Machine Learning (2024) 113:1683–1704
https://doi.org/10.1007/s10994-022-06254-0

Heterogeneous sets in dimensionality reduction and ensemble learning

Henry W. J. Reeve[1] · **Ata Kabán**[2] · **Jakramate Bootkrajang**[3]

Received: 30 May 2022 / Revised: 11 August 2022 / Accepted: 19 September 2022 /
Published online: 28 October 2022

Abstract
We present a general framework for dealing with set heterogeneity in data and learning problems, which is able to exploit low complexity components. The main ingredients are (i) A definition of complexity for elements of a convex union that takes into account the complexities of their individual composition – this is used to cover the heterogeneous convex union; and (ii) Upper bounds on the complexities of restricted subsets. We demonstrate this approach in two different application areas, highlighting their conceptual connection. (1) In random projection based dimensionality reduction, we obtain improved bounds on the uniform preservation of Euclidean norms and distances when low complexity components are present in the union. (2) In statistical learning, our generalisation bounds justify heterogeneous ensemble learning methods that were incompletely understood before. We exemplify empirical results with boosting type random subspace and random projection ensembles that implement our bounds.

Keywords Heterogeneous ensembles · Random projection · Suprema of empirical processes

Editors: Yu-Feng Li and Prateek Jain.

✉ Ata Kabán
axk@cs.bham.ac.uk

Henry W. J. Reeve
henry.reeve@bristol.ac.uk

Jakramate Bootkrajang
jakramate.b@cmu.ac.th

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

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

[3] Department of Computer Science, Chiang Mai University, Muang, Chiang Mai 50200, Thailand

1 Introduction

We are interested in data and learning problems of a heterogeneous nature, which we will describe shortly. Let m be a positive integer, and consider a sequence $\mathbb{S} = (\mathbb{S}_j)_{j\in[m]}$ consisting of bounded subsets of a vector space. The convex union $\overline{\mathbb{S}}$ is defined to be the convex hull of the union of these sets,

$$\overline{\mathbb{S}} := \bigcup_{\tau\in\mathbb{N}} \left\{ \sum_{t\in[\tau]} \alpha_t \cdot s_t \,:\, (s_t)_{t\in[\tau]} \in \bigcup_{j\in[m]} \mathbb{S}_j, (\alpha_t)_{t\in[\tau]} \in \Delta_\tau \right\} = \operatorname{conv}\left(\bigcup_{j\in[m]} \mathbb{S}_j \right), \quad (1)$$

where $\Delta_\tau := \{(\alpha_t)_{t\in[\tau]} \in [0,1]^\tau \,:\, \sum_{t\in[\tau]} \alpha_t = 1\}$ is the simplex for $\tau \in \mathbb{N}$.

In dimensionality reduction, random projection (RP) is a universal and computationally convenient method that enjoys near-isometry. The distortion of Euclidean norms and distances depends on the complexity of the set being projected (see Liaw et al. (2017) and references therein). Now suppose some high dimensional data resides in a set of the form (1). What can be said about simultaneous preservation of norms and distances? The complexity of the union grows with its highest complexity component. We would like to take advantage of heterogeneity to better exploit the presence of low complexity components.

In statistical learning (SL), suppose we want to learn a weighted ensemble where base learners belong to different complexity classes. The ensemble predictor then belongs to a function class of the form (1) – for instance, learning a weighted ensemble of random subspace classifiers, as raised in the future work section of Tian and Feng (2021). What simultaneous (i.e. worst-case) generalisation guarantees can be given?

To tackle these problems, it is helpful to observe their common structure. Both problems can be described by a certain stochastic process – an infinite collection of random variables $\{X_s\}_{s\in S}$, indexed by the elements of a bounded set S. In the RP task, the index-set $S \subset \mathbb{R}^d$ is a set of points in a high dimensional space, and the source of randomness is the RP map R, a random matrix taking values in $\mathbb{R}^{k\times d}$ with independent rows drawn from a known distribution. We are interested in norm-preservation, i.e. the discrepancy between the norm of a point before and after RP, so the collection of random variables of interest is $\{\sqrt{k}\|s\|_2 - \|Rs\|_2\}_{s\in S}$, and we would like to guarantee that all of these discrepancies are small *simultaneously*, with high probability.

In the SL task, the index-set is a set of functions $\mathcal{H}$ (the hypothesis class), and the source of randomness is a training sample $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, drawn i.i.d. from an unknown distribution. We are interested in generalisation, i.e. the discrepancy between true error and sample error, so the infinite collection of random variables of interest is $\{\mathbb{E}_{X,Y}[\mathcal{L}(f(X), Y)] - \frac{1}{n}\sum_{i=1}^n \mathcal{L}(f(X_i), Y_i)\}_{f\in\mathcal{H}}$, where $\mathcal{L}$ is a loss function. Again, we want all of these discrepancies to be small *simultaneously*, with high probability.

This analogy suggests dealing with the problem of index-set heterogeneity in both tasks in a unified way. The index-sets will be of the form (1), and we appeal to empirical process theory to link the processes of interest with canonical processes whose suprema can be bounded.

1.1 Related work

The use of empirical process theory to provide low distortion guarantees for random projections of bounded sets was pioneered by the work of (Klartag and Mendelson 2005), and further refined by others – see Liaw et al. (2017) and references therein for a relatively recent treatment. These results extend the celebrated Johnson-Lindenstrauss (JL) lemma from finite sets to infinite sets. They allow simultaneous high probability statements to be made about Euclidean norm preservation of all points of the set being projected, and these guarantees depend on a notion of metric complexity of the set. For this reason, these bounds are more capable of explaining empirically observed low distortion in application areas where the underlying data support has a low intrinsic dimension, or a simple intrinsic structure, such as images or text data (Bingham and Mannila 2001). However, these existing bounds do not cater to the heterogeneity of the data support, so the presence of any small high-complexity component still renders them lose. This observation will be made precise in the sequel.

Empirical process theory is also a cornerstone in statistical learning, where it is widely used to provide uniform generalisation guarantees for learning problems (Boucheron et al. 2013) via Rademacher and Gaussian complexities. Uniform generalisation bounds ascertain, under certain conditions, that with high probability the training data does not mislead the learning algorithm. However, for complex models like heterogeneous ensembles of interest to practitioners, theory is scarce (Parnell et al. 2020, Cortes et al. 2014), Tian and Feng 2021) and a general unifying treatment is missing.

In both of the above domains, classic theory considers a single homogeneous index-set in the underlying empirical process, ignoring any heterogeneity of its subsets. The complexity of a union of sets of differing complexities grows linearly with the complexity of the most complex one, consequently by this approach one obtains uniform bounds that grow linearly with the complexity of the most complex component set. However, in many natural situations one would expect predominantly lower complexity components – for instance, data may lie mostly (though not exclusively) on low dimensional structures, or the required hypothesis class has mostly (though not exclusively) low-complexity. This class of problems motivates our approach.

1.2 Contributions

In this paper, we develop a general unifying framework that allows us to formulate simultaneous high-probability bounds over all elements of a convex union of sets of differing complexity, taking advantage of any low-complexity components. The main contributions are summarised below.

- We introduce a notion of complexity for elements of the convex union, defined as a weighted average of complexities of constituent sets. This serves to cover the convex union with sets of increasing complexity and treat each individually.
- We bound the supremum of a weighted combination of canonical subgaussian processes, which serves as a tool to bound the complexities of restricted subsets of the convex union.

We demonstrate our approach in two different areas, highlighting their conceptual connection in our framework, namely random projection based dimensionality reduction, and statistical learning of heterogeneous ensembles.

- In dimensionality reduction, heterogeneity of the data support brings improvement in simultaneous norm-preservation guarantees when points have some low complexity constitution, improving on results from Liaw et al. (2017).
- In statistical learning, our bounds justify and guide principled heterogeneous weighted ensemble construction more generally than previous work has, and we exemplify regularised gradient boosting type random subspace & random projection ensembles for high dimensional learning.

2 Theory

We begin with preliminaries, and develop some theory in Sects. 2.1 and 2.2.

Definition 1 *(Sub-Gaussian right tail)* A random variable X is said to have a sub-Gaussian right tail with parameter $\sigma > 0$ if $\mathbb{P}(X > \xi) \leq e^{-\xi^2/2\sigma^2}$ for all $\xi > 0$. Let $\mathcal{R}(\sigma^2)$ denote the collection of such random variables.

A sub-Gaussian right tail implies an expectation upper bound, as integrating the tail inequality yields $\mathbb{E}(X) \leq \int_0^\infty \mathbb{P}(X > \xi) d\xi \leq \int_0^\infty e^{-\xi^2/2\sigma^2} d\xi = \sigma\sqrt{\pi/2}$. Hence, the class $\mathcal{R}(\sigma^2)$ will serve as useful generalisation of sub-Gaussian random variables X, for which both tails decay quickly i.e. $|X| \in \mathcal{R}(\sigma^2)$.

Example 1 A univariate Gaussian random variable X with mean μ and variance σ^2 satisfies $X - \mu \in \mathcal{R}(\sigma^2)$.

In the sequel, we shall be concerned with canonical stochastic processes $\{X_s\}_{s\in S}$ indexed by a bounded set S, and their suprema $Z := \sup_{s\in S} X_s$. In many useful cases these suprema turn out to have sub-Gaussian right tail.

Example 2 *(Suprema of Gaussian processes)* Given a bounded set $S \subset \mathbb{R}^n$ let $\{X_s\}_{s\in S}$ be the Gaussian process obtained by taking a standard normal vector $g \sim \mathbb{N}(0, I_n)$ and setting $X_s := \langle s, g\rangle$. The expectation of the supremum $\mathfrak{G}(S) := \mathbb{E}[\sup_{s\in S} X_s]$ is referred to as the Gaussian width of S. The Borell-TIS inequality (Boucheron et al. 2013), Theorem 5.8) yields $\sup_{s\in S}\{X_s - \mathfrak{G}(S)\} \in \mathcal{R}(\sup_{s\in S} \sum_{i\in[n]} s_i^2)$.

Example 3 *(Suprema of Rademacher processes)* Given a bounded set $S \subset \mathbb{R}^n$, let $\{X_s\}_{s\in S}$ be the Rademacher process obtained by letting $\gamma = (\gamma_i)_{i\in[n]}$ be an i.i.d. random sequence with γ_i chosen uniformly from $\{-1, +1\}$, and $X_s := \langle s, \gamma\rangle$. Then expectation of the supremum $\mathfrak{R}(S) := \mathbb{E}[\sup_{s\in S} X_s]$ is referred to as the Rademacher width of S. By McDiarmard's inequality, $\sup_{s\in S} X_s - \mathfrak{R}(S) \in \mathcal{R}(\sum_{i\in[n]} \sup_{s\in S} s_i^2)$. We also have $\sup_{s\in S}\{X_s - \mathfrak{R}(S)\} \in \mathcal{R}(8 \cdot \sup_{s\in S} \sum_{i\in[n]} s_i^2)$ (Wainwright 2019, Example 3.5). Whilst the latter bound is sometimes tighter, we will rely primarily on the former bound in what follows.

