Journal of Machine Learning Research 25 (2024) 1-29 Submitted 10/21; Revised 4/24; Published 5/24

Fat-Shattering Dimension of k-fold Aggregations*

Idan Attias IDANATTI@POST.BGU.AC.IL
Aryeh Kontorovich KARYEH@CS.BGU.AC.IL
Department of Computer Science
Ben-Gurion University of the Negev, Beer Sheva, Israel

Editor: John Shawe-Taylor

Abstract

We provide estimates on the fat-shattering dimension of aggregation rules of real-valued function classes. The latter consists of all ways of choosing k functions, one from each of the k classes, and computing pointwise an "aggregate" function of these, such as the median, mean, and maximum. The bounds are stated in terms of the fat-shattering dimensions of the component classes. For linear and affine function classes, we provide a considerably sharper upper bound and a matching lower bound, achieving, in particular, an optimal dependence on k. Along the way, we improve several known results in addition to pointing out and correcting a number of erroneous claims in the literature.

Keywords: combinatorial dimension, scale-sensitive dimension, fat-shattering dimension, aggregation rules, k-fold maximum, ensemble methods

1. Introduction

The *fat-shattering dimension*, also known as "scale-sensitive" and the "parametrized variant of the P-dimension", was first defined by Kearns and Schapire (1994); its key role in learning theory lies in characterizing the PAC learnability of real-valued function classes (Alon et al., 1997; Bartlett and Long, 1998).

In this paper, we study the behavior of the fat-shattering dimension under various k-fold aggregations. Let $F_1, \ldots, F_k \subseteq \mathbb{R}^\Omega$ be real-valued function classes, and $G : \mathbb{R}^k \to \mathbb{R}$ be an aggregation rule. We consider the *aggregate* function class $G(F_1, \ldots, F_k)$, which consists of all mappings $x \mapsto G(f_1(x), \ldots, f_k(x))$, for any $f_1 \in F_1, \ldots, f_k \in F_k$. Some natural aggregation rules include the pointwise k-fold maximum, median, mean, and max-min. We seek to bound the fat-shattering complexity of $G(F_1, \ldots, F_k)$ in terms of the fat-shattering dimensions of the constituent F_is. This question naturally arises in the context of ensemble methods, such as boosting and bagging, where the learner's prediction consists of an aggregation of base learners.

The analogous question regarding aggregations of VC classes (VC dimension being the combinatorial complexity controlling the learnability of Boolean function classes) have been studied in detail and largely resolved (Baum and Haussler, 1989; Blumer et al., 1989; Eisenstat and Angluin, 2007; Eisenstat, 2009; Csikós et al., 2019). Furthermore, closure properties were also studied in the context of online classification and private PAC learning (Alon

*. A previous version of this paper was titled "Fat-shattering dimension of k-fold maxima".

et al., 2020; Ghazi et al., 2021) for the Littlestone and Threshold dimensions. However, for real-valued functions, this question remained largely uninvestigated.

1.1 Our Contributions

- For a natural class of aggregation rules that commute with shifts (see definition (7)) and commute with truncation (see definition (21)), assuming $\mathrm{fat}_\gamma(F_i) \le d$, for $1 \le i \le k$, we show that

$$\mathrm{fat}_\gamma(G(F_1,\dots,F_k)) \le O\left(dk\log^2(dk)\right), \qquad \gamma > 0.$$

  In particular, this result holds for the maximum, minimum, median, and max-min aggregations. The formal statement is given in Theorem 1.

- By using an entirely different approach, for aggregations that are L-Lipschitz ($L \ge 1$) in supremum norm (see definition (8)) and for bounded function classes $F_1,\dots,F_k \subset [-R,R]^\Omega$ with $\mathrm{fat}_{\varepsilon\gamma}(F_i) \le d$, for $1 \le i \le k$, we show that

$$\mathrm{fat}_\gamma(G(F_1,\dots,F_k)) \le O\left(dk\log^{1+\varepsilon}\frac{LRk}{\gamma}\right), \qquad 0 < \gamma/L < R \text{ and } 0 < \varepsilon < \log 2.$$

  In particular, this result holds for the maximum, minimum, median, mean, and max-min aggregations. The formal statement is given in Theorem 2.

- For R-bounded affine functions and for aggregations that are L-Lipschitz in supremum norm, we show the following dimension-free bound,

$$\mathrm{fat}_\gamma(G(F_1,\dots,F_k)) \;\le\; O\left(\frac{L^2R^2k\log(k)}{\gamma^2}\right), \qquad 0 < \gamma/L < R.$$

  This result also extends to the hinge-loss class of affine functions. In particular, this result holds for the maximum, minimum, median, mean, and max-min aggregations. We improve by a log factor the estimate of Fefferman et al. (2016, Lemma 6) on the fat-shattering dimension of max-min aggregation of linear functions. The formal statement is given in Theorem 3

  Furthermore, in Corollary 5 we show an upper bound on the Rademacher complexity of the k-fold maximum aggregation of affine functions and hinge-loss affine functions. Our bound scales with $\sqrt{k}$, improving upon Raviv et al. (2018) where the dependence on k is linear.

- For affine functions and the k-fold maximum aggregation, we show tight dimension-dependent bounds (up to constants),

$$\mathrm{fat}_\gamma(G_{\max}(F_1,\dots,F_k)) \;=\; \Theta\left(dk\log k\right), \qquad \gamma > 0,$$

  where d is the Euclidean dimension. For the formal statements, see Theorems 7 and 8.

1.2 Applications

The need to analyze the combinatorial complexity of a k-fold maximum of function classes (see (4) for the formal definition) arises in a number of diverse settings. One natural example is adversarially robust PAC learning to test time attacks for real-valued functions (Attias et al., 2022; Attias and Hanneke, 2023). In this setting, the learner observes an i.i.d. labeled sample from an unknown distribution, and the goal is to output a hypothesis with a small error on unseen examples from the same distribution, with high probability. The difference from the standard PAC learning model is that at test time, the learner only observes a corrupted example, while the prediction is tested on the original label. Formally, (x, y) is drawn from the unknown distribution, and there is an adversary that can map x to k possible corruptions z that are known to the learner. The learner observes only z while its loss is with respect to the original label y. This scenario is naturally captured by the k-fold max: the learner aims to learn the maximum aggregation of the loss classes. Attias et al. (2022) showed that uniform convergence holds in this case, and so the sample complexity of an empirical risk minimization algorithm is determined by the complexity measure of the k-fold maximum aggregation.

Analyzing the k-fold maximum arises also in a setting of learning polyhedra with a margin. Gottlieb et al. (2018) provided a learning algorithm that represents polyhedra as intersections of bounded affine functions. The sample complexity of the algorithm is determined by the complexity measure of the maximum aggregation of affine function classes.

Another natural example of where the k-fold maximum and k-fold max-min play a role is in analyzing the convergence of k-means clustering. Fefferman et al. (2016) bounded the max-min aggregation and Klochkov et al. (2021); Biau et al. (2008); Appert and Catoni (2021); Zhivotovskiy (2022) bounded the max aggregation. The main challenge in this setting is bounding the covering numbers of the aggregation over k function classes which can be obtained by bounding the Rademacher complexity or the fat-shattering dimension.

Finally, there are numerous ensemble methods for regression that output some aggregation of base learners, such as the median or mean. Examples of these methods include boosting (e.g., Freund and Schapire (1997); Kégl (2003)), bagging (bootstrap aggregation) by Breiman (1996), and its extension to the random forest algorithm (Breiman, 2001).

1.3 Related Work

It was claimed in Attias et al. (2019, Theorem 12) that

$$\mathrm{fat}_\gamma(G_{\max}(F_1, \ldots, F_k)) \quad \leq \quad 2\log(3k) \sum_{j=1}^{k} \mathrm{fat}_\gamma(F_i),$$

but the proof had a mistake (see Section 5); our Open Problem (28) asks if the general form of the bound does hold (we conjecture it does at least for the max aggregation). Using the recent disambiguation result of Alon et al. (2022) presented in Lemma 9 here, Attias et al. (2022, Lemma 15) obtained the bound

$$\mathrm{fat}_\gamma(G_{\max}(F_1, \ldots, F_k)) \leq O\left(\log(k)\log^2(|\Omega|)\sum_{j=1}^{k} \mathrm{fat}_\gamma(F_i)\right), \tag{1}$$

where Ω is the domain of the function classes $F_1, \ldots, F_k$. The latter is, in general, incomparable to Theorem 1. However, for large or infinite Ω, Theorem 1 is clearly a considerable improvement over (1).

Using the covering number results of Mendelson and Vershynin (2003); Talagrand (2003) (see Section A.2), Duan (2012, Theorem 6.2) obtained a general result, which, when specialized to k-fold maxima, yields

$$\mathrm{fat}_\gamma(G_{\max}(F_1, \ldots, F_k)) \quad \leq \quad O\left(\log \frac{k}{\gamma} \cdot \sum_{i=1}^{k} \mathrm{fat}_{c\gamma/\sqrt{k}}(F_i)\right) \tag{2}$$

for a universal constant $c > 0$; (2) is an immediate consequence of Theorem 10 (with $p = 2$), Lemma 18, and Lemma 19 in this paper. Our results improve over (2) by removing the dependence on k in the scale of the fat-shattering dimensions; however, Duan's general method is applicable to a wider class of uniformly continuous k-fold aggregations.

Srebro et al. (2010, Lemma A.2) bounded the fat-shattering dimension in terms of the Rademacher complexity. Foster and Rakhlin (2019) bounded the Rademacher complexity of a smooth k-fold aggregate, see also references therein. Inspired by Appert and Catoni (2021), Zhivotovskiy (2022) has obtained the best known upper bound on the Rademacher complexity of k-fold maxima over linear function classes. Raviv et al. (2018) upper bounded the Rademacher complexity of the k-fold maximum aggregation of affine functions and hinge-loss affine functions.

2. Preliminaries

2.1 Aggregation Rules

A k-fold *aggregation* rule is any mapping $G : \mathbb{R}^k \to \mathbb{R}$. Just as G maps k-tuples of reals into reals, it naturally aggregates k-tuples of functions into a single one: for $f_1, \ldots, f_k : \Omega \to \mathbb{R}$, we define $G(f_1, \ldots, f_k) : \Omega \to \mathbb{R}$ as the mapping $x \mapsto G(f_1(x), \ldots, f_k(x))$. Finally, the aggregation extends to k-tuples of function classes: for $F_1, \ldots, F_k \subseteq \mathbb{R}^\Omega$, we define

$$G(F_1, \ldots, F_k) := \{x \mapsto G(f_1(x), \ldots, f_k(x)) : f_i \in F_i, i \in [k]\}. \tag{3}$$

Examples. A canonical example of an aggregation rule is the k-fold max, induced by the mapping

$$G_{\max}(x_1, \ldots, x_k) := \max_{i \in [k]} x_i. \tag{4}$$

The minimum is defined analogously as

$$G_{\min}(x_1, \ldots, x_k) := \min_{i \in [k]} x_i.$$

The mean aggregation defined as

$$G_{\mathrm{mean}}(x_1, \ldots, x_k) := \frac{1}{k} \sum_{i=0}^{k} x_i.$$

Denoting by $x_{(1)}, \ldots, x_{(k)}$ the ascending order of a sequence $x_1, \ldots, x_k$, that is, $x_{(1)} \leq \cdots \leq x_{(k)}$, the (lower[1]) median is defined as

$$G_{\text{med}}(x_1, \ldots, x_k) := x_{(\lceil k/2 \rceil)}. \tag{5}$$

We also define $G_{\text{max-min}} : \mathbb{R}^{k \times \ell} \to \mathbb{R}$ as

$$G_{\text{max-min}}(x_{11}, \ldots, x_{k\ell}) := \max_{j \in [\ell]} \min_{i \in [k]} x_{ij}; \tag{6}$$

Next, we consider some properties that an aggregation rule might possess.

