PAC learning with approximate predictors

Turner, Andrew; Kaban, Ata

DOI:
10.1007/s10994-023-06301-4

Machine Learning
https://doi.org/10.1007/s10994-023-06301-4

PAC-learning with approximate predictors

Andrew J. Turner[1] · **Ata Kabán**[1]

Received: 2 March 2022 / Revised: 2 January 2023 / Accepted: 11 January 2023

Abstract
Approximate learning machines have become popular in the era of small devices, including quantised, factorised, hashed, or otherwise compressed predictors, and the quest to explain and guarantee good generalisation abilities for such methods has just begun. In this paper, we study the role of approximability in learning, both in the full precision and the approximated settings. We do this through a notion of sensitivity of predictors to the action of the approximation operator at hand. We prove upper bounds on the generalisation of such predictors, yielding the following main findings, for any PAC-learnable class and any given approximation operator: (1) We show that under mild conditions, approximable target concepts are learnable from a smaller labelled sample, provided sufficient unlabelled data; (2) We give algorithms that guarantee a good predictor whose approximation also enjoys the same generalisation guarantees; (3) We highlight natural examples of structure in the class of sensitivities, which reduce, and possibly even eliminate the otherwise abundant requirement of additional unlabelled data, and henceforth shed new light onto what makes one problem instance easier to learn than another. These results embed the scope of modern model-compression approaches into the general goal of statistical learning theory, which in return suggests appropriate algorithms through minimising uniform bounds.

Keywords Statistical learning · Generalisation error bounds · Model-compression · Approximate learning algorithms

1 Introduction

The last decade has seen a tremendous increase of interest in complex learning problems, such as deep neural networks, and learning in very high dimensional spaces. This results in a large number of parameters which need to be learned from the data. This is typically very resource-intensive in terms of memory, computation, and labelled training data; and

Editors: João Gama, Alípio Jorge, Salvador García.

✉ Ata Kabán
A.Kaban@bham.ac.uk

Andrew J. Turner
A.J.Turner@bham.ac.uk

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

consequently infeasible to deploy on devices with limited resources such as mobile phones, wearable devices, and the Internet of Things. Therefore, a plethora of model-compression and approximation techniques have been proposed, such as quantisation, pruning, factorisation, random projection, hashing, and others (Choudhary et al., 2020). Rather intriguingly, many empirical findings on realistic benchmark problems seem to indicate that, despite a drastic compression of the complex model, such techniques often perform impressively well, with predictive accuracy comparable to that of full precision models. Below we mention just a few illustrative landmarks.

Quantisation of the weights of deep neural networks was proposed in BinaryConnect (Courbariaux et al., 2015), where a neural network with weights constrained to a single bit (± 1) was proposed and empirically demonstrated to achieve comparable results to a full precision network of the same size. These results were further refined by the Quantised Neural Networks (QNN) training algorithm (Hubara et al., 2017), and the idea was also extended to convolutional networks in Xnor-net (Rastegari et al., 2016). Another compression scheme introduced in Han et al. (2016), called Deep Compression, employed a combination of pruning, quantisation, and Huffman coding to achieve similar results to the original network, with a significant reduction in memory usage.

Factorisation of the weights into low-rank matrices has been another common technique to reduce the size of a deep neural network (DNN), see Denil et al. (2013, 2014) for details. Recent survey articles on a variety of model-compression techniques specific to deep neural networks may be found in Choudhary et al. (2020), Cheng et al. (2017) and Menghani (2021).

In a related work (Ravi, 2019), the author proposes to learn the high and low complexity networks simultaneously through a joint objective function that minimises not only their individual sample errors but also their disagreement. They found experimentally that this approach improves accuracy of both models, *regardless* of the model-compression technique employed. While a theoretical explanation remains elusive, this was among the first attempts to shift focus from the compressed model back to the fuller picture of the original model, and consider these objectives in tandem.

Theoretical studies of model-compression are much scarcer, and the interplay between model approximation and generalisation is not very well understood. Work taking an information theoretic approach (Gao et al., 2019) studied the trade-off between the compression granularity (rate) and the change it induces in the empirical error, using rate distortion theory. Follow-on work (Bu et al., 2021) extended their analysis to show that it is possible (on occasion) for compressed versions of pre-trained models to generalise even better than the original.

Another line of research exploited a notion of compression (Arora et al., 2018; Zhou et al., 2019). In Arora et al. (2018), a new compression framework was introduced for proving generalisation bounds, and their analysis indicated that resilience to noise implies a better generalisation for deep neural networks. A PAC-Bayes bound was then proposed to give a non-vacuous generalisation bound on the compressed network in Zhou et al. (2019). This was further built upon in Baykal et al. (2019), and has inspired a new algorithm along with a generalisation bound for the fully connected network.

In Suzuki et al. (2020b), compression-based bounds on a new pruning method for DNN was established, and more recently the authors also gave bounds for the full network (Suzuki et al., 2020a). This latter work allows the compression-based bound to be converted into a bound for the full network, using the local Rademacher complexity of the Minkowski difference between the loss class of the full networks and the loss class of the compressed networks. This is therefore another instance, entirely complementary of the

work of Ravi (2019), where the performance of the approximate model is linked back in some way to that of the full model, albeit a joint treatment has not been attempted.

In Ashbrock and Powell (2021), a stochastic Markov gradient decent was introduced to learn in memory limited setting directly in the discrete parameter space. They provide convergence analysis for their optimisation algorithm, but generalisation is only demonstrated experimentally.

The general trend and focus on compressing deep neural networks (DNNs) is remarkable. However, we conjecture a more fundamental connection between approximability and generalisation that is not specific to deep networks. Contrary to the increasingly sophisticated and specialised tools being developed for DNNs, our aim here is to study the connection between approximability and generalisation from first principles. To do this, we want to ensure generalisation guarantees for learning with approximate models in general.

We also hypothesise that target concepts that have low sensitivity to approximation may represent a benign trait of learning problems in general, which would imply easier learnability of the full precision model too. To substantiate this, we shall seek learning algorithms whose generalisation ability depends on the approximability of the target concept, irrespective of the form of the learned predictor being used in the full or approximated setting.

1.1 Contributions

In the following roadmap we summarise the main contributions and findings of this paper:

- We define a notion of approximability of a predictor, which quantifies the average extent of sensitivity of its predictions when subjected to a given approximation operator (Sect. 2.1). This quantity will feature heavily in our generalisation bounds.
- In Sect. 2.2 we show that low sensitivity target functions may require less labelled training data, provided we have access to an independent unlabelled set of sufficient size (Theorem 1). This sets the stage for approximability to be viewed as a benign trait for learning.
- In Sect. 2.3 we develop a practical theory, showing that a constrained empirical risk minimisation algorithm with a modified loss function, which enforces approximability up to a given threshold, learns a predictor that is guaranteed to generalise well both in its full precision and its approximate forms (Proposition 2.4). Furthermore, we construct an objective function that also implicitly optimises the trade-off managed by the sensitivity threshold (Proposition 2.5). These results then give rise to a learning algorithm that is able to take advantage of additional unlabelled data without the requirement for it to be independent from the labelled set (Theorem 2).
- For learning a good approximate predictor, we also give two variants of our algorithm that allows the user to control the above trade-off directly (Theorem 3, Remark 2.6). This may be useful in certain settings, for example when low memory requirements prevail over prediction accuracy.
- Section 3 is devoted to studying our unlabelled data requirements. We show that, while the worst case unlabelled sample size requirement is necessarily large (Proposition 3.1), natural examples of structure may arise from the data source interacting with the model, which may reduce, or may even eliminate the requirement for additional unlabeled sample (Propositions 3.3, 3.4). This analysis is independent of the hypothesis class employed, and leads to some general conditions under which sensitivity estima-

tion enjoys favourable convergence (Theorem 4). In addition, we also point out that, the structural restrictions of the hypothesis class in itself can bring further insights – in particular, for generalised linear models, the weight sensitivity turns out to be sufficient for dimension-independent learning (Proposition 3.5).
- We discuss implications of our theoretical results related to real problems, including binarisation with depth-indepedent error bounds and on-device deep network classification in Sect. 4.

Throughout the exposition of the main sections, we only consider deterministic approximation operators, keeping the reasoning and the formalism simple, and rooted in first principles. We discuss extensions in Sect. 4, including the use of stochastic approximation operators.

1.2 Related work

We have already highlighted two existing studies that considered both sides of model-compression, namely the approximate predictor as well as the full predictor. Below we further discuss these in the light of our aims, approach, and findings, along with existing works that relate to ours in terms of either high-level ideas or technical aspects.

In a similar spirit to Ravi (2019), our inquiry concerns simultaneously both the approximate model and the full precision model. However, contrary to the empirical approach taken in Ravi (2019), where the heuristic nature of the algorithms make a theoretical understanding somewhat elusive, our approach is analytic. We employ Rademacher complexity analysis of the generalisation error (Bartlett & Mendelson, 2002) to give algorithm-independent uniform bounds on the generalisation for both approximate and approximable function classes. The uniform nature of these bounds justifies algorithms that minimise them. Therefore, our algorithms come with guarantees of good generalisation. Our framework is general, and can be used to analyse the approximability and generalisation in tandem for any PAC-learnable machine learning problem.

Our findings are consistent with those found in Suzuki et al. (2020a), with a difference in the approach, resulting in a different and more general angle. In particular, their focus is on translating already known bounds on compressed neural networks to the full uncompressed class. In contrast, we focus on showing that having good approximability (i.e. low sensitivity to approximation) improves generalisation bounds in PAC-learnable classes. In addition, we pursue a joint treatment of learning both the approximate and the full predictor simultaneously.

The works in Arora et al. (2018) and Zhou et al. (2019), based on the idea of compression and resilience to noise, are also somewhat related to our work, on a high-level. However, in both Arora et al. (2018) and Zhou et al. (2019) the generalisation bounds are for the compressed model only; whereas, our treatment provides both sides of the coin—algorithms that learn a predictor that generalises both in its full precision and its approximate form. In Arora et al. (2018), the focus is on bounding the classification error of the compressed predictor with the γ-margin loss (with $\gamma > 0$) of the full model for multi-class classification. This corresponds to our general bounded Lipschitz loss function. Moreover, in Zhou et al. (2019) a PAC-Bayes approach is taken and so numerical tightness comes from data-dependent quantities in the bound that do not necessarily identify or shed light on structural traits of the problem responsible for good generalisation. In contrast, by employing Rademacher analysis we are able to highlight structural properties responsible for low

complexity and good generalisation, so our approach and findings are complementary to these works.

Our starting point in Sect. 2.2 is the semi-supervised framework of Bălcan and Blum (2010), where our approximability, or sensitivity of functions to approximation plays the role of an unlabeled error, and we replace VC entropy with Rademacher complexity to facilitate the use of our bounds outside the classification setting. However, from Sect. 2.3 onward we depart from this framework in favour of simpler and more straightforwardly implementable bounds that fit our specific goals at the expense of a negligible additive term. In return, we obtain some advantages: (1) for our purposes, the unlabelled data need not be independent from the labelled set, (2) the sensitivity threshold is optimised implicitly and automatically by our algorithm without appeal to structural risk minimisation, and (3) we are able to study structural regularities that reduce or even eliminate the need of unlabelled data, which was not attempted in the previous work.

2 Generalisation through approximability

2.1 Notations and preliminaries

Consider the input domain $\mathcal{X} \subseteq \mathbb{R}^d$, where d denotes the dimensionality of the feature representation, and output domain $\mathcal{Y} \subseteq \mathbb{R}$. Let $m \in \mathbb{N}$ and consider a sample $S \in (\mathcal{X} \times \mathcal{Y})^m$ of size m drawn i.i.d. from an unknown distribution D. Let $\mathcal{H}$ be the hypothesis class; this is a set of functions mapping from $\mathcal{X}$ to $\mathcal{Y}$. We consider a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$. Then we define the generalisation and empirical error of a function $f \in \mathcal{H}$ as

$$\operatorname{err}(f) := \mathbb{E}_{(x,y)\sim D}[\ell(f(x), y)], \quad \text{and} \quad \widehat{\operatorname{err}}(f) := \frac{1}{m} \sum_{(x,y)\in S} \ell(f(x), y).$$

The best function in the class will be denoted as $f^* := \operatorname{argmin}_{f\in\mathcal{H}}\{\operatorname{err}(f)\}$.

We let $\mathcal{H}_A$ be the set of approximate functions from $\mathcal{X}$ to $\mathcal{Y}$. Note $\mathcal{H}_A$ needs not be a subset of $\mathcal{H}$. We define an approximation operator $A : \mathcal{H} \to \mathcal{H}_A$, which maps a hypothesis to its approximation. Here A is considered to be deterministic; extension to stochastic approximate algorithms is discussed later in Sect. 4.

Definition 2.1 (*Approximation-sensitivity of a function*) Fix $p \geq 1$. Given a sample $S \in \mathcal{X}^m$ of size m drawn i.i.d. from the marginal distribution D_x, we define the true and empirical sensitivity as

$$\mathcal{D}_A^p(f) := \mathbb{E}_{x\sim D_x}[|f(x) - Af(x)|^p]^{\frac{1}{p}}, \quad \text{and} \quad \widehat{\mathcal{D}}_A^p(f) := \left(\frac{1}{m}\sum_{x\in S} |f(x) - Af(x)|^p\right)^{\frac{1}{p}}.$$

The choice of p-norm will be left to the user in our forthcoming bounds. Formally, it is sufficient to work with $p = 1$, as by Jensen's inequality, for all $p \geq 1$, we have $\mathcal{D}_A^1(f) \leq \mathcal{D}_A^p(f)$ and $\widehat{\mathcal{D}}_A^1(f) \leq \widehat{\mathcal{D}}_A^p(f)$, for all $f \in \mathcal{H}$. So the forthcoming bounds will be tightest with the choice $p = 1$. However, sometimes the user might like to specify a constraint on the sensitivity of functions in terms of the more familiar Euclidean norm ($p = 2$), or some other member of the family of p-norms. Our results apply to any specification of p, so we will state results for general p-norms. An example where $p = 2$ is advantageous will

be encountered later in Theorem 4. When the choice of p is arbitrary, we may omit the upper index in our notation.

The approximating class $\mathcal{H}_A$ is typically chosen to be much smaller than the original class $\mathcal{H}$, implying a reduced complexity term in our generalisation bounds, at the expense of a larger empirical error, and the appearance of an additional sensitivity term $\mathcal{D}_A(f)$. We can think of $\mathcal{H}_A$ as a compressed model class whose elements occupy less memory, yet still expressive enough to represent the essence of $\mathcal{H}$. Examples include quantisation and other model-compression schemes. The granularity of approximation that we can afford is considered to be fixed. In memory-constrained settings this is constrained by the available hardware.

We now define sensitivity-restricted hypothesis classes

$$\mathcal{H}_t := \{f \in \mathcal{H} : \mathcal{D}_A(f) \leq t\} \quad \text{and} \quad \widehat{\mathcal{H}}_t := \{f \in \mathcal{H} : \widehat{\mathcal{D}}_A(f) \leq t\}.$$

We also define the class of sensitivities to be

$$\mathcal{D}_A\mathcal{H} := \{x \mapsto |f(x) - Af(x)| : f \in \mathcal{H}\}.$$

We begin by stating the assumptions that we employ throughout the remainder of the paper. The first assumption is that the loss function is bounded and Lipschitz. These allow us to invoke the theory of Rademacher complexity, as well as make the connection between the generalisation error and the sensitivity of a function.

Recall, for a sample S of size m, the empirical Rademacher complexity of the class $\mathcal{H}$ is defined as

$$\widehat{\mathcal{R}}_S(\mathcal{H}) := \mathbb{E}_\sigma \sup_{f \in \mathcal{H}} \frac{1}{m} \sum_{k=1}^{m} \sigma_k f(x_k),$$

where $\sigma \in \{-1, 1\}^m$ is a Rademacher variable, i.e. distributed uniformly on $\{-1, 1\}$. The Rademacher complexity is

$$\mathcal{R}_m(\mathcal{H}) := \mathbb{E}_S \widehat{\mathcal{R}}_S(\mathcal{H}).$$

A classic result (Bartlett & Mendelson, 2002), Theorem 8 [see also Mohri et al. (2018), Lemma 3.3] shows that the generalisation gap scales as the Rademacher complexity – that is, we have with probability at least $1 - \delta$ that

$$|\operatorname{err}(f) - \widehat{\operatorname{err}}(f)| \leq 2\mathcal{R}_m(\mathcal{H}) + \sqrt{\frac{\ln(\frac{2}{\delta})}{m}}, \quad \text{and}$$

$$|\operatorname{err}(f) - \widehat{\operatorname{err}}(f)| \leq 2\widehat{\mathcal{R}}_S(\mathcal{H}) + 3\sqrt{\frac{\ln(\frac{2}{\delta})}{m}}.$$

We make two assumptions that let us leverage the theory of Rademacher complexities. The first one is standard.

Assumption 1 $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ is a bounded and ρ-Lipschitz loss function. That is, there exists $B > 0$ such that

$$\ell(x, y) \leq B \text{ and } |\ell(x, y) - \ell(z, y)| \leq \rho|x - z|,$$

for all $x, z \in \mathcal{X}, y \in \mathcal{Y}$. By re-scaling we may assume without loss of generality that $B = 1$.

Assumption 1 lets us bound the empirical Rademacher complexity of the loss class $\ell \circ \mathcal{H}$ with that of $\mathcal{H}$ using Talagrand's contraction lemma (Mohri et al., 2018), Lemma 5.7, that is $\widehat{\mathcal{R}}_S(\ell \circ \mathcal{H}) \leq \rho \widehat{\mathcal{R}}_S(\mathcal{H})$. Classic examples of Lipschitz loss functions include the (clipped) hinge loss, and the logistic loss ($\rho = 1$ for both). The 0-1 loss for ± 1 valued classifiers also satisfies Assumption 1 ($\rho = 1/2$), and indeed we have $\widehat{\mathcal{R}}_S(l_{01} \circ \mathcal{H}) = \frac{1}{2}\widehat{\mathcal{R}}(\mathcal{H})$ by Mohri et al. (2018), Lemma 3.4.

