

UC Berkeley

UC Berkeley Electronic Theses and Dissertations

Title

Wisdom of the Crowd: The Statistical Benefits of Aggregation

Permalink

<https://escholarship.org/uc/item/5qd7d11j>

ISBN

9798189271830

Author

Aden-Ali, Ishaq

Publication Date

2026-05-01

Peer reviewed|Thesis/dissertation

Wisdom of the Crowd: The Statistical Benefits of Aggregation

by

Ishaq Aden-Ali

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Peter Bartlett, Co-chair

Professor Jelani Nelson, Co-chair

Associate Professor Nika Haghtalab

Assistant Professor Nikita Zhivotovskiy

Spring 2026

Wisdom of the Crowd: The Statistical Benefits of Aggregation

Copyright 2026
by
Ishaq Aden-Ali

Abstract

Wisdom of the Crowd: The Statistical Benefits of Aggregation

by

Ishaq Aden-Ali

Doctor of Philosophy in Computer Science

University of California, Berkeley

Professor Peter Bartlett, Co-chair

Professor Jelani Nelson, Co-chair

Binary classification and mean estimation are among the most fundamental problems in learning theory and statistics, yet the most intuitive estimators can provide suboptimal guarantees. This raises a natural question: can we obtain optimal statistical guarantees without giving up the simplicity of these basic methods?

This dissertation addresses this question through the lens of aggregation. We will show that, in several settings, simply combining natural estimators in the right way is enough to recover optimal guarantees. We first consider classification in the PAC framework, where majority votes of a few empirical risk minimizers can improve on what any single empirical risk minimizer achieves. We then study more general learning problems, where aggregating leave-one-out predictors turns average guarantees into sharp high-probability guarantees. Finally, we study heavy-tailed mean estimation under ℓ_p norms. We show that, assuming only finite covariance, median-based aggregation of sample means achieves the sharp minimax rate. This result also reveals a phase transition in the optimal rate, governed by the interaction between the sample size and the geometry of the norm.

Together, these results show that simple estimators need not be replaced. With the right aggregation principle, they can become statistically optimal.

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

In the name of God, the Most Gracious, the Most Merciful.

Contents

Contents	ii
List of Figures	iv
1 Introduction	1
1.1 Binary Classification	1
1.2 Beyond Uniform Convergence	3
1.3 Heavy Tailed Mean Estimation	4
2 Majority-of-Three: The Simplest Optimal Learner?	7
2.1 Majority-of-Three is optimal in-expectation	12
2.2 A lower bound for certain majorities	14
2.3 High probability upper bound	18
2.4 Omitted proofs from Sections 2.1 and 2.3	24
2.5 Omitted proofs from Section 2.2	29
3 Optimal PAC Bounds without Uniform Convergence	31
3.1 Main results	34
3.2 Related work	44
3.3 Proof of main result	47
3.4 Proof of the lower bound	51
3.5 Proofs from Section 3.1: Classification	59
3.6 Proofs for Section 3.1: Bounded regression	68
3.7 Proofs for Section 3.3: Main result	73
3.8 Function Class for which our Adversarial Construction has Optimal Out-degree	75
4 Heavy-Tailed Mean Estimation under ℓ_p Norms	82
4.1 Preliminaries	87
4.2 In-expectation Results	87
4.3 Tight deviation bounds for the ℓ_∞ norm	94
4.4 Lower bounds	98
4.5 Omitted Proofs	103

Bibliography**115**

List of Figures

- 2.1 An illustration of the partitioning of the interval $(0, 1]$ for a training sample consisting of $m = 18$ points with $d = 2$. The interval $(0, 1]$ is partitioned into 4 intervals $I_1, \dots, I_4$. Each interval I_i is further partitioned into the 4 subintervals $I_{(i,1)}, \dots, I_{(i,4)}$. The red points correspond to the first half of the sample $(X_1, \dots, X_9)$ and the blue points correspond to the second half of the sample $(X_{10}, \dots, X_{18})$. The yellow highlighted regions are the first d intervals I_2 and I_4 that contain no points from $(X_1, \dots, X_9)$. The green highlighted regions are the first d subintervals of I_2 and I_4 that contain no points from $(X_{10}, \dots, X_{18})$. The green intervals are added to the union of intervals used by $\hat{f}_S$ as their indices correspond to the set $L_1(S)$ 16

Acknowledgments

First and foremost, I would like to thank Hassan Ashtiani, who advised me on my master's thesis before I came to Berkeley. Hassan taught me so much about research, academia, and life. Thank you, Hassan, for taking a chance on me as a clueless undergraduate student.

I would like to thank my PhD advisors, Peter Bartlett and Jelani Nelson. Peter and Jelani have been so supportive over the past five years and have consistently encouraged me to explore many different topics. Thank you, Peter and Jelani, for your guidance and encouragement, and for creating an environment that has deeply nurtured my intellectual curiosity.

Nikita Zhivotovskiy has been my closest collaborator for the past four years, and I often jokingly refer to him as my third PhD advisor. Much of my academic taste and style has been inherited from Nikita. We've spent countless hours over the years talking about math, academia, child-rearing, and AI. I have excitedly sent Nikita an uncountable number of incorrect proofs, many of which contained embarrassingly obvious bugs. He has always been patient and willing to try to fix the arguments. Thank you, Nikita, for fostering such a joyful collaboration.

Next, I would like to thank my frequent collaborators Yeshwanth Cherapanmjeri, Abhishek Shetty, Mikael Møller Høgsgaard, and Kasper Green Larsen for all that they have taught me. Yeshwanth has been like a zen older academic sibling to me and still manages to surprise me with his sharp insights and wit. After Nikita, Abhishek has probably been the person I've spoken to the most about research. I am still in awe of his uncanny ability to call me at the most inconvenient times. Mikael has put up with my often unreasonable rewriting of our papers over the years, and I thank him for his patience. Kasper has been a great collaborator and has been kind enough to invite me to Denmark on two separate occasions to work with him and Mikael. These trips have been a very memorable part of my PhD and led to about two-thirds of the content in this dissertation. I am grateful to Kasper for spending so much of his time (and grant money!) on me. Thank you all.

I would also like to thank the numerous other collaborators I've had over the years for teaching me so much of what I know. In rough chronological order, these are: Chris Liaw, Gautam Kamath, Argyris Mouzakis, Huacheng Yu, Yanjun Han, Sidhanth Mohanty, Jan van den Brand, Xin Lyu, Tijana Zrnic, Kewen Wu, Nived Rajaraman, Omar Alrabiah, Omar Montasser, Adam Sealfon, Thomas Steinke, Jon Ullman, Adam Smith, Noah Golowich, Allen Liu, Nika Haghtalab, and Ankur Moitra. Not all of these collaborations led to papers, but all of them taught me a lot about doing research. Thank you all.

A big thank you to Ansh, Sidhanth, Nathan, Nived, Omar, Chris, David, Vamsi, Eric, Simeon, and Seri for so many memorable and hilarious conversations over the years. Many of our hangouts are among my fondest memories of PhD life. I would also like to thank my officemates Luke, Bhaskar, Thiago, Louis, Angelos, Grace, Tarun, and Yinuo for many entertaining (and often unproductive) conversations over the years. I would also like to thank the theory group more generally for making Berkeley such a warm and welcoming place.

A big shoutout to Moiz, Abdullah, and Faizan for all the phone calls that have kept me connected to my Canadian roots. Special thanks to Moiz for all the spontaneous trips he has made out to California: to hike in Yosemite, watch the Golden State Warriors, or to do nothing but fly in with a box of diapers for my son.

I would like to thank my siblings, parents, and in-laws for all their love and support over the years. It is the understatement of the decade to say that I would not have been able to complete my PhD without their encouragement and support. Thank you from the bottom of my heart.

Finally, I would like to thank my beloved wife, Suaad, and our two beautiful boys, Yaqub and Yasin, for all their love. The PhD journey has been far from easy, for so many reasons. Still, knowing that I could come home every day to a home full of love, laughter, and a bit of chaos has been one of my greatest sources of motivation. It is the greatest honour of my life to call them my family. Suaad is the most amazing person I have ever had the pleasure of knowing. Thank you, Suaad, for all the sacrifices you have made over the past nine years while I spent so much time “in school.” You have earned this PhD as much as I did.

Chapter 1

Introduction

Many of the most basic estimation problems in learning theory and statistics can be solved with simple and intuitive estimators. For the problem of binary classification, the canonical procedure is empirical risk minimization: output any classifier in the class that fits the sample. For the problem of mean estimation, the canonical procedure is to output the empirical mean of the observed data. While these estimators are simple, natural, and quite easy to analyze, their strongest provable guarantees are often sub-optimal in many settings.

This dissertation studies a common mechanism for improving these estimators without sacrificing their simplicity: aggregation. In each setting considered, optimal statistical guarantees are obtained not by replacing the natural estimator with a different object, but by combining the simple estimators in the right way.

1.1 Binary Classification

In [Chapter 2](#) we study the problem of binary classification. Specifically, we study classification in the Probably Approximately Correct (PAC) setup [\[81, 84\]](#). Let $\mathcal{X}$ be an instance space and let $\mathcal{H} \subseteq \{0, 1\}^{\mathcal{X}}$ be a hypothesis class consisting of binary classifiers. The learner observes an i.i.d. sample $S = (X_1, \dots, X_n)$ from an unknown distribution P over $\mathcal{X}$, together with labels generated by an unknown target function $f^* \in \mathcal{H}$. The learner's goal is to output a predictor $\hat{f}_S : \mathcal{X} \rightarrow \{0, 1\}$ with small population error

$$\text{err}_P(\hat{f}_S) = \mathbf{Pr}_{X \sim P} [\hat{f}_S(X) \neq f^*(X)],$$

with probability at least $1 - \delta$ over the randomness of the training sample S .

The most natural algorithm for this problem is *Empirical Risk Minimization (ERM)*. Given the labelled sample $(S, f^*(S))$, an ERM outputs *any* $\hat{f}_S \in \mathcal{H}$ satisfying $\hat{f}_S(X_i) = f^*(X_i)$ for every sample point. A fundamental result of Vapnik and Chervonenkis is *uniform convergence*, which tells us that the output of ERM, denoted $\hat{f}_{\text{ERM}}$, satisfies, for any $\delta \in$

$(0, 1)$, the bound

$$\text{err}_P(\hat{f}_{\text{ERM}}) = O\left(\frac{d}{n} \log\left(\frac{n}{d}\right) + \frac{1}{n} \log\left(\frac{1}{\delta}\right)\right),$$

with probability at least $1 - \delta$ over the randomness of the sample. Here, the parameter d is the *Vapnik-Chervonenkis (VC)* dimension, a combinatorial parameter that measures the complexity of the class $\mathcal{H}$. Moreover, it turns out that this bound is unimprovable for ERM in general,

Hanneke [36], building on the work of Simon [77], proposed an algorithm that takes the majority vote of many ERMs trained on different subsamples of S and showed that this estimator that we denote $\hat{f}_{\text{OPT}}$ achieves the bound

$$\text{err}_P(\hat{f}_{\text{OPT}}) = O\left(\frac{d}{n} + \frac{1}{n} \log\left(\frac{1}{\delta}\right)\right), \quad (1.1)$$

removing the log factor. While ERM is arguably the most natural algorithm, it does not achieve the best error guarantee. Moreover, lower bounds for proper learning show that this is not merely a defect of one particular analysis of ERM; in general, any learner that is forced to output a member of the class $\mathcal{H}$ can be suboptimal [22].

It is thus a natural question to ask how little “improperness” is needed to achieve the optimal rate. Continuing the line of work that culminated in optimal improper PAC learners based on aggregating ERMs [77, 36, 52], we consider what is perhaps the simplest improper algorithm, *Majority-of-Three (ERMs)*: partition the sample S into three equal-sized and disjoint pieces S_1, S_2, S_3 , run the same ERM algorithm on each $(S_i, f^*(S_i))$, and combine them via a majority vote to produce the estimator $\text{Maj}(f_{S_1}, f_{S_2}, f_{S_3})$. Our main result shows that this simple aggregation achieves the following in-expectation guarantee:

$$\mathbb{E}_{S_1, S_2, S_3} \left[\text{err}_P(\text{Maj}(\hat{f}_{S_1}, \hat{f}_{S_2}, \hat{f}_{S_3})) \right] = O\left(\frac{d}{n}\right).$$

Moreover, this rate is optimal and it provably cannot be achieved by a single ERM in general.

We also show that Majority-of-Three satisfies the high-probability bound

$$\text{err}_P(\text{Maj}(\hat{f}_{S_1}, \hat{f}_{S_2}, \hat{f}_{S_3})) = O\left(\frac{d}{n} \log \log \left(\min \left\{ \frac{n}{d}, \frac{1}{\delta} \right\} \right) + \frac{1}{n} \log \left(\frac{1}{\delta} \right) \right).$$

This rate is optimal in the low confidence regime (δ constant) and in the very high confidence regime ($\delta \leq d^{-d}$), which is not the case for ERM in general.¹ We conjecture that an improved analysis will show that Majority-of-Three achieves the optimal rate for the entire range of $\delta \in (0, 1)$.

¹In particular, in the very high confidence regime $\delta \leq d^{-d}$, our bound is optimal for any sample size n .

1.2 Beyond Uniform Convergence

In [Chapter 3](#) we move beyond binary classification and study more general realizable learning problems. Let $\mathcal{X}$ be an instance space, $\mathcal{Y}$ be a label space, and let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a hypothesis class. The learner observes an i.i.d. labelled sample $S = ((X_1, Y_1), \dots, (X_n, Y_n))$ from an unknown distribution P over $\mathcal{X} \times \mathcal{Y}$, with the promise that the labels match an unknown target function $f^* \in \mathcal{H}$. The goal is to output a predictor $\hat{f}_S : \mathcal{X} \rightarrow \mathcal{Y}$ with small population risk

$$R_P(\hat{f}_S) = \mathbb{E}_{(X,Y) \sim P} [\ell(\hat{f}_S(X), Y)],$$

under a known loss function $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow [0, 1]$, with probability at least $1 - \delta$ over the randomness of S . This model encompasses general learning settings such as multiclass classification, bounded regression and more.

In the special case of binary classification, uniform convergence gives a natural way to analyze ERM, however in more general settings uniform convergence can be severely sub-optimal or may even fail to hold! A natural alternative to ERM in these settings is the *one-inclusion graph* algorithm and its extensions [\[43, 18, 74\]](#). These algorithms are built with the natural goal of minimizing the *leave-one-out (LOO)* error. Given a realizable sample $S = ((x_1, y_1), \dots, (x_n, y_n))$, write S^{-i} for the sample obtained by removing the i th point. The goal is to construct a predictor $\hat{f}_{\text{LOO}}$ whose leave-one-out error is small:

$$\sum_{i=1}^n \ell(\hat{f}_{\text{LOO}}(x_i; S^{-i}), y_i) \leq M_n,$$

where $\{M_n\}$ is a sequence that depend on the sample size n . Given such a predictor, a simple exchangeability argument immediately implies the in-expectation risk bound

$$\mathbb{E}_S [R_P(\hat{f}_{\text{LOO}})] \leq \frac{M_{n+1}}{n+1}.$$

The one-inclusion graph algorithm and its extension all satisfy simple bounds on M_n based on combinatorial parameters of certain (hyper)graphs constructed from the data. In the special case of binary classification, it is known that $M_n \leq d$ for all n where d is the VC dimension [\[41\]](#). Thus, in the binary setting, the one-inclusion graph algorithm has an optimal in-expectation guarantee.

A natural question is whether these LOO estimators also enjoy sharp high-probability guarantees. In fact, a conjecture [\[90\]](#) asserts that the one-inclusion graph algorithm achieves the optimal PAC guarantee of [\(1.1\)](#). We show that this conjecture is false; there are settings where the one-inclusion graph algorithm has error at least

$$R_P(\hat{f}_{\text{OIG}}) \geq c \frac{d}{n\delta},$$

with probability at least δ , where c is a universal constant. This lower bound shows that Warmuth's conjecture is false in the strongest sense: one cannot improve on the bound obtained by applying Markov's inequality to the in-expectation guarantee.

Thus, while these LOO-based estimators are quite natural and yield non-trivial guarantees in-expectation, we cannot hope to get sharp bounds on their risk with high probability in general. Our main result shows that this obstruction can be remedied by aggregation. Instead of using a single predictor trained on the whole sample, we consider the sequence of predictors trained on prefixes of the data,

$$\hat{f}(\cdot; S_{\leq 1}), \hat{f}(\cdot; S_{\leq 2}), \dots, \hat{f}(\cdot; S_{\leq n}),$$

where $S_{\leq t}$ denotes the first t training samples. We then take the predictors in the suffix of this sequence (that have been trained on a constant fraction of the data) and aggregate them in a loss specific way.² We show that when the predictor $\hat{f}$ satisfies the LOO guarantee and a natural symmetry condition, the above aggregation $\hat{f}_{\text{agg}}$ achieves risk

$$R_P(\hat{f}_{\text{agg}}) = O\left(\frac{M_n}{n} + \frac{1}{n} \log\left(\frac{1}{\delta}\right)\right),$$

with probability at least $1 - \delta$ over the training sample S . Thus, sharp in-expectation results can be “boosted” to sharp high probability bounds on the risk. Moreover, the above bound is sharp for the one-inclusion graph algorithm and its multiclass extension.

1.3 Heavy Tailed Mean Estimation

Chapter 4 focuses on heavy-tailed mean estimation. Let $X_1, \dots, X_n$ be independent samples from a distribution P on $\mathbb{R}^d$ with mean μ and finite covariance matrix Σ . We make no assumptions on higher moments. The goal is to estimate μ under a general ℓ_p norm $\|\cdot\|_p$ for $p \geq 1$. For the Euclidean norm and isotropic data ($\Sigma = I_d$), the empirical mean satisfies

$$\mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^n X_i - \mu \right\|_2 \leq \sqrt{\frac{d}{n}},$$

which is optimal. In this sense, ℓ_2 is unusually benign and the sample mean is an optimal estimator.

The situation changes drastically for ℓ_p norms with $p > 2$. Under only a covariance assumption, the empirical mean can be suboptimal even in-expectation. Our main result is an optimal characterization of the minimax risk for estimating the mean under ℓ_p norms. Interestingly, the minimax risk exhibits a phase transition governed by the interaction between the sample size n , the dimension d , and the truncated parameter

$$\bar{p} = \min\{p, \max\{\log d, 1\}\}.$$

²For example, in classification settings we aggregate by taking the majority vote of the predictors.

Concretely, we show that the minimax expected risk for estimation in ℓ_p norm over distributions with covariance I_d is given by

$$\inf_{\hat{\mu}} \sup_{\substack{\mu \in \mathbb{R}^d \\ P: \mathbb{E}_P X = \mu \\ \text{Cov}_P(X) = I_d}} \asymp \begin{cases} d^{\frac{1}{2\lceil \frac{1}{2} \rceil}}, & \bar{p} > 2 \lceil \frac{n}{2} \rceil, \\ d^{\frac{1}{\bar{p}}} \sqrt{\frac{\bar{p}}{n}}, & \bar{p} \leq 2 \lceil \frac{n}{2} \rceil. \end{cases}$$

The second branch is the familiar Gaussian-width rate: if $G \sim N(0, I_d)$, then

$$\frac{\mathbb{E} \|G\|_p}{\sqrt{n}} \asymp d^{1/\bar{p}} \sqrt{\frac{\bar{p}}{n}}.$$

The first branch is a genuinely heavy-tailed phenomena that appears in the small sample regime and demonstrates the impossibility of matching the gaussian rate for all sample sizes. We show that an estimator that achieves this bound is based on aggregating the natural (but sub-optimal) mean estimator. We use the coordinate-wise median-of-means estimator:

$$\hat{\mu}_j = \text{Med} \left(\bar{X}_j^{(1)}, \dots, \bar{X}_j^{(t)} \right), \quad j = 1, \dots, d,$$

where the number of buckets t is chosen based on the geometry of the norm via the truncated parameter $\bar{p}$:

$$t = 2 \left\lceil \frac{\bar{p}}{2} \right\rceil - 1.$$

Notice that when $p \leq 2$ we have $t = 1$, in which case the estimator is the sample mean.

Our second result determines the sharp high-probability minimax rate for mean estimation in the ℓ_∞ norm under only a covariance assumption. If $\text{Cov}(X) = I_d$, then the optimal error at confidence level $1 - \delta$ is, up to universal constants,

$$\begin{cases} \left(\frac{d}{\delta} \right)^{\frac{1}{2\lceil \frac{1}{2} \rceil}}, & n \lesssim \log(d/\delta), \\ \sqrt{\frac{\log(d/\delta)}{n}}, & n \gtrsim \log(d/\delta). \end{cases}$$

The upper bound is again achieved by the coordinate-wise median-of-means estimator with $t = 2\lceil \log(d/\delta) \rceil$ blocks. Thus, even for the ℓ_∞ norm, the usual sub-Gaussian rate captures only the large-sample regime; under only a covariance assumption, there is an additional finite-sample heavy-tailed phase.

We also show that natural projection-based median of means estimators, of the type used in the general norm theory, can be suboptimal for ℓ_p norms. Such estimators search for a point whose projection in every dual direction agrees with the median of the projected block means:

$$\hat{\mu}_{\text{proj}} \in \arg \min_x \sup_{\|v\|_q \leq 1} \left| \langle v, x \rangle - \text{Med} \left(\{ \langle v, \bar{X}^{(1)} \rangle, \dots, \langle v, \bar{X}^{(t)} \rangle \} \right) \right|, \quad \frac{1}{p} + \frac{1}{q} = 1.$$

Although this construction is natural as a norm-agnostic procedure, it can be too crude for structured norms such as ℓ_p . In particular, for $p = \infty$ and $\log d \lesssim n \ll \sqrt{d}$, the coordinate-wise estimator achieves error of order

$$\sqrt{\frac{\log d}{n}},$$

whereas projection-based median of means estimators can have error of order

$$\frac{\sqrt{d}}{n}$$

with constant probability. Thus, projection-based estimators may pay a polynomial price in the dimension, even in regimes where the coordinate-wise estimator is sharp.

Finally, while the Euclidean norm is the most benign geometry for in-expectation mean estimation, the high-probability problem still contains a limitation for coordinate-wise median of means estimators. While the sample mean is optimal in-expectation for the ℓ_2 setting, we show that for high-probability guarantees, no single distribution-independent choice of the number of buckets in a coordinate-wise median of means estimator is optimal for all distributions with identity covariance. Thus, the coordinate-wise median construction is computationally simple and optimal for the ℓ_p expectation regimes studied here, but it is not a universally optimal high-probability procedure, even in the Euclidean case.

Chapter 2

Majority-of-Three: The Simplest Optimal Learner?

This chapter is based on joint work with Mikael Møller Høgsgaard, Kasper Green Larsen, and Nikita Zhivotovskiy [3].

In the setting of realizable Probably Approximately Correct (PAC) learning [81, 84, 87], the goal is to learn or approximate an unknown target function $f^* \in \{0, 1\}^{\mathcal{X}}$ from a labelled training sample $(S, f^*(S)) = ((X_1, f^*(X_1)), \dots, (X_n, f^*(X_n)))$, where the X_i 's are i.i.d. samples from an unknown distribution P over an instance space $\mathcal{X}$. In the realizable setting, we are furthermore promised that f^* belongs to a known function class $\mathcal{F} \subseteq \{0, 1\}^{\mathcal{X}}$ of Vapnik-Chervonenkis (VC) dimension d .

Given a labelled training sample $(S, f^*(S))$, a learning algorithm produces a function $\hat{f}_S \in \{0, 1\}^{\mathcal{X}}$ with the goal of minimizing the probability of mispredicting the label of a new sample from P , where we denote this error by $\text{err}_P(\hat{f}_S) := \Pr_{X \sim P}[\hat{f}_S(X) \neq f^*(X)]$. The simplest reasonable learning algorithm, known as *empirical risk minimization* (ERM), simply reports an arbitrary function $\hat{f}_S \in \mathcal{F}$ that is consistent with f^* on the training data, i.e. $\hat{f}_S(X_i) = f^*(X_i)$ for all $i = 1, \dots, n$. Classic work by Blumer et al. [21] (the same bound also essentially follows from the earlier works [85, 83]) shows that for any $\delta > 0$, it holds with probability $1 - \delta$ over S that any $\hat{f}_S \in \mathcal{F}$ consistent with f^* on S has

$$\text{err}_P(\hat{f}_S) = O\left(\frac{d}{n} \log\left(\frac{n}{d}\right) + \frac{1}{n} \log\left(\frac{1}{\delta}\right)\right). \quad (2.1)$$

On the lower bound side, there exists an instance space $\mathcal{X}$ and function class $\mathcal{F}$ such that for a certain ERM algorithm, there is a target function $f^* \in \mathcal{F}$ and hard distribution P for which (2.1) is tight [42, 15, 77, 36]. Learning algorithms that always output a function in $\mathcal{F}$ are referred to as *proper* learning algorithms. Generally, it is known that not only ERM, but all proper learners fail to achieve the optimal error bound in the PAC learning framework. See the corresponding lower bounds in [22].

For *improper* learning algorithms — algorithms that are allowed to output an arbitrary function $\hat{f}_S \in \{0, 1\}^{\mathcal{X}}$ — known lower bounds on the error only imply that we must have

$$\text{err}_P(\hat{f}_S) = \Omega\left(\frac{d}{n} + \frac{1}{n} \log\left(\frac{1}{\delta}\right)\right). \quad (2.2)$$

Developing an algorithm with a matching error upper bound, or strengthening the lower bound, was a major open problem for decades. This was finally resolved in 2015 when Hanneke [36], building on the work of Simon [77], proposed the first optimal algorithm with an error upper bound matching (2.2), leading to the optimal error bound

$$\Theta\left(\frac{d}{n} + \frac{1}{n} \log\left(\frac{1}{\delta}\right)\right). \quad (2.3)$$

Hanneke’s algorithm is based on constructing a large number ($\approx n^{0.79}$) of sub-samples $S_1, S_2, \dots \subseteq S$ of the training data. This algorithm then runs ERM on each $(S_i, f^*(S_i))$ to obtain functions $\hat{f}_{S_1}, \hat{f}_{S_2}, \dots$ and finally combines them via a majority vote. The sub-samples S_i are constructed to have a carefully designed overlapping structure, and an intricate inductive argument exploiting this structure is then used to argue optimality. Recent work by Larsen [52] shows that the carefully designed overlap structure may instead be replaced by the significantly simpler strategy of sampling each S_i as $\Theta(n)$ samples with replacement from S . This algorithm is precisely the classic heuristic known as Bagging, or bootstrap aggregation, due to Breiman [23]. Furthermore, the proof shows that a mere $O(\log(n/\delta))$ sub-samples suffice for an optimal sample complexity. The proof is however even more involved than Hanneke’s and uses his analysis at its core.

Another line of work studied an alternative learning algorithm, the one-inclusion graph algorithm of Haussler, Littlestone, and Warmuth [42] that returns a function $\hat{f}_{\text{OIG}}$. This work also introduces the *prediction model* of learning, which focuses on achieving bounds on the *expected error* rather than *high probability* bounds on the error. The one-inclusion graph algorithm was initially shown to have an expected error of

$$\mathbb{E}_{S \sim P^n} \left[\text{err}_P(\hat{f}_{\text{OIG}}) \right] \leq \frac{d}{n+1}, \quad (2.4)$$

which was later proven to be optimal within this prediction model [54]. Because of the tightness of the in-expectation bound (2.4), Warmuth conjectured [90] that the one-inclusion graph algorithm achieves an error upper bound matching the general lower bound (2.2) in the high probability regime. The lower-bound result in Chapter 3 unfortunately refutes this conjecture. While we do show in Chapter 3 that combining the output of $\Omega(n)$ predictions made by one-inclusion algorithms on prefixes of the training data $((X_1, f^*(X_1)), \dots, (X_m, f^*(X_m)))$ for $m = n/2, \dots, n$ via a majority is optimal in the high probability regime, unfortunately, the one-inclusion graph algorithm (and this extension) is much less intuitive than the aforementioned algorithms based on taking majority votes of ERMs.

The simplest possible optimal algorithm?

In light of prior work, we have several provably optimal algorithms for PAC learning in the realizable setting. The algorithms and their analyses vary in complexity and a natural question remains: What is the simplest possible optimal algorithm? We know from lower bounds that the algorithm has to be improper and as such must be more complicated than ERM. Bagging is arguably the simplest algorithm among previous proposed algorithms, but has the most difficult analysis. The voting among one-inclusion algorithms has a somewhat simple proof, but the algorithm is not the simplest. In this work, we consider what is perhaps the simplest imaginable improper algorithm, *Majority-of-Three* (ERMs): Partition S into three equal-sized disjoint pieces S_1, S_2, S_3 , run the same ERM algorithm on each $(S_i, f^*(S_i))$ to obtain $\hat{f}_{S_1}, \hat{f}_{S_2}, \hat{f}_{S_3}$, and combine them via a majority vote to produce the function $\text{Maj}(\hat{f}_{S_1}, \hat{f}_{S_2}, \hat{f}_{S_3})$. Since a majority vote of two functions is undefined when the functions disagree, this is arguably the simplest possible improper algorithm. Our first main result shows that this concrete majority vote of three ERMs, which we will refer to as Majority-of-Three throughout, is optimal in expectation.

Theorem 2.0.1. *Fix a function class $\mathcal{F} \subseteq \{0, 1\}^{\mathcal{X}}$ with VC dimension d . Fix a distribution P over $\mathcal{X}$ and target function $f^* \in \mathcal{F}$. For any ERM algorithm $\hat{f} : \mathcal{X} \times \mathcal{Z}^* \rightarrow \{0, 1\}$ it follows that*

$$\mathbb{E}_{S_1, S_2, S_3 \sim P^n} \left[\text{err}_P \left(\text{Maj}(\hat{f}_{S_1}, \hat{f}_{S_2}, \hat{f}_{S_3}) \right) \right] = O \left(\frac{d}{n} \right).$$

This result shows that Majority-of-Three matches the optimal expectation bound (2.4) achieved by the one-inclusion graph algorithm, up to a universal constant. Furthermore, our proof of Theorem 2.0.1 is in fact quite simple, especially compared to the previous proof that Bagging is optimal.

We note here that a single ERM alone is sub-optimal by a multiplicative $\ln(n/d)$ factor in-expectation (see the well-known lower bound in [42, Theorem 4.2]). We emphasize that in Theorem 2.0.1, the ERMs corresponding to S_1, S_2 and S_3 can be chosen by *any* algorithm $\hat{f}$ that outputs functions consistent with the sample. The only restriction is that it is the same algorithm $\hat{f}$ that is run on each S_i (and that the subsets S_i are disjoint and thus i.i.d.).

We now turn our attention to the high-probability regime, where we prove the following result.