Commuting with shifts. We say that an aggregation rule G commutes with shifts if

$$G(x) - r = G(x - r), \qquad x \in \mathbb{R}^k, r \in \mathbb{R}, \tag{7}$$

where $x - r$ is defined as $(x_1 - r, \ldots, x_k - r)$ for $x = (x_1, \ldots, x_k)$. It is readily verified that the maximum, minimum, max-min, mean, and median commute with shifts.

Lipschitz continuity. The mapping $G : \mathbb{R}^k \to \mathbb{R}$ is L-Lipschitz with respect to $\|\cdot\|_\infty$ if

$$\left|G(x) - G(x')\right| \leq L \left\|x - x'\right\|_\infty = L \max_{i \in [k]} |x_i - x'_i|, \qquad x, x' \in \mathbb{R}^k. \tag{8}$$

In Appendix A.1, we show that maximum, median, and max-min aggregations are 1-Lipschitz (Lemmas 13, 14, 15 respectively). Showing it for the mean is a simple exercise. The proof for the minimum is similar to the one for the maximum.

We also consider aggregations that *commute with truncation*; see Section 4.1 for the formal definition.

2.2 Complexity Measures

Fat-shattering dimension. Let Ω be a set and $F \subset \mathbb{R}^\Omega$. For $\gamma > 0$, a set $S = \{x_1, \ldots, x_m\} \subset \Omega$ is said to be γ-shattered by F if

$$\sup_{r \in \mathbb{R}^m} \min_{y \in \{-1,1\}^m} \sup_{f \in F} \min_{i \in [m]} y_i (f(x_i) - r_i) \geq \gamma. \tag{9}$$

The γ-fat-shattering dimension, denoted by $\text{fat}_\gamma(F)$, is the size of the largest γ-shattered set (possibly ∞).

Fat-shattering dimension at zero. As in Gottlieb et al. (2014), we also define the notion of γ-shattering at 0, where the "shift" r in (9) is constrained to be 0. Formally, the shattering condition is $\min_{y \in \{-1,1\}^m} \sup_{f \in F} \min_{i \in [m]} y_i f(x_i) \geq \gamma$, and we denote the corresponding dimension by $\text{få}\text{t}_\gamma(F)$.

Attias et al. (2019, Lemma 13) showed that for all $F \subset \mathbb{R}^\Omega$,

$$\text{fat}_\gamma(F) = \max_{r \in \mathbb{R}^\Omega} \text{få}\text{t}_\gamma(F - r), \qquad \gamma > 0, \tag{10}$$

where $F - r = \{f - r; f \in F\}$ is the r-shifted class (the maximum is always achieved). Lemma 24 presents another, apparently novel, connection between fat and fåt.

1. Ordinarily, for even k, any $m \in [x_{(k/2)}, x_{(k/2+1)}]$ is a median of x. For the proof of Theorem 1, the median must be a value actually occurring in x.

Rademacher complexity. Let $\mathcal{F}$ be a real-valued function class on the domain space $\mathcal{W}$. Define the empirical Rademacher complexity of $\mathcal{F}$ on a given sequence $(w_1, \ldots, w_n) \in \mathcal{W}^n$ as

$$\mathcal{R}_n(\mathcal{F}|w_1, \ldots, w_n) = \mathbb{E}_\sigma \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(w_i),$$

where $\sigma = (\sigma_1, \ldots \sigma_n)$ are independent random variables uniformly chosen from $\{-1, 1\}$. The Rademacher complexity of $\mathcal{F}$ with respect to a distribution $\mathcal{D}$ is defined as

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_{w_1, \ldots, w_n \sim \mathcal{D}} R_n(\mathcal{F}|w_1, \ldots, w_n).$$

It is a classic fact (Mohri et al., 2012, Theorem 3.1) that the Rademacher complexity controls generalization bounds in a wide range of supervised learning settings.

Covering numbers. We start with some background on covering numbers. Whenever Ω is endowed with a probability measure μ, this induces, for $p \in [1, \infty)$ and $f : \Omega \to \mathbb{R}^k$, the norm

$$\|f\|^p_{L_p^{(k)}(\mu)} = \mathop{\mathbb{E}}_{X \sim \mu} \|f(X)\|_p^p = \int_\Omega \|f(x)\|_p^p \, \mathrm{d}\mu(x)$$

on $L_p^{(k)}(\mu) := \left\{ f \in (\mathbb{R}^k)^\Omega : \|f\|_{L_p^{(k)}(\mu)} < \infty \right\}$. When $k = 1$, we write $L_p(\mu) := L_p^{(1)}(\mu)$. For $p = \infty$, $\|f\|_{L_\infty^{(k)}(\mu)}$ is the essential supremum of f with respect to μ. For $t > 0$ and $H \subset F \subset L_p(\mu)$, we say that H is a t-cover of F under $\|\cdot\|_{L_p(\mu)}$ if $\sup_{f \in F} \inf_{h \in H} \|f - h\|_{L_p(\mu)} \leq t$. The t-covering number of F, denoted by $\mathcal{N}(F, L_p(\mu), t)$, is the cardinality of the smallest t-cover of F (possibly, ∞). We note the obvious relation

$$p > q \quad \Longrightarrow \quad \mathcal{N}(F, L_p(\mu), t) \geq \mathcal{N}(F, L_q(\mu), t), \tag{11}$$

which holds for all probability measures μ and all $t > 0$.

We sometimes overload the notation about aggregations by defining G on k-tuples of functions (instead of k-tuples of reals), $G : (\mathbb{R}^\Omega)^k \to \mathbb{R}^\Omega$. We say that G is L-Lipschitz with respect to $\|\cdot\|_{L_p^{(k)}(\mu)}$, if

$$\left\| G(f_{1:k}) - G(f'_{1:k}) \right\|_{L_p(\mu)} \leq L \left\| (f_{1:k}) - (f'_{1:k}) \right\|_{L_p^{(k)}(\mu)}, \qquad f_{1:k}, f'_{1:k} \in (\mathbb{R}^k)^\Omega.$$

2.3 Notation

We write $\mathbb{N} = \{0, 1, \ldots\}$ to denote the natural numbers. For $n \in \mathbb{N}$, we write $[n] := \{1, 2, \ldots, n\}$. All of our logarithms are base e, unless explicitly denoted otherwise. We use $\max\{u, v\}$ and $u \vee v$ interchangeably, and write $\mathrm{Log}(x) := \log(\mathrm{e} \vee x)$. For any function class F over a set Ω and $E \subset \Omega$, $F(E) = F|_E$ denotes the projection on (restriction to) E. In line with the common convention in functional analysis, absolute numerical constants will be denoted by letters such as C, c, whose value may change from line to line. Any transformation $\varphi : \mathbb{R} \to \mathbb{R}$ may be applied to a function $f \in \mathbb{R}^\Omega$ via $\varphi(f) := \varphi \circ f$, as well as to $F \subset \mathbb{R}^\Omega$ via $\varphi(F) := \{\varphi(f); f \in F\}$. The sign function thresholds at 0: $\mathrm{sign}(t) = \mathbf{1}[t \geq 0]$.

3. Main Results

Our main results involve upper-bounding the fat-shattering dimension of aggregation rules in terms of the dimensions of the component classes. We begin with the simplest (to present):

Theorem 1 (General F and G that commutes with shifts and truncation) *For $F_1, \ldots, F_k \subseteq \mathbb{R}^\Omega$, and an aggregation rule G that commutes with shifts, (see definition (7)) and commutes with truncation (see definition (21)), we have*

$$\mathrm{fat}_\gamma(G(F_1, \ldots, F_k)) \quad \leq \quad 35 D_\gamma \log^2(126 D_\gamma), \qquad \gamma > 0,$$

where $D_\gamma := \sum_{i=1}^k \mathrm{fat}_\gamma(F_i) > 0$. In the degenerate case where $D_\gamma = 0$, $\mathrm{fat}_\gamma(G) = 0$.

In particular, this result holds for the maximum, minimum, max-min, and median aggregation rules.

Remark. We made no attempt to optimize the constants; these are only provided to give a rough order-of-magnitude sense. In the sequel, we forgo numerical estimates and state the results in terms of unspecified universal constants.

The next result provides an alternative bound based on an entirely different technique:

Theorem 2 (Bounded function classes and Lipschitz aggregations) *For $0 < \varepsilon < \log 2$, $F_1, \ldots, F_k \subseteq [-R, R]^\Omega$, and an aggregation rule G that is L-Lipschitz ($L \geq 1$) in supremum norm (see definition (8)), we have*

$$\mathrm{fat}_\gamma(G(F_1, \ldots, F_k)) \quad \leq \quad C D \,\mathrm{Log}^{1+\varepsilon} \frac{LRk}{\gamma}, \qquad 0 < \gamma/L < R,$$

where

$$D \quad = \quad \sum_{i=1}^k \mathrm{fat}_{c\varepsilon\gamma}(F_i)$$

and $C, c > 0$ are universal constants. In particular, this result holds for natural aggregation rules, such as maximum, minimum, max-min, mean, and median.

Remark. The bounds in Theorems 1 and 2 are, in general, incomparable—and not just because of the unspecified constants in the latter. One notable difference is that Theorem 1 only depends on the shattering scale γ, while Theorem 2 additionally features a (weak) explicit dependence on the aspect ratio R/γ. In particular, Theorem 1 is applicable to semi-bounded affine classes (see Section A.4), while Theorem 2 is not. Still, for fixed R, γ and large k, the latter presents a significant asymptotic improvement over the former.

For the special case of affine functions and hinge-loss affine functions, the technique of Theorem 2 yields a considerably sharper estimate:

Theorem 3 (Dimension-free bound for Lipschitz aggregations of affine functions) *Let $B \subset \mathbb{R}^d$ be the d-dimensional Euclidean unit ball and*

$$F_i = \left\{ x \mapsto w \cdot x + b; \|w\| \vee |b| \leq R_i, w \in \mathbb{R}^d, b \in \mathbb{R} \right\}, \qquad R_i \in \mathbb{R},\ i \in [k], \tag{12}$$

be k collections of R_i-bounded affine functions on $\Omega = B$ and G be an aggregation rule that is L-Lipschitz in supremum norm (see definition (8)). Then

$$\mathrm{fat}_\gamma(G(F_1,\ldots,F_k)) \;\leq\; \frac{CL^2\,\mathrm{Log}(k)}{\gamma^2}\sum_{i=1}^k R_i^2, \qquad 0<\gamma/L<\min_{i\in[k]} R_i, \tag{13}$$

where $C>0$ is a universal constant. Further, if

$$F_i^{\mathrm{Hinge}} = \{(x,y)\mapsto \max\{0,1-yf(x)\}\,;f\in F_i\} \tag{14}$$

is a family of R_i-bounded hinge-loss affine functions for $i\in[k]$ and $G_{\mathrm{Hinge}}\equiv G(F_1^{\mathrm{Hinge}},\ldots,F_k^{\mathrm{Hinge}})$ is an aggregation rule that is L-Lipschitz in supremum norm, then the same bound as in (13) holds for $\mathrm{fat}_\gamma(G_{\mathrm{Hinge}})$.