The second assumption we make is the uniform boundedness of the sensitivities. This will let us extend Rademacher analysis to the class of sensitivities $\mathcal{D}_A(\mathcal{H})$, which then allows us to shift the complexity terms from the full models to the approximate models.

Assumption 2 The set of sensitivities, $\mathcal{D}_A\mathcal{H}$, is uniformly bounded. That is, there exists $C > 0$ such that,

$$\|f - Af\|_\infty \leq C$$

for all $f \in \mathcal{H}$.

Assumption 2 is weaker than assuming that the functions in $\mathcal{H}$ and $\mathcal{H}_A$ are bounded. The latter is often assumed in analyses, either by constraining the norms of parameters and taking $\mathcal{X}$ to be bounded, or by passing linear outputs through a bounded nonlinearity—for instance, a sigmoidal function, or a threshold function—in the case of classification.

We start by giving a lemma that compares the true and empirical sensitivity. This is where our estimates for the size of the unlabeled sample are derived. We explore this topic further in Sect. 3.

Lemma 2.2 *With probability at least* $1 - \delta$ *we have*

$$|\mathcal{D}_A^1(f) - \widehat{\mathcal{D}}_A^1(f)| \leq 2\widehat{\mathcal{R}}_S(\mathcal{D}_A\mathcal{H}) + 3C\sqrt{\frac{\ln(\frac{2}{\delta})}{2m}},$$

for all $f \in \mathcal{H}$.

Proof By classic Rademacher bounds (Bartlett & Mendelson, 2002), Theorem 8, it holds with probability at least $1 - \delta$ that

$$\begin{aligned} |\mathcal{D}_A^1(f) - \widehat{\mathcal{D}}_A^1(f)| &= \left| \mathbb{E}_{x \sim D_x}[|f(x) - Af(x)|] - \frac{1}{m}\sum_{x \in S} |f(x) - Af(x)| \right| \\ &\leq 2\widehat{\mathcal{R}}_S(\mathcal{D}_A\mathcal{H}) + 3C\sqrt{\frac{\ln(\frac{2}{\delta})}{2m}}, \end{aligned}$$

as required. □

We now relate the generalisation error of the full model with the generalisation error of the approximate model through our notion of approximation sensitivity. The following is a key lemma as it allows us to shift from the complexity of the full precision models to the low precision models.

Lemma 2.3 *Fix $t \geq 0$. We have the following bound*

$$|\operatorname{err}(f) - \operatorname{err}(Af)| \leq \rho \mathcal{D}_A^1(f),$$

for all $f \in \mathcal{H}_t$.

Proof Let $f \in \mathcal{H}_t$. Then, by Jensen's inequality and using the Lipschitz property of ℓ we have

$$\begin{aligned} |\operatorname{err}(f) - \operatorname{err}(Af)| &\leq \mathbb{E}_{(x,y)\sim D}[|\ell(f(x), y) - \ell(Af(x), y)|] \\ &\leq \rho\, \mathbb{E}_{x\sim D_x}[|f(x) - Af(x)|] = \rho \mathcal{D}_A^1(f). \end{aligned}$$

This completes the proof. □

2.2 Learning of low approximation-sensitive predictors

Learning in high dimensional settings or complex model classes requires enormous training sets in general, or some fairly specific prior knowledge about the problem structure. However, many real-world problems possess benign traits that are hard to know in advance. Inspired by the practical success of approximate algorithms created by various model-compression methods, in this section we investigate approximability as a potential benign trait for learning, by quantifying its effect on the generalisation error. More precisely, we elaborate on our intuition that, if a relatively complex target concept admits a simpler approximation that makes little alteration to its predictive behaviour, then it should be learnable from smaller training set sizes.

The rationale is easy to see, as follows. Fix some approximation operator A and associated sensitivity threshold $t \geq 0$. Then by the classic Rademacher bound (Bartlett & Mendelson, 2002), Theorem 8, for any $\delta > 0$, with probability at least $1 - \delta$ over the draw of the training sample, we have for all $f \in \mathcal{H}_t$ that

$$\operatorname{err}(f) \leq \widehat{\operatorname{err}}(f) + 2\rho \widehat{\mathcal{R}}_S(\mathcal{H}_t) + 3\sqrt{\frac{\log(2/\delta)}{2m}}. \tag{1}$$

Let $f_t^* = \operatorname{argmin}_{f\in\mathcal{H}_t}\{\operatorname{err}(f)\}$. To learn this function, we consider a hypothetical Empirical Risk Minimiser (ERM) in the restricted class $\mathcal{H}_t$ – that is, we define the following minimum:

$$\hat{f} := \operatorname{argmin}_{f\in\mathcal{H}_t}\{\widehat{\operatorname{err}}(f)\}. \tag{2}$$

Applying (1) to the function $\hat{f}$ from (2) with a failure probability of at most $2\delta/3$, we note that $\widehat{\operatorname{err}}(\hat{f}) \leq \widehat{\operatorname{err}}(f_t^*)$ by definition of $\hat{f}$, and further note that $\widehat{\operatorname{err}}(f_t^*) \leq \operatorname{err}(f_t^*) + \sqrt{\frac{\log(3/\delta)}{2m}}$ with probability at least $1 - \delta/3$ by Hoeffding's inequality. Combining these with the use of a union bound yields that, with probability at least $1 - \delta$, $\hat{f}$ satisfies

$$\operatorname{err}(\hat{f}) \leq \operatorname{err}(f_t^*) + 2\rho \widehat{\mathcal{R}}_S(\mathcal{H}_t) + 4\sqrt{\frac{\log(3/\delta)}{2m}}. \tag{3}$$

Clearly, since $\mathcal{H}_t \subseteq \mathcal{H}$, then by a property of Rademacher complexities (Bartlett & Mendelson, 2002), Theorem 12 part 1 we have $\widehat{\mathcal{R}}_S(\mathcal{H}_t) \leq \widehat{\mathcal{R}}_S(\mathcal{H})$. So, whenever the concept we

try to learn is actually in $\mathcal{H}_t$ (i.e. a low-sensitivity target function) then, depending on $t \geq 0$, we can have a tighter guarantee compared to that of an empirical risk minimiser over the larger class $\mathcal{H}$.

Unfortunately, the minimisation in (2) is not implementable, because the specification of the function class $\mathcal{H}_t$ depends on the sensitivity function $\mathcal{D}_A$, which in turn depends on the true marginal distribution of the input data. It is often much easier to specify a larger function class $\mathcal{H}$ independent of the distribution, but this would ignore the sensitivity property and consequently lose out on the tighter guarantee.

The first approach that we consider will be based on observing that the sensitivity function only depends on inputs and is independent of the target values. Hence, we can make use of additional unlabelled data to estimate it, which is typically more widely available in applications. To this end, our first line of attack is similar in flavour with a classic semi-supervised framework proposed in Bălcan and Blum (2010). In that work, the authors augmented the standard PAC model with a notion of compatibility to encode a prior belief about the target function in terms of an expectation over the marginal distribution. As a first approach, we will instantiate their compatibility notion with our notion of approximation-sensitivity. Similarly to Bălcan and Blum (2010), this approach also allows us to use structural risk minimisation (SRM) to adapt the threshold parameter t. Therefore, balancing between the reduced complexity of the class and the potentially increased error of the best function on this reduced class yields the following result.

Theorem 1 *Fix an approximation operator A. Suppose we have an independent i.i.d. unlabelled sample $S'_x \sim D_x^{m_u}$ of size m_u, and let $\epsilon_u > 0$ s.t. $\sup_{f\in\mathcal{H}} |\mathcal{D}_A(f) - \widehat{\mathcal{D}}_A(f)| \leq \epsilon_u$ with probability at least $1-\delta/2$ with respect to the random draw of S'_x. Take an increasing sequence $(t_k)_{k\in\mathbb{N}} \subset \mathbb{R}_+$, and for each $k \in \mathbb{N}$ define $f_k^* := \operatorname{argmin}_{f\in\mathcal{H}_{t_k}} \{\operatorname{err}(f)\}$. Let $w : \mathbb{N} \to \mathbb{R}$ be such that for all $k \in \mathbb{N}$, $w_k \geq 0$ and $\sum_{k\in\mathbb{N}} w_k \leq 1$. Then, for all $k \in \mathbb{N}$ and all $f \in \hat{\mathcal{H}}_{t_k+\epsilon_u}$, with probability at least $1-\delta$, we have*:

$$\operatorname{err}(f) \leq \widehat{\operatorname{err}}(f) + 2\rho\widehat{\mathcal{R}}_S(\hat{\mathcal{H}}_{t_k+\epsilon_u}) + 3\sqrt{\frac{\log(1/w_k)}{2m}} + 3\sqrt{\frac{\log(4/\delta)}{2m}}. \tag{4}$$

Furthermore, for each $f \in \mathcal{H}$ define $\hat{k}(f) := \min\{k \in \mathbb{N} : \widehat{\mathcal{D}}_A(f) \leq t_k + \epsilon_u\}$, and consider the following algorithm

$$\hat{f} := \operatorname{argmin}_{f\in\mathcal{H}} \left\{ \widehat{\operatorname{err}}(f) + 2\rho\widehat{\mathcal{R}}_S\left(\hat{\mathcal{H}}_{t_{\hat{k}(f)}+\epsilon_u}\right) + 3\sqrt{\frac{\log\left(1/w_{\hat{k}(f)}\right)}{2m}} \right\}. \tag{5}$$

Then, with probability at least $1-\delta$ we have

$$\operatorname{err}(\hat{f}) \leq \min_{k\in\mathbb{N}} \left\{ \operatorname{err}(f_k^*) + 2\rho\widehat{\mathcal{R}}_S(\hat{\mathcal{H}}_{t_k+\epsilon_u}) + 4\sqrt{\frac{\log(1/w_k)}{2m}} \right\} + 3\sqrt{\frac{\log(6/\delta)}{2m}}. \tag{6}$$

Before giving the proof, we make a few comments. Firstly, we see that, with a large enough m_u (i.e. sufficient additional unlabelled data), we have by Lemma 2.2 with probability $1-\delta/2$ that, the magnitude of $\epsilon_u \leq 2\widehat{\mathcal{R}}_{S'_x}(\mathcal{D}_A\mathcal{H}) + 3\sqrt{\frac{\log(4/\delta)}{2m_u}}$ can be made arbitrarily small – this is the only role of S'_x. A detailed account of the possible ranges of magnitude of this quantity will be discussed in Sect. 3, along with some natural factors that make it

small. For now, let us point out that, by construction, whenever both $\mathcal{H}$ and $\mathcal{H}_A$ are PAC-learnable, and without further conditions, the complexity of our sensitivity class is determined by the complexities of $\mathcal{H}$ and $\mathcal{H}_A$ [see discussion around (29)]. By contrast, the general setting of semi-supervised learning in Bălcan and Blum (2010) allows arbitrarily complex compatibility classes, which, in a worst case scenario can backfire and blow up the required labelled data size (Bălcan & Blum, 2010), Theorem 22.

The objective of the minimisation algorithm in (5) follows the idea of minimising the uniform bound (4). It finds a good predictor along with the appropriate subclass of $\mathcal{H}$ to which it belongs. The sequence of sensitivity threshold candidates $(t_k)_{k\in\mathbb{N}}$, and the associated weights $(w_k)_{k\in\mathbb{N}}$, with $w_k \geq 0$ for all $k \in \mathbb{N}$ and $\sum_{k\in\mathbb{N}} w_k \leq 1$, must be chosen before seeing any data (for instance, $w_k := 2^{-k}$), with w_k representing an a-priori belief in a particular t_k.

As a further observation, the function classes $\hat{\mathcal{H}}_{t_k+\epsilon_u}$ that feature in the high probability guarantee (6) are dependent on the unlabelled data. This dependence can be removed if desired, by noting that, with high probability, $\widehat{\mathcal{D}}^1_A(f) \leq t + \varepsilon_u$ implies $\mathcal{D}^1_A(f) \leq t + 2\varepsilon_u$ for all $f \in \hat{\mathcal{H}}_{t_k+\epsilon_u}$ – hence we have $\hat{\mathcal{H}}_{t_k+\epsilon_u} \subseteq \mathcal{H}_{t_k+2\epsilon_u}$, which in turn implies $\widehat{\mathcal{R}}_S(\hat{\mathcal{H}}_{t_k+\epsilon_u}) \leq \widehat{\mathcal{R}}_S(\mathcal{H}_{t_k+2\epsilon_u})$ with high probability. In fact, the failure probability of this bound is already accounted for in the proof of (6), so replacing $\widehat{\mathcal{R}}_S(\hat{\mathcal{H}}_{t_k+\epsilon_u})$ by $\widehat{\mathcal{R}}_S(\mathcal{H}_{t_k+2\epsilon_u})$ in (6) holds with the same probability as the stated.

Lastly, but most importantly, since $\hat{\mathcal{H}}_{t_k+\epsilon_u} \subseteq \mathcal{H}$, we have $\widehat{\mathcal{R}}_S(\hat{\mathcal{H}}_{t_k+\epsilon_u}) \leq \widehat{\mathcal{R}}_S(\mathcal{H})$. The extent of this reduction of complexity depends on several factors even for specific approximation choices, including the sensitivity the unknown target function, the magnitude of ϵ_u and the threshold estimate $t_k + \epsilon_u$, the original class $\mathcal{H}$, and the data distribution. For the sake of intuition, suppose that availability of unlabelled data is not a barrier, so the potential gain is down to the interaction between the unknown data distribution and the unknown target function. A low sensitivity asserts that, for the particular approximation A, only a small mass fraction of the input points is affected by subjecting a predictor to A. If the target function satisfies this, and the marginal distribution is such that most functions of $\mathcal{H}$ do not satisfy this, then $\mathcal{H}_t$ (i.e. the remaining set of functions that have low sensitivity) will be small. Let us consider some informal examples.

Example 1. We can think of a model approximation as a perturbation of the model. In classification, this induces a perturbation of the decision boundary. If the true classes are well separated by a large margin, then there is leeway for such perturbation. Hence, just as in the framework in Bălcan and Blum (2010), dense classes separated by a large margin will rule out all functions that cut across dense regions, leaving a handful few – especially if $\mathcal{H}$ was a simple class, such as linear predictors.

Furthermore, in the extreme case of zero sensitivity we can use the simpler class $\mathcal{H}_A$ instead of $\mathcal{H}$, as in the following.

Example 2. Consider a relatively complex parametric class $\mathcal{H}$, and a coarse quantisation as A. Then $\mathcal{H}_A$ simply becomes a finite hypothesis class. If the target function is insensitive to this approximation, then it is enough to work with $\mathcal{H}_A$ and have the guarantees enjoyed by the finite class.

Example 3. Suppose the functions in $\mathcal{H}$ have a large number of parameters so $\mathcal{H}$ has high complexity, but the data distribution is supported in a simple restricted set that makes much of the representational capacity of $\mathcal{H}$ remain dormant. Then the effect of a model-compression will spread out among both relevant and irrelevant parameters, making less of a noticeable difference to the function values.

While these examples are both simplistic and informal, ample empirical evidence in the literature demonstrates that many model approximation methods do work surprisingly well

in practice. In the next section we aim to develop an approach that helps to untangle and shed more light onto the various contributing factors that influence the error when learning involves approximate predictors. But first we prove Theorem 1.

Proof of Theorem 1 For a fixed $t \geq 0$, by the definition of ϵ_u, with probability $1-\delta/2$, we have that $f \in \mathcal{H}_t$ implies $f \in \hat{\mathcal{H}}_{t+\epsilon_u}$. We shall pursue SRM by exploiting the independent unlabelled sample to define a nested sequence of function classes $\hat{\mathcal{H}}_{t_1+\epsilon_u} \subseteq \hat{\mathcal{H}}_{t_2+\epsilon_u} \cdots \subseteq \hat{\mathcal{H}}_{t_k+\epsilon_u} \subseteq \hat{\mathcal{H}}_{t_{k+1}+\epsilon_u} \subseteq \cdots \subseteq \mathcal{H}$ where $k \in \mathbb{N}$. These classes depend on the unlabelled sample, but not on the labelled sample. For any fixed $k \in \mathbb{N}$, the classic Rademacher bound (Bartlett & Mendelson, 2002), Theorem 8 implies with probability at least $1-(w_k\delta/2)$ that all $f \in \hat{\mathcal{H}}_{t_k+\epsilon_u}$ satisfy

$$\operatorname{err}(f) \leq \widehat{\operatorname{err}}(f) + 2\rho\widehat{\mathcal{R}}_S(\hat{\mathcal{H}}_{t_k+\epsilon_u}) + 3\sqrt{\frac{\log(4/(\delta w_k))}{2m}}.$$

Since $k \in \mathbb{N}$ is arbitrary, and the non-negative weights satisfy $\sum_{k\in\mathbb{N}} w(k) \leq 1$, we take a union bound and it follows with probability at least $1-\frac{\delta}{2}$ that, uniformly for all $k \in \mathbb{N}$ and all $f \in \hat{\mathcal{H}}_{t_k+\epsilon_u}$ we have

$$\operatorname{err}(f) \leq \widehat{\operatorname{err}}(f) + 2\rho\widehat{\mathcal{R}}_S(\hat{\mathcal{H}}_{t_k+\epsilon_u}) + 3\sqrt{\frac{\log(1/w_k)}{2m}} + 3\sqrt{\frac{\log(4/\delta)}{2m}}.$$

This proves (4).