Theorem 2.0.2. *Fix a function class $\mathcal{F} \subseteq \{0, 1\}^{\mathcal{X}}$ with VC dimension d . Fix a distribution P over $\mathcal{X}$ and target function $f^* \in \mathcal{F}$. Fix any ERM algorithm $\hat{f} : \mathcal{X} \times \mathcal{Z}^* \rightarrow \{0, 1\}$. For any parameter $\delta \in (0, 1/2]$ it holds with probability at least $1 - \delta$ over the randomness of $S_1, S_2, S_3 \sim P^n$ that*

$$\text{err}_P \left(\text{Maj}(\hat{f}_{S_1}, \hat{f}_{S_2}, \hat{f}_{S_3}) \right) = O \left(\frac{d}{n} \log \left(\log \left(\min \left\{ \frac{n}{d}, \frac{1}{\delta} \right\} \right) \right) + \frac{1}{n} \log \left(\frac{1}{\delta} \right) \right).$$

This bound is sub-optimal due to the $\log(\log(\min\{n/d, 1/\delta\}))$ term, however the additive $\log(1/\delta)$ term dominates for $\delta \leq d^{-d}$. Thus, Majority-of-Three is optimal both in the constant (Theorem 2.0.1) and high-probability regimes (Theorem 2.0.2). Because of this, we conjecture that Majority-of-Three is in fact optimal for all δ and leave this as an open question for future research.

Conjecture 2.0.3. *Fix a function class $\mathcal{F} \subseteq \{0, 1\}^{\mathcal{X}}$ with VC dimension d . Fix a distribution P over $\mathcal{X}$ and target function $f^* \in \mathcal{F}$. Fix any ERM algorithm $\hat{f} : \mathcal{X} \times \mathcal{Z}^* \rightarrow \{0, 1\}$. For any parameter $\delta \in (0, 1)$ it holds with probability at least $1 - \delta$ over the randomness of $S_1, S_2, S_3 \sim P^n$ that*

$$\text{err}_P \left(\text{Maj}(\hat{f}_{S_1}, \hat{f}_{S_2}, \hat{f}_{S_3}) \right) = O \left(\frac{d}{n} + \frac{1}{n} \log \left(\frac{1}{\delta} \right) \right).$$

An alternative by Simon

In his breakthrough work, Simon [77] proposed taking majority votes of three ERMs trained on certain sub-samples of the training sample.¹ However, his algorithm is slightly different than ours. Concretely, he proposed the following algorithm: given an ERM algorithm and labelled training sample, partition S into three equal-sized disjoint pieces S_1, S_2, S_3 and for $i = 1, 2, 3$, run any ERM algorithm on $((S_1, \dots, S_i), f^*((S_1, \dots, S_i)))$ to obtain $\hat{f}_{S_1}, \hat{f}_{(S_1, S_2)}, \hat{f}_{(S_1, S_2, S_3)}$, and combine them via a majority vote to produce the function $\text{Maj}(\hat{f}_{S_1}, \hat{f}_{(S_1, S_2)}, \hat{f}_{(S_1, S_2, S_3)})$. Intuitively, more training data for the ERM should be better and Simon also proved the following high-probability upper bound on his algorithm's error:

$$\text{err}_P \left(\text{Maj} \left(\hat{f}_{S_1}, \hat{f}_{(S_1, S_2)}, \hat{f}_{(S_1, S_2, S_3)} \right) \right) = O \left(\frac{d}{n} \log \left(\log \left(\frac{n}{d} \right) \right) + \frac{1}{n} \log \left(\frac{1}{\delta} \right) \right). \quad (2.5)$$

This bound is asymptotically smaller than the tight bound (2.1) that holds for a single ERM.

We note that Simon also discusses the applicability of his analysis to more general majorities of ERMs including the Majority-of-Three function $\text{Maj}(\hat{f}_{S_1}, \hat{f}_{S_2}, \hat{f}_{S_3})$ analyzed in this work.² However, adopting the approach in [77], the error of Majority-of-Three is controlled by the same upper bound as expressed in (2.5), which is suboptimal as demonstrated by Theorem 2.0.1. Furthermore, we additionally remark that a similar in spirit construction based on the majority of three functions has been extensively studied in Schapire's PhD thesis [76]. However, his approach (inspired by what we now know as boosting) works with essentially any learning algorithm and is not necessarily limited to ERM.

In the same work [77], Simon further showed that for a specific function class $\mathcal{F}$ for which there is a choice of target function $f^* \in \mathcal{F}$ and hard distribution P that certify the tightness

¹Simon studied majority votes over any odd number L of ERMs trained on specific sub-samples of the data. He also proved bounds on the error of these majority votes that shrunk as L increased.

²Simon's analysis applies to any majority where each of the participating ERMs is trained on an independent constant fraction of the training sample.

of (2.1) for a certain choice ERM, his algorithm can actually achieve an optimal upper bound matching (2.2) for $\mathcal{F}$ regardless of the choice of $f^* \in \mathcal{F}$ and P . Unfortunately, we prove the following lower bound that shows that the upper bound (2.5) cannot be improved in general, answering a question posed by Simon.

Theorem 2.0.4. *For any sample size n that is divisible by 6 and positive integer $d \leq n$, there is a function class $\mathcal{F} \subseteq \{0, 1\}^{[0,1]}$ with VC dimension $4d$, distribution P over $[0, 1]$, target function $f^* \in \mathcal{F}$, and an ERM algorithm $\hat{f} : \mathcal{X} \times \mathcal{Z}^* \rightarrow \{0, 1\}$ such that the following holds: given i.i.d. training samples $S_1, S_2, S_3 \sim P^n$,*

$$\text{err}_P \left(\text{Maj} \left(\hat{f}_{S_1}, \hat{f}_{(S_1, S_2)}, \hat{f}_{(S_1, S_2, S_3)} \right) \right) = \Omega \left(\frac{d}{n} \log \left(\log \left(\frac{n}{d} \right) \right) \right),$$

with probability at least $2/3$ over the randomness of $S = (S_1, S_2, S_3)$.

This result shows that Simon's algorithm unfortunately cannot achieve the optimal bound (2.3) in general. This indicates that it is important that the ERM algorithm used in Majority-of-Three is instantiated on disjoint subsets of the training sample.

Notation

We use $\mathcal{X}$ to denote the *instance space*, $\mathcal{F} \subseteq \{0, 1\}^{\mathcal{X}}$ to denote a function class, and let $\mathcal{Z} = \mathcal{X} \times \{0, 1\}$. Throughout, P is a distribution over $\mathcal{X}$ and $f^* \in \mathcal{F}$ is the unknown *target function* in the class. For $n \in \mathbb{N}$ and a distribution P , we denote by P^n the product distribution of P . We say that a sequence $S = (X_1, \dots, X_n)$ is a *training sample* of size n where X_i are i.i.d. samples from a distribution P . For a training sample $S = (X_1, \dots, X_n)$, we find it useful to write $(S, f^*(S)) = ((X_1, f^*(X_1)), \dots, (X_n, f^*(X_n)))$, and we call this the *labelled training sample*. For training samples $S_1 = (X_1, \dots, X_n)$ and $S_2 = (X_{n+1}, \dots, X_{n+m})$ we let $(S_1, S_2) = (X_1, \dots, X_n, X_{n+1}, \dots, X_{n+m})$, and for S_1, S_2 and S_3 we take $(S_1, S_2, S_3) = ((S_1, S_2), S_3)$. We define the error of a binary function f under distribution P and target function f^* to be $\text{err}_P(f) = \mathbf{Pr}_{X \sim P}[f(X) \neq f^*(X)]$. For any measurable set $R \subseteq \mathcal{X}$, we define P_R to be the conditional distribution of P restricted to R , i.e. for $X \sim P_R$ we have that for any measurable function g that $\mathbb{E}_{X \sim P_R}[g(X)] = \mathbb{E}_{X \sim P}[g(X)\mathbf{1}\{X \in R\}] / \mathbf{Pr}_{X \sim P}[X \in R]$.

For a function class $\mathcal{F}$ and subset $U = \{x_1, \dots, x_d\} \subseteq \mathcal{X}$ of d points we let $\mathcal{F}|_U$ denote the set $\{y \in \{0, 1\}^d \mid \exists f \in \mathcal{F} : \forall i \in [d], f(x_i) = y_i\}$. The *Vapnik-Chervonenkis (VC) dimension* is then defined as the largest number d such that there exists a point set $U \subseteq \mathcal{X}$ of size d such that the cardinality of $\mathcal{F}|_U$ is 2^d . We use $\log(x)$ and $\ln(x)$ to denote $\log_2(x)$ and $\log_e(x)$ respectively and we also use $\text{Log}(x) := \max\{2, \log_2(x)\}$ to denote a truncated logarithm.

Let $\mathcal{Z}^* = \cup_{i=1}^{\infty} \mathcal{Z}^i$ be the set of all possible labelled training samples. We define a *learning algorithm* $\hat{f}$ to be a mapping $\hat{f} : \mathcal{X} \times \mathcal{Z}^* \rightarrow \{0, 1\}$. That is, given a labelled training sample $(S, f^*(S))$ as input, $\hat{f}(\cdot; (S, f^*(S))) : \mathcal{X} \rightarrow \{0, 1\}$ is the function that is learned from $(S, f^*(S))$. For ease of reading, we often denote the learned function by $\hat{f}_S := \hat{f}(\cdot; (S, f^*(S)))$.

A learning algorithm $\hat{f}$ is an *Empirical Risk Minimizer (ERM)* for the class $\mathcal{F}$ if, given a labelled training sample $(S, f^*(S))$ as input, it output a function $\hat{f}_S$ in $\mathcal{F}$ that satisfies $\hat{f}_S(X_i) = f^*(X_i)$ for every X_i that appears in S . We define the majority vote of k binary functions $f_1, \dots, f_k : \mathcal{X} \rightarrow \{0, 1\}$ to be the function

$$\text{Maj}(f_1, \dots, f_k)(x) := \mathbf{1}\{f_1(x) + \dots + f_k(x) > k/2\}.$$

2.1 Majority-of-Three is optimal in-expectation

In this section, we prove the main in-expectation result for the Majority-of-Three algorithm. Before we prove our result, we will find it helpful to introduce some auxiliary notation. Throughout this section, we set $\hat{f} : \mathcal{X} \times \mathcal{Z}^* \rightarrow \{0, 1\}$ to be a fixed (but arbitrary) ERM algorithm. Fix a distribution P over $\mathcal{X}$ and let $f^* \in \mathcal{F}$ be the target function. For any $x \in \mathcal{X}$ we let

$$p_x(n, f^*, P) = \mathbf{Pr}_{S \sim P^n} [\hat{f}_S(x) \neq f^*(x)].$$

In words, $p_x(n, f^*, P)$ is the chance that $\hat{f}_S$ errs on the point x for an *average* sample $S \sim P^n$. We now define a partition of $\mathcal{X}$ based on $p_x(n, f^*, P)$. Consider the following sets for any $i \in \mathbb{N}$:

$$R_i(n, f^*, P) = \{x \in \mathcal{X} : p_x(n, f^*, P) \in (2^{-i}, 2^{-i+1}]\}.$$

We often write $R_i = R_i(n, f^*, P)$ and $p_x = p_x(n, f^*, P)$ since n , P , and f^* will always be clear from the context. With this notation in place, we are now ready to prove that Majority-of-Three has an optimal in-expectation upper bound on its error.

Theorem 2.0.1. *Fix a function class $\mathcal{F} \subseteq \{0, 1\}^{\mathcal{X}}$ with VC dimension d . Fix a distribution P over $\mathcal{X}$ and target function $f^* \in \mathcal{F}$. For any ERM algorithm $\hat{f} : \mathcal{X} \times \mathcal{Z}^* \rightarrow \{0, 1\}$ it follows that*

$$\mathbb{E}_{S_1, S_2, S_3 \sim P^n} \left[\text{err}_P \left(\text{Maj}(\hat{f}_{S_1}, \hat{f}_{S_2}, \hat{f}_{S_3}) \right) \right] = O \left(\frac{d}{n} \right).$$

To prove [Theorem 2.0.1](#), we require the following lemma which says that two ERMs trained on 2 i.i.d. training samples of the same size rarely makes a mistake on the same point.

Lemma 2.1.1. *Fix a function class $\mathcal{F} \subseteq \{0, 1\}^{\mathcal{X}}$ with VC dimension d . Fix a distribution P over $\mathcal{X}$ and target function $f^* \in \mathcal{F}$. For any ERM algorithm $\hat{f} : \mathcal{X} \times \mathcal{Z}^* \rightarrow \{0, 1\}$ it follows that*

$$\mathbb{E}_{S_1, S_2 \sim P^n} \left[\mathbf{Pr}_{X \sim P} \left[\hat{f}_{S_1}(X) \neq f^*(X) \wedge \hat{f}_{S_2}(X) \neq f^*(X) \right] \right] \leq c \frac{d}{n},$$

where c is a universal constant.

We postpone the proof of [Lemma 2.1.1](#) for now and show how it implies [Theorem 2.0.1](#).

Proof of Theorem 2.0.1. For any fixed $x \in \mathcal{X}$ and fixed samples S_1, S_2, S_3 , if $\text{Maj}(\hat{f}_{S_1}, \hat{f}_{S_2}, \hat{f}_{S_3})(x) \neq f^*(x)$, then there must be at least two distinct indices $i, j \in [3]$ such that $\hat{f}_{S_i}(x) \neq f^*(x)$ and $\hat{f}_{S_j}(x) \neq f^*(x)$. So,

$$\begin{aligned} \text{err}_P \left(\text{Maj}(\hat{f}_{S_1}, \hat{f}_{S_2}, \hat{f}_{S_3}) \right) &= \Pr_{X \sim P} \left[\text{Maj}(\hat{f}_{S_1}, \hat{f}_{S_2}, \hat{f}_{S_3})(X) \neq f^*(X) \right] \\ &\leq \sum_{\substack{i, j \in [3] \\ i < j}} \Pr_{X \sim P} \left[\hat{f}_{S_i}(X) \neq f^*(X) \wedge \hat{f}_{S_j}(X) \neq f^*(X) \right]. \end{aligned}$$

Combining the above and Lemma 2.1.1 gives us

$$\mathbb{E}_{S_1, S_2, S_3 \sim P^n} \left[\text{err}_P \left(\text{Maj}(\hat{f}_{S_1}, \hat{f}_{S_2}, \hat{f}_{S_3}) \right) \right] \leq 3c \frac{d}{n}.$$

This concludes the proof. $\square$

We now move on to proving Lemma 2.1.1, where we will use the following lemma.

Lemma 2.1.2. Fix a function class $\mathcal{F} \subseteq \{0, 1\}^{\mathcal{X}}$ with VC dimension d . Fix a distribution P over $\mathcal{X}$, target function $f^* \in \mathcal{F}$, and $R \subseteq \mathcal{X}$ such that $\Pr_{X \sim P} [X \in R] \neq 0$. For any ERM algorithm $\hat{f}: \mathcal{X} \times \mathcal{Z}^* \rightarrow \{0, 1\}$ it follows that

$$\mathbb{E}_{S \sim P^n} \left[\text{err}_{P_R} \left(\hat{f}_S \right) \right] \leq 20 \frac{d \text{Log}(e \Pr_{X \sim P} [X \in R] n / d)}{\Pr_{X \sim P} [X \in R] n}.$$

Lemma 2.1.2 is an immediate consequence of the celebrated uniform convergence principle and a simple proof can be found in Section 2.4. We now prove Lemma 2.1.1.

Proof Lemma 2.1.1. Let S_1 and S_2 be independent samples from P^n . By the independence of S_1 and S_2 and the definition of p_x we have, for any $x \in \mathcal{X}$, that

$$\begin{aligned} \Pr_{S_1, S_2 \sim P^n} \left[\hat{f}_{S_1}(x) \neq f^*(x) \wedge \hat{f}_{S_2}(x) \neq f^*(x) \right] &= \prod_{i=1}^2 \Pr_{S_i \sim P^n} \left[\hat{f}_{S_i}(x) \neq f^*(x) \right] \\ &= \Pr_{S_1 \sim P^n} \left[\hat{f}_{S_1}(x) \neq f^*(x) \right]^2 = p_x^2. \end{aligned}$$

Using the above, the law of total expectation with partitioning $(R_i)_{i \in \mathbb{N}}$, and swapping the order of expectations (X and (S_1, S_2) being independent), we get that

$$\begin{aligned} &\mathbb{E}_{S_1, S_2 \sim P^n} \left[\Pr_{X \sim P} \left[\hat{f}_{S_1}(X) \neq f^*(X) \wedge \hat{f}_{S_2}(X) \neq f^*(X) \right] \right] \\ &= \sum_{i=1}^{\infty} \Pr_{X \sim P} [X \in R_i] \mathbb{E}_{X \sim P} [p_X^2 | X \in R_i] \leq \sum_{i=1}^{\infty} \Pr_{X \sim P} [X \in R_i] 2^{-2i+2}, \end{aligned}$$

where the inequality follows from the fact that $p_x \leq 2^{-i+1}$ for every $x \in R_i$. We will now show that $\mathbf{Pr}_{X \sim P}[X \in R_i] \leq cdi2^i/n$ for every $i \in \mathbb{N}$ (for a universal constant $c \geq 1$ chosen below), which combined with the above gives us

$$\mathbb{E}_{S_1, S_2 \sim P^n} \left[\mathbf{Pr}_{X \sim P} \left[\widehat{f}_{S_1}(X) \neq f^*(X) \wedge \widehat{f}_{S_2}(X) \neq f^*(X) \right] \right] \leq \frac{4cd}{n} \sum_{i=1}^{\infty} i2^{-i} \leq 8c \frac{d}{n}.$$

This yields the claim with the constant $8c$. Towards a contradiction, assume there is an $i \in \mathbb{N}$ such that $\mathbf{Pr}_{X \sim P}[X \in R_i] > cdi2^i/n$, which is equivalent to $\mathbf{Pr}_{X \sim P}[X \in R_i]n/d > ci2^i \geq 1$. Using this assumption, the fact that $x \rightarrow \text{Log}(ex)/x$ is decreasing for $x > 0$, and [Lemma 2.1.2](#), we have

$$\mathbb{E}_{S_1 \sim P^n} \left[\text{err}_{P_{R_i}}(\widehat{f}_{S_1}) \right] \leq 20 \frac{\text{Log}(e \mathbf{Pr}_{X \sim P}[X \in R_i]n/d)}{(\mathbf{Pr}_{X \sim P}[X \in R_i]n/d)} \leq 20 \frac{\text{Log}(eci2^i)}{ci2^i}. \quad (2.6)$$

By changing the order of expectations of the left hand side of the above and using the fact that $p_x > 2^{-i}$ for every $x \in R_i$, we also have

$$\mathbb{E}_{S_1 \sim P^n} \left[\text{err}_{P_{R_i}}(\widehat{f}_{S_1}) \right] = \mathbb{E}_{X \sim P_{R_i}}[p_X] > 2^{-i}. \quad (2.7)$$

Combining the upper bound (2.6), the lower bound (2.7), and the fact that the function $x \rightarrow \text{Log}(ex)/x$ is decreasing for $x > 0$, we get

$$1 < 20 \frac{\text{Log}(eci2^i)}{ci} \leq 20 \left(\frac{\text{Log}(eci)}{ci} + \frac{2}{c} \right) \leq 20 \frac{\text{Log}(ec) + 2}{c}.$$

However, for c large enough, the right hand side of the above is strictly less than 1. This gives us the desired contradiction and concludes the proof. $\square$

2.2 A lower bound for certain majorities

In this section, we prove that not all majorities of 3 ERMs trained on subsets of the data are optimal. In particular, we show that Simon's [77] original partitioning scheme of the training sample into 3 sub-samples can produce a majority of 3 ERMs with sub-optimal error. Recall Simon's algorithm: partition the training sample $S = (S_1, S_2, S_3)$ into 3 equal pieces S_1, S_2, S_3 , train 3 ERMs $\widehat{f}_{S_1}, \widehat{f}_{(S_1, S_2)}, \widehat{f}_{(S_1, S_2, S_3)}$, and return the majority vote $\text{Maj}(\widehat{f}_{S_1}, \widehat{f}_{(S_1, S_2)}, \widehat{f}_{(S_1, S_2, S_3)})$. Simon proved that this algorithm enjoys the PAC upper bound

$$\text{err}_P \left(\text{Maj} \left(\widehat{f}_{S_1}, \widehat{f}_{(S_1, S_2)}, \widehat{f}_{(S_1, S_2, S_3)} \right) \right) = O \left(\frac{d}{n} \log \left(\log \left(\frac{n}{d} \right) \right) + \frac{1}{n} \log \left(\frac{1}{\delta} \right) \right).$$

The next theorem shows that this algorithm unfortunately has a matching lower bound on its error.

Theorem 2.0.4. *For any sample size n that is divisible by 6 and positive integer $d \leq n$, there is a function class $\mathcal{F} \subseteq \{0, 1\}^{[0,1]}$ with VC dimension $4d$, distribution P over $[0, 1]$, target function $f^* \in \mathcal{F}$, and an ERM algorithm $\hat{f} : \mathcal{X} \times \mathcal{Z}^* \rightarrow \{0, 1\}$ such that the following holds: given i.i.d. training samples $S_1, S_2, S_3 \sim P^n$,*

$$\text{err}_P \left(\text{Maj} \left(\hat{f}_{S_1}, \hat{f}_{(S_1, S_2)}, \hat{f}_{(S_1, S_2, S_3)} \right) \right) = \Omega \left(\frac{d}{n} \log \left(\log \left(\frac{n}{d} \right) \right) \right),$$

with probability at least $2/3$ over the randomness of $S = (S_1, S_2, S_3)$.

Comparing the above bound with the upper bound in [Theorem 2.0.1](#), we see that if the ERMs did not overlap in their sub-samples the log factor would not be present. The construction we use in our lower bound is a modification of the usual construction used to prove a lower bound on the error of a single ERM (see [\[15, 77\]](#)). In these constructions, one takes the domain $\mathcal{X}$ to be a finite set of size roughly $n/\log(n/d)$ where $n \geq d$ is the sample size³ and the function class $\mathcal{F}$ is taken to be all functions that assign the value 1 to at most d points on $\mathcal{X}$. Furthermore, the target function is set to be the 0 function, and the sampling distribution is the uniform distribution over $\mathcal{X}$. Finally, the “bad” ERM algorithm returns any function that assigns as many 1’s to the domain as possible, while being consistent on the observed samples. The error of this ERM is tightly connected to the number of unique elements we sample from the domain. One can then use a coupon collector argument to show that the error is $\Omega(d \log(n/d)/n)$ with constant probability.

Simon noticed that we cannot directly use this “hard instance” to prove a lower bound on his algorithm due to the structure of the class $\mathcal{F}$ [\[77, Theorem 7\]](#). We get around this by considering a version of this construction that uses a continuous domain (instead of finite) and a function class consisting of functions that are unions of intervals (instead of points).

Before we prove [Theorem 2.0.4](#), it will be convenient to introduce the following notation. For a positive integer d and non-empty set A , we define $A_{[d]}$ to be the set consisting of the smallest d elements of A with respect to an ordering of the elements of A . The ordering we use will be clarified when needed. We now prove [Theorem 2.0.4](#).

Proof of Theorem 2.0.4. Fix a sample size n divisible by 6 and positive integer $d \leq n$. Throughout, we will assume any interval considered is left-open and right-closed. A collection of intervals $I_1, \dots, I_t \subseteq [0, 1]$ can be viewed as the binary function $f_{I_1 \cup \dots \cup I_t}$ that satisfies $f_{I_1 \cup \dots \cup I_t}(x) = 1$ if and only if there is an index j such that $x \in I_j$. We will consider the function class $\mathcal{F}$ that is the collection of all functions corresponding to the union of at most $2d$ interval. It is not hard to show that this class has VC dimension $4d$. We take P to be the uniform distribution on the domain $[0, 1]$ and choose the target function f^* to be the 0 function on the domain $[0, 1]$.

We now describe the “bad” ERM algorithm $\hat{f} : \mathcal{X} \times \mathcal{Z}^* \rightarrow \{0, 1\}$. For the remainder of the proof, $C > 0$ is a large universal constant that we will determine below. For a training sample size m , we define three collections of sets:

³These results are often stated as lower bounds on the *sample complexity* for some target error ϵ .

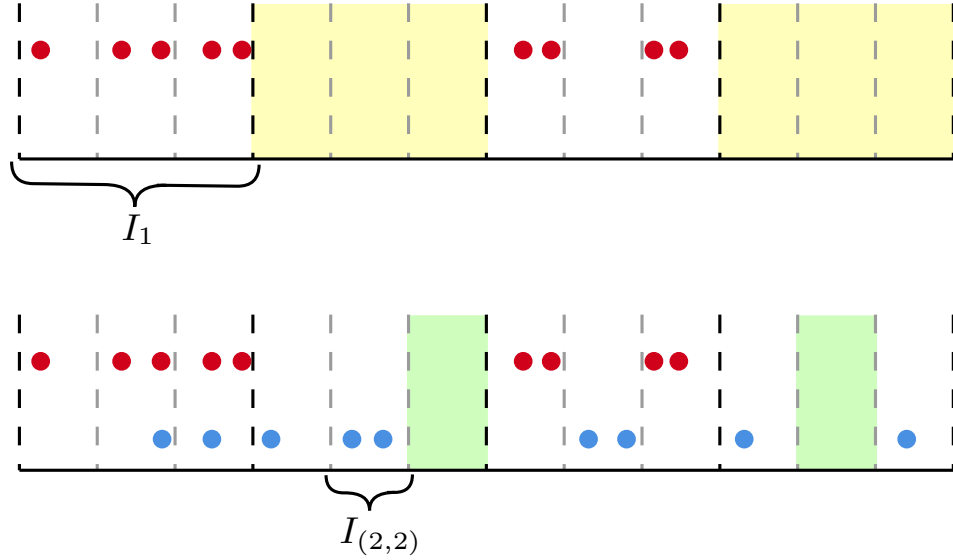


Figure 2.1: An illustration of the partitioning of the interval $(0, 1]$ for a training sample consisting of $m = 18$ points with $d = 2$. The interval $(0, 1]$ is partitioned into 4 intervals $I_1, \dots, I_4$. Each interval I_i is further partitioned into the 4 subintervals $I_{(i,1)}, \dots, I_{(i,4)}$. The red points correspond to the first half of the sample $(X_1, \dots, X_9)$ and the blue points correspond to the second half of the sample $(X_{10}, \dots, X_{18})$. The yellow highlighted regions are the first d intervals I_2 and I_4 that contain no points from $(X_1, \dots, X_9)$. The green highlighted regions are the first d subintervals of I_2 and I_4 that contain no points from $(X_{10}, \dots, X_{18})$. The green intervals are added to the union of intervals used by $\hat{f}_S$ as their indices correspond to the set $L_1(S)$.

1. $\{I_i(m) : i \in [m_1]\}$ is the unique partition of $(0, 1]$ into $m_1 := \lceil Cm / \ln(Cm/d) \rceil$ intervals of the same length.
2. $\{I_{i,j}(m) : i \in [m_1], j \in [m_2]\}$ where, for a fixed i , $\{I_{i,j}(m) : j \in [m_2]\}$ is the unique partition of $I_i(m)$ into $m_2 := \lceil 4Cm / (m_1 \ln(\ln(Cm/d))) \rceil$ intervals of the same length.
3. $\{J_i(m) : i \in [m_3]\}$ is the unique partition of $(0, 1]$ into $m_3 := \lceil 2Cm / \ln(2Cm/d) \rceil$ intervals of the same length.

Given a labelled training sample $(S, f^*(S)) = ((X_1, 0), \dots, (X_m, 0))$ as input, the ERM algorithm $\hat{f}$ constructs the function $\hat{f}_S = \hat{f}(\cdot; (S, f^*(S)))$ in the following way:⁴

⁴This defines $\hat{f}_S$ when $(S, f^*(S))$ contains only 0 labels. On any $(S, f^*(S))$ that contains a 1 label we return an arbitrary consistent function.

1. For $i \in [m_1]$, $j \in [m_2]$, and $k \in [m_3]$ define the sets

$$\begin{aligned}\tilde{I}_i(S) &= \{x_1, \dots, x_{\lfloor m/2 \rfloor}\} \cap I_i(m), \\ \tilde{I}_{(i,j)}(S) &= \{x_{\lfloor m/2 \rfloor + 1}, \dots, x_m\} \cap I_{(i,j)}(m), \\ \tilde{J}_k(S) &= \{x_1, \dots, x_m\} \cap J_k(m).\end{aligned}$$

2. Define the index sets

$$\begin{aligned}L_1(S) &= \{(i, j) : i \in \{i' : \tilde{I}_{i'}(S) = \emptyset\}_{[d]}, \tilde{I}_{(i,j)}(S) = \emptyset\}_{[d]},^5 \\ L_2(S) &= \{k : \tilde{J}_k(S) = \emptyset\}_{[d]}.\end{aligned}$$

3. Define the union of intervals

$$I_S = \left(\bigcup_{(i,j) \in L_1(S)} I_{i,j}(m) \right) \cup \left(\bigcup_{i \in L_2(S)} J_i(m) \right).$$

4. Finally, define the function $\hat{f}_S = f_{I_S}$.

Observe that I_S is the union of at most $2d$ disjoint intervals, so $\hat{f}_S$ will always be in the class $\mathcal{F}$. Furthermore, $\hat{f}_S$ is always consistent with the sample S by construction. See Fig. 2.1 for an example of the resulting intervals considered by the set $L_1(S)$. Let $m = n/3$. From now on we use m_1 and m_2 to denote the number of intervals of the form $I_i(2m)$ and $I_{(i,j)}(2m)$ considered by $\hat{f}_{(S_1, S_2)}$ respectively. Consider the unions of intervals I_{S_1} and $I_{(S_1, S_2)}$ corresponding to the ERM functions $\hat{f}_{S_1}$ and $\hat{f}_{(S_1, S_2)}$. The number m is divisible by 2 from our choice of n , so it follows that $J_i(m) = I_i(2m)$ and $\tilde{J}_i(S_1) = \tilde{I}_i(S_1, S_2)$, which implies $L_2(S_1) = \{k : \tilde{I}_k((S_1, S_2)) = \emptyset\}_{[d]}$. Thus, $\hat{f}_{S_1}$ and $\hat{f}_{(S_1, S_2)}$ agree, and simultaneously err, on every subinterval $I_{(i,j)}(2m)$ with $(i, j) \in L_1(S_1, S_2)$. Because P is the uniform distribution and every interval $I_{(i,j)}(2m)$ has length $1/(m_1 m_2) = \Theta(\ln(\ln(n/d))/n)$, it follows that the error of the majority vote satisfies

$$\text{err}_P \left(\text{Maj} \left(\hat{f}_{S_1}, \hat{f}_{(S_1, S_2)}, \hat{f}_{(S_1, S_2, S_3)} \right) \right) \geq \frac{|L_1(S_1, S_2)|}{m_1 m_2} = \Omega \left(\frac{|L_1(S_1, S_2)|}{n} \ln \left(\ln \left(\frac{n}{d} \right) \right) \right).$$

Thus, if $|L_1(S_1, S_2)| = d$, we have

$$\text{err}_P \left(\text{Maj} \left(\hat{f}_{S_1}, \hat{f}_{(S_1, S_2)}, \hat{f}_{(S_1, S_2, S_3)} \right) \right) = \Omega \left(\frac{d}{n} \log \left(\log \left(\frac{n}{d} \right) \right) \right),$$

⁵The ordering used for pairs (i, j) and (i', j') is the natural one: $(i, j) \leq (i', j')$ if $i < i'$ or $i = i'$ and $j \leq j'$.

so the claim of the theorem follows once we prove that

$$\Pr_{(S_1, S_2) \sim P^{2m}} [|L_1(S_1, S_2)| = d] \geq 2/3.$$

To this end, let $E_1 = E_1(S_1)$ be the event that S_1 satisfies $|L_2(S_1)| = d$ and let $E_2 = E_2((S_1, S_2))$ be the event that (S_1, S_2) satisfies $|L_1(S_1, S_2)| = d$. Using the law of total probability we get,

$$\Pr_{(S_1, S_2) \sim P^{2n/3}} [|L_1(S_1, S_2)| = d] = \Pr_{(S_1, S_2) \sim P^{2m}} [E_2] \geq \Pr_{(S_1, S_2) \sim P^{2m}} [E_2 \mid E_1] \Pr_{S_1 \sim P^m} [E_1],$$

so it suffices to prove that $\Pr_{(S_1, S_2) \sim P^{2m}} [E_2 \mid E_1] \geq \sqrt{2/3}$ and $\Pr_{S_1 \sim P^m} [E_1] \geq \sqrt{2/3}$. We omit the proof of the later inequality since it is very similar to the proof of the former inequality.

