אוניברסיטת בן-גוריון בנגב

הצעת תוכנית מחקר ללימודי דוקטורט

אלגוריתמי דחיסה עבור פונקציות ממשיות

Compression-Schemes for Real-Valued Learners

מנחם סדיגורסקי

אפריל 2018

שם המנחה: ______________________________חתימת המנחה: ________________________

חתימת יו״ר ועדת מוסמכים מחלקתי: ________________________

BEN-GURION UNIVERSITY OF THE NEGEV
THE FACULTY OF NATURAL SCIENCES
DEPARTMENT OF COMPUTER SCIENCE

Compression-Schemes using Real-Valued Learners

This dissertation proposal is submitted in partial fulfillment of the requirements for Doctor in Philosophy degree

in

Computer Science

by

Menachem Sadigurschi

April 2018

תקציר

במחקר זה התמקדתי במושג ה"דחיסה", כפי שהוא מתואר בספרות הלמידה, את הקשר שלו לתחום ובפרט בהקשר של בעיות רגרסיה.

הדבר נעשה בשני מישורים –

הכיוון הראשון החל בגרסה יעילה של אלגוריתם הדחיסה המתואר במאמרם של מורן ויהודייוף (2016). לאחר מכן הרחבנו גישה זו מבעיות סיווג אל בעיות רגרסיה, וכך השגנו אלגוריתם דחיסה גנרי עבור המקרים הללו שגודל הדחיסה המתקבלת בו הינה בעלת גודל חסום. הדבר נעשה בעזרת רדוקציה ללומד גנרי עבור המחלקות המדוברות. ככל הידוע לנו זוהי הבניה הגנרית הראשונה (ללא קשר ליעילות או גודל הדחיסה), המבטיחה שחזור עם קירוב אחיד.

במסגרת בניה זו פיתחנו תהליך גנרי ליצירת לומד-רגרסיה-מוחלש. תהליך זה הינו בעל חשיבות בפני עצמו, מעבר לשימוש שנעשה בו במסגרת זו. בפרט, תוצאה זו שופכת אור על בעיה פתוחה שהוצגה על ידי סימון (1997).

בנוסף אנו מדגימים את השימוש באלגוריתם עבור שתי בעיות רגרסיה: למידה של פונקציות ליפשיץ ופונקציות עם השתנות חסומה.

הכיוון השני נובע לדחיסה אגנוסטית. במסגרת זו אנו מספקים את התוצאה החיובית הראשונה עבור דחיסה אגנוסטית בעלת גודל חסום. אנו מראים שעבור $p \in \{1, \infty\}$ קיימת לרגרסיה לינארית אגנוסטית אלגוריתם דחיסה מגודל חסום שתלוי רק במימד המרחב (תלות לינארית).

בניגוד לתוצאה זו אנו מראים כי לכל שגיאה ℓ_p אחרת $(1 < p < \infty)$ לא קיים אלגוריתם דחיסה מגודל חסום (שגודלו לא תלוי בגודל המדגם).

תוצאה זו מעדנת ומכלילה את תוצאת אי-ההיתכנות של דויד ושות' עבור $\ell 2$

Abstract

In my research we focued on the notion of *Compression-Scheme* and its relation to *Learning Theory* and in particular to the problem of regression. This was done in two directions -

The first one was to give an algorithmically efficient version of the learner-to-compression scheme conversion in Moran and Yehudayoff (2016). We further extend this technique to real-valued hypotheses, to obtain a bounded-size sample compression scheme via an efficient reduction to a certain generic real-valued learning strategy. To our knowledge, this is the first general compressed regression result (regardless of efficiency or boundedness) guaranteeing uniform approximate reconstruction. Along the way, we develop a generic procedure for constructing weak real-valued learners out of abstract regressors; this result is also of independent interest. In particular, this result sheds new light on an open question of H. Simon (1997). We show applications to two regression problems: learning Lipschitz and bounded-variation functions.

The second direction is the *Agnostic-Compression* setting. We obtain the first positive results for bounded sample compression in this setting. We show that for $p \in \{1, \infty\}$, agnostic linear regression admits a bounded sample compression scheme. Specifically, we exhibit efficient sample compression schemes for agnostic linear regression in $\mathbb{R}^d$ of size $d+1$ under the ℓ_1 loss and size $d+2$ under the ℓ_∞ loss. We further show that for every other ℓ_p loss $(1 < p < \infty)$, there does not exist an agnostic compression scheme of bounded size. This refines and generalizes a negative result of David et al. [2016] for the ℓ_2 loss.

Table of Contents

Chapter 1

Introduction

> We may assume the superiority ceteris paribus of the demonstration which derives from fewer postulates or hypotheses
>
> Aristotle, Posterior Analytics

The study of Machine Learning Theory has been for three decades a growing field both in the statistical and in the algorithmic research areas. Learning algorithms are used this days on a wide range of topics, from image-segmentation and natural-language processing to data-science and bioinformatics. Since the beginning, several notions of learning were proposed, trying to capture the characteristics of learnable problems. Two of the most important and dominant notions are the VC-Dimension by Vapnik and Chervonenkis and the PAC learning by Valiant.

The problem of compressing data dates back to the beginning of the field of coding-theory and information-theory by Shannon. As more and more novel learning algorithms had been designed, one of the common aspects that where noted is that at the core of some of them lays some kind of compression, the principle of finding "representative" subsets of the data. as part of a more general *Occam learning* paradigm. Most notable is the SVM algorithm, which derives its name from the set of supporting vectors which uniquely defines the linear separator returned by the algorithm.

Following this path, Littlestone and Warmuth established a formal framework for discussion of compression scheme from the learning point of view. In addition they showed that for the case of binary-labeled classes - compression implies learnability [1].

A fundamental question, posed by Littlestone and Warmuth [1986] on the same paper, concerns the reverse implication: Can every learner be converted into a sample compression scheme? Or, in a more quantitative formulation: Does every VC class admit a constant-size sample compression scheme? A series of partial results [Floyd, 1989, Helmbold et al., 1992, Floyd and Warmuth, 1995, Ben-David and Litman, 1998, Kuzmin and Warmuth, 2007, Rubinstein et al., 2009, Rubinstein and Rubinstein, 2012, Chernikov and Simon, 2013, Livni and Simon, 2013, Moran et al., 2017] culminated in Moran and Yehudayoff [2016] which resolved the latter question[2].

The usefulness of this link is that while learning is a statistical notion, compression is a combinatorial one. Thus by linking the two, by such an equivalence, can help moving questions about learning to the combinatorial world, open the research to other directions and to a wide range of tools previously not relevant to this area.

Moran and Yehudayoff's solution involved a clever use of von Neumann's minimax theorem, which allows one to make the leap from the existence of a weak learner uniformly over all *distributions on examples* to the existence of a *distribution on weak hypotheses* under which they achieve a certain performance simultaneously over all of the examples. Although their paper can be understood without any knowledge of boosting, Moran and Yehudayoff note the well-known connection between boosting and compression. Indeed, boosting may be used to obtain a constructive proof of the minimax theorem [Freund and Schapire, 1996, 1999] and [Floyd and Warmuth, 1995, Section 9.1] — and this connection was what motivated us to seek an efficient algorithm implementing Moran and Yehudayoff's existence proof. Having obtained an efficient conversion procedure from consistent PAC learners to bounded-size sample compression schemes, we turned our attention to the case of real-valued hypotheses, a case which had almost no results on this area. It turned out that a virtually identical boosting framework could be made to work for this case as well, although a novel analysis was required.

[1] Lately there a growing study on the properties and the generalization bounds of compressing-based learning algorithms, see for example Gottlieb et al. [2017c]Graepel et al. [2005]Cummings et al. [2016].

[2] The refined conjecture of Littlestone and Warmuth [1986], that any concept class with VC-dimension d admits a compression scheme of size $O(d)$, remains open.

Our second path of investigation is focused on the notion of *agnostic-compression scheme*. In a recent paper, David, Moran, and Yehudayoff [2016] generalized the definition of *compression scheme* to the agnostic case, where it is required that the function reconstructed from the compression set obtains an average loss on the full data set nearly as small as the function in the class that minimizes this quantity. Below, we give a strong motivation for this criterion by arguing an equivalence to the generalization ability of the compression-based learning algorithm. Under this definition, David et al. [2016] extended the realizable-case result for VC classes to cover the agnostic case as well: a bounded-size compression scheme for the former implies such a scheme (in fact, of the same size) for the latter. They also generalized from binary to multiclass concept families, with the graph dimension in place of VC-dim. Proceeding to real-valued function classes, David et al. came to a starkly negative conclusion: they established that there is *no* constant-size agnostic sample compression scheme for linear functions under the ℓ_2 loss. (*Realizable* linear regression in $\mathbb{R}^d$ trivially admits sample compression of size $d+1$, under any loss, by selecting a minimal subset that spans the data.)

Those results led us to try and find a more precise characterization of the loss functions which can be *agnostic-compressed* effectively. As a first step we turned our attention to the ℓ_p-losses. We extend the impossibility results of David et al. to the ℓ_p-losses for $p \in (1, \infty)$, and on the other hand construct an efficient agnostic-compression scheme for ℓ_1 and ℓ_∞ losses, which is independent on the sample size. Resulting is an interesting separation between those two cases and the rest of the ℓ_p family, which offers a hint towards characterizing the loss functions amenable to compression.

1.1 Definitions and notation

We will write $[k] := \{1, \ldots, k\}$. An *instance space* is an abstract set $\mathcal{X}$. For a concept class $\mathcal{C} \subset \{0,1\}^{\mathcal{X}}$, if say that $\mathcal{C}$ *shatters* a set $\{x_1, \ldots, x_k\} \subset \mathcal{X}$ if

$$\mathcal{C}(S) = \{(f(x_1), f(x_2), \ldots, f(x_k)) : f \in \mathcal{C}\} = \{0,1\}^k .$$

The VC-dimension $d = d_{\mathcal{C}}$ of $\mathcal{C}$ is the size of the largest shattered set (or ∞ if $\mathcal{C}$ shatters sets of arbitrary size) [Vapnik and Červonenkis, 1971]. When the roles of $\mathcal{X}$ and $\mathcal{C}$ are exchanged — that is, an $x \in \mathcal{X}$ acts on $f \in \mathcal{C}$ via $x(f) = f(x)$, — we refer to $\mathcal{X} = \mathcal{C}^*$ as the *dual* class of $\mathcal{C}$. Its VC-dimension is then $d^* = d^*_{\mathcal{C}} := d_{\mathcal{C}^*}$, and referred to as the *dual VC dimension*. Assouad [1983] showed that $d^* \le 2^{d+1}$.

For $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ and $t > 0$, For $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ and $t > 0$, we say that $\mathcal{F}$ t-shatters a set $\{x_1, \dots, x_k\} \subset \mathcal{X}$ if there is an $r \in \mathbb{R}^m$ such that for all $y \in \{-1, 1\}^m$ there is an $f \in \mathcal{F}$ such that $\min_{i \in [k]} y_i(f(x_i) - r_i) \ge t$. The t-fat-shattering dimension $d(t) = d_{\mathcal{F}}(t)$ is the size of the largest t-shattered set (possibly ∞) [Alon et al., 1997]. Again, the roles of $\mathcal{X}$ and $\mathcal{F}$ may be switched, in which case $\mathcal{X} = \mathcal{F}^*$ becomes the dual class of $\mathcal{F}$. Its t-fat-shattering dimension is then $d^*(t)$, and Assouad's argument shows that $d^*(t) \le 2^{d(t)+1}$.

A *sample compression scheme* (κ, ρ) for a hypothesis class $\mathcal{F} \subset \mathcal{Y}^{\mathcal{X}}$ is defined as follows. A k-*compression* function κ maps sequences $((x_1, y_1), \dots, (x_m, y_m)) \in \bigcup_{\ell \ge 1} (\mathcal{X} \times \mathcal{Y})^\ell$ to elements in $\mathcal{K} = \bigcup_{\ell \le k'} (\mathcal{X} \times \mathcal{Y})^\ell \times \bigcup_{\ell \le k''} \{0,1\}^\ell$, where $k' + k'' \le k$. A *reconstruction* is a function $\rho : \mathcal{K} \to \mathcal{Y}^{\mathcal{X}}$. We say that (κ, ρ) is a k-size sample compression scheme for $\mathcal{F}$ if κ is a k-compression and for all $h^* \in \mathcal{F}$ and all $S = ((x_1, h^*(x_1)), \dots, (x_m, h^*(y_m)))$, we have $\hat{h} := \rho(\kappa(S))$ satisfies $\hat{h}(x_i) = h^*(x_i)$ for all $i \in [m]$.

For real-valued functions, there are several notions of compression-schemes. We say it is a *uniformly* ε-*approximate* compression scheme if

$$\max_{1 \le i \le m} |\hat{h}(x_i) - h^*(x_i)| \le \varepsilon.$$

Note that David et al. [2016] proposed the following definitions:
Let $S = (x_1, y_i), \dots, (x_m, y_m)$ be a tagged sample drawn i.i.d from some unknown distribution, an let $l : \mathcal{X} \times \mathbb{R} \to \mathbb{R}$ be some loss function. We say that (κ, ρ) is an *agnostic sample compression scheme* for $\mathcal{H}$ if, for every sample S, $f_S := \rho(\kappa(S))$, achieves $\mathcal{F}$-competitive empirical loss:

$$\frac{1}{m} \sum_{i=1}^{m} l(f_S(x_i), y_i) \le \inf_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} l(f_S(x_i), y_i),$$

and we say that it is ϵ-*Approximate Agnostic Sample Compression Scheme* for $\mathcal{H}$ if for every sample S

$$\frac{1}{m} \sum_{i=1}^{m} l(f_S(x_i), y_i) \le \inf_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} l(f_S(x_i), y_i) + \epsilon.$$

1.2 Main results

Throughout the paper, we implicitly assume that all hypothesis classes are *admissible* in the sense of satisfying mild measure-theoretic conditions, such as those specified in Dudley [1984, Section 10.3.1] or Pollard [1984, Appendix C]. We begin with an algorithmically efficient version of the learner-to-compression scheme conversion in Moran and Yehudayoff [2016]:

Theorem 1.1 (Efficient compression for classification)**.** *Let $\mathcal{C}$ be a concept class over some instance space $\mathcal{X}$ with VC-dimension d, dual VC-dimension d^*, and suppose that $\mathcal{A}$ is a (proper, consistent) PAC-learner for $\mathcal{C}$: For all $0 < \varepsilon, \delta < 1/2$, all $f^* \in \mathcal{C}$, and all distributions D over $\mathcal{X}$, if $\mathcal{A}$ receives $m \geq m_{\mathcal{C}}(\varepsilon, \delta)$ points $S = \{x_i\}$ drawn iid from D and labeled with $y_i = f^*(x_i)$, then $\mathcal{A}$ outputs an $\hat{f} \in \mathcal{C}$ such that*

$$\mathbb{P}_{S \sim D^m}\left(\mathbb{P}_{X \sim D}\left(\hat{f}(X) \neq f^*(X) \,|\, S\right) > \varepsilon\right) < \delta.$$

For every such $\mathcal{A}$, there is a randomized sample compression scheme for $\mathcal{C}$ of size $O(k \log k)$, where $k = O(dd^*)$. Furthermore, on a sample of any size m, the compression set may be computed in expected time

$$O\left((m + T_{\mathcal{A}}(cd)) \log m + m T_{\mathcal{E}}(cd)(d^* + \log m)\right),$$

where $T_{\mathcal{A}}(\ell)$ is the runtime of $\mathcal{A}$ to compute $\hat{f}$ on a sample of size ℓ, $T_{\mathcal{E}}(\ell)$ is the runtime required to evaluate $\hat{f}$ on a single $x \in \mathcal{X}$, and c is a universal constant.

Although for our purposes the existence of a distribution-free sample complexity $m_{\mathcal{C}}$ is more important than its concrete form, we may take $m_{\mathcal{C}}(\varepsilon, \delta) = O(\frac{d}{\varepsilon} \log \frac{1}{\varepsilon} + \frac{1}{\varepsilon} \log \frac{1}{\delta})$ [Vapnik and Chervonenkis, 1974, Blumer et al., 1989], known to bound the sample complexity of empirical risk minimization; indeed, this loses no generality, as there is a well-known efficient reduction from empirical risk minimization to any proper learner having a polynomial sample complexity [Pitt and Valiant, 1988, Haussler et al., 1991]. We allow the evaluation time of $\hat{f}$ to depend on the size of the training sample in order to account for non-parametric learners, such as nearest-neighbor classifiers. A naive implementation of the Moran and Yehudayoff [2016] existence proof yields a runtime of order $m^{cd} T_{\mathcal{A}}(c'd) + m^{cd^*}$ (for some universal constants c, c'), which can be doubly exponential when $d^* = 2^d$; this is without taking into account the cost of computing the minimax distribution on the $m^{cd} \times m$ game matrix.