In particular, this result holds for the maximum, minimum, max-min, mean, and median aggregation rules.

Theorem 3 improves by a log factor the estimate of Fefferman et al. (2016), on the fat-shattering dimension of max-min aggregation (defined in Section 2) of linear functions:[2]

Lemma 4 (Fefferman et al. (2016), Lemma 6) *Let $B\subset\mathbb{R}^d$ be the d-dimensional Euclidean unit ball and*

$$F_{ij} = \left\{x\mapsto w\cdot x; \|w\|\leq\|1\|, w\in\mathbb{R}^d\right\}, \qquad i\in[k], j\in[\ell],$$

be $k\ell$ (identical) linear function classes defined on $\Omega=B$. If $G_{\text{max-min}}$ is the max-min aggregation rule (6)*, then*

$$\mathrm{fat}_\gamma(G_{\text{max-min}}(F_{11},\ldots,F_{k\ell})) \;\leq\; C\frac{k\ell}{\gamma^2}\,\mathrm{Log}^2\left(\frac{k\ell}{\gamma^2}\right),$$

where $C>0$ is a universal constant.

Our Theorem 3 improves the latter by a log factor:

$$\mathrm{fat}_\gamma(G_{\text{max-min}}(F_{11},\ldots,F_{k\ell})) \;\leq\; C\frac{k\ell\log(k\ell)}{\gamma^2}.$$

Corollary 5 (Rademacher complexity for k-Fold Maximum of Affine Functions) *Let F_i be an R_i-bounded affine function class as in (12) or a hinge loss affine function class as in (14), let $G_{\max}$ be the maximum aggregation rule, and let $\tilde{R}=\max_i R_i$, then*

$$\mathcal{R}_n(G_{\max}(F_1,\ldots,F_k)) \leq C\sqrt{\frac{\mathrm{Log}(k)\,\mathrm{Log}^3(\tilde{R}n)\tilde{R}^2\sum_{i=1}^k R_i^2}{n}}.$$

where $\mathcal{R}_n$ is the Rademacher complexity and $C>0$ is a universal constant.

2. The max-min aggregation is shown to be 1-Lipschitz in supremum norm in Lemma 15 of Section A.1.

Corollary 5 improves upon Raviv et al. (2018, Theorem 7). Their upper bound scales linearly with k, whereas ours scales as $\sqrt{k \log k}$.

Note, however, that for linear classes a better bound is known:

Theorem 6 (Zhivotovskiy (2022)) *Let $B \subset \mathbb{R}^d$ be the d-dimensional Euclidean unit ball and*

$$F_i = \left\{ x \mapsto w \cdot x; \|w\| \leq 1, w \in \mathbb{R}^d \right\}, \qquad i \in [k]$$

be k (identical) linear function classes defined on $\Omega = B$. If $G_{\max}$ is the maximum aggregation rule, then

$$\mathcal{R}_n(G_{\max}(F_1, \ldots, F_k)) \leq C \log\left(\frac{n}{k}\right) \sqrt{\frac{k \log k}{n}},$$

where $\mathcal{R}_n$ is the Rademacher complexity and $C > 0$ is a universal constant.

The estimate in Theorem 3 is *dimension-free* in the sense of being independent of d. In applications where a dependence on d is admissible, an optimal bound can be obtained:

Theorem 7 (Dimension-dependent bound for k-fold maximum of affine functions) *Let $\Omega = \mathbb{R}^d$ and $F_i \subset \mathbb{R}^\Omega$ be k (identical) function classes consisting of all real-valued affine functions:*

$$F_i = \left\{ x \mapsto w \cdot x + b; w \in \mathbb{R}^d, b \in \mathbb{R} \right\}, \qquad i \in [k]$$

and let $G_{\max}$ be the k-fold maximum (see definition (4)). Then

$$\mathrm{fat}_\gamma(G_{\max}(F_1, \ldots, F_k)) \quad \leq \quad Cdk \operatorname{Log} k, \qquad \gamma > 0,$$

where $C > 0$ is a universal constant.

The optimality of the upper bound in Theorem 7 is witnessed by the matching lower bound:

Theorem 8 (Dimension-dependent lower bound for k-fold maximum of affine functions) *For $k \geq 1$ and $d \geq 4$, let $F_1 = F_2 = \ldots = F_k$ be the collection of all affine functions over $\Omega = \mathbb{R}^d$ and let $G_{\max}$ be the k-fold maximum (see definition (4)). Then*

$$\mathrm{fat}_\gamma(G_{\max}(F_1, \ldots, F_k)) \quad \geq \quad C \log(k) \sum_{i=1}^{k} \mathrm{fat}_\gamma(F_i) = Cdk \log k, \qquad \gamma > 0,$$

where $C > 0$ is a universal constant.

The scaling argument employed in the proof of Theorem 8 can be invoked to show that the claim continues to hold for $\Omega = B$.

Together, Theorems 7 and 8 show that the dependence on k is optimal.

4. Proofs

We start with upper-bounding the fat-shattering dimension of aggregation rules that commute with shifts (definition (7)) and commute with truncation (defined below), in terms of the dimensions of the component classes.

4.1 Proof of Theorem 1

Partial concept classes and disambiguation. We say that $F^\star \subseteq \{0,1,\star\}^\Omega$ is a *partial* concept class over Ω; this usage is consistent with Alon et al. (2022), while Attias et al. (2019, 2022) used the descriptor *ambiguous*. Define the *disambiguation operator* $\mathscr{D} : \{0,1,\star\} \to 2^{\{0,1\}}$ as

$$\mathscr{D}(0) = \{0\}\,; \quad \mathscr{D}(1) = \{1\}\,; \quad \mathscr{D}(\star) = \{0,1\}\,. \tag{15}$$

For $f^\star \in F^\star$, define its *disambiguation set* $\mathscr{D}(f^\star) \subseteq \{0,1\}^\Omega$ as

$$\mathscr{D}(f^\star) = \Big\{ g \in \{0,1\}^\Omega : \forall x \in \Omega,\ g(x) \in \mathscr{D}(f^\star(x)) \Big\}\,; \tag{16}$$

in words, $\mathscr{D}(f^\star)$ consists of the *total* concepts $g : \Omega \to \{0,1\}$ that agree pointwise with $f^\star$, whenever the latter takes a value in $\{0,1\}$. We say that $\bar{F} \subseteq \{0,1\}^\Omega$ disambiguates $F^\star$ if for all $f^\star \in F^\star$, we have $\bar{F} \cap \mathscr{D}(f^\star) \neq \emptyset$; in words, every $f^\star \in F^\star$ must have a disambiguated representative in $\bar{F}$.[3]

As in Alon et al. (2022); Attias et al. (2022), we say[4] that $S \subset \Omega$ is VC-shattered by $F^\star$ if $F^\star(S) \supseteq \{0,1\}^S$. We write $\mathrm{vc}(F^\star)$ to denote the size of the largest VC-shattered set (possibly, ∞). The obvious relation $\mathrm{vc}(F^\star) \leq \mathrm{vc}(\bar{F})$ always holds between a partial concept class and any of its disambiguations. Alon et al. (2022, Theorem 13) proved the following variant of the Sauer-Shelah-Perles Lemma for partial concept classes:

Lemma 9 (Alon et al. (2022)) *For every $F^\star \subseteq \{0,1,\star\}^\Omega$ with $d = \mathrm{vc}(F^\star) < \infty$ and $|\Omega| < \infty$, there is an $\bar{F}$ disambiguating $F^\star$ such that*

$$|\bar{F}(\Omega)| \ \leq \ (|\Omega|+1)^{(d+1)\log_2 |\Omega| + 2}. \tag{17}$$

For $d > 0$ and $|\Omega| > 1$, this implies the somewhat more wieldy estimate[5]

$$|\bar{F}(\Omega)| \ \leq \ |\Omega|^{7d \log_2 |\Omega|}. \tag{18}$$

We will make use of the elementary fact

$$x \leq A \log_2 x \quad \Longrightarrow \quad x \leq 3A \log(3A), \qquad x, A \geq 1$$

and its corollary

$$y \leq A(\log_2 y)^2 \quad \Longrightarrow \quad y \leq 5A \log^2(18A), \qquad y, A \geq 1. \tag{19}$$

3. Attias et al. (2022) additionally required that $\bar{F} \subseteq \bigcup_{f^\star \in F^\star} \mathscr{D}(f^\star)$, but this is an unnecessary restriction, and does not affect any of the results.
4. Attias et al. (2019) had incorrectly given $F^\star(S) = \{0,1\}^S$ as the shattering condition.
5. The estimate (18) does not appear in Alon et al. (2022), but is an elementary consequence of (17).

Aggregation rules commuting with truncation. Fix $\gamma > 0$ and define the truncation operator $[\cdot]^\star_\gamma : \mathbb{R} \to \{0, 1, \star\}$ as

$$[t]^\star_\gamma \;=\; \begin{cases} 0, & t \le -\gamma \\ 1, & t \ge \gamma \\ \star, & \text{else.} \end{cases} \tag{20}$$

Let $x_i \in \mathbb{R}$, $i \in [k]$. Let the γ-truncation $[x_i]^\star_\gamma \in \{0, 1, \star\}$, and $\bar{x}_i \in \mathscr{D}([x_i]^\star_\gamma) \subseteq \{0, 1\}$ be a disambiguation. We say that an aggregation rule $G : \mathbb{R}^k \to \mathbb{R}$ *commutes with truncation* if for any $\gamma > 0$,

$$G(\bar{x}_1, \ldots, \bar{x}_k) \in \mathscr{D}([G(x_1, \ldots, x_k)]^\star_\gamma) \tag{21}$$

for *all* disambiguations $\bar{x}_i$, $i \in [k]$ (see definitions in (15) and (16)). In Appendix A.1, we show that median and max-min aggregations commute with truncations (Lemmas 16, 17 respectively). Showing it for the maximum and minimum is a simple exercise. We note that the mean aggregation does not satisfy this property.

Proof [of Theorem 1] We follow the basic techniques of discretization and r-shifting, employed in Attias et al. (2019, 2022).