2.1 Empirical processes over heterogeneous sets

Here we give a general result that will allow us to bound the complexity of certain subsets of a convex union. Consider the canonical stochastic process whose index set is the convex union of our interest. The next lemma bounds the supremum and its expectation for the resulting mixture process, subject to constraints, by showing that this supremum has sub-Gaussian right tail.

Lemma 1 *(Supremum of mixture process) Suppose that for each $j \in [m]$ we have a real-valued stochastic process $\{X_s^j\}_{s\in\mathbb{S}_j}$ with supremum $Z_j := \sup_{s\in\mathbb{S}_j} X_s^j$, and that for each $j \in [m]$ there exist $\mu_j, \sigma_j \in (0,\infty)$ with $Z_j - \mu_j \in \mathcal{R}(\sigma_j^2)$. Given $\mu, \sigma > 0$ we consider the following random variable*

$$\overline{Z}_{\mu,\sigma} := \sup\left\{ \sum_{t\in[\tau]} \alpha_t \cdot X_{s_t}^{j_t} \,:\, \sum_{t\in[\tau]} \alpha_t \cdot \mu_{j_t} \le \mu \text{ and } \sum_{t\in[\tau]} \alpha_t \cdot \sigma_{j_t} \le \sigma \right\},$$

where the supremum runs over all $\tau \in \mathbb{N}$, $(j_t)_{t\in[\tau]} \in [m]^\tau$, $(s_t)_{t\in[\tau]} \in \prod_{t\in[\tau]} \mathbb{S}_{j_t}$ and $(\alpha_t)_{t\in[\tau]} \in (0,\infty)^\tau$ satisfying the specified constraints. It follows that $\overline{Z}_{\mu,\sigma} - (\mu + \sigma\sqrt{2\log m}) \in \mathcal{R}(\sigma^2)$ and $\mathbb{E}\left(\overline{Z}_{\mu,\sigma}\right) \le \mu + \sigma \cdot (\sqrt{2\log m} + \sqrt{\pi/2})$.

Proof For each $\delta \in (0,1)$ we define $\mathcal{E}_\delta := \bigcap_{j\in[m]} \left\{ Z_j \le \mu_j + \sigma_j\sqrt{2\log(m/\delta)} \right\}$. By $Z_j - \mu_j \in \mathcal{R}(\sigma_j^2)$, combined with the union bound, we have $\mathbb{P}\left(\mathcal{E}_\delta\right) \ge 1 - \delta$. Next we observe that, on the event $\mathcal{E}_\delta$, for any $\tau \in \mathbb{N}$, $(\alpha_t)_{t\in[\tau]} \in (0,\infty)^\tau$, $(j_t)_{t\in[\tau]} \in [m]^\tau$ and $(s_t)_{t\in[\tau]} \in \prod_{t\in[\tau]} \mathbb{S}_{j_t}$ satisfying $\sum_{t\in[\tau]} \alpha_t \cdot \mu_{j_t} \le \mu$ and $\sum_{t\in[\tau]} \alpha_t \cdot \sigma_{j_t} \le \sigma$ we have,

$$\begin{aligned}\sum_{t\in[\tau]} \alpha_t \cdot X_{s_t}^{j_t} &\le \sum_{t\in[\tau]} \alpha_t \cdot Z_{j_t} \le \sum_{t\in[\tau]} \alpha_t \left(\mu_j + \sigma_j\sqrt{2\log(m/\delta)}\right) \\ &\le \mu + \sigma\sqrt{2\log m} + \sigma\sqrt{2\log(1/\delta)}.\end{aligned}$$

Hence, on the event $\mathcal{E}_\delta$, we have $\overline{Z}_{\mu,\sigma} \le \mu + \sigma\sqrt{2\log m} + \sigma\sqrt{2\log(1/\delta)}$. Since $\mathbb{P}\left(\mathcal{E}_\delta\right) \ge 1 - \delta$ for each $\delta \in (0,1)$ we deduce that $\overline{Z}_{\mu,\sigma} - (\mu + \sigma\sqrt{2\log m}) \in \mathcal{R}(\sigma^2)$. The expectation bound follows by integrating the tail inequality. □

2.2 Element-wise complexity-restricted subsets

We define a notion of complexity for elements of a convex union as follows.

Definition 2 *(Gaussian widths for elements of the convex hull of a union)* Given $\mathbb{S} = (\mathbb{S}_j)_{j\in[m]}$ consisting of bounded sets $\mathbb{S}_j \subset \mathbb{R}^d$ and $s \in \overline{\mathbb{S}}$, we define

$$\mathfrak{G}_{\mathbb{S}}(s) := \inf\left\{ \sum_{t\in[\tau]} \alpha_t \cdot \mathfrak{G}(\mathbb{S}_{j_t}) \,:\, s = \sum_{t\in[\tau]} \alpha_t \cdot s_{j_t} \right\},$$

where the infimum is over all $\tau \in \mathbb{N}$, $(j_t)_{t\in[\tau]} \in [m]^\tau$, $(s_t)_{t\in[\tau]} \in \prod_{t\in[\tau]} \mathbb{S}_{j_t}$ and $(\alpha_t)_{t\in[\tau]} \in \Delta_\tau$. Similarly, we define $\mathfrak{R}_{\mathbb{S}}(s)$ with $\mathfrak{R}(\mathbb{S}_{j_t})$ in place of $\mathfrak{G}(\mathbb{S}_{j_t})$.

This will be useful in obtaining high probability bounds that hold simultaneously for all elements of the convex union, yet provide individual guarantees for each – a key idea in our approach. Note that an element $s \in \overline{\mathbb{S}}$ may have multiple representations as a convex combination; the infimum breaks ties in favour of the most parsimonious one. The convex coefficients $(\alpha_t)_{t\in[\tau]}$ that realise the infimum in this definition depend on the individual element s. Note also that the complexity $\mathfrak{G}_{\mathbb{S}}(s)$ of an element $s \in \overline{\mathbb{S}}$ depends crucially upon the sequence of sets $\mathbb{S}$ with respect to which the complexity is quantified. Indeed, if the sequence of sets contains $\{s\}$, we would have $\mathfrak{G}_{\mathbb{S}}(s) = 0$.

The following result shows the utility of element-wise complexities.

Theorem 1 *(Element-wise complexity bounds) Suppose we have a sequence $\mathbb{S} = (\mathbb{S}_j)_{j\in[m]}$ consisting of sets $\mathbb{S}_j \subset \mathbb{R}^d$ with $\max_{j\in[m]} \sup_{s\in\mathbb{S}_j} \|s\|_2 \le b$ for some $b > 0$. Then, for all $\kappa \in (0, \infty)$, we have*

$$\mathfrak{G}\left(\left\{s \in \overline{\mathbb{S}} : \mathfrak{G}_{\mathbb{S}}(s) < \kappa\right\}\right) \le \kappa + b \cdot \left(\sqrt{2\log m} + \sqrt{\pi/2}\right). \tag{2}$$

Moreover, if $\mathbb{S}_j \subseteq [-r, r]^n$ for all $j \in [n]$, then for all $\kappa > 0$, then

$$\mathfrak{R}\left(\left\{s \in \overline{\mathbb{S}} : \mathfrak{R}_{\mathbb{S}}(s) < \kappa\right\}\right) \le \kappa + r \cdot (\sqrt{2n\log m} + \sqrt{\pi n/2}). \tag{3}$$

Eq. (3) also holds with $\mathfrak{G}$ in place of $\mathfrak{R}$, using (2), since $\sup_{s\in[-r,r]^n} \le \sqrt{n}r$.

Proof Both bounds are instances of Lemma 1, using the sub-Gaussian right tail properties described in Examples 2 and 3. Take $g \sim \mathcal{N}(0, I_n)$; for each $t \in [\tau]$ and $j_t \in [m]$, take the canonical process $\{X^{j_t}_{s_t}\}_{s_t\in\mathbb{S}_{j_t}}$ with $X^{j_t}_{s_t} := \langle s_t, g\rangle$. By Example 2, $\sup_{s_t\in\mathbb{S}_{j_t}} X^{j_t}_{s_t}$ has sub-Gaussian right tail with parameter $\sigma_{j_t} = \sup_{s\in\mathbb{S}_{j_t}} \|s\|_2$. By Definition 2 and Lemma 1 with $\sigma := b$, $\mu_{j_t} := \mathfrak{G}(\mathbb{S}_{j_t})$, $\mu := \kappa$, (2) follows. Now, let γ be a sequence of n i.i.d. Rademacher variables. For each $t \in [\tau]$, $j_t \in [m]$ take the canonical process $\{X^{j_t}_{s_t}\}_{s_t\in\mathbb{S}_{j_t}}$ with $X^{j_t}_{s_t} := \langle \gamma, s_t\rangle$. By Example 3, $\sup_{s_t\in\mathbb{S}_{j_t}} X^{j_t}_{s_t}$ has sub-Gaussian right tail with parameter $\sigma^2_{j_t} = \sup_{s\in\mathbb{S}_{j_t}} \sum_{i\in[n]} s_i^2 \le \sum_{i\in[n]} \sup_{s\in\mathbb{S}_j} s_i^2 \le n \cdot r^2$. Hence, Definition 2 combined with Lemma 1 with $\sigma := r\sqrt{n}$, $\mu_{j_t} := \mathfrak{G}(\mathbb{S}_{j_t})$, $\mu := \kappa$, gives (3). □

Furthermore, using element-wise complexities we can cover the convex union with sets of increasing complexity, allowing us to deal with each in turn.

Lemma 2 (Covering the convex union) *Take $\epsilon > 0$, $L := \max_{j\in[m]} \lceil \mathfrak{G}(\mathbb{S}_j)/\epsilon \rceil$ and let $T_l := \{s \in \overline{\mathbb{S}} : (l-1)\cdot\epsilon \le \mathfrak{G}_{\mathbb{S}}(s) \le l \cdot \epsilon\}$ for $l \in [L]$. Then, $\overline{\mathbb{S}} \subseteq \bigcup_{l=1}^{L} T_l$.*

A similar result holds with $\mathfrak{R}$ in place of $\mathfrak{G}$.

Proof By Definition 2, for $s \in \overline{\mathbb{S}}$, $0 \le \mathfrak{G}_{\mathbb{S}}(s) \le \max_{j\in[m]} \mathfrak{G}(\mathbb{S}_j) \le L \cdot \epsilon$, so $s \in \bigcup_{l=1}^{L} T_l$. □

The next sections rely on Theorem 1 combined with the covering approach of Lemma 2.

3 Dimension reduction for heterogeneous sets

Here we consider random projection (RP) based dimensionality reduction of sets of the form (1) in some high dimensional Euclidean ambient space, with component regions each having their own predominantly simple structure together with various higher complexity noise components. This is a realistic scenario in real world data (Wright and Ma 2022). Dimensionality reduction is often desirable before a time-consuming processing of the data, and RP is a convenient approach, oblivious to the data, with useful distance-preservation guarantees. However, there is a gap in understanding what makes RP preserve structure more accurately. We apply our theory to this problem.