To obtain (6) for $\hat{f}$ defined in (5), we apply (4) to $\hat{f}$. By construction, $\hat{f} \in \hat{\mathcal{H}}_{t_{\hat{k}(\hat{f})}+\epsilon_u}$. Recall also that with probability at least $1-\frac{\delta}{2}$ we have $f_k^* \in \hat{\mathcal{H}}_{t_k+\epsilon_u}$ as $f_k^* \in \mathcal{H}_{t_k}$. Therefore, with probability at least $1-\delta/3$,

$$\operatorname{err}(\hat{f}) \leq \widehat{\operatorname{err}}(\hat{f}) + 2\rho\widehat{\mathcal{R}}_S(\hat{\mathcal{H}}_{t_{\hat{k}(\hat{f})}+\epsilon_u}) + 3\sqrt{\frac{\log(1/w_{\hat{k}(\hat{f})})}{2m}} + 3\sqrt{\frac{\log(2\cdot 3/\delta)}{2m}} \tag{7}$$

$$\leq \widehat{\operatorname{err}}(f_k^*) + 2\rho\widehat{\mathcal{R}}_S(\hat{\mathcal{H}}_{t_k+\epsilon_u}) + 3\sqrt{\frac{\log(1/w_k)}{2m}} + 3\sqrt{\frac{\log(6/\delta)}{2m}}, \tag{8}$$

for all $k \in \mathbb{N}$. In the last inequality we used the definition of $\hat{f}$ noting that the right hand side of (7) is minimised by $\hat{f}$. In addition, by Hoeffding's inequality, we also have $\widehat{\operatorname{err}}(f_k^*) \leq \operatorname{err}(f_k^*) + \sqrt{\frac{\log(6/(w_k\delta))}{2m}}$ with probability at least $1-(w_k\delta)/6$. Combining with (8) and using the union bound, it follows with probability at least $1-\delta$ that

$$\operatorname{err}(\hat{f}) \leq \operatorname{err}(f_k^*) + 2\rho\widehat{\mathcal{R}}_S(\hat{\mathcal{H}}_{t_k+\epsilon_u}) + 4\sqrt{\frac{\log(1/w_k)}{2m}} + 3\sqrt{\frac{\log(6/\delta)}{2m}},$$

for all $k \in \mathbb{N}$. Finally, choosing k to minimise the bound concludes the proof. □

2.3 A joint approach to sensitivity and generalisation

The conceptually straightforward approach of the previous subsection implies that a target concept that is robust to the effects of approximation by a low-complexity predictor,

may require less labelled examples to be learned. In particular, the regularised ERM algorithm defined in (5) can accomplish this learning task, the regulariser being the empirical Rademacher complexity of the restricted class $\hat{\mathcal{H}}_{t_{\hat{k}(f)}}$, along with a penalty for estimating $\hat{k}(f)$. In effect, this algorithm adaptively trims the original function class to the relevant subset of low-sensitivity predictors, and consequently returns a low-sensitivity element of an otherwise potentially much larger function class.

The appeal of this finding lies not only to serve as a possible explanation towards the question of what makes some instances of a learning problem easier than others. Also, by the low-sensitivity property, such predictor should be usable in its approximated form in memory-constrained settings. Indeed, for any $t \geq 0$, if $\hat{f} \in \mathcal{H}_t$, then by Lemma 2.3 we have

$$\operatorname{err}(A\hat{f}) = \operatorname{err}(\hat{f}) + \left(\operatorname{err}(A\hat{f}) - \operatorname{err}(\hat{f})\right) \leq \operatorname{err}(\hat{f}) + \rho \mathcal{D}_A(\hat{f}) \leq \operatorname{err}(\hat{f}) + \rho t. \tag{9}$$

In other words, for a predictor with low approximation-sensitivity, using $A\hat{f}$ instead of $\hat{f}$ will only incur an additive error of up to ρt. This additional term is the price to pay for predicting with the simplified function $A\hat{f}$ instead of the full-precision function $\hat{f}$ – it will not improve with more data, but t is small precisely when the target function we try to learn has a low-sensitivity.

In this section we are interested in a more practical formulation of the tandem of learning an approximate predictor as well as a full precision predictor. The approach presented so far, beyond its conceptual elegance, has some practical drawbacks: (1) it requires an additional independent unlabelled data set; and (2) it requires computing the empirical Rademacher complexity of the restricted class. Computing empirical Rademacher complexities is known to be typically a hard combinatorial optimisation problem for interesting hypothesis classes (Bartlett & Mendelson, 2002), as it amounts to computing an empirical risk minimiser under the 0–1 loss.

To get around these limitations, we shall take a different approach. We start by modifying the loss function to explicitly encode the fact that we are interested in a good low-complexity approximate predictor. More precisely, for a given threshold value $t \geq 0$, we start by defining the minimiser of the following constrained function:

$$\hat{f}_t := \operatorname{argmin}_{f \in \mathcal{H}_t} \{ \widehat{\operatorname{err}}(Af) \} \tag{10}$$

This is defined for the purpose of theoretical analysis, it is not computable, since checking $f \in \mathcal{H}_t$ would require knowledge of the data distribution.

The following result shows that the function $\hat{f}_t$ in (10) achieves two different functionalities simultaneously, as it not only produces a good approximate predictor with quantified error guarantee, including the price to pay for the approximation, but $\hat{f}_t$ itself is a good predictor whenever the problem admits an approximable target function.

Proposition 2.4 *Fix an approximation operator A and $t \geq 0$. Define $f_t^* := \operatorname{argmin}_{f \in \mathcal{H}_t} \{ \operatorname{err}(f) \}$ and $g_t^* := \operatorname{argmin}_{g \in A\mathcal{H}_t} \{ \operatorname{err}(g) \}$. Then, with probability at least $1 - \delta$, the function $\hat{f}_t$ from (10) satisfies all of the following simultaneously*:

$$\operatorname{err}(A\hat{f}_t) \leq \min\{ \operatorname{err}(Af_t^*), \operatorname{err}(g_t^*) \} + 2\rho \widehat{\mathcal{R}}_S(\mathcal{H}_A) + 4\sqrt{\frac{\ln(\frac{9}{\delta})}{2m}}, \tag{11}$$

$$\operatorname{err}(A\hat{f}_t) \leq \operatorname{err}(f_t^*) + \rho t + 2\rho\widehat{\mathcal{R}}_S(\mathcal{H}_A) + 4\sqrt{\frac{\ln(\frac{9}{\delta})}{2m}} \tag{12}$$

$$\operatorname{err}(\hat{f}_t) \leq \operatorname{err}(f_t^*) + 2\rho t + 2\rho\widehat{\mathcal{R}}_S(\mathcal{H}_A) + 4\sqrt{\frac{\ln(\frac{9}{\delta})}{2m}}. \tag{13}$$

We note that $\rho t \to 0$ as $t \to 0$; however, as t decreases, the choice of predictors in $\mathcal{H}_t$ decreases too, and so $\operatorname{err}(f_t^*)$ would be expected to increase. That is, the choice of t balances the trade-off between the sensitivity term ρt, and the error term, $\operatorname{err}(f_t^*)$.

Proposition 2.4 allows us to view learning and model-compression as two sides of the same coin. Eq. (13) suggests that low-sensitivity target functions are easier to learn, and a *constrained* ERM algorithm is able to learn it up to a constant factor of its sensitivity. Indeed, suppose $f^* = f_t^*$, i.e. the target function has sensitivity below t. Then the error of $\hat{f}_t$ is guaranteed to be much smaller than the worst case error of finding f^* in the whole class $\mathcal{H}$. At the same time, (12) provides a guarantee for the approximate predictor $A\hat{f}$ that can potentially be deployed in low-memory settings by paying the additive term ρt proportional to the extent of approximability. Remarkably, both of these two seemingly different goals are accomplished by the same function $\hat{f}_t$ defined in (10). Moreover (11) gives guarantees for $A\hat{f}_t$ relative to both Af_t^* (the approximation of the best predictor in $\mathcal{H}$ with sensitivity of at most t) and g_t^* (the best approximate predictor in $A\mathcal{H}_t$).

Proof of Proposition 2.4 By Rademacher bounds (Bartlett & Mendelson, 2002), Theorem 8 and Talagrand's contraction lemma (Mohri et al., 2018), Lemma 5.7, we have with probability at least $1 - \frac{2\delta}{9}$, that

$$\begin{aligned}\operatorname{err}(A\hat{f}_t) &\leq \widehat{\operatorname{err}}(A\hat{f}_t) + 2\widehat{\mathcal{R}}_S(\ell \circ A\mathcal{H}_t) + 3\sqrt{\frac{\ln(\frac{2\cdot 9}{2\delta})}{2m}} \\ &\leq \widehat{\operatorname{err}}(A\hat{f}_t) + 2\rho\widehat{\mathcal{R}}_S(A\mathcal{H}_t) + 3\sqrt{\frac{\ln(\frac{9}{\delta})}{2m}}.\end{aligned} \tag{14}$$

By definition of $\hat{f}_t$ we have $\widehat{\operatorname{err}}(A\hat{f}_t) \leq \min\{\widehat{\operatorname{err}}(Af_t^*), \widehat{\operatorname{err}}(g_t^*)\}$. Using this together with Hoeffding's inequality, with probability $1 - \frac{\delta}{9}$ both

$$\widehat{\operatorname{err}}(A\hat{f}_t) \leq \widehat{\operatorname{err}}(Af_t^*) \leq \operatorname{err}(Af_t^*) + \sqrt{\frac{\ln(\frac{9}{\delta})}{2m}}, \text{ and}$$

$$\widehat{\operatorname{err}}(A\hat{f}_t) \leq \widehat{\operatorname{err}}(g_t^*) \leq \operatorname{err}(g_t^*) + \sqrt{\frac{\ln(\frac{9}{\delta})}{2m}}$$

hold separately. Therefore, by the union bound and the fact that $\widehat{\mathcal{R}}_S(A\mathcal{H}_t) \leq \widehat{\mathcal{R}}_S(\mathcal{H}_A)$ we have with probability at least $1 - \frac{4\delta}{9}$, that (11) holds. Similarly, as $\widehat{\operatorname{err}}(A\hat{f}_t) \leq \widehat{\operatorname{err}}(Af_t^*)$ and by Lemma 2.3, we have with probability at least $1 - \frac{\delta}{9}$, that

$$\widehat{\operatorname{err}}(A\hat{f}_t) \leq \widehat{\operatorname{err}}(Af_t^*) \leq \operatorname{err}(Af_t^*) + \sqrt{\frac{\ln(\frac{9}{\delta})}{2m}} \leq \operatorname{err}(f_t^*) + \rho\mathcal{D}_A^1(f_t^*) + \sqrt{\frac{\ln(\frac{9}{\delta})}{2m}} \tag{15}$$

$$\leq \operatorname{err}(f_t^*) + \rho t + \sqrt{\frac{\ln(\frac{9}{\delta})}{2m}}. \tag{16}$$

Combining the above three inequalities and the fact that $\widehat{\mathcal{R}}_S(A\mathcal{H}_t) \leq \widehat{\mathcal{R}}_S(\mathcal{H}_A)$ we have with probability at least $1 - \frac{2\delta}{9}$, that

$$\operatorname{err}(A\hat{f}_t) \leq \operatorname{err}(f_t^*) + 2\rho t + 2\rho\widehat{\mathcal{R}}_S(\mathcal{H}_A) + 4\sqrt{\frac{\ln(\frac{9}{\delta})}{2m}}. \tag{17}$$

This proves (12). The second part follows by using Lemma 2.3, Jensen's inequality and $\hat{f}_t \in \mathcal{H}_t$, so we have

$$\operatorname{err}(\hat{f}_t) \leq \rho\mathcal{D}_A^1(\hat{f}_t) + \operatorname{err}(A\hat{f}_t) \leq \rho\mathcal{D}_A(\hat{f}_t) + \operatorname{err}(A\hat{f}_t) \leq \rho t + \operatorname{err}(A\hat{f}_t).$$

Taking the union bound for each of the equations completes the proof.

Next, we show that in this formulation we can relax the fixed parameter t that constrains the function class, without the use of SRM. To avoid clutter, here we suppose the functional form of $f \mapsto \mathcal{D}_A(f)$ is known – this can be estimated from an independent unlabelled data set as in the previous section.

To this end, consider the minimiser of the following (hypothetical) objective function, used for theoretical analysis.

$$\hat{f} := \operatorname{argmin}_{f\in\mathcal{H}}\{\widehat{\operatorname{err}}(Af) + \rho\mathcal{D}_A(f)\} \tag{18}$$

Here the first term is our modified loss function as before, and the second term acts as a regulariser that implicitly constrains the function class. The following result shows that $\hat{f}$ from (18) behaves as the previous minimiser from (10), while it also automatically adapts the class-constraining sensitivity threshold t.

Proposition 2.5 *Fix an approximation operator A. For $t \geq 0$, let $f_t^* := \operatorname{argmin}_{f\in\mathcal{H}_t}\{\operatorname{err}(f)\}$. For the function $\hat{f}$ defined in (18), with probability at least $1-\delta$ we have both of the following*

$$\operatorname{err}(A\hat{f}) \leq \min_{t\geq 0}\left\{\operatorname{err}(f_t^*) + 2\rho t\right\} + 2\rho\widehat{\mathcal{R}}_S(\mathcal{H}_A) + 4\sqrt{\frac{\ln(\frac{8}{\delta})}{2m}}, \text{ and} \tag{19}$$

$$\operatorname{err}(\hat{f}) \leq \min_{t\geq 0}\left\{\operatorname{err}(f_t^*) + 2\rho t\right\} + 2\rho\widehat{\mathcal{R}}_S(\mathcal{H}_A) + 4\sqrt{\frac{\ln(\frac{8}{\delta})}{2m}}, \tag{20}$$

simultaneously.

Proof of Proposition 2.5 Using Lemma 2.3 and Rademacher bounds (Bartlett & Mendelson, 2002), Theorem 8, we have with probability at least $1 - \frac{\delta}{4}$, that

$$\operatorname{err}(\hat{f}) \leq \operatorname{err}(A\hat{f}) + \rho\mathcal{D}_A^1(\hat{f}) \tag{21}$$

$$\leq \widehat{\mathrm{err}}\,(A\hat{f}) + \rho\mathcal{D}_A^1(\hat{f}) + 2\widehat{\mathcal{R}}_S(\ell\circ\mathcal{H}_A) + 3\sqrt{\frac{\ln(\frac{2\cdot 4}{\delta})}{2m}}. \tag{22}$$

Let $g := \mathrm{argmin}_{t\geq 0}\{\,\mathrm{err}\,(f_t^*) + 2\rho\mathcal{D}_A^1(f_t^*)\}$. Then by the definition of $\hat{f}$ and the Hoeffding bound we obtain with a probability of at least $1-\frac{\delta}{8}$, that

$$\widehat{\mathrm{err}}\,(A\hat{f}) + \rho\mathcal{D}_A^1(\hat{f}) \leq \widehat{\mathrm{err}}\,(Ag) + \rho\mathcal{D}_A^1(g) \leq \mathrm{err}\,(Ag) + \rho\mathcal{D}_A^1(g) + \sqrt{\frac{\ln(\frac{8}{\delta})}{2m}}. \tag{23}$$

Then, by Lemma 2.3 and definition of g we have

$$\mathrm{err}\,(Ag) + \rho\mathcal{D}_A^1(g) \leq \mathrm{err}\,(g) + 2\rho\mathcal{D}_A^1(g) \leq \mathrm{err}\,(f_t^*) + 2\rho\mathcal{D}_A^1(f_t^*),$$

for all $t \geq 0$. Hence, $\mathrm{err}\,(Ag) + \rho\mathcal{D}_A^1(g) \leq \min_{t\geq 0}\{\,\mathrm{err}\,(Af_t^*) + 2\rho\mathcal{D}_A^1(f_t^*)\}$, and substituting into (23) yields

$$\widehat{\mathrm{err}}\,(A\hat{f}) + \rho\mathcal{D}_A^1(\hat{f}) \leq \min_{t\geq 0}\{\,\mathrm{err}\,(Af_t^*) + 2\rho\mathcal{D}_A^1(f_t^*)\} + \sqrt{\frac{\ln(\frac{8}{\delta})}{2m}}.$$

with probability at least $1-\frac{\delta}{8}$. By the Talagrand contraction lemma (Mohri et al., 2018), Lemma 5.7 we have $\widehat{\mathcal{R}}_S(\ell\circ\mathcal{H}_A) \leq \rho\widehat{\mathcal{R}}_S(\mathcal{H}_A)$, and so combining with (22) and then by a union bound we have with probability at least $1-\frac{\delta}{2}$ that

$$\mathrm{err}\,(\hat{f}) \leq \min_{t\geq 0}\{\,\mathrm{err}\,(Af_t^*) + 2\rho\mathcal{D}_A^1(f_t^*)\} + \sqrt{\frac{\ln(\frac{8}{\delta})}{2m}} + 2\rho\widehat{\mathcal{R}}_S(\mathcal{H}_A) + 4\sqrt{\frac{\ln(\frac{8}{\delta})}{2m}}.$$

Noting that $\mathcal{D}_A(f_t^*) \leq t$ completes the proof of (20). Eq. (19) also follows, with probability at least $1-\frac{\delta}{2}$, since $\mathrm{err}\,(A\hat{f})$ is upper bounded by the right hand side of (21), by adding the non-negative term $\rho\mathcal{D}_A(f)$.

From Proposition 2.5 we see again that, for any fixed approximation function A such that $\mathcal{H}_A$ has smaller complexity than $\mathcal{H}$, if the target function has a low sensitivity (i.e. $\mathcal{D}_A(f^*)$ is small), then it is learnable from fewer labels than an arbitrary target from $\mathcal{H}$ would be. Of course, there may be learning problems where f^* has low error but high sensitivity for the pre-defined A, but the minimiser in (18) is a function that automatically balances between generalisation error and sensitivity.

It is now straightforward to use an estimate of $\mathcal{D}_A(f)$, giving rise to a learning algorithm that is an implementable version of the construct analysed in Proposition 2.5.