Next, we extend the result in Theorem 1.1 from classification to regression:

Theorem 1.2 (Efficient compression for regression)**.** *Let $\mathcal{F} \subset [0,1]^{\mathcal{X}}$ be a function class with t-fat-shattering dimension $d(t)$, dual t-fat-shattering dimension $d^*(t)$, and suppose that $\mathcal{A}$ is an ERM (i.e., proper, almost consistent) learner for $\mathcal{F}$: For all $f^* \in \mathcal{C}$, and all distributions D over $\mathcal{X}$, if $\mathcal{A}$ receives m points $S = \{x_i\}$ drawn iid from D and labeled with $y_i = f^*(x_i)$, then $\mathcal{A}$ outputs an $\hat{f} \in \mathcal{F}$ such that $\max_{i\in[m]}|\hat{f}(x_i) - f^*(x_i)| \leq \alpha\eta$ for $\alpha \in [0,1)$]. For every such $\mathcal{A}$, there is a randomized uniformly ε-approximate sample compression scheme for $\mathcal{F}$ of size $O(k\tilde{m}\log(k\tilde{m}))$, where $\tilde{m} = O(d(c\varepsilon)\log(1/\varepsilon))$ and $k = O(d^*(c\varepsilon)\log(d^*(c\varepsilon)/\varepsilon))$. Furthermore, on a sample of any size m, the compression set may be computed in expected time*

$$O(mT_{\mathcal{E}}(\tilde{m})(k + \log m) + T_{\mathcal{A}}(\tilde{m})\log(m)),$$

where $T_{\mathcal{A}}(\ell)$ is the runtime of $\mathcal{A}$ to compute $\hat{f}$ on a sample of size ℓ, $T_{\mathcal{E}}(\ell)$ is the runtime required to evaluate $\hat{f}$ on a single $x \in \mathcal{X}$, and c is a universal constant.

A key component in the above result is our construction of a generic (η,γ)-weak learner.

Definition 1.1. For $\eta \in [0,1]$ and $\gamma \in [0,1/2]$, we say that $f : \mathcal{X} \to \mathbb{R}$ is an an (η,γ)-*weak hypothesis* (with respect to distribution D and target $f^* \in \mathcal{F}$) if

$$\mathbb{P}_{X\sim D}(|f(X) - f^*(X)| > \eta) \leq \frac{1}{2} - \gamma.$$

Theorem 1.3 (Generic weak learner)**.** *Let $\mathcal{F} \subset [0,1]^{\mathcal{X}}$ be a function class with t-fat-shattering dimension $d(t)$. For some universal numerical constants $c_1, c_2, c_3 \in (0,\infty)$, for any $\eta, \delta \in (0,1)$ and $\gamma \in (0,1/4)$, any $f^* \in \mathcal{F}$, and any distribution D, letting $X_1, \ldots, X_m$ be drawn iid from D, where*

$$m = \left\lceil c_1 \left(d(c_2\eta)\ln\left(\frac{c_3}{\eta}\right) + \ln\left(\frac{1}{\delta}\right)\right)\right\rceil,$$

with probability at least $1-\delta$, every $f \in \mathcal{F}$ with $\max_{i\in[m]} |f(X_i) - f^*(X_i)| \leq \alpha\eta$ for $\alpha \in [0,1)$, is an (η,γ)-weak hypothesis with respect to D and f^*.

As one can see, our results allow us to use any hypothesis $f \in \mathcal{F}$ with $\max_{i\in[m]} |f(X_i) - f^*(X_i)|$ bounded below η: for instance, bounded by $\eta/2$.

Following this we give applications to sample compression for nearest-neighbor and bounded-variation regression. In order to provide those applications we had to provide upper-bounds on the dual-t-shattering dimension for both cases.

For the *agnostic-compression scheme* setting - the negative result of David et al. [2016] raises a general doubt over whether sample compression is ever a viable approach to agnostic learning of real-valued functions. In this work, we address this concern by proving that

Theorem 1.4. *There exists an efficiently computable compression scheme for agnostic linear regression in* $\mathbb{R}^d$ *under the* ℓ_1 *loss of size* $d+1$.

The upshot is that if we replace the ℓ_2 loss with the ℓ_1 loss, then there *is* a simple agnostic compression scheme of size $d+1$ for linear regression in $\mathbb{R}^d$. This is somewhat surprising, given the above negative result for the ℓ_2 loss. We also prove a similar result can be done under the ℓ_∞ loss.

Theorem 1.5. *There exists an efficiently computable compression scheme for agnostic linear regression in* $\mathbb{R}^d$ *under the* ℓ_∞ *loss of size* $d+2$.

This construction is somewhat different then the ℓ_1 case. However, interestingly, we also generalize the argument of David et al. [2016] to show that these are the *only two* ℓ_p losses ($1 \le p \le \infty$) for which there exists a constant-size compression scheme. Specifically we prove that

Theorem 1.6. *There is no agnostic sample compression scheme for zero-dimensional linear regression under* ℓ_p *loss,* $1 < p < \infty$*, with size* $k(m) < \log(m)$.

Computationally, our compression schemes for ℓ_1 and ℓ_∞ amount to solving a polynomial (in fact, linear) size linear program. These appear to be the first positive results for bounded agnostic sample compression for real-valued function classes.

1.3 Related work

It appears that generalization bounds based on sample compression were independently discovered by Littlestone and Warmuth [1986] and Devroye et al. [1996] and further elaborated upon by Graepel et al. [2005]; see Floyd and Warmuth [1995] for background and discussion. A more general kind of Occam learning was discussed in Blumer et al. [1989]. Computational lower bounds on sample compression were obtained in Gottlieb et al. [2014], and some communication-based lower bounds were given in Kane et al. [2017].

Beginning with Freund and Schapire [1997]'s `AdaBoost.R` algorithm, there have been numerous attempts to extend AdaBoost to the real-valued case [Bertoni et al., 1997, Drucker, 1997, Avnimelech and Intrator, 1999, Karakoulas and Shawe-Taylor, 2000, Duffy and Helmbold, 2002, Kégl, 2003, Nock and Nielsen, 2007] along with various theoretical and heuristic constructions of particular weak regressors [Mason et al., 1999, Friedman, 2001, Mannor and Meir, 2002]; see also the survey Mendes-Moreira et al. [2012].

An explanation for the challenge of defining a good weak-learner was given by Duffy and Helmbold [2002, Remark 2.1] we discuss this issue on 2.2.1. The (η, γ)-weak learner, which has appeared, among other works, in Anthony et al. [1996], Simon [1997], Avnimelech and Intrator [1999], Kégl [2003], gets around this difficulty — but provable general constructions of such learners have been lacking. Likewise, the heart of our sample compression engine, `MedBoost`, has been widely in use since Freund and Schapire [1997] in various guises. Our Theorem 1.3 supplies the remaining piece of the puzzle: *any* sample-(almost-))consistent regressor applied to some random sample of bounded size yields an (η, γ)-weak hypothesis. The closest analogue we were able to find was Anthony et al. [1996, Theorem 3], which is non-trivial only for function classes with finite pseudo-dimension, and is inapplicable, e.g., to classes of 1-Lipschitz or bounded variation functions (see 3.3).

The literature on general sample compression schemes for real-valued functions is quite sparse. There are well-known narrowly tailored results on specifying functions or approximate versions of functions using a finite number of points, such as the classical fact that a polynomial of degree p can be perfectly recovered from $p + 1$ points. To our knowledge, the only *general* results on sample compression for real-valued functions (applicable to *all* learnable function classes) is Theorem 4.3 of David, Moran, and Yehudayoff [2016]. They propose a general technique to convert any learning algorithm achieving an arbitrary sample complexity $M(\varepsilon, \delta)$ into a compression scheme of size $O(M(\varepsilon, \delta) \log(M(\varepsilon, \delta)))$, where δ may approach 1. However, their notion of compression scheme is significantly weaker than ours: namely, they allow $\hat{h} = \rho(\kappa(S))$ to satisfy merely $\frac{1}{m} \sum_{i=1}^{m} |\hat{h}(x_i) - h^*(x_i)| \leq \varepsilon$, rather than our *uniform* ε-approximation requirement $\max_{1 \leq i \leq m} |\hat{h}(x_i) - h^*(x_i)| \leq \varepsilon$. In particular, in the special case of $\mathcal{F}$ a family of *binary*-valued functions, their notion of sample compression does *not* recover the usual notion of sample compression schemes for classification, whereas our uniform ε-approximate compression notion *does* recover it as a special case. We therefore consider our notion to be a more fitting generalization of the definition of sample compression to the real-valued case.

For the problem of *agnostic-compression scheme* David et al. [2016, Theorem 4.1] obtained the aforementioned negative result for ℓ_2 agnostic linear regression, as well as an $\tilde{O}(\log(d/\varepsilon))$-size compression scheme for *approximate* ℓ_2 agnostic linear regression (the latter model is not considered here, although connections to this setting are discussed on 5.2.3).

Ashtiani et al. [2018] adapted the notion of a compression scheme to the distribution learning problem. They showed that if a class of distributions admits robust compressibility then it is agnostically learnable. They used those results in order to provide state-of-the-art sample-complexity bounds for learning mixture of Gaussians.

1.4 Overview of Techniques

Our point of departure is the simple but powerful observation [Schapire and Freund, 2012] that many boosting algorithms (e.g., AdaBoost, α-Boost) are capable of outputting a family of $O(\log(m)/\gamma^2)$ hypotheses such that not only does their (weighted) majority vote yield a sample-consistent classifier, but in fact a $\approx (\frac{1}{2} + \gamma)$ super-majority does as well. This fact implies that after boosting, we can sub-sample a constant (i.e., independent of sample size m) number of classifiers and thereby efficiently recover the sample compression bounds of Moran and Yehudayoff [2016].

Our chief technical contribution, however, is in the real-valued case. As we discuss below, extending the boosting framework from classification to regression presents a host of technical challenges, and there is currently no off-the-shelf general-purpose analogue of AdaBoost for real-valued hypotheses. One of our insights is to impose distinct error metrics on the weak and strong learners: a "stronger" one on the latter and a "weaker" one on the former. This allows us to achieve two goals simultaneously:

(a) We give apparently the first generic construction for our weak learner, demonstrating that the object is natural and abundantly available. This is in contrast with many previous proposed weak regressors, whose stringent or exotic definitions made them unwieldy to construct or verify as such. The construction is novel and may be of independent interest.

(b) We show that the output of a certain real-valued boosting algorithm may be sparsified so as to yield a constant size sample compression analogue of the Moran and Yehudayoff result for classification. This gives the first general constant-size sample compression scheme having uniform approximation guarantees on the data.

Chapter 2

Boosting Real-Valued Functions

In the study of machine learning theory, the standard definitions of learning, as PAC-learning for the binary case, require the learner to achieve arbitrary small accuracy. It is often hard to be able to supply such strong requirement, but nevertheless it may be much easier, for a large set of problems, to construct learners which are somewhat better than a random labeling. Those learners are called *weak-learner* as opposed to the standard *strong-learners*. The idea of leveraging or boost weak-learners in order to achieve stronger learning guarantees started as a question proposed by Kearns, and got to a positive result in the seminal works by Schapire [1990] and Freund and Schapire [1997]. The latter contained the well known Adaboost algorithm which is widely used in practice.

2.1 The MedBoost Algorithm

In the context of boosting for real-valued functions, the notion of an (η, γ)-weak hypothesis plays a role analogous to the usual notion of a weak hypothesis in boosting for classification. Using this notion, the following boosting algorithm was proposed by Kégl [2003] as an extension to the classic Adaboost algorithm.

Algorithm 1 `MedBoost`$(\{(x_i, y_i)\}_{i\in[m]}, T, \gamma, \eta)$

Define P_0 as the uniform distribution over $\{1, \ldots, n\}$
for $t = 0, \ldots, T$ **do**
  Call weak learner to get h_t and $(\eta/2, \gamma)$-weak hypothesis
   w.r.t. $(x_i, y_i) : i \sim P_t$ (repeat until it succeeds)
  for $i = 1, \ldots, m$ **do**
    $\theta_i^{(t)} \leftarrow 1 - 2\mathbb{I}[|h_t(x_i) - y_i| > \eta/2]$
  end for
  $\alpha_t \leftarrow \frac{1}{2}\ln\left(\frac{(1-\gamma)\sum_{i=1}^m P_t(i)\mathbb{I}[\theta_i^{(t)}=1]}{(1+\gamma)\sum_{i=1}^m P_t(i)\mathbb{I}[\theta_i^{(t)}=-1]}\right)$
  if $\alpha_t = \infty$ **then**
    Return T copies of h_t, and $(1, \ldots, 1)$
  end if
  for $i = 1, \ldots, m$ **do**
    $P_{t+1}(i) \leftarrow P_t(i)\frac{\exp\{-\alpha_t \theta_i^{(t)}\}}{\sum_{j=1}^m P_t(j)\exp\{-\alpha_t\theta_j^{(t)}\}}$
  end for
end for
Return $(h_1, \ldots, h_T)$ and $(\alpha_1, \ldots, \alpha_T)$

As it will be convenient for our later results, we expressed the algorithm's output as a sequence of functions and weights; the boosting guarantee from Kégl [2003] applies to the weighted quantiles (and in particular, the weighted median) of these function values.

Here we define the weighted median as

$$\mathrm{Median}(y_1, \ldots, y_T; \alpha_1, \ldots, \alpha_T) = \min\left\{ y_j : \frac{\sum_{t=1}^T \alpha_t \mathbb{I}[y_j < y_t]}{\sum_{t=1}^T \alpha_t} < \frac{1}{2} \right\}.$$

Also define the weighted *quantiles*, for $\gamma \in [0, 1/2]$, as

$$Q_\gamma^+(y_1, \ldots, y_T; \alpha_1, \ldots, \alpha_T) = \min\left\{ y_j : \frac{\sum_{t=1}^T \alpha_t \mathbb{I}[y_j < y_t]}{\sum_{t=1}^T \alpha_t} < \frac{1}{2} - \gamma \right\}$$
$$Q_\gamma^-(y_1, \ldots, y_T; \alpha_1, \ldots, \alpha_T) = \max\left\{ y_j : \frac{\sum_{t=1}^T \alpha_t \mathbb{I}[y_j > y_t]}{\sum_{t=1}^T \alpha_t} < \frac{1}{2} - \gamma \right\},$$

and abbreviate $Q_\gamma^+(x) = Q_\gamma^+(h_1(x), \ldots, h_T(x); \alpha_1, \ldots, \alpha_T)$ and $Q_\gamma^-(x) = Q_\gamma^-(h_1(x), \ldots, h_T(x); \alpha_1, \ldots, \alpha_T)$ for $h_1, \ldots, h_T$ and $\alpha_1, \ldots, \alpha_T$ the values returned by `MedBoost`.

2.1.1 Analysis

After proposing the algorithm, Kégl [2003] proves the following result.

Lemma 2.1. *(Kégl [2003]) For a training set $Z = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ of size m, the return values of* `MedBoost` *satisfy*

$$\frac{1}{m}\sum_{i=1}^{m} \mathbb{I}\Big[\max\Big\{\Big|Q^{+}_{\gamma/2}(x_i) - y_i\Big|, \Big|Q^{-}_{\gamma/2}(x_i) - y_i\Big|\Big\} > \eta/2\Big] \le \prod_{t=1}^{T} e^{\gamma\alpha_t} \sum_{i=1}^{m} P_t(i) e^{-\alpha_t \theta_i^{(t)}}.$$

We note that, in the special case of binary classification, `MedBoost` is closely related to the well-known AdaBoost algorithm [Freund and Schapire, 1997], and the above results correspond to a standard margin-based analysis of Schapire et al. [1998].

For our purposes, we will need the following corollary of this, which we prove below.