Recall that a $k \times d$ random matrix R, is said to be isotropic if every row $R_{i:}$ of R satisfies $\mathbb{E}\left[R_{i:}^{\top} R_{i:}\right] = I_d$. The sub-Gaussian norm $\|\cdot\|_{\psi_2}$ of a random matrix R is defined as

$$\|R\|_{\psi_2} := \max_{i\in[k]} \left\{ \sup_{u\in\mathbb{R}^d\,:\,\|u\|_2=1} \left\{ \inf\left\{\sigma \in (0,\infty)\,:\, |R_{i:}u| \in \mathcal{R}(\sigma^2)\right\}\right\}\right\}.$$

We shall make use of the following result.

Lemma 3 (Liaw et al. 2017) *There exists a universal constant $C_{\mathfrak{L}} > 0$ such that for any isotropic $k \times d$ random matrix R, any set $S \subset \mathbb{R}^d$ and $\delta \in (0,1)$, the following holds with probability at least $1-\delta$,*

$$\sup_{s\in S}\left\{ \left| \|Rs\|_2 - \sqrt{k}\|s\|_2 \right| \right\} \le C_{\mathfrak{L}} \cdot \|R\|_{\psi_2}^2 \cdot \left(\mathfrak{G}(S) + \sqrt{\log(1/\delta)} \cdot \sup_{s\in S} \|s\|_2 \right).$$

The main result of this section is the following simultaneous bound for norm preservation.

Theorem 2 (Norm preservation in the convex union) *Suppose we have an isotropic $k \times d$ random matrix R and a sequence of sets $\mathbb{S} = (\mathbb{S}_j)_{j\in[m]}$ with $\mathbb{S}_j \subseteq \mathbb{R}^d$ and let $\overline{\mathbb{S}}$ denote the convex union (1). Suppose further that $\max_{j\in[m]} \sup_{s\in\mathbb{S}_j} \|s\|_2 \le b$ for some $b > 0$. Given any $\delta \in (0,1)$, with probability at least $1-\delta$, the following holds simultaneously for all points $s \in \overline{\mathbb{S}}$,*

$$\left| \|Rs\|_2 - \sqrt{k}\|s\|_2 \right| \le C_{\mathfrak{L}} \|R\|_{\psi_2}^2 \left(\mathfrak{G}_{\mathbb{S}}(s) + 2b\sqrt{\log\left(m\lceil \max_{j\in[m]} \mathfrak{G}(\mathbb{S}_j)/b \rceil / \delta \right) + 2\pi} \right).$$

The two dominant terms are in a tradeoff in the above bound; these are the elementwise complexity $\mathfrak{G}_{\mathbb{S}}(s)$ (cf. Definition 2), and a logarithmic function of the number of component sets m in the union. Indeed, if the union consists of many low complexity sets, then the latter quantity will increase, while if it consists of fewer high complexity sets then the former will increase.

Proof of Theorem 2 Let $\epsilon > 0$ (to be chosen later), and $L = \max_{j\in[m]} \lceil \mathfrak{G}(\mathbb{S}_j)/\epsilon \rceil$ and define sets $T_l := \{s \in \overline{\mathbb{S}} : (l-1)\cdot\epsilon \le \mathfrak{G}_{\mathbb{S}}(s) \le l\cdot\epsilon\}$, for $l \in [L]$, as in Lemma 2, so $\overline{\mathbb{S}} \subseteq \bigcup_{l=1}^{L} T_l$. By the first part of Theorem 1, for each $l \in [L]$ we have

$$\mathfrak{G}(T_l) \le l \cdot \epsilon + b \cdot \left(\sqrt{2\log m} + \sqrt{\pi/2}\right).$$

We apply Lemma 3 to each T_l and take union bound, so the following holds w.p. $1-\delta$ for all $l \in [L]$ and $s \in T_l$,

$$\begin{aligned} |\|Rs\|_2 - \sqrt{k}\|s\|_2| - C_{\mathfrak{L}}\|R\|_{\psi_2}^2 \cdot b\sqrt{\log\frac{L}{\delta}} &\le C_{\mathfrak{L}}\|R\|_{\psi_2}^2 \cdot \mathfrak{G}(T_l) \\ &\le C_{\mathfrak{L}}\|R\|_{\psi_2}^2 \cdot \left(l \cdot \epsilon + b \cdot \left(\sqrt{2\log m} + \sqrt{\pi/2}\right)\right) \\ &\le C_{\mathfrak{L}}\|R\|_{\psi_2}^2 \cdot \left(\epsilon + \mathfrak{G}_{\mathbb{S}}(s) + b \cdot \left(\sqrt{2\log m} + \sqrt{\pi/2}\right)\right). \end{aligned}$$

Finally, we take $\epsilon = b$, note $1 < \sqrt{\pi/2}$ and use $\sqrt{x} + \sqrt{y} \le \sqrt{2(x+y)}$ twice. □

The $\log(m)$ term in Theorem 2 is the price to pay for a bound which holds simultaneously over all convex combinations. Let us compare the obtained bound with the alternative of applying Liaw et al. (2017) directly to the convex union, which would give us

$$\begin{aligned} |\|Rs\|_2 - \sqrt{k}\|s\|_2| &\le C_{\mathfrak{L}}\|R\|_{\psi_2}^2 \left\{\mathfrak{G}\left(\overline{\mathbb{S}}\right) + b\sqrt{\log(1/\delta)}\right\} \\ &\le C_{\mathfrak{L}}\|R\|_{\psi_2}^2 \left\{\max_{j\in[m]} \mathfrak{G}(\mathbb{S}_j) + 2b\sqrt{\log(m/\delta) + 2\pi}\right\}, \end{aligned}$$

where the latter bound follows from (2). Crucially, in Theorem 2 the maximal complexity $\max_{j\in[m]} \mathfrak{G}(\mathbb{S}_j)$ only appears under a log in our bound. By contrast, the above bound scales linearly with this quantity.

Figure 1 exemplifies the tightening of our bound in low complexity regions of the data support in comparison with the previous uniform bound of Liaw et al. (2017).

4 Learning in heterogeneous function classes

In this section we apply the second part of Theorem 1 to heterogeneous function classes. Let $\mathcal{X}$ be the instance space (a measurable space). Throughout, we denote by $\mathcal{M}(\mathcal{X}, \mathcal{V})$ the set of (measurable) functions with domain $\mathcal{X}$ and co-domain $\mathcal{V}$. First, let us recall some classic complexity measures for function classes. Given a class $\mathcal{H} \subseteq \mathcal{M}(\mathcal{X}, \mathbb{R})$, and a sequence of points $x = (x_1, \cdots x_n)_{i\in[n]} \in \mathcal{X}^n$, the *empirical Gaussian width* $\hat{\mathfrak{G}}_n(\mathcal{H}, \boldsymbol{x})$ and *empirical Rademacher width* $\hat{\mathfrak{R}}_n(\mathcal{H}, \boldsymbol{x})$ are defined as

$$\hat{\mathfrak{G}}_n(\mathcal{H}, \boldsymbol{x}) := \mathbb{E}_g\left(\sup_{h\in\mathcal{H}}\left\{\frac{1}{n}\sum_{i\in[n]} g_i h(x_i)\right\}\right); \hat{\mathfrak{R}}_n(\mathcal{H}, \boldsymbol{x}) := \mathbb{E}_\gamma\left(\sup_{h\in\mathcal{H}}\left\{\frac{1}{n}\sum_{i\in[n]} \gamma_i h(x_i)\right\}\right)$$

where g and γ are n-dimensional standard Gaussian and Rademacher random variables respectively. The *uniform (worst-case) Gaussian width* $\mathfrak{G}_n^*(\mathcal{H})$ and *uniform Rademacher width* $\mathfrak{R}_n^*(\mathcal{H})$ are defined as

$$\mathfrak{G}_n^*(\mathcal{H}) := \sup_{\boldsymbol{x}\in\mathcal{X}^n} \hat{\mathfrak{G}}_n(\mathcal{H}, \boldsymbol{x}) \text{ and } \mathfrak{R}_n^*(\mathcal{H}) := \sup_{\boldsymbol{x}\in\mathcal{X}^n} \hat{\mathfrak{R}}_n(\mathcal{H}, \boldsymbol{x}).$$

The uniform complexities are useful in obtaining faster rates than $\mathcal{O}(n^{1/2})$, see Theorem 4.

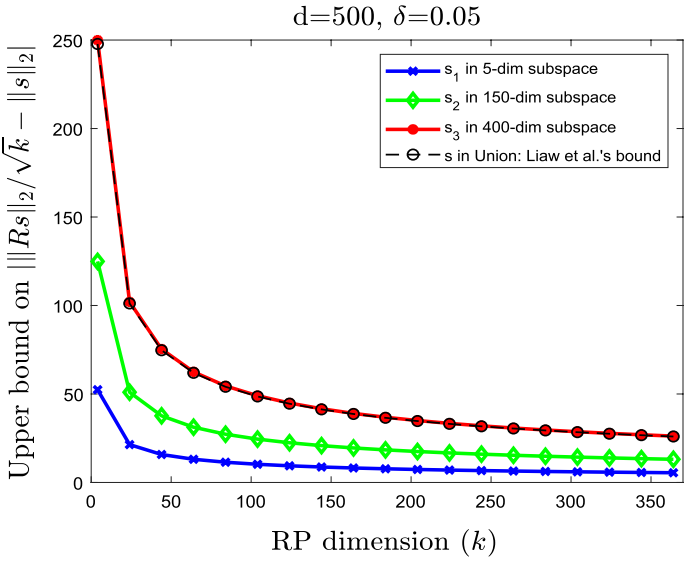

Fig. 1 Example comparison of bounds on norm preservation in a union of three linear subspaces of dimensions 5, 150 and 400 respectively in the ambient space $\mathbb{R}^{500}$ using Gaussian RP. With probability at least 0.95, we have simultaneously holding low distortion guarantees in the union of these, such that the guarantee is tighter in lower complexity subspaces at the expense of a negligible increase in the highest complexity subspace. In contrast, the previous bound (Liaw et al.2017) gives the same guarantee everywhere in the union

Finally, given a distribution P_X on $\mathcal{X}$, and $\mathcal{H} \subseteq \mathcal{M}(\mathcal{X}, \mathbb{R})$, the *Gaussian width* $\mathfrak{G}_n(\mathcal{H}, P_X)$ and *Rademacher width* $\mathfrak{R}_n(\mathcal{H}, P_X)$ are

$$\mathfrak{G}_n(\mathcal{H}, P_X) := \mathbb{E}\left[\hat{\mathfrak{G}}_n(\mathcal{H}, \boldsymbol{X})\right] \text{ and } \mathfrak{R}_n(\mathcal{H}, P_X) := \mathbb{E}\left[\hat{\mathfrak{R}}_n(\mathcal{H}, \boldsymbol{X})\right],$$

where the expectation is taken over a random sample $\boldsymbol{X} = (X_1, \cdots, X_n)$, consisting of n independent random variables X_i with distribution P_X.