Theorem 2 (Joint learning of full and approximate predictors) *Fix an approximation operator A, and consider the following algorithm.*

$$\hat{f} := \mathrm{argmin}_{f\in\mathcal{H}}\{\,\widehat{\mathrm{err}}\,(Af) + \rho\widehat{\mathcal{D}}_A(f)\}. \tag{24}$$

Let $\epsilon_u > 0$ be such that $\sup_{f\in\mathcal{H}}|\mathcal{D}_A(f) - \widehat{\mathcal{D}}_A(f)| \leq \epsilon_u$ with probability at least $1-\frac{\delta}{2}$ with respect to $D_x^{m_u}$ where $m_u \geq m$. For $t \geq 0$, let $f_t^* := \mathrm{argmin}_{f\in\mathcal{H}_t}\{\,\mathrm{err}\,(f)\}$. Then with probability at least $1-\delta$, the function $\hat{f}$ satisfies both

$$\operatorname{err}(A\hat{f}) \leq \min_{t\geq 0} \left\{ \operatorname{err}(f_t^*) + 2\rho t \right\} + 2\rho \widehat{\mathcal{R}}_S(\mathcal{H}_A) + (4+\rho)\sqrt{\frac{\ln(\frac{16}{\delta})}{2m}} + \rho\epsilon_u, \text{ and}$$

$$\operatorname{err}(\hat{f}) \leq \min_{t\geq 0} \left\{ \operatorname{err}(f_t^*) + 2\rho t \right\} + 2\rho \widehat{\mathcal{R}}_S(\mathcal{H}_A) + (4+\rho)\sqrt{\frac{\ln(\frac{16}{\delta})}{2m}} + \rho\epsilon_u,$$

simultaneously.

Proof This follows by the same steps as the proof of Proposition 2.5 combined with Lemma 2.2. □

Let us compare Theorem 2 with Theorem 1. With sufficient unlabelled data ϵ_u can be made arbitrarily small, in both theorems. However, in Theorem 1 the unlabelled sample for estimating ϵ_u must be independent of the labelled sample; this is because in that construction the function class depends on the unlabelled data through the sensitivity estimate. By contrast, in Theorem 2 we have an implicit adaptation of t, so the function class does not depend on the unlabelled sample. This enables us to reuse the labelled points also for estimating the sensitivity, and any additional unlabelled data just contributes to further shrinking ϵ_u. Hence, in Theorem 2 whenever ϵ_u is already small enough using the m training points of S, we do not even require any additional unlabelled points. In later sections we will see natural conditions where this is easily the case.

The advantage of the algorithm analysed in Theorem 1 is its statistical consistency, since given enough labelled data the generalisation error converges to that of the best predictor of the class. However, if the goal is to obtain an approximate predictor, we pay the price of an additive sensitivity term (9), and Theorem 2 shows that allowing such term enables a much more implementation-friendly algorithm without sacrificing the essence of the theoretical guarantee on generalisation.

Comparing the algorithm from Theorem 2 with that of Theorem 1, observe the difference in the regularisation term. Regularising with the sensitivity estimate was not justified in the formulation of Theorem 1, and indeed the authors of Bălcan and Blum (2010) have pointed out that regularising with their general compatibility estimate was not theoretically justified – despite it being used in practice (Chapelle et al., 2006). By contrast, in the formulation of Theorem 2, we have been able to justify it within our approximability objective.

2.4 Managing the trade-off between sample error and sensitivity for the approximate predictor

The analysis from Proposition 2.5 and Theorem 2 have shown that the associated algorithm has an implicit ability to realise the optimal trade-off between the sample error of $A\hat{f}$ and the sensitivity term, t, without any effort or tuning parameter from the user.

However, there may be situations when a different trade-off is desired, and in such a case we want to manage this trade-off as a tuning parameter. This is especially relevant for practical applications in memory-constrained settings, where obtaining a good approximate predictor $A\hat{f}$ is the sole interest. For instance, we may only care about very low sensitivity functions at the expense of a slightly raised error, or vice-versa. Or we might like to explore multiple trade-offs as in a multi-objective approach. Another

instance of this is when unlabelled data is also scarce but an analytic upper bound can be derived on the sensitivity function up to an unknown constant.

Conceptually, a good way to address this sort of issues would be to take back control over the threshold parameter t using the learning algorithm in (10) (with or without estimating the sensitivity). However, the constrained optimisation formulation can be awkward to perform in practice. Below we suggest a more user-friendly form of the algorithm, and show that its solution is close to that of (10).

For each $\lambda \geq 0$ consider the following algorithm

$$\tilde{f}_\lambda := \operatorname{argmin}_{f \in \mathcal{H}} \{ \widehat{\operatorname{err}}\,(Af) + \lambda \widehat{\mathcal{D}}_A(f) \}. \tag{25}$$

Algorithms of this form, including the exploitation of unlabelled data in the regularisation term, have been in use in practice for a long time (Chapelle et al., 2006), see also (van Engelen & Hoos, 2020). The regularisation parameter λ balances the two terms of the objective function, and in addition to potential availability of prior knowledge, there is a wide range of well-established model selection methods available to set this parameter in practice.

To this end, we shall compare the error of $\tilde{f}_\lambda$ from algorithm (25) with that for $\hat{f}_t$ from the algorithm given in (10). The following proposition shows that, for any specification of λ, there is a value of $t \geq 0$ such that the errors of these two predictors are close, up to additive terms that decay with the sample size.

Theorem 3 (Balancing sample error & sensitivity) *Let $\epsilon_u > 0$ be such that $\sup_{f \in \mathcal{H}} |\mathcal{D}_A(f) - \widehat{\mathcal{D}}_A(f)| \leq \epsilon_u$ with probability at least $1 - \delta/4$ with respect to $D_x^{m_u}$, where $m_u \geq m$. For any $\lambda > 0$, there exists $t > 0$ such that with probability at least $1 - \delta$ we have*

$$\operatorname{err}(A\tilde{f}_\lambda) - \operatorname{err}(A\hat{f}_t) \leq 4\rho \widehat{\mathcal{R}}_S(\mathcal{H}_A) + 6\sqrt{\frac{\ln(\frac{8}{\delta})}{2m}} + 2\lambda\epsilon_u. \tag{26}$$

Proof of Theorem 3 Take $t \leq \mathcal{D}_A(\tilde{f}_\lambda)$. Then from the definition of algorithm (10) we have $\mathcal{D}_A(\hat{f}_t) \leq t \leq \mathcal{D}_A(\tilde{f}_\lambda)$. Using this, the definition of $\tilde{f}_\lambda$, and Lemma 2.2, it follows with probability at least $1 - \frac{\delta}{4}$ that

$$\begin{aligned} \widehat{\operatorname{err}}\,(A\tilde{f}_\lambda) + \lambda \widehat{\mathcal{D}}_A(\tilde{f}_\lambda) &\leq \widehat{\operatorname{err}}\,(A\hat{f}_t) + \lambda \widehat{\mathcal{D}}_A(\hat{f}_t) \\ &\leq \widehat{\operatorname{err}}\,(A\hat{f}_t) + \lambda \mathcal{D}_A(\hat{f}_t) + \lambda\epsilon_u \\ &\leq \widehat{\operatorname{err}}\,(A\hat{f}_t) + \lambda \mathcal{D}_A(\tilde{f}_\lambda) + \lambda\epsilon_u. \end{aligned}$$

Rearranging, and using Lemma 2.2 again, we have with probability at least $1 - \delta/2$ that

$$\widehat{\operatorname{err}}\,(A\tilde{f}_\lambda) - \widehat{\operatorname{err}}\,(A\hat{f}_t) \leq \lambda(\mathcal{D}_A(\tilde{f}_\lambda) - \widehat{\mathcal{D}}_A(\tilde{f}_\lambda)) + \lambda\epsilon_u \leq 2\lambda\epsilon_u.$$

This shows that the sample errors of the two predictors are close.

Now, to prove (26) we again use Rademacher bounds (Bartlett & Mendelson, 2002), Theorem 8 with probability at least $1 - \delta/2$ on $\mathcal{H}_A$ twice, combined with (24) and the union bound. We have with probability $1 - \delta$,

$$\begin{aligned}\operatorname{err}(A\tilde{f}_\lambda) - \operatorname{err}(A\hat{f}_t) &= (\operatorname{err}(A\tilde{f}_\lambda) - \widehat{\operatorname{err}}(A\tilde{f}_\lambda)) + (\widehat{\operatorname{err}}(A\tilde{f}_\lambda)) - \widehat{\operatorname{err}}(A\hat{f}_t)) \\ &\quad + (\widehat{\operatorname{err}}(A\hat{f}_t) - \operatorname{err}(A\hat{f}_t)) \\ &\leq 2\lambda\epsilon_u + 2\left(2\rho\widehat{\mathcal{R}}_S(\mathcal{H}_A) + 3\sqrt{\frac{\ln\left(\frac{8}{\delta}\right)}{2m}}\right),\end{aligned}$$

as required.

The comments we made on Theorem 2 also apply to Theorem 3. In particular, the training points contribute to estimating the sensitivity also, unlike the approach in Theorem 1 which required a separate independent unlabelled sample.

As a further remark, let us also address the case when instead of estimating the sensitivity from unlabelled data we have an analytic upper bound on this function, in the case of some specific choice of function class and approximation operator, up to some unknown absolute constant. The constant will be subsumed into the tuning parameter λ.

Let $\overline{\mathcal{D}_A}(\cdot)$ be a mapping from $\mathcal{H}$ to $\mathbb{R}_+$ where there exists $c > 0$ such that for all $f \in \mathcal{H}$, we have $\mathcal{D}_A(f) \leq c \cdot \overline{\mathcal{D}_A}(f)$. Note that, $\overline{\mathcal{D}_A}(\cdot)$ does not depend on the sample. Now, for each $\lambda \geq 0$ define the following algorithm

$$\bar{f}_\lambda := \operatorname{argmin}_{f\in\mathcal{H}}\{\widehat{\operatorname{err}}(Af) + \lambda\overline{\mathcal{D}_A}(f)\}. \tag{27}$$

Furthermore, let $\hat{f}_t$ be the predictor returned by algorithm (10), and $\breve{f}_t$ the predictor from a version of the same algorithm (10) that replaces the unknown $\mathcal{D}_A(\cdot)$ with $\overline{\mathcal{D}_A}(\cdot)$. Then $\breve{f}_t$ will have a guarantee of the same form as before in Proposition 2.4 where t is now a threshold on $\overline{\mathcal{D}_A}(\cdot)$ rather than $\mathcal{D}_A(\cdot)$. The following remark shows that the error of $\bar{f}_\lambda$ is close to that of $\breve{f}_t$.

Remark 2.6 For any $\lambda > 0$, there exists $t > 0$ such that with probability at least $1 - \delta$ we have

$$\operatorname{err}(A\bar{f}_\lambda) - \operatorname{err}(A\bar{f}_t) \leq 4\rho\widehat{\mathcal{R}}_S(\mathcal{H}_A) + 6\sqrt{\frac{\ln(\frac{8}{\delta})}{2m}}. \tag{28}$$

Proof Let $t \geq 0$ be such that $t \leq \overline{\mathcal{D}_A}(\bar{f}_\lambda)$. Then $\overline{\mathcal{D}_A}(\breve{f}_t) \leq \overline{\mathcal{D}_A}(\bar{f}_\lambda)$. Consequently, by the definition of $\bar{f}_\lambda$, we have

$$\widehat{\operatorname{err}}(A\bar{f}_\lambda) + \lambda\overline{\mathcal{D}_A}(\bar{f}_\lambda) \leq \widehat{\operatorname{err}}(A\bar{f}_t) + \lambda\overline{\mathcal{D}_A}(\bar{f}_t) \leq \widehat{\operatorname{err}}(A\breve{f}_t) + \lambda\overline{\mathcal{D}_A}(\bar{f}_\lambda).$$

Therefore, $\widehat{\operatorname{err}}(A\bar{f}_\lambda)) \leq \widehat{\operatorname{err}}(A\breve{f}_t)$. Using this, we have

$$\begin{aligned}\operatorname{err}(A\bar{f}_\lambda) - \operatorname{err}(A\breve{f}_t) &= (\operatorname{err}(A\bar{f}_\lambda) - \widehat{\operatorname{err}}(A\bar{f}_\lambda)) + (\widehat{\operatorname{err}}(A\bar{f}_\lambda)) - \widehat{\operatorname{err}}(A\breve{f}_t)) \\ &\quad + (\widehat{\operatorname{err}}(A\breve{f}_t) - \operatorname{err}(A\breve{f}_t)) \leq 2\left(2\rho\widehat{\mathcal{R}}_S(\mathcal{H}_A) + 3\sqrt{\frac{\ln\left(\frac{8}{\delta}\right)}{2m}}\right).\end{aligned}$$

with probability at least $1-\delta$, by applying the usual Rademacher bounds (Bartlett & Mendelson, 2002), Theorem 8 to the class $\mathcal{H}_A$ twice. □

We should note that Theorem 3 and Remark 2.6 require that λ is specified before seeing the data. However, we can use SRM to allow an exploration of a countable number of different values for this parameter before making this choice for a small additional error term. Specifically, take a sequence of candidate values $\{\lambda_k\}_{k\in\mathbb{N}}$ weighted by $\{w_k\}_{k\in\mathbb{N}}$ with $\sum_{k\in\mathbb{N}} w_k \leq 1$. Then the same bounds hold for all λ_k, where $k \in \mathbb{N}$, simultaneously at the expense of an additional term of $3\sqrt{\frac{\log(1/w_k)}{2m}}$.

3 Rademacher complexity of the class of sensitivities

The generalisation bounds of Sect. 2 that include estimated values of the sensitivity, rely on the empirical Rademacher complexity of the class of sensitivities $\mathcal{D}_A\mathcal{H}$. In Theorem 1 this was estimated on a separate unlabelled set independent of the labelled sample, while in Theorems 2 and 3 it was estimated on the input points of the labelled training set, possibly augmented with further unlabelled data. To unify notations, in this section we will write S for a (generic) sample in both cases, and m for its cardinality – with a view that, if the empirical Rademacher complexity of $\mathcal{D}_A\mathcal{H}$ converges sufficiently fast with the cardinality of the labelled sample m, then the labelled data S may actually be sufficient.

However, arguably, the complexity of the sensitivity class, $\widehat{\mathcal{R}}_S(\mathcal{D}_A\mathcal{H})$, can be at least as large as that of the original function class $\mathcal{H}$ in the worst case, so one may wonder whether the bounds are actually useful. In this section we look at this quantity more closely. Indeed, using a property of the empirical Rademacher complexities (Bartlett & Mendelson, 2002), Theorem 12, part 7 gives

$$\widehat{\mathcal{R}}_S(\mathcal{D}_A\mathcal{H}) \leq \widehat{\mathcal{R}}_S(\mathcal{H}) + \widehat{\mathcal{R}}_S(\mathcal{H}_A). \tag{29}$$

Moreover, this bound is tight, since equality holds when the approximating class $\mathcal{H}_A$ is a singleton—however, the use of a singleton $\mathcal{H}_A$ is quite contrived, and far from what approximate algorithms are designed for.

For a fixed (possibly unlabelled) sample S, the set of interest in this section is the restriction of $\mathcal{D}_A\mathcal{H}$ to S,

$$\mathcal{D}_A\mathcal{H}|_S := \left\{ \begin{pmatrix} |f(x_1) - Af(x_1)| \\ \vdots \\ |f(x_m) - Af(x_m)| \end{pmatrix} : f \in \mathcal{H} \right\}$$

We use $R_p := \sup_{f\in\mathcal{H}} \widehat{\mathcal{D}}_A^p(f)$ for the worst sensitivity in the chosen p-norm on the sample S. Note that from Assumption 2 we have $R_p \leq C$ for all $p > 0$. We shall also use the shorthand

$$u_k = u(x_k) = |f(x_k) - Af(x_k)| \text{ and } u = (u_k)_{k\in[m]}.$$

Note that $\mathcal{D}_A\mathcal{H}|_S \subseteq B_p(0, m^{1/p}R_p)$ for all $p \geq 1$, where $B_p(c, r)$ denotes the p-ball centered at c with radius r.

We start by putting a crude magnitude bound on $\widehat{\mathcal{R}}_S(\mathcal{D}_A\mathcal{H})$, which holds irrespective of the choices of $\mathcal{H}$ and $\mathcal{H}_A$, and is tight up to a constant factor. The following proposition shows that, whenever R_p is small, the empirical Rademacher complexity of the sensitivity

class must be small in magnitude, and this bound is also tight up to a constant factor, for all choices of $p \geq 1$. This magnitude bound will not imply a decay as m increases, as we make no assumptions beyond an i.i.d. sample at this point. However this magnitude bound will be a useful reference in our later subsections, and it can also be taken in conjunction with other bounds, since one can always take the minimum of all upper bounds.

Proposition 3.1 (Crude magnitude bound) *For any $p \geq 1$, we have $\widehat{\mathcal{R}}_S(\mathcal{D}_A\mathcal{H}) \leq R_p$. Moreover, a lower bound of the same order holds as follows. Given p as chosen above, suppose that $\mathcal{D}_A\mathcal{H}|_S$ nearly fills the p-ball of radius $R_p m^{1/p}$, in the sense that the convex hull of $\mathcal{D}_A\mathcal{H}|_S$ contains the p-ball of radius $\frac{m^{1/p}}{2}R_p$ intersected with the positive orthant. Then there exists a constant $C_p > 0$ that only depends on the choice of the p-norm, such that $\widehat{\mathcal{R}}_S(\mathcal{D}_A\mathcal{H}) \geq C_p \cdot R_p$.*

Proof By Hölder's inequality,

$$\begin{aligned}
\widehat{\mathcal{R}}_S(\mathcal{D}_A\mathcal{H}) &= \frac{1}{m}\,\mathbb{E}_\sigma \sup_{f\in\mathcal{H}} \sum_{k=1}^{m} \sigma_k |f(x_k) - Af(x_k)| \\
&\leq \frac{1}{m} \sup_{f\in\mathcal{H}} \sum_{k=1}^{m} |f(x_k) - Af(x_k)| \\
&\leq \sup_{f\in\mathcal{H}} \left(\frac{1}{m} \sum_{k=1}^{m} |f(x_k) - Af(x_k)|^p \right)^{\frac{1}{p}}, \\
&= \sup_{f\in\mathcal{H}} \widehat{\mathcal{D}}_A^p(f) = R_p
\end{aligned}$$

for all $p \in [1, \infty)$. This proves the upper bound.