Corollary 2.2. *For $T = \Theta\Big(\frac{1}{\gamma^2}\ln(m)\Big)$, every $i \in \{1, \ldots, m\}$ has*

$$\max\Big\{\Big|Q^{+}_{\gamma/2}(x_i) - y_i\Big|, \Big|Q^{-}_{\gamma/2}(x_i) - y_i\Big|\Big\} \le \eta/2.$$

In the proof we use the following technical lemma

Lemma 2.3. *For $x \ge \frac{1}{2} + \gamma$ it holds that*

$$x^{1+\gamma}(1-x)^{1-\gamma} \le \left(\frac{1}{2} + \gamma\right)^{1-\gamma} \left(\frac{1}{2} - \gamma\right)^{1+\gamma}.$$

Proof. Denote the left side as a function f and take log of f

$$\log(f(x)) = (1+\gamma)\log(x) + (1-\gamma)\log(1-x).$$

Observe that the derivative with respect to x which is $(\log(f(x)))' = (1+\gamma)/x - (1-\gamma)/(1-x)$ is negative for $x \ge (1+\gamma)/2$. Since $x \ge \frac{1}{2} + \gamma > (1+\gamma)/2$ this condition holds. So the function $\log(f(a)) := \log\left(a^{1+\gamma}(1-a)^{1-\gamma}\right)$ is monotonically decreasing and by that also f itself is monotonically decreasing. Hence

$$x^{1+\gamma}(1-x)^{1-\gamma} \le (\frac{1}{2} + \gamma)^{1+\gamma}(1 - \frac{1}{2} + \gamma)^{1-\gamma}.$$

□

Proof of Corollary 2.2. By the definition of α_t we know that

$$e^{\alpha_t} = \left(\frac{(1-\gamma)\sum_{\theta_i(t)=1} P_t(i)}{(1+\gamma)\sum_{\theta_i(t)=-1} P_t(i)} \right)^{1/2}.$$

Split the sum within the RHS into $\{i \mid \theta_i(t) = 1\}$ and $\{i \mid \theta_i(t) = -1\}$ to get that

$$\begin{aligned}
&\prod_{t=1}^{T} e^{\gamma\alpha_t} \sum_{i=1}^{m} P_t(i) e^{-\alpha_t \theta_i(t)} a \\
&= \prod_{t=1}^{T} e^{\gamma\alpha_t} \left[\sum_{\theta_i(t)=1} P_t(i) e^{-\alpha_t} + \sum_{\theta_i(t)=-1} P_t(i) e^{\alpha_t} \right] \\
&= \prod_{t=1}^{T} e^{\gamma\alpha_t} \left[e^{-\alpha_t} \sum_{\theta_i(t)=1} P_t(i) + e^{\alpha_t} \sum_{\theta_i(t)=-1} P_t(i) \right] \\
&= \prod_{t=1}^{T} \left[e^{-\alpha_t(1-\gamma)} \sum_{\theta_i(t)=1} P_t(i) + e^{\alpha_t(1+\gamma)} \sum_{\theta_i(t)=-1} P_t(i) \right].
\end{aligned}$$

Plug-in e^{α_t}

$$\begin{aligned}
&= \prod_{t=1}^{T} \Bigg[\left(\frac{(1+\gamma)\sum_{\theta_i(t)=-1} P_t(i)}{(1-\gamma)\sum_{\theta_i(t)=1} P_t(i)} \right)^{\frac{1-\gamma}{2}} \sum_{\theta_i(t)=1} P_t(i) \\
&\quad + \left(\frac{(1-\gamma)\sum_{\theta_i(t)=1} P_t(i)}{(1+\gamma)\sum_{\theta_i(t)=-1} P_t(i)} \right)^{\frac{1+\gamma}{2}} \sum_{\theta_i(t)=-1} P_t(i) \Bigg] \\
&= \prod_{t=1}^{T} \Bigg[\left(\sum_{\theta_i(t)=1} P_t(i) \right)^{\frac{1+\gamma}{2}} \left(\sum_{\theta_i(t)=-1} P_t(i) \right)^{\frac{1-\gamma}{2}} \left(\frac{1+\gamma}{1-\gamma} \right)^{\frac{1-\gamma}{2}} \\
&\quad + \left(\sum_{\theta_i(t)=1} P_t(i) \right)^{\frac{1+\gamma}{2}} \left(\sum_{\theta_i(t)=-1} P_t(i) \right)^{\frac{1-\gamma}{2}} \left(\frac{1-\gamma}{1+\gamma} \right)^{\frac{1+\gamma}{2}} \Bigg].
\end{aligned}$$

By the (ε, γ)-weak-learning guarantee we know that $\sum_{\theta_i(t)=1} P_t(i) \geq \frac{1}{2} + \gamma$ and

$\sum_{\theta_i(t)=-1} P_t(i) < \frac{1}{2} - \gamma$ and by Lemma 2.3

$$\leq \prod_{t=1}^{T} \left[\left(\frac{1+\gamma}{1-\gamma}\right)^{\frac{1-\gamma}{2}} + \left(\frac{1-\gamma}{1+\gamma}\right)^{\frac{1+\gamma}{2}} \right] \left(\frac{1}{2}+\gamma\right)^{\frac{1-\gamma}{2}} \left(\frac{1}{2}-\gamma\right)^{\frac{1+\gamma}{2}}$$

$$= \prod_{t=1}^{T} \frac{1}{2} \left(\frac{1-\gamma}{1+\gamma}\right)^{\frac{\gamma}{2}} \left(\frac{1+2\gamma}{1-2\gamma}\right)^{\frac{\gamma}{2}} (1-4\gamma^2)^{1/2} \left(\left(\frac{1+\gamma}{1-\gamma}\right)^{1/2} + \left(\frac{1-\gamma}{1+\gamma}\right)^{1/2} \right),$$

noting that for every $\gamma \in (0, 1/3)$.

$$\frac{1}{2} \left(\frac{1-\gamma}{1+\gamma}\right)^{\frac{\gamma}{2}} \left(\frac{1+2\gamma}{1-2\gamma}\right)^{\frac{\gamma}{2}} (1-4\gamma^2)^{1/2} \left(\left(\frac{1+\gamma}{1-\gamma}\right)^{1/2} + \left(\frac{1-\gamma}{1+\gamma}\right)^{1/2} \right) < e^{-\gamma^2/4},$$

we get that

$$\frac{1}{m} \sum_{i=1}^{m} \mathbb{I}\Big[\max\Big\{ \Big| Q^{+}_{\gamma/2}(x_i) - y_i \Big|, \Big| Q^{-}_{\gamma/2}(x_i) - y_i \Big| \Big\} > \eta/2 \Big]$$
$$\leq \prod_{t=1}^{T} e^{\gamma \alpha_t} \sum_{i=1}^{m} P_t(i) e^{-\alpha_t \theta_i^{(t)}} < e^{-T\gamma^2/4}.$$

Finally for $T = \frac{4}{\gamma^2} \ln(m)$ the last bound is equal to $\frac{1}{m}$ and hence the corollary holds. □

2.2 The Sample Complexity Weak Learning

This section reveals our intention in choosing this notion of weak hypothesis, rather than using, say, an ε-good strong learner under absolute loss. In addition to being a strong enough notion for boosting to work, we show here that it is also a weak enough notion for the sample complexity of weak learning to be of reasonable size: namely, a size quantified by the fat-shattering dimension. This result is also relevant to an open question posed by Simon [1997], which we discuss on Subsection 2.2.3.

2.2.1 The Notion of "Weak Learning"

As mentioned above, the notion of a *weak learner* for learning real-valued functions must be formulated carefully. The naïve thought that we could take any learner guaranteeing, say, absolute loss at most $\frac{1}{2} - \gamma$ is known to not be strong enough to enable boosting to ε loss. However, if we make the requirement too strong, such as in Freund and Schapire [1997] for `AdaBoost.R`, then the sample complexity of weak learning will be so high that weak learners cannot be expected to exist for large classes of functions.

Starting with Kearns and Schapire, the notion of weak learning was tied to the notion of PAC learnability. Weak learning is, as on may expect, the weak version of PAC learning. This relation meant that also weak-learning was defined using a loss-function and a (weak) upper-bound on the loss of the resulting hypothesis, namely a fixed, yet bounded away from 1/2, bound on the expected loss.

Normally when extending the PAC paradigm to the real-valued/continuous case we just replace the loss-function. So we get the following

Definition 2.1 ("Standard"-Weak-Hypothesis)**.** For $\gamma \in [0, 1/2]$, we say that $f : \mathcal{X} \to \mathbb{R}$ is an an γ-*weak hypothesis* (with respect to distribution D and target $f^* \in \mathcal{F}$) if

$$\mathbb{E}_{X\sim D}\left[l(f_S(x), f^*(x))\right] \leq \frac{1}{2} - \gamma.$$

Unfortunately, this extension for the problem of boosting essentially fails. Duffy and Helmbold [2002, Remark 2.1] points out that, using this notion of weak learning, one can't guarantee that using the method of modifying the distribution over the sample will force the learner to establish a good hypothesis. This is due to the fact that, unlike the binary-case. the error can be spread evenly over all the sample, meaning that the error remains the same regardless of the distribution on the sample. This might result in the learner outputting the same hypothesis on each iteration and hence not improving the error of the final output regressor. Some lines of work, including Freund and Schapire's `AdaBoost.R`, used more complex boosting ideas in order to bypass this problem. Those algorithms are either problematic in their runtime, or, as in the `AdaBoost.R` case, based on weak learners whose sample complexity depends on the Pseudo-dimension of the class [1], which tends to be so high that weak learners cannot be expected to exist for large classes of functions.

For this reason we use a different notion. Recall the definition

Definition 2.2 ((η, γ)-Weak-Hypothesis)**.** For $\eta \in [0, 1]$ and $\gamma \in [0, 1/2]$, we say that $f : \mathcal{X} \to \mathbb{R}$ is an an (η, γ)-*weak hypothesis* (with respect to distribution

[1] For the definition of the Pseudo-dimension see 5.2

D and target $f^* \in \mathcal{F}$) if

$$\mathbb{P}_{X\sim D}(|f(X) - f^*(X)| > \eta) \leq \frac{1}{2} - \gamma.$$

The (η, γ)-weak-learner, which has appeared, among other works, in Anthony et al. [1996], Simon [1997], Avnimelech and Intrator [1999], Kégl [2003], gets around this difficulty by demanding a bound on the measure of the points in which the hypothesis has "big" local error. Furthermore this notion was in fact proved useful in various, quite simple, boosting mechanisms, but, to our knowledge, provable general constructions of such learners have been lacking. Note that, as in other definitions of weak-learning, this definition also uses a "strong" definition of learning, which was proposed by Simon.

Definition 2.3 ((ε, γ)-good-model)**.** For $\varepsilon, \eta \in [0, 1]$ and $\gamma \in [0, 1/2]$, we say that $f : \mathcal{X} \to \mathbb{R}$ is an an (ε, γ)-*good model* (with respect to distribution D and target $f^* \in \mathcal{F}$) if

$$\mathbb{P}_{X\sim D}(|f(X) - f^*(X)| > \eta) \leq \varepsilon.$$

and a $\mathcal{A}$ is γ-learner if *for every* ε, δ and sample S of size $m = m(\varepsilon, \delta)$, with probability at least $1 - \delta$, $f = \mathcal{A}(S)$ is a (ε, γ)-good-model. So (η, γ)-weak-learner is simply a γ-learner with the error parameter ε fixed, and bounded away from $1/2$.

Although there exist several used of this type of "weak-learning" to our knowledge, there exist no provable constructions of such algorithms. We now present a provable and very natural, namely ERM based, (η, γ)-learner. From this result we are also able to construct our (η, γ)-weak-learner, which was used by our compression-boosting mechanism.

2.2.2 Upper Bound on The Sample Complexity of (ε, γ)-Good-Learning

The following result is stated in the notion of the more general case of (ε, γ)-good-model, in order to apply it into our boosting mechanism we later fix the error parameter ε as was previously discussed, which then yields an Upper Bound on the sample complexity of (ε, γ)-weak-learner.

Define $\rho_\eta(f, g) = P_{2m}(x : |f(x) - g(x)| > \eta)$, where P_{2m} is the empirical measure induced by $X_1, \ldots, X_{2m}$ iid P-distributed random variables (the m data points and m ghost points). Define $N_\eta(\beta)$ as the β-covering numbers of $\mathcal{F}$ under the ρ_η pseudo-metric.

Theorem 2.4. *Fix any* $\eta, \beta \in (0,1)$*,* $\alpha \in [0,1)$*, and* $m \in \mathbb{N}$*. For* $X_1, \ldots, X_m$ *iid* P*-distributed, with probability at least* $1 - \mathbb{E}\left[N_{\eta(1-\alpha)/2}(\beta/8)\right] 2e^{-m\beta/96}$*, every* $f \in \mathcal{F}$ *with* $\max_{1 \leq i \leq m} |f(X_i) - f^*(X_i)| \leq \alpha\eta$ *satisfies* $P(x : |f(x) - f^*(x)| > \eta) \leq \beta$*.*

Proof. This proof roughly follows the usual symmetrization argument for uniform convergence Vapnik and Červonenkis [1971], Haussler [1992], with a few important modifications to account for this (η, β)-based criterion. If $\mathbb{E}\left[N_{\eta(1-\alpha)/2}(\beta/8)\right]$ is infinite, then the result is trivial, so let us suppose it is finite for the remainder of the proof. Similarly, if $m < 8/\beta$, then $2e^{-m\beta/96} > 1$ and hence the claim trivially holds, so let us suppose $m \geq 8/\beta$ for the remainder of the proof. Without loss of generality, suppose $f^*(x) = 0$ everywhere and every $f \in \mathcal{F}$ is non-negative (otherwise subtract f^* from every $f \in \mathcal{F}$ and redefine $\mathcal{F}$ as the absolute values of the differences; note that this transformation does not increase the value of $N_{\eta(1-\alpha)/2}(\beta/8)$ since applying this transformation to the original $N_{\eta(1-\alpha)/2}(\beta/8)$ functions remains a cover).

Let $X_1, \ldots, X_{2m}$ be iid P-distributed. Denote by P_m the empirical measure induced by $X_1, \ldots, X_m$, and by P'_m the empirical measure induced by $X_{m+1}, \ldots, X_{2m}$. We have

$$
\begin{aligned}
&\mathbb{P}(\exists f \in \mathcal{F} : P'_m(x : f(x) > \eta) > \beta/2 \text{ and } P_m(x : f(x) \leq \alpha\eta) = 1) \\
&\geq \mathbb{P}\left(\exists f \in \mathcal{F} : P(x : f(x) > \eta) > \beta \text{ and } P_m(x : f(x) \leq \alpha\eta) = 1 \text{ and } P'_m(x : f(x) > \eta) > \beta/2\right).
\end{aligned}
$$

Denote by A_m the event that there exists $f \in \mathcal{F}$ satisfying $P(x : f(x) > \eta) > \beta$ and $P_m(x : f(x) \leq \alpha\eta) = 1$, and on this event let $\tilde{f}$ denote such an $f \in \mathcal{F}$ (chosen solely based on $X_1, \ldots, X_m$); when A_m fails to hold, take $\tilde{f}$ to be some arbitrary fixed element of $\mathcal{F}$. Then the expression on the right hand side above is at least as large as

$$
\mathbb{P}\left(A_m \text{ and } P'_m(x : \tilde{f}(x) > \eta) > \beta/2\right),
$$

and noting that the event A_m is independent of $X_{m+1}, \ldots, X_{2m}$, this equals

$$
\mathbb{E}\left[\mathbb{I}_{A_m} \cdot \mathbb{P}\left(P'_m(x : \tilde{f}(x) > \eta) > \beta/2 \middle| X_1, \ldots, X_m\right)\right]. \tag{2.1}
$$

Then note that for any $f \in \mathcal{F}$ with $P(x : f(x) > \eta) > \beta$, a Chernoff bound implies

$$\mathbb{P}\Big(P'_m(x : f(x) > \eta) > \beta/2\Big)$$
$$= 1 - \mathbb{P}\Big(P'_m(x : f(x) > \eta) \leq \beta/2\Big) \geq 1 - \exp\{-m\beta/8\} \geq \frac{1}{2},$$

where we have used the assumption that $m \geq \frac{8}{\beta}$ here. In particular, this implies that the expression in (2.1) is no smaller than $\frac{1}{2}\mathbb{P}(A_m)$. Altogether, we have established that

$$\mathbb{P}(\exists f \in \mathcal{F} : P(x : f(x) > \eta) > \beta \text{ and } P_m(x : f(x) \leq \alpha\eta) = 1)$$
$$\leq 2\mathbb{P}(\exists f \in \mathcal{F} : P'_m(x : f(x) > \eta) > \beta/2 \text{ and } P_m(x : f(x) \leq \alpha\eta) = 1)\,. \quad (2.2)$$