When E_1 occurs, we have $|L_2(S_1)| = |\{k : \tilde{J}_k(S_1) = \emptyset\}_{[d]}| = |\{k : \tilde{I}_k((S_1, S_2)) = \emptyset\}_{[d]}| = d$. So, showing that the event E_2 occurs conditioned on E_1 is equivalent to showing that *at least* d subintervals in the collection $\{I_{(i,j)} : i \in L_2(S_1), j \in [m_2]\}$ do not contain any points from the sample S_2 . Let $Y \sim Q$ be the random variable that counts the number of points required to sample from P until $m_2 d - d$ subintervals in $\{I_{(i,j)} : i \in L_2(S_1), j \in [m_2]\}$ contain one of the sampled points. Furthermore, let $Y_t \sim Q_t$ denote the random variable that counts the number of trials required to cover $(t + 1)$ subintervals given that we have covered t . Notice that Y_t is a geometric random variable with parameter $p_t = \frac{m_2 d - t}{m_1 m_2} = \frac{d}{m_1} - \frac{t}{m_1 m_2}$ and $Y = \sum_{t=0}^{m_2 d - d - 1} Y_t$. It follows that

$$\Pr_{(S_1, S_2) \sim P^{2m}} [E_2 \mid E_1] \geq \Pr_{Y \sim Q} [Y \geq m] = \Pr_{Y_t \sim Q_t} \left[\sum_{t=0}^{m_2 d - d - 1} Y_t \geq m \right].$$

We can use a concentration inequality for sums of geometric random variables together with some simple calculations to show that

$$\Pr_{Y_t \sim Q_t} \left[\sum_{t=0}^{m_2 d - d - 1} Y_t \geq m \right] \geq \sqrt{2/3}, \quad (2.8)$$

when C is large enough. We defer these calculations to [Section 2.5](#). This concludes the proof. $\square$

2.3 High probability upper bound

In this section we prove our high-probability upper bound for the Majority-of-Three algorithm which we now restate for convenience.

Theorem 2.0.2. Fix a function class $\mathcal{F} \subseteq \{0, 1\}^{\mathcal{X}}$ with VC dimension d . Fix a distribution P over $\mathcal{X}$ and target function $f^* \in \mathcal{F}$. Fix any ERM algorithm $\hat{f} : \mathcal{X} \times \mathcal{Z}^* \rightarrow \{0, 1\}$. For any parameter $\delta \in (0, 1/2]$ it holds with probability at least $1 - \delta$ over the randomness of $S_1, S_2, S_3 \sim P^n$ that

$$\text{err}_P(\text{Maj}(\hat{f}_{S_1}, \hat{f}_{S_2}, \hat{f}_{S_3})) = O\left(\frac{d}{n} \log\left(\log\left(\min\left\{\frac{n}{d}, \frac{1}{\delta}\right\}\right)\right) + \frac{1}{n} \log\left(\frac{1}{\delta}\right)\right).$$

In this section it will be convenient to use the following notation: for a probability distribution P over $\mathcal{X}$ and set $R \subseteq \mathcal{X}$, we define $P(R) = \Pr_{X \sim P}[X \in R]$. [Theorem 2.0.2](#) is a consequence of the following technical lemma.

Lemma 2.3.1. Fix a function class $\mathcal{F} \subseteq \{0, 1\}^{\mathcal{X}}$ with VC dimension d . Fix a distribution P over $\mathcal{X}$ and target function $f^* \in \mathcal{F}$. Fix an ERM algorithm $\hat{f} : \mathcal{X} \times \mathcal{Z}^* \rightarrow \{0, 1\}$. For any parameter $\delta \in (0, 1/2]$ it holds with probability at least $1 - \delta$ over the randomness of $S_1, S_2 \sim P^n$ that

$$\Pr_{X \sim P} \left[\hat{f}_{S_1}(X) \neq f^*(X) \wedge \hat{f}_{S_2}(X) \neq f^*(X) \right] \leq c \left(\frac{d}{n} \text{Log} \left(\text{Log} \left(\min \left\{ \frac{n}{d}, \frac{1}{\delta} \right\} \right) \right) + \frac{1}{n} \text{Log} \left(\frac{1}{\delta} \right) \right),$$

where c is a universal constant.

We now prove [Theorem 2.0.2](#) using [Lemma 2.3.1](#) and postpone the proof of [Lemma 2.3.1](#).

Proof of Theorem 2.0.2. Since $\text{Maj}(\hat{f}_{S_1}, \hat{f}_{S_2}, \hat{f}_{S_3})(x) \neq f^*(x)$ happens if and only if there exists two distinct indices $i, j \in [3]$ such that $\hat{f}_{S_i}(X) \neq f^*(X)$ and $\hat{f}_{S_j}(X) \neq f^*(X)$, we get that

$$\text{err}_P(\text{Maj}(\hat{f}_{S_1}, \hat{f}_{S_2}, \hat{f}_{S_3})) \leq \sum_{\substack{i, j \in [3] \\ i < j}} \Pr_{X \sim P} \left[\hat{f}_{S_i}(X) \neq f^*(X) \wedge \hat{f}_{S_j}(X) \neq f^*(X) \right].$$

Using [Lemma 2.3.1](#) with confidence parameter $\delta/3$ for every distinct pair $i, j \in [3]$ together with a union bound gives us, with probability at least $1 - \delta$ over the randomness of (S_1, S_2, S_3) , that

$$\text{err}_P(\text{Maj}(\hat{f}_{S_1}, \hat{f}_{S_2}, \hat{f}_{S_3})) = O\left(\frac{d}{n} \text{Log} \left(\text{Log} \left(\min \left\{ \frac{n}{d}, \frac{1}{\delta} \right\} \right) \right) + \frac{1}{n} \text{Log} \left(\frac{1}{\delta} \right) \right).$$

This concludes the proof. $\square$

Before we prove [Lemma 2.3.1](#), we provide a short overview of the proof. Our first step is to reuse the idea from [Section 2.1](#) to partition the instance space $\mathcal{X}$ into sets $\{R_i\}_{i \in \mathbb{N}}$ based on the chance that an average ERM errs on a point in $x \in \mathcal{X}$. However, we use a different way to quantify the errors defining R_i by incorporating the failure parameter δ . For $i \geq 2$, we can actually reuse our in-expectation analysis from [Section 2.1](#) together with a simple

application of Markov's inequality and a sequence of union bounds. This gives us an upper bound on the joint error of two ERMs on the conditional distributions for all $\{R_i\}_{i \geq 2}$, with high probability. The major technical work of the proof lies in controlling the joint error of two ERMs on the conditional distribution of R_1 . To do this, we borrow an idea from Simon [77] that views the probability of $\widehat{f}_{S_1}$ and $\widehat{f}_{S_2}$ jointly erring as the probability that $\widehat{f}_{S_1}$ errs times the probability that $\widehat{f}_{S_2}$ errs conditioned on $\widehat{f}_{S_1}$ erring.⁶ A crucial technicality that differentiates our setting from Simon's is that the probability that $\widehat{f}_{S_1}$ and $\widehat{f}_{S_2}$ jointly err is taken over a *conditional distribution* P_R rather than the distribution P from which the samples S_1 and S_2 are drawn.

The following lemma formalizes how we can control the joint error of two ERMs under P_R .

Lemma 2.3.2. *Fix a function class $\mathcal{F} \subseteq \{0, 1\}^{\mathcal{X}}$ with VC dimension d . Fix a distribution P over $\mathcal{X}$, target function $f^* \in \mathcal{F}$ and $R \subseteq \mathcal{X}$ such that $P(R) \neq 0$. Fix an ERM algorithm $\widehat{f} : \mathcal{X} \times \mathcal{Z}^* \rightarrow \{0, 1\}$. For any parameter $\delta \in (0, 1/2]$ it holds with probability at least $1 - \delta$ over the randomness of $S_1, S_2 \sim P^n$ that*

$$\Pr_{X \sim P_R} \left[\widehat{f}_{S_1}(X) \neq f^*(X) \wedge \widehat{f}_{S_2}(X) \neq f^*(X) \right] \leq 8 \max \left\{ \frac{d \log(8e \log(eP(R)n/d))}{P(R)n}, \frac{\log(8/\delta)}{P(R)n} \right\}.$$

We now prove Lemma 2.3.1 using Lemma 2.3.2 and postpone the proof of Lemma 2.3.2.

Proof of Lemma 2.3.1. We use the same definition for p_x (see Section 2.1) but redefine the sets $\{R_i\}_{i \in \mathbb{N}}$ to be

$$R_1 = \{x \in \mathcal{X} : p_x \in (2^{-1}\delta / \log(1/\delta), 1]\},$$

and for any integer $i \geq 2$,

$$R_i = \{x \in \mathcal{X} : p_x \in (2^{-i}\delta / \log(1/\delta), 2^{-i+1}\delta / \log(1/\delta)]\}.$$

Using the law of total probability we have

$$\begin{aligned} \Pr_{X \sim P} \left[\widehat{f}_{S_1}(X) \neq f^*(X) \wedge \widehat{f}_{S_2}(X) \neq f^*(X) \right] &= P(R_1) \Pr_{X \sim P_{R_1}} \left[\widehat{f}_{S_1}(X) \neq f^*(X) \wedge \widehat{f}_{S_2}(X) \neq f^*(X) \right] \\ &\quad + \sum_{i=2}^{\infty} P(R_i) \Pr_{X \sim P_{R_i}} \left[\widehat{f}_{S_1}(X) \neq f^*(X) \wedge \widehat{f}_{S_2}(X) \neq f^*(X) \right]. \end{aligned}$$

We will prove that there is a universal constant $c > 0$ such that the events

$$\begin{aligned} E_1 = E_1((S_1, S_2)) &:= \left\{ P(R_1) \Pr_{X \sim P_{R_1}} \left[\widehat{f}_{S_1}(X) \neq f^*(X) \wedge \widehat{f}_{S_2}(X) \neq f^*(X) \right] \leq \right. \\ &\quad \left. c \max \left\{ \frac{d \log(\log(\min\{n/d, 1/\delta\}))}{n}, \frac{\log(1/\delta)}{n} \right\} \right\}, \end{aligned}$$

⁶This idea used by Simon in fact builds upon even earlier work of Hanneke [37] in the context of active learning. It was also applied in the context of PAC learning by Darnstädt [31].

and

$$E_2 = E_2((S_1, S_2)) := \left\{ \sum_{i=2}^{\infty} P(R_i) \Pr_{X \sim P_{R_i}} \left[\widehat{f}_{S_1}(X) \neq f^*(X) \wedge \widehat{f}_{S_2}(X) \neq f^*(X) \right] \leq c \frac{d}{n} \right\}$$

each happen with probability at least $1 - \delta/2$ over the randomness of (S_1, S_2) . The claim of [Lemma 2.3.1](#) then follows from a union bound. Define the set $I = \{i \geq 2 : P(R_i) \neq 0\}$. To prove that E_1 and E_2 each occur with high probability, we will use the following proposition.

Proposition 2.3.3. *In the setting of [Lemma 2.3.1](#) we have the following:*

1. *There is a universal constant c' such that for any $i \in \mathbb{N}$*

$$P(R_i) \leq \frac{c' 2^i d \log(2^i \log(1/\delta)/\delta) \log(1/\delta)}{\delta n}.$$

2. *With probability at least $1 - \delta/2$ over the randomness of (S_1, S_2) we have, simultaneously for all $i \in I = \{i \geq 2 : P(R_i) \neq 0\}$, that*

$$\Pr_{X \sim P_{R_i}} \left[\widehat{f}_{S_1}(X) \neq f^*(X) \wedge \widehat{f}_{S_2}(X) \neq f^*(X) \right] \leq \frac{5 \cdot 2^{-1.1i} \delta}{\log^2(1/\delta)}.$$

We defer the proof of [Proposition 2.3.3](#) to [Section 2.4](#) as its proof is similar to that of [Lemma 2.1.1](#).

We first prove that the event E_1 occurs with high probability. If $P(R_1) = 0$, then we immediately have that $\mathbf{Pr}_{(S_1, S_2)}[E_1] = 1$. We now consider the case that $P(R_1) \neq 0$. From [Item 1](#) of [Proposition 2.3.3](#) we can conclude there is a universal constant $\tilde{c}$ such that $P(R_1) \leq \min\{1, \tilde{c} \frac{d \log^2(1/\delta)}{\delta n}\}$. Using this combined with [Lemma 2.3.2](#) we have, with probability at least $1 - \delta/2$ over the randomness of (S_1, S_2) , that

$$\begin{aligned} & P(R_1) \Pr_{X \sim P_{R_1}} \left[\widehat{f}_{S_1}(X) \neq f^*(X) \wedge \widehat{f}_{S_2}(X) \neq f^*(X) \right] \\ & \leq 8 \max \left\{ \frac{d \log(8e \log(\min\{en/d, e\tilde{c} \log^2(1/\delta)/\delta\}))}{n}, \frac{\log(16/\delta)}{n} \right\} \\ & \leq c \max \left\{ \frac{d \log(\log(\min\{n/d, 1/\delta\}))}{n}, \frac{\log(1/\delta)}{n} \right\}, \end{aligned}$$

where the last inequality holds for c large enough.

We now prove that the event E_2 occurs with high probability. Combining [Items 1](#) and [2](#) of [Proposition 2.3.3](#) we have, with probability at least $1 - \delta/2$ over the randomness of (S_1, S_2) ,

that

$$\begin{aligned}
 & \sum_{i=2}^{\infty} P(R_i) \Pr_{X \sim P_{R_i}} \left[\widehat{f}_{S_1}(X) \neq f^*(X) \wedge \widehat{f}_{S_2}(X) \neq f^*(X) \right] \\
 & \leq \sum_{i \notin I} 0 + \sum_{i \in I} \frac{2^i c' d \operatorname{Log}(2^i \operatorname{Log}(1/\delta)/\delta) \operatorname{Log}(1/\delta)}{\delta n} \cdot \frac{5 \cdot 2^{-1.1i} \delta}{\operatorname{Log}^2(1/\delta)} \\
 & \leq \frac{5c'd}{n} \sum_{i=2}^{\infty} \frac{2^{-0.1i} \operatorname{Log}(2^i \operatorname{Log}(1/\delta)/\delta)}{\operatorname{Log}(1/\delta)} \\
 & \leq \frac{5c'd}{n} \sum_{i=2}^{\infty} \frac{2^{-0.1i} \cdot (i \operatorname{Log}(2) + \operatorname{Log}(\operatorname{Log}(1/\delta)) + \operatorname{Log}(1/\delta))}{\operatorname{Log}(1/\delta)} \\
 & \leq c \frac{d}{n},
 \end{aligned}$$

where the last inequality holds for c large enough. This concludes the proof. $\square$

We now move on to prove [Lemma 2.3.2](#). To do so, we will need the following lemma which is a simple consequence of uniform convergence. We defer the proof of this lemma to [Section 2.4](#).

Lemma 2.3.4. *Fix a function class $\mathcal{F} \subseteq \{0, 1\}^{\mathcal{X}}$ with VC dimension d . Fix a distribution P over $\mathcal{X}$, target function $f^* \in \mathcal{F}$, and $R \subseteq \mathcal{X}$ such that $P(R) \neq 0$. Fix an ERM algorithm $\widehat{f} : \mathcal{X} \times \mathcal{Z}^* \rightarrow \{0, 1\}$. For any parameter $\delta \in (0, 1/2]$ it holds with probability at least $1 - \delta$ over the randomness of $S \sim P^n$ that*

$$\operatorname{err}_{P_R}(\widehat{f}_S) \leq 8 \max \left\{ \frac{d \operatorname{Log}(eP(R)n/d)}{P(R)n}, \frac{\operatorname{Log}(4/\delta)}{P(R)n} \right\}.$$

We are now ready to prove [Lemma 2.3.2](#).

Proof of Lemma 2.3.2. We will prove that the event

$$\begin{aligned}
 E &= E((S_1, S_2)) \\
 &:= \left\{ \Pr_{X \sim P_R} \left[\widehat{f}_{S_1}(X) \neq f^*(X) \wedge \widehat{f}_{S_2}(X) \neq f^*(X) \right] \leq 8 \max \left\{ \frac{d \operatorname{Log}(8e \operatorname{Log}(eP(R)n/d))}{P(R)n}, \frac{\operatorname{Log}(8/\delta)}{P(R)n} \right\} \right\}
 \end{aligned}$$

occurs with high probability. Let B_1 denote the (random) set $\{x \in \mathcal{X} : \widehat{f}_{S_1}(x) \neq f^*(x)\}$ and define the event $E_1 = E_1(S_1) := \{P(R \cap B_1) \neq 0\}$. By the law of total probability, we have

$$\Pr_{(S_1, S_2) \sim P^{2n}}[E] = \Pr_{(S_1, S_2) \sim P^{2n}}[E_1 \cap E] + \Pr_{(S_1, S_2) \sim P^{2n}}[\bar{E}_1 \cap E]. \quad (2.9)$$

Furthermore, we can write the second term on the right hand side of (2.9) as

$$\Pr_{(S_1, S_2) \sim P^{2n}}[\bar{E}_1 \cap E] = \Pr_{(S_1, S_2) \sim P^{2n}}[E \mid \bar{E}_1] \Pr_{(S_1, S_2) \sim P^{2n}}[\bar{E}_1] = \Pr_{(S_1, S_2) \sim P^{2n}}[\bar{E}_1].$$

Combining the identities above, it suffices to show that

$$\mathbf{Pr}_{(S_1, S_2) \sim P^{2n}} [E \cap E_1] \geq \mathbf{Pr}_{(S_1, S_2) \sim P^{2n}} [E_1] - \delta.$$

Notice that when E_1 occurs, then for any measurable set $C \subseteq \mathcal{X}$, the distribution $(P_R)_{B_1}$ (which is the conditional distribution of P_R restricted to B_1) satisfies

$$(P_R)_{B_1}(C) = \frac{P_R(C \cap B_1)}{P_R(B_1)} = \frac{P(C \cap B_1 \cap R)}{P(B_1 \cap R)} = P_{R \cap B_1}(C),$$

i.e., $(P_R)_{B_1} = P_{R \cap B_1}$. Thus on E_1 , the probability that both $\hat{f}_{S_1}$ and $\hat{f}_{S_2}$ simultaneously err on a new data point drawn from P_R can be written as

$$\Pr_{X \sim P_R} [\hat{f}_{S_1}(X) \neq f^*(X) \wedge \hat{f}_{S_2}(X) \neq f^*(X)] = \text{err}_{P_R}(\hat{f}_{S_1}) \text{err}_{P_{R \cap B_1}}(\hat{f}_{S_2}). \quad (2.10)$$

We now bound the right side of (2.10). To do this, we define the following events over (S_1, S_2) :

$$E_2 = E_2(S_1) := \left\{ \text{err}_{P_R}(\hat{f}_{S_1}) \leq 8 \max \left\{ \frac{d \text{Log}(eP(R)n/d)}{P(R)n}, \frac{\text{Log}(8/\delta)}{P(R)n} \right\} \right\}$$

and for outcomes of S_1 in E_1

$$E_3 = E_3((S_1, S_2)) := \left\{ \text{err}_{P_{R \cap B_1}}(\hat{f}_{S_2}) \leq 8 \max \left\{ \frac{d \text{Log}(eP(R)n \text{err}_{P_R}(\hat{f}_{S_1})/d)}{P(R)n \text{err}_{P_R}(\hat{f}_{S_1})}, \frac{\text{Log}(8/\delta)}{P(R)n \text{err}_{P_R}(\hat{f}_{S_1})} \right\} \right\}.$$

We now show that the event $E_1 \cap E_2 \cap E_3$ happens with probability at least $P[E_1] - \delta$ and that it implies the event $E_1 \cap E$. Assume that $E_1 \cap E_2 \cap E_3$ occurs. If

$$8 \max \left\{ \frac{d \text{Log}(eP(R)n/d)}{P(R)n}, \frac{\text{Log}(8/\delta)}{P(R)n} \right\} = 8 \frac{d \text{Log}(eP(R)n/d)}{P(R)n},$$

we have

$$\begin{aligned} \text{err}_{P_R}(\hat{f}_{S_1}) \text{err}_{P_{R \cap B_1}}(\hat{f}_{S_2}) &\leq 8 \max \left\{ \frac{d \text{Log}(eP(R)n \text{err}_{P_R}(\hat{f}_{S_1})/d)}{P(R)n}, \frac{\text{Log}(8/\delta)}{P(R)n} \right\} \\ &\leq 8 \max \left\{ \frac{d \text{Log}(8e \text{Log}(eP(R)n/d))}{P(R)n}, \frac{\text{Log}(8/\delta)}{P(R)n} \right\}. \end{aligned}$$

Otherwise, if

$$8 \max \left\{ \frac{d \operatorname{Log}(eP(R)n/d)}{P(R)n}, \frac{\operatorname{Log}(8/\delta)}{P(R)n} \right\} = 8 \frac{\operatorname{Log}(8/\delta)}{P(R)n},$$

we have

$$\operatorname{err}_{P_R}(\hat{f}_{S_1}) \operatorname{err}_{P_{R \cap B_1}}(\hat{f}_{S_2}) \leq 8 \frac{\operatorname{Log}(8/\delta)}{P(R)n} \cdot 1.$$

We can thus conclude that $E_1 \cap E_2 \cap E_3$ implies $E_1 \cap E$. Towards showing the bound $\mathbf{Pr}_{(S_1, S_2) \sim P^{2n}}[E_1 \cap E_2 \cap E_3] \geq P[E_1] - \delta$, notice that the bound $\mathbf{Pr}_{(S_1, S_2) \sim P^{2n}}[\bar{E}_2] \leq \delta/2$ can be established from [Lemma 2.3.4](#) directly. Furthermore, for *any fixed realization* of S_1 such that $P(R \cap B_1) \neq 0$, [Lemma 2.3.4](#) implies that

$$\operatorname{err}_{P_{R \cap B_1}}(\hat{f}_{S_2}) \leq 8 \max \left\{ \frac{d \operatorname{Log}(eP(R) \operatorname{err}_{P_R}(\hat{f}_{S_1})n/d)}{P(R) \operatorname{err}_{P_R}(\hat{f}_{S_1})n}, \frac{\operatorname{Log}(8/\delta)}{P(R) \operatorname{err}_{P_R}(\hat{f}_{S_1})n} \right\},$$

with probability at least $1 - \delta/2$ over the randomness of S_2 . Using the independence of S_1 and S_2 we have

$$\begin{aligned} \mathbf{Pr}_{(S_1, S_2) \sim P^{2n}}[E_1 \cap E_2 \cap E_3] &= \mathbb{E}_{S_1 \sim P^n} \left[\mathbf{1}_{E_1} \mathbf{1}_{E_2} \mathbf{Pr}_{S_2 \sim P^n}[E_3] \right] \\ &\geq \mathbb{E}_{S_1 \sim P^n} [\mathbf{1}_{E_1} \mathbf{1}_{E_2}] (1 - \delta/2) \\ &\geq (1 - \mathbf{Pr}_{S_1 \sim P^n}[\bar{E}_1] - \mathbf{Pr}_{S_1 \sim P^n}[\bar{E}_2]) (1 - \delta/2) \\ &\geq (\mathbf{Pr}_{S_1 \sim P^n}[E_1] - \delta/2) (1 - \delta/2) \\ &\geq \mathbf{Pr}_{S_1 \sim P^n}[E_1] - \delta. \end{aligned}$$

This completes the proof. □

2.4 Omitted proofs from [Sections 2.1](#) and [2.3](#)

In this appendix we prove [Lemma 2.1.2](#), [Proposition 2.3.3](#), and [Lemma 2.3.4](#). These results are by-products of the classic *uniform convergence* result which uniformly bounds the error of any function in $\mathcal{F}$ that is consistent with the training sample. To state the result, we first introduce some notation. For a training sample $S = ((X_1, f^*(X_1)), \dots, (X_n, f^*(X_n)))$, let $\mathcal{F}_S$ denote the functions in $\mathcal{F}$ that are consistent with S , i.e., $f \in \mathcal{F}_S$ if and only if $f(X_i) = f^*(X_i)$ for every $i \in [n]$.

Lemma 2.4.1 (Uniform convergence [21]). *Fix a function class $\mathcal{F} \subseteq \{0, 1\}^{\mathcal{X}}$ with VC dimension d . Fix a distribution P over $\mathcal{X}$ and target function $f^* \in \mathcal{F}$. For any parameter $\delta \in (0, 1/2]$ it holds with probability at least $1 - \delta$ over $S \sim P^n$ that*

$$\sup_{f \in \mathcal{F}_S} \text{err}_P(f) \leq 2 \left(\frac{d \log(2en/d) + \log(2/\delta)}{n} \right).$$

In what follows, we will use the slightly weaker bound

$$\sup_{f \in \mathcal{F}_S} \text{err}_P(f) \leq 4 \max \left\{ \frac{d \text{Log}(2en/d)}{n}, \frac{\text{Log}(2/\delta)}{n} \right\}. \quad (2.11)$$

Proof of Lemma 2.1.2

We now prove Lemma 2.1.2 which we restate here for convenience.

Lemma 2.1.2. *Fix a function class $\mathcal{F} \subseteq \{0, 1\}^{\mathcal{X}}$ with VC dimension d . Fix a distribution P over $\mathcal{X}$, target function $f^* \in \mathcal{F}$, and $R \subseteq \mathcal{X}$ such that $\Pr_{X \sim P}[X \in R] \neq 0$. For any ERM algorithm $\hat{f} : \mathcal{X} \times \mathcal{Z}^* \rightarrow \{0, 1\}$ it follows that*

$$\mathbb{E}_{S \sim P^n} \left[\text{err}_{P_R}(\hat{f}_S) \right] \leq 20 \frac{d \text{Log}(e \Pr_{X \sim P}[X \in R]n/d)}{\Pr_{X \sim P}[X \in R]n}.$$

Proof. Consider the case that $\Pr_{X \sim P}[X \in R]n \leq 4d$. In this case the claim follows easily since $\text{err}_{P_R}(\hat{f}_S) \leq 1$ and $x \rightarrow \text{Log}(ex)/x$ is decreasing in x for $x > 0$, so $20 \frac{d \text{Log}(e \Pr_{X \sim P}[X \in R]n/d)}{\Pr_{X \sim P}[X \in R]n} > 1$. We now consider the case that $\Pr_{X \sim P}[X \in R]n > 4d$. For any $m \in \mathbb{N}$ we define the event $E_m = E_m(S) = \{|\{i \in [n] : X_i \in R\}| = m\}$. Similarly, we define the event

$$E = E(S) = \bigcup_{m \geq \Pr_{X \sim P}[X \in R]n/2} E_m.$$

It follows from a Chernoff bound and our assumption that $\Pr_{X \sim P}[X \in R]n > 4d$ that

$$\Pr_{S \sim P^n}[E] \geq 1 - \exp \left(- \frac{\Pr_{X \sim P}[X \in R]n}{8} \right) \geq 1 - \frac{8}{\Pr_{X \sim P}[X \in R]n}.$$

Using the law of total probability we have

$$\mathbb{E}_{S \sim P^n} \left[\text{err}_{P_R}(\hat{f}_S) \right] \leq \mathbb{E}_{S \sim P^n} \left[\text{err}_{P_R}(\hat{f}_S) \mid E \right] + \frac{8}{\Pr_{X \sim P}[X \in R]n}. \quad (2.12)$$

So, if we show that for any $m \geq \Pr_{X \sim P}[X \in R]n/2$ that

$$\mathbb{E}_{S \sim P^n} \left[\text{err}_{P_R}(\hat{f}_S) \mid E_m \right] \leq \frac{12d \text{Log}(e \Pr_{X \sim P}[X \in R]n/d)}{\Pr_{X \sim P}[X \in R]n}, \quad (2.13)$$

the claim follows from one more application of the law of total probability applied to the first term on the right hand side of (2.12).

We now prove (2.13). Using the non-negativity of $\text{err}_{P_R}(\hat{f}_S)$ we have

$$\begin{aligned} \mathbb{E}_{S \sim P^n} \left[\text{err}_{P_R}(\hat{f}_S) \mid E_m \right] &= \int_0^\infty \mathbf{Pr}_{S \sim P^n} \left[\text{err}_{P_R}(\hat{f}_S) > x \mid E_m \right] dx \\ &\leq \frac{4d \text{Log}(2em/d)}{m} + \int_{\frac{4d \text{Log}(2em/d)}{m}}^1 \mathbf{Pr}_{S \sim P^n} \left[\text{err}_{P_R}(\hat{f}_S) > x \mid E_m \right] dx. \end{aligned} \quad (2.14)$$

Notice that conditioned on E_m , the m samples that land in R form an i.i.d. sample from the conditional distribution P_R . Thus, any ERM trained on S is also consistent with m i.i.d. samples from P_R , so we can apply uniform convergence (Lemma 2.4.1) to control the error of any ERM when measured with respect to the conditional distribution P_R . Setting $\delta = 2^{1-\frac{mx}{4}}$ we have

$$\begin{aligned} \int_{\frac{4d \text{Log}(2em/d)}{m}}^1 \mathbf{Pr}_{S \sim P^n} \left[\text{err}_{P_R}(\hat{f}_S) > x \mid E_m \right] dx &= \int_{\frac{4d \text{Log}(2em/d)}{m}}^1 \mathbf{Pr}_{S \sim P^n} \left[\text{err}_{P_R}(\hat{f}_S) > \frac{4 \text{Log}(2/\delta)}{m} \mid E_m \right] dx \\ &\leq 2 \int_{\frac{4d \text{Log}(2em/d)}{m}}^1 2^{-\frac{mx}{4}} dx \\ &\leq 2 \left(\frac{4 \cdot 2^{-d \text{Log}(2em/d)}}{m \ln(2)} \right) \leq \frac{2d \text{Log}(2em/d)}{m}. \end{aligned}$$

Here, the first equality follows from the fact that $m \geq \mathbf{Pr}_{X \sim P}[X \in R]n/2 \geq 2d$ and our choice of δ . The second inequality follows from (2.11) and the final inequality follows from the fact that $d \text{Log}(2em/d) \geq 2$. Now, using the fact that $x \rightarrow \text{Log}(2ex)/x$ is decreasing for $x > 0$ together with the fact that $m \geq \mathbf{Pr}_{X \sim P}[X \in R]n/2 \geq 2d$, we conclude

$$\mathbb{E}_{S \sim P^n} \left[\text{err}_{P_R}(\hat{f}_S) \mid E_m \right] \leq \frac{6d \text{Log}(2em/d)}{m} < \frac{12d \text{Log}(e \mathbf{Pr}_{X \sim P}[X \in R]n/d)}{\mathbf{Pr}_{X \sim P}[X \in R]n},$$

as claimed. □

Proof of Proposition 2.3.3

We now prove Proposition 2.3.3 which we restate here for convenience.