Fix $\gamma > 0$, recall the truncation operator $[\cdot]^\star_\gamma : \mathbb{R} \to \{0, 1, \star\}$ defined in (20). We also define the truncation operator over functions $[\cdot]^\star_\gamma : \mathbb{R}^\Omega \to \{0, 1, \star\}^\Omega$, as $[f]^\star_\gamma = f^\star$ where $f^\star(x) = [f(x)]^\star_\gamma$, for $x \in \Omega$. Observe that for all $F \subseteq \mathbb{R}^\Omega$ and $[F]^\star_\gamma := \left\{[f]^\star_\gamma; f \in F\right\}$, we have $\mathring{\text{fat}}_\gamma(F) = \text{vc}([F]^\star_\gamma)$. Let $G : \mathbb{R}^k \to \mathbb{R}$ be a k-fold aggregation rule and $F_1, \ldots, F_k \subseteq \mathbb{R}^\Omega$ be real-valued function classes. Suppose that some $S = \{x_1, \ldots, x_\ell\} \subset \Omega$ is γ-shattered by $G \equiv G(F_1, \ldots, F_k)$. Proving the claim amounts to upper-bounding ℓ appropriately. By (10), there is an $r \in \mathbb{R}^\Omega$ such that $\text{fat}_\gamma(G) = \mathring{\text{fat}}_\gamma(G - r) = \text{vc}([G - r]^\star_\gamma)$. Put $F'_i := F_i - r$ and since G commutes with r-shift, as defined in (7), we have

$$G' := G(F'_1, \ldots, F'_k) = G(F_1 - r, \ldots, F_k - r) = G(F_1 \ldots, F_k) - r. \tag{22}$$

Hence, S is VC-shattered by $[G']^\star_\gamma$ and

$$v_i := \text{vc}([F'_i]^\star_\gamma) = \mathring{\text{fat}}_\gamma(F'_i) \le \text{fat}_\gamma(F'_i) = \text{fat}_\gamma(F_i), \qquad i \in [k]. \tag{23}$$

Let us assume for now that each $v_i > 0$; in this case, there is no loss of generality in assuming $\ell > 1$. Let $\bar{F}_i$ be a "good" disambiguation of $[F'_i]^\star_\gamma$ on S, as furnished by Lemma 9:

$$|\bar{F}_i(S)| \;\le\; \ell^{7 v_i \log_2 \ell}.$$

Observe that $\bar{G} := G(\bar{F}_1, \ldots, \bar{F}_k)$ is a valid disambiguation of $[G']^\star_\gamma$ since we assume that G commutes with truncation. It follows that

$$2^\ell \;=\; |\bar{G}(S)| \;\le\; \prod_{i=1}^{k} |\bar{F}_i(S)| \;\le\; \ell^{7 \log_2 \ell \sum_{i=1}^k v_i}. \tag{24}$$

Thus, (19) implies that $\ell \le 35(\sum v_i) \log^2(126 \sum v_i)$, and the latter is an upper bound on $\text{vc}(\bar{G})$ — and hence, also on $\text{vc}([G']^\star_\gamma) = \text{fat}_\gamma(G)$. The claim now follows from (23).

If any one given $v_i = 0$, we claim that (24) is unaffected. This is because any $C^\star \subset \{0,1,\star\}^\Omega$ with $\mathrm{vc}(C^\star) = 0$ has a singleton disambiguation $\bar{C} = \{c\}$. Indeed, any given $x \in \Omega$ can receive at most one of $\{0,1\}$ as a label from the members of C (otherwise, it would be shattered, forcing $\mathrm{vc}(C^\star) \geq 1$). If *any* $c^\star \in C^\star$ labels x with 0, then *all* members of $C^\star$ are disambiguated to label x with 0 (and, *mutatis mutandis*, 1). Any x labeled with $\star$ by *every* $c^\star \in C_i^\star$ can be disambiguated arbitrarily (say, to 0). Disambiguating the degenerate $[F_i']_\gamma^\star$ to the singleton $\bar{F}_i(S)$ has no effect on the product in (24).

The foregoing argument continues to hold if more than one $v_i = 0$. In particular, in the degenerate case where $\mathrm{fat}_\gamma(F_1) = \mathrm{fat}_\gamma(F_2) = \ldots = \mathrm{fat}_\gamma(F_k) = 0$, we have $\prod |\bar{F}_i(S)| = 1$, which forces $\ell = 0$. ■

4.2 Proof of Theorem 2

First, we upper bound the covering numbers of Lipschitz aggregations as a function of the covering numbers of the component classes.

Theorem 10 (Covering number of L-Lipschitz aggregations) *Let $t > 0$, $p \in [1,\infty]$, and $F_1, \ldots, F_k \subset L_p(\mu)$. Let G be an aggregation rule that is L-Lipschitz. Then, for all probability measures μ on Ω,*

$$\mathcal{N}(G(F_1,\ldots,F_k), L_p(\mu), t) \;\leq\; \begin{cases} \prod_{i=1}^k \mathcal{N}(F_i, L_p(\mu), t/Lk^{1/p}), & p < \infty \\ \prod_{i=1}^k \mathcal{N}(F_i, L_p(\mu), t/L), & p = \infty. \end{cases}$$

We proceed to the main proof.

Proof [of Theorem 2]. Let $G : \mathbb{R}^k \to \mathbb{R}$ be an aggregation rule that is L-Lipschitz ($L \geq 1$) in supremum norm, as defined in (8), and let $F_1, \ldots, F_k \subseteq [-R,R]^\Omega$ be real-valued function classes. Suppose that some $\Omega_\ell = \{x_1, \ldots, x_\ell\} \subset \Omega = B$ is a maximal set that is γ-shattered by G, let $F_i(\Omega_\ell) = F_i|_{\Omega_\ell}$, and μ_ℓ be the uniform distribution on Ω_ℓ. We upper bound the covering number with the fat-shattering dimension as in Lemma 21 (see Section A.2), with $n = \ell$ and $p = \infty$,

$$\log \mathcal{N}(F_i(\Omega_\ell), L_\infty(\mu_\ell), \gamma) \;\leq\; C v_i \log(R\ell/v_i\gamma) \log^\varepsilon(\ell/v_i), \qquad 0 < \gamma < R,$$

where $v_i = \mathrm{fat}_{c\varepsilon\gamma}(F_i)$. Then Theorem 10 implies that

$$\begin{aligned} \log \mathcal{N}(G(\Omega_\ell), L_\infty(\mu_\ell), \gamma/2) &\leq \sum_{i=1}^k \log \mathcal{N}(F_i(\Omega_\ell), L_\infty(\mu_\ell), \gamma/2L) \\ &\leq C \sum_{i=1}^k v_i \log(LR\ell/v_i\gamma) \log^\varepsilon(\ell/v_i) \\ &\overset{(a)}{\leq} C \sum_{i=1}^k v_i \log^{1+\varepsilon}(LR\ell/v_i\gamma) \\ &\overset{(b)}{\leq} CD \log^{1+\varepsilon} \frac{LR\ell k}{D\gamma}, \end{aligned}$$

where $D := \sum_{i=1}^k v_i$, (a) follows since $R/\gamma > 1$ and assuming $L \geq 1$, and (b) follows by the concavity of $x \log^{1+\varepsilon}(u/x)$ (see Lemma 28 in Section A.5). We can assume $\ell \geq 2$ without loss of generality. Combining the monotonicity of the covering number (see (11)), a lower bound on the covering number in terms of the fat-shattering dimension (see Lemma 18 in Section A.2), and the fact the Ω_ℓ is a maximal set that is γ-shattered by G yields

$$\log \mathcal{N}(G(\Omega_\ell), L_\infty(\mu_\ell), \gamma/2) \quad \geq \quad C \operatorname{fat}_\gamma(G) = C\ell,$$

whence

$$\ell \quad \leq \quad CD \log^{1+\varepsilon} \frac{LR\ell k}{D\gamma}.$$

Using the elementary fact

$$x \leq A \operatorname{Log}^{1+\varepsilon} x \quad \Longrightarrow \quad x \leq cA \operatorname{Log}^{1+\varepsilon} A \qquad x, A \geq 1$$

(with $x = LR\ell k/D\gamma$ and $A = cLRk/\gamma$), we get

$$\ell \quad \leq \quad CD \operatorname{Log}^{1+\varepsilon} \frac{LRk}{\gamma},$$

which implies the claim. ■

4.3 Proof of Theorem 3

We use the notation and results from the Appendix, and in particular, from Section A.3.

Proof [of Theorem 3] A bound of this form for the k-fold maximum aggregation was claimed in Kontorovich (2018), however the argument there was flawed, see Section 5.

Let $G : \mathbb{R}^k \to \mathbb{R}$ be an aggregation rule that is L-Lipschitz in supremum norm, as defined in (8), and let $F_1, \ldots, F_k$ be bounded affine function classes, as defined in (12). Suppose that some $\Omega_\ell = \{x_1, \ldots, x_\ell\} \subset \Omega = B$ is a maximal set that is γ-shattered by G, let $F_i(\Omega_\ell) = F_i|_{\Omega_\ell}$, and μ_ℓ be the uniform distribution on Ω_ℓ. We upper bound the covering number as in Lemma 23 (with $m = \ell$),

$$\log \mathcal{N}(F_i(\Omega_\ell), L_\infty(\mu_\ell), \gamma) \quad \leq \quad C \frac{R_i^2}{\gamma^2} \operatorname{Log} \frac{\ell \gamma^2}{R_i^2}, \qquad 0 < \gamma < R_i.$$

Denote $v_i := L^2 R_i^2/\gamma^2$, and consider the L_∞ covering number of $F_i(\Omega_\ell)$ at scale γ/L:

$$\log \mathcal{N}(F_i(\Omega_\ell), L_\infty(\mu_\ell), \gamma/L) \quad \leq \quad C v_i \operatorname{Log} \frac{\ell}{v_i}.$$

Then Theorem 10 implies that

$$\begin{aligned}
\log \mathcal{N}(G(\Omega_\ell), L_\infty(\mu_\ell), \gamma/2) \quad &\leq \quad \sum_{i=1}^k \log \mathcal{N}(F_i(\Omega_\ell), L_\infty(\mu_\ell), \gamma/2L) \\
&\leq \quad C \sum_{i=1}^k v_i \operatorname{Log} \frac{\ell}{v_i} \\
&\overset{(a)}{\leq} \quad CD \operatorname{Log} \frac{k\ell}{D},
\end{aligned}$$

where $D := \sum_{i=1}^{k} v_i$ and (a) follows by the concavity of $x \log(u/x)$ (see Corollary 27 in Section A.5). Combining the monotonicity of the covering number (see (11)), a lower bound on the covering number in terms of the fat-shattering dimension (see Lemma 18 in Section A.2), and the fact the Ω_ℓ is a maximal set that is γ-shattered by G yields

$$\log \mathcal{N}(G(\Omega_\ell), L_\infty(\mu_\ell), \gamma/2) \quad \geq \quad C \operatorname{fat}_\gamma(G) = C\ell,$$

whence

$$\ell \quad \leq \quad CD \operatorname{Log} \frac{k\ell}{D}.$$

Using the elementary fact

$$x \leq A \operatorname{Log} x \quad \Longrightarrow \quad x \leq cA \operatorname{Log} A, \qquad x, A \geq 1$$

(with $x = k\ell/D$ and $A = ck$) we get $\ell \leq cD \operatorname{Log} k$, which implies the claim.

The result can easily be generalized to hinge-loss affine classes. Let F_i be an affine function class as in (12), define F_i' as the function class on $B \times \{-1, 1\}$ given by $F_i' = \{(x, y) \mapsto yf(x); f \in F_i\}$, and the *hinge-loss affine class* F_i^{Hinge} as the function class on $B \times \{-1, 1\}$ given by $F_i^{\text{Hinge}} = \{(x, y) \mapsto \max\{0, 1 - f(x, y)\}; f \in F_i'\}$. One first observes that the restriction of F_i' to any $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, as a body in $\mathbb{R}^n$, is identical to the restriction of F_i to $\{x_1, \ldots, x_n\}$. Interpreting F_i^{Hinge} as a 2-fold maximum over the singleton class $H = \{h \equiv 0\}$ and the bounded affine class F_i' lets us invoke Theorem 10 to argue that F_i and F_i^{Hinge} have the same L_∞ covering numbers. Hence, the argument we deployed here to establish (13) for affine classes also applies to k-fold L-Lipschitz aggregations hinge-loss classes. ■

4.4 Proof of Corollary 5

Proof [of Corollary 5] Raviv et al. (2018, Theorem 7) upper-bounded the Rademacher complexity of the maximum aggregation of k hinge loss affine functions by $k/\sqrt{n}$.

For R_i-bounded affine functions or hinge loss affine functions, the analysis above, combined with the calculation in Kontorovich (2018) yields a bound of $O\left(\sqrt{\frac{\operatorname{Log}(k) \operatorname{Log}^3(n) \sum_{i=1}^{k} R_i^2}{n}}\right)$. For completeness, we provide the full proof.