We can now define our element-wise complexities. Suppose we have a sequence $\boldsymbol{\mathcal{H}} = (\mathcal{H}_j)_{j\in[m]}$ of function classes $\mathcal{H}_j \subseteq \mathcal{M}(\mathcal{X}, \mathbb{R})$ and let $\overline{\boldsymbol{\mathcal{H}}} := \text{conv}(\bigcup_{j\in[m]} \mathcal{H}_j)$ be the convex union. Given a function $f \in \overline{\boldsymbol{\mathcal{H}}}$,

$$\hat{\mathfrak{R}}_{\boldsymbol{\mathcal{H}},n}(f, \boldsymbol{x}) := \inf\left\{ \sum_{t\in[\tau]} \alpha_t \cdot \hat{\mathfrak{R}}_n(\mathcal{H}_{j_t}, \boldsymbol{x}) \ : f = \sum_{t\in[\tau]} \alpha_t \cdot h_{j_t} \right\},$$

$$\mathfrak{R}_{\boldsymbol{\mathcal{H}},n}(f, P_X) := \inf\left\{ \sum_{t\in[\tau]} \alpha_t \cdot \mathfrak{R}_n(\mathcal{H}_{j_t}, P_X) \ : f = \sum_{t\in[\tau]} \alpha_t \cdot h_{j_t} \right\},$$

$$\mathfrak{R}^*_{\boldsymbol{\mathcal{H}},n}(f) := \inf\left\{ \sum_{t\in[\tau]} \alpha_t \cdot \mathfrak{R}^*_n(\mathcal{H}_{j_t}) \ : f = \sum_{t\in[\tau]} \alpha_t \cdot h_{j_t} \right\},$$

where each infimum runs over all $\tau \in \mathbb{N}$, $(j_t)_{t\in[\tau]} \in [m]^\tau$, all $(h_t)_{t\in[\tau]} \in \prod_{t\in[\tau]} \mathcal{H}_{j_t}$ and $(\alpha_t)_{t\in[\tau]} \in \Delta_\tau$. We can also make corresponding definitions for $\hat{\mathfrak{G}}_{\mathcal{H},n}(f,\boldsymbol{x})$, $\mathfrak{G}_{\mathcal{H},n}(f,P_X)$ and $\mathfrak{G}^*_{\mathcal{H},n}(f)$; the results that follow hold unchanged.

The following lemma extends Theorem 1 to these element-complexities.

Lemma 4 *(Element-wise complexity bounds for function classes) Take $n, m \in \mathbb{N}$ and $\beta > 0$. Given $\boldsymbol{x} = (x_i)_{i\in[n]} \in \mathcal{X}^n$,*

$$\hat{\mathfrak{R}}_n\left(\left\{f \in \overline{\mathcal{H}} : \hat{\mathfrak{R}}_{\mathcal{H},n}(f,\boldsymbol{x}) < \kappa\right\}, \boldsymbol{x}\right) \leq \kappa + \beta \cdot \left(\sqrt{\frac{2\log m}{n}} + \sqrt{\frac{\pi}{2n}}\right). \tag{4}$$

Moreover, the bound (4) also holds with any one of $\hat{\mathfrak{G}}_{\mathcal{H},n}(\cdot,\boldsymbol{x})$, $\mathfrak{R}^*_{\mathcal{H},n}(\cdot)$, $\mathfrak{G}^*_{\mathcal{H},n}(\cdot)$ in place of $\hat{\mathfrak{R}}_{\mathcal{H},n}(\cdot,\boldsymbol{x})$. In addition, given any distribution P_X on $\mathcal{X}$,

$$\mathfrak{R}_n\left(\left\{f \in \overline{\mathcal{H}} : \mathfrak{R}_{\mathcal{H},n}(f,P_X) < \kappa\right\}, P_X\right) \leq \kappa + 2\beta \cdot \left(\sqrt{\frac{2\log m}{n}} + \sqrt{\frac{\pi}{2n}}\right). \tag{5}$$

Moreover, the bound (5) also holds with $\mathfrak{G}_{\mathcal{H},n}(\cdot,P_X)$ in place of $\mathfrak{R}_{\mathcal{H},n}(\cdot,P_X)$.

The proof is given in the Appendix. The bound for empirical widths follows directly from Theorem 1, and the others will be reduced to these by using concentration of the empirical widths around its expectation, and for the uniform complexities this reduction will follow simply from its definition.

4.1 Learning with a Lipschitz loss

In this section we focus on the problem of supervised learning. Suppose we have a measurable input data space $\mathcal{X}$ and an output space $\mathcal{Y} \subseteq \mathbb{R}$. Suppose further that we have a tuple of random variables (X, Y), where X takes values in $\mathcal{X}$, and Y takes values in $\mathcal{Y}$, with joint distribution P, and marginal P_X over X. The learning task is defined in terms of a loss function $\mathcal{L} : \mathbb{R} \times \mathcal{Y} \to [0, B]$. The goal of the learner is to obtain a measurable mapping $f : \mathcal{X} \to \mathbb{R}$ with low risk, $\mathcal{E}_\mathcal{L}(f) \equiv \mathcal{E}_\mathcal{L}(f,P) := \mathbb{E}_{(X,Y)\sim P}\big[\mathcal{L}(f(X),Y)\big]$. Whilst the distribution P is unknown, the learner does have access to a data set $\mathcal{D} := \{(X_i,Y_i)\}_{i\in[n]}$, where (X_i,Y_i) are independent copies of (X, Y), and computes the empirical risk, $\hat{\mathcal{E}}_\mathcal{L}(f) \equiv \hat{\mathcal{E}}_\mathcal{L}(f,\mathcal{D}) := \frac{1}{n}\sum_{i\in[n]} \mathcal{L}(f(X_i),Y_i)$. This setting includes both binary classification, where $\mathcal{Y} = \{-1,+1\}$ and regression where $\mathcal{Y} = \mathbb{R}$.

The main result of this section is the following simultaneous upper bound for weighted heterogeneous ensembles, given in terms of our element-wise Rademacher width of individual predictors.

Theorem 3 *Suppose we have a bounded, Λ-Lipschitz loss function $\mathcal{L} : \mathbb{R} \times \mathcal{Y} \to [0, B]$ along with a sequence of function classes $\boldsymbol{\mathcal{H}} = (\mathcal{H}_j)_{j\in[m]}$ with each $\mathcal{H}_j \subseteq \mathcal{M}(\mathcal{X}, [-\beta,\beta])$. Given $n \in \mathbb{N}$, $\delta \in (0,1)$, with probability at least $1-\delta$, both of the following holds for all $f \in \overline{\mathcal{H}}$,*

$$\mathcal{E}_{\mathcal{L}}(f) - \hat{\mathcal{E}}_{\mathcal{L}}(f) \leq 2\Lambda \mathfrak{R}_{\mathcal{H},n}\big(f, P_X\big) + 4\Lambda\beta\sqrt{\frac{2\,(\log(m)+1)}{n}} + B\sqrt{\frac{2\log(en/\delta)}{n}}$$

$$\mathcal{E}_{\mathcal{L}}(f) - \hat{\mathcal{E}}_{\mathcal{L}}(f) \leq 2\Lambda \hat{\mathfrak{R}}_{\mathcal{H},n}(f, \boldsymbol{X}) + 2\Lambda\beta\sqrt{\frac{2\,(\log(m)+1)}{n}} + 3B\sqrt{\frac{2\log(4n/\delta)}{n}}.$$

Proof For each $l \in [n]$, let $\mathcal{F}_l := \left\{ f \in \overline{\mathcal{H}} : (l-1)\cdot\epsilon \leq \mathfrak{R}_{\mathcal{H},n}\big(f, P_X\big) \leq l\cdot\epsilon \right\}$ where $\epsilon := B/(2\Lambda n)$. By Talagrand's contraction lemma (Mohri et al. 2012, Lemma 5.1), we have $\hat{\mathfrak{R}}_n(\mathcal{L}\circ\mathcal{H}, \mathcal{D}) \leq \Lambda \cdot \hat{\mathfrak{R}}_n(\mathcal{H}, \boldsymbol{X})$. Moreover, by Lemma 4 for each $l \in [n]$,

$$\mathfrak{R}_n\big(\mathcal{F}_l, P_X\big) \leq l\cdot\epsilon + 2\beta\cdot\left(\sqrt{\frac{2\log m}{n}} + \sqrt{\frac{\pi}{2n}}\right).$$

Thus, by the classic Rademacher bound (Mohri et al. 2012, Theorem 3.3) combined with a union bound, the following holds with probability at least $1-\delta$ for all $l \in [n]$ and $f \in \mathcal{F}_l$,

$$\begin{aligned}
\mathcal{E}_{\mathcal{L}}(f) - \hat{\mathcal{E}}_{\mathcal{L}}(f) - B\cdot\sqrt{\frac{\log(n/\delta)}{2n}} &\leq 2\cdot\mathfrak{R}_n\big(\mathcal{L}\circ\mathcal{F}_l, P\big) \leq 2\Lambda\cdot\mathfrak{R}_n\big(\mathcal{F}_l, P_X\big) \\
&\leq 2\Lambda\cdot\left(l\cdot\epsilon + 2\beta\cdot\left(\sqrt{\frac{2\log m}{n}} + \sqrt{\frac{\pi}{2n}}\right)\right) \\
&\leq 2\Lambda\cdot\left(\mathfrak{R}_{\mathcal{H},n}\big(f, P_X\big) + \epsilon + 2\beta\cdot\left(\sqrt{\frac{2\log m}{n}} + \sqrt{\frac{\pi}{2n}}\right)\right) \\
&= 2\Lambda\cdot\left(\mathfrak{R}_{\mathcal{H},n}\big(f, P_X\big) + 2\beta\cdot\left(\sqrt{\frac{2\log m}{n}} + \sqrt{\frac{\pi}{2n}}\right)\right) + \frac{B}{n} \\
&\leq 2\Lambda\cdot\mathfrak{R}_{\mathcal{H},n}\big(f, P_X\big) + 4\Lambda\beta\cdot\sqrt{\frac{2\cdot(1+\log m)}{n}} + \frac{B}{n}
\end{aligned}$$

and $\sqrt{\log(n/\delta)/(2n)} + 1/n \leq \sqrt{2\log(en/\delta)/n}$ for all $n \geq 3$. This proves the first bound in Theorem 3 for all $f \in \bigcup_{l=1}^{n}\mathcal{F}_l$ and $n \geq 3$. On the other hand, if $f \in \overline{\mathcal{H}}\setminus\bigcup_{l=1}^{n}\mathcal{F}_l$ or $n \leq 2$ then $\max\{2\Lambda\cdot\mathfrak{R}_{\mathcal{H},n}\big(f, P_X\big), B\sqrt{2\log(en/\delta)/n}\} \geq B$, in which case the bound in Theorem 3 follows from $\sup_{(u,y)\in\mathbb{R}\times\mathcal{Y}}\mathcal{L}(u,y) \leq B$, which completes the proof of the first bound in Theorem 3. The second bound may be proved by a similar argument exploiting (4). □

4.2 Learning with a self-bounding Lipschitz loss

To further demonstrate the generality of our theory, here we apply Theorem 4 to multi-output learning, and show how to obtain a heterogeneous ensemble with good generalisation as well as favourable rates.