Proposition 2.3.3. *In the setting of Lemma 2.3.1 we have the following:*

1. *There is a universal constant c' such that for any $i \in \mathbb{N}$*

$$P(R_i) \leq \frac{c' 2^i d \text{Log}(2^i \text{Log}(1/\delta)/\delta) \text{Log}(1/\delta)}{\delta n}.$$

2. With probability at least $1 - \delta/2$ over the randomness of (S_1, S_2) we have, simultaneously for all $i \in I = \{i \geq 2 : P(R_i) \neq 0\}$, that

$$\Pr_{X \sim P_{R_i}} \left[\widehat{f}_{S_1}(X) \neq f^*(X) \wedge \widehat{f}_{S_2}(X) \neq f^*(X) \right] \leq \frac{5 \cdot 2^{-1.1i} \delta}{\text{Log}^2(1/\delta)}.$$

Proof. We first prove [Item 1](#). Towards a contradiction, assume that there is an $i \in \mathbb{N}$ such that

$$\frac{P(R_i) n}{d} \geq \frac{c' 2^i \text{Log}(2^i \text{Log}(1/\delta)/\delta) \text{Log}(1/\delta)}{\delta}$$

for a constant c' that we will choose below. By changing the order of expectations we have

$$\mathbb{E}_{S_1 \sim P^n} \left[\text{err}_{P_{R_i}}(\widehat{f}_{S_1}) \right] = \mathbb{E}_{X \sim P_{R_i}} \left[\mathbf{Pr}_{S_1 \sim P^n} \left[\widehat{f}_{S_1}(X) \neq f^*(X) \right] \right] = \mathbb{E}_{X \sim P_{R_i}} [p_X].$$

Using the above together with the fact that $p_X \geq 2^{-i} \delta / \text{Log}(1/\delta)$ for any $X \in R_i$, we can conclude that

$$\mathbb{E}_{S_1 \sim P^n} \left[\text{err}_{P_{R_i}}(\widehat{f}_{S_1}) \right] > \frac{\delta}{2^i \text{Log}(1/\delta)}. \quad (2.15)$$

On the other hand, using [Lemma 2.1.2](#) we conclude that there is a universal constant $\hat{c}$ such that

$$\mathbb{E}_{S_1 \sim P^n} \left[\text{err}_{P_{R_i}}(\widehat{f}_{S_1}) \right] \leq \hat{c} \frac{d \text{Log}(P(R_i) n / d)}{P(R_i) n}.$$

Combining this inequality with the fact that $\text{Log}(x)/x$ is a decreasing function for $x > 0$, we have

$$\begin{aligned} \mathbb{E}_{S_1 \sim P^n} \left[\text{err}_{P_{R_i}}(\widehat{f}_{S_1}) \right] &\leq \hat{c} \frac{\delta \text{Log} \left(\frac{c' 2^i \text{Log}(2^i \text{Log}(1/\delta)/\delta) \text{Log}(1/\delta)}{\delta} \right)}{c' 2^i \text{Log}(2^i \text{Log}(1/\delta)/\delta) \text{Log}(1/\delta)} \\ &= \hat{c} \frac{2^{-i} \delta}{\text{Log}(1/\delta)} \cdot \frac{\text{Log} \left(\frac{c' 2^i \text{Log}(2^i \text{Log}(1/\delta)/\delta) \text{Log}(1/\delta)}{\delta} \right)}{c' \text{Log} \left(\frac{2^i \text{Log}(1/\delta)}{\delta} \right)} \\ &\leq \hat{c} \frac{2^{-i} \delta}{\text{Log}(1/\delta)} \cdot \frac{\text{Log}(c') + \text{Log} \left(\text{Log} \left(\frac{2^i \text{Log}(1/\delta)}{\delta} \right) \right) + \text{Log} \left(2^i \frac{\text{Log}(1/\delta)}{\delta} \right)}{c' \text{Log} \left(\frac{2^i \text{Log}(1/\delta)}{\delta} \right)}. \end{aligned}$$

However, for c' large enough, the above is less than $2^{-i} \delta / \text{Log}(1/\delta)$ which contradicts the lower bound (2.15). Thus, we have shown that there is a constant c' such that

$$\frac{P(R_i) n}{d} \leq \frac{c' 2^i \text{Log}(2^i \text{Log}(1/\delta)/\delta) \text{Log}(1/\delta)}{\delta},$$

which proves [Item 1](#).

We now prove [Item 2](#). We will show that for each $i \in I$ with probability at least $1 - 2^{-0.9i+2}\delta/5$ we have

$$\Pr_{X \sim P_{R_i}} \left[\widehat{f}_{S_1}(X) \neq f^*(X) \wedge \widehat{f}_{S_2}(X) \neq f^*(X) \right] \leq \frac{5 \cdot 2^{-1.1i}\delta}{\text{Log}^2(1/\delta)}. \quad (2.16)$$

Applying a union bound implies that the above holds simultaneously for every $i \in I$ with probability at least $1 - \sum_{i \geq 2} 2^{-0.9i+2}\delta/5 \geq 1 - \delta/2$. To see that (2.16) holds for each $i \in I$, notice that we can use the fact that S_1 and S_2 are i.i.d. samples together with the fact that $p_X \leq 2^{-i+1}\delta/\text{Log}(1/\delta)$ for $X \in R_i$ to conclude that

$$\mathbb{E}_{X \sim P_{R_i}} \left[\Pr_{S_1, S_2 \sim P^n} \left[\widehat{f}_{S_1}(X) \neq f^*(X) \wedge \widehat{f}_{S_2}(X) \neq f^*(X) \right] \right] = \mathbb{E}_{X \sim P_{R_i}} p_X^2 \leq \frac{2^{-2i+2}\delta^2}{\text{Log}^2(1/\delta)}.$$

Combining this with Markov's inequality we have

$$\begin{aligned} & \Pr_{S_1, S_2 \sim P^n} \left[\Pr_{X \sim P_{R_i}} \left[\widehat{f}_{S_1}(X) \neq f^*(X) \wedge \widehat{f}_{S_2}(X) \neq f^*(X) \right] > \frac{5 \cdot 2^{-1.1i}\delta}{\text{Log}^2(1/\delta)} \right] \\ & \leq \frac{2^{-2i+2}\delta^2}{\text{Log}^2(1/\delta)} \frac{\text{Log}^2(1/\delta)}{5 \cdot 2^{-1.1i}\delta} = 2^{-0.9i+2}\delta/5, \end{aligned}$$

which proves the claim. $\square$

Proof of [Lemma 2.3.4](#)

We now prove [Lemma 2.3.4](#) which we restate here for convenience.

Lemma 2.3.4. *Fix a function class $\mathcal{F} \subseteq \{0, 1\}^{\mathcal{X}}$ with VC dimension d . Fix a distribution P over $\mathcal{X}$, target function $f^* \in \mathcal{F}$, and $R \subseteq \mathcal{X}$ such that $P(R) \neq 0$. Fix an ERM algorithm $\widehat{f} : \mathcal{X} \times \mathcal{Z}^* \rightarrow \{0, 1\}$. For any parameter $\delta \in (0, 1/2]$ it holds with probability at least $1 - \delta$ over the randomness of $S \sim P^n$ that*

$$\text{err}_{P_R}(\widehat{f}_S) \leq 8 \max \left\{ \frac{d \text{Log}(eP(R)n/d)}{P(R)n}, \frac{\text{Log}(4/\delta)}{P(R)n} \right\}.$$

Proof. If $8 \text{Log}(4/\delta)/(P(R)n) \geq 1$ we are done as $\text{err}_{P_R}(f) \leq 1$. Thus, for the remainder of the proof we will assume that $8 \text{Log}(4/\delta)/(P(R)n) < 1$, which is equivalent to $P(R)n \geq 8 \text{Log}(4/\delta)$. Define the event

$$E = E(S) = \{ |\{i \in [n] : X_i \in R\}| \geq P(R)n/2 \}.$$

Using the Chernoff bound and our assumption that $P(R)n \geq 8 \text{Log}(4/\delta)$, we have

$$\Pr_{S \sim P^n} [E] \geq 1 - \exp \left(-\frac{P(R)n}{8} \right) \geq 1 - \delta/2.$$

Notice that conditioned on E , the $M \geq P(R)n/2 \geq 1$ samples⁷ that land in R form an i.i.d. sample from the conditional distribution P_R . Thus when E occurs, any ERM trained on S is also consistent with $M \geq P(R)n/2 \geq 1$ i.i.d. samples from P_R , so [Lemma 2.4.1](#) yields, with probability at least $1 - \delta/2$ over the randomness of S , that

$$\begin{aligned} \text{err}_{P_R}(\hat{f}_S) &\leq 4 \max \left\{ \frac{d \log(2eM/d)}{M}, \frac{\log(4/\delta)}{M} \right\} \\ &\leq 8 \max \left\{ \frac{d \log(eP(R)n/d)}{P(R)n}, \frac{\log(4/\delta)}{P(R)n} \right\}. \end{aligned}$$

Here, the second inequality follows from the fact that $x \rightarrow \log(2ex)/x$ is decreasing for $x > 0$. Using the law of total probability we get that

$$\begin{aligned} &\mathbf{Pr}_{S \sim P^n} \left[\text{err}_{P_R}(\hat{f}_S) > 8 \max \left\{ \frac{d \log(eP(R)n/d)}{P(R)n}, \frac{\log(4/\delta)}{P(R)n} \right\} \right] \\ &\leq \mathbf{Pr}_{S \sim P^n} [\bar{E}] + \mathbf{Pr}_{S \sim P^n} \left[\text{err}_{P_R}(\hat{f}_S) > 8 \max \left\{ \frac{d \log(eP(R)n/d)}{P(R)n}, \frac{\log(4/\delta)}{P(R)n} \right\} \mid E \right] \\ &\leq \delta/2 + \delta/2 = \delta. \end{aligned}$$

This concludes the proof. $\square$

2.5 Ommited proofs from [Section 2.2](#)

In this appendix we prove [\(2.8\)](#). We will show that

$$\mathbf{Pr}_{Y \sim Q} [Y \geq m] = \mathbf{Pr}_{Y_t \sim Q_t} \left[\sum_{t=0}^{m_2 d - d - 1} Y_t \geq m \right] \geq \sqrt{\frac{2}{3}}.$$

Let $p^* = p_{m_2 d - d - 1} = \frac{d+1}{m_1 m_2}$ be the smallest parameter p_t of the geometric random variables $\{Y_t\}_{t=0}^{m_2 d - d - 1}$ that we consider. We make use of the following well known concentration inequality for sums of geometric random variables:

$$\mathbf{Pr}_{Y \sim Q} \left[Y \leq \lambda \mathbb{E}_{Y \sim Q} [Y] \right] \leq \exp \left(-p^* \mathbb{E}_{Y \sim Q} [Y] (\lambda - 1 - \ln(\lambda)) \right), \quad (2.17)$$

which holds for any $0 < \lambda \leq 1$ (see [\[46, Theorem 3.1\]](#)). Let $\lambda = 1/4$. We will show $\mathbb{E}_{Y \sim Q} [Y] \geq 4m$ and $p^* \mathbb{E}_{Y \sim Q} [Y] \geq 4$ which combined with $1/4 - 1 - \ln(1/4) \geq 1/2$ and [\(2.17\)](#) gives us

$$\mathbf{Pr}_{Y \sim Q} [Y \leq m] \leq \mathbf{Pr}_{Y \sim Q} \left[Y \leq \mathbb{E}_{Y \sim Q} [Y] / 4 \right] \leq \exp \left(-p^* \mathbb{E}_{Y \sim Q} [Y] / 2 \right) \leq \exp(-2) \leq 1 - \sqrt{2/3}.$$

⁷The number of samples M is random.

This implies $\Pr_{Y \sim Q}[Y \geq m] \geq \sqrt{2/3}$ as required. We first show that $\mathbb{E}_{Y \sim Q}[Y] \geq 4m$. We have

$$\mathbb{E}_{Y \sim Q}[Y] = \sum_{t=0}^{m_2 d - d - 1} \frac{m_1 m_2}{m_2 d - t} = \sum_{i=d+1}^{m_2 d} \frac{m_1 m_2}{i} \geq m_1 m_2 \ln \left(\frac{m_2}{2} \right). \quad (2.18)$$

Plugging in the definition of m_1 and m_2 into (2.18) and using the fact that $\lceil x \rceil \leq 2x$ for $x \geq 0.5$ gives us

$$\mathbb{E}_{Y \sim Q}[Y] \geq \frac{8Cm}{\ln(\ln(2Cm/d))} \ln \left(\frac{\ln(2Cm/d)}{\ln(\ln(2Cm/d))} \right).$$

For C large enough we have $\ln(\ln(2Cm/d)) > 0$, $\frac{\ln(2Cm/d)}{\ln(\ln(2Cm/d))} \geq \sqrt{\ln(2Cm/d)}$ and $C > 1$, so

$$\mathbb{E}_{Y \sim Q}[Y] \geq \frac{8Cm}{\ln(\ln(2Cm/d))} \ln \left(\sqrt{\ln(2Cm/d)} \right) \geq 4m.$$

We now show that $p^* \mathbb{E}_{Y \sim Q}[Y] \geq 4$. Using the fact that $p^* \geq 2/(m_1 m_2)$ together with (2.18) gives us

$$p^* \mathbb{E}_{Y \sim Q}[Y] \geq 2 \ln \left(\frac{m_2}{2} \right) \geq 2 \ln \left(\frac{\ln(2Cm/d)}{\ln(\ln(2Cm/d))} \right) \geq 4,$$

where the last inequality holds for C large enough.

Chapter 3

Optimal PAC Bounds without Uniform Convergence

This chapter is based on joint work with Yeshwanth Cherapanamjeri, Abhishek Shetty, and Nikita Zhivotovskiy [6, 4].

The study of the statistical complexity of prediction is a central question in statistical learning theory. In the simplest setting of realizable prediction, the predication task is as follows: for an instance space $\mathcal{X}$, label space $\mathcal{Y}$ and hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, one is given access to n i.i.d. labelled training points $S = ((X_1, Y_1), \dots, (X_n, Y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$ drawn from a distribution P with the promise that there exists a target hypothesis $f^* \in \mathcal{H}$ that perfectly labels the data. We refer to S as the *training sample*. The goal is to use the training sample S to design a predictor, denoted by $\hat{f}(\cdot; S)$, with small prediction error on a held-out sample drawn from the same distribution (which we refer to as the *risk*) with high probability over the samples. Concretely, we would like to minimize

$$\mathbb{E}_{(X,Y) \sim P} \left[\ell(\hat{f}(X; S), Y) \right]$$

for a suitable loss function $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow [0, 1]$ with probability at least $1 - \delta$ over S .

A vast body of literature is dedicated to characterizing the optimal achievable risk as a function of the number data points n , the failure probability δ , and the complexity of hypothesis class $\mathcal{H}$. Even for the simplest setting of binary classification, where $\mathcal{Y} = \{0, 1\}$, $\ell(\hat{y}, y) = \mathbf{1}\{\hat{y} \neq y\}$ and the complexity of the function class is captured by its Vapnik-Chervonenkis (VC) dimension, this remained a challenging open problem. The VC-dimension of a class $\mathcal{H}$ is the largest integer d such that there exists a set of d points that can be labeled in all 2^d possible ways using functions from $\mathcal{H}$. In this specific setting, seminal early works [85, 87, 21] established that an Empirical Risk Minimizer (ERM), i.e.,

any predictor that minimizes the loss on the training samples, achieves the risk bound¹

$$\mathbb{E}_{(X,Y) \sim P} \left[\ell(\hat{f}_{\text{ERM}}(X; S), Y) \right] \leq C \left(\frac{d}{n} \log \left(\frac{n}{d} \right) + \frac{1}{n} \log \left(\frac{1}{\delta} \right) \right).$$

High probability risk bounds of this form are usually referred to as Probably Approximately Correct (PAC) bounds, a notion introduced in the classical paper of Valiant [81].

Meanwhile, a complementary line of work in the *transductive* setting and the related *prediction model of learning* [42], led to the development of predictors with strong leave-one-out (LOO) performance with direct implications for the distributional setting. In a prototypical example of these settings, there is a fixed realizable sample $S = ((x_i, y_i))_{i=1}^n$ and the learner is given access to the labels y_i of a randomly chosen $n - 1$ of them with the task of predicting accurately on the remaining point. Denoting by $S^{-i} = ((x_j, y_j))_{j \neq i}$, the goal is to design a predictor $\hat{f}$ which minimizes the LOO error

$$\sum_{i=1}^n \ell \left(\hat{f}(x_i; S^{-i}), y_i \right).$$

For binary classification, the famous one-inclusion graph strategy of Haussler, Littlestone, and Warmuth [41, 42] achieves LOO error at most d .² An exchangeability argument yields a predictor $\hat{f}_{\text{OIG}}$ with the bound

$$\mathbb{E}_{\substack{(X,Y) \sim P \\ S \sim P^n}} \left[\ell \left(\hat{f}_{\text{OIG}}(X; S), Y \right) \right] \leq \frac{d}{n+1}$$

for the distributional setting *in-expectation* over the training samples. Interestingly, this approach eliminates the extra logarithmic factor typically present in bounds for empirical risk minimization, leading to the conjecture that the logarithmic factor might not be necessary, even in the PAC setting. Further, it suggests that the one-inclusion graph predictor could be optimal [90]. Since then, these predictors have been adapted for a variety of contexts where uniform convergence, a crucial aspect needed for analyzing ERM predictors, is suboptimal or fails to hold. Examples include multiclass learning [74, 78, 29], partial hypothesis learning [10], and bounded regression [18].

In a breakthrough result, Hanneke [36], building upon the work of Simon [77], devised an optimal PAC predictor $\hat{f}_{\text{OPT}}$ satisfying

$$\mathbb{E}_{(X,Y) \sim P} \left[\ell(\hat{f}_{\text{OPT}}(X; S), Y) \right] \leq C \left(\frac{d}{n} + \frac{1}{n} \log \left(\frac{1}{\delta} \right) \right),$$

affirming the conjectured optimal rate. Remarkably, the analysis does not leverage the optimal in-expectation performance of the leave-one-out predictors but instead considers a

¹Throughout the paper we use $\log$ to denote the natural logarithm.

²Their bound is optimal even up to the leading constant [54].

careful aggregation of suboptimal ERM predictors. Unfortunately, the use of ERM precludes the application of this approach in settings where uniform convergence does not hold, but accurate prediction is nevertheless possible as evidenced by the success of leave-one-out predictors.

The key technical question tackled by our work is the extension of optimal bounds obtained by Hanneke [36] to the other learning settings just described. As previously discussed, the resolution of Warmuth’s conjecture offers a natural path toward such guarantees. Indeed, it would suggest that, at least in some cases, an optimal in-expectation guarantee could immediately yield an optimal high-probability guarantee. Unfortunately, we show that this is not the case by proving a series of results demonstrating that the one-inclusion graph algorithm does not generally lead to optimal PAC bounds with existing methods for orienting the one-inclusion graph. See [Section 3.4](#) for the technical details.

These results prompt the following natural question:

Can optimal in-expectation guarantees *ever* be extended to corresponding optimal PAC guarantees without uniform convergence?

Our main contribution is a general technique that transforms a broad class of predictors with optimal leave-one-out performance (such as the one-inclusion graph strategy), into ones with optimal PAC bounds, thus, affirmatively answering the above question. Remarkably, while a single such predictor may not yield optimal guarantees (as is the case with the one-inclusion graph strategy), we show that a simple aggregation of them suffices to achieve optimality.

More concretely, our reliance on the existence of optimal leave-one-out predictors allows us to completely bypass uniform convergence, enabling applications to a range of settings where uniform convergence may not hold. Consider the general realizable prediction setting with a predictor, $\hat{f}_{\text{LOO}}$, satisfying

$$\sum_{i=1}^n \ell \left(\hat{f}_{\text{LOO}}(x_i; S^{-i}), y_i \right) \leq M_n$$

for any realizable sample S . Our framework now yields a predictor, $\hat{f}_{\text{PAC}}$ satisfying

$$\mathbb{E}_{(X,Y) \sim P} \left[\ell \left(\hat{f}_{\text{PAC}}(X; S), Y \right) \right] \leq C \left(\frac{M_n}{n} + \frac{1}{n} \log \left(\frac{1}{\delta} \right) \right).$$

for the distributional setting.

In several cases of interest, there exist leave-one-out predictors which achieve optimal in-expectation performance for the first term. Our results show that these may be extended to the PAC setup paying only *additively* in $\log(1/\delta)$ which is also known to be optimal. We now illustrate some applications of our result.

1. **Multiclass Classification:** There exist leave-one-out predictors with error scaling with the graph density, $\mu_{\mathcal{H}}(n)$, of the associated one-inclusion hypergraph leading to

optimal in-expectation performance. Our results now immediately yield an optimal PAC bound in this setting ([Theorem 3.1.7](#)). Moreover, in the context of binary classification, we recover the optimal PAC bound of Hanneke [\[36\]](#) with a simplified analysis, improving the constant factors by several orders of magnitude.

2. **Learning Partial Hypotheses:** Here, the optimal bound on M_n scales with the VC-dimension of the *partial* hypothesis class, a generalization of the VC-dimension in the binary setting. Our result yields an optimal PAC bound answering an open problem in the literature [\[10\]](#). See [Corollary 3.1.13](#).
3. **Bounded Regression:** In the setting of bounded ℓ_1 regression, we obtain the first predictor in the PAC setting whose error depends on the scale-sensitive complexity measure $V_{\mathcal{H}}(\gamma)$ which in some cases represents a significant improvement over those obtained by the classic uniform convergence-based arguments which depend instead on the *fat-shattering* dimension $P_{\mathcal{H}}(\gamma)$. See [Section 3.1](#) and [Theorem 3.1.15](#).

The remainder of the paper is organized as follows. The main results of our paper, including a formal description of our framework and its application to several settings of interest are discussed in [Section 3.1](#). We present related work in [Section 3.2](#). Subsequently, we prove our main result in [Section 3.3](#) and our lower bound in [Section 3.4](#).

3.1 Main results

Notation

We use $\mathcal{X}$ to denote our instance space, $\mathcal{Y}$ to denote the label space, and sometimes use $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Let $\mathcal{U} = \bigcup_{n=1}^{\infty} \mathcal{Z}^n$ denote the set of possible observable training samples. A predictor $\hat{f}: \mathcal{X} \times \mathcal{U} \rightarrow \mathcal{Y}$ takes as input a test point and training sample and outputs a prediction. For a training sample $S = ((X_1, Y_1), \dots, (X_n, Y_n))$, we will use S^{-i} to denote the training sample with the i^{th} data point removed, i.e., $S^{-i} = ((X_1, Y_1), \dots, (X_{i-1}, Y_{i-1}), (X_{i+1}, Y_{i+1}), \dots, (X_n, Y_n))$ for all $i \in [n]$. For $j \in [n]$, we will use $S_{\leq j}$ to denote the training sample $((X_1, Y_1), \dots, (X_j, Y_j))$. We will also find it convenient at times to write $Z_i = (X_i, Y_i)$ for the i^{th} point in a training sample S . Throughout the paper, we will use the upper case X, Y and Z to denote random variables and lower case x, y and z to denote the realizations.

In the learning setting, the algorithm is given access to samples from an unknown distribution P over $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. The goal of the learning algorithm is to produce a hypothesis $\hat{f}$ that has low *risk* under P with respect to a *loss function* $\ell: \mathcal{Y} \times \mathcal{Y} \rightarrow [0, 1]$, where we define the risk to be $R_P(\hat{f}) := \mathbb{E}_{(X,Y) \sim P}[\ell(\hat{f}(X), Y)]$. Throughout this paper we will make the assumption that the unknown distribution P is *realizable* by $\mathcal{H}$. This means there is some $f^* \in \mathcal{H}$ such that $Y = f^*(X)$ almost surely. Similarly, we say a training sample $S = ((x_1, y_1), \dots, (x_n, y_n))$ is *realizable* by $\mathcal{H}$ (or simply *realizable*) if there is some $f^* \in \mathcal{H}$ such that $f^*(x_i) = y_i$ for all $(x_i, y_i) \in S$. Given a training sample $S = ((x_1, y_1), \dots, (x_n, y_n))$,

we will find it convenient to define $\mathcal{X}_S \subseteq \mathcal{X}$ to be the *set* of *unique* x_i terms in the training sample S .

Assumptions

Our main result is applicable to a broad class of predictors satisfying two natural assumptions. The main assumption pertains to the *leave-one-out* error. As described in [Chapter 3](#), the predictor is first evaluated on each point in the training sample with the predictor trained on the training sample with that point removed. The sum of these errors over all points in the training sample yields the leave-one-out error. We assume that the predictor has bounded leave-one-out error.

Assumption 3.1.1. The predictor $\hat{f}$ has bounded leave-one-out error for any realizable sample. Formally, for any realizable $S = ((x_i, y_i))_{i=1}^n$, we have

$$\sum_{i=1}^n \ell(\hat{f}(x_i; S^{-i}), y_i) \leq M_n.$$

Further, we will assume that M_n is monotone in the sample size n , that is, $M_n \leq M_{n+1}$ for all $n \in \mathbb{N}$.³

Remark 3.1.2. Note that by averaging over a sample S of size n drawn from P , the leave-one-out error from [Assumption 3.1.1](#) translates to an expected risk bound of

$$\mathbb{E}_{S \sim P^n} R_P(\hat{f}(\cdot; S)) \leq \frac{M_{n+1}}{n+1}. \quad (3.1)$$

The second assumption necessitates that our classifier is symmetric with respect to the training data. This is intuitive in the batch setting where the training data is drawn i.i.d. from a distribution and is satisfied by numerous algorithms in both theoretical and practical contexts. We formally state it as the following assumption.

Assumption 3.1.3. The predictor $\hat{f}$ is symmetric in the training data: for all $n \in \mathbb{N}$, for any permutation π , $S \in \mathcal{Z}^n$, $x \in \mathcal{X}$, we have

$$\hat{f}(x; S) = \hat{f}(x; S_\pi),$$

where S_π is the sample S permuted by π .

The main motivation to consider these assumptions is the *one-inclusion graph algorithm* which is symmetric and, as previously discussed, has optimal leave-one-out error in various settings. For example, in the case of realizable binary classification with hypothesis classes with finite VC-dimension, this algorithm has M_n that is independent of the sample size. We will discuss our applications subsequently and the one-inclusion graph algorithm in detail in [Section 3.5](#).

³This is largely without loss of generality since in most natural settings, we can replace M_n with $\max_{i \leq n} M_i$ to enforce monotonicity.

Main results

To provide context for our main result, we first highlight key findings that will appear later in the paper. Our results establish lower bounds showing that existing orientation strategies for the one-inclusion graph do not yield optimal PAC bounds. More broadly, our findings contribute another example to the literature illustrating that optimal in-expectation guarantees do not always directly translate to optimal PAC guarantees. We defer a formal statement and the relevant details regarding our lower bound to [Section 3.4](#).

We now turn to our main result, which addresses these shortcomings and examines the performance of a sequence of hypotheses generated by training a predictor, $\hat{f}$, on prefixes of the training data, $(S_{\leq t})_{t=1}^n$. When $\hat{f}$ is symmetric and has bounded leave-one-out error, our result asserts that the *average* risk of the predictors in the *suffix* of the sequence $(\hat{f}(\cdot; S_{\leq t}))_{t=1}^n$ is bounded by M_n , with high probability. For the sake of simplicity, we assume for the rest of the paper that $n/4$ is an integer.

Theorem 3.1.4. *Fix a hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, loss function $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow [0, 1]$, and a predictor $\hat{f} : \mathcal{X} \times \mathcal{U} \rightarrow \mathcal{Y}$ satisfying [Assumptions 3.1.1](#) and [3.1.3](#). Then, for any realizable distribution P over $\mathcal{X} \times \mathcal{Y}$ and confidence parameter $\delta \in (0, 1)$, given a training sample $S \sim P^n$ we have*

$$\frac{4}{3n} \cdot \sum_{t=n/4}^{n-1} R_P(\hat{f}(\cdot; S_{\leq t})) \leq 4.82 \left(\frac{M_n}{n} + \frac{1}{n} \log \left(\frac{2}{\delta} \right) \right),$$

with probability at least $1 - \delta$ over the randomness of S .

Our proof technique is based on *online-to-batch conversion*, which transforms any online learning algorithm — making M mistakes on a finite sample — into a batch algorithm utilizing the entire sample in a statistical setting. A key technical aspect of our work is demonstrating that this technique remains applicable even if M is a random variable dependent on the specific realization of the training sample.

Another technical element of our proof is the use of suffix averaging. To illustrate this, consider an algorithm that, upon processing an i.i.d. realizable sample of size t , makes an error on a newly sampled point with probability at most d/t , where d is, for example, the VC dimension. The expected total number of mistakes for a sample of size n is then bounded by $\sum_{t=1}^n d/t = O(d \log n)$, which cannot lead to the desired accuracy. Interestingly, when utilizing only the suffix of the sample, we have $\sum_{t=n/2}^n d/t = O(d)$, effectively eliminating the logarithmic factor. Using martingale arguments, we essentially convert the last computation into a high probability risk bound that scales as

$$O \left(\frac{d}{n} + \frac{1}{n} \log \left(\frac{1}{\delta} \right) \right).$$

Remarkably, the two observations discussed above emerged almost simultaneously in the late 1980s. Haussler, Littlestone, and Warmuth [\[41\]](#) discovered that suffix averaging does

not introduce logarithmic factors in sequential errors, while Littlestone established optimal online-to-batch conversions for algorithms with a deterministic number of errors [55]. [Theorem 3.1.4](#) connects these two insights.

Classification

The first major application of [Theorem 3.1.4](#) is to the setting of multiclass classification. In this setting, we assume the label space $\mathcal{Y}$ is finite and we take the loss function to be the 0/1 loss. In this section, we use $\text{err}_P(f)$ to denote the prediction risk of f . Before we state the main result, we introduce *the one-inclusion hypergraph density* which is the central complexity measure we use. To do so, we introduce some basic definitions and notation.