Let $G_{\max} : \mathbb{R}^k \to \mathbb{R}$ be the k-fold maximum aggregation rule, as defined in (4), and let $F_1, \ldots, F_k \subseteq \mathbb{R}^\Omega$ be R_i-bounded affine function classes as in (12) or hinge loss affine function classes as in (14). Since this aggregation is 1-Lipschitz in the supremum norm, Theorem 3 implies that

$$\operatorname{fat}_\gamma(G_{\max}) \quad \leq \quad \frac{C \operatorname{Log}(k)}{\gamma^2} \sum_{i=1}^{k} R_i^2, \qquad 0 < \gamma < \min_{i \in [k]} R_i,$$

where $C > 0$ is a universal constant.

From fat-shattering to Rademacher. The fat-shattering estimate above can be used to upper-bound the Rademacher complexity by converting the former to a covering number bound and plugging it into Dudley's chaining integral (Dudley, 1967):

$$\mathcal{R}_n(F) \leq \inf_{\alpha \geq 0} \left(4\alpha + 12 \int_\alpha^\infty \sqrt{\frac{\log \mathcal{N}(F, \|\cdot\|_2, t)}{n}} dt \right), \tag{25}$$

where $\mathcal{N}(\cdot)$ are the L_2 covering numbers.

It remains to bound the covering numbers. A simple way of doing so is to invoke Lemmas 2.6, 3.2, and 3.3 in Alon et al. (1997) — but this incurs superfluous logarithmic factors in n. Instead, we use the sharper estimate of Mendelson and Vershynin (2003), stated here in Lemma 19. Putting $\tilde{R} = \max_i R_i$, the latter yields

$$\begin{aligned}
\mathcal{R}_n(G_{\max}) &\leq \inf_{\alpha \geq 0} \left(4\alpha + 12 \int_\alpha^1 \sqrt{\frac{\log \mathcal{N}(G_{\max}, \|\cdot\|_2, t)}{n}} dt \right) \\
&\leq \inf_{\alpha \geq 0} \left(4\alpha + 12c' \int_\alpha^1 \sqrt{\frac{\mathrm{fat}_{ct/\tilde{R}}(G_{\max}) \log \frac{2\tilde{R}}{t}}{n}} dt \right) \\
&\leq \inf_{\alpha \geq 0} \left(4\alpha + 12c'' \sqrt{\frac{\mathrm{Log}(k) \sum_{i=1}^k R_i^2}{n}} \int_\alpha^1 \frac{\tilde{R}}{t} \sqrt{\log \frac{2\tilde{R}}{t}} dt \right).
\end{aligned}$$

Now

$$\int_\alpha^1 \frac{\tilde{R}}{t} \sqrt{\log \frac{2\tilde{R}}{t}} dt = \frac{2\tilde{R}}{3} \left(\log(2\tilde{R}/\alpha)^{3/2} - (\log 2\tilde{R})^{3/2} \right)$$

and choosing $\alpha = 1/\sqrt{n}$ yields

$$\begin{aligned}
\mathcal{R}_n(G_{\max}) &\leq \frac{4}{\sqrt{n}} + 12c'' \sqrt{\frac{\mathrm{Log}(k) \sum_{i=1}^k R_i^2}{n}} \frac{2\tilde{R}}{3} \left(\log(2\tilde{R}\sqrt{n})^{3/2} - (\log 2\tilde{R})^{3/2} \right) \\
&= O\left(\sqrt{\frac{\mathrm{Log}(k) \log^3(\tilde{R}n) \tilde{R}^2 \sum_{i=1}^k R_i^2}{n}} \right).
\end{aligned}$$

■

4.5 Proof of Theorem 7

Proof [of Theorem 7] Let $G_{\max} : \mathbb{R}^k \to \mathbb{R}$ be the k-fold maximum aggregation rule, as defined in (4), and let $F_1, \ldots, F_k \subseteq \mathbb{R}^\Omega$ be identical function classes consisting of all real-valued affine functions. Note that $G_{\max}$ is an aggregation that commutes with shift, as defined in (7).

By (10), there is an $r \in \mathbb{R}^\Omega$ such that $\mathrm{fat}_\gamma(G_{\max}) = \mathring{\mathrm{fat}}_\gamma(G_{\max} - r)$. As in (22), put $F_i' := F_i - r$ and $G'_{\max} := G_{\max} - r = G_{\max}(F_1', \ldots, F_k')$. Define $\bar{G}_{\max} = \mathrm{sign}(G'_{\max})$ and $\bar{F}_i = \mathrm{sign}(F_i')$.

Since sign and max commute, we have $\bar{G}_{\max} = \max_{i\in[k]}(\bar{F}_i)$. We claim that

$$\mathring{\mathrm{fat}}_\gamma(G'_{\max}) \quad \leq \quad \mathrm{vc}(\bar{G}_{\max}). \tag{26}$$

Indeed, any $S \subset \Omega$ that is γ-shattered with shift $r = 0$ by any $G \subset \mathbb{R}^\Omega$ is also VC-shattered by $\mathrm{sign}(G)$. (See Section 4.1, and notice that the converse implication—and the reverse inequality—do not hold.) It holds that

$$d + 1 \overset{(a)}{=} \mathrm{vc}(\bar{F}_i) \overset{(b)}{=} \mathring{\mathrm{fat}}_\gamma(F_i') \overset{(c)}{=} \mathrm{fat}_\gamma(F_i') \overset{(d)}{=} \mathrm{fat}_\gamma(F_i),$$

where (a) follows from a standard argument (e.g., Mohri et al. (2012, Example 3.2)), (b) holds because any $S \subset \mathbb{R}^d$ that is VC-shattered by $\mathrm{sign}(F_i')$ is also γ-shattered by F_i' with shift $r = 0$, (c) follows from Lemma 24, since the class satisfies the closure property (34), and (d) holds since the shattering remains the same for the shifted class.

Now the argument of Blumer et al. (1989, Lemma 3.2.3) applies:

$$\mathrm{vc}(\bar{G}_{\max}) \quad \leq \quad 2(d+1)k\log(3k) \tag{27}$$

(this holds for any k-fold aggregation function, not just the maximum). Combining (26) with (27) proves the claim. ■

4.6 Proof of Theorem 8

Proof [of Theorem 8] It follows from Mohri et al. (2012, Example 3.2) that $\mathrm{vc}(\mathrm{sign}(F_i)) = d + 1$. Since F_i is closed under scalar multiplication, a scaling argument shows that any $S \subset \mathbb{R}^d$ that is VC-shattered by $\mathrm{sign}(F_i)$ is also γ-shattered by F_i with shift $r = 0$, whence $\mathring{\mathrm{fat}}_\gamma(F_i) = d + 1$ for all $\gamma > 0$; invoking Lemma 24 extends this to $\mathrm{fat}_\gamma(F_i)$ as well. Now Csikós et al. (2019, Theorem 1) shows that the k-fold unions of half-spaces necessarily shatter some set $S \subset \mathbb{R}^d$ of size at least $cdk\log k$. Since union is a special case of the max operator, and the latter commutes with sign, the scaling argument shows that this S is γ-shattered by $G_{\max}$ with shift $r = 0$. Hence, $\mathrm{fat}_\gamma(G_{\max}) \geq \mathring{\mathrm{fat}}_\gamma(G_{\max}) \geq |S|$, which proves the claim. ■

5. Discussion

In this paper, we proved upper and lower bounds on the fat-shattering dimension of aggregation rules as a function of the fat-shattering dimension of the component classes. We leave some remaining gaps for future work. First, for aggregation rules that commute with shifts and commute with truncation, assuming $\mathrm{fat}_\gamma(F_i) \leq d$, for $1 \leq i \leq k$, we show in Theorem 1 that

$$\mathrm{fat}_\gamma(G(F_1, \ldots, F_k)) \leq Cdk\,\mathrm{Log}^2(dk), \qquad \gamma > 0,$$

$C > 0$ is a universal constant. We pose the following

Open problem. Let G be an aggregation rule with the properties as in Theorem 1. Is it the case that for all $F_i \subseteq \mathbb{R}^\Omega$ with $\mathrm{fat}_\gamma(F_i) \leq d$, $i \in [k]$, we have

$$\mathrm{fat}_\gamma(G(F_1, \ldots, F_k)) \quad \leq \quad Cdk \operatorname{Log}(k), \qquad \gamma > 0, \tag{28}$$

for some universal $C > 0$?

In light of Theorem 8, this is the best one could hope for in general. We pose also the following conjecture about bounded affine functions.

Conjecture 11 *Theorem 3 is tight up to constants. For R_i-bounded affine functions and an aggregation rule G that is 1-Lipschitz in supremum norm,*

$$\mathrm{fat}_\gamma(G(F_1, \ldots, F_k)) \quad \geq \quad \frac{C \operatorname{Log}(k)}{\gamma^2} \sum_{i=1}^{k} R_i^2, \qquad 0 < \gamma < \min_{i \in [k]} R_i, \tag{29}$$

where $C > 0$ is a universal constant.

Throughout the paper, we mentioned several mistaken claims in the literature. In this section, we briefly discuss the nature of these mistakes—which are, in a sense, variations on the same kind of error. We begin with Attias et al. (2019, Lemma 14), which incorrectly claimed that any partial function class $F^\star$ has a disambiguation $\bar{F}$ such that $\mathrm{vc}(\bar{F}) \leq \mathrm{vc}(F^\star)$ (see Section 4.1 for the definitions). The mistake was pointed out to us by Yann Guermeur, and later, Alon et al. (2022, Theorem 11) showed that there exist partial classes $F^\star$ with $\mathrm{vc}(F^\star) = 1$ for which every disambiguation $\bar{F}$ has $\mathrm{vc}(\bar{F}) = \infty$.

Kontorovich (2018) attempted to prove the bound stated in our Theorem 3 (up to constants, and only for linear classes). The argument proceeded via a reduction to the Boolean case, as in our proof of Theorem 7. It was correctly observed that if, say, some finite $S \subset \Omega$ is 1-shattered by F_i with shift $r = 0$, then it is also VC-shattered by $\mathrm{sign}(F_i)$. Neglected was the fact that $\mathrm{sign}(F_i)$ might shatter additional points in $\Omega \setminus S$—and, in sufficiently high dimension, it necessarily will. The crux of the matter is that (26) holds in the dimension-dependent but not the dimension-free setting; again, this may be seen as a variant of the disambiguation mistake.

Finally, the proof of Hanneke and Kontorovich (2019, Lemma 6) claims, in the first display, that the shattered set can be classified with large margin, which is incorrect — yet another variant of mistaken disambiguation.

Acknowledgments

We thank Steve Hanneke and Ramon van Handel for very helpful discussions; the latter, in particular, patiently explained to us how to prove Lemma 23. Roman Vershynin kindly gave us permission to share his example in Remark 20. We deeply thank the anonymous reviewers for their insightful comments and suggestions.

This research was partially supported by the Israel Science Foundation (grant No. 1602/19), an Amazon Research Award, the Ben-Gurion University Data Science Research Center, Cyber Security Research Center, Prime Minister's Office, and the Vatat Scholarship from the Israeli Council for Higher Education.

References

Noga Alon, Shai Ben-David, Nicolò Cesa-Bianchi, and David Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM*, 44(4):615–631, 1997.