We denote by K_+ the positive orthant, and let $B_p^+\left(0, \frac{m^{1/p}}{2}R_p\right) := K_+ \cap B_p\left(0, \frac{m^{1/p}}{2}R_p\right)$. To prove the lower bound, we recall Moreau's decomposition theorem (Moreau, 1965) (see also (Wei et al., 2019), Sec. 2.1 & Sec. 3.1.5), which is the following: Given a closed convex cone $K \subset \mathbb{R}^m$, denote its polar cone by $K^* = \{u \in \mathbb{R}^m : \langle u, u'\rangle \leq 0 \text{ for all } u' \in K\}$. Then, every vector $v \in \mathbb{R}^m$ can be decomposed as

$$v = \Pi_K(v) + \Pi_{K^*}(v) \text{ such that } \langle \Pi_K(v), \Pi_{K^*}(v)\rangle = 0, \tag{30}$$

where $\Pi_K(u) := \operatorname{argmin}_{u'\in K} \|u - u'\|_2$ is the orthogonal projection of u into K. Hence we have

$$\widehat{\mathcal{R}}_S(\mathcal{D}_A\mathcal{H}) = \frac{1}{m}\,\mathbb{E}_\sigma \sup_{f\in\mathcal{H}} \sum_{k=1}^{m} \sigma_k |f(x_m) - Af(x_m)| \tag{31}$$

$$= \frac{1}{m}\,\mathbb{E}_\sigma \sup_{u\in \mathrm{conv}(\mathcal{D}_A\mathcal{H}|_S)} \sum_{k=1}^{m} \sigma_k u_k \tag{32}$$

$$\geq \frac{1}{m}\,\mathbb{E}_\sigma \sup_{u\in B_p^+\left(0, \frac{m^{1/p}}{2}R_p\right)} \sum_{k=1}^{m} \sigma_k u_k \tag{33}$$

$$= \frac{1}{m}\,\mathbb{E}_{\sigma} \sup_{u\in B_p^+\left(0,\frac{m^{1/p}}{2}R_p\right)} u^T(\Pi_{K_+}(\sigma)+\Pi_{K_+^*}(\sigma)) \tag{34}$$

$$= \frac{1}{m}\cdot\frac{m^{1/p}}{2}R_p\cdot\,\mathbb{E}_{\sigma}\|\Pi_{K_+}\sigma\|_{p'} \tag{35}$$

$$= \frac{1}{m}\cdot\frac{m^{1/p}}{2}R_p\cdot\left(\frac{m}{2}\right)^{1/p'} \tag{36}$$

$$= m^{1/p+1/p'-1}\cdot 2^{-1-1/p}\cdot R_p = \frac{R_p}{2\sqrt[p]{2}}. \tag{37}$$

where p' is the Hölder conjugate of p, i.e. $1/p+1/p'=1$. In line (34) we applied (30) to σ, and (35) follows from the fact that u is in the positive orthant K_+ so $\langle u,\Pi_{K_+^*}(\sigma)\rangle\le 0$ and because the supremum equality is attained when u is a nonnegative scalar multiple of $\Pi_{K_+}(\sigma)$ – in which case $\langle u,\Pi_{K_+^*(\sigma)}\rangle = 0$. This completes the proof of the lower bound.

□

The lower bound highlights the fact that one cannot tighten the complexity bound by more than a constant factor without making extra assumptions. In addition, we also see that non-negativity of the elements of $\mathcal{D}_A\mathcal{H}$ only affects this constant. Therefore in the next few sections we set out to find and exploit other structures in order to gain more transparency and insight on the effective magnitude of this quantity in some natural settings. Specifically, we shall discuss examples of some non-restrictive structural models from which one can read off benign conditions that give better bounds on $\widehat{\mathcal{R}}_S(\mathcal{D}_A\mathcal{H})$. A lower magnitude of this complexity implies a smaller unlabelled data set size requirement for accurate estimation of the sensitivity, and in the case of our bounds in Sects. 2.3 and 2.4 this may even permit solving the learning problem without the need of an additional unlabelled sample.

3.1 Exploiting structural models of the sensitivity set

Throughout this section we make no assumption about either the function class $\mathcal{H}$ or the approximating class $\mathcal{H}_A$. So the results of this section are equally relevant to very rich classes like deep neural networks, all the way to very restricted ones like linear classes. We also make no assumption about the form of the approximating function, and indeed the approximating class is not required to be of the same architectural type as the original class.

We demonstrate the benign effects of some structural traits that the set $\mathcal{D}_A\mathcal{H}$ may naturally exhibit regardless of the linear or nonlinear nature of the actual predictors. Such benign structures will manifest themselves by explaining a reduced complexity $\widehat{\mathcal{R}}_S(\mathcal{D}_A\mathcal{H})$ —which in turn allow the bounds of Sect. 2 to provide a better understanding of what makes some instances of a learning problem easier than others.

Our strategy in the next subsections will be to study the complexity of the set $\mathcal{D}_A\mathcal{H}$ restricted to the sample S (as it appears in the empirical Rademacher bounds presented in Sect. 2) by inscribing it into various parametrised geometric shapes. These include natural structures such as the points of $\mathcal{D}_A\mathcal{H}_{|S}$ being near-sparse, or exhibiting clusters,

or having some structured sparsity type model. For this we will not actually impose any extra conditions, instead our strategy is to use these constructs to reveal how the Rademacher complexity depends on the parameters of these models. In other words, our bounds will always hold with some parameter values, as in the worst case we just recover the crude magnitude bound in Proposition 3.1, while at the same time the effects of parameters convey more insight.

3.1.1 Near-sparse sensitivity set

A very natural situation is when some points in S have little effect on the sensitivity of the approximation, or in other words the approximation has little effect on the predictions for part of the points of S. For instance in classification, points that are far from the boundary will often have the approximating function Af predict in agreement with the original f.

A simple way to model this situation is by having the vectors in $\mathcal{D}_A\mathcal{H}|_S$ lie near the axes corresponding to the points in S which are less affected by the approximation, such as taking a shape of an axis-aligned ellipsoid in some Minkowski norm, defined as

$$\mathcal{E}_p(\mu) := \left\{ x \in \mathbb{R}^m \,:\, \sum_{k=1}^{m} \frac{|u_k|^p}{\mu_k^p} \leq 1 \right\}, \tag{38}$$

for $p \geq 1$, where $\mu := (\mu_1, \ldots, \mu_m) \in (0, \infty)^m$ are the semi-axes of the ellipsoid.

Note, this model is not restrictive, since we have $\mathcal{D}_A\mathcal{H}|_S \subset B_p^m(0, R_p m^{1/p})$, therefore $\mu_k \leq R_p m^{1/p}$ for all $k \in [m]$. However, the added flexibility of this model allows us to infer the effect of the magnitudes of the semi-axes, yielding some simple and natural conditions that improve on the worst-case magnitude guarantee in Proposition 3.1.

The following lemma gives the exact expression for the Rademacher complexity of an ellipsoid in any p-norm.

Lemma 3.2 *Let $\mu \in (0, \infty)^m$ and $p \geq 1$, and consider $\mathcal{E}_p(\mu)$ as defined in (38). Then,*

$$\widehat{\mathcal{R}}_S(\mathcal{E}_p(\mu)) = \frac{\|\mu\|_{\frac{p}{p-1}}}{m}.$$

Proof Using Hölder's inequality, $\sigma_k \in \{-1, 1\}$, and the definition of $\mathcal{E}_p(\mu)$ we have

$$\widehat{\mathcal{R}}_S(\mathcal{E}_p(\mu)) = \frac{1}{m} \mathbb{E}_\sigma \sup_{u \in \mathcal{E}_p(\mu)} \sum_{k=1}^{m} \sigma_k u_k \tag{39}$$

$$= \frac{1}{m} \mathbb{E}_\sigma \sup_{u \in \mathcal{E}_p(\mu)} \sum_{k=1}^{m} (\sigma_k \mu_k) \frac{u_k}{\mu_k} \tag{40}$$

$$= \frac{1}{m} \mathbb{E}_\sigma \sup_{u \in \mathcal{E}_p(\mu)} \left(\sum_{k=1}^{m} |\sigma_k \mu_k|^{p'} \right)^{\frac{1}{p'}} \left(\sum_{k=1}^{m} \frac{|u_k|^p}{\mu_k^p} \right)^{\frac{1}{p}} \tag{41}$$

$$= \frac{1}{m}\left(\sum_{k=1}^{m} \mu_k^{p'}\right)^{\frac{1}{p'}}, \tag{42}$$

where p' is the Hölder conjugate of p, i.e. $1/p + 1/p' = 1$. The identities (41) and (42) hold due to the supremum. This completes the proof. □

For more intuition, consider the case when $p = 2$, which corresponds to the usual Euclidean norm ellipsoid, and we can relate the right hand side of the bound in Lemma 3.2 to the volume of the ellipsoid. Indeed, using the relation between the arithmetic and geometric mean,

$$\frac{1}{m}\|\mu\|_2 = \frac{1}{\sqrt{m}}\left(\frac{1}{m}\sum_{k=1}^{m}\mu_k^2\right)^{\frac{1}{2}} \geq \frac{1}{\sqrt{m}}\left(\prod_{k=1}^{m}\mu_k\right)^{\frac{1}{m}} = C_m \mathrm{Vol}(\mathcal{E}_2(\mu))^{\frac{1}{m}},$$

where $C_m > 0$ is a constant depending only on m. Hence, for a fixed sample size m, if the quadratic mean of the μ_k's is small then the ellipsoid has a small volume.

If $\mathcal{D}_A\mathcal{H}|_S \subseteq \mathcal{E}_p(\mu)$, then in the worst case $\mu_k = R_p m^{1/p}$ for all $k \in [m]$, and so

$$\frac{\|\mu\|_{\frac{p}{p-1}}}{m} \leq \frac{1}{m}\left(\sum_{k\in[m]}(R_p m^{1/p})^{\frac{p}{p-1}}\right)^{\frac{p-1}{p}} = R_p.$$

Hence it is clear that the bound in Lemma 3.2 recovers the bound in Proposition 3.1 in the worst case. Thus, if $\mathcal{D}_A\mathcal{H}|_S \subseteq \mathcal{E}_p(\mu)$, then Lemma 3.2 is already an improvement on Proposition 3.1.

As a model of the sensitivity set, an ellipsoid with high excentricity posits that most sensitivity vectors reside in a linear subspace of $\mathbb{R}^m$. Interesting to note that this has no implication on the form of the predictors. Indeed, even with highly nonlinear predictors (nonlinear classification boundaries for example), the fraction of points for which the predictions are distorted under the action of approximation may be expected to be small.

However, it might be unrealistic to expect of all good functions of $\mathcal{H}$ that the approximation should change the prediction for the same points and should leave alone the same points. Hence, instead of assuming that $\mathcal{D}_A\mathcal{H}$ is contained in a single ellipsoid, for a more realistic model, we consider a union of multiple axes-aligned ellipsoids that cover $\mathcal{D}_A\mathcal{H}|_S$. This allows the set of points for which predictions are relatively unaffected by the approximation of some $f \in \mathcal{H}$ be different for all $f \in \mathcal{H}$.

The following proposition shows that in this model the Rademacher complexity of $\mathcal{D}_A\mathcal{H}|_S$ is bounded by the Rademacher complexity of the largest ellipsoid from the union and, remarkably, it does not depend on the number of ellipsoids in the union – we can have countably many in this model, so the diversity of sensitivity profiles of the predictors of $\mathcal{H}$ in the span of the sample is accounted for at no expense. The vector of axis lengths for the i-th ellipsoid will be denoted by μ_i. We refer to individual components of this vector by adding a second index, for example $\mu_{i,k}$ for the kth semi-axis of the ith ellipsoid.

Proposition 3.3 (Complexity of near-sparse sensitivity set) *Let $S \subseteq \mathcal{X}$ be an i.i.d. unlabeled sample drawn from D_x, of size m. Let $l \in \mathbb{N}$, suppose that there exist $\mu_i \in (0,\infty)^m, i \in [l]$ with $\mu_{i,k} \leq R_p m^{1/p}$, and $\mathcal{D}_A\mathcal{H}|_S \subset \bigcup_{i=1}^{l}\mathcal{E}_p(\mu_i)$ for ellipsoids $\mathcal{E}_p(\mu_i)$. Then we have the following bound*

$$\widehat{\mathcal{R}}_S(\mathcal{D}_A\mathcal{H}) \leq \frac{1}{m}\max_i \|\mu_i\|_{\frac{p}{p-1}}.$$

The proof makes use of similar steps as the proof of Lemma 3.2, but it does not apply the result of Lemma 3.2, as it turns out that a direct approach yields the exact Rademacher complexity of the union of axis-aligned ellipsoids.

Proof of Proposition 3.3 As $\mathcal{D}_A\mathcal{H}|_S \subset \bigcup_{i=1}^{l} \mathcal{E}_p(\mu_i)$, then using the fact that for two bounded sets A and B we have $\sup(A \cup B) = \max\{\sup A, \sup B\}$, taking absolute value, the Hölder inequality, $\sigma_k \in \{-1, 1\}$, and the definition of $\mathcal{E}_p(\mu_i)$ give

$$\widehat{\mathcal{R}}_S(\mathcal{D}_A\mathcal{H}) \leq \widehat{\mathcal{R}}_S\left(\bigcup_{i=1}^{l} \mathcal{E}_p(\mu_i)\right) \tag{43}$$

$$= \frac{1}{m}\,\mathbb{E}_\sigma \sup_{u\in\bigcup_{i=1}^{l}\mathcal{E}_p(\mu_i)} \sum_{k=1}^{m} \sigma_k u_k \tag{44}$$

$$= \frac{1}{m}\,\mathbb{E}_\sigma \max_i \sup_{u\in\mathcal{E}_p(\mu_i)} \sum_{k=1}^{m} \sigma_k u_k \tag{45}$$

$$= \frac{1}{m}\max_i \sup_{u\in\mathcal{E}_p(\mu_i)} \sum_{k=1}^{m} |u_k| \tag{46}$$

$$= \frac{1}{m}\max_i \sup_{u\in\mathcal{E}_p(\mu_i)} \sum_{k=1}^{m} \mu_{i,k}\frac{|u_k|}{\mu_{i,k}} \tag{47}$$

$$= \frac{1}{m}\max_i \sup_{u\in\mathcal{E}_p(\mu_i)} \left(\sum_{k=1}^{m} \mu_{i,k}^{p'}\right)^{\frac{1}{p'}} \left(\sum_{k=1}^{m} \frac{|u_k|^p}{\mu_{i,k}^{p}}\right)^{\frac{1}{p}} \tag{48}$$

$$= \frac{1}{m}\max_i \left(\sum_{k=1}^{m} \mu_{i,k}^{p'}\right)^{\frac{1}{p'}}, \tag{49}$$

where p' is the Hölder conjugate of p. The equality in (46) is due to the symmetry of the set $\bigcup_{i=1}^{l} \mathcal{E}_p(\mu_i)$ around each axis, and in (48) Hölder's inequality holds with equality due to the supremum.

We remark that the above proposition is true for a countably infinite number of ellipsoids by noticing that the sequence

$$\left(\sup_{u\in\bigcup_{i=1}^{l}\mathcal{E}_p(\mu_i)} \sum_{k=1}^{m} \sigma_k u_k\right)_{l\in\mathbb{N}}$$

is non-decreasing in l. Thus, by the monotone convergence theorem we have

$$\begin{aligned}\widehat{\mathcal{R}}_S\left(\bigcup_{i=1}^{\infty}\mathcal{E}_p(\mu_i)\right) &= \lim_{l\to\infty}\widehat{\mathcal{R}}_S\left(\bigcup_{i=1}^{l}\mathcal{E}_p(\mu_i)\right)\\ &= \lim_{l\to\infty}\frac{1}{m}\max_{i\in[l]}\left(\sum_{k=1}^{m}\mu_{i,k}^{p'}\right)^{\frac{1}{p'}}\\ &= \frac{1}{m}\sup_{i\in\mathbb{N}}\left(\sum_{k=1}^{m}\mu_{i,k}^{p'}\right)^{\frac{1}{p'}}.\end{aligned}$$

Thus Proposition 3.3 is true for countably infinite ellipsoids.

It may be interesting to note that the model of a union of axis-aligned ellipsoids has an intuitive meaning of near-sparsity of sensitivities. This may also be interpreted as a kind-of near-compression bound, since Proposition 3.3 tells us that, when fewer points are affected by the approximation, the guarantee on the sensitivity estimation quality will be tighter, hence the generalisation bound will be tighter as well.

However, beyond the intuitive meaning above, our structural modelling approach has potential to reveal additional benign conditions that might be harder to find by intuition alone. To see this, we shall modify Proposition 3.3 to get an upper bound for a non-axis aligned union of ellipsoids. As long as the ellipsoids share the same center (for instance, at the origin), the upper bound will still be independent of the number of ellipsoids in the union.