We begin with the problem-specific preliminaries. The main result of this section is Theorem 5.

The label space is $\mathcal{Y} \subseteq \{0,1\}^Q$, where Q, the number of classes, can be very large in applications, but the number of simultaneous positive labels for an instance is typically much smaller, resulting in q-sparse binary vectors $\mathbb{Y}(q) := \{(y_j)_{j\in[Q]} \in \{0,1\}^Q : \sum_{j\in[Q]} y_j \leq q\}$, where $q \leq Q$. The following definition from Reeve and Kabán (2020) was shown to explain

favourable rates for learning multi-output problems, ranging from slow rate $n^{-1/2}$, in the case of general Lipschitz losses, to fast rates n^{-1}.

Definition 3 *(Self-bounding Lipschitz condition)* A loss function $\mathcal{L} : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ is said to be (λ, θ)-*self-bounding Lipschitz* for $\lambda > 0, \theta \in [0, 1/2]$ if for all $y \in \mathcal{Y}$ and $u, u' \in \mathbb{R}$, $|\mathcal{L}(u, y) - \mathcal{L}(u', y)| \leq \lambda \cdot \max\{\mathcal{L}(u, y), \mathcal{L}(u', y)\}^{\theta} \cdot \|u - u'\|_{\infty}$.

A nice example associated with fast rates is the pick-all-labels loss (Menon et al. 2019), which generalises the multinomial logistic loss to multi-label problems.

Example 4 *(Pick-all-labels)* Given $\mathcal{Y} = \mathbb{Y}(q)$, the *pick-all-labels loss* $\mathcal{L} : \mathbb{R} \times \mathcal{Y} \to [0, \infty)$ is defined by $\mathcal{L}(u, y) := \sum_{l \in [Q]} y_l \log\left(\sum_{j \in [Q]} \exp(u_j - u_l)\right)$, where $u = (u_j)_{j \in [Q]} \in \mathbb{R}$ and $y = (y_j)_{j \in [Q]} \in \mathcal{Y}$. As shown in (Reeve and Kabán, 2020), $\mathcal{L}$ is (λ, θ)-*self-bounding Lipschitz* with $\lambda = 2\sqrt{q}$ and $\theta = 1/2$.

To capture the complexity of a multi-output function class $\mathcal{H} \subseteq \mathcal{M}(\mathcal{X}, \mathbb{R}^Q)$, its projected class is defined as $\Pi \circ \mathcal{H} := \{\Pi \circ f : f \in \mathcal{H}\} \subseteq \mathcal{M}(\mathcal{X} \times [Q], \mathbb{R})$, where $\Pi \circ f : \mathcal{X} \times [Q] \to \mathbb{R}$ defined by $(\Pi \circ f)(x, \ell) = \pi_\ell(f(x))$, and $\pi_\ell : \mathbb{R}^Q \to \mathbb{R}$ is the ℓ-th coordinate projection.

We shall make use of the following optimistic-rate bound from Reeve and Kabán (2020).

Theorem 4 (Reeve and Kabán, 2020) *Suppose we have a multi-output function class $\mathcal{H} \subseteq \mathcal{M}(\mathcal{X}, [-\beta, \beta]^Q)$ along with a (λ, θ)-self-bounding Lipschitz loss $\mathcal{L} : \mathbb{R}^Q \times \mathcal{Y} \to [0, B]$ for some $\lambda > 0, \theta \in [0, 1/2]$. Given any $\delta \in (0, 1)$, with probability at least $1 - \delta$, the following bounds hold for all $f \in \mathcal{H}$*

$$\mathcal{E}_{\mathcal{L}}(f) - \hat{\mathcal{E}}_{\mathcal{L}}(f) \leq K\left(\sqrt{\hat{\mathcal{E}}_{\mathcal{L}}(f, \mathcal{D}) \cdot \Gamma^{\lambda,\theta}_{n,Q,\delta}(\mathcal{H})} + \Gamma^{\lambda,\theta}_{n,Q,\delta}(\mathcal{H})\right),$$

where K is a numerical constant, and $\Gamma^{\lambda,\theta}_{n,Q,\delta}(\mathcal{H}) :=$

$$\left(\lambda\left(\sqrt{Q} \cdot \log^{3/2}(e\beta nQ) \cdot \mathfrak{R}^*_{nQ}(\Pi \circ \mathcal{H}) + \frac{1}{\sqrt{n}}\right)\right)^{\frac{1}{1-\theta}} + \frac{B \log(\log(n)/\delta)}{n}.$$

With these preliminaries in place, we consider convex combinations of multi-output functions. Let us suppose $\mathcal{H} := (\mathcal{H}_j)_{j \in [m]}$ consists of multi-output functions $\mathcal{H}_j \subseteq \mathcal{M}(\mathcal{X}, [-\beta, \beta]^Q)$. Note that Π is linear, so if $f \in \mathrm{conv}\left(\bigcup_{j \in [m]} \mathcal{H}_j\right)$ then we also have $\Pi \circ f \in \mathrm{conv}\left(\bigcup_{j \in [m]} \Pi \circ \mathcal{H}_j\right)$, and hence we can quantify the complexity of f through $\mathfrak{R}^*_{\Pi \circ \mathcal{H}, nQ}(\Pi \circ f)$, where $\Pi \circ \mathcal{H} := (\Pi \circ \mathcal{H}_j)_{j \in [m]}$, which leads to the following result.

Theorem 5 *Consider a (λ, θ)-self-bounding Lipschitz loss $\mathcal{L} : \mathbb{R}^Q \times \mathcal{Y} \to [0, B]$ for some $\lambda \in (0, \infty)$, $\theta \in [0, 1/2]$ and $B \in [1, \infty)$, along with multi-output function classes $\mathcal{H}_j \subseteq \mathcal{M}(\mathcal{X}, [-\beta, \beta]^Q)$ for $j \in [m]$ and $\overline{\mathcal{H}} := \mathrm{conv}\left(\bigcup_{j \in [m]} \mathcal{H}_j\right)$. Given any $\delta \in (0, 1)$, with probability at least $1 - \delta$, for all $f \in \overline{\mathcal{H}}$,*

$$\mathcal{E}_{\mathcal{L}}(f) - \hat{\mathcal{E}}_{\mathcal{L}}(f) \leq K\left(\sqrt{\hat{\mathcal{E}}_{\mathcal{L}}(f) \cdot \Gamma^{\lambda,\theta}_{n,Q,\delta}(f)} + \Gamma^{\lambda,\theta}_{n,Q,\delta}(f)\right),$$

where K is a numerical constant and $\Gamma^{\lambda,\theta}_{n,Q,\delta}(f) :=$

$$\left(\lambda\left(Q^{\frac{1}{2}}\log^{\frac{3}{2}}(e\beta nQ)\left(\mathfrak{R}^{*}_{\Pi\circ\mathcal{H},nQ}(\Pi\circ f) + \beta\sqrt{\frac{\log m}{nQ}}\right) + \frac{1}{\sqrt{n}}\right)\right)^{\frac{1}{1-\theta}} + \frac{4B\log(n/\delta)}{n}.$$

We note that often $\mathfrak{R}^{*}_{nQ}(\Pi\circ\mathcal{F}) = \tilde{\mathcal{O}}(\{nQ\}^{-1/2})$, so the dependence on the number of classes Q is, up to the mild factor $\log^{\frac{3}{2(1-\theta)}}(Q)$, only though the self-bounding Lipschitz constant λ. Hence in Example 4 there is no further dependence on Q, but only q. Since $\theta = 1/2$, we also have fast rates for multi-label heterogeneous ensembles with very large numbers of labels, provided the individual label vectors are sufficiently sparse.

Proof (Proof of Theorem 5) Take $\epsilon > 0, L \in \mathbb{N}$ (to be determined later), and for each $l \in [L]$,

$$\mathcal{F}_l = \{f \in \overline{\mathcal{H}} : (l-1)\cdot\epsilon \leq \mathfrak{R}_{\Pi\circ\mathcal{H},nQ}(\Pi\circ f) \leq l\cdot\epsilon\} \tag{6}$$

By Theorem 4 combined with the union bound, the following holds with probability at least $1-\delta$ for all $l \in [L]$ and $f \in \mathcal{F}_l$,

$$\mathcal{E}_{\mathcal{L}}(f,P) - \hat{\mathcal{E}}_{\mathcal{L}}(f,\mathcal{D}) \leq K\left(\sqrt{\hat{\mathcal{E}}_{\mathcal{L}}(f,\mathcal{D}) \cdot \Gamma^{\lambda,\theta}_{n,Q,\delta/L}(\mathcal{F}_l)} + \Gamma^{\lambda,\theta}_{n,Q,\delta/L}(\mathcal{F}_l)\right).$$

Moreover, by Lemma 4, for each $l \in [L]$ we have $\mathfrak{R}^{*}_{\Pi\circ\mathcal{H},nQ}\left(\Pi\circ\mathcal{F}_l\right) \leq l\cdot\epsilon + \beta\cdot\left(\sqrt{\frac{2\log m}{nQ}} + \sqrt{\frac{\pi}{2nQ}}\right)$. Hence, for any $\ell \in [L]$ and $f \in \mathcal{F}_\ell$ we have $\Gamma^{\lambda,\theta}_{n,Q,\delta/L}(\mathcal{F}_l) \leq$

$$\left(\lambda\left(Q^{\frac{1}{2}}\log^{\frac{3}{2}}(e\beta nQ)\cdot\left(l\cdot\epsilon + \beta\cdot\left(\sqrt{\frac{2\log m}{nQ}} + \sqrt{\frac{\pi}{2nQ}}\right)\right) + \frac{1}{\sqrt{n}}\right)\right)^{\frac{1}{1-\theta}} + \frac{B\log(L\log(n)/\delta)}{n}.$$

Moreover, since $f \in \mathcal{F}_l$, we have $l\cdot\epsilon \leq \mathfrak{R}^{*}_{\Pi\circ\mathcal{H},nQ}(\Pi\circ f) + \epsilon$. Hence, choosing $L = n$ and $\epsilon = B/n$ yields the required bound when n is sufficiently large that $\epsilon \leq \epsilon^{1-\theta}$. On the other hand, if $\epsilon > 1$, so $n < B$, then the bound is immediate. □

4.3 Algorithmic consequences and numerical experiments

We exemplify and assess the use of our generalisation bounds empirically by turning Theorems 3 and 5 into learning algorithms for binary and multi-label classification problems, by minimising the bounds. We implement these as regularised gradient boosting with random subspace and random projection based base learners. Such ensembles are heterogeneous, since each base class is defined on a different subspace of the ambient input space.