Let $G = (V, E)$ be a hypergraph. Given a subset of vertices $U \subseteq V$, we define the induced hypergraph $G[U]$ to be the hypergraph with vertex set U and edge set $E' = \{e \cap U : e \in E, |e \cap U| > 1\}$. For a hypergraph $G = (V, E)$, we define the density of G to be

$$\text{Dens}(G) = \frac{1}{|V|} \sum_{e \in E} (|e| - 1).$$

We define the maximum density of G to be

$$\mu(G) = \max_{U \subseteq V} \text{Dens}(G[U]).$$

In words, $\mu(G)$ is the largest density of an induced hypergraph of G . We will also find it useful to define the average degree of a hypergraph $G = (V, E)$ to be

$$\text{Avg}(G) = \frac{1}{|V|} \sum_{e \in E} |e|.$$

Given a hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ and subset $U = \{x_1, \dots, x_n\}$ of the instance space $\mathcal{X}$, we define the projection of $\mathcal{H}$ onto U to be $\mathcal{H}|_U = \{(f(x_1), \dots, f(x_n)) : f \in \mathcal{H}\}$. For a projection $\mathcal{H}|_U$ onto an n -element set U , we will often write $f(i) = f(x_i)$ where $i \in [n]$.

We can now define one-inclusion hypergraphs which were first studied in the setting of binary classification by Haussler, Littlestone, and Warmuth [42],⁴ and generalized to the multiclass setting by Rubinstein, Bartlett, and Rubinstein [74]. Daniely and Shalev-Shwartz [29] initiated the study of these objects for infinite $\mathcal{Y}$.

Definition 3.1.5 (One-inclusion hypergraph). Fix a hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ and an n -element subset of the domain $U \subseteq \mathcal{X}$. The one inclusion hypergraph $G(\mathcal{H}|_U) = (V, E)$ has its vertex set as $V := \mathcal{H}|_U$. For every $i \in [n]$ and $f : [n] \setminus \{i\} \rightarrow \mathcal{Y}$, let $e_{i,f}$ be the set of all $f' \in \mathcal{H}|_U$ such that $f'(j) = f(j)$ for all $j \in [n] \setminus \{i\}$. The edge set of the one-inclusion graph $G(\mathcal{H}|_U)$ is given by

$$E = \{e_{i,f} : i \in [n], f : [n] \setminus \{i\} \rightarrow \mathcal{Y}, |e_{i,f}| > 1\}.$$

⁴In the binary setting, the one-inclusion hypergraph is simply a graph.

In words, we create a hyperedge for all vertices (projected hypotheses) that are the same in all but one coordinate $i \in [n]$, and we do this for every coordinate.

The one-inclusion hypergraph density of a hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ is defined as follows.⁵

Definition 3.1.6 (One-inclusion hypergraph density of $\mathcal{H}$ [29]). Fix a hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$. We define the *one-inclusion hypergraph density* of the class $\mathcal{H}$ for the sample size n to be

$$\mu_{\mathcal{H}}(n) = \max_{S \in \mathcal{Z}^n} \mu(G(\mathcal{H}|_{\mathcal{X}_S})).$$

In words, $\mu_{\mathcal{H}}(n)$ is the largest maximum density of any one-inclusion graph formed from $\mathcal{H}$ and a training sample S of size n (possibly containing repetitions).

Haussler, Littlestone, and Warmuth [42] defined the one-inclusion *graph* algorithm for the setting of binary classification, and Rubinstein, Bartlett, and Rubinstein [74] generalized this algorithm to the multiclass setting and dubbed it the one-inclusion *hypergraph* algorithm. The one-inclusion hypergraph algorithm satisfies Assumptions 3.1.1 and 3.1.3, and it is the predictor that we use in our results. The details of this algorithm are not essential to prove the new results in this paper, however we include a description in Section 3.5 for the interested reader. Finally, we note that the one-inclusion hypergraph algorithm collapses to the original one-inclusion graph algorithm when $\mathcal{Y} = \{0, 1\}$.

Our main PAC risk bound for multiclass classification scales with $\mu_{\mathcal{H}}(n)$.

Theorem 3.1.7. Fix a hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$. There is a predictor $\hat{f} : \mathcal{X} \times \mathcal{U} \rightarrow \mathcal{Y}$ which, for any $\delta \in (0, 1)$, and $S \sim P^n$ sampled from any realizable distribution P , satisfies

$$\text{err}_P(\hat{f}(\cdot; S)) \leq 9.64 \left(\frac{\lceil \mu_{\mathcal{H}}(n) \rceil}{n} + \frac{1}{n} \log \left(\frac{2}{\delta} \right) \right),$$

with probability at least $1 - \delta$ over the randomness of S .

We include the proof of Theorem 3.1.7 in Section 3.5. This bound provides a sharp characterization of the prediction risk in terms of $\mu_{\mathcal{H}}(n)$, boosting the *optimal in-expectation* risk bound of Daniely and Shalev-Shwartz [29] to the canonical PAC setup. In particular, our optimal PAC bound does not exhibit an *explicit* dependence on the number of classes which is *not* the case for the bounds achieved by ERM classifiers based on uniform convergence. We refer to Section 3.2 for additional details.

We now briefly discuss the optimality of the bound in Theorem 3.1.7. The $\frac{1}{n} \log \left(\frac{2}{\delta} \right)$ dependence is optimal up to absolute constant factors (already for binary classification)

⁵Our definition of this complexity measure is slightly different than the original definition introduced by Daniely and Shalev-Shwartz [29]. In their paper, they defined this complexity measure using the *average degree* instead of the *edge density*, which differ by a factor of at most 2. Their definition is also more general and handles infinite $\mathcal{Y}$.

under the standard non-triviality assumption that $\mathcal{H}$ contains two hypotheses that disagree on at least one instance and agree on at least one instance; see [34]. It remains to justify that the dependence on $\mu_{\mathcal{H}}(n)$ cannot, in general, be removed. Daniely and Shalev-Shwartz [29, Theorem 2] proved a realizable transductive minimax in-expectation lower bound of order $\mu_{\mathcal{H}}(n)/n$. Unfortunately, this is an in-expectation obstruction and does not immediately yield a constant-failure-probability lower bound of order $\mu_{\mathcal{H}}(n)/n$. Our next result, whose proof is deferred to Section 3.5 shows the desired lower bound.

Theorem 3.1.8. *There exists an absolute constant $C > 0$ such that the following holds. Let $\mathcal{X}, \mathcal{Y}$, and $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ be a sample space, label space, and hypothesis class respectively. Furthermore, suppose that $\mathcal{H}$ is non-trivial (Definition 3.5.6). Then, for any predictor $\hat{f}(\cdot; \cdot)$, $n \in \mathbb{N}$, and $\delta \in (\exp(-n/4), 1/16]$, there exists a realizable distribution P such that:*

$$\Pr_{S \sim P^n} \left[\text{err}_P \left(\hat{f}(\cdot; S) \right) \geq C \left(\sup_{k \leq 2n} \frac{\mu_{\mathcal{H}}(k)}{n} + \frac{\log(1/\delta)}{n} \right) \right] \geq \delta.$$

A characterization of learnability with respect to the sequence $(\mu_{\mathcal{H}}(n))_{n \in \mathbb{N}}$ is a bit unsatisfying due to its implicit dependence on n . This prompted Daniely and Shalev-Shwartz to investigate whether there is a *single number* that can replace this sequence of complexity measures in these risk bound. They proposed a notion of a dimension which was later named the Daniely Shalev-Shwartz (DS) dimension [24].⁶

Definition 3.1.9 (DS dimension [29]). Fix a hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$. The *Daniely Shalev-Shwartz dimension* of $\mathcal{H}$, denoted d_{DS} , is the largest integer d such that there is a subset $\mathcal{H}' \subseteq \mathcal{H}$ satisfying

$$\max_{S \in \mathcal{Z}^d} \text{Avg}(G(\mathcal{H}'|_{\mathcal{X}_S})) = d.$$

To better appreciate this definition, recall the definition of the VC dimension in the binary setting. Fix a binary hypothesis class $\mathcal{H} \subseteq \{0, 1\}^{\mathcal{X}}$. We say an n -element set $U \subseteq \mathcal{X}$ is *shattered* by $\mathcal{H}$ if $\mathcal{H}|_U = \{0, 1\}^n$. The VC dimension of $\mathcal{H}$ is the largest integer d such that there exists a d -element subset of $\mathcal{X}$ that is shattered by $\mathcal{H}$. When $\mathcal{Y} = \{0, 1\}$, the DS dimension is equivalent to the VC dimension. Daniely and Shalev-Shwartz made the following conjecture that relates the DS dimension to the sequence $(\mu_{\mathcal{H}}(n))_{n \in \mathbb{N}}$.

Conjecture 3.1.10 ([29]). *Fix a hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ with DS dimension d_{DS} . There is an absolute constant $c > 0$ such that for any $n \geq d_{\text{DS}}$ we have*

$$\mu_{\mathcal{H}}(n) \leq c \cdot d_{\text{DS}}.$$

A positive resolution of this combinatorial conjecture would imply, via Theorem 3.1.7, the PAC bound

$$\text{err}_P(\hat{f}) = O \left(\frac{d_{\text{DS}}}{n} + \frac{1}{n} \log \left(\frac{1}{\delta} \right) \right),$$

⁶This dimension was originally defined to allow for infinite $\mathcal{Y}$. Our results easily extend to infinite $\mathcal{Y}$, however we choose to consider finite $\mathcal{Y}$ for simplicity.

which would significantly improve upon the recent work of Bruckhim, Carmon, Dinur, Moran, and Yehudayoff [24] who proved the bound⁷

$$\tilde{O} \left(\frac{d_{\text{DS}}^{3/2}}{n} + \frac{1}{n} \log \left(\frac{1}{\delta} \right) \right).$$

The main motivation for [Conjecture 3.1.10](#) is an elegant result of Haussler, Littlestone, and Warmuth [42] (see also [40]) which shows that for every *binary* hypothesis class $\mathcal{H}$ with VC dimension d , the one-inclusion hypergraph density is at most d .

Theorem 3.1.11 ([42, 40]). *Fix a hypothesis class $\mathcal{H} \subseteq \{0, 1\}^{\mathcal{X}}$ with VC dimension d . For every $n \geq 1$ we have*

$$\mu_{\mathcal{H}}(n) \leq d.$$

In fact, using [Theorems 3.1.7](#) and [3.1.11](#) in the binary setting, we can get a sharp PAC bound that scales with a single parameter, the VC dimension.

Corollary 3.1.12. *Fix a hypothesis class $\mathcal{H} \subseteq \{0, 1\}^{\mathcal{X}}$ with VC dimension d . There is a predictor $\hat{f} : \mathcal{X} \times \mathcal{U} \rightarrow \{0, 1\}$ which, for any $\delta \in (0, 1)$, and $S \sim P^n$ sampled from any realizable distribution P , satisfies*

$$\text{err}_P \left(\hat{f}(\cdot; S) \right) \leq 9.64 \left(\frac{d}{n} + \frac{1}{n} \log \left(\frac{2}{\delta} \right) \right),$$

with probability at least $1 - \delta$ over the randomness of S .

[Corollary 3.1.12](#) recovers the known optimal risk upper bound for binary classification first proven by Hanneke [36]. Recently, Larsen [52] showed that an implementation of the natural bagging heuristic also achieves an optimal risk bound. Our proof of the optimal bound is remarkably simpler than both of their proofs. Furthermore, both Hanneke and Larsen state that the constant factors in their upper bounds are very large and explicitly ask whether these constants can be reduced. Our new analysis reduces the constant factors by a few orders of magnitude.

In addition, our result can be seen as a partial answer to a question of Warmuth [90] who asked whether the one-inclusion graph algorithm can achieve an optimal PAC risk bound. The lower bound in [Section 3.4](#) shows that the conjecture is not true in the strict sense. That is, the one-inclusion graph algorithm can have a large risk bound in the PAC setting. While the one-inclusion graph algorithm itself is not optimal, our result shows that a simple aggregation of such predictors can achieve an optimal PAC bound.

The final classification setting we apply our result to is partial hypothesis classification. In this setting, the label set is $\mathcal{Y} = \{0, 1, \star\}$ where $\star$ label is interpreted as an “I don’t know” response. The loss function ℓ is again the 0/1 loss. We can extend the notion of shattering

⁷Here, the tilde notation hides poly-logarithmic factors in the parameters n and d_{DS} .

and VC dimension used for binary hypothesis classes to partial hypothesis classes. We say an n -element set $U = \{x_1, \dots, x_n\} \subseteq \mathcal{X}$ is *shattered* by a partial hypothesis class $\mathcal{H} \subseteq \{0, 1, \star\}^{\mathcal{X}}$ if $\{0, 1\}^n \subseteq \mathcal{H}|_U$. The VC dimension of a partial hypothesis class $\mathcal{H}$ is now defined as the largest integer d such that there exists a d -element subset of $\mathcal{X}$ that is shattered by $\mathcal{H}$.⁸ We say that a distribution P over $\mathcal{X} \times \{0, 1\}$ (importantly, not over $\mathcal{X} \times \{0, 1, \star\}$) is realizable by $\mathcal{H}$ if the following holds: for every $n \in \mathbb{N}$, $S = ((X_1, Y_1), \dots, (X_n, Y_n)) \sim P^n$ almost surely there exists $f^* \in \mathcal{H}$ realizing S , i.e., $f^*(X_i) = Y_i$ for all $i \in [n]$.

In this setting, the predictor we use is a simple modification of the one-inclusion hypergraph algorithm to exclude the $\star$ labels. One can show, as in the binary classification setting (Theorem 3.1.11), that the one-inclusion hypergraph density $\mu_{\mathcal{H}}(n)$ of $\mathcal{H}$ is at most d , the VC dimension of $\mathcal{H}$. This leads to the following risk bound which implies a sample complexity upper bound that matches the lower bound in [10] up to universal constants. This resolves a question posed in [10].⁹ We elaborate more on this standard modification of the one-inclusion hypergraph in Section 3.5.

Corollary 3.1.13. *Fix a partial hypothesis class $\mathcal{H} \subseteq \{0, 1, \star\}^{\mathcal{X}}$ with VC dimension d . There is a predictor $\hat{f} : \mathcal{X} \times \mathcal{U} \rightarrow \{0, 1, \star\}$ which, for any $\delta \in (0, 1)$, and $S \sim P^n$ sampled from a realizable distribution P , satisfies*

$$\text{err}_P(\hat{f}(\cdot; S)) \leq 9.64 \left(\frac{d}{n} + \frac{1}{n} \log \left(\frac{2}{\delta} \right) \right),$$

with probability at least $1 - \delta$ over the randomness of S .

In addition to resolving these questions regarding optimal rates, our technique can be seen as a unifying perspective for these notions of classification. This should be contrasted with the fact that the previous algorithms and proof of optimality in the binary case [36, 52] were built on empirical risk minimization which can provably fail in the more general settings. Due to the generality of our main result, the list of potential applications in this section is not exhaustive. For instance, one could extend Corollary 3.1.13 to the multiclass partial hypothesis setting. This would improve some recent bounds appearing in [48].

Bounded regression

Moving from a discrete set $\mathcal{Y}$, we now explore bounded regression with $\mathcal{Y} = [0, 1]$ and the absolute loss $\ell(\hat{y}, y) = |\hat{y} - y|$. The key reference for this section is the work of Bartlett and Long [18], who observed the applicability of the one-inclusion graph algorithm in this setting. We begin by revisiting some standard definitions. Let $\mathcal{X}$ be an instance space with

⁸Note that the complexity of a class does not depend directly on $\star$ labels distinguishing this setting from just being an instance of the multiclass setting with three labels.

⁹The definition of realizable learning used in [10] is more general than the standard notion that we use here. This difference is immaterial in our analysis and our bounds hold under their more general definition as well.

label space $\mathcal{Y} = [0, 1]$ and $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ denote a class of real-valued functions mapping $\mathcal{X}$ to $\mathcal{Y}$. Let $\gamma \geq 0$ be a margin parameter. Define $V_{\mathcal{H}}(\gamma)$ to be the largest integer for which there exists $\tau \in [0, 1]$ and a subset $S \subseteq X$ such that $|S| = V_{\mathcal{H}}(\gamma)$ and, for any $A \subseteq S$, there is an $f_A \in \mathcal{H}$ satisfying $f_A(x) \geq \tau + \gamma$ for all $x \in A$ and $f_A(x) \leq \tau - \gamma$ for $x \in S \setminus A$. This complexity measure, which is a scale-sensitive version of the original shattering notion of Vapnik and Chervonenkis [88] for real-valued functions, is explicitly defined in [11] with the name V_{γ} -dimension. We refer to [50] where this complexity measure is discussed in detail.

A real-valued predictor using the one-inclusion graph algorithm is constructed in the following lemma. Unfortunately, the authors were unable to verify the correctness of the proof of Theorem 1 in [18] where such a construction was originally proposed. We instead devise an alternative predictor that achieves an even sharper in-expectation risk bound than the one claimed in [18]. The proof of the lemma may be found in [Section 3.6](#).

Lemma 3.1.14. *Fix a hypothesis class $\mathcal{H} \subseteq [0, 1]^{\mathcal{X}}$. There is a predictor $\hat{f} : \mathcal{X} \times \mathcal{U} \rightarrow \mathcal{Y}$ that, for any realizable sample $S = (x_i, y_i)_{i=1}^n$, satisfies [Assumptions 3.1.1](#) and [3.1.3](#) with a leave-one-out error bound*

$$M_n \leq (n + 1)\gamma + V_{\mathcal{H}}(\gamma).$$

The main result of this section is the following PAC bound.

Theorem 3.1.15. *Fix hypothesis class $\mathcal{H} \subseteq [0, 1]^{\mathcal{X}}$ and margin $\gamma \in (0, 1)$. There is a predictor $\hat{f} : \mathcal{X} \times \mathcal{U} \rightarrow [0, 1]$ which, for any $\delta \in (0, 1)$, and $S \sim P^n$ sampled from any realizable distribution P , satisfies*

$$\mathbb{E}_{(X, Y) \sim P} \left[|\hat{f}(X; S) - Y| \right] \leq 4.82 \left(\gamma + \frac{V_{\mathcal{H}}(\gamma)}{n} + \frac{1}{n} \log \left(\frac{2e}{\delta} \right) \right),$$

with probability at least $1 - \delta$ over the randomness of S .

[Theorem 3.1.15](#) presents a bound that is optimal in the following sense: Bartlett and Long [18] demonstrated a certain hypothesis class $\mathcal{H}$ for which the in-expectation version of [Theorem 3.1.15](#) is optimal up to multiplicative constant factors.

Relations to uniform convergence results

Characterization of uniform convergence for real valued sets of functions has been a central question in empirical processes theory motivated by problems in statistical learning theory and convex geometry. The work of Vapnik and Chervonenkis [88] initiated the study of uniform convergence for real valued functions in terms of a scale-insensitive version of the complexity measure $V_{\mathcal{H}}(\gamma)$. The work of Alon et al. [11] studied the complexity measure $V_{\mathcal{H}}(\gamma)$ and showed that its finiteness for all $\gamma > 0$ is a necessary and sufficient condition for a class of functions to be a uniform Glivenko-Cantelli (GC) class. Similar conclusions appeared in [79, 80]. See also the related results in [86].

The main problem with the complexity measure $V_{\mathcal{H}}(\gamma)$ in the context of the uniform GC property is that it poorly characterizes the covering numbers of the corresponding class. According to Talagrand [80], using this complexity measure in quantitative bounds leads to a “loss of accuracy that is devastating when dealing with precise rates”. In fact, most of literature works with a different complexity measure that we call the P_{γ} -dimension. Let $\mathcal{H}$ be a class of $[0, 1]$ -valued functions on some domain set $\mathcal{X}$ and let γ be a margin. We say that $\mathcal{H}$ P_{γ} -shatters a set $S \subseteq \mathcal{X}$ if there exists a function $s : S \rightarrow [0, 1]$ such that, for every $A \subseteq S$, there exists some $f_A \in \mathcal{H}$ satisfying: For every $x \in S \setminus A$, $f_A(x) \leq s(x) - \gamma$ and, for every $x \in A$, $f_A(x) \geq s(x) + \gamma$. Let the P_{γ} -dimension of $\mathcal{H}$, denoted by $P_{\mathcal{H}}(\gamma)$, be the maximal integer such that the $S \subseteq \mathcal{X}$ of size $P_{\mathcal{H}}(\gamma)$ is P_{γ} -shattered by $\mathcal{H}$. This complexity measure is typically referred to as the *fat-shattering* dimension [49, 18, 13]. The main distinction between complexity measures $V_{\mathcal{H}}(\gamma)$ and $P_{\mathcal{H}}(\gamma)$ lies in the level at which shattering occurs. In the case of $V_{\mathcal{H}}(\gamma)$, shattering takes place at a single level, denoted by τ . On the other hand, the introduction of the function s in the second definition allows for shattering at various levels, represented by τ_i , for distinct points $x \in S$. A straightforward relationship between these measures of complexity can be expressed through the following result.

Lemma 3.1.16 ([11]). *For any hypothesis class $\mathcal{H} \subseteq [0, 1]^{\mathcal{X}}$ and $\gamma > 0$, we have*

$$V_{\mathcal{H}}(\gamma) \leq P_{\mathcal{H}}(\gamma) \leq \left(\frac{1}{\gamma} + 1 \right) V_{\mathcal{H}}(\gamma/2).$$

Given a training sample $S = ((x_i, y_i))_{i=1}^n$, recall that an ERM algorithm selects a hypothesis $\hat{f}_{\text{ERM}}$ as

$$\hat{f}_{\text{ERM}} \in \operatorname{argmin}_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n |f(x_i) - y_i|.$$

To place the result of Theorem 3.1.15 into better perspective, we present a more standard risk bound valid for any ERM classifier via a uniform convergence argument. This bound is expressed in terms of $P_{\mathcal{H}}(\gamma)$. We define the covering numbers with respect to the empirical L_1 distance. Given the sample $x_1, \dots, x_n$, we define, for any $f, g \in \mathcal{H}$, $\rho_n(f, g) = \frac{1}{n} \sum_{i=1}^n |f(x_i) - g(x_i)|$. Let $\mathcal{N}_1(t, \mathcal{H}, n)$ denote the maximal (with respect to $x_1, \dots, x_n$) covering number of $\mathcal{H}$ with respect to ρ_n at scale t . As we could not find an explicit reference in the literature,¹⁰ we provide a standard proof in Section 3.6. To simplify the bound, we introduce a mild regularity assumption on the behavior of $P_{\mathcal{H}}(\cdot)$ which allows us to use the powerful result of Rudelson and Vershynin [75].

Proposition 3.1.17. *There are absolute $c_1, c_2 > 0$ such that the following holds. Fix a hypothesis class $\mathcal{H} \subseteq [0, 1]^{\mathcal{X}}$ and margin parameter $\gamma \in (0, 1)$. Assume that there is a constant $c_3 > 1$ such that for all $\theta > 0$, it holds that $P_{\mathcal{H}}(c_3\theta) \leq P_{\mathcal{H}}(\theta)/2$. For any $\delta \in (0, 1)$,*

¹⁰There are closely related bounds in [13, Section 21.4].

every ERM algorithm that receives a training sample $S \sim P^n$ from a realizable distribution P as input and outputs a hypothesis $\widehat{f}_{\text{ERM}}$ satisfies

$$\mathbb{E}_{(X,Y) \sim P} [|\widehat{f}_{\text{ERM}}(X) - Y|] \leq c_1 \left(\gamma + \frac{P_{\mathcal{H}}(c_2 \gamma)}{n} + \frac{1}{n} \log \left(\frac{1}{\delta} \right) \right),$$

with probability at least $1 - \delta$ over the randomness of S .

[Lemma 3.1.16](#) demonstrates that the bound of [Theorem 3.1.15](#) is never worse than the bound of [Proposition 3.1.17](#), up to constant multiplicative factors. However, the difference between the corresponding complexity measures can be quite significant, as shown by the following example which pertains to the classic problem of isotonic regression. Our example also illustrates the result from [\[75\]](#), emphasizing the equivalence between the logarithm of covering numbers and the P_γ -dimension for general hypothesis classes.

Example 3.1.18 (Isotonic regression with absolute loss). Let $\mathcal{H} \subseteq [0, 1]^{[0,1]}$ be a set of all monotonic functions. Then, for any $\gamma, t \in (0, 1]$,

$$V_{\mathcal{H}}(\gamma) \leq 2, \quad \text{but} \quad P_{\mathcal{H}}(\gamma) = \Theta \left(\frac{1}{\gamma} \right), \quad \text{and} \quad \log \mathcal{N}_1(t, \mathcal{H}, n) = \Theta \left(\frac{1}{t} \right).$$

Specifically, in this case, the upper bound of [Theorem 3.1.15](#) scales as $O \left(\frac{1}{n} \log \left(\frac{1}{\delta} \right) \right)$, whereas the covering-number-based upper bound of [Proposition 3.1.17](#) scales as $O \left(\frac{1}{\sqrt{n}} + \frac{1}{n} \log \left(\frac{1}{\delta} \right) \right)$.

The bounds in this example are based on several existing results. The upper bound for $P_{\mathcal{H}}(\gamma)$ follows from [Lemma 3.1.16](#), while the lower bound can be directly constructed from the definition. The upper bound for $\log \mathcal{N}_1(t, \mathcal{H}, n)$ follows from [\[82, Theorem 2.7.5\]](#), while the lower bound is implied by [\[13, Theorem 12.10\]](#).

One might think that employing a version of [Proposition 3.1.17](#) with localized covering numbers could improve the ERM convergence rates in [Example 3.1.18](#). Unfortunately, this does not help because global and local covering numbers have the same behavior for non-parametric classes [\[92, 64\]](#). Finally, we mention that the recent work of Kotłowski, Koolen, and Malek [\[51\]](#) explores in-expectation risk bounds for the class considered in [Example 3.1.18](#) in a similar learning setting.

3.2 Related work

The optimal sample complexity of binary classification. Determining the optimal sample complexity, or equivalently the optimal risk bound, of realizable (noiseless) binary classification is a classic problem in statistical learning theory. The early work of Vapnik and Chervonenkis [\[84\]](#) introduced what we call the PAC model in this paper (a terminology due to Valiant [\[81\]](#), who also studied these questions from a computational perspective), providing matching upper and lower risk bounds for finite hypothesis classes $\mathcal{H}$. Subsequently,

Vapnik and Chervonenkis presented the first upper bounds on the error rate for an infinite class, specifically the class of half-spaces in $\mathbb{R}^p$ [85]. Their next groundbreaking work [83] generalized the shattering properties of half-spaces to general classes and introduced the complexity measure now known as the VC dimension. The monograph [87] and the paper [21] provide the canonical risk bound applicable to any classifier consistent with the training sample:

$$O\left(\frac{d}{n} \log\left(\frac{n}{d}\right) + \frac{1}{n} \log\left(\frac{1}{\delta}\right)\right). \quad (3.2)$$

This convergence rate is known to be unimprovable for certain sample consistent classification rules and hypothesis classes [42, 15, 77, 36]. However, the existing lower bound [34] that applies to any predictor

$$\Omega\left(\frac{d}{n} + \frac{1}{n} \log\left(\frac{1}{\delta}\right)\right) \quad (3.3)$$

did not match the risk bound (3.2) by a multiplicative logarithmic factor. The pivotal work of Haussler, Littlestone, and Warmuth [42] (originally presented as [41]) took a significant step towards proving an upper bound that matches the rate of (3.3). They introduced the one-inclusion graph algorithm and presented a simple confidence boosting strategy that occasionally provides a better risk bound

$$O\left(\frac{d}{n} \log\left(\frac{1}{\delta}\right)\right), \quad (3.4)$$

raising the question of what the optimal sample complexity of PAC learning is, and whether the one-inclusion graph algorithm is itself optimal [90]. Since then, numerous authors have provided bounds matching the rate of the lower bound (3.3) for specific classes [20, 55, 14, 15, 31, 36] and/or under additional assumptions on the distribution of the data [57, 25, 16, 36]. The problem of obtaining a rate matching (3.3) up to constant factors was ultimately resolved by Hanneke [36], who improved upon an earlier result by Simon [77] which contained a slowly growing additional multiplicative term.

Leave-one-out/exchangeability arguments. The leave-one-out analysis is a standard approach to control the *expected* risk of classification rules. Some early applications of this method can be traced back to the monograph of Vapnik and Chervonenkis [87], where they utilized it to derive optimal risk bounds for both the Perceptron algorithm and the hard margin SVM. However, these results were valid only in-expectation, prompting Vapnik and Chervonenkis [87, Chapter VI, Section 7] to ask whether the same could be achieved with high probability for some practically relevant algorithms. Lunts and Brailovsky [62] showed that this is not possible in general by presenting an example where the leave-one-out argument failed to yield a low variance risk bound, thus excluding the possibility of attaining high-probability upper bounds without additional assumptions. The leave-one-out argument is also central to the analysis of the one-inclusion graph algorithm introduced in [42]. Warmuth [90] conjectured that this algorithm consistently provides optimal PAC upper

bounds. However, [Section 3.4](#) shows that, without additional assumptions, the one-inclusion graph algorithm is unable to achieve the optimal PAC bound. Lastly, we draw attention to the recent analysis of *stable compression schemes* in [\[93, 22, 38\]](#), which reveals that a leave-one-out argument can indeed result in sharp high probability bounds when stability assumptions are incorporated.

Confidence boosting and online to batch conversions. We first discuss the *confidence boosting* approach, a natural method for deriving high-probability risk bounds from in-expectation ones. This technique involves running the in-expectation optimal algorithm on $O(\log(1/\delta))$ independent partitions of the original sample and selecting the best performer based on an independent hold-out sample. Confidence boosting was employed by Haussler, Littlestone, and Warmuth [\[42\]](#), however this approach only leads to a bound with a multiplicative $\log(1/\delta)$ as in [\(3.4\)](#).

A more pertinent approach is the online-to-batch conversion, with its earliest relevant applications tracing back to the works [\[7, 87\]](#). Littlestone [\[55\]](#) demonstrated that if a conservative online algorithm (an algorithm that only changes its state when a point is misclassified) makes a finite number M of mistakes on any sample, it can be converted into a classifier in a statistical setting with a risk bound

$$O\left(\frac{M}{n} + \frac{1}{n} \log\left(\frac{1}{\delta}\right)\right). \quad (3.5)$$

Despite providing a correct additive term containing $\log(1/\delta)$, Littlestone’s bound has two major limitations, both addressed in this paper. The first limitation is the requirement of a finite number of mistakes, characterized by the so-called *Littlestone dimension* [\[56\]](#), which is infinite for most practically relevant hypothesis classes. Haussler, Littlestone, and Warmuth [\[42\]](#) attempted to address this issue by considering the *expected* number of mistakes made by one-inclusion graph algorithms run on the sample in an online manner, resulting in an expected number of mistakes scaling as $O(d \log n)$. The second limitation of the bound [\(3.5\)](#) is the necessity for the algorithm to be conservative. Although any algorithm achieving a deterministic mistake bound can be converted to a conservative one [\[35\]](#), this is not known to hold for algorithms achieving a mistake bound *in-expectation* as is the case in our work.