Now let $\sigma(1), \ldots, \sigma(m)$ be independent random variables (also independent of the data), with $\sigma(i) \sim \mathit{Uniform}(\{i, m+i\})$, and denote $\sigma(m+i)$ as the sole element of $\{i, m+i\} \setminus \{\sigma(i)\}$ for each $i \leq m$. Also denote by $P_{m,\sigma}$ the empirical measure induced by $X_{\sigma(1)}, \ldots, X_{\sigma(m)}$, and by $P'_{m,\sigma}$ the empirical measure induced by $X_{\sigma(m+1)}, \ldots, X_{\sigma(2m)}$. By exchangeability of $(X_1, \ldots, X_{2m})$, the right hand side of (2.2) is equal

$$\mathbb{P}\big(\exists f \in \mathcal{F} : P'_{m,\sigma}(x : f(x) > \eta) > \beta/2 \text{ and } P_{m,\sigma}(x : f(x) \leq \alpha\eta) = 1\big)\,.$$

Now let $\hat{\mathcal{F}} \subseteq \mathcal{F}$ be a minimal subset of $\mathcal{F}$ such that $\max\limits_{f \in \mathcal{F}} \min\limits_{\hat{f} \in \hat{\mathcal{F}}} \rho_{\eta(1-\alpha)/2}(\hat{f}, f) \leq \beta/8$. The size of $\hat{\mathcal{F}}$ is at most $N_{\eta(1-\alpha)/2}(\beta/8)$, which is finite almost surely (since we have assumed above that its expectation is finite). Then note that (denoting by $X_{[2m]} = (X_1, \ldots, X_{2m})$) the above expression is at most

$$\mathbb{P}\Big(\exists f \in \hat{\mathcal{F}} : P'_{m,\sigma}(x : f(x) > \eta(1+\alpha)/2) > (3/8)\beta \text{ and } P_{m,\sigma}(x : f(x) > \eta(1+\alpha)/2) \leq \beta/8\Big)$$
$$\leq \mathbb{E}\Bigg[N_{\eta(1-\alpha)/2}(\beta/8) \max_{f \in \hat{\mathcal{F}}} \mathbb{P}\big(P'_{m,\sigma}(x\!:\!f(x)\!>\!\eta(1+\alpha)/2) > (3/8)\beta$$
$$\text{and } P_{m,\sigma}(x\!:\!f(x)\!>\!\eta(1+\alpha)/2) \leq \beta/8 \big| X_{[2m]}\big)\Bigg]. \quad (2.3)$$

Then note that for any $f \in \mathcal{F}$, we have almost surely

$$\begin{aligned}
&\mathbb{P}\big(P'_{m,\sigma}(x : f(x) > \eta(1+\alpha)/2) > (3/8)\beta \text{ and } P_{m,\sigma}(x : f(x) > \eta(1+\alpha)/2) \leq \beta/8 \big| X_{[2m]}\big) \\
&\leq \mathbb{P}\big(P_{2m}(x : f(x) > \eta(1+\alpha)/2) > (3/16)\beta \text{ and } P_{m,\sigma}(x : f(x) > \eta(1+\alpha)/2) \leq \beta/8 \big| X_{[2m]}\big) \\
&\leq \exp\{-m\beta/96\}\,,
\end{aligned}$$

where the last inequality is by a Chernoff bound, which (as noted by Hoeffding [1963]) remains valid even when sampling without replacement. Together with (2.2) and (2.3), we have that

$$\begin{aligned}
&\mathbb{P}(\exists f \in \mathcal{F} : P(x : f(x) > \eta) > \beta \text{ and } P_m(x : f(x) \leq \alpha\eta) = 1) \\
&\leq 2\mathbb{E}\big[N_{\eta(1-\alpha)/2}(\beta/8)\big]\, e^{-m\beta/96}.
\end{aligned}$$

□

Lemma 2.5. *There exist universal numerical constants $c, c' \in (0, \infty)$ such that $\forall \eta, \beta \in (0,1)$,*

$$N_\eta(\beta) \leq \left(\frac{2}{\eta\beta}\right)^{cd(c'\eta\beta)},$$

where $d(\cdot)$ is the fat-shattering dimension.

Proof. Mendelson and Vershynin [2003, Theorem 1] establishes that the $\eta\beta$-covering number of $\mathcal{F}$ under the $L_2(P_{2m})$ pseudo-metric is at most

$$\left(\frac{2}{\eta\beta}\right)^{cd(c'\eta\beta)} \tag{2.4}$$

for some universal numerical constants $c, c' \in (0, \infty)$. Then note that for any $f, g \in \mathcal{F}$, Markov's and Jensen's inequalities imply $\rho_\eta(f,g) \leq \frac{1}{\eta}\|f-g\|_{L_1(P_{2m})} \leq \frac{1}{\eta}\|f-g\|_{L_2(P_{2m})}$. Thus, any $\eta\beta$-cover of $\mathcal{F}$ under $L_2(P_{2m})$ is also a β-cover of $\mathcal{F}$ under ρ_η, and therefore (2.4) is also a bound on $N_\eta(\beta)$. □

Combining the above two results yields the following theorem.

Theorem 2.6. *For some universal numerical constants $c_1, c_2, c_3 \in (0, \infty)$, for any $\eta, \delta, \beta \in (0,1)$ and $\alpha \in [0,1)$, letting $X_1, \ldots, X_m$ be iid P-distributed, where*

$$m = \left\lceil \frac{c_1}{\beta}\left(d(c_2\eta\beta(1-\alpha)) \ln\!\left(\frac{c_3}{\eta\beta(1-\alpha)}\right) + \ln\!\left(\frac{1}{\delta}\right)\right)\right\rceil,$$

with probability at least $1-\delta$*, every* $f \in \mathcal{F}$ *with* $\max_{i\in[m]} |f(X_i) - f^*(X_i)| \leq \alpha\eta$ *satisfies* $P(x : |f(x) - f^*(x)| > \eta) \leq \beta$*.*

Proof. The result follows immediately from combining Theorem 2.4 and Lemma 2.5. $\square$

In particular, the specific case of weak-learners, as stated in Theorem 1.3, follows immediately from this result by taking $\beta = 1/2 - \gamma$ and $\alpha = \gamma/2$.

2.2.3 Tightness of The Upper Bound

To discuss tightness of Theorem 2.6, we note that in addition to the definition of a (β, η)-*good model* Simon [1997] also proved the following lower bound

Theorem 2.7 (Simon [1997])**.** *Let* A *be an algorithm which learns function class* F *with an* (β, η)*-good model*

1. *If* F *is nontrivial* [2] *,* $\beta < 1/2$ *and* $\eta < \Delta(F)/2$*. then* A *needs* $\Omega(\ln(1/\delta)/\beta)$ *examples.*
2. *If* $\beta \leq 1/8$*,* $0 < \delta \leq 1/100$*. then* A *needs* $\Omega((d_F^N(\eta) - 1)/\beta)$ *examples.*

When $\Delta(F) = sup\{\|g - f\|_\infty \mid \exists x \in X : f(x) = g(x)\}$.

Combining the two we get that a sample complexity lower bound for the same criterion of

$$\Omega\left(\frac{d_F^N(c\eta)}{\beta} + \frac{1}{\beta}\log\frac{1}{\delta}\right),$$

where $d_F^N(\cdot)$ is a quantity somewhat smaller than the fat-shattering dimension, essentially representing a fat Natarajan dimension.

Simon showed that this lower bound is tight and placed an open question

Open Problem: For every function class F there exist an algorithm A which learns F with an (β, η)-good model using

$$O\left(\frac{d_F^N(\eta)}{\beta} + \frac{1}{\beta}\ln(1/\delta)\right)$$

examples.

Thus, aside from the differences in the complexity measure (and a logarithmic factor), we establish an upper bound of a similar form to Simon's lower

[2] *Meaning: there exist* $f, g \in F$ *which are not pairwise disjoin, namely* $\exists x \in X : f(x) = g(x)$*.*

bound and hence making a significant progress towards solving Simon's open question.

Chapter 3

From Boosting to Compression

Generally, our strategy for converting the boosting algorithm `MedBoost` into a sample compression scheme of smaller size follows a strategy of Moran and Yehudayoff for binary classification, based on arguing that because the ensemble makes its predictions with a *margin* (corresponding to the results on *quantiles* in Corollary 2.2), it is possible to recover the same proximity guarantees for the predictions while using only a smaller *subset* of the functions from the original ensemble. Specifically, we use the following general *sparsification* strategy.

For $\alpha_1, \ldots, \alpha_T \in [0,1]$ with $\sum_{t=1}^{T} \alpha_t = 1$, denote by $Cat(\alpha_1, \ldots, \alpha_T)$ the *categorical distribution*: i.e., the discrete probability distribution on $\{1, \ldots, T\}$ with probability mass α_t on t.

Algorithm 2 `Sparsify`$(\{(x_i, y_i)\}_{i\in[m]}, \gamma, T, n)$

Run `MedBoost`$(\{(x_i, y_i)\}_{i\in[m]}, T, \gamma, \eta)$
Let $h_1, \ldots, h_T$ and $\alpha_1, \ldots, \alpha_T$ be its return values
Denote $\alpha'_t = \alpha_t / \sum_{t'=1}^{T} \alpha_{t'}$ for each $t \in [T]$
repeat
    Sample $(J_1, \ldots, J_n) \sim Cat(\alpha'_1, \ldots, \alpha'_T)^n$
    Let $F = \{h_{J_1}, \ldots, h_{J_n}\}$
until $\max_{1\le i\le m} |\{f \in F : |f(x_i) - y_i| > \eta\}| < n/2$
Return F

For any values $a_1, \ldots, a_n$, denote the (unweighted) median

$$Med(a_1, \ldots, a_n) = Median(a_1, \ldots, a_n; 1, \ldots, 1).$$

Our intention in dicussing the above algorithm is to argue that, for a sufficiently large choice of n, the above procedure returns a set $\{f_1, \ldots, f_n\}$ such that

$$\forall i \in [m], |Med(f_1(x_i), \ldots, f_n(x_i)) - y_i| \leq \eta.$$

We analyze this strategy separately for binary classification and real-valued functions, since the argument in the binary case is much simpler (and demonstrates more directly the connection to the original argument of Moran and Yehudayoff), and also because we arrive at a tighter result for binary functions than for real-valued functions.

3.1 Binary Classification

We begin with the simple observation about binary classification (i.e., where the functions in $\mathcal{F}$ all map into $\{0, 1\}$). The technique here is quite simple, and follows a similar line of reasoning to the original argument of Moran and Yehudayoff. The argument for real-valued functions below will diverge from this argument in several important ways, but the high level ideas remain the same.

The compression function is essentially the one introduced by Moran and Yehudayoff, except applied to the classifiers produced by the above `Sparsify` procedure, rather than a set of functions selected by a minimax distribution over all classifiers produced by $O(d)$ samples each. The weak hypotheses in `MedBoost` for binary classification can be obtained using samples of size $O(d)$. Thus, if the `Sparsify` procedure is successful in finding n such classifiers whose median predictions are within η of the target y_i values for all i, then we may encode these n classifiers as a compression set, consisting of the set of $k = O(nd)$ samples used to train these classifiers, together with $k \log k$ extra bits to encode the order of the samples.[1] To obtain Theorem 1.1, it then suffices to argue that $n = \Theta(d^*)$ is a sufficient value. The proof follows.

Proof of Theorem 1.1. Recall that d^* bounds the VC dimension of the class of sets $\{\{h_t : t \leq T, h_t(x_i) = 1\} : 1 \leq i \leq m\}$. Thus for the iid samples

[1] In fact, $k \log n$ bits would suffice if the weak learner is permutation-invariant in its data set.

$h_{J_1}, \ldots, h_{J_n}$ obtained in `Sparsify`, for $n = 64(2309 + 16d^*) > \frac{2304+16d^*+\log(2)}{1/8}$, by the VC uniform convergence inequality of Vapnik and Červonenkis [1971], with probability at least $1/2$ we get that

$$\max_{1 \le i \le m} \left| \left(\frac{1}{n} \sum_{j=1}^{n} h_{J_j}(x_i) \right) - \left(\sum_{t=1}^{T} \alpha' h_t(x_i) \right) \right| < 1/8.$$

In particular, if we choose $\gamma = 1/8$, $\eta = 1$, and $T = \Theta(\log(m))$ appropriately, then Corollary 2.2 implies that every $y_i = \mathbb{I}\left[\sum_{t=1}^{T} \alpha' h_t(x_i) \ge 1/2\right]$ and $\left|\frac{1}{2} - \sum_{t=1}^{T} \alpha' h_t(x_i)\right| \ge 1/8$ so that the above event would imply every $y_i = \mathbb{I}\left[\frac{1}{n}\sum_{j=1}^{n} h_{J_j}(x_i) \ge 1/2\right] = Med(h_{J_1}(x_i), \ldots, h_{J_n}(x_i))$. Note that the `Sparsify` algorithm need only try this sampling $\log_2(1/\delta)$ times to find such a set of n functions. Combined with the description above (from Moran and Yehudayoff, 2016) of how to encode this collection of h_{J_i} functions as a sample compression set plus side information, this completes the construction of the sample compression scheme. □

3.2 Real-Valued Functions

Next we turn to the general case of real-valued functions (where the functions in $\mathcal{F}$ may generally map into $[0, 1]$). We have the following result, which says that the `Sparsify` procedure can reduce the ensemble of functions from one with $T = O(\log(m)/\gamma^2)$ functions in it, down to one with a number of functions *independent of* m.

Theorem 3.1. *Choosing*

$$n = \Theta\left(\frac{1}{\gamma^2} d^*(c\eta) \log^2(d^*(c\eta)/\eta)\right)$$

suffices for the ***`Sparsify`*** *procedure to return* $\{f_1, \ldots, f_n\}$ *with*

$$\max_{1 \le i \le m} |Med(f_1(x_i), \ldots, f_n(x_i)) - y_i| \le \eta.$$

Proof. Recall from Corollary 2.2 that `MedBoost` returns functions $h_1, \ldots, h_T \in \mathcal{F}$ and $\alpha_1, \ldots, \alpha_T \ge 0$ such that $\forall i \in \{1, \ldots, m\}$,

$$\max\left\{ \left| Q^+_{\gamma/2}(x_i) - y_i \right|, \left| Q^-_{\gamma/2}(x_i) - y_i \right| \right\} \le \eta/2,$$

where $\{(x_i, y_i)\}_{i=1}^m$ is the training data set. We use this property to sparsify $h_1, \ldots, h_T$ from $T = O(\log(m)/\gamma^2)$ down to k elements, where k will depend on η, γ, and the dual fat-shattering dimension of $\mathcal{F}$ (actually, just of $H = \{h_1, \ldots, h_T\} \subseteq \mathcal{F}$) — but **not** sample size m.