To this end, in addition to the axis-length parameters, for each ellipsoid in the union, take a rotation matrix $V_i \in \mathbb{R}^{m\times m}$ where $V_i^T V_i = V_i V_i^T = I_m$ for $i \in [l]$. The columns of V_i are the principal directions for the i-th ellipsoid. We will refer to the k-th column of V_i by $(V_i)_k$, and $(V_i)_{k,k'}$ will denote its (k,k')-th element. The i-th ellipsoid is then defined as

$$\mathcal{E}_p^{V_i}(\mu_i) := \left\{ u \in \mathbb{R}^m \,:\, \sum_{k=1}^{m}\frac{|(V_i)_k^T u|^p}{\mu_{i,k}^p} \le 1 \right\}. \tag{50}$$

By a change of variables, we have that $u \in \mathcal{E}_p^{V_i}(\mu_i)$ is equivalent to $V_i^T u \in \mathcal{E}_p(\mu_i)$. Let Λ_i be the diagonal matrix with elements $\mu_{i,k} \in (0,\infty)$ for $k \in [m]$, so $\Lambda_i^{-1} V_i^T u \in B_p(0,1)$.

We no longer have symmetry around the axes, so (46) becomes an inequality, and we have

$$\widehat{\mathcal{R}}_S\Big(\bigcup_{i=1}^{l}\mathcal{E}_p^{V_i}(\mu_i)\Big) = \frac{1}{m}\,\mathbb{E}_\sigma \max_{i\in[l]}\sup_{u\in\mathcal{E}_p^{V_i}(\mu_i)}\sum_{k=1}^{m}\sigma_k u_k \tag{51}$$

$$\le \frac{1}{m}\max_{i}\sup_{u\in\mathcal{E}_p^{V_i}(\mu_i)}\sum_{k=1}^{m}|u_k| \tag{52}$$

$$= \frac{1}{m}\max_{i}\sup_{u\in\mathcal{E}_p^{V_i}(\mu_i)}\|u\|_1 \tag{53}$$

$$= \frac{1}{m}\max_{i}\sup_{u\in\mathcal{E}_p^{V_i}(\mu_i)}\|(V_i\Lambda_i)(\Lambda_i^{-1}V_i^T u)\|_1 \tag{54}$$

$$= \frac{1}{m} \max_i \sup_{\Lambda_i^{-1} V_i^T u \in B_p(0,1)} \|(V_i \Lambda_i)(\Lambda_i^{-1} V_i^T u)\|_1 \tag{55}$$

$$= \frac{1}{m} \max_i \|V_i \Lambda_i\|_{p \to 1}. \tag{56}$$

Equation (55) used the assumption that Λ_i and V_i are full rank square matrices. The last line (56) holds by the definition of $\|\cdot\|_{p\to 1}$, called the operator norm (or induced matrix norm) with domain p and co-domain 1. Such norms can only be computed explicitly in a few special cases. In particular,

1. Whenever $V_i = I_m$, then $\|V_i \Lambda_i\|_{p\to 1} = \|\mu_i\|_{\frac{p}{p-1}}$, since Λ_i is the diagonal matrix with elements $\mu_{i,k}$. This recovers precisely the axis-aligned setting.
2. With $p = 1$, the expression of the induced norm is known to be $\|V_i \Lambda_i\|_{1\to 1} = \max_{k\in[m]} \mu_{i,k} \|(V_i)_k\|_1$.

We see, the non-axis alignment has led to somewhat less intuitive expressions, but nevertheless the main quantity that governs the empirical Rademacher complexity remains some notion of size of the largest ellipsoid. To interpret this in the context of interest here, it is enough if the sensitivities mainly reside in linear subspaces of $\mathcal{D}_A \mathcal{H}|_S \subset \mathbb{R}^m$ for the Rademacher complexity of $\mathcal{D}_A \mathcal{H}$ to be small. Equivalently, for the estimation of sensitivities this means to require less unlabelled points and still getting accurate sensitivity estimates (not to be confused with small sensitivity values).

3.1.2 Clustered sensitivity set

In this section we consider another natural structure, namely when the elements of $\mathcal{D}_A \mathcal{H}|_S$ form clusters. A cluster is a subset of $\mathcal{H}$ with similar sensitivity profile on the sample S. We can model each cluster with a p-norm ellipsoid, each having its own center as the following

$$\mathcal{E}_p(c_i, \mu_i, V_i) := \left\{ u \in \mathbb{R}^m : \sum_{k=1}^m \frac{|(V_i)_k^T (u - c_i)|^p}{\mu_{i,k}^p} \le 1 \right\}.$$

The components of the vector μ_i are the semi-axes, and the vector c_i is the center of the i-th cluster. This model is again non-restrictive, as there exist worst case parameter values ($c_i = 0$, $\mu_i = R_p m^{1/p}$, $V_i = I_m$ for all $i \in [l]$) that recover the ball $B_p(0, R_p m^{1/p})$ used previously in the crude bound of Proposition 3.1.

The following proposition shows that in this model, $\widehat{\mathcal{R}}_S(\mathcal{D}_A \mathcal{H})$ is bounded by the Rademacher complexity of the largest cluster plus an additive term that grows logarithmically with the number of clusters and linearly with the largest displacement of a cluster from the origin.

Proposition 3.4 (Complexity of clustered sensitivity set) *Let $S \subset \mathcal{X}$ be an unlabeled sample of size m drawn i.i.d. from D_x. Let $l \in \mathbb{N}$, suppose that there exist $\mu_i \in (0, \infty)^m, c_i \in \mathbb{R}^m$ and $V_i \in \mathbb{R}^{m\times m}$ such that $\mu_{i,k} \le R_p m^{1/p}$, and $V_i^T V_i = V_i V_i^T = I_m$ with $\mathcal{D}_A \mathcal{H}|_S \subseteq \bigcup_{i=1}^l \mathcal{E}_p(c_i, \mu_i, V_i)$ for p-ellipsoids. Then,*

$$\widehat{\mathcal{R}}_S(\mathcal{D}_A\mathcal{H}) \leq \frac{1}{m}\max_i \|V_i\Lambda_i\|_{p\to 1} + \max_i\{\|c_i\|_2\}\frac{\sqrt{2\ln l}}{m}.$$

where Λ_i is the diagonal matrix with elements $\mu_{i,k} \in (0,\infty)$ for $k \in [m]$.

This cluster model highlights a trade-off about the effect of large sensitivities: If a cluster only contains functions whose approximation leads to large sensitivity values, then the first term of the bound can still be small, but a penalty is incurred in the second term if not all function fit in the same cluster.

Proof Let $c : \mathcal{D}_A\mathcal{H} \to \{c_1,\ldots,c_l\}$ be defined as the function that sends $u \in \mathcal{D}_A\mathcal{H}$ to its best fitting ellipsoid, $c(u) := \operatorname{argmin}_{c_i : i\in[l]} \sum_{k\in[m]} (V_i)_k^T(u-c_i)/\mu_k$. Ties are broken arbitrarily.

Now, adding and subtracting $c(u_k)$ and noting that, by construction, $\{c(u) : u \in \mathcal{D}_A\mathcal{H}\} = \{c_1,\ldots,c_l\}$, we have

$$\widehat{\mathcal{R}}_S(\mathcal{D}_A\mathcal{H}) \leq \widehat{\mathcal{R}}_S\left(\bigcup_{i=1}^{l}\mathcal{E}_p(c_i,\mu_i,V_i)\right) \tag{57}$$

$$= \frac{1}{m}\mathbb{E}_\sigma \sup_{u\in\bigcup_{i=1}^{l}\mathcal{E}_p(c_i,\mu_i,V_i)} \sum_{k=1}^{m}\sigma_k u_k \tag{58}$$

$$= \frac{1}{m}\mathbb{E}_\sigma \max_{i\in[l]} \sup_{u\in\mathcal{E}_p(c_i,\mu_i,V_i)} \sum_{k=1}^{m}\sigma_k u_k \tag{59}$$

$$\begin{aligned}\leq \frac{1}{m}\mathbb{E}_\sigma \max_{i\in[l]} \sup_{u\in\mathcal{E}_p(c_i,\mu_i,V_i)} \sum_{k=1}^{m}\sigma_k (V_i)_k^T(u-c_i)\\ + \frac{1}{m}\mathbb{E}_\sigma \max_{i\in[l]} \sup_{u\in\mathcal{E}_p(c_i,\mu_i,V_i)} \sum_{k=1}^{m}\sigma_k (V_i)_k^T c_i\end{aligned} \tag{60}$$

$$\begin{aligned}= \frac{1}{m}\mathbb{E}_\sigma \max_{i\in[l]} \sup_{V_i^T(u-c_i)\in\mathcal{E}_p(0,\mu_i)} \sum_{k=1}^{m}\sigma_k (V_i)_k^T(u-c_i)\\ + \frac{1}{m}\mathbb{E}_\sigma \max_{i\in[l]} \sum_{k=1}^{m}\sigma_k (V_i)_k^T c_i.\end{aligned} \tag{61}$$

We proceed by bounding the above two terms separately.

We bound the first term by applying Proposition 3.3, or its extension, Eq. (56). To bound the second term, we use Massart's lemma to get

$$\frac{1}{m}\mathbb{E}_\sigma \max_{i\in[l]} \sum_{k=1}^{m}\sigma_k (V_i)_k^T c_i \leq \max_{i\in[l]}\{\|c_i\|_2\}\frac{\sqrt{2\ln l}}{m},$$

since V_i is a rotation matrix, so $\|V_i^T c_i\|_2 = \|c_i\|$. Combining the two bounds together completes the proof.

This bound is similar in flavour to that of the complexity of a union given in Golowich et al. (2020), Lemma 7.4 in the sense that there is a logarithmic price to pay for the number of clusters. However, by contrast, here we have an explicit constant in the second term with clear relation to the position of the ellipsoids, and our bound reduces to that from Proposition 3.3 if all $c_i = 0$ for all $i \in [l]$. Therefore, the above bound gives more information as to what helps decrease the Rademacher complexity. More specifically, the benign structures identified are: small number of clusters, cluster centers close to the origin, and highly concentrated (low volume) clusters. We will summarise these positive findings and discuss their implications in Sect. 4.1.

3.2 Effect of the structural form of predictors

Our analysis so far was completely independent of the specification of $\mathcal{H}$ and $\mathcal{H}_A$, and applies to any PAC-learnable hypothesis class. From the crude bound in (29) we know that a low complexity $\mathcal{H}$ always implies a low complexity $\mathcal{D}_A\mathcal{H}$. In this section we give a worked example of how this effect plays out in the case of hypothesis classes that are linear in the parameters. Linear models represent a well-weathered object of study at the foundation of machine prediction (Vapnik, 1998), whose high-dimensional / low sample size version has been of much interest for the puzzle of over-parameterisation, see e.g. Bartlett et al. (2020). These models also allow for nonlinearity effortlessly through a feature map or a kernel.

Let $\mathbb{H}$ be a reproducing kernel Hilbert space with reproducing kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ associated with the feature map $\Phi : \mathcal{X} \to \mathbb{H}$, so for any $x_1, x_2 \in \mathcal{X}$, we have $k(x_1, x_2) = \langle \Phi(x_1), \Phi(x_2) \rangle_{\mathbb{H}}$. Then our hypothesis class is

$$\mathcal{H} := \{x \mapsto \langle w, \Phi(x) \rangle_{\mathbb{H}} : w \in \mathbb{H}\}.$$

The familiar Euclidean space setting corresponds to Φ being the identity map and $\mathbb{H} = \mathbb{R}^d$.

We define our approximation operator to be $A : \mathcal{H} \to \mathcal{H}_A$ defined by $Af_w(x) = \langle Q(w), \Phi(x) \rangle_{\mathbb{H}}$ where $f_w(x) = \langle w, \Phi(x) \rangle_{\mathbb{H}}$ and $Q : \mathbb{H} \to \mathbb{H}$ is some approximation of the weights w of the predictor f_w.

Proposition 3.5 *Let $m \in \mathbb{N}$ and $S = \{x_1, \ldots, x_m\} \subset \mathcal{X}$. Then we have the following bound*

$$\widehat{\mathcal{R}}_S(\mathcal{D}_A\mathcal{H}) \leq \frac{\sup_{w \in \mathbb{H}} \|w - Q(w)\|_{\mathbb{H}}}{\sqrt{m}} \sqrt{\sum_{k=1}^{m} k(x_k, x_k)}. \tag{62}$$

This is of course upper bounded by the sum of familiar bounds for linear classes $\mathcal{H}$ and $\mathcal{H}_A$ by the triangle inequality, as already implied indeed by the crude bound (29); however, the important observation from the special-case analysis of Proposition 3.5 is that (62) does not explicitly depend on the norm of the weight vectors, but instead it only depends on how the approximation A (through Q) distorts the weights. In other words, we do not need the norms $\|w\|_{\mathbb{H}}$ for $\widehat{\mathcal{R}}_S(\mathcal{D}_A\mathcal{H})$ to be bounded as long as the weight sensitivity $\|w - Q(w)\|_{\mathbb{H}}$ is bounded for the chosen operator Q.

Therefore the finding we conclude from Proposition 3.5 is that, in the generalised-linear model class considered, small weight-sensitivity is sufficient for dimension-independent learning when the approximating class $\mathcal{H}_A$ has dimension-free complexity. This is in contrast with existing dimension-free bounds that required a bounded norm constraint.

We have not found an analogous property for other hypothesis classes, and it remains an open question as to whether analyses of the sensitivity class tailored to specific classes would unearth additional insights.

Proof of Proposition 3.5 Since σ_k are uniform on $\{-1, 1\}$, we can remove the absolute value, and by the linearity of inner products, and the Cauchy-Schwarz inequality we have

$$\begin{aligned}
\widehat{\mathcal{R}}_S(\mathcal{D}_A\mathcal{H}) &= \frac{1}{m}\,\mathbb{E}_\sigma \sup_{f\in\mathcal{H}} \sum_{k=1}^{m} \sigma_k |f(x_k) - Af(x_k)| \\
&= \frac{1}{m}\,\mathbb{E}_\sigma \sup_{w\in\mathbb{H}} \sum_{k=1}^{m} \sigma_k (\langle w, \Phi(x_k)\rangle_{\mathbb{H}} - \langle Q(w), \Phi(x_k)\rangle_{\mathbb{H}}) \\
&= \frac{1}{m}\,\mathbb{E}_\sigma \sup_{w\in\mathbb{H}} \langle w - Q(w), \sum_{k=1}^{m} \sigma_k \Phi(x_k)\rangle_{\mathbb{H}} \\
&\leq \frac{1}{m} \sup_{w\in\mathbb{H}} \|w - Q(w)\|_{\mathbb{H}}\, \mathbb{E}_\sigma \left\| \sum_{k=1}^{m} \sigma_k \Phi(x_k) \right\|_{\mathbb{H}}.
\end{aligned}$$

Finally it is know that (Mohri et al., 2018), Theorem 6.12

$$\mathbb{E}_\sigma \left\| \sum_{k=1}^{m} \sigma_k \Phi(x_k) \right\|_{\mathbb{H}} \leq \left[\sum_{k=1}^{m} k(x_k, x_k) \right]^{\frac{1}{2}}.$$

This completes the proof.

4 Discussion of implications, and potential extensions

In this section we elaborate on the significance of our theoretical results. The next Sect. 4.1 shows how to use our analysis of the sensitivity set to obtain a natural and very general structural condition that yields favourable convergence rates on the sensitivity estimation error. Hence, under this condition, the generalisation error bounds in our previous sections become dominated by the complexity of the reduced approximate class, irrespective of the form or size of the original class.

In Sect. 4.2 we discuss consequences related to real problems by revisiting the original motivation of understanding model compression in deep networks. In particular, we consider a concrete case of approximation by weight-binarisation, as in BinaryConnect (Courbariaux et al., 2015), or parameter quantisation in deep network classifiers (Hubara et al., 2017), where applying our results yields a depth-independent bound. We also discuss a potential way to relate our approach to a previously successful but theoretically unjustified on-device deep net approach, Neural Projections (Ravi, 2019), which brings insights into its working.

Finally, in Sect. 4.3 we describe how our framework can be extended to stochastic approximation schemes.

4.1 Favourable rates for sensitivity estimation in approximable hypothesis classes

We already commented that whenever the target function admits a small sensitivity threshold t, this can usefully restrict the hypothesis class in favourable data distributions. Here we show that a uniformly small t, with approximation sensitivity specified in the $p = 2$ norm, can even obtain a speed-up of the convergence rate of sensitivity estimation, based on the findings of Sect. 3.

First, we extract the fortuitous conditions that arose from our analysis in Sect. 3 that enable a fast convergence of the Rademacher complexity of the class of sensitivities. More precisely, if the sample sensitivity set $\mathcal{D}_A\mathcal{H}_{|S}$ resides in a countable union of near-sparse sets and a finite number $l \in \mathbb{N}$ of dense clusters, then the Rademacher complexity of the sensitivity set decays at a fast rate $1/m$, up to a logarithmic factor.

Condition 4.1 (*Structured sensitivity condition*) Suppose that $\mathcal{D}_A\mathcal{H}_{|S} \subset \left(\bigcup_{i=1}^{\infty} \mathcal{E}_p(\mu_i)\right) \bigcup \left(\bigcup_{i=1}^{l} \mathcal{E}_p(\tilde{c}_i, \tilde{\mu}_i, V_i)\right)$, where $\mathcal{E}_p(\tilde{c}_i, \tilde{\mu}_i, V_i)$ are ellipsoids centered at $\tilde{c}_i$ having side-lengths concatenated in the vector $\tilde{\mu}_i$ and orientation V_i; and $\mathcal{E}_p(\mu_i), i \geq 1$, are ellipsoids centered at the origin, having side-lengths concatenated in μ_i. Let Λ_i be a diagonal matrix with elements $\tilde{\mu}_{i,k} \in (0, \infty)$ for $k \in [m]$. If there are constants $\kappa, \kappa_1, \kappa_2 \geq 0$, independent of m, such that $\max_{i \geq 1} \|\mu_i\|_2 \leq \kappa$, $\max_i \|V_i\Lambda_i\|_{2\to 1} \leq \kappa_1$ and $\max_i\{\|\tilde{c}_i\|_2\} \leq \kappa_2$, we say that the sensitivity set $\mathcal{D}_A\mathcal{H}_{|S}$ is structured, with parameters κ, κ_1 and κ_2.