A pertinent reference is the work of Wu, Heidari, Grama, and Szpankowski [\[91\]](#). They proved, via a reverse martingale argument, that *high-probability online bounds* analogous to the in-expectation bound of Haussler et al. [\[42\]](#) can be attained, showing that $O(d \log n + \log(1/\delta))$ mistakes are made with probability at least $1 - \delta$. Another observation stems from [\[42\]](#), who noted that when the one-inclusion graph algorithm is executed in an online fashion, and its error is measured solely on the second half of the sample — referred to as the *suffix* — the expected number of mistakes diminishes to $O(d)$. This idea of utilizing the suffix of the sample is quite general and has led to recent sharp high-probability bounds for stochastic gradient descent [\[39\]](#).

Multiclass classification and partial hypotheses. Multiclass classification is a natural extension of the binary case. The works [70, 69] extended the PAC model to multiclass problems and introduced an analog of the VC dimension, now commonly referred to as the Natarajan dimension. Later work asked whether this dimension characterizes learnability and highlighted the insufficiency of standard uniform convergence techniques for answering this question. Subsequent works exploring the Natarajan dimension include [43, 19]. While the Natarajan dimension enables the uniform convergence property with a finite number of classes, ERM's may fail to learn the class when the number of classes is unbounded [30]. The authors of [74] adapted the one-inclusion-graph algorithm to multiclass classification, while Daniely and Shalev-Shwartz [29] characterized the expected optimal risk of multiclass problems by the one-inclusion hypergraph density. Recent results and a detailed survey on complexity measures in multiclass classification can be found in [24].

A partial hypothesis class consists of $\{0, 1, \star\}$ -valued hypotheses. A key distinction from multiclass classification is the unique treatment of $\star$ -values, which do not affect the notion of shattering or corresponding complexity measures. This idea traces back to the work of Haussler and Long [43] and is further explored in the work of Bartlett and Long [18]. The authors in [18] recognized that uniform convergence should not be relied upon for partial hypotheses and adapted the one-inclusion graph algorithm accordingly. Subsequent work by Long [58] extended the analysis of partial hypotheses to the agnostic setting and derived corresponding risk bounds. The recent resurgence of interest in this topic has been sparked by the work of Alon, Hanneke, Holzman, and Moran [10]. Their contributions include determining sharper bounds for both the realizable and agnostic cases, while also extending these ideas to online learning.

3.3 Proof of main result

In this section, we prove the main technical result of the paper, [Theorem 3.1.4](#). Recall that we aim to bound the average risk

$$\frac{4}{3n} \cdot \sum_{t=n/4}^{n-1} R_P(\hat{f}(\cdot; S_{\leq t})) \quad (3.6)$$

of a sequence of predictors constructed from the data with high probability. As a first step, we analyze the empirical quantity

$$\frac{4}{3n} \cdot \sum_{t=n/4}^{n-1} \ell(\hat{f}(X_{t+1}; S_{\leq t}), Y_{t+1}). \quad (3.7)$$

Intuitively, (3.7) approximates (3.6) since the test point X_{t+1} is independent of the training sample $S_{\leq t}$ used to construct the predictor. That is, $\ell(\hat{f}(X_{t+1}; S_{\leq t}), Y_{t+1})$ is an unbiased

estimator of $R_P(\hat{f}(\cdot; S_{\leq t}))$. We then convert the in-expectation bound to a PAC bound using a martingale concentration argument in [Lemma 3.3.2](#).

The main technical challenge is to bound (3.7) with high probability. This step makes crucial use of the fact that the predictors under consideration are permutation invariant as captured by [Assumption 3.1.3](#). Crucially, our analysis conditions on the values and labels of the inputs $((X_i, Y_i))_{i=1}^n$ but not on their *permutation* in the input sequence. The key insight in the proof is the following conditional independence structure. For any $i < j$, $\ell(\hat{f}(X_{i+1}; S_{\leq i}), Y_{i+1})$ has small expected value even when conditioned on $\ell(\hat{f}(X_{j+1}; S_{\leq j}), Y_{j+1})$. This is due to the fact that $\ell(\hat{f}(x_{j+1}; S_{\leq j}), y_{j+1})$ is determined by the *set* of training samples $S_{\leq j}$ and (X_{j+1}, Y_{j+1}) while $\ell(\hat{f}(x_{i+1}; S_{\leq i}), y_{i+1})$ depends on the *set* of training samples $S_{\leq i}$ and (X_{i+1}, Y_{i+1}) . Due to the exchangeability of $S_{\leq i}$ even given the set $S_{\leq j}$, we can relate $\ell(\hat{f}(X_{j+1}; S_{\leq j}), Y_{j+1})$ has small expected value from [Assumption 3.1.1](#). This intuition is formalized in [Lemma 3.3.3](#) by constructing an appropriate filtration and using a simple martingale concentration bound.

To prove our results, we need the following martingale Chernoff bounds. We include a proof of [Lemma 3.3.1](#) in [Section 3.7](#) for completeness.

Lemma 3.3.1. *Let $W_1, \dots, W_T$ be a stochastic process adapted to the filtration $(\mathcal{F}_i)_{i \leq T}$. Suppose that $0 \leq W_t \leq 1$ almost surely. Then, for any $\delta, \lambda, \eta \in (0, 1)$, we have*

$$\Pr \left[\sum_{t=1}^T W_t \geq \frac{e^\lambda - 1}{\lambda} \sum_{t=1}^T \mathbb{E}[W_t | \mathcal{F}_{t-1}] + \frac{\log(1/\delta)}{\lambda} \right] \leq \delta, \quad (3.8)$$

and

$$\Pr \left[\sum_{t=1}^T \mathbb{E}[W_t | \mathcal{F}_{t-1}] \geq \frac{\eta e^\eta}{e^\eta - 1} \sum_{t=1}^T W_t + \frac{e^\eta \log(1/\delta)}{e^\eta - 1} \right] \leq \delta. \quad (3.9)$$

We first formally look at the argument connecting the empirical loss to the risk.

Lemma 3.3.2 (Forward martingale risk bound). *Fix a realizable distribution P and let ℓ be a loss bounded by 1. Let $\hat{f} : \mathcal{X} \times \mathcal{U} \rightarrow \mathcal{Y}$ be a predictor. Given a training sample $S = ((X_1, Y_1), \dots, (X_T, Y_T)) \sim P^T$, for any $\delta, \eta \in (0, 1)$, we have that*

$$\sum_{t=T/4}^{T-1} R_P(\hat{f}(\cdot; S_{\leq t})) \leq \frac{\eta e^\eta}{e^\eta - 1} \sum_{t=T/4}^{T-1} \ell(\hat{f}(X_{t+1}; S_{\leq t}), Y_{t+1}) + \frac{e^\eta \log(1/\delta)}{e^\eta - 1},$$

with probability at least $1 - \delta$ over the randomness of S .

Proof. For $t = 1, \dots, T$ define the usual filtration $(\mathcal{F}_t)_{t=1}^T$ where $\mathcal{F}_t = \sigma(X_1, \dots, X_t)$. Here, $\sigma(\cdot)$ denotes the smallest σ -algebra with respect to which the arguments are measurable. With slight abuse of notation, we write $\hat{f}_t(\cdot) = \hat{f}(\cdot; S_{\leq t})$. By definition $R_P(\hat{f}_t) = \mathbb{E}[\ell(\hat{f}_t(X_{t+1}), Y_{t+1}) | \mathcal{F}_t]$, so applying [Lemma 3.3.1](#) to the random variables

$$\ell(\hat{f}_{T/4}(X_{T/4+1}), Y_{T/4+1}), \dots, \ell(\hat{f}_{T-1}(X_T), Y_T)$$

gives us that

$$\sum_{t=T/4}^{T-1} R_P(\hat{f}_t) \leq \frac{\eta e^\eta}{e^\eta - 1} \sum_{t=T/4}^{T-1} \ell(\hat{f}_t(X_{t+1}), Y_{t+1}) + \frac{e^\eta \log(1/\delta)}{e^\eta - 1},$$

with probability at least $1 - \delta$ over the randomness of S . $\square$

As noted earlier, [Lemma 3.3.2](#) tells us that to bound the average risk of the predictors $\hat{f}(\cdot, S_{\leq t})$, it suffices to bound the loss when evaluated on the points X_t . In order to do this, we use a reverse martingale argument. These arguments are standard in the study of empirical processes since the work of Pollard [72] and have found recent applications in sequential estimation and uncertainty quantification. The results of Manole and Ramdas [63, Proposition 5] make a connection between leave-one-out arguments and reverse martingales. Wu, Heidari, Grama, and Szpankowski [91] used similar arguments specifically for the study of sequential covering numbers using the one-inclusion graph algorithm. The proof below could be shortened by translating it to the language of reverse martingales and the exchangeable filtration, but we chose an elementary presentation in terms of random permutations for clarity.

Lemma 3.3.3 (Reverse martingale bound). *Fix a realizable distribution P . Let $\hat{f} : \mathcal{X} \times \mathcal{U} \rightarrow \mathcal{Y}$ be a predictor satisfying [Assumptions 3.1.1](#) and [3.1.3](#) with leave-one-out error M_n . Given a training sample $S = (Z_1, \dots, Z_T) = ((X_1, Y_1), \dots, (X_T, Y_T)) \sim P^T$, for any $\delta, \lambda \in (0, 1)$, we have that*

$$\sum_{t=T/4}^{T-1} \ell(\hat{f}(X_{t+1}; S_{\leq t}), Y_{t+1}) \leq \left(\ln(4) + \frac{1}{2} \right) \frac{e^\lambda - 1}{\lambda} M_T + \frac{\log(1/\delta)}{\lambda},$$

with probability at least $1 - \delta$ over the randomness of S .

Proof. Let $I_t = \ell(\hat{f}(X_{T-t+1}; S_{\leq T-t}), Y_{T-t+1})$. Note that I_t is a function of $Z_1, \dots, Z_{T-t+1}$ and when required we will make this dependence explicit by writing $I_t(Z_1, \dots, Z_{T-t+1})$. Further, observe that since $Z_1, \dots, Z_{T-t+1}$ are exchangeable random variables, we have that the joint distribution of $Z_{\pi(1)}, \dots, Z_{\pi(T)}$ does not depend on the permutation π . Consider a uniformly random permutation π independent of $Z_1, \dots, Z_T$. That is, throughout the proof we may assume that the values $Z_1, \dots, Z_T$ are fixed. Consider the filtration

$$\mathcal{F}_t = \sigma(\{Z_{\pi(1)}, \dots, Z_{\pi(T-t)}\}, Z_{\pi(T-t+1)}, \dots, Z_{\pi(T)}).$$

Here, $\{Z_{\pi(1)}, \dots, Z_{\pi(T-t)}\}$ denotes the set of values taken by $Z_{\pi(1)}, \dots, Z_{\pi(T-t)}$ (excluding their order) and $\sigma(\cdot)$ denotes the smallest σ -algebra with respect to which the arguments are measurable. Let $S_\pi = ((X_{\pi(1)}, Y_{\pi(1)}), \dots, (X_{\pi(n)}, Y_{\pi(n)}))$ denote the permutation of the training sample with respect to π . Define

$$I'_t = I_t(Z_{\pi(1)}, \dots, Z_{\pi(T-t+1)}) = \ell(\hat{f}(X_{\pi(T-t+1)}; (S_\pi)_{\leq T-t}), Y_{\pi(T-t+1)}).$$

Note that I'_t is adapted to the filtration $(\mathcal{F}_t)_{t=1}^T$ since, by [Assumption 3.1.3](#), $\hat{f}$ is permutation invariant on the training sample $(S_\pi)_{\leq T-t}$. Applying (3.8) gives us that with probability at least $1 - \delta$, we have

$$\sum_{t=1}^{3T/4} I'_t \leq \frac{e^\lambda - 1}{\lambda} \sum_{t=1}^{3T/4} \mathbb{E}[I'_t | \mathcal{F}_{t-1}] + \frac{\log(1/\delta)}{\lambda}. \quad (3.10)$$

To use this bound, we need to control

$$\mathbb{E}[I'_t | \mathcal{F}_{t-1}] = \mathbb{E}[I(Z_{\pi(1)}, \dots, Z_{\pi(T-t+1)}) | \mathcal{F}_{t-1}]. \quad (3.11)$$

Notice that at time $t-1$ (with respect to the filtration $(\mathcal{F}_t)_{t=1}^T$) the random indices $\pi(1), \dots, \pi(T-t+1)$ correspond to a uniformly random permutation of the elements of the set $[T] \setminus \{\pi(i)\}_{i=T-t+2}^T$. A standard computation relates averaging over uniformly random permutations to the leave-one-out error from which we can conclude together with (3.11) and [Assumption 3.1.1](#) the bound

$$\mathbb{E}[I'_t | \mathcal{F}_{t-1}] \leq \frac{M_{T-t+1}}{T-t+1}.$$

Plugging the above into (3.10) along with the exchangeability of $Z_1, \dots, Z_T$ gives us

$$\begin{aligned} \sum_{t=T/4}^{T-1} \ell(\hat{f}(X_{t+1}; S_{\leq t}), Y_{t+1}) &\leq \frac{e^\lambda - 1}{\lambda} \sum_{t=1}^{3T/4} \frac{M_{T-t+1}}{T-t+1} + \frac{\log(1/\delta)}{\lambda} \\ &\leq \frac{e^\lambda - 1}{\lambda} M_T \sum_{t=T/4+1}^T \frac{1}{i} + \frac{\log(1/\delta)}{\lambda} \end{aligned} \quad (3.12)$$

$$\leq \frac{e^\lambda - 1}{\lambda} M_T \left(\ln 4 + \frac{1}{2} \right) + \frac{\log(1/\delta)}{\lambda}. \quad (3.13)$$

The inequality (3.12) follows from the monotonicity of the function M_t . Finally, (3.13) follows from a simple approximation to the harmonic series considered. $\square$

We are now ready to prove our main result by piecing together the two results above. We restate the theorem for convenience.

Theorem 3.1.4. *Fix a hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, loss function $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow [0, 1]$, and a predictor $\hat{f} : \mathcal{X} \times \mathcal{U} \rightarrow \mathcal{Y}$ satisfying [Assumptions 3.1.1](#) and [3.1.3](#). Then, for any realizable distribution P over $\mathcal{X} \times \mathcal{Y}$ and confidence parameter $\delta \in (0, 1)$, given a training sample $S \sim P^n$ we have*

$$\frac{4}{3n} \cdot \sum_{t=n/4}^{n-1} R_P(\hat{f}(\cdot; S_{\leq t})) \leq 4.82 \left(\frac{M_n}{n} + \frac{1}{n} \log \left(\frac{2}{\delta} \right) \right),$$

with probability at least $1 - \delta$ over the randomness of S .

Proof of Theorem 3.1.4. Using Lemmas 3.3.2 and 3.3.3 with confidence parameter $\delta/2$ together with a union bound implies, with probability at least $1 - \delta$ over the randomness of S , that

$$\frac{4}{3n} \sum_{t=n/4}^{n-1} R_P(\hat{f}(\cdot; S_{\leq t})) \leq \frac{4}{3n} \left(\frac{\eta e^\eta}{e^\eta - 1} \left((\ln(4) + 1/2) \frac{e^\lambda - 1}{\lambda} M_n + \frac{\log(1/\delta)}{\lambda} \right) + \frac{e^\eta \log(2/\delta)}{e^\eta - 1} \right).$$

Setting the parameters¹¹ to $\lambda = 0.82$ and $\eta = 0.78$ followed by a simple calculation gives us the upper bound

$$\frac{4}{3n} \sum_{t=n/4}^{n-1} R_P(\hat{f}(\cdot; S_{\leq t})) \leq 4.82 \left(\frac{M_n}{n} + \frac{1}{n} \log \left(\frac{2}{\delta} \right) \right).$$

This completes the proof. □

3.4 Proof of the lower bound

In this section, we begin by presenting negative results on the performance of the one-inclusion graph algorithm. Throughout the text, we provide several results, with arguably the strongest—but also the most technically involved—deferred to Section 3.8. Importantly, all proofs share a unifying technical argument. To set the stage, let us revisit Warmuth’s conjecture [90]. Formally, Warmuth posits the following:

Conjecture 3.4.1 ([90]). *The one-inclusion graph algorithm of [42] always achieves the (PAC) lower bound of [34]. That is, after receiving $O\left(\frac{1}{\varepsilon}(d + \log \frac{1}{\delta})\right)$ examples, its error is at most ε with probability at least $1 - \delta$.*

As mentioned above, we establish several results showing that there exist one-inclusion prediction strategies that are *optimal* in expectation, yet whose high-probability guarantees are no better than those obtained via *Markov’s* inequality. We begin with the following definition. A one-inclusion graph strategy is said to be *valid* if it always produces graphs with an out-degree bounded by $O(d)$. We also need the following definition.

Definition 3.4.2 (Star sets). A concept class $\mathcal{H}$ defined on an infinite domain $\mathcal{X}$ has a *star set of arbitrary size* if for any integer n , there exist $U = \{x_1, \dots, x_n\} \subseteq \mathcal{X}$ and classifiers $f_0, f_1, \dots, f_n \in \mathcal{H}$ such that for all $i = 1, \dots, n$,

$$\{x \in U : f_i(x) \neq f_0(x)\} = \{x_i\}.$$

For any n such a set U is a *star set*. The corresponding classifiers $\{f_i\}_{i=0}^n$ witness the star set U .

¹¹The choices of the parameters have been optimized to get a small constant.

We are ready to formulate the central result of this section.

Theorem 3.4.3. *Let $d \geq 1$ be an arbitrary integer, and let $\mathcal{H}$ be a hypothesis class defined on a countable set $\mathcal{X}$ with finite VC dimension d . Assume that $\mathcal{H}$ has a star set of arbitrary size. There exist positive absolute constants C, c_1, c_2 , and c_3 such that for any sample size $n \geq Cd$ and confidence parameter δ satisfying $\delta \in (c_1 d/n, c_2)$, there is a valid one-inclusion graph algorithm that outputs the hypothesis $\hat{f}_{\text{OIG}}$ such that, for some distribution $P = P(n, \delta)$ over $\mathcal{X} \times \{0, 1\}$, we have with probability at least δ :*

$$\text{err}_P(\hat{f}_{\text{OIG}}) \geq c_3 \frac{d}{n\delta}.$$

Some remarks are in order. First, we emphasize that the out-degree of all one-inclusion graphs constructed in the above theorem is bounded by $d + 1$, even though the statement refers to valid orientations. Of course, one may interpret Warmuth’s conjecture very literally, assuming it claims that the original maximum flow algorithm used in [42] to minimize the maximum out-degree is optimal in the PAC setting, rather than the broader interpretation that any orientation strategy yielding optimal (up to constants) in-expectation guarantees is also optimal in the PAC setting. We provide a negative result even for this interpretation of the conjecture by slightly modifying the argument used in this section to prove [Theorem 3.4.3](#). Specifically, this lower bound applies to an orientation that is a feasible output of the maximum flow algorithm proposed by [42] that may yield *fractional* orientations for the one-inclusion graph. Since this argument is very similar but slightly more involved, we defer the statement and proof of this result to [Section 3.8](#). Note that our lower bounds do not rule out the existential possibility that, for a given class, there exists a (possibly class-dependent) orientation rule that achieves the optimal PAC rate.

To prove [Theorem 3.4.3](#), we first introduce some new notation and concepts. Let $A = \{x_1, \dots, x_m\}$ be a set of size m with a fixed ordering of its elements. For any k -element subset $B \subseteq A$ with $k \leq m$, we will slightly overload notation by simultaneously referring to B as a subset of elements of A , and the length m binary vector that is 1 in the i -th entry if and only if x_i is in B . We will clarify the definition of the set A such that its correspondence to length m binary vectors is clear. For a finite ordered set $A = \{x_1, \dots, x_m\}$ and $k \in \{0, \dots, m\}$, define

$$\mathcal{U}_k(A) := \{B \subseteq A : |B| = k\}.$$

When A is clear from context we write $\mathcal{U}_k$. When convenient, we identify $B \in \mathcal{U}_k(A)$ with its indicator vector in $\{0, 1\}^m$ (with respect to the fixed ordering of A).

We begin by recalling VT codes [89] and stating a useful property they enjoy. VT codes of length m and parameter $a \in \{0, 1, \dots, m\}$ are given by the following family of vectors:

$$\mathcal{C}_a(m) = \left\{ C \in \{0, 1\}^m : \sum_{i=1}^m i \cdot C(i) \equiv a \pmod{m+1} \right\}.$$

We will find the following notion of coverage to be useful.

Definition 3.4.4 (Coverage). Fix an integer m . We say a vector $U \in \{0, 1\}^m$ is covered by the vector $U' \in \{0, 1\}^m$ if there is an index $i \in [m]$ such that $U(i) = 0$ but $U'(i) = 1$, and for every $j \neq i$ we have $U(j) = U'(j)$. We will denote this by $U \prec U'$.

With this notation, VT codes satisfy a “uniqueness” property crucial for our analysis.

Lemma 3.4.5 (Unique neighborhoods). *Let A be a set of m elements. For any $k \leq m$, binary vector $U' \in \mathcal{U}_k(A)$ and $a \in \{0, 1, \dots, m\}$, there exists at most one binary vector $U \in \mathcal{C}_a(m)$ with $U \prec U'$.*

Proof. Towards a contradiction assume that for some $U' \in \mathcal{U}_k(A)$ there were two vectors $U_1, U_2 \in \mathcal{C}_a(m)$ that satisfied the property above with corresponding indices i_1 and i_2 , i.e., $U_1 \prec U'$ and $U_2 \prec U'$. Then, we have that U_1 and U_2 disagree only on i_1 and i_2 . So

$$0 \equiv \sum_{i=1}^m i \cdot U_1(i) - \sum_{i=1}^m i \cdot U_2(i) \equiv i_2 - i_1 \pmod{m+1}. \quad (3.14)$$

This is a contradiction since $1 \leq i_1, i_2 \leq m$. $\square$

Let $\mathcal{H}$ be the hypothesis class with star sets of arbitrary size and VC dimension d for which we would like to construct the bad one-inclusion graph algorithm. Let n denote the sample size and let $\mathcal{W}_m \subseteq \mathcal{X}$ be a star set of size $2n$. Fix and keep an ordering $\mathcal{W}_m = \{x_1, \dots, x_{2n}\}$ throughout the proof. Let $f_0, f_1, \dots, f_{2n}$ denote the corresponding functions that witness the star set. We will not assume that f_0 is constant on $\mathcal{W}_m$; instead, we take the realizable target to be $f^* := f_0$, so guaranteeing an error at a point x_i amounts to forcing the predictor to output the label $f_i(x_i) \neq f_0(x_i)$.

Our hard distribution $P = P(n, \delta)$ will simply be the uniform distribution over $\mathcal{W}_m$ with labels generated by $f^* = f_0$. We will first prove a lower bound on the error conditioned on the size of $\mathcal{X}_S$, the set of unique X_i observed in $S \sim P^n$. Formally, throughout this proof we define $\mathcal{X}_S$ by *deduplication*: if $S = ((X_1, Y_1), \dots, (X_n, Y_n))$, then $\mathcal{X}_S := \{X_1, \dots, X_n\}$. Let $k = |\mathcal{X}_S|$. Note that $\mathcal{X}_S$ may be associated with an element of $\mathcal{U}_k(\mathcal{W}_m)$. Slightly abusing notation we denote $\mathcal{U}_k(\mathcal{W}_m)$ by $\mathcal{U}_k$, and $\mathcal{U}_{k+1}(\mathcal{W}_m)$ by $\mathcal{U}_{k+1}$. We will now define a bipartite graph with one set of vertices a subset of $\mathcal{U}_k$ (and hence, possible realizations of $\mathcal{X}_S$) and the other $\mathcal{U}_{k+1}$.

Before proceeding with the key technical details of the proof, we elaborate on how we address duplicates in our training sample and how they are related to corresponding one-inclusion graphs. Since the one-inclusion graph is usually defined for an ordered sample, we make explicit how our orientations (constructed for sets of distinct instances) induce orientations for general samples with repetitions. For an ordered instance sequence $S' = (x_1, \dots, x_t)$ (allowing repetitions), define

$$\mathcal{H}|_{S'} := \{(f(x_1), \dots, f(x_t)) : f \in \mathcal{H}\},$$

and let $G(\mathcal{H}|_{S'})$ be the corresponding one-inclusion graph (two vertices adjacent iff they differ in exactly one coordinate). Let $U(S') := \{x_1, \dots, x_t\}$ be the set of distinct instances.

Fix any ordering of $U(S')$ (in our application we will always use the ordering inherited from $\mathcal{W}_m$), and define $\mathcal{H}|_{U(S')}$ and $G(\mathcal{H}|_{U(S')})$ with respect to this ordering. The next lemma shows the monotonicity property needed for the lower bound: repetitions can only *remove* edges, so restricting an orientation from $U(S')$ to S' never increases the maximum out-degree.

Lemma 3.4.6. *Let $S' = (x_1, \dots, x_t)$ be a sequence in $\mathcal{X}$ (allowing repetitions) and let $U(S') = \{x_1, \dots, x_t\}$ be the set of distinct points, equipped with a fixed ordering. Then, under the natural identification of vertices induced by restriction to $U(S')$ and repetition of coordinates, the one-inclusion graph $G(\mathcal{H}|_{S'})$ is a subgraph of $G(\mathcal{H}|_{U(S')})$. Consequently, any orientation $\sigma_{U(S')}$ of $G(\mathcal{H}|_{U(S')})$ induces an orientation $\sigma_{S'}$ of $G(\mathcal{H}|_{S'})$ by restriction, and $\text{out}(\sigma_{S'}) \leq \text{out}(\sigma_{U(S')})$.*

Proof. Fix $S' = (x_1, \dots, x_t)$ and let $U = U(S')$. Since hypotheses are functions, for any $f \in \mathcal{H}$ and any indices i, j with $x_i = x_j$ we have $f(x_i) = f(x_j)$. Thus the map that takes a labeling of U and repeats its value on each occurrence of the corresponding point in S' gives a canonical identification between vertices in $\mathcal{H}|_U$ and vertices in $\mathcal{H}|_{S'}$. Now consider any edge $\{v, w\}$ in $G(\mathcal{H}|_{S'})$. By definition, v and w differ in exactly one coordinate position $i \in [t]$. Let $x = x_i$. If x appears more than once in S' , then v and w must agree on *all* positions where x appears (because the label of x is the same across all occurrences), contradicting that they differ at position i . Hence x appears exactly once in S' . Under the above vertex identification, v and w therefore differ on exactly the single point x in U , so they are adjacent in $G(\mathcal{H}|_U)$ as well. This shows every edge of $G(\mathcal{H}|_{S'})$ also appears in $G(\mathcal{H}|_U)$, i.e., $G(\mathcal{H}|_{S'})$ is a subgraph of $G(\mathcal{H}|_U)$. Finally, restricting any orientation σ_U from $G(\mathcal{H}|_U)$ to the edge set of the subgraph $G(\mathcal{H}|_{S'})$ yields an orientation $\sigma_{S'}$ whose out-degree at each vertex is no larger than under σ_U , proving $\text{out}(\sigma_{S'}) \leq \text{out}(\sigma_U)$. $\square$

In what follows, we specify orientations σ_U for graphs induced by sets U of distinct instances. This also defines the required orientation rule for general ordered datasets: given an ordered sample $S' = (x_1, \dots, x_{n+1})$ (allowing repetitions), let $U = U(S') = \{x_1, \dots, x_{n+1}\}$ be the set of distinct instances, and define $\sigma_{S'}$ to be the restriction of σ_U to the (sub)graph $G(\mathcal{H}|_{S'})$ via Lemma 3.4.6. In particular, $\text{out}(\sigma_{S'}) \leq \text{out}(\sigma_U)$.

This graph will then be used to construct an orientation rule with poor performance. Formally, the vertex sets of the graph are defined below

$$\begin{aligned} V_1 &= \bigcup_{0 \leq i < 4\lceil \delta n \rceil} T_{(i)}, \\ V_2 &= \mathcal{U}_{k+1}, \end{aligned} \tag{VERT-SETS}$$

where $T_i := \mathcal{C}_i(2n) \cap \mathcal{U}_k$ is the intersection of the code $\mathcal{C}_i(2n)$ with the set of vectors containing exactly k ones and $T_{(0)}, \dots, T_{(2n)}$ is a re-ordering such that $|T_{(0)}| \geq |T_{(1)}| \geq \dots \geq |T_{(2n)}|$. In particular, since $\mathcal{C}_i(2n) \cap \mathcal{C}_j(2n) = \emptyset$ for $i \neq j$, we have for any $\ell \in [2n] \cup \{0\}$,

$$\sum_{i=0}^{\ell} |T_{(i)}| \geq \frac{\ell+1}{2n+1} \cdot |\mathcal{U}_k| \geq \frac{\ell+1}{3n} \cdot |\mathcal{U}_k|. \tag{3.15}$$

We will construct our family of one-inclusion graphs such that for a substantial fraction of the $\mathcal{X}_S$ in V_1 , the corresponding one-inclusion strategy incurs large error. We now construct the edge set of our (undirected) bipartite graph, $G = (V_1, V_2, E)$. Define for all $v \in V_1 \cup V_2$,

$$N(v) = \{u \in V_1 \cup V_2 : \{u, v\} \in E\},$$

that is, the set of neighbors of v in G . We will use the probabilistic method to construct an edge set, E , satisfying certain cardinality constraints. The first constraint ensures that the family of one-inclusion graphs we construct from G satisfies the appropriate out-degree constraints while the second will be used to show that they incur large error.

Lemma 3.4.7. *There exists an edge set E such that the bipartite graph $G = (V_1, V_2, E)$ satisfies:*

1. *for all $v \in V_1$ and $v' \in V_2$, if $\{v, v'\} \in E$ then $v \prec v'$;*
2. *for all $v' \in V_2$, we have $|N(v')| \leq d$;*
3. *$|\{v \in V_1 : |N(v)| \geq d/(8\delta)\}| \geq \frac{3}{4}|V_1|$.*

Proof. Our proof will utilize the probabilistic method. Defining $M(v') = \{v \in V_1 : v \prec v'\}$ for all $v' \in V_2$, E is constructed according to the following random process. For any $v' \in V_2$:

1. With $\ell = \min\{d, |M(v')|\}$, pick $\{v_i\}_{i \in [\ell]}$ uniformly at random without replacement from $M(v')$.
2. Add all edges $\{\{v', v_i\}\}_{i \in [\ell]}$ to E .