Noga Alon, Amos Beimel, Shay Moran, and Uri Stemmer. Closure properties for private classification and online prediction. In *Conference on Learning Theory*, pages 119–152. PMLR, 2020.

Noga Alon, Steve Hanneke, Ron Holzman, and Shay Moran. A theory of PAC learnability of partial concept classes. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 658–671, 2022.

Gautier Appert and Olivier Catoni. New bounds for k-means and information k-means. *arXiv preprint arXiv:2101.05728*, 2021.

S. Artstein, V. Milman, and S. J. Szarek. Duality of metric entropy. *Annals of Mathematics*, 159(3):1313–1328, 2004.

Idan Attias and Steve Hanneke. Adversarially robust PAC learnability of real-valued functions. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 1172–1199, 2023.

Idan Attias, Aryeh Kontorovich, and Yishay Mansour. Improved generalization bounds for robust learning. In *Algorithmic Learning Theory, ALT 2019*, volume 98 of *Proceedings of Machine Learning Research*, pages 162–183. PMLR, 2019.

Idan Attias, Aryeh Kontorovich, and Yishay Mansour. Improved generalization bounds for adversarially robust learning. *The Journal of Machine Learning Research*, 23(1):7897–7927, 2022.

Peter Bartlett and John Shawe-Taylor. *Generalization performance of support vector machines and other pattern classifiers*, pages 43–54. MIT Press, Cambridge, MA, USA, 1999. ISBN 0-262-19416-3.

Peter L. Bartlett and Philip M. Long. Prediction, learning, uniform convergence, and scale-sensitive dimensions. *J. Comput. Syst. Sci.*, 56(2):174–190, 1998.

Eric B. Baum and David Haussler. What size net gives valid generalization? *Neural Comput.*, 1(1):151–160, 1989.

Gérard Biau, Luc Devroye, and Gábor Lugosi. On the performance of clustering in hilbert spaces. *IEEE Transactions on Information Theory*, 54(2):781–790, 2008.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *J. Assoc. Comput. Mach.*, 36(4):929–965, 1989.

Leo Breiman. Bagging predictors. *Machine learning*, 24:123–140, 1996.

Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

Doron Cohen, Aryeh Kontorovich, Aaron Koolyk, and Geoffrey Wolfer. Dimension-free empirical entropy estimation. *IEEE Transactions on Information Theory*, 69(5):3190–3202, 2023. doi: 10.1109/TIT.2022.3232739.

Mónika Csikós, Nabil H. Mustafa, and Andrey Kupavskii. Tight lower bounds on the vc-dimension of geometric set systems. *J. Mach. Learn. Res.*, 20:81:1–81:8, 2019.

Hubert Haoyang Duan. *Bounding the Fat Shattering Dimension of a Composition Function Class Built Using a Continuous Logic Connective*. PhD thesis, University of Waterloo, 2012.

Richard M Dudley. The sizes of compact subsets of hilbert space and continuity of gaussian processes. *Journal of Functional Analysis*, 1(3):290–330, 1967.

David Eisenstat. k-fold unions of low-dimensional concept classes. *Inf. Process. Lett.*, 109 (23-24):1232–1234, 2009.

David Eisenstat and Dana Angluin. The VC dimension of k-fold union. *Inf. Process. Lett.*, 101(5):181–184, 2007.

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

Dylan J. Foster and Alexander Rakhlin. ℓ_∞ vector contraction for rademacher complexity. *CoRR*, abs/1911.06468, 2019.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

Badih Ghazi, Noah Golowich, Ravi Kumar, and Pasin Manurangsi. Near-tight closure bounds for the littlestone and threshold dimensions. In *Algorithmic Learning Theory*, pages 686–696. PMLR, 2021.

Lee-Ad Gottlieb, Aryeh Kontorovich, and Robert Krauthgamer. Efficient classification for metric data (extended abstract: COLT 2010). *IEEE Transactions on Information Theory*, 60(9):5750–5759, 2014.

Lee-Ad Gottlieb, Eran Kaufman, Aryeh Kontorovich, and Gabriel Nivasch. Learning convex polytopes with margin. In *Neural Information Processing Systems (NIPS)*, 2018.

Steve Hanneke and Aryeh Kontorovich. Optimality of SVM: novel proofs and tighter bounds. *Theor. Comput. Sci.*, 796:99–113, 2019.

Michael J. Kearns and Robert E. Schapire. Efficient distribution-free learning of probabilistic concepts. *J. Comput. Syst. Sci.*, 48(3):464–497, 1994.

Balázs Kégl. Robust regression by boosting the median. In *Learning Theory and Kernel Machines: 16th Annual Conference on Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003. Proceedings*, pages 258–272. Springer, 2003.

Yegor Klochkov, Alexey Kroshnin, and Nikita Zhivotovskiy. Robust k-means clustering for distributions with two moments. *The Annals of Statistics*, 49(4):2206–2230, 2021.

Aryeh Kontorovich. Rademacher complexity of k-fold maxima of hyperplanes. 2018.

S. Mendelson and R. Vershynin. Entropy and the combinatorial dimension. *Invent. Math.*, 152(1):37–55, 2003.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations Of Machine Learning*. The MIT Press, 2012.

Dolev Raviv, Tamir Hazan, and Margarita Osadchy. Hinge-minimax learner for the ensemble of hyperplanes. *J. Mach. Learn. Res.*, 19:62:1–62:30, 2018.

M. Rudelson and R. Vershynin. Combinatorics of random processes and sections of convex bodies. *Annals of Mathematics*, 164(2):603–648, 2006.

Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. *Advances in neural information processing systems*, 23, 2010.

Michel Talagrand. Vapnik–chervonenkis type conditions and uniform donsker classes of functions. *The Annals of Probability*, 31(3):1565–1582, 2003.

Roman Vershynin. *High-dimensional probability*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2018.

Roman Vershynin, 2021. Private communication.

Tong Zhang. Covering number bounds of certain regularized linear function classes. *The Journal of Machine Learning Research*, 2:527–550, 2002.

Nikita Zhivotovskiy. A bound for k-fold maximum. 2022.

Appendix A. Auxiliary results

A.1 Properties of Aggregation Rules

Lemma 12 *If $G : \mathbb{R}^k \to \mathbb{R}$ is L-Lipschitz under $\|\cdot\|_p$, then $G : (\mathbb{R}^\Omega)^k \to \mathbb{R}^\Omega$ is L-Lipschitz in $\|\cdot\|_{L_p^{(k)}(\mu)}$.*

Proof

$$\begin{aligned}\left\|G(f_1,\dots,f_k) - G(f_1',\dots,f_k')\right\|_{L_p(\mu)}^p &= \int_\Omega |G(f_1,\dots,f_k)(x) - G(f_1',\dots,f_k')(x)|^p \mathrm{d}\mu(x)\\ &= \int_\Omega |G(f_1(x),\dots,f_k(x)) - G(f_1'(x),\dots,f_k'(x))|^p \mathrm{d}\mu(x)\\ &\le \int_\Omega L^p \left\|(f_1(x),\dots,f_k(x)) - (f_1'(x),\dots,f_k'(x))\right\|_p^p \mathrm{d}\mu(x),\end{aligned}$$

where the inequality follows from the assumption that $G : \mathbb{R}^k \to \mathbb{R}$ is L-Lipschitz in $\|\cdot\|_p$. This proves

$$\left\|G(f_1,\dots,f_k) - G(f_1',\dots,f_k')\right\|_{L_p(\mu)} \le L \left\|(f_1,\dots,f_k) - (f_1',\dots,f_k')\right\|_{L_p^{(k)}(\mu)},$$

and hence the claim. ■

Proof [of Theorem 10] Suppose $p < \infty$, and let $g = G(f_1,\dots,f_k) \in G(F_1,\dots,F_k)$. For each $i \in [k]$, let $\hat{F}_i \subset F_i$ be a $t/Lk^{1/p}$-cover of F_i. Let each f_i be "$t/Lk^{1/p}$-covered" by some $\hat{f}_i \in \hat{F}_i$, in the sense that $\left\|f_i - \hat{f}_i\right\|_{L_p(\mu)} \le t/Lk^{1/p}$. Assuming that $G : \mathbb{R}^k \to \mathbb{R}$ is L-Lipschitz in $\|\cdot\|_p$, Lemma 12 implies that $G : (\mathbb{R}^\Omega)^k \to \mathbb{R}^\Omega$ is L-Lipschitz in $\|\cdot\|_{L_p^{(k)}(\mu)}$. Then it follows that g is t-covered by $G(\hat{f}_1,\dots,\hat{f}_k)$, since

$$\begin{aligned}\left\|G(f_1,\dots,f_k) - G(\hat{f}_1,\dots,\hat{f}_k)\right\|_{L_p(\mu)}^p &\le L^p \left\|(f_1,\dots,f_k) - (\hat{f}_1,\dots,\hat{f}_k)\right\|_{L_p^{(k)}(\mu)}^p\\ &= L^p \int_\Omega \left\|(f_1(x),\dots,f_k(x)) - (\hat{f}_1(x),\dots,\hat{f}_k(x))\right\|_p^p \mathrm{d}\mu(x)\\ &= L^p \int_\Omega \sum_{i=1}^k \left|f_i(x) - \hat{f}_i(x)\right|^p \mathrm{d}\mu(x)\\ &= L^p \sum_{i=1}^k \int_\Omega \left|f_i(x) - \hat{f}_i(x)\right|^p \mathrm{d}\mu(x)\\ &= L^p \sum_{i=1}^k \left\|f_i - \hat{f}_i\right\|_{L_p(\mu)}^p\\ &\le L^p k \left(\frac{t}{Lk^{1/p}}\right)^p\\ &= t^p,\end{aligned}$$

and so $\left\|G(f_1,\ldots,f_k) - G(\hat{f}_1,\ldots,\hat{f}_k)\right\|_{L_p(\mu)} \le t$.

We conclude that $G(F_1,\ldots,F_k)$ has a t-cover of size $|\hat{F}_1 \times \hat{F}_2 \times \ldots \times \hat{F}_k|$, which proves the claim. The case $p=\infty$ is proved analogously (or, alternatively, as a limiting case of $p<\infty$). ■

We show that natural aggregations are Lipschitz in $\|\cdot\|_p$ norms, $p \in [1,\infty)$, and in supremum norm. The following facts are elementary:

$$|a \vee b - c \vee d| \le |a-c| \vee |b-d|, \qquad a,b,c,d \in \mathbb{R}; \tag{30}$$
$$|a \wedge b - c \wedge d| \le |a-c| \vee |b-d|, \qquad a,b,c,d \in \mathbb{R}, \tag{31}$$

where $s \vee t := \max\{s,t\}$ and $s \wedge t := \min\{s,t\}$.