Lemma 4.2 *If $\mathcal{D}_A\mathcal{H}_{|S}$ satisfies Condition 4.1 with parameters κ, κ_1, and κ_2, then we have*

$$\widehat{\mathcal{R}}_S(\mathcal{D}_A\mathcal{H}) \leq \frac{\kappa + \kappa_1 + \kappa_2\sqrt{\log l}}{m} = \mathcal{O}\left(\sqrt{\log(l)}/m\right). \tag{63}$$

Proof In the near-sparse subset of $\mathcal{D}_A\mathcal{H}_{|S}$ that resides in $\bigcup_{i=1} \mathcal{E}_p(\mu_i)$, we have by Proposition 3.3 that $\widehat{\mathcal{R}}_S(\mathcal{D}_A\mathcal{H}) \leq \kappa/m$.

In the remaining subset that resides in the elliptic clusters $\bigcup_{i=1}^{l} \mathcal{E}_p(\tilde{c}_i, \tilde{\mu}_i, V_i)$, we have by Proposition 3.4 that $\widehat{\mathcal{R}}_S(\mathcal{D}_A\mathcal{H}) \leq (\kappa_1 + \kappa_2\sqrt{\log l})/m$. The union has complexity no larger than the sum of complexities of its constituent subsets. □

This implies the following for the estimation of sensitivities.

Theorem 4 (Sensitivity estimation bound for uniformly approximable classes) *Let $S \in \mathcal{X}^m$ be a sample drawn i.i.d. from the marginal distribution D_x of size m. Suppose there exists $t \geq 0$ such that $\mathcal{D}_A^2(f) \leq t$ for all $f \in \mathcal{H}$. Then, with probability at least $1 - \delta$, we have*

$$\sup_{f\in\mathcal{H}} |\mathcal{D}_A^1(f) - \widehat{\mathcal{D}}_A^1(f)| \leq 6\widehat{\mathcal{R}}_S(\mathcal{D}_A\mathcal{H}) + t\sqrt{\frac{2\ln(\frac{1}{\delta})}{m}} + \frac{6C\ln(\frac{1}{\delta})}{m}. \tag{64}$$

Furthermore, if Condition 4.1 holds, then $\widehat{\mathcal{R}}_S(\mathcal{D}_A\mathcal{H}) = \tilde{\mathcal{O}}(1/m)$.

Proof First note that from Assumption 2 we have $\|f - Af\|_\infty < C$ and we have the following bound on the variance of the function $f - Af$,

$$\begin{aligned}\mathrm{Var}_X[f(X) - Af(X)] &= \mathbb{E}_X[(f(X) - Af(X))^2] - \mathbb{E}_X[f(X) - Af(X))]^2 \\ &\leq \mathbb{E}_X[(f(X) - Af(X))^2] \\ &\leq (\mathcal{D}_A^2(f))^2 = t^2,\end{aligned}$$

where the last line is due to Jensen's inequality. The result then follows from Bartlett et al. (2005), Theorem 2.1 by setting $\alpha = \frac{1}{2}$. The second statement is proved in Lemma 4.2.

□

Theorem 4 bounds the deviation between the true sensitivity and its sample estimate in terms of the global sensitivity threshold t of functions in $\mathcal{H}$. Whenever t is sufficiently small, then the last term will dominate the t-dependent term, which in turn decays with m at a faster rate.

The observation that the sensitivity threshold t acts as a variance to control the rate could also be further refined using localisation to replace the global sensitivity threshold with the sensitivities of individual functions and relax the requirement that the entire class $\mathcal{H}$ is well approximable, at the expense of a more involved machinery of local Rademacher complexities (Bartlett et al., 2005), which we do not pursue here, and which would likely need a specialised treatment to bound the local complexity for particular choices of $\mathcal{H}$, similarly to the approach taken in Suzuki et al. (2020a).

The key difference from the approach of Suzuki et al. (2020a) is the following. Their bounds depend on the local Rademacher complexity of the Minkowski difference between the loss classes of the full and the approximate predictors, which they are able to bound for some specific hypothesis classes; whereas, our bounds depend on the Rademacher complexity of the *set of sensitivities* of predictors from the hypothesis class. The Minkowski difference loses the coupling between the full and approximate predictor pairs which, in our approach is the key to taking advantage of structure in the set of sensitivities. These structures that we identified and exploited are not specific to the form of functions in the hypothesis class chosen, and instead uncover new general insights, as well as tighten our bounds effortlessly, with elementary tools.

Indeed, we highlighted that even in the simple global analysis of Theorem 4, from the findings of Sect. 3 we were able to readily extract some general favourable rate conditions for sensitivity estimation. Note that Lemma 4.2 is general, and holds for any PAC-learnable class. It says that whenever the target function admits a small t and the interplay of data and model satisfies Condition 4.1, the error from sensitivity estimation becomes negligible very quickly (even without any additional unlabelled data), hence the dominant term of our generalisation bounds (Theorems 1, 2, 3) now becomes the complexity of the approximate class, regardless of how big the original class $\mathcal{H}$ was.

4.2 Implications related to real problems

In this section we discuss the significance of our theoretical results by revisiting some of our motivating examples related to real problems.

4.2.1 From BinaryConnect to a depth-independent bound

We consider a specific example. Take $\mathcal{H}$ to be the class of L-layer feed-forward neural network classifiers with ReLu activations in the hidden layers and binary output. Let $|W|$

be the total number of parameters (including all weights and bias terms). It was shown in Bartlett et al. (2019) that the VC dimension of this class is $\mathcal{O}(|W|L\log(|W|))$, and this is near-tight with a lower bound of $\Omega(|W|L\log(|W|/L))$. A well-known relation between Rademacher complexity and VC dimension (Bartlett & Mendelson, 2002), Theorem 6 implies that for this class we have

$$\widehat{\mathcal{R}}_S(\mathcal{H}) = \mathcal{O}\left(\sqrt{\frac{|W|L\log(|W|)}{m}}\right). \tag{65}$$

Let us consider the approximation operator A of parameter binarisation – that is, we retain only the signs of all parameters while keeping the same architecture, as it has been done in practice in BinaryConnect (Courbariaux et al., 2015). Hence $\mathcal{H}_A$ is a finite class of cardinality $|\mathcal{H}_A| = 2^{|W|}$. By Massart's finite lemma (Mohri et al., 2018), Theorem 3.7 the Rademacher complexity of this class of approximate classifiers is

$$\widehat{\mathcal{R}}_S(\mathcal{H}_A) \leq \sqrt{\max_{f\in\mathcal{H}_A}\frac{1}{n}\sum_{i\in[n]} f^2(x_i)}\sqrt{\frac{|W\log(2)|}{m}} = \mathcal{O}\left(\sqrt{\frac{|W|}{m}}\right) \tag{66}$$

Observe, this is independent of the network depth L. The same reasoning holds if quantisation in pursued into q bins, since then $|\mathcal{H}_A| = q^{|W|}$. By contrast, the complexity of the original class $\mathcal{H}$ grows with L. Applying our Theorem 2 combined with Lemma 4.2 under condition 4.1, we obtain the following depth-independent error bound for both the full-precision and the binarised network (i.e. a guarantee on $\max\{\operatorname{err}(A\hat{f}), \operatorname{err}(\hat{f})\}$). Condition 4.1 in this setting is implied whenever the number of points on which the binarised network disagrees with the full-precision network is of constant order with respect to the training set size.

Corollary 4.3 (Learning with binarised deep nets) *Let $\mathcal{H}$ be the class of arbitrary depth neural network classifiers having $|W|$ parameters, ReLu activations in the hidden layers, 0–1 outputs, and 0–1 loss. Let the approximation operator A be parameter binarisation. Suppose the structured sensitivity condition 4.1 holds. For $t \geq 0$, let $f_t^* := \operatorname{argmin}_{f\in\mathcal{H}_t}\{\operatorname{err}(f)\}$, and let $t^* := \operatorname{argmin}_{t\geq 0}\{\operatorname{err}(f_t^*) + 2t\}$. Then, with probability at least $1-\delta$, the network $\hat{f}$ trained by minimising (24) on a labelled sample of size m satisfies*

$$\max\{\operatorname{err}(A\hat{f}), \operatorname{err}(\hat{f})\} \leq \operatorname{err}(f_{t^*}^*) + 2t^* + 2\sqrt{\frac{|W|}{m}} + 5\sqrt{\frac{\ln(\frac{16}{\delta})}{2m}} + \frac{c}{m} + t^*\sqrt{\frac{2\ln(\frac{2}{\delta})}{m}} + \frac{6C\ln(\frac{2}{\delta})}{m},$$

where $c > 0$ is a constant depending only on the parameters of condition 4.1.

Proof In Theorem 2 we use $m_u = m$, replace (66) for $\widehat{\mathcal{R}}_S(\mathcal{H}_A)$, use Lemma 4.2 for ϵ_u, and $\rho = 1$. □

We included in Appendix a numerical illustration of algorithm (24) with binarised deep nets as in Corollary 4.3.

Here we would like to discuss a potential interpretation of the result of Corollary 4.3. BinaryConnect and quantised deep nets are known empirically to be successful from the previous literature (Courbariaux et al., 2015; Hubara et al., 2017), e.g. in image classification problems. Our theory suggests that there must be something fortuitous about many natural *data sources* that, in our context, makes the complex function class of deep nets behave as a low complexity class. We can only speculate on this and to this end we identified Condition 4.1. It is intriguing that the same condition also turned out to explain depth independence of error in this example. There have been many attempts to depth independent error bounds for deep nets in the literature, by making various assumptions, for instance by imposing norm constraints on the weights (Golowich et al., 2020). Our above interpretation provides a complementary view on this, simply as a byproduct of our general pursuit to understand approximate predictors.

4.2.2 Towards understanding neural projections

Having developed an analytic approach to the twofold problem of learning a good full-precision predictor and a good approximate predictor, it may now be interesting to relate the training objective function we obtained in (25) to the training objective of Neural Projections proposed in Ravi (2019). The latter has been a practical approach to on-device deep networks. It has no theoretical backing, however ample empirical evidence demonstrated its impressive success in real world image classification problems (Ravi, 2019). It minimises a weighted sum of three terms – the empirical errors of full and approximate models plus their disagreement – with the ultimate goal to deploy the approximate model on-device.

Take any $\eta \in [0, 1]$. Our training objective can be written as the following.

$$\begin{aligned}
\widehat{\mathrm{err}}\,(Af) + \lambda \widehat{\mathcal{D}}_A(f) &= \frac{\eta}{2}\,\widehat{\mathrm{err}}\,(Af) + \frac{1-\eta}{2}\,\widehat{\mathrm{err}}\,(Af) + \lambda \widehat{\mathcal{D}}_A(f) \\
&\leq \frac{\eta}{2}\,\widehat{\mathrm{err}}\,(f) + \frac{\eta}{2}\lambda \widehat{\mathcal{D}}_A(f) + \frac{1-\eta}{2}\,\widehat{\mathrm{err}}\,(Af) + \lambda \widehat{\mathcal{D}}_A(f) \\
&= \frac{\eta}{2}\,\widehat{\mathrm{err}}\,(f) + \frac{1-\eta}{2}\,\widehat{\mathrm{err}}\,(Af) + \frac{1-\eta+2\lambda}{2}\widehat{\mathcal{D}}_A(f) \\
&\propto \widehat{\mathrm{err}}\,(f) + \lambda_1\,\widehat{\mathrm{err}}\,(Af) + \lambda_2 \widehat{\mathcal{D}}_A(f),
\end{aligned}$$

where $\lambda_1 = 1/\eta - 1 \geq 0, \lambda_2 = 1/\eta - 1 + 2\lambda/\eta \geq 0$.

Now, if we relax Af in $\mathcal{H}_A$, i.e. replace it with some $g \in \mathcal{H}_A$, then we arrive precisely at the training objective of Neural Projections. Indeed, in Ravi (2019) this modified objective is minimised in the parameters of f and g along with tuning both λ_1 and λ_2 independently. Thus, we may interpret the training objective function of Neural Projections as an approximate version of our objective function (25). While this has no theoretical justification, our objective function has a similar flavour, and it follows from a rigorous theory. Hence, while we reckon this is not a complete explanation of why Neural Projections (Ravi, 2019) are so effective in practice, nevertheless we believe this interpretation still brings some insights into its working.

4.3 Potential extensions to stochastic approximate predictors

The approximation schemes assumed so far were deterministic. Many approximation schemes are in fact stochastic in nature, therefore, in this section we discuss how to straightforwardly adapt our framework to stochastic approximation schemes.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Then we define a Stochastic approximation scheme by $A : \Omega \times \mathcal{H} \to \mathcal{H}_{\mathcal{A}}$, where $\mathcal{H}_{\mathcal{A}} := \{A_\omega f : \omega \in \Omega \text{ and } f \in \mathcal{H}\}$. Then for a fixed $\omega \in \Omega$ we have an approximation operator $A_\omega : \mathcal{H} \to \mathcal{H}_\omega$ where $\mathcal{H}_\omega := \{A_\omega f : f \in \mathcal{H}\}$; that is, for a fixed ω we have one approximation operator. Thus, when $|\Omega| = 1$ we reduce to the deterministic setting. Also, for a fixed $f \in \mathcal{H}$ we have the collection of possible approximations to f the set $\{A_\omega f : \omega \in \Omega\}$.

Now we define $\mathcal{D}_\omega(f) := \mathcal{D}_{A_\omega}(f)$, and then for a fixed arbitrary $\omega \in \Omega$ we have with probability at least $1 - \delta$, that

$$\operatorname{err}(f) \leq \widehat{\operatorname{err}}(A_\omega f) + \rho \mathcal{D}_\omega(f) + 2\rho \widehat{\mathcal{R}}_S(\mathcal{H}_\omega) + 3\sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2m}}, \tag{67}$$

for all $f \in \mathcal{H}$. This uniform bound follows directly from Lemma 2.3 combined with a standard Rademacher bound, and for fixed ω the first two terms on its right hand side correspond to the objective function of the Algorithm (18) in Sect. 2.3.

We can make this independent of a particular random instance, e.g. by considering expectation. Although we cannot take expectation on both sides as this would incur a union bound over infinitely many sets, we can simply write

$$\begin{aligned}
\operatorname{err}(f) &= \operatorname{err}(f) - \mathbb{E}_\omega \operatorname{err}(A_\omega f) + \mathbb{E}_\omega \operatorname{err}(A_\omega f) \\
&\leq \rho \mathbb{E}_\omega \mathcal{D}_\omega(f) + \mathbb{E}_\omega \operatorname{err}(A_\omega f) - \mathbb{E}_\omega \widehat{\operatorname{err}}(A_\omega f) + \mathbb{E}_\omega \widehat{\operatorname{err}}(A_\omega f) \\
&\leq \rho \mathbb{E}_\omega \mathcal{D}_\omega(f) + \sup_{f \in \mathcal{H}} \left[\mathbb{E}_\omega \operatorname{err}(A_\omega f) - \mathbb{E}_\omega \widehat{\operatorname{err}}(A_\omega f) \right] + \mathbb{E}_\omega \widehat{\operatorname{err}}(A_\omega f).
\end{aligned}$$

Now applying Jensen's inequality, we have

$$\sup_{f \in \mathcal{H}} \left[\mathbb{E}_\omega \operatorname{err}(A_\omega f) - \mathbb{E}_\omega \widehat{\operatorname{err}}(A_\omega f) \right] \leq \mathbb{E}_\omega \sup_{f \in \mathcal{H}} \left[\operatorname{err}(A_\omega f) - \widehat{\operatorname{err}}(A_\omega f) \right],$$

and the argument of the expectation can be bounded in terms of the Rademacher complexity $\mathcal{R}_m(\mathcal{H}_\omega)$. Thus, we have the following uniform bound expressed in terms of the expected sensitivity, the expected Rademacher complexity of the small approximating class, and a new empirical error term that, due to the expectation may be interpreted as a data augmentation loss. That is, we have, with probability at least $1 - \delta$, the following

$$\operatorname{err}(f) \leq \mathbb{E}_\omega \widehat{\operatorname{err}}(A_\omega f) + \rho \mathbb{E}_\omega \mathcal{D}_\omega(f) + 2\rho \mathbb{E}_\omega \mathcal{R}_m(\mathcal{H}_\omega) + \sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{2m}}. \tag{68}$$

Minimising the first two terms on its right hand side could be used to justify a regularised data augmentation algorithm in analogy with our previous algorithm in (18).

Likewise, one can introduce estimates of the expected distortion $\mathcal{D}_\omega(f)$ from unlabeled data. Alternatively, if the approximation operator A satisfies a variance condition, namely that $\left[\mathbb{E}_\omega \|A_w f - f\|_{L^2}^2 \right]^{\frac{1}{2}} \leq \alpha \mathcal{C}(f)$ for all $f \in \mathcal{H}$, where $\mathcal{C}(f)$ is some property of $f \in \mathcal{H}$, then we have, by Jensen's inequality and the variance condition, $\mathbb{E}_\omega[\mathcal{D}_\omega(f)] \leq \mathbb{E}_\omega[\mathcal{D}_\omega(f)^2]^{\frac{1}{2}} = \left[\mathbb{E}_\omega \mathbb{E}_{x \sim D_x} |A_\omega f(x) - f(x)|^2 \right]^{\frac{1}{2}} = \left[\mathbb{E}_\omega \|A_\omega f - f\|_{L^2}^2 \right]^{\frac{1}{2}} \leq \alpha \mathcal{C}(f)$. So we see this variance condition on A provides another instance where the need for additional unlabelled data is eliminated in the case of stochastic approximation operators. A similar condition, formulated on the level of parameters, is frequently

encountered in the literature of quantisation for learning and optimisation, such as in stochastic rounding (Alistarh et al., 2017; Wen et al., 2017).