For concreteness and simplicity, we consider generalised linear model base learners. Denoting by $\Theta := \{a, b, v, w\}$ the parameters, a base learner has the form $h(x,\Theta) = a\tanh(x^{\top}w + v) + b$, where $a, b, v \in \mathbb{R}$, and $w \in \mathcal{X}$. For multi-label problems, $w \in \mathcal{X}^{\mathcal{Q}}$ and $\tanh(\cdot)$ is computed component-wise. To ensure that h has bounded outputs, we constrain the magnitudes of a and b. We do not regularise the weight vectors w, as the

Table 1 Characteristics of the binary classification data sets used

Data	# features	# examples	Train set sizes	Description
Mice	77	1079	200, 800	Protein expression
Musk	166	6598	200, 1000	Shape measurements of molecules

random dimensionality reduction itself performs a regularisation role. Thus, with k-dimensional inputs, a binary classification base class of this form has Rademacher width of order $(k/n)^{1/2}$, a multi-label base class has its $\Gamma^{\lambda,\theta}_{n,Q,\delta}$ of order $(\lambda k/n)^{\frac{1}{2(1-\theta)}}$, and neither the exponents nor n affect the minimisation. This translates into easy-to-compute individual penalties for each base learner. The pseudo-code of the resulting algorithm is given in Algorithm 1. Other base learners are of course possible, and their Rademacher width would then be replacing this penalty term. However our goal is to assess in principle the ability of our bounds to turn into competitive learning algorithms. We generated k_t for $t \in [\tau]$ for the base learners independently from a skew distribution proportional to $-\log(U)$ where $U \sim \text{Uniform}(0, 1)$, re-scaled these between 1 and half of the rank of the data matrix, and rounded them to the closest integers. This favours simpler base models, both for efficiency and to avoid large penalty terms.

Algorithm 1 Heterogeneous gradient boosting with compressive learners

Require: Loss function $\mathcal{L}$, training set $D = \{x_i, y_i\}_{i=1}^n$, regularisation parameter η, shrinkage ϵ, number of rounds T.

$F_0 \leftarrow 0$; $\tilde{\Theta}_0 \equiv [\tilde{a}_0, \tilde{b}_0, \tilde{v}_0, \tilde{w}_0] \leftarrow$ initialise randomly
for $t = 1$ to T **do**
  $k_t \leftarrow$ generate from a distribution skewed towards lower values
  $R \in \mathbb{R}^{k_t \times d} \leftarrow$ generate RP matrix or RSubspace selector
  $\{\tilde{x}_i\}_{i\in[n]} \leftarrow \{Rx_i\}_{i\in[n]}$
  $\tilde{\Theta}_t \leftarrow \arg\min_{\tilde{\Theta}} \frac{1}{n}\sum_{i=1}^n \mathcal{L}(F + h(\tilde{x}_i, \tilde{\Theta}), y_i) + \eta \cdot \sqrt{k_t} \cdot (\tilde{a}_t^2 + \tilde{b}_t^2)$
  $\Theta_t \leftarrow [\tilde{a}_t, \tilde{b}_t, \tilde{v}_t, R^\top \tilde{w}_t]$
  $F_t(\cdot) \leftarrow F_{t-1}(\cdot) + \epsilon \cdot h(\cdot, \Theta_t)$
end for
return F_T.

In addition to our heterogeneous ensembles, we also tested regularised gradient boosting on the original data; this is a homogeneous ensemble that performs all computations in the original high dimensional space. For comparisons we chose the closest related existing methods as follows. For binary classification we compare with adaboost, logitboost, and with the top results obtained by Tian and Feng (2021) by the methods RASE_1-LDA, RASE_1-kNN, RP-ens-LDA, RP-ens-kNN, as well as the classic Random Forest. For multi-label classification, we compare with existing multi-label ensembles: COCOA (Zhang et al. 2015), ECC (Read et al. 2011), and fRAkEL (Kimura et al. 2016) provided by the MLC-Toolbox (Kimura et al. 2017).We use data sets previously employed by our competitors: the largest two real-world data sets from Tian and Feng (2021), and 5 benchmark multi-label data sets from Zhang et al. (2015), Read et al. (2011) and Kimura et al. (2016).

The data characteristics are given in Tables 1 and 2. We stantardised all data sets to zero mean and unit variance. In binary problems we tested different training set sizes, following Tian and Feng (2021), leaving the rest of the data for testing. In multi-label problems we used 80% of the data for training and 20% for testing. We did not do any feature selection, to avoid external effects in assessing the informativeness of our bounds, while the RASE methods do so and hence might have some advantage in comparisons. In particular, the RASE algorithms use 200 evenly weighted base learners each selected from 500 trained candidates and meanwhile collecting information for feature selection – this totals 10,000 trained base learners – while we just train 1000 base learners in gradient boosting fashion. We have set η by 5-fold cross-validation in $\{10^{-7}, 10^{-5}, 10^{-3}, 10^{-1}, 0\} \cdot n^{-1/2}$.

The misclassification rates obtained on the binary problems are summarised in Table 3 with both exponential and logistic loss functions. The shrinkage parameter was set to 0.1, which is a common choice in gradient boosting algorithms. The multi-label results are given in Table 4, with the pick-all-labels loss function – here the values represent the average area under the ROC curve (AUC) over the labels (higher is better). We present results with shrinkage $\epsilon = 0.1$ as well as without shrinkage ($\epsilon = 1$); our heterogeneous ensembles appear more robust to the setting of this parameter than homogeneous gradient boosting, where shrinkage is known to have a role in preventing overfitting.

From Tables 3 and 4 we see that our regularised heterogeneous ensembles (s reg = regularised random subspace gradient boosting; g reg = regularised random projection gradient boosting) consistently display good performance, even best performance in several cases. The regularised high-dimensional gradient boosting (HD reg) is only sometimes better and only marginally – despite it performs the computations to train all base learners in the full dimensional input space. The logistic loss worked better than exponential on these data, likely because of noise. Interestingly, the random subspace setting of our ensembles tended to work better than random projections, which is good news both computationally and from interpretability considerations. We also see that un-regularised models (Adaboost and Logitboost) sometimes display erratic behaviour, especially in the small sample regime. RASE performs very well in general, as its in-built feature selection also has a regularisation effect. One could mimic this with our boosting-type random subspace ensemble, especially when interpretability is at premium, although we have not pursued this here. Based on these results, we conclude that our heterogeneous random subspace ensemble is a safe-bet competitive approach.

5 Relation to previous work and discussion

The following corollary shows that, with a specific example loss function, our Theorem 3 recovers a result of Cortes et al. (2014), termed as "deep boosting".

Example 5 *(The margin loss and the ramp loss)* Consider binary classifcation, we have $\mathcal{Y} = \{-1, +1\}$. The ramp loss is a Lipschitz upper bound for the zero–one loss, which is in turn upper bounded by the margin loss. More precisely, for $\rho \in [0, 1]$, the ρ-ramp loss is defined by $\mathcal{L}^r_\rho(u, y) := \min\{1, \max\{0, 1 - (1/\rho) \cdot u \cdot y\}\}$ and the margin loss is defined by $\mathcal{L}_\rho(u, y) := \mathbf{1}\{u \cdot y \leq \rho\}$. The $\mathcal{L}^r_\rho$ is $1/\rho$-Lipschitz and $\mathcal{L}_{0,1}(u, y) \leq \mathcal{L}^r_\rho(u, y) \leq \mathcal{L}_\rho(u, y)$ for all $u \in \mathbb{R}$ and $y \in \mathcal{Y}$.

Table 2 Multi-label classification data sets used

Data	# features	# examples	Label dim	Avg. labels	Description
Birds	260	645	19	1.014	Audio
Emotions	72	593	6	1.8685	Music
Flags	19	194	7	3.3918	Flag descriptors
Scene	294	2407	6	1.074	Images
Yeast	103	2417	14	4.237	Gene profiles

Table 3 First 6 rows: Binary classification error rates (average ± standard deviation computed from 10 independent repetitions) on the Mice and Musk data sets after 1000 regularised gradient-boosting rounds. Our regularised heterogeneous ensembles are 's reg' and 'g reg' (underlined). The method descriptors specify the loss function used (exp=exponential; log=logistic), the input type (s = subspace ensemble; g = random projection ensemble with Gaussian RPs in base learners; HD = original uncompressed inputs). The competing methods do not regularise their base learners. The last 5 rows are taken from Tian and Feng (2021) for comparison. Bold font indicates best performance, the second best is marked in italic if its performance is within one standard deviation of the best performer

		Mice $n = 200$	Mice $n = 800$	Musk $n = 200$	Musk $n = 1000$
Exp	s reg	*3.87 ± 0.32*	1.51 ± 0.73	*8.42 ± 1.64*	5.57 ± 0.47
	g reg	4.55 ± 0.32	1.36 ± 0.28	8.44 ± 1.20	5.54 ± 0.41
	HD reg	4.94 ± 0.95	0.97 ± 0.63	9.23 ± 0.98	5.93 ± 0.35
Log	s reg	**3.85 ± 0.47**	0.86 ± 0.82	**8.25 ± 1.26**	**4.70 ± 0.47**
	g reg	4.41 ± 0.47	0.79 ± 0.54	8.49 ± 0.98	*4.74 ± 0.44*
	HD reg	5.37 ± 1.04	*0.72 ± 0.76*	8.80 ± 1.08	5.35 ± 0.47
Adaboost		11.39 ± 1.83	1.04 ± 0.51	12.06 ± 1.71	5.52 ± 0.47
Logitboost		10.07 ± 2.09	1.25 ± 0.69	12.20 ± 1.71	5.51 ± 0.41
$RASE_1$-LDA		7.24 ± 1.10	4.49 ± 1.23	10.56 ± 1.19	7.82 ± 0.46
RP-ens-LDA		24.84 ± 2.91	22.34 ± 2.55	12.58 ± 1.86	9.50 ± 0.46
$RASE_1$-kNN		7.43 ± 2.00	**0.60 ± 0.56**	10.52 ± 1.95	5.71 ± 0.78
RP-ens-kNN		11.77 ± 2.54	0.92 ± 0.68	10.01 ± 1.66	6.89 ± 0.84
Random Forest		8.32 ± 1.71	1.04 ± 0.73	10.83 ± 1.44	5.71 ± 0.48

Table 4 AUC results in multi-label classification problems (higher values are better)

	Birds	Emotions	Flags	Scene	Yeast
$\epsilon = 1$ s reg	0.29 ±.02	0.86 ±.01	0.71 ±.01	0.77 ±.03	0.74 ±.05
g reg	0.26 ±.03	0.85 ±.01	0.71 ±.01	0.78 ±.03	0.71 ±.06
HD reg	0.24 ±.03	0.83 ±.01	0.70 ±.02	0.75 ±.03	0.70 ±.04
$\epsilon = .1$ s reg	**0.30 ±.02**	**0.89 ±.01**	0.72 ±.01	0.79 ±.02	**0.75 ±.05**
g reg	0.29 ± 0.03	0.88 ±.01	0.72 ±.01	**0.80 ±.02**	0.74 ±.06
HD reg	0.26 ±.03	0.88 ±.01	0.72 ±.01	0.79 ±.03	0.73 ±.05
COCOA	**0.30 ±.03**	0.84 ±.01	**0.74 ±.01**	0.79 ±.03	0.73 ±.06
ECC	0.28 ±.02	0.84 ±.01	**0.74 ±.01**	0.78 ±.04	0.74 ±.06
fRAkEL	0.25 ±.03	0.85 ±.01	0.73 ±.01	0.79 ±.03	0.73 ±.05

The loss function used for training was the pick-all-labels function

Corollary 1 *Consider a sequence of function classes* $\boldsymbol{\mathcal{H}} = (\mathcal{H}_j)_{j\in[m]}$ *with* $\mathcal{H}_j \subseteq \mathcal{M}(\mathcal{X}, [-1,1])$. *Given* $\delta \in (0,1)$, *with probability* $1-\delta$, *the following holds for all* $f = \sum_{t\in[\tau]} \alpha_t \cdot h_{j_t} \in \overline{\boldsymbol{\mathcal{H}}}$ *with* $(\alpha_t)_{t\in[\tau]} \in \Delta_\tau$ *and* $h_{j_t} \in \mathcal{H}_{j_t}$,

$$\mathcal{E}_{\mathcal{L}_{0,1}}(f) - \hat{\mathcal{E}}_{\mathcal{L}_\rho}(f) \leq \frac{2}{\rho} \cdot \sum_{t\in[\tau]} \alpha_t \cdot \mathfrak{R}(\mathcal{H}_{j_t}, P_X) + \frac{8}{\rho} \cdot \sqrt{\frac{2\log m}{n}} + \sqrt{\frac{2\log(n/\delta)}{n}}.$$

A similar result holds with $\hat{\mathfrak{R}}_n(\mathcal{H}_{j_t}, P_X)$ in place of $\mathfrak{R}(\mathcal{H}_{j_t}, P_X)$.