Letting $\alpha'_j = \alpha_j / \sum_{t=1}^T \alpha_t$ for each $j \leq T$, we will sample k hypotheses $\left\{\tilde{h}_1, \ldots, \tilde{h}_k\right\} =: \tilde{H} \subseteq H$ with each $\tilde{h}_i = h_{J_i}$, where $(J_1, \ldots, J_k) \sim Cat(\alpha'_1, \ldots, \alpha'_T)^k$ as in `Sparsify`. Define a function $\hat{h}(x) = Med(\tilde{h}_1(x), \ldots, \tilde{h}_k(x))$. We claim that for any fixed $i \in [m]$, with high probability

$$|\hat{h}(x_i) - f^*(x_i)| \leq \eta/2. \tag{3.1}$$

Indeed, partition the indices $[T]$ into the disjoint sets

$$\begin{aligned}
L(x) &= \left\{j \in [T] : h_j(x) < Q_\gamma^-(h_1(x), \ldots, h_T(x); \alpha_1, \ldots, \alpha_T)\right\}, \\
M(x) &= \left\{j \in [T] : Q_\gamma^-(h_1(x), ..., h_T(x); \alpha_1, ..., \alpha_T) \leq h_j(x) \leq Q_\gamma^+(h_1(x), ..., h_T(x); \alpha_1, ..., \alpha_T)\right\}, \\
R(x) &= \left\{j \in [T] : h_j(x) > Q_\gamma^+(h_1(x), \ldots, h_T(x); \alpha_1, \ldots, \alpha_T)\right\}.
\end{aligned}$$

Then the only way (3.1) can fail is if half or more indices $J_1, \ldots, J_k$ sampled fall into $R(x_i)$ — or if half or more fall into $L(x_i)$. Since the sampling distribution puts mass less than $1/2 - \gamma$ on each of $R(x_i)$ and $L(x_i)$, Chernoff's bound puts an upper estimate of $\exp(-2k\gamma^2)$ on either event. Hence,

$$\mathbb{P}\left(|\hat{h}(x_i) - f^*(x_i)| > \eta/2\right) \leq 2\exp(-2k\gamma^2). \tag{3.2}$$

Next, our goal is to ensure that with high probability, (3.1) holds simultaneously for all $i \in [m]$. Define the map $\boldsymbol{\xi} : [m] \to \mathbb{R}^k$ by $\boldsymbol{\xi}(i) = (\tilde{h}_1(x_i), \ldots, \tilde{h}_k(x_i))$. Let $G \subseteq [m]$ be a minimal subset of $[m]$ such that

$$\max_{i \in [m]} \min_{j \in G} \|\boldsymbol{\xi}(i) - \boldsymbol{\xi}(j)\|_\infty \leq \eta/2.$$

This is just a minimal ℓ_∞ covering of $[m]$. Then

$$\begin{aligned}
&\mathbb{P}\left(\exists i \in [m] : |Med(\boldsymbol{\xi}(i)) - f^*(x_i)| > \eta\right) \leq \\
&\sum_{j \in G} \mathbb{P}\left(\exists i : |Med(\boldsymbol{\xi}(i)) - f^*(x_i)| > \eta, \|\boldsymbol{\xi}(i) - \boldsymbol{\xi}(j)\|_\infty \leq \eta/2\right) \leq \\
&\sum_{j \in G} \mathbb{P}\left(|Med(\boldsymbol{\xi}(j)) - f^*(x_j)| > \eta/2\right) \leq 2N_\infty([m], \eta/2)\exp(-2k\gamma^2),
\end{aligned}$$

where $N_\infty([m], \eta/2)$ is the $\eta/2$-covering number (under ℓ_∞) of $[m]$, and we used the fact that

$$|Med(\boldsymbol{\xi}(i)) - Med(\boldsymbol{\xi}(j))| \le \|\boldsymbol{\xi}(i) - \boldsymbol{\xi}(j)\|_\infty .$$

Finally, to bound $N_\infty([m], \eta/2)$, note that $\boldsymbol{\xi}$ embeds $[m]$ into the dual class $\mathcal{F}^*$. Thus, we may apply the bound in [Rudelson and Vershynin, 2006, Display (1.4)]:

$$\log N_\infty([m], \eta/2) \le Cd^*(c\eta)\log^2(k/\eta),$$

where C, c are universal constants and $d^*(\cdot)$ is the dual fat-shattering dimension of $\mathcal{F}$. It now only remains to choose a k that makes $\exp\left(Cd^*(c\eta)\log^2(k/\eta) - 2k\gamma^2\right)$ as small as desired. □

To establish Theorem 1.2, we use the weak learner from above, with the booster `MedBoost` from Kégl, and then apply the `Sparsify` procedure. Combining the corresponding theorems, together with the same technique for converting to a compression scheme discussed above for classification (i.e., encoding the functions with the set of training examples they were obtained from, plus extra bits to record the order and which examples which weak hypothesis was obtained by training on), this immediately yields the result claimed in Theorem 1.2, which represents our main new result for sample compression of general families of real-valued functions.

3.3 Examples

As an example for the generality and usefulness of the above schemes, we present two interesting and efficient compression schemes than can be derived from it. the main technical result needed in order to apply our method to those cases was to find and prove and dual Fat-Shattering dimension of the function-classes at hand, a problem which isn't trivial most of the time, required using tools from various domains. Leveraging novel and relatively-new algorithmic results from learning theory yields the final wanted compression-schemes.

3.3.1 Sample compression for BV functions

The function class $\mathrm{BV}(v)$ consists of all $f : [0, 1] \to \mathbb{R}$ for which

$$V(f) := \sup_{n\in\mathbb{N}} \sup_{0=x_0<x_1<\ldots<x_n=1} \sum_{i=1}^{n-1} |f(x_{i+1}) - f(x_i)| \le v.$$

It is known [Anthony and Bartlett, 1999, Theorem 11.12] that $d_{\mathrm{BV}(v)}(t) = 1 + \lfloor v/(2t) \rfloor$. In Theorem 3.3 below, we show that the dual class has $d^*_{\mathrm{BV}(v)}(t) = \Theta\left(\log(v/t)\right)$. Long [2004] presented an efficient, proper, consistent learner for the class $\mathcal{F} = \mathrm{BV}(1)$ with range restricted to $[0,1]$, with sample complexity $m_{\mathcal{F}}(\varepsilon,\delta) = O(\frac{1}{\varepsilon}\log\frac{1}{\delta})$. Combined with Theorem 1.2, this yields

Corollary 3.2. *Let $\mathcal{F} = \mathrm{BV}(1) \cap [0,1]^{[0,1]}$ be the class $f : [0,1] \to [0,1]$ with $V(f) \le 1$. Then the proper, consistent learner $\mathcal{L}$ of Long [2004], with target generalization error ε, admits a sample compression scheme of size $O(k \log k)$, where*

$$k = O\left(\frac{1}{\varepsilon}\log^2\frac{1}{\varepsilon} \cdot \log\left(\frac{1}{\varepsilon}\log\frac{1}{\varepsilon}\right)\right).$$

The compression set is computable in expected runtime

$$O\left(n\frac{1}{\varepsilon^{3.38}}\log^{3.38}\frac{1}{\varepsilon}\left(\log n + \log\frac{1}{\varepsilon}\log\left(\frac{1}{\varepsilon}\log\frac{1}{\varepsilon}\right)\right)\right).$$

The remainder of this section is devoted to proving

Theorem 3.3. *For $\mathcal{F} = \mathrm{BV}(v)$ and $t < v$, we have $d^*_{\mathcal{F}}(t) = \Theta\left(\log(v/t)\right)$.*

First, we define some preliminary notions:

Definition 3.1. For a binary $m \times n$ matrix M, define

$$\begin{aligned} V(M,i) &:= \sum_{j=1}^{m} \mathbb{I}[M_{j,i} \neq M_{j+1,i}], \\ G(M) &:= \sum_{i=1}^{n} V(M,i), \\ V(M) &:= \max_{i\in[n]} V(M,i). \end{aligned}$$

Lemma 3.4. *Let M be a binary $2^n \times n$ matrix. If for each $b \in \{0,1\}^n$ there is a row j in M equal to b, then*

$$V(M) \ge \frac{2^n}{n}.$$

In particular, for at least one row i, we have $V(M,i) \ge 2^n/n$.

Proof. Let M be a $2^n \times n$ binary such that for each $b \in \{0,1\}^n$ there is a row j in M equal to b. Given M's dimensions, every $b \in \{0,1\}^n$ appears exactly in one row of M, and hence the minimal Hamming distance between two rows is 1. Summing over the $2^n - 1$ adjacent row pairs, we have

$$G(M) = \sum_{i=1}^{n} V(M,i) = \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbb{I}[M_{j,i} \neq M_{j+1,i}] \geq 2^n - 1,$$

which averages to

$$\frac{1}{n} \sum_{i=1}^{n} V(M,i) = \frac{G(M)}{n} \geq \frac{2^n - 1}{n}.$$

By the pigeon-hole principle, there must be a row $j \in [n]$ for which $V(M,i) \geq \frac{2^n-1}{n}$, which implies $V(M) \geq \frac{2^n-1}{n}$. □

We split the proof of Theorem 3.3 into two estimates:

Lemma 3.5. *For* $\mathcal{F} = \mathrm{BV}(v)$ *and* $t < v$, $d^*_{\mathcal{F}}(t) \leq 2\log_2(v/t)$.

Lemma 3.6. *For* $\mathcal{F} = \mathrm{BV}(v)$ *and* $4t < v$, $d^*_{\mathcal{F}}(t) \geq \lfloor \log_2(v/t) \rfloor$.

Proof of Lemma 3.5. Let $\{f_1, \ldots, f_n\} \subset \mathcal{F}$ be a set of functions that are t-shattered by $\mathcal{F}^*$. In other words, there is an $r \in \mathbb{R}^n$ such that for each $b \in \{0,1\}^n$ there is an $x_b \in \mathcal{F}^*$ such that

$$\forall i \in [n], x_b(f_i) \begin{cases} \geq r_i + t, & b_i = 1 \\ \leq r_i - t, & b_i = 0 \end{cases}.$$

Let us order the x_bs by magnitude $x_1 < x_2 < \ldots < x_{2^n}$, denoting this sequence by $(x_i)_{i=1}^{2^n}$. Let $M \in \{0,1\}^{2^n \times n}$ be a matrix whose ith row is b_j, the latter ordered arbitrarily.

By Lemma 3.4, there is $i \in [n]$ s.t.

$$\sum_{j=1}^{2^n} \mathbb{I}[M(j,i) \neq M(j+1,i)] \geq \frac{2^n}{n}.$$

Note that if $M(j,i) \neq M(j+1,i)$ shattering implies that

$$x_j(f_i) \geq r_i + t \text{ and } x_{j+1}(f_i) \leq r_i - t$$

or

$$x_j(f_i) \leq r_i - t \text{ and } x_{j+1}(f_i) \geq r_i + t;$$

either way,

$$|f_i(x_j) - f_i(x_{j+1})| = |x_j(f_i) - x_{j+1}(f_i)| \geq 2t.$$

So for the function f_i, we have

$$\sum_{j=1}^{2^n} |f_i(x_j) - f_i(x_{j+1})| = \sum_{j=1}^{2^n} |x_j(f_i) - x_{j+1}(f_i)| \geq \sum_{j=1}^{2^n} \mathbb{I}[b_{j_i} \neq b_{j+1_i} \cdot 2t \geq \frac{2^n}{n} \cdot 2t.$$

As $\{x_j\}_{j=1}^{2^n}$ is a partition of $[0,1]$ we get

$$v \geq \sum_{j=1}^{2^n} |f_i(x_j) - f_i(x_{j+1})| \geq \frac{t2^{n+1}}{n} \geq t2^{n/2}$$

and hence

$$v/t \geq 2^{n/2}$$

$$\Rightarrow 2\log_2(v/t) \geq n.$$

□

Proof of Lemma 3.6. We construct a set of $n = \lfloor \log_2(v/t) \rfloor$ functions that are t-shattered by $\mathcal{F}^*$. First, we build a balanced Gray code [Flahive and Bose, 2007] with n bits, which we arrange into the rows of M. Divide the unit interval into 2^n segments and define, for each $j \in [2^n]$,

$$x_j := \frac{j}{2^n}.$$

Define the functions $f_1, \ldots, , f_{\lfloor \log_2(v/t) \rfloor}$ as follows:

$$f_i(x_j) = \begin{cases} t, & M(j,i) = 1 \\ -t, & M(j,i) = 0 \end{cases}.$$

We claim that each $f_i \in \mathcal{F}$. Since M is balanced Gray code,

$$V(M) = \frac{2^n}{n} \leq \frac{v}{t\log_2(v/t)} \leq \frac{v}{2t}.$$

Hence, for each f_i, we have

$$V(f_i) \leq 2tV(M,i) \leq 2t\frac{v}{2t} = v.$$

Next, we show that this set is shattered by $\mathcal{F}^*$. Fix the trivial offest $r_1 = ... = r_n = 0$ For every $b \in \{0,1\}^n$ there is a $j \in [2^n]$ s.t. $b = b_i$. By construction, for every $i \in [n]$, we have

$$x_j(f_i) = f_i(x_j) = \begin{cases} t \geq r_i + t, & M(j,i) = 1 \\ -t \leq r_i - t, & M(j,i) = 0 \end{cases}.$$

□

3.3.2 Sample compression for nearest-neighbor regression

Let $(\mathcal{X}, \rho)$ be a metric space and define, for $L \geq 0$, the collection $\mathcal{F}_L$ of all $f : \mathcal{X} \to [0,1]$ satisfying

$$|f(x) - f(x')| \leq L\rho(x, x');$$

these are the L-Lipschitz functions. Gottlieb et al. [2017b] showed that

$$d_{\mathcal{F}_L}(t) = O\left(\lceil L \operatorname{diam}(X)/t \rceil^{\operatorname{ddim}(\mathcal{X})}\right),$$

where $\operatorname{diam}(\mathcal{X})$ is the diameter and ddim is the *doubling dimension*, defined therein. The proof is achieved via a packing argument, which also shows that the estimate is tight. Below we show that $d^*_{\mathcal{F}_L}(t) = \Theta(\log(M(\mathcal{X}, 2t/L)))$, where $M(\mathcal{X}, \cdot)$ is the packing number of $(\mathcal{X}, \rho)$. Applying this to the efficient nearest-neighbor regressor[2] of Gottlieb et al. [2017a], we obtain

Corollary 3.7. *Let $(\mathcal{X}, \rho)$ be a metric space with hypothesis class $\mathcal{F}_L$, and let $\mathcal{L}$ be a consistent, proper learner for $\mathcal{F}_L$ with target generalization error ε. Then $\mathcal{L}$ admits a compression scheme of size $O(k \log k)$, where*

$$k = O\left(D(\varepsilon) \log \frac{1}{\varepsilon} \cdot \log D(\varepsilon) \log\left(\frac{1}{\varepsilon} \log D(\varepsilon)\right)\right)$$

and

$$D(\varepsilon) = \left\lceil \frac{L \operatorname{diam}(\mathcal{X})}{\varepsilon} \right\rceil^{\operatorname{ddim}(\mathcal{X})}.$$

We now prove our estimate on the dual fat-shattering dimension of $\mathcal{F}$:

[2] In fact, the technical machinery in Gottlieb et al. [2017a] was aimed at achieving *approximate* Lipschitz-extension, so as to gain a considerable runtime speedup. An *exact* Lipschitz extension is much simpler to achieve. It is more computationally costly but still polynomial-time in sample size.

Lemma 3.8. *For* $\mathcal{F} = \mathcal{F}_L$*,* $d^*_{\mathcal{F}}(t) \leq \log_2(\mathcal{M}(\mathcal{X}, 2t/L))$.

Proof. Let $\{f_1, \ldots, f_n\} \subset \mathcal{F}_L$ a set that is t-shattered by $\mathcal{F}^*_L$. For $b \neq b' \in \{0,1\}^n$, let i be the first index for which $b_i \neq b'_i$, say, $b_i = 1 \neq 0 = b'$. By shattering, there are points $x_b, x_{b'} \in \mathcal{F}^*_L$ such that $x_b(f_i) \geq r_i + t$ and $x_{b'}(f_i) \leq r_i - t$, whence

$$f_i(x_b) - f_i(x_{b'}) \geq 2t$$

and

$$L\rho(x_b, x_{b'}) \geq f_i(x_b) - f_i(x_{b'}) \geq 2t.$$

It follows that for $b \neq b' \in \{0,1\}^n$, we have $\rho(x_b, x_{b'}) \geq 2t/L$. Denoting by $M(\mathcal{X}, \varepsilon)$ the ε-packing number of $\mathcal{X}$, we get

$$2^n = |\{x_b \mid b \in \{0,1\}^n\}| \leq \mathcal{M}(\mathcal{X}, 2t/L).$$

□

Lemma 3.9. *For* $\mathcal{F} = \mathcal{F}_L$ *and* $t < L$*,* $d^*_{\mathcal{F}}(t) \geq \log_2(\mathcal{M}(\mathcal{X}, 2t/L))$.

Proof. Let $S = \{x_1, ..., x_m\} \subseteq \mathcal{X}$ be a maximal $2t/L$-packing of $\mathcal{X}$. Suppose that $c : S \to \{0,1\}^{\lfloor \log_2 m \rfloor}$ is one-to-one. Define the set of function $F = \{f_1, \ldots, f_{\lfloor \log_2(m) \rfloor}\} \subseteq \mathcal{F}_L$ by

$$f_i(x_j) = \begin{cases} t, & c(x_j)_i = 1 \\ -t, & c(x_j)_i = 0 \end{cases}.$$

For every $f \in F$ and every two points $x, x' \in S$ it holds that

$$|f(x) - f(x')| \leq 2t = L \cdot 2t/L \leq L\rho(x, x').$$

This set of functions is t-shattered by S and is of size $\lfloor \log_2 m \rfloor = \lfloor \log_2(\mathcal{M}(\mathcal{X}, 2t/L)) \rfloor$.