Lemma 13 (Maximum aggregation is 1-Lipschitz) *Let $G_{\max} : \mathbb{R}^k \to \mathbb{R}$ be the maximum aggregation, then for any $x, x' \in \mathbb{R}^k$ and $p \in [1,\infty]$,*

$$\left|G_{\max}(x) - G_{\max}(x')\right| \le \left\|x - x'\right\|_p.$$

Proof For $k=2$ and $p=\infty$, the claim follows from the stronger, pointwise inequality (30). The proof follows by simple induction on k. Since $\|\cdot\|_\infty \le \|\cdot\|_p$, we conclude the proof for $p \in [1,\infty]$. ■

Lemma 14 (Median aggregation is 1-Lipschitz) *Let $G_{\mathrm{med}} : \mathbb{R}^k \to \mathbb{R}$ be the median aggregation, then for any $x, x' \in \mathbb{R}^k$ and $p \in [1,\infty]$,*

$$\left|G_{\mathrm{med}}(x) - G_{\mathrm{med}}(x')\right| \le \left\|x - x'\right\|_p.$$

Proof Denote by $x_{(1)},\ldots,x_{(k)}$ the ascending order of a sequence $x_1,\ldots,x_k$, that is, $x_{(1)} \le \cdots \le x_{(k)}$. For all $x, x' \in \mathbb{R}^k$ we have

$$\begin{aligned}\left|G_{\mathrm{med}}(x_1,\ldots,x_k) - G_{\mathrm{med}}(x'_1,\ldots,x'_k)\right| &= \left|x_{(\lceil k/2\rceil)} - x'_{(\lceil k/2\rceil)}\right| \\ &\le \max_{i\in[k]}\left|x_{(i)} - x'_{(i)}\right| \le \max_{i\in[k]}\left|x_i - x'_i\right|,\end{aligned}$$

where the last inequality follows from Cohen et al. (2023, Eq. (16))[6] Since $\|\cdot\|_\infty \le \|\cdot\|_p$, we conclude the proof for $p \in [1,\infty]$. ■

Lemma 15 (Max-Min aggregation is 1-Lipschitz) *Let $G_{\text{max-min}} : \mathbb{R}^{k\times\ell} \to \mathbb{R}$ be the max-min aggregation, then for any $x, x' \in \mathbb{R}^{k\times\ell}$ and $p \in [1,\infty]$,*

$$\left|G_{\text{max-min}}(x) - G_{\text{max-min}}(x')\right| \le \left\|x - x'\right\|_p.$$

6. stated there for distributions but true for all vectors, by the same argument

Proof The inequalities (30), (31) imply that the k-fold max and min aggregations are both 1-Lipschitz with respect to $\|\cdot\|_\infty$. Hence, for all $x, y \in \mathbb{R}^{k\times\ell}$, we have

$$\left|\min_{i\in[k]} x_{ij} - \min_{i\in[k]} y_{ij}\right| \le \max_{i\in[k]} |x_{ij} - y_{ij}|, \qquad j \in [\ell]$$

and further,

$$\left|\max_{j\in[\ell]}\min_{i\in[k]} x_{ij} - \max_{j\in[\ell]}\min_{i\in[k]} y_{ij}\right| \le \max_{j\in[\ell]}\max_{i\in[k]} |x_{ij} - y_{ij}|.$$

This proves that $|G_{\text{max-min}}(x) - G_{\text{max-min}}(x')| \le \|x - x'\|_\infty$. Since $\|\cdot\|_\infty \le \|\cdot\|_p$, the claim holds for all $p \in [1, \infty]$. ■

Lemma 16 (Median aggregation commutes with truncation) *Let $G_{\text{med}} : \mathbb{R}^k \to \mathbb{R}$ be the median aggregation, then G_{med} commutes with truncation. That is, for any $\gamma > 0$ and $x \in \mathbb{R}^d$,*

$$G_{\text{med}}(\bar{x}_1, \dots, \bar{x}_k) \in \mathscr{D}([G_{\text{med}}(x_1, \dots, x_k)]^\star_\gamma)$$

for all disambiguations $\bar{x}_i \in \mathscr{D}([x_i]^\star_\gamma)$, $i \in [k]$.

Proof Fix $\gamma > 0$ and denote by $x_{(1)}, \dots, x_{(k)}$ the ascending order of a sequence $x_1, \dots, x_k$. Now for $\bar{x}_i \in \mathscr{D}([x_i]^\star_\gamma) \subseteq \{0, 1\}$, our definition of the median (5) implies that $G_{\text{med}}(\bar{x}_1, \dots, \bar{x}_k) \in \{0, 1\}$. It remains to perform an exhaustive verification of the possible cases.

If $[G_{\text{med}}(x_1, \dots, x_k)]^\star_\gamma = \star$ then any value in $\{0, 1\}$ is valid. If $[G_{\text{med}}(x_1, \dots, x_k)]^\star_\gamma = 0$ it means that $G_{\text{med}}(x_1, \dots, x_k)$ outputs a value smaller than $-\gamma$, which means that at least half of inputs $x_1, \dots, x_k$ have a value smaller than $-\gamma$. Let these values be $x_{(1)}, \dots, x_{(m)}$ where $m \ge k/2$. We have $[x_{(j)}]^\star_\gamma = 0$ for $j \in [m]$ and $[x_{(\ell)}]^\star_\gamma \subseteq \{\star, 1\}$ for $\ell \in \{m+1, \dots, k\}$. For any disambiguation $\bar{x}_{(\ell)}$, $G_{\text{med}}(\bar{x}_1, \dots, \bar{x}_k)$ would still output 0 and is a valid disambiguation of $[G_{\text{med}}(x_1, \dots, x_k)]^\star_\gamma$. The case $[G_{\text{med}}(x_1, \dots, x_k)]^\star_\gamma = 1$ follows from the same argument. ■

Lemma 17 (Max-Min aggregation commutes with truncation) *Let $G_{\text{max-min}} : \mathbb{R}^{k\times\ell} \to \mathbb{R}$ be the max-min aggregation, then $G_{\text{max-min}}$ commutes with truncation. That is, for any $\gamma > 0$ and $x \in \mathbb{R}^{k\times\ell}$,*

$$G_{\text{max-min}}(\bar{x}_{11}, \dots, \bar{x}_{k\ell}) \in \mathscr{D}([G_{\text{max-min}}(x_{11}, \dots, x_{k\ell})]^\star_\gamma),$$

for all disambiguations $\bar{x}_{ij} \in \mathscr{D}([x_{ij}]^\star_\gamma)$, $i \in [k], j \in [\ell]$.

Proof Fix $\gamma > 0$. For any $i \in [k]$ denote by $x_{i(1)}, \dots, x_{i(\ell)}$ the ascending order of a sequence $x_{i1}, \dots, x_{i\ell}$. We assume $\bar{x}_{ij} \in \mathscr{D}([x_{ij}]^\star_\gamma) \subseteq \{0, 1\}$ and $G_{\text{max-min}}(\bar{x}_{11}, \dots, \bar{x}_{k\ell})$ outputs a value in $\{0, 1\}$ by our definition of the max-min. We check all possible outputs of $[G_{\text{max-min}}(x_{11}, \dots, x_{k\ell})]^\star_\gamma$ and verify that $G_{\text{max-min}}(\bar{x}_{11}, \dots, \bar{x}_{k\ell})$ is a valid disambiguation.

If $[G_{\text{max-min}}(x_{11}, \dots, x_{k\ell})]^\star_\gamma = \star$ then any value in $\{0, 1\}$ is valid. If $[G_{\text{max-min}}(x_{11}, \dots, x_{k\ell})]^\star_\gamma = 0$ it means that $G_{\text{max-min}}(x_{11}, \dots, x_{k\ell})$ outputs a value smaller than $-\gamma$. This means that

all values that minimize each row $x_{1(1)}, \ldots, x_{k(1)}$ are smaller than $-\gamma$ since the maximum of them is smaller than $-\gamma$. We have $[x_{i(1)}]^\star_\gamma = 0$ for $i \in [k]$. For any disambiguation $\bar{x}_{ij}$ $G_{\text{max-min}}(\bar{x}_{11}, \ldots, \bar{x}_{k\ell})$ would still output 0 and is a valid disambiguation of $[G_{\text{max-min}}(x_{11}, \ldots, x_{k\ell})]^\star_\gamma$. The case$[G_{\text{max-min}}(x_{11}, \ldots, x_{k\ell})]^\star_\gamma = 1$ follows from the same argument. ■

A.2 Covering numbers and the fat-shattering dimension

In this section, we summarize some known results connecting the covering numbers of a bounded function class to its fat-shattering dimension.

Lemma 18 (Talagrand (2003), Proposition 1.4) *For any $F \subseteq [-R, R]^\Omega$, there exists a probability measure μ on Ω such that*

$$\mathcal{N}(F, L_2(\mu), t) \;\geq\; 2^{C \operatorname{fat}_{2t}(F)}, \qquad 0 < t < R, \tag{32}$$

where $C > 0$ is a universal constant. Moreover, μ may be taken to be the uniform distribution on any $2t$-shattered subset of Ω.

Remark. The tightness of (32) is trivially demonstrated by the example $F = \{-\gamma, \gamma\}^n$.

Lemma 19 (Mendelson and Vershynin (2003), Theorem 1) *For all $F \subseteq [-1, 1]^\Omega$ and all probability measures μ,*

$$\mathcal{N}(F, L_2(\mu), t) \;\leq\; \left(\frac{2}{t}\right)^{C \operatorname{fat}_{ct}(F)}, \qquad 0 < t < 1, \tag{33}$$

where $C, c > 0$ are universal constants.

Remark 20 *The following example due to Vershynin (2021) shows that (33) is tight. Take $\Omega = [n]$ and $F = [-1, 1]^\Omega$. Then, for all sufficiently small $t > 0$, we have $\operatorname{fat}_t(F) = n$. However, a simple volumetric calculation shows that $\mathcal{N}(F, L_2(\mu), t)$ behaves as $(C/t)^n$ for small t, where $C > 0$ is a constant.*

Lemma 21 (Rudelson and Vershynin (2006)) *Suppose that $p \in [2, \infty)$, μ is a probability measure on Ω, and $R > 0$. If $F \subset L_p(\Omega, \mu)$ satisfies $\sup_{f \in F} \|f\|_{L_{2p}(\mu)} \leq R$, then*

$$\log \mathcal{N}(F, L_p(\mu), t) \;\leq\; Cp^2 \operatorname{fat}_{ct}(F) \log \frac{R}{ct}, \qquad 0 < t < R;$$

furthermore, for all $\varepsilon > 0$, if $\sup_{f \in F} \|f\|_{L_\infty(\mu)} \leq R$, then

$$\log \mathcal{N}(F, L_\infty(\mu), t) \;\leq\; Cv \log(Rn/vt) \log^\varepsilon(n/v), \qquad 0 < t < R,$$

where $n = |\Omega|$, $v = \operatorname{fat}_{c\varepsilon t}(F)$, and $C, c > 0$ are universal constants.

A.3 Covering numbers of linear and affine classes

Let $B \subset \mathbb{R}^d$ be the d-dimensional Euclidean unit ball and

$$F = \{x \mapsto w \cdot x + b; \|w\| \vee |b| \leq R\}$$

be the collection of R-bounded affine functions on $\Omega = B$.

Remark 22 *There is a trivial reduction from an R-bounded affine class in d dimensions to a $2R$-bounded* linear *class in $d+1$ dimensions, via the standard trick of adding an extra dummy dimension. This only affects the covering number bounds up to constants.*

For $\Omega_n \subset B$, $|\Omega_n| = n$, define $F(\Omega_n) = F|_{\Omega_n}$, and endow Ω_n with the uniform measure μ_n. Zhang (2002, Theorem 4) implies the covering number estimate

$$\log \mathcal{N}(F(\Omega_n), L_\infty(\mu_n), t) \quad \leq \quad C\frac{R^2}{t^2} \operatorname{Log} \frac{nR}{t}, \qquad t > 0,$$

where $C > 0$ is a universal constant (Zhang's result is more general and allows to compute explicit constants). We will use the following sharper bound:

Lemma 23

$$\log \mathcal{N}(F(\Omega_n), L_\infty(\mu_n), t) \quad \leq \quad C\frac{R^2}{t^2} \operatorname{Log} \frac{mt^2}{R^2}, \qquad 0 < t < R,$$

where $m = \min\{n, d\}$ and $C > 0$ is a universal constant.