By construction, every added edge satisfies $v_i \prec v'$, and each $v' \in V_2$ receives exactly $\ell \leq d$ neighbors, proving the first two claims. For the third claim, fix $v \in V_1$ and define

$$Q(v) = \{v' \in V_2 : v \prec v'\}.$$

For $v' \in V_2$, set $p_{v'} := \min\{d, |M(v')|\} / \max\{1, |M(v')|\}$. We have by the above definition and linearity of expectation

$$\mathbb{E}[|N(v)|] = \sum_{v' \in Q(v)} p_{v'}.$$

Moreover, since the neighbor-selection is performed independently across different $v' \in V_2$, the indicators $\mathbf{1}\{v' \in N(v)\}$ are independent over $v' \in Q(v)$. From the independence of the indicators $\mathbf{1}\{v' \in N(v)\}$,

$$\text{Var}(|N(v)|) = \text{Var}\left(\sum_{v' \in Q(v)} \mathbf{1}\{v' \in N(v)\}\right) = \sum_{v' \in Q(v)} \text{Var}(\mathbf{1}\{v' \in N(v)\}) \leq \sum_{v' \in Q(v)} p_{v'}.$$

By [Lemma 3.4.5](#) and the definition of V_1 , we have $|M(v')| \leq 4\lceil \delta n \rceil$ for all $v' \in V_2$. When c_1 ([Theorem 3.4.3](#)) is large enough, we have $4\lceil \delta n \rceil \leq 6\delta n$ and $n\delta \geq c_1 d$, which yields

$$p_{v'} \geq \frac{1}{6} \cdot \frac{d}{\delta n}.$$

We will only apply the following bounds for $k \in [n]$, in which case $|Q(v)| = 2n - k \geq n$. Additionally, we have when c_2 ([Theorem 3.4.3](#)) is small enough and using the fact that $|Q(v)| \geq n$,

$$\frac{\sqrt{\text{Var}(|N(v)|)}}{\mathbb{E}[|N(v)|]} \leq \frac{\sqrt{\sum_{v' \in Q(v)} p_{v'}}}{\sum_{v' \in Q(v)} p_{v'}} = \sqrt{\frac{1}{\sum_{v' \in Q(v)} p_{v'}}} \leq \sqrt{\frac{6\delta n}{d|Q(v)|}} \leq \sqrt{\frac{6\delta}{d}} \leq \frac{1}{10}.$$

This yields via Chebyshev's inequality,

$$\Pr \left[|N(v)| \geq \frac{d}{8\delta} \right] \geq \Pr \left[|N(v)| \geq \frac{|Q(v)|}{8} \cdot \frac{d}{\delta n} \right] \geq \Pr \left[|N(v)| \geq \frac{3}{4} \sum_{v' \in Q(v)} p_{v'} \right] \geq \frac{3}{4}.$$

Taking expectation over $v \in V_1$ and applying the probabilistic method establishes the third claim. $\square$

With the bipartite graph, $G = (V_1, V_2, E)$, defined in [VERT-SETS](#) and [Lemma 3.4.7](#), we will now construct our family of one-inclusion graphs for the sets in $\mathcal{U}_{k+1}$.

We think of an element of $U \in \mathcal{U}_{k+1}$ as the union of a possible $\mathcal{X}_S$ and a new test point x_{n+1} . Informally, for any set $U \in \mathcal{U}_{k+1}$, we pick any low out-degree orientation, and then reorient the edges of the center vertex f_0 . We prove that for most of the k -sized sets $\mathcal{X}_S \in V_1$ and a significant fraction of their “extensions” to $(k+1)$ -sized sets $U \in V_2$, the one-inclusion graph algorithm predicts incorrectly on $x_{n+1} \in U \setminus \mathcal{X}_S$.

Before stating the construction formally, we clarify notation in [Algorithm 1](#). For each $U \subseteq \mathcal{W}_m$, the one-inclusion graph $G(\mathcal{H}|_U)$ has vertex set $\mathcal{H}|_U$. We will abuse notation and write $f_0, f_i \in \mathcal{H}|_U$ for the vertices corresponding to the restrictions of the star witnesses f_0, f_i to U . In particular, if $x_i \in U$, then f_0 and f_i are distinct vertices in $\mathcal{H}|_U$ and are adjacent in $G(\mathcal{H}|_U)$; we denote this undirected edge by $\{f_0, f_i\}$. An orientation σ_U maps each edge to one of its endpoints, and writing $\sigma_U(\{f_0, f_i\}) = f_i$ means the edge is oriented from f_0 to f_i .

Our construction is formally described in [Algorithm 1](#). Note that [Algorithm 1](#) only constructs an orientation rule for V_2 (which corresponds to $\mathcal{U}_{k+1}$ in this context).

Algorithm 1 Constructing sub-optimal one-inclusion graphs

Input: Bipartite Graph $G = (V_1, V_2, E)$.

Output: Orientation rule σ for V_2 .

- 1: Set $\sigma = \{\}$.
 - 2: **for** $U \in V_2$ **do**
 - 3: Pick any orientation σ_U of $G(\mathcal{H}|_U)$ that has max out-degree at most d .
 - 4: For every $\mathcal{X}_S \in N(U)$, let $i = i(U, \mathcal{X}_S)$ be the unique index such that $U \setminus \mathcal{X}_S = \{x_i\}$ (well-defined since $\mathcal{X}_S \prec U$), and reorient the edge $\{f_0, f_i\}$ so that $\sigma_U(\{f_0, f_i\}) = f_i$.
 - 5: For every edge $\{f_0, g\}$ of $G(\mathcal{H}|_U)$ incident to f_0 that was not modified in the previous step, set $\sigma_U(\{f_0, g\}) = f_0$.
 - 6: Add σ_U to σ .
 - 7: **end for**
 - 8: **return** σ .
-

Before we proceed, we first show that the orientation rule constructed in [Algorithm 1](#) is valid. For this, we require a beautiful result of Haussler, Littlestone, and Warmuth that shows that there is always a valid one-inclusion graph algorithm for any hypothesis class with finite VC dimension [\[42\]](#).

Theorem 3.4.8. *[42, Theorem 2.2] For any hypothesis class $\mathcal{H}$ over instance space $\mathcal{X}$ with VC dimension at most d and any finite $U \subseteq \mathcal{X}$, there exists an orientation σ_U of $G(\mathcal{H}|_U)$ with*

$$\text{out}(\sigma_U) \leq d.$$

Lemma 3.4.9. *When G satisfies the conclusion of [Lemma 3.4.7](#) the orientation rule for V_2 , σ , constructed in [Algorithm 1](#) satisfies*

$$\max_{\sigma_U \in \sigma} \text{out}(\sigma_U) \leq d + 1.$$

Remark 3.4.10. For the sake of generality, we prove [Lemma 3.4.9](#) with an upper bound of $d + 1$. If we consider more structured classes we can improve this bound to d , which would match the out-degree bound of [Theorem 3.4.8](#). For example, if we take the class of functions that take on the value 1 at most d times, we can take the original orientation in [Algorithm 1](#) to be the one that orients all edges towards the lower hamming weight vertex (akin to the orientation of the OIG that corresponds to the Closure algorithm; see the discussion in [\[90\]](#)). It is easy to see that this strategy originally has maximum out-degree d and that the modification to the orientation applied in [Algorithm 1](#) does not change the maximum out-degree.

Remark 3.4.11. The bound $d + 1$ on the max out-degree only slightly changes the risk bound for the one-inclusion graph algorithm. In particular, our orientation rule gives

$$\mathbb{E} \operatorname{err}_P(\widehat{f}_{\text{OIG}}) \leq \frac{d+1}{n+1}, \quad \text{and by Markov's inequality:} \quad \operatorname{err}_P(\widehat{f}_{\text{OIG}}) \leq \frac{d+1}{(n+1)\delta},$$

with probability at least $1 - \delta$. Note that by Lemma 3.4.6, the same out-degree bound holds for the induced orientation on the one-inclusion graph of any $(n+1)$ -sample with repetitions whose set of distinct instances is U .

Proof of Lemma 3.4.9. Let $U \in \mathcal{U}_{k+1}$. By construction, the only edges oriented *out of* f_0 in σ_U are those reoriented in Line 3 and below for neighbors $\mathcal{X}_S \in N(U)$; all other edges incident to f_0 are oriented into f_0 . Hence the out-degree of f_0 is exactly $|N(U)| \leq d$ by Lemma 3.4.7. For any vertex $g \in \mathcal{H}|_U$ with $g \neq f_0$, its out-degree is at most d in the original orientation σ_U selected in Line 3 of Algorithm 1 (such an orientation exists by Theorem 3.4.8). The only possible increase in the out-degree of g comes from the orientation of its (unique, if it exists) edge to f_0 in Line 5, so $\operatorname{out}(g)$ increases by at most 1. $\square$

Our last lemma will show that the one-inclusion strategy defined by the orientation rule σ constructed in Algorithm 1 incurs large error for most $\mathcal{X}_S$ in V_1 . In fact, this will be shown for the subset whose existence was established in Lemma 3.4.7. Formally, define W_1^k as

$$W_1^k = \left\{ v \in V_1 : |N(v)| \geq \frac{d}{8\delta} \right\}.$$

The following lemma shows that any training sample S with $\mathcal{X}_S$ in W_1^k incurs large error.

Lemma 3.4.12. *For any training sample S with $\mathcal{X}_S \in W_1^k$, the one-inclusion graph algorithm that predicts using the orientation rule σ defined by Algorithm 1 on input G satisfying the conclusion of Lemma 3.4.7 has*

$$\operatorname{err}_P(\widehat{f}_{\sigma,S}) \geq \frac{1}{16} \cdot \frac{d}{\delta n}.$$

Proof. Fix S with $\mathcal{X}_S \in W_1^k$, and let $x \in \mathcal{W}_m$ be such that $\mathcal{X}_S \cup \{x\} \in N(\mathcal{X}_S)$. By the first claim of Lemma 3.4.7, $\mathcal{X}_S \cup \{x\} \in N(\mathcal{X}_S)$ implies $\mathcal{X}_S \prec \mathcal{X}_S \cup \{x\}$, hence $x \notin \mathcal{X}_S$. In particular, x does not appear among the training instances, so the test point is not a duplicate of any training example.

Let $U := \mathcal{X}_S \cup \{x\} \in V_2$ and let i be the unique index with $x = x_i$. Since the labels are generated by $f^* = f_0$, the labeled sample S assigns label $f_0(z)$ to each $z \in \mathcal{X}_S$. By the star-set property, f_i and f_0 agree on every point of $\mathcal{X}_S$ and differ only on x , so both f_0 and f_i are consistent with the labeled training sample S when viewed on the instance set U . Thus, when the one-inclusion rule is applied to the augmented sample (S, x) , the relevant one-inclusion edge is exactly the edge $\{f_0, f_i\}$ (the two consistent vertices differ on the single point x). By construction of Algorithm 1, because $\mathcal{X}_S \in N(U)$ we have oriented this edge

away from f_0 , i.e., $\sigma_U(\{f_0, f_i\}) = f_i$. Therefore, the one-inclusion predictor outputs the label $f_i(x)$, which satisfies $f_i(x) \neq f_0(x)$, and hence makes an error at x . Moreover, since $x \notin \mathcal{X}_S$, the point x appears exactly once in the $(n+1)$ -sample (S, x) , so the edge $\{f_0, f_i\}$ is present in the one-inclusion graph induced by the full ordered sample; by Lemma 3.4.6, the induced orientation on (S, x) coincides with σ_U on this edge. Thus the predictor errs on every such x . Therefore,

$$\text{err}_P(\hat{f}_{\sigma, S}) \geq \frac{|N(\mathcal{X}_S)|}{2n} \geq \frac{1}{16} \cdot \frac{d}{\delta n}.$$

□

We will now prove Theorem 3.4.3. Recall that our distribution $P = P(n, \delta)$ is the uniform distribution over the participating elements of a star set of size $2n$ (with labels generated by $f^* = f_0$).

For each $k \in [n]$ denoting the possible number of unique observed elements of the star set $\mathcal{W}_m$, consider the family of one-inclusion prediction strategies constructed in Algorithm 1 with the bipartite graph defined in VERT-SETS and Lemma 3.4.7. By Lemma 3.4.12, any S with $\mathcal{X}_S \in W_1^k$ incurs large error. Hence, we only need to lower bound the probability of observing a training sample S with $\mathcal{X}_S \in W_1 := \cup_{k \in [n]} W_1^k$. By symmetry of the uniform distribution over $\mathcal{W}_m$, conditioned on $|\mathcal{X}_S| = k$, the random set $\mathcal{X}_S$ is uniformly distributed over $\mathcal{U}_k$. Hence, we have by Lemma 3.4.7 and (3.15) that

$$\Pr_{S \sim P^n} [\mathcal{X}_S \in W_1^k \mid |\mathcal{X}_S| = k] = \frac{|W_1^k|}{|\mathcal{U}_k|} \geq \frac{3}{4} \cdot \frac{|V_1|}{|\mathcal{U}_k|} = \frac{3}{4} \cdot \sum_{i=0}^{4\lceil \delta n \rceil - 1} \frac{|T_{(i)}|}{|\mathcal{U}_k|} \geq \frac{3}{4|\mathcal{U}_k|} \cdot \frac{4\lceil \delta n \rceil |\mathcal{U}_k|}{3n} \geq \delta.$$

To conclude the proof of the theorem, we now have

$$\Pr_{S \sim P^n} [\mathcal{X}_S \in W_1] = \sum_{k=1}^n \Pr_{S \sim P^n} [\mathcal{X}_S \in W_1^k \mid |\mathcal{X}_S| = k] \cdot \Pr_{S \sim P^n} [|\mathcal{X}_S| = k] \geq \delta.$$

3.5 Proofs from Section 3.1: Classification

In this section we provide relevant details on the one-inclusion hypergraph algorithm [42, 74]. We also provide proofs for the results claimed in Section 3.1.

We begin with some preliminary definitions for hypergraphs. We define an *orientation* of hypergraph $G = (V, E)$ to be a function $\sigma_G : E \rightarrow V$ such that for any $e \in E$, $\sigma_G(e) \in e$. Given a hypergraph $G = (V, E)$ and orientation σ_G , define the *out-degree* of a vertex $v \in V$ to be

$$\text{out}(v; \sigma_G) = |\{e \in E : v \in e, \sigma_G(e) \neq v\}|.$$

For a hypergraph G oriented using σ_G , the *max out-degree* of G is defined to be $\text{out}(\sigma_G) = \max_{v \in V} \text{out}(v; \sigma_G)$.

Algorithm 2 One-inclusion hypergraph algorithm.

Inputs: Training sample S .

Output: Hypothesis $\hat{f}_{\text{OIG}}(\cdot; S) : \mathcal{X} \rightarrow \mathcal{Y}$.

For any point $x \in \mathcal{X}$, $\hat{f}_{\text{OIG}}(\cdot; S)$ predicts as follows:

- 1: Build the hypergraph $G(\mathcal{H}|_{\mathcal{X}_S \cup \{x\}})$ and pick an orientation $\sigma_{\mathcal{X}_S \cup \{x\}}$ *minimizing* $\text{out}(\sigma_{\mathcal{X}_S \cup \{x\}})$.
 - 2: If there is a unique label y for x consistent with S and $\mathcal{H}$, predict y .
 - 3: Otherwise, let e be the edge in $G(\mathcal{H}|_{\mathcal{X}_S \cup \{x\}})$ with hypotheses consistent with S but not on x .
 - 4: Predict according to $\sigma_{\mathcal{X}_S \cup \{x\}}(e)$, i.e., the hypothesis pointed to in the orientation of e .
-

We are now ready to describe the one-inclusion hypergraph algorithm as pseudocode in [Algorithm 2](#). The one-inclusion (hyper)graph algorithm [\[42, 74\]](#) was originally defined for the *transductive* model of learning, i.e., it predicts a label for a single test point x . The algorithm as presented above is the standard extension that defines a predictor for the entire domain $\mathcal{X}$. When predicting on a test point x , the algorithm minimizes $\text{out}(G(\mathcal{H}|_{\mathcal{X}_S \cup \{x\}}))$ which upper bounds its leave-one-out error ([Lemma 3.5.2](#)). Before we explore the leave-one-out error of this algorithm, we connect $\text{out}(G(\mathcal{H}|_{\mathcal{X}_S \cup \{x\}}))$ to $\mu_{\mathcal{H}}(n)$.

Every hypergraph G has an orientation σ_G for which the max out-degree $\text{out}(\sigma_G)$ is at most $\lceil \mu(G) \rceil$. This is a standard fact for graphs (e.g., see [\[9, Lemma 3.1\]](#), [\[42, Theorem 2.2\]](#)). We expect the generalized statement for hypergraphs to also be known, but since we could not find an explicit reference, we include a proof based on a maximum flow argument generalizing a known proof for the graph case [\[42\]](#).

Lemma 3.5.1. *For any hypergraph $G = (V, E)$ and any integer $d \geq \mu(G)$, there is an orientation σ_G such that the max out-degree satisfies $\text{out}(\sigma_G) \leq d$.*

Proof. We build an auxiliary graph that contains a source s , a sink t , and two vertex sets V_1 and V_2 between s and t . V_1 contains a vertex for each edge in E , i.e., $V_1 = E$. V_2 is equal to the vertex set of G , i.e., $V_2 = V$. There is a weighted and directed edge from s to every $e \in V_1$. For each $e \in V_1$, the weight of the edge to e from s is equal to $|e| - 1$. For every $e \in V_1$, there is a weighted directed edge from e to every $v \in V_1$ that participates in the hyperedge e in G . Each such edge has weight 1. Finally, there is an edge from every $v \in V_2$ to t , each with weight d . We will prove that the maximum flow of this graph from s to t is equal to $m := \sum_{e \in V_1} |e| - 1$.

Since all the weights are integers, there exists an integral maximum flow. Consider such a flow. There is a simple way to obtain the desired orientation of G from this flow. Fix an $e \in V_1$. Since e gets $|e| - 1$ units of flow, $|e| - 1$ of its $|e|$ neighbours gets a unit of flow. Thus, there is a single neighbour $v \in V_2$ of e that does not get a unit of flow, and we orient the edge e in the original hypergraph G towards this v . Repeat this for every $e \in V_1$. Every unit of flow some $v \in V_2$ receives corresponds to an edge in G that it is not selected to be the head of. Since the capacity of the edge out of every $v \in V_2$ to t is d , we can conclude that the orientation satisfies that the out-degree of every $v \in V$ is at most d .

We now prove the maximum flow is m . It is easy to see the flow is at most m since the sum of the edge weights from s to V_1 is equal to m . It remains to prove the maximum flow is at least m . Recall that the maximum flow is equal to the minimum cut (set of edges) that separates s from t , so it suffices to prove a lower bound for the minimum cut.

The following is a useful observation we will make use of. Notice that every path from s to t consists of 3 edges passing through some $e \in V_1$ and $v \in V_2$. Fix one such path. Since we consider a minimum cut, if the first edge (connecting s to e) is cut, there is no need to cut the edge between e and v since this would only increase the cut value. Similarly, if the third edge is cut (connecting v to t) there is no need to cut the edge between e and v .

Let $U \subseteq V_1$ be the set of vertices whose edge from s are *not* cut. Let $W \subseteq V_2$ be the set of vertices whose edges to t are cut. Fix an $e \in U$. There are $|e|$ paths from e to t and each of these must be cut in either the middle layer of the edges or the third layer of edges, but not both. Let $W_{e,2} \subseteq V_2$ be the set of vertices for which the paths from e to t are cut in the second layer and let $W_{e,3} \subseteq V_2$ be the set of vertices for which the paths from e to t are cut in the third layer. Clearly $W_{e,3} \subseteq W$. Since we consider the minimum cut, we can conclude that $W_{e,3} \cap W_{e,2} = \emptyset$ from our earlier observation.

We now upper bound the sum of the weight of the edges from s to U . Write this sum as $a = \sum_{e \in U} |e| - 1$ which can be written $\sum_{e \in U} |e \cap W_{e,3}| + |e \cap W_{e,2}| - 1$. Let $W' = \cup_{e \in U} W_{e,3}$ and notice that

$$\sum_{e \in U} |e \cap W'| - 1 \leq \sum_{e \in V_1} |e \cap W'| - 1 \leq \mu(G)|W'| \leq d|W'|.$$

Thus, $a \leq d|W'| + \sum_{e \in V_1} |e \cap W_{e,2}|$, so the sum of the weights of edges in the first layer that are cut is at least $m - a \geq m - d|W'| - \sum_{e \in V_1} |e \cap W_{e,2}|$. Since the weight of every edge in the second layer is 1, the sum of weights of edges in the second layer that are cut can be written as $\sum_{e \in U} |e \cap W_{e,2}|$ as a result of the earlier observation. Finally, since every edge in the final layer has weight d , the total weight of the cut edges in the third layer is equal to $d|W|$. Since $W' \subseteq W$ we can conclude that the total weight of cut edges is at least

$$m - d|W'| - \sum_{e \in U} |e \cap W_{e,2}| + d|W| + \sum_{e \in U} |e \cap W_{e,2}| \geq m.$$

Thus, the maximum flow is indeed m . □

We can now use [Lemma 3.5.1](#) to prove a bound on the LOO error of the one-inclusion graph algorithm.

Lemma 3.5.2. *Fix a hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$. The leave-one-out error of the one-inclusion graph algorithm $\widehat{f}_{\text{OIG}} : \mathcal{X} \times \mathcal{U} \rightarrow \mathcal{Y}$ on any realizable training sample $S = ((x_1, y_1), \dots, (x_n, y_n))$ satisfies*

$$\sum_{i=1}^n \mathbf{1} \left\{ \widehat{f}_{\text{OIG}}(x_i; S^{-i}) \neq y_i \right\} \leq \lceil \mu_{\mathcal{H}}(n) \rceil.$$

Proof. Let $U = \{x_1, \dots, x_n\}$, $G(\mathcal{H}|_U) = (V, E)$ be the one-inclusion hypergraph built using U , and σ_U be an orientation that minimizes $\text{out}(G(\mathcal{H}|_U))$. Furthermore, let $f^* \in V$ be the hypothesis that labels all points in U correctly, i.e., $f^*(i) = y_i$ for all $i \in [n]$. Notice that Algorithm 2 builds the same one-inclusion hypergraph and orients it the same way for every choice of held out x_i . So, the number of times the predictor makes a mistake is equal to the number of times it does not select f^* . We have

$$\sum_{i=1}^n \mathbf{1} \left\{ \widehat{f}_{\text{OIG}}(x_i; S^{-i}) \neq y_i \right\} = \sum_{i=1}^n \mathbf{1} \left\{ \sigma_U(e_{i,f^*}) \neq f^* \right\} = \text{out}(f^*; \sigma_U) \leq \lceil \mu_{\mathcal{H}}(n) \rceil,$$

where the inequality follows from Lemma 3.5.1. $\square$

For a set of hypotheses $\{f_1, \dots, f_t\} \subseteq \mathcal{H}$, we define their plurality vote $\text{Maj}(f_1, \dots, f_t)$ to be the function that, on input $x \in \mathcal{X}$, outputs the label $y \in \mathcal{Y}$ that is most common out of $f_1(x), \dots, f_t(x)$. We now prove our main multiclass PAC bound. We restate Theorem 3.1.7 below before its proof for ease of reading.

Theorem 3.1.7. *Fix a hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$. There is a predictor $\widehat{f} : \mathcal{X} \times \mathcal{U} \rightarrow \mathcal{Y}$ which, for any $\delta \in (0, 1)$, and $S \sim P^n$ sampled from any realizable distribution P , satisfies*

$$\text{err}_P(\widehat{f}(\cdot; S)) \leq 9.64 \left(\frac{\lceil \mu_{\mathcal{H}}(n) \rceil}{n} + \frac{1}{n} \log \left(\frac{2}{\delta} \right) \right),$$

with probability at least $1 - \delta$ over the randomness of S .

Proof. Given the full training sample S , let $(\widehat{f}_{\text{OIG}}(\cdot; S_{\leq t}))_{t=1}^{n-1}$ be the one-inclusion hypergraph predictors trained on the sequence of growing training samples $(S_{\leq t})_{t=1}^{n-1}$. Because these predictors are symmetric in their training sample and they have a leave-one-out error guarantee (Lemma 3.5.2), we can apply Theorem 3.1.4 to the suffix of the sequence to get that

$$\frac{4}{3n} \sum_{t=n/4}^{n-1} \text{err}_P(\widehat{f}_{\text{OIG}}(\cdot; S_{\leq t})) \leq 4.82 \left(\frac{\lceil \mu_{\mathcal{H}}(n) \rceil}{n} + \frac{1}{n} \log \left(\frac{2}{\delta} \right) \right),$$

with probability at least $1 - \delta$ over the randomness of S . Whenever this bound holds, the plurality vote of the one-inclusion hypergraph predictors in the suffix, $\widehat{f}(\cdot) = \text{Maj}(\widehat{f}_{\text{OIG}}(\cdot; S_{\leq \frac{n}{4}}), \dots, \widehat{f}_{\text{OIG}}(\cdot; S_{\leq n}))$

satisfies

$$\begin{aligned}
 \text{err}_P(\hat{f}) &= \mathbb{E}_{(X,Y) \sim P} [\mathbf{1}\{\hat{f}(X) \neq Y\}] \\
 &\leq \mathbb{E}_{(X,Y) \sim P} \left[2 \left(\frac{4}{3n} \sum_{t=n/4}^{n-1} \mathbf{1}\{\hat{f}_{\text{OIG}}(X; S_{\leq t}) \neq Y\} \right) \right] \\
 &= 2 \cdot \frac{4}{3n} \sum_{t=n/4}^{n-1} \text{err}_P(\hat{f}_{\text{OIG}}(\cdot; S_{\leq t})) \\
 &\leq 9.64 \left(\frac{\lceil \mu_{\mathcal{H}}(n) \rceil}{n} + \frac{1}{n} \log \left(\frac{2}{\delta} \right) \right).
 \end{aligned}$$

In the first inequality above, we used the fact that no more than half the predictors can be correct when the majority vote is wrong. This completes the proof. $\square$

The remaining results of this section lead to the proof of [Theorem 3.1.8](#).

Theorem 3.5.3. *Let $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ be a sample space, label space, and hypothesis class, respectively. Suppose, there exists a learning algorithm $\hat{f}(\cdot; \cdot)$ which satisfies for any realizable distribution P over $\mathcal{X} \times \mathcal{Y}$ and $n \in \mathbb{N}$:*

$$\Pr_{S \sim P^n} \left[\text{err}_P(\hat{f}(\cdot; S)) \leq B_n \right] \geq 0.9.$$

Then, there exists a predictor, $\hat{g}$, which satisfies for any $n, m > 4$, realizable set $T = ((x_i, y_i))_{i=1}^m$:

$$\sum_{i=1}^m \ell(\hat{g}(x_i, T^{-i}), y_i) \leq 1 + 30 \begin{cases} mB_n & \text{if } m \geq 2n \\ nB_n & \text{otherwise} \end{cases}.$$

Proof. For a realizable set $T = ((x_i, y_i))_{i=1}^m$, we define the distributions, D_T , where for $m < 2n$, D_T is defined by a canonical ordering over the transductive inputs $(x_i)_{i=1}^m$:

$$D_T = \begin{cases} \text{Uniform}((x_i, y_i)_{i=1}^m) & \text{if } m \geq 2n \\ \left(1 - \frac{m}{4n}\right) \delta((x_1, y_1)) + \frac{m}{4n} \text{Uniform}((x_i, y_i)_{i=1}^m) & \text{otherwise.} \end{cases}$$

We now define our transductive predictor as follows:

$$\hat{g}(x_i; T^{-i}) := \arg \max_{y \in \mathcal{Y}} \mathbb{E}_{S \sim D_T^n | (x_i, y_i) \notin S} \left[\mathbf{1}\{\hat{f}(x_i; S) = y\} \right].$$

Observe now from our definition of D_T and our guarantees on $\hat{f}$:

$$\begin{aligned}
 \forall i \neq 1 : \Pr_{S \sim D_T^n} [(x_i, y_i) \notin S] &\geq \frac{1}{2} \\
 \Pr_{S \sim D_T^n} \left[\text{err}_{D_T}(\hat{f}(\cdot; S)) \leq B_n \right] &\geq \frac{9}{10}.
 \end{aligned}$$

Now, through the following observations:

$$\forall a' \geq a, b, \mathbf{1}\{a \geq b\} \leq \mathbf{1}\{a' \geq b\} \quad \text{and} \quad \forall a \geq 0, b > 0, \mathbf{1}\{a \geq b\} \leq \frac{a}{b},$$

and a repeated application of Bayes' rule, we have for any $i \neq 1$;

$$\begin{aligned} & \mathbf{1}\{\widehat{g}(x_i; T^{-i}) \neq y_i\} \\ & \leq \mathbf{1}\left\{ \Pr_{S \sim D_T^n | (x_i, y_i) \notin S} \left[\widehat{f}(x_i; S) \neq y_i \right] \geq \frac{1}{2} \right\} \\ & = \mathbf{1}\left\{ \Pr_{S \sim D_T^n | (x_i, y_i) \notin S} \left[\widehat{f}(x_i; S) \neq y_i \text{ and } \text{err}_{D_T}(\widehat{f}(\cdot; S)) \leq B_n \right] \right. \\ & \quad \left. + \Pr_{S \sim D_T^n | (x_i, y_i) \notin S} \left[\widehat{f}(x_i; S) \neq y_i \text{ and } \text{err}_{D_T}(\widehat{f}(\cdot; S)) > B_n \right] \geq \frac{1}{2} \right\} \\ & \leq \mathbf{1}\left\{ \Pr_{S \sim D_T^n | (x_i, y_i) \notin S} \left[\widehat{f}(x_i; S) \neq y_i \text{ and } \text{err}_{D_T}(\widehat{f}(\cdot; S)) \leq B_n \right] \right. \\ & \quad \left. + \Pr_{S \sim D_T^n | (x_i, y_i) \notin S} \left[\text{err}_{D_T}(\widehat{f}(\cdot; S)) > B_n \right] \geq \frac{1}{2} \right\} \\ & \leq \mathbf{1}\left\{ \Pr_{S \sim D_T^n | (x_i, y_i) \notin S} \left[\widehat{f}(x_i; S) \neq y_i \text{ and } \text{err}_{D_T}(\widehat{f}(\cdot; S)) \leq B_n \right] + \frac{\Pr_{S \sim D_T^n} \left[\text{err}_{D_T}(\widehat{f}(\cdot; S)) > B_n \right]}{\Pr_{S \sim D_T^n} [(x_i, y_i) \notin S]} \geq \frac{1}{2} \right\} \\ & \leq \mathbf{1}\left\{ \Pr_{S \sim D_T^n | (x_i, y_i) \notin S} \left[\widehat{f}(x_i; S) \neq y_i \text{ and } \text{err}_{D_T}(\widehat{f}(\cdot; S)) \leq B_n \right] + \frac{1}{5} \geq \frac{1}{2} \right\} \\ & \leq \mathbf{1}\left\{ \Pr_{S \sim D_T^n | (x_i, y_i) \notin S} \left[\widehat{f}(x_i; S) \neq y_i \text{ and } \text{err}_{D_T}(\widehat{f}(\cdot; S)) \leq B_n \right] \geq \frac{3}{10} \right\} \\ & \leq \frac{10}{3} \cdot \Pr_{S \sim D_T^n | (x_i, y_i) \notin S} \left[\widehat{f}(x_i; S) \neq y_i \text{ and } \text{err}_{D_T}(\widehat{f}(\cdot; S)) \leq B_n \right] \\ & \leq \frac{10}{3} \cdot \frac{\Pr_{S \sim D_T^n} \left[\widehat{f}(x_i; S) \neq y_i \text{ and } \text{err}_{D_T}(\widehat{f}(\cdot; S)) \leq B_n \right]}{\Pr_{S \sim D_T^n} [(x_i, y_i) \notin S]} \\ & \leq \frac{20}{3} \cdot \Pr_{S \sim D_T^n} \left[\widehat{f}(x_i; S) \neq y_i \text{ and } \text{err}_{D_T}(\widehat{f}(\cdot; S)) \leq B_n \right] \\ & = \frac{20}{3} \cdot \Pr_{S \sim D_T^n} \left[\text{err}_{D_T}(\widehat{f}(\cdot; S)) \leq B_n \right] \cdot \Pr_{S \sim D_T^n} \left[\widehat{f}(x_i; S) \neq y_i \mid \text{err}_{D_T}(\widehat{f}(\cdot; S)) \leq B_n \right] \\ & \leq \frac{20}{3} \cdot \Pr_{S \sim D_T^n} \left[\widehat{f}(x_i; S) \neq y_i \mid \text{err}_{D_T}(\widehat{f}(\cdot; S)) \leq B_n \right]. \end{aligned}$$