□

Chapter 4

Agnostic-Compressable loss functions

4.1 Problem setting, definitions and notation

Our instance space is $\mathcal{X} = \mathbb{R}^d$, label space is $\mathcal{Y} = \mathbb{R}$, and hypothesis class is $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$, consisting of all $h_{\mathbf{a},b} : \mathcal{X} \to \mathcal{Y}$ given by $h_{\mathbf{a},b}(\mathbf{x}) = \langle \mathbf{a}, \mathbf{x} \rangle + b$, indexed by $\mathbf{a} \in \mathbb{R}^d, b \in \mathbb{R}$. For $1 \leq p < \infty$, the loss incurred by a hypothesis $h \in \mathcal{F}$ on a labeled sample $S = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m))$ is given by

$$L_p(h, S) := \frac{1}{m} \sum_{i=1}^{m} |h(\mathbf{x}_i) - y_i|^p,$$

while for $p = \infty$,

$$L_\infty(h, S) := \max_{1 \leq i \leq m} |h(\mathbf{x}_i) - y_i|.$$

Following David et al. [2016] , let $S = (x_1, y_i), \ldots, (x_m, y_m)$ be a tagged sample drawn i.i.d from some unknown distribution, an let $l : \mathcal{X} \times \mathbb{R} \to \mathbb{R}$ be some loss function. We say that (κ, ρ) is an *agnostic sample compression scheme* for $\mathcal{H}$ if, for every sample S, $f_S := \rho(\kappa(S))$, achieves $\mathcal{F}$-competitive empirical loss:

$$L_p(f_S, S) \quad \leq \quad \inf_{f \in \mathcal{F}} L_p(f, S).$$

In principle, the *size* k of an agnostic compression scheme may depend on the data set size m, in which case we may denote this dependence by $k(m)$. However, in this work we are primarily interested in the case when $k(m)$ is *bounded*: that is, $k(m) \leq k$ for some m-independent value k. Note that the above definition is fully general, in that it defines a notion of agnostic compression scheme for *any* function class $\mathcal{F}$ and loss function L, though in the present work we focus on $\mathcal{F}$ as linear functions in $\mathbb{R}^d$ and the loss as L_p for $1 \leq p \leq \infty$.

Remark. At first, it might seem unclear why this is an appropriate generalization of sample compression to the agnostic setting. To see that it is so, we note that one of the main interests in sample compression schemes is their ability to *generalize*. More formally: Denoting the *excess risk* of a learner to be

$$R := \mathbb{E}_S[L_p(f_S, S)] - \inf_{f \in \mathcal{F}} \mathbb{E}_S[L_p(f_S, S)],$$

we can say that sample-compression-schemes based learners achieve low excess-risk under a *distribution* P on $\mathcal{X} \times \mathcal{Y}$ when the data S are sampled iid according to P [Littlestone and Warmuth, 1986, Floyd and Warmuth, 1995, Graepel, Herbrich, and Shawe-Taylor, 2005]. Also, as mentioned, in this work we are primarily interested in sample compression schemes that have *bounded size*: $k(m) \leq k$ for an m-independent value k. Furthermore, we are also focusing on the most-general case, where this size bound should be independent of everything else in the scenario, such as the data S or the underlying distribution P. Given these interests, we claim that the above definition is essentially the only reasonable choice. More specifically, for L_p loss with $1 \leq p < \infty$, any compression scheme with $k(m)$ bounded such that its expected excess risk under any P converges to 0 as $m \to \infty$ necessarily satisfies the above condition (or is easily converted into one that does). To see this, note that for any data set S for which such a compression scheme fails to satisfy the above $\mathcal{F}$-competitive empirical loss criterion, we can define a distribution P that is simply uniform on S, and then the compression scheme's selection function would be choosing a bounded number of points from S and a bounded number of bits, while guaranteeing that excess risk under P approaches 0, or equivalently, excess empirical loss approaches 0. To make this argument fully formal, only a slight modification is needed, to handle having multiple copies of points from S in the compression set; given that the size is bounded, these repetitions can be encoded in a bounded number of extra bits, so that we can stick to strictly distinct points in the compression set.

In the converse direction, we also note that any bounded-size agnostic compression scheme (in the sense of the above definition) will be guaranteed to have excess risk under P converging to 0 as $m \to \infty$, in the case that S is sampled iid according to P, for losses L_p with $1 \leq p < \infty$, as long as P guarantees that $(X, Y) \sim P$ has Y bounded (almost surely). This follows from classic arguments about the generalization ability of compression schemes, which includes results for the agnostic case [Graepel, Herbrich, and Shawe-Taylor, 2005]. For unbounded Y one cannot, in general, obtain distribution-free generalization bounds. However, one can still obtain generalization under certain broader restrictions (see, e.g., Mendelson, 2015 and references therein). The generalization problem becomes more subtle for the L_∞ loss: this cannot be expressed as a sum of pointwise losses and there are no standard techniques for bounding the deviation of the sample risk from the true risk. Our above results, in particular the "hybrid-error" analysis on Theorem 2.4, can produce such some insight about the guarantee achieved by minimizing empirical L_∞ loss. We leave this connection for our future research.

We denote set cardinality by $|\cdot|$ and $[m] := \{1, \ldots, m\}$. Vectors $\mathbf{v} \in \mathbb{R}^d$ are denoted by boldface, and their jth coordinate is indicated by $\mathbf{v}(j)$. (Thus, $\mathbf{v}_i(j)$ indicates the jth coordinate of the ith vector in a sequence.)

4.2 Impossibility results for ℓ_p, $1 < p < \infty$

David et al. [2016, Theorem 4.1] proved an impossibility result for the ℓ_2 loss:

Theorem 4.1 (David et al. [2016])**.** *There is no agnostic sample compression scheme for zero-dimensional linear regression with size $k(m) \leq m/2$.*

We show that constant-size compression is impossible for all ℓ_p losses with $1 < p < \infty$:

Theorem 4.2. *There is no agnostic sample compression scheme for zero-dimensional linear regression under ℓ_p loss, $1 < p < \infty$, with size $k(m) < \log(m)$.*

Proof. Consider a sample $(y_1, \ldots, y_m) \in \{0, 1\}^m$. Partition the indices $i \in [m]$ into $S_0 := \{i \in [m] : y_i = 0\}$ and $S_1 := \{i \in [m] : y_i = 1\}$. The empirical risk minimizer is given by

$$\hat{r} := \operatorname*{argmin}_{s \in \mathbb{R}} \sum_{i=1}^{m} |y_i - s|^p.$$

To obtain an explicit expression for $\hat{r}$, define

$$F(s) = \sum_{i=1}^{m} |y_i - s|^p = |S_1|(1-s)^p + |S_0|s^p =: N_1(1-s)^p + N_0 s^p.$$

We then compute

$$F'(s) = pN_0 s^{p-1} - pN_1(1-s)^{p-1}$$

and find that $F'(s) = 0$ occurs at

$$\hat{s} = \frac{\mu^{1/(p-1)}}{1+\mu^{1/(p-1)}},$$

where $\mu = N_1/N_0$. A straightforward analysis of the second derivative shows that $\hat{s} = \hat{r}$ is indeed the unique minimizer of F.

Thus, given a sample of size m, the unique minimizer $\hat{r}$ is uniquely determined by N_0 — which can take on any of integer $m+1$ values between 0 and m. On the other hand, every output of a k-selection function κ outputs a multiset $\hat{S} \subseteq S$ of size k' and a binary string of length $k'' = k - k'$. Thus, the total number of values representable by a k-selection scheme is at most

$$\sum_{k'=0}^{k} k' 2^{k-k'} < 2^{k+1} - k,$$

which, for $k < \log m$, is less than m. □

Remark. A more refined analysis, along the lines of David et al. [2016, Theorem 4.1], should yield a lower bound of $k = \Omega(m)$. A technical complication is that unlike the $p = 2$ case, whose empirical risk minimizer has a simple explicit form, the general ℓ_p loss does not admit a closed-form solution and uniqueness must be argued from general convexity principles. We leave this for our future research.

4.3 Compressibility results for ℓ_1 and ℓ_∞

In sharp contrast with the $1 < p < \infty$ case, we show that in $\mathbb{R}^d$, agnostic linear regression admits a compression scheme of size $d+1$ under ℓ_1 and $d+2$ under ℓ_2.

Theorem 4.3. *There exists an efficiently computable compression scheme for agnostic linear regression in* $\mathbb{R}^d$ *under the* ℓ_1 *loss of size* $d+1$.

Proof. We start with $d = 0$. The sample then consists of $(y_1, \ldots, y_m)$ [formally: pairs (x_i, y_i), where $x_i \equiv 0$], and $\mathcal{F} = \mathbb{R}$ [formally, all functions $h : 0 \mapsto \mathbb{R}$]. We define f_S to be the median of $(y_1, \ldots, y_m)$, which for odd m is defined uniquely and for even m can be taken arbitrarily as the smaller of the two midpoints. It is well-known that such a choice minimizes the empirical ℓ_1 risk, and it clearly constitutes a compression scheme of size 1.

The case $d = 1$ will require more work. The sample consists of $(x_i, y_i)_{i \in [m]}$, where $x_i, y_i \in \mathbb{R}$, and $\mathcal{F} = \{\mathbb{R} \ni x \mapsto ax + b : a, b \in \mathbb{R}\}$. Let $(a^\star, b^\star)$ be a (possibly non-unique) minimizer of

$$L(a,b) := \sum_{i \in [m]} |(ax_i + b) - y_i|, \tag{4.1}$$

achieving the value $L^\star$. We claim that we can always find two indices $\hat{\imath}, \hat{\jmath} \in [m]$ such that the line determined by $(x_{\hat{\imath}}, y_{\hat{\imath}})$ and $(x_{\hat{\jmath}}, y_{\hat{\jmath}})$ also achieves the optimal empirical risk $L^\star$. More precisely, the line $(\hat{a}, \hat{b})$ induced by $((x_{\hat{\imath}}, y_{\hat{\imath}}), (x_{\hat{\jmath}}, y_{\hat{\jmath}}))$ via[1] $\hat{a} = (y_{\hat{\jmath}} - y_{\hat{\imath}})/(x_{\hat{\jmath}} - x_{\hat{\imath}})$ and $\hat{b} = y_{\hat{\imath}} - \hat{a}x_{\hat{\imath}}$, verifies $L(\hat{a}, \hat{b}) = L^\star$.

To prove this claim, we begin by recasting (5.1) as a linear program:

$$\begin{aligned} \min_{(\varepsilon_1, \ldots, \varepsilon_m, a, b) \in \mathbb{R}^{m+2}} \quad & \sum_{i=1}^m \varepsilon_i \quad \text{s.t.} \\ \forall i \in [m] \quad & \varepsilon_i \geq 0 \\ \forall i \in [m] \quad & ax_i + b - y \leq \varepsilon_i \\ \forall i \in [m] \quad & -ax_i - b + y \leq \varepsilon_i. \end{aligned} \tag{4.2}$$

We observe that the linear program in (4.2) is feasible with a finite solution (and actually, the constraints $\varepsilon_i \geq 0$ are redundant). Furthermore, any optimal value is achievable at one of the extreme points of the constraint-set polytope $\mathcal{P} \subset \mathbb{R}^{m+2}$. Next, we claim that the extreme points of the polytope $\mathcal{P}$ are all of the form $v \in \mathcal{P}$ with two (or more) of the ε_is equal to 0. This suffices to prove our main claim, since $\varepsilon_i = 0$ in $v \in \mathcal{P}$ iff the (a,b) induced by v verifies $ax_i + b = y_i$; in other words, the line induced by (a,b) contains the point (x_i, y_i). If a line contains two data points, it is uniquely determined by them:

[1] We ignore the degenerate possibility of vertical lines, which reduces to the 0-dimensional case.

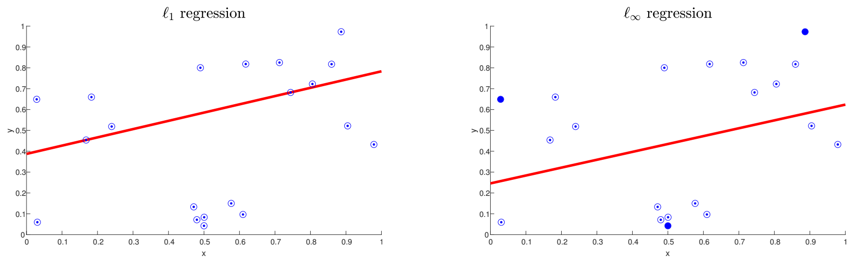

Figure 4.1: *A sample S of $m = 20$ points (x_i, y_i) was drawn iid uniformly from $[0,1]^2$. On this sample, ℓ_1 regression was performed by solving the LP in (4.2), shown on the left, and ℓ_∞ regression was performed by solving the LP in (4.3), on the right. In each case, the regressor provided by the LP solver is indicated by the thick (red) line. Notice that for ℓ_1, the line contains exactly 2 datapoints. For ℓ_∞, the regressor contains no datapoints; rather, the $d + 2 = 3$ "support vectors" are indicated by* ●.

these constitute a compression set of size 2. (See illustration in Figure 4.1.)

Now we prove our claimed property of the extreme points. First, we claim that any extreme point of $\mathcal{P}$ must have least one ε_i equal to 0. Indeed, let (a, b) define a line. Define

$$b^+ := \min\left\{\tilde{b} \in [b, \infty) : \exists i \in [m], ax_i + \tilde{b} = y_i)\right\}$$

and analogously,

$$b^- := \max\left\{\tilde{b} \in (-\infty, b] : \exists i \in [m], ax_i + \tilde{b} = y_i)\right\}.$$

In words, (a, b^+) is the line obtained by increasing b to a maximum value of b^+, where the line (a, b^+) touches a datapoint, and likewise, (a, b^-) is the line obtained by decreasing b to a minimum value of b^-, where the line (a, b^-) touches a datapoint.

Define by $S^+_{a,b} := \{i : |ax_i + b < y_i|\}$ the points above the line defined by (a, b) and $S^-_{a,b} := \{i : |ax_i + b > y_i|\}$ the points below the line defined by (a, b). For a line (a, b) which does not contain a data point we can rewrite the sample loss as

$$
\begin{aligned}
L(a,b) &= \sum_{i\in S^+_{a,b}} (y_i-(ax_i+b)) + \sum_{i\in S^-_{a,b}} ((ax_i+b)-y_i) \\
&= \left(\sum_{i\in S^-_{a,b}} x_i - \sum_{i\in S^+_{a,b}} x_i\right) a + \left(|S^-_{a,b}|-|S^+_{a,b}|\right) b + \left(\sum_{i\in S^+_{a,b}} y_i - \sum_{i\in S^-_{a,b}} y_i\right) \\
&=: \lambda a + \mu b + \nu.
\end{aligned}
$$

Since for fixed a and $b \in [b^-, b^+]$, the quantities $S^-_{a,b}, S^+_{a,b}$ are constant, it follows that the function $L(a,\cdot)$ is affine in b, and hence minimized at $b^\pm \in \{b^-, b^+\}$. Thus, there is no loss of generality in taking $b^\star = b^\pm$, which implies that the optimal solution's line $(a^\star, b^\star)$ contains a data point $(x_{\hat\imath}, y_{\hat\imath})$. If the line $(a^\star, b^\pm)$ contains other data points then we are done, so assume to the contrary that $\varepsilon_{\hat\imath}$ is the only ε_i that vanishes in the corresponding solution $v^\star \in \mathcal{P}$.

Let $\mathcal{P}_{\hat\imath} \subset \mathcal{P}$ consist of all v for which $\varepsilon_{\hat\imath} = 0$, corresponding to all feasible solutions whose line contains the data point $(x_{\hat\imath}, y_{\hat\imath})$. Let us say that two lines $(a_1,b_1),(a_2,b_2)$ are *equivalent* if they induce the same partition on the data points, in the sense of linear separation in the plane. The formal condition is $S^-_{a_1,b_1} = S^-_{a_1,b_1}$, which is equivalent to $S^+_{a_1,b_1} = S^+_{a_1,b_1}$.

Define $\mathcal{P}^\star_{\hat\imath} \subset \mathcal{P}_{\hat\imath}$ to consist of those feasible solutions whose line is equivalent to $(a^\star, b^\pm)$. Denote by $a^+ := \max\{a : (\varepsilon_1,..,\varepsilon_m,a,b) \in \mathcal{P}^\star_{\hat\imath}\}$ and define v^+ to be a feasible solution in $\mathcal{P}^\star_{\hat\imath}$ with slope a^+, and analogously, $a^- := \min\{a : (\varepsilon_1,..,\varepsilon_m,a,b) \in \mathcal{P}^\star_{\hat\imath}\}$ and $v^- \in \mathcal{P}^\star_{\hat\imath}$ with slope a^-. Geometrically this corresponds to rotating the line $(a^\star, b^\star)$ about the point $(x_{\hat\imath}, y_{\hat\imath})$ until it encounters a data point above and below.

