance-based classification with Lipschitz functions. *Journal of Machine Learning Research, 5*, 669–695.
Wolf, M. M. (2020). Mathematical foundations of supervised learning. Retrieved July 22, 2022.