5 Conclusions

We end our study with a high-level summary. Inspired by the recent surge of interest in model-compression and approximate learning algorithms in the context of small device settings, we studied the role of approximability in generalisation, both in the full precision and in the approximated settings. Our main findings can be summarised as follows: (1) For any given PAC-learnable problem, and any approximation scheme, target concepts that have low sensitivity to the approximation can be learned from a smaller labelled sample, provided sufficient unlabelled data. This is achieved by using approximation to modify the loss function and isolating a sensitivity term in the generalisation error. The modified loss function has a lower complexity in comparison with the original, pushing the complexity of the learning problem onto the class of sensitivity functions – which in turn only requires unlabeled data for estimation whenever the original loss is Lipschitz. (2) Our analysis yielded algorithms showing that it is possible to learn a good predictor whose approximation has the same generalisation guarantee as the full precision predictor. Owing to the generality of our approach, such provably accurate approximate predictors can be used with a variety of model-compression and approximation schemes, and potentially deployed in memory-constrained settings. (3) Our algorithms use unlabelled data to estimate the sensitivity of predictors to the given approximation operator, and this needs not be independent from the labelled training set. Moreover, while the required unlabelled sample complexity can be large in general, we highlighted several examples of natural structure in the class of sensitivities that significantly reduce, and possibly even eliminate, the need of additional unlabelled data. At the same time, structural properties of the sensitivity class shed new light onto the question of what makes certain instances of learning problems easier than others.

Several open questions remain. As our upper bounds highlighted structural traits that explain good performance in model-compression settings, it will be interesting to develop lower bounds under the same structural traits, to assess the tightness of our bounds. From the practical perspective, it will be interesting to develop efficient implementations, and study their computational complexity. Another line of interesting future work is to explore adversarial settings (Chowdhury et al., 2022; Montasser et al., 2019), where the approximation operator A is in the hands of an adversary, and the learner needs to find a predictor that is robust to it. Furthermore, it would be interesting to study model-compression and approximate algorithms in other learning theory frameworks such as PAC-Bayes, and perhaps even non-uniform frameworks.

Appendix 1: Numerical illustration

We presented a theory for learning with model approximation / model compression. Our algorithm (24), and its refinement (25), were of theoretical interest in that pursuit, and we should note that, for certain choices of approximation—such as weight-binarisation or

quantisation—our minimisation objective is not differentiable. However, it may still be interesting to illustrate its working in numerical experiments, which we do in this section.

Appendix 1.1: A multi-objective optimisation approach

We employ a general-purpose assumption-free method, known as NSGA-II, a multi-objective evolutionary algorithm based on non-dominated sorting (Deb et al., 2002). This is an iterative population-based heuristic approach that has had many successful applications in practice. Its computation complexity is of order $\mathcal{O}(MN^2)$ per iteration, where M is the number of objectives (in our case, $M = 2$), and N is the population size. The latter is a parameter representing the number of candidate solutions at any given iteration. NSGA-II returns a set of non-dominated solutions (classifiers in our case) that estimate the Pareto front, each solution representing a different tradeoff between the objectives.

NSGA-II was previously demonstrated to work well in regularised machine learning problems (Chen & Yao, 2010), as it alleviates the need for tuning the balance between competing terms. The user can choose from the returned solutions *a-posteriori*—for instance based on validation errors, or some other criteria depending on the application context.

Our objective function breaks up naturally in two components, which we will minimise simultaneously:

$$E_1 \equiv \widehat{\mathrm{err}}\,(Af) \qquad \text{and} \qquad E_2 \equiv \widehat{\mathcal{D}}_A(f)$$

We shall demonstrate the working of this approach with the approximation operator A taken to be weight-binarisation, as in BinaryConnect (Courbariaux et al., 2015) – a non-differentiable problem.

Appendix 1.2: Implementation and parameter setup

Our implementation is based on the Python package PyMOO[1] (Blank & Deb, 2020). The candidate predictors are fully connected 3-layer feed-forward neural network classifiers with ReLu activation on the hidden nodes, thresholded sigmoid on the output node, and no regularisation (other than the implicit effect of the approximation operator). We employ the 0–1 loss directly, since the optimiser allows non-differentiable objectives.

We have set the population size to $N = 100$, following (Chen & Yao, 2010), and we stop when the change in both objectives becomes less than 10^{-3} for 100 consecutive iterations, or when the allotted computing time is exhausted. We use default settings (simulated binary crossover, and polynomial mutation) with default parameters, with two additions[2] that enhance efficiency for our problem, as follows. Firstly, we added a constraint to ensure the sample error $\widehat{\mathrm{err}}\,(Af)$ shrinks throughout iterations to no larger than 0.3. This speeds up the process in our experience, as new candidates are then able to explore more promising regions of the hypothesis space. Secondly, at each iteration, we eliminate candidates with identical objective values, even if they are different in the parameter space,

[1] https://pymoo.org/

[2] credit to Yangfan Peng

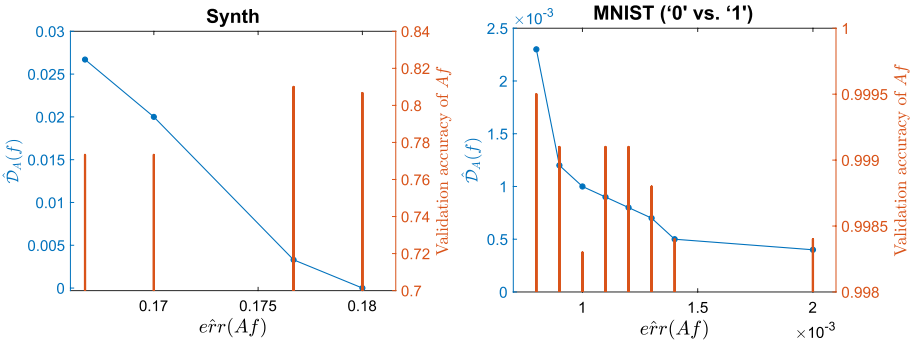

Fig. 1 Estimated Pareto fronts (blue curves) obtained in a typical run of the multi-objective minimiser on Synth and MNIST. Each marker on the curve represents one non-dominated classifier obtained. For each of these, we also show the validation accuracy of Af (vertical lines) on the scale on the right axis

breaking ties randomly. This encourages diversity of the candidate set, leading to a better coverage of the Parento front.

Appendix 1.3: Data sets and protocol

We created a 2D synthetic data set (Synth) by sampling two 0-mean axis-aligned Gaussians with variances (1, 0.3) and (0.3, 1) respectively, following the 'Bumpy' data set description from (Chen & Yao, 2010). The training set has 600 points, of which we used half for validation, and the independent test set has another 300 points. The validation set is only used to select one classifier from the non-dominated set of solutions returned by the multi-objective optimiser.

The second data set we use is MNIST. This consists of 28×28 pixel images of handwritten numerals, of which we took two classes representing '0' versus '1'. We have $4702 + 5430$ images in these two classes for training, a further 2533 for validation, and an independent test set of size 2115. No pre-processing was applied to either data sets.

We use the same number of hidden nodes in all candidate classifiers which, in the reported experiments, we set to 10 per hidden later in the case of Synth, and to 100 per hidden layer in the case of MNIST.

Appendix 1.4: Results

In Fig. 1 we show the estimated Pareto fronts (blue curves) obtained on Synth and on MNIST after one full run (300 iterations) of the algorithm. The markers on these curves represent the two objective values of the non-dominated solutions found in the last generation. As we can see from the figure, each of these classifiers exhibit a different tradeoff between their sample error and sensitivity. For each of these classifiers $\hat{f}$, we then computed the validation-set accuracy of their weight-binarised version, $A\hat{f}$ (vertical lines). The specific tradeoff at which the validation accuracy is highest is data set dependent, and this is one of the reasons that a multi-objective approach capable of capturing multiple tradeoffs is well suited. Indeed in the case of MNIST, the highest validation accuracy happens to

Table 1 Test accuracy (%) results for our trained full-precision classifier ($\hat{f}$) and its weight-binarised version ($A\hat{f}$), compared with classic training of a full-precision network using Adam Kingma and Ba (2015)

Dataset	Hidden layers	$A\hat{f}$	$\hat{f}$	Classic
Synth	{10, 10}	80.27 ± 1.67	80.13 ± 1.15	80.00 ± 1.01
MNIST (0 vs. 1)	{100,100}	99.89 ± 0.04	99.80 ± 0.12	99.91 ± 0.03

Averages and standard deviations are reported from 5 independent repetitions. Note that Af operates with parameter values of ±1, yet it performs comparably

correspond to the classifier that also has the lowest sample error (second plot of Fig. 1)—in contrast, for Synth (first plot of Fig. 1, which has overlapping classes, the highest validation accuracy occurs in a classifier with relatively high sample error but relatively low sensitivity—so here a lower sensitivity in fact appears to help prevent overfitting.

From the non-dominated classifiers obtained (as in Fig. 1), we select the one with the highest validation accuracy for $A\hat{f}$, and take the corresponding $\hat{f}$ forward for evaluation on the independent test set. We performed 5 independent repetitions of this entire procedure, and report the average and standard deviation of test set accuracy in Table 1. As predicted from our theory, we see that $\hat{f}$ and $A\hat{f}$ perform very similarly. For reference and comparison, following the experiment protocol in Courbariaux et al. (2015), we also trained a classic, differentiable version of the full-precision network, having the same 2-hidden layer architecture, but with sigmoid output, cross-entropy loss, and L_2-weight decay, using Adam (Kingma & Ba, 2015). From Table 1 we see the performances are very similar, and our weight-binarised $A\hat{f}$ performs on-par with the classic full-precision network. These experimental findings are similar to those in Courbariaux et al. (2015), but in contrast our algorithm has a principled theoretical foundation.

Appendix 1.5: Computing effort

The main computational burden is at the training phase. While the per-iteration computational complexity is polynomial, convergence can take a large number of iterations on larger data sets. However, the test-time computation speed is not hindered; the binarised network runs at the same speed as BinaryConnect (Courbariaux et al., 2015) at test time. One can devise more efficient optimisation procedures, for instance by exploiting parallelisation, or by developing specialised optimisation methods for particular model-approximation schemes.

Acknowledgements This work was funded by EPSRC Fellowship EP/P004245/1 "FORGING: Fortuitous Geometries and Compressive Learning". We thank the anonymous reviewers for they thoughtful comments and suggestions, which greatly improved the presentation. We thank Yangfan Peng (MSc student at Birmingham) for implementing the example presented in the Appendix.

Author contributions Conception and design: AK and AT. Manuscript preparation: AT and AK. Manuscript revisions: AK and AT. Supervision: AK.

Funding The funding was provided by EPSRC Fellowship (Grand No. 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.

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

References

Alistarh, D., Grubic, D., Li, J. Z., Tomioka, R., & Vojnovic, M.: Qsgd: Communication-efficient SGD via gradient quantization and encoding. In *Proceedings of the 31st international conference on neural information processing systems (NIPS'17)* (pp. 1707–1718). Curran Associates Inc.

Arora, S., Ge, R., Neyshabur, B., & Zhang, Y. (2018). Stronger generalization bounds for deep nets via a compression approach. In *International conference on machine learning*, PMLR (pp. 254–263).

Ashbrock, J., & Powell, A. M. (2021). Stochastic Markov gradient descent and training low-bit neural networks. *Sampling Theory, Signal Processing, and Data Analysis, 19*(15), 1.

Bălcan, M.-F., & Blum, A. (2010). A discriminative model for semi-supervised learning. *Journal of the ACM, 57*(3), 1.

Bartlett, P. L., Bousquet, O., & Mendelson, S. (2005). Local Rademacher complexities. *The Annals of Statistics, 33*(4), 1497–1537.

Bartlett, P. L., Harvey, N., Liaw, C., & Mehrabian, A. (2019). Nearly-tight VC-dimension and pseudo dimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research, 20*(63), 1–17.

Bartlett, P. L., Long, P. M., Lugosi, G., & Tsigler, A. (2020). Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences, 117*(48), 30063–30070.

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

Baykal, C., Liebenwein, L., Gilitschenski, I., Feldman, D., & Rus, D. (2019). Data-dependent coresets for compressing neural networks with applications to generalization bounds. In *7th International conference on learning representations (ICLR)*.

Blank, J., & Deb, K. (2020). pymoo: Multi-objective optimization in python. *IEEE Access, 8,* 89497–89509.

Bu, Y., Gao, W., Zou, S., & Veeravalli, V. V. (2021). Population risk improvement with model compression: An information-theoretic approach. *Entropy (Basel), 23*(10), 1.

Chapelle, O., Schölkopf, B., & Zien, A. (2006). In *Semi-supervised learning (adaptive computation and machine learning)*, The MIT Press.

Cheng, Y., Wang, D., Zhou, P., & Zhang, T. (2017). A survey of model compression and acceleration for deep neural networks. arXiv. preprint.

Chen, H., & Yao, X. (2010). Multiobjective neural network ensembles based on regularized negative correlation learning. *IEEE Transactions on Knowledge and Data Engineering, 22*(12), 1738–1751.

Choudhary, T., Mishra, V. K., Goswami, A., & Jagannathan, S. (2020). A comprehensive survey on model compression and acceleration. *Artificial Intelligence Review, 53,* 1–43.

Chowdhury, S., & Urner, R. (2022). Robustness should not be at odds with accuracy. In L. E. Celis (Ed.), *3rd Symposium on foundations of responsible computing, FORC 2022* (Vol. 218, pp. 1–20), June 6–8, 2022, Cambridge, MA. LIPIcs. Schloss Dagstuhl - Leibniz-Zentrum für Informatik.

Courbariaux, M., Bengio, Y., & David, J.-P. (2015). BinaryConnect: Training deep neural networks with binary weights during propagations. In *Advances in neural information processing systems* (pp. 3123–3131).

Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: Nsga-II. *IEEE Transactions on Evolutionary Computation, 6*(2), 182–197.

Denil, M., Shakibi, B., Dinh, L., Ranzato, M., & de Freitas, N. (2013). Predicting parameters in deep learning. In *Advances on neural information processing systems*.

Denton, E. L., Zaremba, W., Bruna, J., LeCun, Y., & Fergus, R. (2014). Exploiting linear structure within convolutional networks for efficient evaluation. In *Advances in neural information processing systems* (pp. 1269–1277).

Gao, W., Liu, Y.-H., Wang, C., & Oh, S.: Rate distortion for model compression: From theory to practice. In K. Chaudhuri, & R. Salakhutdinov (Eds.), *Proceedings of the 36th international conference on machine learning. Proceedings of machine learning research* (Vol. 97, pp. 2102–2111).

Golowich, N., Rakhlin, A., & Shamir, O. (2020). Size-independent sample complexity of neural networks. *Information and Inference: A Journal of the IMA, 9*(2), 473–504.

Han, S., Mao, H., & Dally, W. J. (2016). Deep compression: Compressing deep neural network with pruning, trained quantization and Huffman coding. In Y. Bengio, & Y. LeCun (Eds.), *4th International conference on learning representations, (ICLR)*.

Hubara, I., Courbariaux, M., Soudry, D., El-Yaniv, R., & Bengio, Y. (2017). Quantized neural networks: Training neural networks with low precision weights and activations. *The Journal of Machine Learning Research, 18*(1), 6869–6898.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In Y. Bengio, & Y. LeCun (Eds.), *3rd International conference on learning representations, ICLR 2015, conference track proceedings*, San Diego, CA, May 7–9.

Menghani, G. (2021). Efficient deep learning: A survey on making deep learning models smaller, faster, and better.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). In *Foundations of machine learning*, MIT press.

Montasser, O., Hanneke, S., & Srebro, N. (2019). Vc classes are adversarially robustly learnable, but only improperly. In A. Beygelzimer, & D. Hsu (Eds.), *Proceedings of the thirty-second conference on learning theory. Proceedings of machine learning research, PMLR* (Vol. 99, pp. 2512–2530).

Moreau, J. J. (1965). Proximité et dualité dans un espace Hilbertien. *Bulletin de la Société Mathématique de France, 93,* 273–299.

Rastegari, M., Ordonez, V., Redmon, J., & Farhadi, A. (2016). Xnor-net: Imagenet classification using binary convolutional neural networks. In *European conference on computer vision* (pp. 525–542), Springer.

Ravi, S. (2019). Efficient on-device models using neural projections. In K. Chaudhuri, & R. Salakhutdinov (Eds.), *Proceedings of the 36th international conference on machine learning. Proceedings of machine learning research* (Vol. 97, pp. 5370–5379).

Suzuki, T., Abe, H., & Nishimura, T. (2020a). Compression based bound for non-compressed network: unified generalization error analysis of large compressible deep neural network. In *8th International conference on learning representations (ICLR)*.

Suzuki, T., Abe, H., Murata, T., Horiuchi, S., Ito, K., Wachi, T., Hirai, S., Yukishima, M., & Nishimura, T. (2020b). Spectral pruning: Compressing deep neural networks via spectral analysis and its generalization error. In *Proceedings of the twenty-ninth international joint conference on artificial intelligence, IJCAI-20* (pp. 2839–2846).

van Engelen, J. E., & Hoos, H. H. (2020). A survey on semi-supervised learning. *Machine Learning, 109*(2), 373–440.

Vapnik, V. (1998). *Statistical learning theory*, Wiley.

Wei, Y., Wainwright, M. J., & Guntuboyina, A. (2019). The geometry of hypothesis testing over convex cones: Generalized likelihood tests and minimax radii. *The Annals of Statistics, 47*(2), 994–1024.

Wen, W., Xu, C., Yan, F., Wu, C., Wang, Y., Chen, Y., & Li, H. (2017). Terngrad: Ternary gradients to reduce communication in distributed deep learning. In I. Guyon, U. V. Luxburg, U. V. Bengio, S. Wallach, H. Fergus, R. Vishwanathan, & S. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc.

Zhou, W., Veitch, V., Austern, M., Adams, R. P., & Orbanz, P. (2019). Non-vacuous generalization bounds at the imagenet scale: A PAC-Bayesian compression approach. In *7th International conference on learning representations (ICLR)*.