Writing, as above, the sample loss in the form $L(a,b)$, we see that $L(\cdot, b^\pm)$ is affine in a over the range $a \in [a^-, a^+]$ and hence is minimized at one of the endpoints. This furnishes another datapoint $(x_{\hat\jmath}, y_{\hat\jmath})$ verifying $\hat a x_{\hat\jmath} + \hat b = y_{\hat\jmath}$ for $L(\hat a, \hat b) = L^\star$, and hence proves compressibility into two points for $d = 1$.

Generalizing to $d > 1$ is quite straightforward. We define

$$
L(\mathbf{a}, b) = \sum_{i\in[m]} |(\langle \mathbf{a}, \mathbf{x}_i\rangle + b) - y_i|
$$

and express it as a linear program analogous to (4.2), where the minimization is over $(\varepsilon_1, \ldots, \varepsilon_m, \mathbf{a}, b) \in \mathbb{R}^{m+d+1}$ and the expression ax_i in the constraints is replaced by $\langle \mathbf{a}, \mathbf{x}_i\rangle$. Given an optimal solution $(\mathbf{a}^\star, b^\star)$, we argue exactly as above that $b^\star$ may be chosen so that the optimal regressor contains some datapoint — say, $(\mathbf{x}_1, y_1)$. Holding $b^\star$ and $\mathbf{a}(j)$, $j \neq 1$ fixed, we argue, as above, that $\mathbf{a}(1)$ may be chosen so that the optimal regressor contains another datapoint (say, $(\mathbf{x}_2, y_2)$). Proceeding in this fashion, we inductively argue that the optimal regressor may be chosen to contain some $d+1$ datapoints, which provides the requisite compression scheme.

□

Theorem 4.4. *There exists an efficiently computable compression scheme for agnostic linear regression in* $\mathbb{R}^d$ *under the* ℓ_∞ *loss of size* $d+2$.

Proof. Given m labeled points in $\mathbb{R}^d \times \mathbb{R}$, $S = (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$ and any $\mathbf{a} \in \mathbb{R}^d$, $b \in \mathbb{R}$ define the empirical risk

$$L(\mathbf{a}, b) \quad := \quad \max\left\{|\langle \mathbf{a}, \mathbf{x}_i \rangle + b - y_i| : i \in [m]\right\}.$$

We cast the risk minimization problem as a linear program:

$$\begin{aligned} \min_{(\varepsilon,\mathbf{a},b)\in\mathbb{R}^{d+2}} \quad &: \quad \varepsilon \\ s.t. \quad \forall i \quad &: \quad \varepsilon - \langle \mathbf{a}, \mathbf{x}_i \rangle - b + y_i \geq 0 \\ & \quad\;\; \varepsilon + \langle \mathbf{a}, \mathbf{x}_i \rangle + b - y_i \geq 0. \end{aligned} \tag{4.3}$$

(As before, the constraint $\varepsilon \geq 0$ is implicit in the other constraints.) Introducing the Lagrange multipliers $\lambda_i, \mu_i \geq 0$, $i \in [m]$, we cast the optimization problem in the form of a Lagrangian:

$$\mathcal{L}(\varepsilon, \mathbf{a}, b, \mu_1 \ldots, \mu_m, \lambda_1 \ldots, \lambda_m) \quad = \quad \varepsilon - \sum_{i=1}^{m} \lambda_i \left(\varepsilon - \langle a, x_i \rangle - b + y_i\right) - \sum_{i=1}^{m} \mu_i \left(\varepsilon + \langle a, x_i \rangle + b - y_i\right).$$

The KKT conditions imply, in particular, that

$$\begin{aligned} \forall i: \quad & \lambda_i(\varepsilon - \langle \mathbf{a}, \mathbf{x}_i \rangle - b + y_i) = 0 \\ & \mu_i(\varepsilon + \langle \mathbf{a}, \mathbf{x}_i \rangle + b - y_i) = 0. \end{aligned}$$

Geometrically, this means that either the constraints corresponding to the ith datapoint are inactive — in which case, omitting the datapoint does not affect the solution — or otherwise, the ith datapoint induces the active constraint

$$\langle \mathbf{a}, \mathbf{x}_i \rangle + b - y_i = \varepsilon. \tag{4.4}$$

On analogy with SVM, let us refer to the datapoints satisfying (4.4) as the *support vectors*; clearly, the remaining sample points may be discarded without affecting the solution. Solutions to (4.3) lie in $\mathbb{R}^{d+2}$ and hence $d + 2$ linearly independent datapoints suffice to uniquely pin down an optimal $(\varepsilon, \mathbf{a}, b)$ via the equations (4.4).

$\square$

Chapter 5

Future Research

It's difficult to make predictions,
especially about the future.

Niels Bohr

5.1 Expanding Warmuth's Conjecture into Real-Valued Classes

Recall the fundamental question posed by Warmuth:

Do every class with finite VC-dimension admits a constant-size compression scheme which size is *linear* in the dimension?

As mentioned above, Moran and Yehudayoff proved the existence constant-size compression scheme which size is *exponential* in the dimension. Meaning the linear possibility is still open. Warmuth linearity conjecture was based on several previews lines of work, constructing compression schemes of *linear* size for specific classes of binary-functions. In particular: classes with $VCdim = 1$, Maximum classes and Duddley classes.

The original conjecture concerns, the basic, specific, case of binary-function classes, on which the original notion of compression scheme and VC-dimension was defined. Once we established the extension of the reduction for real-valued function classes it is natural to propose the following question:

Open Problem: Do every class with finite Fat-shattering-dimension admits a constant-size compression scheme which size is *linear* in the dimension?

Our work proves a partial, qualitative, result, namely: Every class with finite Fat-shattering-dimension admits a constant-size compression scheme which size is *exponential* in the dimension.

In order to base the possibility of linear-sized compression scheme, a natural initial goal might be extending the known results for the binary case. This direction, beside the difficulties which might rise as the real-valued case is more complex then the binary one, is to define a proper extension to the above notion for each family of classes.

5.1.1 Real-Maximum classes compression

During their investigation of the connection between PAC-learning and sample-compression schemes, Floyd and Warmuth recall a definition by Welzl of *maximum class.* Let $\Phi_d(m)$ called the *growth-function* be defined as

$$\Phi_d(m) = \begin{cases} \sum_{i=0}^{d} \binom{m}{i}, & m \geq d \\ 2^m, & m < d \end{cases}.$$

The fundamental combinatorial result for VC classes known as The Sauer's Lemma is the following

Lemma 5.1 (Sauer's Lemma)**.** *Let $d = VC(\mathcal{F})$, Then for any $Y \subseteq \mathcal{X}$ the for restriction of $\mathcal{F}$ to Y, denoted by $\mathcal{F}|_Y$,*

$$\mathcal{F}|_Y \leq \Phi_d(|Y|).$$

Definition 5.1 (Maximum class)**.** A concept class with $VC(\mathcal{F}) = d$ is called *maximum* if for every finite subset Y of the instance space, contains exactly $\Phi_D(|Y|)$ concepts. More formally

$$\mathcal{F}|_Y = \Phi_d(|Y|).$$

Thus a maximum class $\mathcal{F}$ restricted to a finite subset Y, is of maximum size.

Using some of Welzl's results, Floyd and Warmuth provide a sample-compression scheme of size $O(d)$ for maximum classes with $VCdim(\mathcal{F}) = d$.

In order to extend this to the real-valued setting, we first need to define what is the right notion of *maximum* for such classes. One option is to reduce the problem into a binary one. First recall another combinatorial dimension for real-valued function classes - the Pseudo-dimension, first defined by Pollard:

Definition 5.2 (Pseudo-Dimension)**.** Let $\mathcal{F} \subset [0,1]^{\mathcal{X}}$ we say that $\mathcal{F}$ shatters a set $x = x_1, \ldots, x_m \subseteq$ if there exist $r = r_1, \ldots, r_m \in \mathbb{R}^m$ s.t. for all $b \in \{0,1\}^m$ there exist $f_b \in \mathcal{F}$ s.t.

$$\forall i \in [m] : \operatorname{sign}(f_b(x_i) - r_i) = b_i.$$

The pseudo-dimension of $\mathcal{F}$, denoted by $Pdim(\mathcal{F})$, is the cardinally of the largest set of points in $\mathcal{X}$ that can be pseudo-shattered by $\mathcal{F}$

Now for a function-class $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$, define the class of indicators of the epigraphs

$$\mathcal{H}_{\mathcal{F}} := \{(x,y) \to \operatorname{sign}(f(x) - y) \mid f \in \mathcal{F}\}.$$

It is not hard to prove that $VCdim(\mathcal{H}_{\mathcal{F}}) = Pdim(\mathcal{F})$. Using this reduction a possible definition for *real*-maximum class might be a class $\mathcal{F}$ such that $\mathcal{H}_{\mathcal{F}}$ is maximum class of dimension $Pdim(\mathcal{F})$.

Another direction might be a more direct one - substitute the growth-function Φ by equivalent notion relevant for the real-valued case, e.g. *covering numbers*. The main problem with this direction is that although there are many bounds on the covering-numbers no tight results are known for the general case, in the way Φ is used in Floyd and Warmuth work.

Either way, whatever extended definition chosen, the main issue would be to try and prove tighter upper bound for those classes.

5.1.2 Real-Dudley classes compression

Before the definition of sample-copmression schemes, on the wide and then young research of learning theroy Dudley [1984] defined the following definition:

Definition 5.3. For a function class $\mathcal{F}$ which is a vector space over $\mathbb{R}$, and any $h : \mathcal{X} \to \mathbb{R}$ Denote

$$\mathcal{H}_{\mathcal{F},h} := \{(x,f) \mid x \in \mathcal{X}, f \in \mathcal{F}, f(x) + h(x) \geq 0\}.$$

Or in other words if we denote $pos(g) := \{x \in \mathcal{X} \mid g(x) \geq 0\}$ then

$$\mathcal{H}_{\mathcal{F},h} := \{pos(f+h) \mid f \in \mathcal{F}\}.$$

A concept class which can be construed in such way is called *Dudley-Class*

Dudley proved that for such a class $VC(\mathcal{H}_{\mathcal{F},h}) = dim(\mathcal{F})$ when $dim(\mathcal{F})$ is the dimension of $\mathcal{F}$ as a vector space. Dudley classes where proved to be in fact maximum, under minor assumptions, by Floyd.

In a through work, Ben-David and Litman, used this notion in order to prove some *universality* properties for a collection of natural geometric classes as hyperplanes. Using those properties they then go and prove than the dual VC dimension of Dudley classes are bounded by the primal VC dimension of the same class. Leveraging this results and some, Independently important, result they prove that every Dudley class admits a sample-compression scheme which is linear in the VC-dimension.

As in the case of maximum classes, first we will need to understand what is the most suitable definition for the real-valued case. Here the main candidate is just dropping the $pos(\cdot)$ operator, namely

Definition 5.4. For a function class $\mathcal{F}$ which is a vector space over $\mathbb{R}$, and any $h : \mathcal{X} \to \mathbb{R}$ Denote

$$\mathcal{H}_{\mathcal{F},h} := \{f+h \mid f \in \mathcal{F}\}.$$

A concept class which can be construed in such way is called *real-Dudley-Class*

After selecting the proper notion, there are two possible directions

1. Extend Floyd's result, under modified assumptions and regarding $Pdim$ instead of VC.

2. Extend Ben-David and Litman's embeddings system in to construct their scheme or at least recover the bound on the dual-Pdim, since using such bound and pluging it into our algorithm guarantees, results in a *uniformly ε-approximate* compression schme with linear-dependence on Pdim.

5.2 Agnostic Compressability

The case of agnostic-compression schemes is somewhat different in first sight, combined with the past negative results it is not surprising that the results for this regime our very sparse. Yet the above positive results give rise to couple of basic and more wide questions which can be of high importance to the better known areas of learning theory as the classic notions of compression schemes.

5.2.1 Open Problem: Compressing to Pseudo-dimension Number of Points

The above positive results for ℓ_1 loss may also lead us to wonder how general of a result might be possible. In particular, noting that the pseudo-dimension [Pollard, 1984, 1990, Anthony and Bartlett, 1999] of linear functions in $\mathbb{R}^d$ is precisely $d+1$ [Anthony and Bartlett, 1999], there is an intriguing possibility for the following generalization.

Open Problem: Under the ℓ_1 loss, does every class $\mathcal{F}$ of real-valued functions admit an agnostic compression scheme of size $Pdim(\mathcal{F})$?

It is also interesting, and perhaps more approachable as an initial aim, to ask whether there is an agnostic compression scheme of size at most *proportional to* $Pdim(\mathcal{F})$. Even falling short of this, one can ask the more-basic question of whether classes with $Pdim(\mathcal{F}) < \infty$ always have *bounded* agnostic compression schemes (i.e., independent of sample size m), and more specifically whether the bound is expressible purely as a function of $Pdim(\mathcal{F})$ (Moran and Yehudayoff, 2016 have shown this is always possible in the realizable classification setting).

These questions are directly related to (and inspired by) the well-known long-standing conjecture of Warmuth [2003], which asks whether, for realizable-case binary classification, there is always a compression scheme of size at most linear in the VC dimension of the concept class. Indeed, it is clear that a positive solution of our open problem above would imply a positive solution to the original sample compression conjecture, since in the realizable case with a function class $\mathcal{F}$ of $\{0,1\}$-valued functions, the minimal empirical ℓ_1 loss on the data is zero, and any function obtaining zero empirical ℓ_1 loss on a data set labeled with $\{0,1\}$ values must be $\{0,1\}$-valued on that data set, and thus can be thought of as a sample-consistent classifier.[1] Noting that, for $\mathcal{F}$ containing $\{0,1\}$-valued functions, $Pdim(\mathcal{F})$ is equal the VC dimension, the implication is clear.

The converse of this direct relation is not necessarily true. Specifically, for a

[1] To make such a function actually binary-valued everywhere, it suffices to threshold at 1/2.

set $\mathcal{F}$ of real-valued functions, consider the set $\mathcal{H}$ of subgraph sets: $h_f(x,y) = \mathbb{I}[y \le f(x)]$, $f \in \mathcal{F}$. In particular, note that the VC dimension of $\mathcal{H}$ is precisely $Pdim(\mathcal{F})$. It is *not* true that any realizable classification compression scheme for $\mathcal{H}$ is also an agnostic compression scheme for $\mathcal{F}$ under ℓ_1 loss. Nevertheless, this reduction-to-classification approach seems intuitively appealing, and it might possibly be the case that there is some way to *modify* certain types of compression schemes for $\mathcal{H}$ to convert them into agnostic compression schemes for $\mathcal{F}$. Following up on this line of investigation seems the natural next step toward resolving the above general open question.

5.2.2 Characterization of Agnostic Compressibility

Consider the following proof sketch for Theorem 4.3:

A sample consists of $(x_i, y_i)_{i\in[m]}$, where $x_i, y_i \in \mathbb{R}$ (for simplicity we treat the $d = 1$ case), and $\mathcal{H} = \{\mathbb{R} \ni x \mapsto ax + b : a, b \in \mathbb{R}\}$.

Let $(a^\star, b^\star)$ be a (possibly non-unique) minimizer of

$$L(a,b) := \sum_{i\in[m]} |(ax_i + b) - y_i|, \tag{5.1}$$

achieving the value $L^\star$. We claim that we can always find two indices $\hat{\imath}, \hat{\jmath} \in [m]$ such that the line determined by $(x_{\hat{\imath}}, y_{\hat{\imath}})$ and $(x_{\hat{\jmath}}, y_{\hat{\jmath}})$ also achieves the optimal empirical risk $L^\star$. More precisely, the line $(\hat{a}, \hat{b})$ induced by $((x_{\hat{\imath}}, y_{\hat{\imath}}), (x_{\hat{\jmath}}, y_{\hat{\jmath}}))$ via[2] $\hat{a} = (y_{\hat{\jmath}} - y_{\hat{\imath}})/(x_{\hat{\jmath}} - x_{\hat{\imath}})$ and $\hat{b} = y_{\hat{\imath}} - \hat{a}x_{\hat{\imath}}$, verifies $L(\hat{a}, \hat{b}) = L^\star$.