Proof The result is folklore knowledge, but we provide a proof for completeness.

Let $B = B_2^d$ be the Euclidean unit ball and $X = \{x_1, \ldots, x_n\} \subset B$. This induces the set $F = \{(w \cdot x_1, w \cdot x_2, \ldots, w \cdot x_n); w \in B\} \subset \mathbb{R}^n$. We argue that there is no loss of generality in assuming $d \geq n$. Indeed, if $n > d$, then X is spanned by some $X' = \{x'_1, \ldots, x'_d\} \subset B$ and $F \subset \operatorname{span}(X')$ is also a d-dimensional set. Thus, we assume $d \geq n$ henceforth.

Via a standard infinitesimal perturbation, we can assume that X is a linearly independent set (i.e., spans $\mathbb{R}^n$). If we treat X as an $n \times d$ matrix, then $F = XB$, which means that F is an ellipsoid. We are interested in estimating the ℓ_∞ covering numbers of F.

Let $K \subset \mathbb{R}^d$ be such that $XK = L$, where $L = B_\infty^n$ is the unit cube. (The existence of a K such that $XK \subset L$ is obvious, but because we assumed that X spans $\mathbb{R}^n$, every point in $[-1, 1]^n$ has a pre-image under X.) Let us compute the polar body K°, defined as

$$K^\circ = \left\{ u \in \mathbb{R}^d : \sup_{v \in K} v \cdot u \leq 1 \right\}.$$

We claim that

$$K^\circ = \operatorname{absconv}(X) =: \left\{ \sum_{i=1}^n \alpha_i x_i; \sum |\alpha_i| \leq 1 \right\}.$$

Indeed, consider a $z = \sum_{i=1}^n \alpha_i x_i \in \text{absconv}(X)$. Then, for any $v \in K$, we have

$$\begin{aligned} v \cdot z &= v \cdot \sum_{i=1}^n \alpha_i x_i \\ &= \sum_{i=1}^n \alpha_i (v \cdot x_i) \\ &\leq \sum_{i=1}^n |\alpha_i| \leq 1 \qquad \implies z \in K^\circ, \end{aligned}$$

where we have used $|v \cdot x_i| \leq 1$, since $XK = L = B_\infty^n = [-1,1]^n$. This shows that $\text{absconv}(X) \subseteq K^\circ$. On the other hand, consider any $u \in K^\circ$. There is no loss of generality in assuming that u is in the span of X, that is, $u = \sum_{i=1}^m \alpha_i x_i$, for $\alpha_i \in \mathbb{R}$. By definition of $u \in K^\circ$, we have

$$\sup_{v \in K} v \cdot u = \sup_{v \in K} v \cdot \sum_{i=1}^n \alpha_i x_i = \sup_{v \in K} \sum_{i=1}^n \alpha_i (v \cdot x_i) \leq 1.$$

Now because $XK = [-1,1]^n$, for each choice of $\alpha \in \mathbb{R}^n$, there is a $v \in K$ such that $|v \cdot x_i| = \text{sign}(\alpha_i)$ for all $i \in [n]$. This shows that we must have $\sum_{i=1}^n |\alpha_i| \leq 1$, and proves $K^\circ \subseteq \text{absconv}(X)$.

It is well-known (and easy to verify) that covering numbers enjoy an affine invariance:

$$N(F, L) := N(XB, XK) = N(B, K),$$

where $N(A, B)$, for two sets A, B, is the smallest number of copies of B necessary to cover A. Now the seminal result of Artstein et al. (2004) applies: for all $t > 0$,

$$\log N(B, tK) \leq a \log N(K^\circ, btB),$$

where $a, b > 0$ are universal constants.

This reduces the problem to estimating the ℓ_2-covering numbers of $\text{absconv}(X)$. The latter may be achieved via Maurey's method (Vershynin, 2018, Corollary 0.0.4 and Exercise 0.0.6): the t-covering number of $\text{absconv}(rX)$ under ℓ_2 is at most

$$(c + cmt^2/r^2)^{\lceil r^2/t^2 \rceil},$$

where $c > 0$ is a universal constant.

■

A.4 Fat-shattering dimension of linear and affine classes

In this section, $\Omega = \mathbb{R}^d$ and $B \subset \mathbb{R}^d$ denotes the Euclidean unit ball. A function $f : \Omega \to \mathbb{R}$ is said to be *affine* if it is of the form $f(x) = w \cdot x + b$, for some $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$, where $\cdot$ denotes the Euclidean inner product.

Throughout the paper, we have have referred to R-bounded affine function classes as those for which $\|w\| \vee |b| \leq R$. In this section, we define the larger class of *R-semi-bounded* affine functions, as those for which $\|w\| \leq R$, but b may be unbounded. In particular, the covering-number results (and the reduction to linear classes spelled out in Remark 22) do not apply to semi-bounded affine classes.

The following simple result may be of independent interest.

Lemma 24 *Let $F \subset \mathbb{R}^\Omega$ be some collection of functions with the closure property*

$$f, g \in F \implies (f - g)/2 \in F. \tag{34}$$

Then, for all $\gamma > 0$, we have $\mathrm{fat}_\gamma(F) = \mathrm{f\mathring{a}t}_\gamma(F)$.

Proof

Suppose that some set $\{x_1, \ldots, x_k\}$ is γ-shattered by F. That means that there is an $r \in \mathbb{R}^k$ such that for all $y \in \{-1, 1\}^k$, there is an $f = f_y \in F$ for which

$$\gamma \leq y_i(f(x_i) - r_i), \qquad i \in [k]. \tag{35}$$

Now for any $y \in \{-1, 1\}^k$, let $\hat{f} = f_y$ and $\check{f} = f_{-y}$. Then, for each $i \in [k]$, we have

$$\begin{aligned} \gamma &\leq y_i(\hat{f}(x_i) - r_i), \\ \gamma &\leq -y_i(\check{f}(x_i) - r_i). \end{aligned}$$

It follows that $f = (\hat{f} - \check{f})/2$ achieves (35), for the given y, with $r \equiv 0$. Now (34) implies that the function defined by f belongs to F, which completes the proof. ■

Now it is well-known (Bartlett and Shawe-Taylor, 1999, Theorem 4.6) that bounded *linear* functions — i.e., function classes on B of the form $F = \{x \mapsto w \cdot x; \|w\| \leq R\}$, also known as *homogeneous hyperplanes* — satisfy $\mathrm{fat}_\gamma(F) \leq (R/\gamma)^2$. The discussion in Hanneke and Kontorovich (2019, p. 102) shows that the common approach of reducing of the general (affine) case to the linear (homogeneous, $b = 0$) case, via the addition of a "dummy" coordinate, incurs a large suboptimal factor in the bound. Hanneke and Kontorovich (2019, Lemma 6) is essentially an analysis of the fat-shattering dimension of bounded affine functions. Although this result contains a mistake (see Section 5), much of the proof technique can be salvaged:

Lemma 25 *The semi-bounded affine function class on B defined by $F = \{x \mapsto w \cdot x + b; \|w\| \leq R\}$ in d dimensions satisfies*

$$\mathrm{fat}_\gamma(F) \leq \min\left\{d + 1, \left(\frac{\left(1 + \sqrt{\frac{8}{\pi}}\right) R}{\gamma}\right)^2\right\}, \qquad 0 < \gamma \leq R.$$

Proof Since F satisfies (34), it suffices to consider $\mathrm{f\mathring{a}t}_\gamma(F)$, and so the shattering condition simplifies to

$$\gamma \leq y_i(w \cdot x_i + b), \qquad i \in [k]. \tag{36}$$

Now $\mathfrak{fat}_\gamma(F)$ is always upper-bounded by the VC-dimension of the corresponding class thresholded at zero, i.e., $\mathrm{sign}(F)$. For d-dimensional inhomogeneous hyperplanes, the latter is exactly $d+1$ (Mohri et al., 2012, Example 3.2). Having dispensed with the dimension-dependent part in the bound, we how focus on the R-dependent one.

Let us observe, as in Hanneke and Kontorovich (2019, Lemma 6), that for $\|x_i\| \leq 1$ and $\|w\|$, $\gamma \leq R$, one can always realize (36) with $|b| \leq 2R$; which is what we shall assume, without loss of generality, henceforth. Summing up the k inequalities in (36) yields

$$k\gamma \;\leq\; w \cdot \sum_{i=1}^k y_i x_i + b \sum_{i=1}^k y_i \leq R \left\| \sum_{i=1}^k y_i x_i \right\| + 2R \left| \sum_{i=1}^k y_i \right|.$$

Letting y be drawn uniformly from $\{-1,1\}^k$ and taking expectations, we have

$$\begin{aligned} k\gamma &\leq R\,\mathbb{E} \left\| \sum_{i=1}^k y_i x_i \right\| + 2R\,\mathbb{E} \left| \sum_{i=1}^k y_i \right| \leq R \sqrt{\mathbb{E} \left\| \sum_{i=1}^k y_i x_i \right\|^2} + 2R \sqrt{\mathbb{E} \left(\sum_{i=1}^k y_i \right)^2} \\ &= R \sqrt{\sum_{i=1}^k \|x_i\|^2} + 2R \sqrt{\sum_{i=1}^k \mathbb{E} y_i^2} \leq 3R\sqrt{k}. \end{aligned}$$

Isolating k on the left-hand side of the inequality proves the claim $k \leq \left(\frac{3R}{\gamma}\right)^2$.

Following a referee's suggestion, we improve the constant as follows. Note that

$$\mathbb{E} \left| \sum_{i=1}^k y_i \right| = \frac{1}{2^k} \sum_{i=0}^k \binom{k}{i} |k-2i| = \frac{k}{2^{k-1}} \binom{k-1}{\lfloor \frac{k}{2} \rfloor} \leq \sqrt{\frac{2}{\pi}} \frac{k}{\sqrt{k+\frac{1}{2}}},$$

where the inequality follows from a binomial coefficient estimate via Stirling's approximation. Thus,

$$k\gamma \;\leq\; R\sqrt{k} + 2R\sqrt{\frac{2}{\pi}} \frac{k}{\sqrt{k+\frac{1}{2}}} \leq R\sqrt{k} + 2R\sqrt{\frac{2}{\pi}}\sqrt{k},$$

which proves that $k \leq \left(\frac{\left(1+\sqrt{\frac{8}{\pi}}\right)R}{\gamma} \right)^2$. ∎

A.5 Concavity miscellanea

The results below are routine exercises in differentiation and Jensen's inequality.

Lemma 26 *For $u > 0$, the function $x \mapsto x \log(u/x)$ is concave on $(0, \infty)$.*

Corollary 27 *For all $u > 0$ and $v_i > 0$, $i \in [k]$,*

$$\sum_{i=1}^k v_i \log(u/v_i) \;\leq\; \Big(\sum v_i\Big) \log \frac{uk}{\sum v_i}.$$

Lemma 28 *For $0 \leq \varepsilon \leq \log 2$ and $u \geq 2$, the function $x \mapsto x \log^{1+\varepsilon}(u/x)$ is concave on $[1, \infty)$. It follows that for ε, u as above and $v_i \geq 1$, $i \in [k]$,*

$$\sum_{i=1}^{k} v_i \log^{1+\varepsilon}(u/v_i) \quad \leq \quad \Big(\sum v_i\Big) \log^{1+\varepsilon} \frac{uk}{\sum v_i}.$$