Proof Follows straightforwardly from Theorem 3 applied to Example 5 and relaxing the infimum in our definition of element-wise complexities. □

Corollary 1 is closely related to Theorem 1 of Cortes, (2014), which contains a similar result with a different proof. We can also relate our Theorem 5 to multi-class "deep boosting" given in Kuznetsov et al. (2014) in the special case of $q = 1$. Their bound grows linearly with the number of classes Q, while ours can exploit label-sparsity; their rate is $n^{1/2}$, while ours allow significantly tighter bounds when the empirical error is sufficiently low and the sample size sufficiently large.

Foremost, our theoretical framework is general and widely applicable whenever heterogeneous geometric sets are of interest. The main benefit of our approach is to allow for a unified analysis which can be straightforwardly extended, and it justifies heterogeneous ensemble constructions beyond the previous theory. For instance SnapBoost (Parnell et al. 2020) considered a mix of trees and kernel methods in gradient boosting and was empirically found very successful.

The bound suggests a regularisation should be included in the training of each base learner, proportional to the Rademacher complexity of its class. Of course the more data we have for training the less the effect of this will be – SnapBoost did not include a regularisation but trained on very large data sets. In relatively small sample settings (as we consider in Sect. 4.3) the regularisation suggested by the bound is expected to be more essential. However, we need to reckon that Rademacher complexity is hard to compute in practice, one typically resorts to upper bounds, therefore over-regularising can be a concern. This may be somewhat countered by including a balancing regularisation parameter that may be tuned by cross-validation.

6 Conclusions

We presented a general approach to deal with set heterogeneity in high probability uniform bounds, which is able to exploit low complexity components. We applied this to tighten norm preservation guarantees in random projections, and to justify and guide heterogeneous ensemble construction in statistical learning. We also exemplified concrete use cases by turning our generalisation bounds into a practical learning algorithms with competitive performance.

Appendix

Proof of Lemma 4 Given $\boldsymbol{x} \in \mathcal{X}^n$, the bound for $\hat{\mathfrak{R}}_n(\cdot, \boldsymbol{x})$ holds by the second part of Theorem 1 with $\mathbb{S}_j = \left\{ n^{-1} \cdot \big(h(x_i)\big)_{i\in[n]} \right\}_{h\in\mathcal{H}_j}$ and $r = \beta/n$.

To prove the bound for $\mathfrak{R}_n^*$ we observe that, for any fixed $\boldsymbol{x} \in \mathcal{X}^n$, we have

$$\left\{ f \in \overline{\mathcal{H}} : \mathfrak{R}^*_{\mathcal{H},n}(f) < \kappa \right\} \subseteq \left\{ f \in \overline{\mathcal{H}} : \hat{\mathfrak{R}}_{\mathcal{H},n}(f, \boldsymbol{x}) < \kappa \right\}. \tag{7}$$

Indeed, take $f \in \overline{\mathcal{H}}$ with $\mathfrak{R}^*_{\mathcal{H},n}(f) < \kappa$. It follows that $f = \sum_{t\in[\tau]} \alpha_t \cdot h_t$ with $\sum_{t\in[\tau]} \alpha_t \cdot \mathfrak{R}^*_n\big(\mathcal{H}_{j_t}\big) < \kappa$ where $(j_t)_{t\in[\tau]} \in [m]^\tau$, $(h_t)_{t\in[\tau]} \in \prod_{t\in[\tau]} \mathcal{H}_{j_t}$ and $(\alpha_t)_{t\in[\tau]} \in \Delta_\tau$. Given that $\hat{\mathfrak{R}}_n\big(\mathcal{H}_{j_t}, \boldsymbol{x}\big) \le \mathfrak{R}^*_n\big(\mathcal{H}_{j_t}\big)$ it follows that $\sum_{t\in[\tau]} \alpha_t \cdot \hat{\mathfrak{R}}_n\big(\mathcal{H}_{j_t}, \boldsymbol{x}\big) \le \sum_{t\in[\tau]} \alpha_t \cdot \mathfrak{R}^*_n\big(\mathcal{H}_{j_t}\big) < \kappa$, and so $\hat{\mathfrak{R}}_{\mathcal{H},n}(f, \boldsymbol{x}) < \kappa$, which proves the claim (7). Hence, applying the bound for $\hat{\mathfrak{R}}_n(\cdot, \boldsymbol{x})$ we have

$$\begin{aligned}\hat{\mathfrak{R}}_n\Big(\left\{ f \in \overline{\mathcal{H}} : \mathfrak{R}^*_{\mathcal{H},n}(f) < \kappa \right\}, \boldsymbol{x}\Big) &\le \hat{\mathfrak{R}}_n\Big(\left\{ f \in \overline{\mathcal{H}} : \hat{\mathfrak{R}}_{\mathcal{H},n}(f, \boldsymbol{x}) < \kappa \right\}, \boldsymbol{x}\Big) \\ &\le \kappa + \beta \cdot \left(\sqrt{\frac{2\log m}{n}} + \sqrt{\frac{\pi}{2n}} \right).\end{aligned}$$

Taking a supremum over all $\boldsymbol{x} \in \mathcal{X}^n$ we deduce the bound

$$\mathfrak{R}^*_n\Big(\left\{ f \in \overline{\mathcal{H}} : \mathfrak{R}^*_{\mathcal{H},n}(f) < \kappa \right\}\Big) \le \kappa + \beta \cdot \left(\sqrt{\frac{2\log m}{n}} + \sqrt{\frac{\pi}{2n}} \right).$$

The corresponding bound with $\mathfrak{G}^*_n$ in place of $\mathfrak{R}^*_n$ may be proved similarly.

To prove the bound for $\mathfrak{R}_n$ (5) we first apply McDiarmid's inequality (cf. (Mohri et al., 2012), (3.14)) combined with the union bound to deduce that with probability at least $1 - \delta$ the following holds for all $j \in [m]$,

$$\hat{\mathfrak{R}}_n\big(\mathcal{H}_j, \boldsymbol{X}\big) \le \mathfrak{R}_n\big(\mathcal{H}_j, P\big) + \beta \cdot \sqrt{\frac{2\log(m/\delta)}{n}}. \tag{8}$$

Let $\xi(\delta) := \beta \cdot \sqrt{2\log(m/\delta)/n}$. Now suppose $\boldsymbol{X}$ satisfies (8) and take $f \in \overline{\mathcal{H}}$ with $\mathfrak{R}_{\mathcal{H},n}(f, P) < \kappa$. It follows that $f = \sum_{t\in[\tau]} \alpha_t \cdot h_t$ with $\sum_{t\in[\tau]} \alpha_t \cdot \mathfrak{R}_n\big(\mathcal{H}_{j_t}, P\big) < \kappa$ where $\tau \in \mathbb{N}$, $(j_t)_{t\in[\tau]} \in [m]^\tau$, $(h_t)_{t\in[\tau]} \in \prod_{t\in[\tau]} \mathcal{H}_{j_t}$ and $(\alpha_t)_{t\in[\tau]} \in \Delta_\tau$. By (8) we deduce that $\sum_{t\in[\tau]} \alpha_t \cdot \hat{\mathfrak{R}}_n\big(\mathcal{H}_{j_t}, \boldsymbol{X}\big) < \kappa + \xi(\delta)$. Hence, (8) implies that

$$\left\{ f \in \overline{\mathcal{H}} : \mathfrak{R}_{\mathcal{H},n}(f, P) < \kappa \right\} \subseteq \left\{ f \in \overline{\mathcal{H}} : \hat{\mathfrak{R}}_{\mathcal{H},n}(f, \boldsymbol{X}) < \kappa + \xi(\delta) \right\}.$$

Now applying again the bound in (4), with probability at least $1 - \delta$ we have

$$\hat{\mathfrak{R}}_n\Big(\Big\{f \in \overline{\mathcal{H}} : \mathfrak{R}_{\mathcal{H},n}(f,P) < \kappa\Big\}, \boldsymbol{X}\Big) \le \hat{\mathfrak{R}}_n\Big(\Big\{f \in \overline{\mathcal{H}} : \hat{\mathfrak{R}}_{\mathcal{H},n}(f,\boldsymbol{X}) < \kappa + \xi(\delta)\Big\}, \boldsymbol{X}\Big)$$
$$\le \kappa + \xi(\delta) + \beta \cdot \left(\sqrt{\frac{2\log m}{n}} + \sqrt{\frac{\pi}{2n}}\right)$$
$$\le \kappa + \beta \cdot \left(2\sqrt{\frac{2\log m}{n}} + \sqrt{\frac{\pi}{2n}} + \sqrt{\frac{2\log(1/\delta)}{n}}\right).$$

Hence, if we define a random variable Z by

$$Z := \hat{\mathfrak{R}}_n\Big(\Big\{f \in \overline{\mathcal{H}} : \mathfrak{R}_{\mathcal{H},n}(f,P) < \kappa\Big\}, \boldsymbol{X}\Big) - \kappa - \beta \cdot \left(2\sqrt{\frac{2\log m}{n}} + \sqrt{\frac{\pi}{2n}}\right),$$

we have $Z \in \mathcal{R}(\beta^2/n)$. By integrating the tail bound we deduce $\mathbb{E}(Z) \le \beta \cdot \sqrt{\pi/2n}$. It follows from the definition of the average Rademacher width that

$$\mathfrak{R}_n\Big(\Big\{f \in \overline{\mathcal{H}} : \mathfrak{R}_{\mathcal{H},n}(f,P) < \kappa\Big\}, P\Big) = \mathbb{E}_{\boldsymbol{X}}\Big[\hat{\mathfrak{R}}_n\Big(\Big\{f \in \overline{\mathcal{H}} : \mathfrak{R}_{\mathcal{H},n}(f,P) < \kappa\Big\}, \boldsymbol{X}\Big)\Big]$$
$$\le \kappa + 2\beta \cdot \left(\sqrt{\frac{2\log m}{n}} + \sqrt{\frac{\pi}{2n}}\right),$$

as required. The proof of the corresponding bound with $\mathfrak{G}_n(\cdot, P)$ in place of $\mathfrak{G}_n(\cdot, P)$ is similar, except for replacing McDiarmid's inequality with Borell-TIS. □