To prove this claim, we begin by recasting (5.1) as a linear program:

$$\begin{aligned} \min_{(\varepsilon_1,\dots,\varepsilon_m,a,b)\in\mathbb{R}^{m+2}} \quad & \sum_{i=1}^{m} \varepsilon_i \quad \text{s.t.} \\ \forall i \in [m] \quad & \varepsilon_i \ge 0 \\ \forall i \in [m] \quad & ax_i + b - y \le \varepsilon_i \\ \forall i \in [m] \quad & -ax_i - b + y \le \varepsilon_i. \end{aligned} \tag{5.2}$$

We observe that the linear program in (5.2) is feasible with a finite solution (and actually, the constraints $\varepsilon_i \ge 0$ are redundant). Furthermore, any optimal value is achievable at one of the extreme points of the constraint-set polytope $\mathcal{P} \subset \mathbb{R}^{m+2}$. Next, we claim that the extreme points of the polytope $\mathcal{P}$ are

[2] We ignore the degenerate possibility of vertical lines, which reduces to the 0-dimensional case.

all of the form $v \in \mathcal{P}$ with two (or more) of the ε_is equal to 0. This suffices to prove our main claim, since $\varepsilon_i = 0$ in $v \in \mathcal{P}$ iff the (a, b) induced by v verifies $ax_i + b = y_i$; in other words, the line induced by (a, b) contains the point (x_i, y_i). If a line contains two data points, it is uniquely determined by them: these constitute a compression set of size 2.

With this formulation in mind, it might be possible to extend the above result into every loss-function whose transformation into linear-program yields constant number of constrains which determinate each extreme point. Furthermore we conjecture that this isn't only a sufficient condition but also a necessary one. For this reason, we also conjure that r-piecewise linear loss functions are the only ones that admit bounded agnostic compression schemes.

5.2.3 From Agnostic-Compression to Approximate-Agnostic-Compression

Another direction that can be taken using the linear-programming paradigm, is - using agnostic-compression schemes in order to construct approximate-agnostic-compression schemes.

Following David et al. [2016], we say that (κ, ρ) is a k-size ε-approximate-agnostic sample compression scheme for $\mathcal{F}$ if κ is a k-selection and for all $S = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m))$, $f_S := \rho(\kappa(S))$ achieves $\mathcal{F}$-competitive empirical loss:

$$L_p(f_S, S) \quad \leq \quad \inf_{f \in \mathcal{F}} L_p(f, S) + \varepsilon.$$

According to our conjecture from Subsection 5.2.2, loss-function which are not piecewise-linear can't admit agnostic-compression schemes. Yet, as proven by David et al. [2016, Theorem 4.3], if we relax the requirements and replace the agnostic-compression with approximate-agnostic-compression, we get that "Learning implies approximate compressing". For this reason it is interesting to try and find approximate-compression schemes for different loss-function.

One strategy of constructing such schemes can be through (non-approximate-)agnostic-compression schemes. The idea is to approximate the loss function with a piecewise-linear function, and then apply agnostic-compression scheme regarding that loss-function. See for example Figure 5.1

The resulting approximation will, of course depend on the number linear-pieces. The compression size might depend on the same parameter and in addition on the specific compression scheme used for the linear-approximated

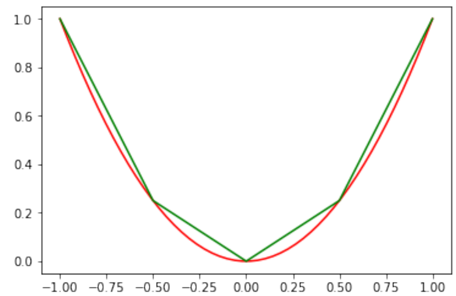

Figure 5.1: *Piecewise approximation of the ℓ_2 loss function, using 5 linear pieces.*

loss.

This approach may be used to a wide range of loss-functions and it is interesting how will it compare to the state-of-the-art approximate-agnostic-compression schemes. Also, using David et al. [2016, Theorem 4.2] idea or a more specific results, one might derive generalization bounds using approximate-agnostic-compression-schemes, and it is hence interesting to try and see what quality of generalization bounds can this approach yield.

Bibliography

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. URL `citeseer.ist.psu.edu/alon97scalesensitive.html`.

Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, Cambridge, 1999. ISBN 0-521-57353-X. doi: 10.1017/CBO9780511624216. URL `http://dx.doi.org/10.1017/CBO9780511624216`.

Martin Anthony, Peter L. Bartlett, Yuval Ishai, and John Shawe-Taylor. Valid generalisation from approximate interpolation. *Combinatorics, Probability & Computing*, 5:191–214, 1996. doi: 10.1017/S096354830000198X. URL `https://doi.org/10.1017/S096354830000198X`.

Hassan Ashtiani, Shai Ben-David, Nick Harvey, Christopher Liaw, Abbas Mehrabian, and Yaniv Plan. Settling the sample complexity for learning mixtures of gaussians. In *NIPS*, 2018.

Patrice Assouad. Densité et dimension. *Ann. Inst. Fourier (Grenoble)*, 33(3):233–282, 1983. ISSN 0373-0956. URL `http://www.numdam.org/item?id=AIF_1983__33_3_233_0`.

Ran Avnimelech and Nathan Intrator. Boosting regression estimators. *Neural computation*, 11(2):499–520, 1999.

Shai Ben-David and Ami Litman. Combinatorial variability of vapnik-chervonenkis classes with applications to sample compression schemes. *Discrete Applied Mathematics*, 86(1):3–25, 1998.

Alberto Bertoni, Paola Campadelli, and M Parodi. A boosting algorithm for regression. In *International Conference on Artificial Neural Networks*, pages 343–348. Springer, 1997.

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. ISSN 0004-5411.

Artem Chernikov and Pierre Simon. Externally definable sets and dependent pairs. *Israel J. Math.*, 194(1):409–425, 2013. ISSN 0021-2172. URL `https://doi.org/10.1007/s11856-012-0061-9`.

Rachel Cummings, Katrina Ligett, Kobbi Nissim, Aaron Roth, and Zhiwei Steven Wu. Adaptive learning with robust generalization guarantees. In *Conference on Learning Theory*, pages 772–814, 2016.

Ofir David, Shay Moran, and Amir Yehudayoff. Supervised learning through the lens of compression. In *Advances in Neural Information Processing Systems*, pages 2784–2792, 2016.

Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 1996. ISBN 0-387-94618-7.

Harris Drucker. Improving regressors using boosting techniques. In *Proceedings of the Fourteenth International Conference on Machine Learning*, ICML '97, pages 107–115, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc. ISBN 1-55860-486-3. URL `http://dl.acm.org/citation.cfm?id=645526.657132`.

Richard M. Dudley. A course on empirical processes. In *École d'été de probabilités de Saint-Flour, XII—1982*, volume 1097 of *Lecture Notes in Math.*, pages 1–142. Springer, Berlin, 1984.

Nigel Duffy and David Helmbold. Boosting methods for regression. *Machine Learning*, 47:153–200, 2002. ISSN 0885-6125.

Mary Flahive and Bella Bose. Balancing cyclic r-ary gray codes. *the electronic journal of combinatorics*, 14(1):R31, 2007.

Sally Floyd. Space-bounded learning and the vapnik-chervonenkis dimension. In *Proceedings of the second annual workshop on Computational learning theory*, pages 349–364. Morgan Kaufmann Publishers Inc., 1989.

Sally Floyd and Manfred K. Warmuth. Sample compression, learnability, and the Vapnik-Chervonenkis dimension. *Machine Learning*, 21(3):269–304, 1995.

Yoav Freund and Robert E Schapire. Game theory, on-line prediction and boosting. In *Proceedings of the ninth annual conference on Computational learning theory*, pages 325–332. ACM, 1996.

Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55(1):119–139, 1997. ISSN 0022-0000. doi: http://dx.doi.org/10.1006/jcss.1997.1504.

Yoav Freund and Robert E. Schapire. Adaptive game playing using multiplicative weights. *Games Econom. Behav.*, 29(1-2):79–103, 1999. ISSN 0899-8256. URL `https://doi.org/10.1006/game.1999.0738`. Learning in games: a symposium in honor of David Blackwell.

Jerome H. Friedman. Greedy function approximation: a gradient boosting machine. *Ann. Statist.*, 29(5):1189–1232, 2001. ISSN 0090-5364. URL `https://doi.org/10.1214/aos/1013203451`.

Lee-Ad Gottlieb, Aryeh Kontorovich, and Pinhas Nisnevitch. Near-optimal sample compression for nearest neighbors. In *Advances in Neural Information Processing Systems*, pages 370–378, 2014.

Lee-Ad Gottlieb, Aryeh Kontorovich, and Robert Krauthgamer. Efficient regression in metric spaces via approximate lipschitz extension. *IEEE Trans. Information Theory*, 63(8):4838–4849, 2017a. doi: 10.1109/TIT.2017.2713820. URL `https://doi.org/10.1109/TIT.2017.2713820`.

Lee-Ad Gottlieb, Aryeh Kontorovich, and Robert Krauthgamer. Efficient regression in metric spaces via approximate lipschitz extension. *IEEE Transactions on Information Theory*, 63(8):4838–4849, Aug 2017b. ISSN 0018-9448. doi: 10.1109/TIT.2017.2713820.

Lee-Ad Gottlieb, Aryeh Kontorovich, and Pinhas Nisnevitch. Nearly optimal classification for semimetrics. *Journal of Machine Learning Research*, 18:37:1–37:22, 2017c. URL `http://jmlr.org/papers/v18/papers/v18/16-217.html`.

Thore Graepel, Ralf Herbrich, and John Shawe-Taylor. PAC-bayesian compression bounds on the prediction error of learning algorithms for classification. *Machine Learning*, 59(1-2):55–76, 2005.

David Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Inf. Comput.*, 100(1):78–150, 1992. doi: 10.1016/0890-5401(92)90010-D. URL `http://dx.doi.org/10.1016/0890-5401(92)90010-D`.

David Haussler, Michael Kearns, Nick Littlestone, and Manfred K Warmuth. Equivalence of models for polynomial learnability. *Information and Computation*, 95(2):129–161, 1991.

David Helmbold, Robert Sloan, and Manfred K Warmuth. Learning integer lattices. *SIAM Journal on Computing*, 21(2):240–266, 1992.

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

Daniel M. Kane, Roi Livni, Shay Moran, and Amir Yehudayoff. On communication complexity of classification problems. *CoRR*, abs/1711.05893, 2017. URL `http://arxiv.org/abs/1711.05893`.

Grigoris Karakoulas and John Shawe-Taylor. Towards a strategy for boosting regressors. In Alexander J. Smola, Peter L. Bartlett, and Schölkopf, editors, *Advances in Large Margin Classifiers*, Advances in Neural Information Processing Systems, pages 43–54. MIT Press, Cambridge, MA, USA, 2000. ISBN 0-262-19448-1.

M. Kearns. Thoughts on hypothesis boosting. Unpublished, December 1988.

Balázs Kégl. Robust regression by boosting the median. In *Learning Theory and Kernel Machines*, pages 258–272. Springer, 2003.

Dima Kuzmin and Manfred K. Warmuth. Unlabeled compression schemes for maximum classes. *Journal of Machine Learning Research*, 8:2047–2081, 2007. URL `http://dl.acm.org/citation.cfm?id=1314566`.

Nick Littlestone and Manfred K. Warmuth. Relating data compression and learnability. Technical report, Department of Computer and Information Sciences, Santa Cruz, CA, Ju, 1986.

Roi Livni and Pierre Simon. Honest compressions and their application to compression schemes. In *Conference on Learning Theory*, pages 77–92, 2013.

Philip M. Long. Efficient algorithms for learning functions with bounded variation. *Inf. Comput.*, 188(1):99–115, 2004. doi: 10.1016/S0890-5401(03)00164-0. URL `https://doi.org/10.1016/S0890-5401(03)00164-0`.

Shie Mannor and Ron Meir. On the existence of linear weak learners and applications to boosting. *Machine Learning*, 48(1-3):219–251, 2002. doi: 10.1023/A:1013959922467. URL `https://doi.org/10.1023/A:1013959922467`.

Llew Mason, Jonathan Baxter, Peter Bartlett, and Marcus Frean. Boosting algorithms as gradient descent. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, NIPS'99, pages 512–518, Cambridge, MA, USA, 1999. MIT Press. URL `http://dl.acm.org/citation.cfm?id=3009657.3009730`.

S. Mendelson and R. Vershynin. Entropy and the combinatorial dimension. *Invent. Math.*, 152(1):37–55, 2003. ISSN 0020-9910. doi: 10.1007/s00222-002-0266-3. URL `http://dx.doi.org/10.1007/s00222-002-0266-3`.

Shahar Mendelson. Learning without concentration. *J. ACM*, 62(3):21:1–21:25, 2015. doi: 10.1145/2699439. URL `http://doi.acm.org/10.1145/2699439`.

João Mendes-Moreira, Carlos Soares, Alípio Mário Jorge, and Jorge Freire De Sousa. Ensemble approaches for regression: A survey. *ACM Comput. Surv.*, 45(1):10:1–10:40, December 2012. ISSN 0360-0300. doi: 10.1145/2379776.2379786. URL `http://doi.acm.org/10.1145/2379776.2379786`.

Shay Moran and Amir Yehudayoff. Sample compression schemes for VC classes. *J. ACM*, 63(3):21:1–21:10, 2016. doi: 10.1145/2890490. URL `http://doi.acm.org/10.1145/2890490`.

Shay Moran, Amir Shpilka, Avi Wigderson, and Amir Yehudayoff. Teaching and compressing for low vc-dimension. In *A Journey Through Discrete Mathematics*, pages 633–656. Springer, 2017.

Richard Nock and Frank Nielsen. A real generalization of discrete adaboost. *Artificial Intelligence*, 171(1):25 – 41, 2007. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2006.10.014. URL `http://www.sciencedirect.com/science/article/pii/S0004370206001111`.

Leonard Pitt and Leslie G Valiant. Computational limitations on learning from examples. *Journal of the ACM (JACM)*, 35(4):965–984, 1988.

David Pollard. *Convergence of Stochastic Processes*. Springer-Verlag, 1984.

David Pollard. *Empirical processes: theory and applications*. NSF-CBMS Regional Conference Series in Probability and Statistics, 2. Institute of Mathematical Statistics, Hayward, CA, 1990. ISBN 0-940600-16-1.

Benjamin I. P. Rubinstein and J. Hyam Rubinstein. A geometric approach to sample compression. *Journal of Machine Learning Research*, 13:1221–1261, 2012. URL `http://dl.acm.org/citation.cfm?id=2343686`.

Benjamin I. P. Rubinstein, Peter L. Bartlett, and J. Hyam Rubinstein. Shifting: One-inclusion mistake bounds and sample compression. *J. Comput. Syst. Sci.*, 75(1):37–59, 2009. doi: 10.1016/j.jcss.2008.07.005. URL `https://doi.org/10.1016/j.jcss.2008.07.005`.

M. Rudelson and R. Vershynin. Combinatorics of random processes and sections of convex bodies. *Ann. of Math. (2)*, 164(2):603–648, 2006. ISSN 0003-486X. URL `https://doi.org/10.4007/annals.2006.164.603`.

Robert E. Schapire. The strength of weak learnability. *Machine Learning*, 5: 197–227, 1990.

Robert E. Schapire and Yoav Freund. *Boosting*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2012. ISBN 978-0-262-01718-3. Foundations and algorithms.

Robert E. Schapire, Yoav Freund, Peter Bartlett, and Wee Sun Lee. Boosting the margin: a new explanation for the effectiveness of voting methods. *Ann. Statist.*, 26(5):1651–1686, 1998. ISSN 0090-5364. doi: 10.1214/aos/1024691352. URL `http://dx.doi.org/10.1214/aos/1024691352`.

Claude E. Shannon. A mathematical theory of communication. *Mobile Computing and Communications Review*, 5:3–55, 2001.

Hans Ulrich Simon. Bounds on the number of examples needed for learning functions. *SIAM J. Comput.*, 26(3):751–763, 1997. doi: 10.1137/S0097539793259185. URL `https://doi.org/10.1137/S0097539793259185`.

Leslie G. Valiant. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, 1984.

V. N. Vapnik and A. Ja. Červonenkis. The uniform convergence of frequencies of the appearance of events to their probabilities. *Teor. Verojatnost. i Primenen.*, 16:264–279, 1971. ISSN 0040-361x.

V. N. Vapnik and A. Ya. Chervonenkis. *Teoriya raspoznavaniya obrazov. Statisticheskie problemy obucheniya.* Izdat. "Nauka", Moscow, 1974.

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

Manfred K. Warmuth. Compressing to VC dimension many points. In *Proceedings of the 16th Conference on Learning Theory*, 2003.