List of symbols

S	A generic geometric set
$\{X_s\}_{s\in S}$	A stochastic process indexed by S
$\mathbb{S}$	A sequence $\mathbb{S} = (\mathbb{S}_j)_{j\in[m]}$ of bounded sets $\mathbb{S}_j$ where, $j \in [m]$
$\overline{\mathbb{S}}$	convex hull from $\mathbb{S}$, i.e. $\mathrm{conv}(\bigcup_{j\in[m]} \mathbb{S}_j)$
m	Number of sets in $\mathbb{S}$
τ	Number of points defining a convex hull
Δ_τ	τ-dimensional simplex
$(\alpha_t)_{t\in[\tau]}$	An element of Δ_τ
$\mathcal{R}(\cdot)$	Set of all random variables with sub-Gaussian right tail
g	Standard Gaussian vector
γ	i.i.d. Rademacher vector
$\mathfrak{G}(\cdot)$	Gaussian width
$\mathfrak{R}(\cdot)$	Rademacher width
Z	Supremum of a stochastic process
$\overline{Z}_{\mu,\sigma}$	Supremum of mixture process s.t. μ, σ dependent constraints
$\mathfrak{G}_{\mathbb{S}}(s)$	Element-wise Gaussian complexity of $s \in \overline{\mathbb{S}}$
$\mathfrak{R}_{\mathbb{S}}(s)$	Element-wise Rademacher complexity of $s \in \overline{\mathbb{S}}$
b	Largest diameter of sets in $\mathbb{S}$

$[-r, r]^n$	Hypercube shaped set used in Theorem 1
κ	Element-complexity constraint parameter in Theorem 1
L	An integer
$(T_l)_{l\in[L]}$	Sets of increasing complexity that cover $\overline{\mathbb{S}}$
d	Data dimensionality
k	Target dimension of RP, $k \leq d$
R	$k \times d$ random matrix (RP map)
$\|R\|_{\psi_2}$	Sub-Gaussian norm of R
$C_{\mathfrak{L}}$	Constant introduced in Lemma 3
$\mathcal{X}$	Instance space (a measurable space)
$\mathcal{Y}$	Label or target space, e.g. {-1,1} or $\mathbb{R}$ or $\mathbb{R}^Q$
$\mathcal{M}(\mathcal{X}, \mathcal{V})$	All measurable functions with domain $\mathcal{X}$ & co-domain $\mathcal{V}$
$\mathcal{H}$	A generic hypothesis class
$[-\beta, \beta]$	Range of values of hypothesis functions
$\boldsymbol{x}$	Non-random sequence of points, $(x_i)_{i\in[n]}$
P_X	Probability distribution on $\mathcal{X}$
$\boldsymbol{X}$	Random sequence of n points drawn i.i.d. from P_X
$\hat{\mathfrak{G}}_n(\cdot, \boldsymbol{x})$	Empirical Gaussian width of a function class
$\hat{\mathfrak{R}}_n(\cdot, \boldsymbol{x})$	Empirical Rademacher width of a function class
$\mathfrak{G}_n(\cdot, P)$	Gaussian width of a function class
$\mathfrak{R}_n(\cdot, P)$	Rademacher width of a function class
$\mathfrak{G}^*_n(\cdot)$	Uniform Gaussian complexity of a function class
$\mathfrak{R}^*_n(\cdot)$	Uniform Rademacher complexity of a function class
$\boldsymbol{\mathcal{H}}$	A sequence of hypothesis classes $(\mathcal{H}_j)_{i\in[m]}$
$\overline{\boldsymbol{\mathcal{H}}}$	Convex hull from $\boldsymbol{\mathcal{H}}$, i.e. conv($\bigcup_{j\in[m]} \mathcal{H}_j$)
$\hat{\mathfrak{R}}_{\boldsymbol{\mathcal{H}},n}(f, \boldsymbol{x})$	Element-wise empirical Rademacher complexity of $f \in \overline{\boldsymbol{\mathcal{H}}}$
$\mathfrak{R}_{\boldsymbol{\mathcal{H}},n}(f, P)$	Element-wise Rademacher complexity of $f \in \overline{\boldsymbol{\mathcal{H}}}$
$\mathfrak{R}^*_{\boldsymbol{\mathcal{H}},n}(f)$	Element-wise uniform Rademacher complexity of $f \in \overline{\boldsymbol{\mathcal{H}}}$
$\hat{\mathfrak{G}}_{\boldsymbol{\mathcal{H}},n}(f, \boldsymbol{x})$	Element-wise empirical Gaussian complexity of $f \in \overline{\boldsymbol{\mathcal{H}}}$
$\mathfrak{G}_{\boldsymbol{\mathcal{H}},n}(f, P)$	Element-wise Gaussian complexity of $f \in \overline{\boldsymbol{\mathcal{H}}}$
$\mathfrak{G}^*_{\boldsymbol{\mathcal{H}},n}(f)$	Element-wise uniform Gaussian complexity of $f \in \overline{\boldsymbol{\mathcal{H}}}$
P	Probability distribution on $\mathcal{X} \times \mathcal{Y}$
(X, Y)	A random tuple from $\mathcal{X} \times \mathcal{Y}$ drawn from P
$\mathcal{L}$	Loss function
B	Largest value of $\mathcal{L}$
Λ	Lipschitz constant of $\mathcal{L}$
$\mathcal{E}_{\mathcal{L}}(f)$	Generalisation error (risk) of f
n	Sample size
$\mathcal{D}$	Training set drawn i.i.d. from P
$\hat{\mathcal{E}}_{\mathcal{L}}(f)$	Training error (empirical risk) of f
$(\mathcal{F}_l)_{l\in[L]}$	Sets of increasing complexity that cover $\overline{\mathcal{H}}$
Q	Number of classes in multi-label problems
q	Maximum number of non-zero labels for an instance
$\mathbb{Y}(q)$	Set of all label vectors with at most $q \leq Q$ non-zeros
(λ, θ)	Self-bounding Lipschitz parameters

$\pi_\ell(f)$	f_ℓ, the ℓ-th coordinate projection of a multi-output f
$\Pi \circ f$	Projection of f: $\mathcal{X} \times [Q] \to \mathbb{R}$, $(\Pi \circ f)(x, \ell) = \pi_\ell(f(x))$
$\Pi \circ \mathcal{H}$	Projected multi-output class, $\{\Pi \circ f : f \in \mathcal{H}\}$
$\Gamma^{\lambda,\theta}_{n,Q,\delta}(\mathcal{H})$	Complexity of $\mathcal{L} \circ \mathcal{H}$ when $\mathcal{L}$ is self-(λ, θ)-Lipschitz
$\Gamma^{\lambda,\theta}_{n,Q,\delta}(f)$	Element-complexity of $f \in \overline{\mathcal{H}}$ with self-(λ, θ)-Lipschitz loss
η	Regularisation parameter in the algorithm
ϵ	Shrinkage parameter in the gradient boosting algorithm
Θ	$\Theta = \{a, b, v, w\}$ base learner's parameters
$\mathcal{L}_{0,1}$	0-1 loss
ρ	Margin parameter, a value in [0, 1]
$\mathcal{L}^r_\rho$	Ramp loss
$\mathcal{L}_\rho$	Margin loss

Acknowledgements AK & HR acknowledge the generous support of EPSRC, though the Fellowship grant EP/P004245/1. Part of the computations for Sect. 4.3 were performed using the University of Birmingham's BlueBEAR HPC service.

Author Contributions Conception and design - HR & AK; software and data analysis - JB & AK; supervision - AK; writing - HR & AK

Data availability Not applicable.

Code availability Research code is available from https://github.com/jakramate/rpgboost.

Declarations

Conflicts of interests 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 or animals performed by any of the authors.

Consent to participate Not applicable.

Consent for publication Not applicable.

References

Bingham, E., & Mannila, H. (2001) . Random projection in dimensionality reduction: applications to image and text data. In: Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 245–250 . ACM

Boucheron, S., Lugosi, G., & Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. UK: Oxford University Press.

Cannings, T.I., & Samworth, R.J. (2017). Random–projection ensemble classification series B statistical methodology. Journal of the Royal Statistical Society

Cortes, C., Mohri, M., & Syed, U. (2014). Deep boosting. In: International Conference on Machine Learning, pp. 1179–1187

Kimura, K., Kudo, M., Sun, L., & Koujaku, S. (2016). Fast random k-labelsets for large-scale multi-label classification. In: ICPR, pp. 438–443. IEEE

Kimura, K., Sun, L., & Kudo, M. (2017). MLC Toolbox: A MATLAB/OCTAVE Library for Multi-Label Classification. arXiv

Klartag, B., & Mendelson, S. (2005). Empirical processes and random projections. *Journal of Functional Analysis, 225*(1), 229–245.

Kuznetsov, V., Mohri, M., & Syed, U. (2014). Multi-class deep boosting. *Advances in Neural Information Processing Systems, 27*, 2501–2509.

Liaw, C., Mehrabian, A., Plan, Y., & Vershynin, R. (2017) . A simple tool for bounding the deviation of random matrices on geometric sets. In: Geometric Aspects of Functional Analysis, pp. 277–299. Springer,

Menon, A.K., Rawat, A.S., Reddi, S., & Kumar, S. (2019). Multilabel reductions: What is my loss optimising? In: Advances in Neural Information Processing Systems, pp. 10599–10610

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). *Foundations of Machine Learning*. UK: MIT press.

Parnell, T.P., et al. (2020) . Snapboost: A heterogeneous boosting machine. In: Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020

Read, J., Pfahringer, B., Holmes, G., & Frank, E. (2011). Classifier chains for multi-label classification. *Machine Learning, 85*, 333–359.

Reeve, H.W.J., & Kabán, A. (2020). Optimistic bounds for multi-output learning. In: Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event. Proceedings of Machine Learning Research, vol. 119, pp. 8030–8040. PMLR,

Tian, Y., & Feng, Y. (2021). Rase: Random subspace ensemble classification. *Journal of Machine Learning Research, 22*(45), 1–93.

Wainwright, M.J. (2019) . High-dimensional Statistics: A Non-asymptotic Viewpoint vol. 48. Cambridge University Press

Wright, J., & Ma, Y. (2022). *High-Dimensional Data Analysis with Low-Dimensional Models: Principles, Computation, and Applications*. UK: Cambridge University Press.

Zhang, M.-L., Li, Y.-K., & Liu, X.-Y. (2015). Towards class-imbalance aware multi-label learning. In: Proceedings of the 24th International Conference on Artificial Intelligence. IJCAI'15, pp. 4041–4047. AAAI Press