We conclude the proof in two cases:

Case 1: $m \geq 2n$. In this case, we get:

$$\begin{aligned}
 \sum_{i=2}^m \mathbf{1} \{ \hat{g}(x_i; T^{-i}) \neq y_i \} &\leq \frac{20}{3} \sum_{i=2}^m \mathbf{Pr}_{S \sim D_T^n} \left[\hat{f}(x_i; S) \neq y_i \mid \text{err}_{D_T}(\hat{f}(\cdot; S)) \leq B_n \right] \\
 &\leq \frac{20}{3} \sum_{i=1}^m \mathbf{Pr}_{S \sim D_T^n} \left[\hat{f}(x_i; S) \neq y_i \mid \text{err}_{D_T}(\hat{f}(\cdot; S)) \leq B_n \right] \\
 &= \frac{20}{3} \sum_{i=1}^m \mathbb{E}_{S \sim D_T^n} \left[\mathbf{1} \{ \hat{f}(x_i; S) \neq y_i \} \mid \text{err}_{D_T}(\hat{f}(\cdot; S)) \leq B_n \right] \\
 &= \frac{20}{3} \mathbb{E}_{S \sim D_T^n} \left[m \cdot \text{err}_{D_T}(\hat{f}(\cdot; S)) \mid \text{err}_{D_T}(\hat{f}(\cdot; S)) \leq B_n \right] \leq \frac{20}{3} \cdot m B_n.
 \end{aligned}$$

Case 2: $m < 2n$. In this case, we get:

$$\begin{aligned}
 \sum_{i=2}^m \mathbf{1} \{ \hat{g}(x_i; T^{-i}) \neq y_i \} &\leq \frac{20}{3} \sum_{i=2}^m \mathbf{Pr}_{S \sim D_T^n} \left[\hat{f}(x_i; S) \neq y_i \mid \text{err}_{D_T}(\hat{f}(\cdot; S)) \leq B_n \right] \\
 &\leq \frac{20}{3} \sum_{i=1}^m \mathbf{Pr}_{S \sim D_T^n} \left[\hat{f}(x_i; S) \neq y_i \mid \text{err}_{D_T}(\hat{f}(\cdot; S)) \leq B_n \right] \\
 &= \frac{20}{3} \sum_{i=1}^m \mathbb{E}_{S \sim D_T^n} \left[\mathbf{1} \{ \hat{f}(x_i; S) \neq y_i \} \mid \text{err}_{D_T}(\hat{f}(\cdot; S)) \leq B_n \right] \\
 &\leq \frac{20}{3} \mathbb{E}_{S \sim D_T^n} \left[4n \cdot \text{err}_{D_T}(\hat{f}(\cdot; S)) \mid \text{err}_{D_T}(\hat{f}(\cdot; S)) \leq B_n \right] \leq \frac{80}{3} \cdot n B_n
 \end{aligned}$$

where the last line follows from the observation for any predictor $\hat{f}$:

$$\begin{aligned}
 \text{err}_{D_T}(\hat{f}(\cdot; S)) &= \left(1 - \frac{m}{4n}\right) \mathbf{1} \{ \hat{f}(x_1; S) \neq y_1 \} + \frac{m}{4n} \mathbb{E}_{(X, Y) \sim \text{Uniform}((x_i, y_i)_{i=1}^m)} \left[\mathbf{1} \{ \hat{f}(X; S) \neq Y \} \right] \\
 &\geq \frac{m}{4n} \sum_{i=1}^m \frac{1}{m} \mathbf{1} \{ \hat{f}(x_i; S) \neq y_i \} = \frac{1}{4n} \sum_{i=1}^m \mathbf{1} \{ \hat{f}(x_i; S) \neq y_i \}
 \end{aligned}$$

by ignoring the first term in the equality. Observing the inequality

$$\mathbf{1} \{ \hat{g}(x_i; T^{-i}) \neq y_i \} \leq 1,$$

concludes the proof. □

We now include a result of [29].

Theorem 3.5.4 ([29]). *For every class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$:*

$$\frac{1}{2} \frac{\mu_{\mathcal{H}}(m)}{m} \leq \varepsilon_{\mathcal{H}}(m) \leq \frac{\mu_{\mathcal{H}}(m)}{m}.$$

Theorems 3.5.3 and 3.5.4, yield the following corollary:

Corollary 3.5.5. *There exists an absolute constant $C > 0$ such that the following holds. Let $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ be a sample space, label space, and hypothesis class, respectively. Suppose further that $\mu_{\mathcal{H}}(\ell) > 4$ for some $\ell \in \mathbb{N}$. Then, for any learning algorithm $\hat{f}(\cdot; \cdot)$ and all $n \geq \ell$, there exists a realizable distribution P such that:*

$$\Pr_{S \sim P^n} \left[\text{err}_P \left(\hat{f}(\cdot; S) \right) \geq C \sup_{k \leq 2n} \frac{\mu_{\mathcal{H}}(k)}{n} \right] \geq 0.1.$$

Proof. We have from Theorems 3.5.3 and 3.5.4:

$$1 + 30nB_n \geq \sup_{k \leq 2n} \frac{\mu_{\mathcal{H}}(k)}{2}.$$

Rearranging the above inequality yields the corollary. $\square$

Next, we turn our focus to the high-probability term and show that it is unavoidable for any non-trivial function class; essentially, any class that is not identifiable from a single sample. We define this notion formally below:

Definition 3.5.6. For a sample space $\mathcal{X}$ and label space $\mathcal{Y}$, we say a function class $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ is non-trivial if there exist $h_1, h_2 \in \mathcal{H}$ and $x_1, x_2 \in \mathcal{X}$ such that:

$$h_1(x_1) = h_2(x_1) \text{ and } h_1(x_2) \neq h_2(x_2).$$

Note that for any $\mathcal{H}$ not satisfying Definition 3.5.6 and any $h \in \mathcal{H}$, observing $h(x)$ for any $x \in \mathcal{X}$ suffices to identify it. We now establish a high-probability lower bound for any non-trivial class.

Lemma 3.5.7. *Let $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ be a sample space, label space, and hypothesis class, respectively. Suppose furthermore that $\mathcal{H}$ is non-trivial (Definition 3.5.6). Then, for any predictor $\hat{f}(\cdot; \cdot)$ for $\mathcal{H}$, $n \in \mathbb{N}$ and $\delta \in (\exp(-n/4), 1/16]$, there exists a realizable distribution P such that:*

$$\Pr_{S \sim P^n} \left[\text{err}_P \left(\hat{f}(\cdot; S) \right) \geq \frac{\log(1/\delta)}{4n} \right] \geq \delta.$$

Proof. We have from Definition 3.5.6 that there exists $x_1, x_2 \in \mathcal{X}$ and $h_1, h_2 \in \mathcal{H}$ such that:

$$y_0 = h_1(x_1) = h_2(x_1) \text{ and } y_1 = h_1(x_2) \neq h_2(x_2) = y_2.$$

We construct two realizable distributions and show through a coupling argument that the error on at least one of them is large. First, the distribution $P_{\mathcal{X}}$ over $\mathcal{X}$ is defined as follows:

$$\Pr_{P_{\mathcal{X}}} [X = x] = \begin{cases} 1 - \frac{\log(1/\delta)}{2n} & \text{if } x = x_1 \\ \frac{\log(1/\delta)}{2n} & \text{if } x = x_2 \end{cases}.$$

Now, P_1 and P_2 are defined by first drawing $X \sim P_{\mathcal{X}}$ and observing $(X, h_1(X))$ for P_1 and $(X, h_2(X))$ for P_2 . Consider the following coupled sample sets S_1 and S_2 :

$$(X_1, \dots, X_n) \sim P_{\mathcal{X}}^n, \forall i \in \{1, 2\} : S_i = ((X_1, h_i(X_1)), \dots, (X_n, h_i(X_n))).$$

Observe that S_1 and S_2 are distributed marginally as P_1^n and P_2^n respectively. Furthermore, we have:

$$\Pr[S_1 = S_2] = \left(1 - \frac{\log(1/\delta)}{2n}\right)^n \geq 2\delta.$$

Also, note that for any $S = ((x_i, y_i))_{i=1}^n$:

$$\begin{aligned} \text{err}_{P_1}(\hat{f}(\cdot; S)) + \text{err}_{P_2}(\hat{f}(\cdot; S)) &= \mathbb{E}_{X \sim P_{\mathcal{X}}} \left[\mathbf{1} \left\{ \hat{f}(X; S) \neq h_1(X) \right\} + \mathbf{1} \left\{ \hat{f}(X; S) \neq h_2(X) \right\} \right] \\ &\geq \frac{\log(1/\delta)}{2n} \cdot \left(\mathbf{1} \left\{ \hat{f}(x_2; S) \neq h_1(x_2) \right\} + \mathbf{1} \left\{ \hat{f}(x_2; S) \neq h_2(x_2) \right\} \right) \\ &\geq \frac{\log(1/\delta)}{2n}. \end{aligned}$$

The result now follows from the following computation:

$$\begin{aligned} &\Pr_{S_1} \left[\text{err}_{P_1}(\hat{f}(\cdot; S_1)) \geq \frac{\log(1/\delta)}{4n} \right] + \Pr_{S_2} \left[\text{err}_{P_2}(\hat{f}(\cdot; S_2)) \geq \frac{\log(1/\delta)}{4n} \right] \\ &\geq \Pr_{S_1, S_2} \left[\text{err}_{P_1}(\hat{f}(\cdot; S_1)) \geq \frac{\log(1/\delta)}{4n} \text{ or } \text{err}_{P_2}(\hat{f}(\cdot; S_2)) \geq \frac{\log(1/\delta)}{4n} \right] \\ &\geq \Pr_{S_1, S_2} \left[\text{err}_{P_1}(\hat{f}(\cdot; S_1)) + \text{err}_{P_2}(\hat{f}(\cdot; S_2)) \geq \frac{\log(1/\delta)}{2n} \right] \\ &\geq \Pr_{S_1, S_2} [S_1 = S_2] \cdot \Pr_{S_1, S_2} \left[\text{err}_{P_1}(\hat{f}(\cdot; S_1)) + \text{err}_{P_2}(\hat{f}(\cdot; S_2)) \geq \frac{\log(1/\delta)}{2n} \mid S_1 = S_2 \right] \\ &\geq \Pr_{S_1, S_2} [S_1 = S_2] \geq 2\delta, \end{aligned}$$

concluding the proof. $\square$

[Theorem 3.5.3](#) and [Lemma 3.5.7](#) now imply the desired lower bound of [Theorem 3.1.8](#).

Theorem 3.1.8. *There exists an absolute constant $C > 0$ such that the following holds. Let $\mathcal{X}, \mathcal{Y}$, and $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ be a sample space, label space, and hypothesis class respectively. Furthermore, suppose that $\mathcal{H}$ is non-trivial ([Definition 3.5.6](#)). Then, for any predictor $\hat{f}(\cdot; \cdot)$, $n \in \mathbb{N}$, and $\delta \in (\exp(-n/4), 1/16]$, there exists a realizable distribution P such that:*

$$\Pr_{S \sim P^n} \left[\text{err}_P(\hat{f}(\cdot; S)) \geq C \left(\sup_{k \leq 2n} \frac{\mu_{\mathcal{H}}(k)}{n} + \frac{\log(1/\delta)}{n} \right) \right] \geq \delta.$$

Proof. The theorem follows from the observation

$$\log(1/\delta) \geq 1 + \frac{\log(1/\delta)}{2},$$

the first inequality of [Corollary 3.5.5](#), and [Lemma 3.5.7](#) and noting that for $n \leq 4$, $\mu_{\mathcal{H}}(n)$ is at most a constant, concluding the proof of the theorem. $\square$

Partial hypothesis classes

As mentioned in [Section 3.1](#), we can apply main multiclass risk bound in partial hypothesis classification by slightly modifying the one-inclusion hypergraph algorithm. We would also like to emphasize that this result is standard [\[18, 10\]](#). We expand on this further. Fix a partial hypothesis class $\mathcal{H} \subseteq \{0, 1, \star\}^{\mathcal{X}}$. If we were to follow the definition of the one-inclusion hypergraph, given any subset of the domain $U \subseteq \mathcal{X}$, we would first compute the projection $\mathcal{H}|_U$ and build the appropriate set of hyperedges. For partial hypothesis classes, we will slightly change the definition of the one-inclusion hypergraph by removing every vertex (hypothesis) in the projection $\mathcal{H}|_U$ that labels a training point $x \in U$ with the $\star$ label, and then build the set of edges in the usual way from this truncated set of vertices. This is motivated by the fact that an optimal hypothesis $f^* \in \mathcal{H}$ would never produce a $\star$ label on points sampled from a realizable distribution P . It is easy to see that this change makes our one-inclusion hypergraph a graph. We can also redefine the one-inclusion hypergraph density $\mu_{\mathcal{H}}(n)$ of $\mathcal{H}$ after making this change. It now follows immediately from the arguments in the binary classification setting ([Theorem 3.1.11](#)) that the one-inclusion hypergraph density $\mu_{\mathcal{H}}(n)$ of $\mathcal{H}$ is at most d , the VC dimension of $\mathcal{H}$.

3.6 Proofs for Section 3.1: Bounded regression

Proof of [Proposition 3.1.17](#)

Proposition 3.1.17. *There are absolute $c_1, c_2 > 0$ such that the following holds. Fix a hypothesis class $\mathcal{H} \subseteq [0, 1]^{\mathcal{X}}$ and margin parameter $\gamma \in (0, 1)$. Assume that there is a constant $c_3 > 1$ such that for all $\theta > 0$, it holds that $P_{\mathcal{H}}(c_3\theta) \leq P_{\mathcal{H}}(\theta)/2$. For any $\delta \in (0, 1)$, every ERM algorithm that receives a training sample $S \sim P^n$ from a realizable distribution P as input and outputs a hypothesis $\hat{f}_{\text{ERM}}$ satisfies*

$$\mathbb{E}_{(X,Y) \sim P} \left[|\hat{f}_{\text{ERM}}(X) - Y| \right] \leq c_1 \left(\gamma + \frac{P_{\mathcal{H}}(c_2\gamma)}{n} + \frac{1}{n} \log \left(\frac{1}{\delta} \right) \right),$$

with probability at least $1 - \delta$ over the randomness of S .

Proof. Without loss of generality, we assume that in the realizable case, it holds that $Y = f^*(X)$ for some $f^* \in \mathcal{H}$. First, we use a standard ratio-type bound on empirical processes.

Theorem 19.7 in [13] implies that for any $t \geq 0$,

$$\Pr \left(\sup_{f \in \mathcal{H}} \mathbb{E}_{X \sim P} [|f(X) - f^*(X)|] - \frac{2}{n} \sum_{i=1}^n |f(X_i) - f^*(X_i)| \geq t \right) \leq 2\mathcal{N}_1 \left(\frac{t}{12}, \mathcal{H}', 2n \right) \exp \left(-\frac{2nt}{9} \right), \quad (3.16)$$

where $\mathcal{H}' = \{x \mapsto |f(x) - f^*(x)| : f \in \mathcal{H}\}$. For any distribution P' we have

$$\mathbb{E}_{X \sim P'} [||f(X) - f^*(X)| - |g(X) - f^*(X)||] \leq \mathbb{E}_{X \sim P'} [|f(X) - g(X)|].$$

Thus, it is sufficient to bound $\mathcal{N}_1 \left(\frac{t}{12}, \mathcal{H}, 2n \right)$ in terms of the P_γ -dimension. The result of Theorem 1.3 in [75] does exactly this.¹² In fact, under the regularity assumption in our statement, this result implies

$$\log 2\mathcal{N}_1 \left(\frac{t}{12}, \mathcal{H}, 2n \right) \leq c_4 P_{\mathcal{H}}(c_2 t),$$

where $c_2, c_4 > 0$ are absolute constants. Observe that $P_{\mathcal{H}}(c_2 \gamma)$ takes only integer values and decreases when γ increases. Thus, one may find $\gamma^* \in [0, 1]$ such that $P_{\mathcal{H}}(c_2 \gamma^*) \leq \frac{\gamma^* n}{9c_4} \leq P_{\mathcal{H}}(c_2 \gamma^*) + 1$. Fix $t = \gamma^*$. Using the definition of ERM and (3.16) we have that, with probability at least $1 - \exp(-n\gamma^*/9)$,

$$\mathbb{E}_{X \sim P} [|\widehat{f}_{\text{ERM}}(X) - f^*(X)|] \leq \gamma^*.$$

If $n\gamma^*/9 \geq \log(1/\delta)$, then the above inequality holds with probability at least $1 - \delta$. Otherwise, we can instead consider $t = \frac{9}{n} \log(1/\delta)$, which satisfies $t \geq \gamma^*$, and thus by monotonicity $c_4 P_{\mathcal{H}}(c_2 t) \leq c_4 P_{\mathcal{H}}(c_2 \gamma^*) \leq \frac{\gamma^* n}{9} \leq \log(1/\delta)$. This implies that whenever $n\gamma^*/9 < \log(1/\delta)$, we have, with probability at least $1 - \delta$,

$$\mathbb{E}_{X \sim P} [|\widehat{f}_{\text{ERM}}(X) - f^*(X)|] \leq \frac{9}{n} \log \left(\frac{1}{\delta} \right).$$

Combining these bounds, we show that, with probability at least $1 - \delta$,

$$\mathbb{E}_{X \sim P} [|\widehat{f}_{\text{ERM}}(X) - f^*(X)|] \leq \gamma^* + \frac{9}{n} \log \left(\frac{1}{\delta} \right).$$

Now take any $\gamma \in [0, 1]$. If $\gamma > \gamma^*$, then we obviously have $\gamma^* \leq \gamma + \frac{P_{\mathcal{H}}(c_2 \gamma)}{n}$ and the claim follows. Otherwise, if $\gamma \leq \gamma^*$, then $\gamma^* \leq \frac{9c_4(P_{\mathcal{H}}(c_2 \gamma^*)+1)}{n} \leq \frac{9c_4(P_{\mathcal{H}}(c_2 \gamma)+1)}{n}$. The claim follows by adjusting the constants. $\square$

¹²The result of Theorem 1.3 is formulated for L_2 distance, but according to the discussion on page 607 in [75], it also holds for all L_p , $1 \leq p < \infty$ distances with appropriate changes in absolute constants. Similar generalizations are made in [65].

Proof of Lemma 3.1.14

Before going into details, we provide some definitions. An alternative and equivalent definition of $V_{\mathcal{H}}(\gamma)$, suggested in [18], naturally leads to the notion of partial hypothesis classes and facilitates connections with the one-inclusion graph algorithm. We define the thresholding operator for a fixed margin γ and threshold τ as follows:

$$\psi_{\gamma,\tau}(z) = \begin{cases} 0 & \text{if } z \leq \tau - \gamma, \\ 1 & \text{if } z \geq \tau + \gamma, \\ \star & \text{otherwise.} \end{cases}$$

We now formally reintroduce the complexity measure from this perspective.

Definition 3.6.1. For $\gamma > 0$, a sequence $(x_i)_{i=1}^d \subseteq \mathcal{X}$ is V_γ -shattered by a function class $\mathcal{H}$ if there exists some $\tau \in [0, 1]$ such that

$$\{0, 1\}^d \subseteq \{(\psi_{\gamma,\tau}(f(x_1)), \dots, \psi_{\gamma,\tau}(f(x_d))) : f \in \mathcal{H}\}.$$

We define $V_{\mathcal{H}}(\gamma)$ as the length of the largest sequence V_γ -shattered shattered by $\mathcal{F}$.

Contrasting with the partial hypothesis setup, we choose a specific loss function over the discrete predictions $\{0, 1, \star\}$ that is insensitive to misclassification on $\star$. We define

$$\ell_b(\hat{y}, y) = \mathbf{1} \{\hat{y} \neq y \text{ and } y \neq \star\}. \quad (3.17)$$

We can now define the VC dimension of the partial hypothesis class $\mathcal{G} \subseteq \{0, 1, \star\}^{\mathcal{X}}$ the same way as in Section 3.1. We recall the properties of a one-inclusion predictor from [18], which extends the classical one-inclusion graph algorithm of Haussler, Littlestone, and Warmuth [42].

Lemma 3.6.2 ([18]). *Fix a hypothesis class $\mathcal{H} \subseteq \{0, 1, \star\}^{\mathcal{X}}$ with VC dimension d . Let $f^* \in \mathcal{G}$, and consider the loss function (3.17). There exists a predictor $\hat{f}$ such that, for any realizable sample $S = (x_i, f^*(x_i))_{i=1}^n$, it satisfies Assumptions 3.1.1 and 3.1.3 with $M_n \leq d$.*

Instead of delving into the details, we provide an informal explanation of how the predictor $\hat{f}$ from Lemma 3.6.2 operates. Upon observing a realizable sample of size $n - 1$, we focus solely on the indices where $f^*(x_i) \neq \star$. We project our class onto these indices only. Consequently, the problem is transformed into a $\{0, 1\}$ -valued classification problem, and we predict the label of the n -th observation using the standard one-inclusion graph prediction strategy of Haussler, Littlestone, and Warmuth [42]. Note that our loss is always equal to zero if $y_n = \star$, so that we indeed restrict ourselves to a binary classification problem.

We now restate and prove Lemma 3.1.14.

Lemma 3.1.14. *Fix a hypothesis class $\mathcal{H} \subseteq [0, 1]^{\mathcal{X}}$. There is a predictor $\hat{f} : \mathcal{X} \times \mathcal{U} \rightarrow \mathcal{Y}$ that, for any realizable sample $S = (x_i, y_i)_{i=1}^n$, satisfies Assumptions 3.1.1 and 3.1.3 with a leave-one-out error bound*

$$M_n \leq (n + 1)\gamma + V_{\mathcal{H}}(\gamma).$$

Proof. Define $m = \lceil 4n/\gamma \rceil$ and thresholds $\tau_i = i\rho$ for $i \in [m]$ and $\rho = 1/m$. Now, let $\widehat{g}_i$ denote the predictors satisfying the conditions of [Lemma 3.6.2](#) for the thresholded function classes $\mathcal{G}_i = \{\psi_{\gamma, \tau_i}(f) : f \in \mathcal{H}\}$. Hence, for a realizable sample $S = ((x_i, f^*(x_i)))_{i=1}^n$, we have for any $i \in [m]$,

$$\sum_{j=1}^n \ell_b(\widehat{g}_i(x_j; S^{-j}), \psi_{\gamma, \tau_i}(f^*(x_j))) \leq V_{\mathcal{H}}(\gamma).$$

Our predictor is defined by

$$\widehat{f}(x; S) = \rho \sum_{i=1}^m \mathbf{1}\{\widehat{g}_i(x; S) = 1\}.$$

Note that for any $z \in [0, 1]$, we have

$$\rho \sum_{i=1}^m \mathbf{1}\{z \geq \tau_i\} \leq z \leq \rho + \rho \sum_{i=1}^m \mathbf{1}\{z \geq \tau_i\}$$

Now, fix $x \in \mathcal{X}$ and let $\mathcal{I} = \{i \in [m] : \psi_{\gamma, \tau_i}(f^*(x)) = \star\}$ and $i_l = \min \mathcal{I}$ and $i_h = \max \mathcal{I}$. Noting that for any $i \notin \mathcal{I}$, it holds that

$$\mathbf{1}\{f^*(x) \geq \tau_i\} = \mathbf{1}\{\psi_{\gamma, \tau_i}(f^*(x)) = 1\} \text{ and } \mathcal{I} = \{i : i_l \leq i \leq i_h\},$$

we have

$$\begin{aligned} f^*(x) - \widehat{f}(x; S) &\leq \rho + \rho \left(\sum_{i=1}^{i_l-1} \mathbf{1}\{f^*(x) \geq \tau_i\} - \mathbf{1}\{\widehat{g}_i(x; S) = 1\} \right) \\ &\quad + \rho \left(\sum_{i=i_l}^{i_h} \mathbf{1}\{f^*(x) \geq \tau_i\} - \mathbf{1}\{\widehat{g}_i(x; S) = 1\} \right) \\ &\quad + \rho \left(\sum_{i=i_h+1}^m \mathbf{1}\{f^*(x) \geq \tau_i\} - \mathbf{1}\{\widehat{g}_i(x; S) = 1\} \right) \\ &\leq \rho + \rho \sum_{i=1}^m \mathbf{1}\{\widehat{g}_i(x; S) \neq \psi_{\gamma, \tau_i}(f^*(x)) \text{ and } \psi_{\gamma, \tau_i}(f^*(x)) \neq \star\} \\ &\quad + \rho \left(\sum_{i=i_l}^{i_h} \mathbf{1}\{f^*(x) \geq \tau_i\} - \mathbf{1}\{\widehat{g}_i(x) = 1\} \right) \\ &\leq \rho + \rho \sum_{i=1}^m \ell_b(\widehat{g}_i(x; S), \psi_{\gamma, \tau_i}(f^*(x))) + \rho \sum_{i=i_l}^{i_h} \mathbf{1}\{f^*(x) \geq \tau_i\} \\ &\leq 2\rho + \gamma + \rho \sum_{i=1}^m \ell_b(\widehat{g}_i(x; S), \psi_{\gamma, \tau_i}(f^*(x))). \end{aligned}$$

where the last inequality follows from the observation

$$\tau_{i_l} \geq f^*(x) - \gamma \implies \sum_{i=i_l}^{i_h} \mathbf{1}\{f^*(x) \geq \tau_i\} \leq \frac{\gamma}{\rho} + 1.$$

Similarly, for the lower bound, defining $i_m = \max\{i : f^*(x) \geq \tau_i\}$, we have

$$\begin{aligned} f^*(x) - \widehat{f}(x; S) &\geq \rho \left(\sum_{i=1}^{i_l-1} \mathbf{1}\{f^*(x) \geq \tau_i\} - \mathbf{1}\{\widehat{g}_i(x; S) = 1\} \right) \\ &\quad + \rho \left(\sum_{i=i_l}^{i_h} \mathbf{1}\{f^*(x) \geq \tau_i\} - \mathbf{1}\{\widehat{g}_i(x; S) = 1\} \right) \\ &\quad + \rho \left(\sum_{i=i_h+1}^m \mathbf{1}\{f^*(x) \geq \tau_i\} - \mathbf{1}\{\widehat{g}_i(x; S) = 1\} \right) \\ &\geq -\rho - \rho \sum_{i=1}^m \mathbf{1}\{\widehat{g}_i(x; S) \neq \psi_{\gamma, \tau_i}(f^*(x)) \text{ and } \psi_{\gamma, \tau_i}(f^*(x)) \neq \star\} \\ &\quad + \rho \left(\sum_{i=i_l}^{i_h} \mathbf{1}\{f^*(x) \geq \tau_i\} - \mathbf{1}\{\widehat{g}_i(x) = 1\} \right) \\ &\geq -\rho - \rho \sum_{i=1}^m \ell_b(\widehat{g}_i(x; S), \psi_{\gamma, \tau_i}(f^*(x))) - \rho(i_h - i_m) \\ &\geq -3\rho - \gamma - \rho \sum_{i=1}^m \ell_b(\widehat{g}_i(x; S), \psi_{\gamma, \tau_i}(f^*(x))), \end{aligned}$$

where the last inequality follows from

$$\tau_{i_h} - \tau_{i_m} \leq \gamma + \rho \implies i_h - i_m \leq \frac{\gamma}{\rho} + 2.$$

From the above two bounds and [Lemma 3.6.2](#), we have

$$\sum_{i=1}^n \ell(\widehat{f}(x; S^{-i}), f^*(x_i)) \leq n\gamma + 3n\rho + \rho m V_{\mathcal{H}}(\gamma) \leq (n+1)\gamma + V_{\mathcal{H}}(\gamma).$$

This concludes the proof. □

Proof of [Theorem 3.1.15](#)

Theorem 3.1.15. *Fix hypothesis class $\mathcal{H} \subseteq [0, 1]^{\mathcal{X}}$ and margin $\gamma \in (0, 1)$. There is a predictor $\widehat{f} : \mathcal{X} \times \mathcal{U} \rightarrow [0, 1]$ which, for any $\delta \in (0, 1)$, and $S \sim P^n$ sampled from any*

realizable distribution P , satisfies

$$\mathbb{E}_{(X,Y) \sim P} \left[|\hat{f}(X; S) - Y| \right] \leq 4.82 \left(\gamma + \frac{V_{\mathcal{H}}(\gamma)}{n} + \frac{1}{n} \log \left(\frac{2e}{\delta} \right) \right),$$

with probability at least $1 - \delta$ over the randomness of S .

Proof. Without loss of generality, we assume that in the realizable case, it holds that $Y = f^*(X)$ for some $f^* \in \mathcal{H}$. Using the convexity of the absolute loss, [Theorem 3.1.4](#) and [Lemma 3.1.14](#), we have, with probability at least $1 - \delta$,

$$\begin{aligned} \mathbb{E}_X \left[\left| \frac{4}{3n} \sum_{t=n/4}^{n-1} \hat{f}(X; S_{\leq t}) - f^*(X) \right| \right] &\leq \frac{4}{3n} \sum_{t=n/4}^{n-1} \mathbb{E}_X \left[|\hat{f}(X; S_{\leq t}) - f^*(X)| \right] \\ &\leq 4.82 \left(\gamma + \frac{\gamma}{n} + \frac{V_{\mathcal{H}}(\gamma)}{n} + \frac{1}{n} \log \left(\frac{2}{\delta} \right) \right). \end{aligned}$$

The claim follows if we choose the predictor $\hat{f}(\cdot) = \frac{4}{3n} \sum_{t=n/4}^{n-1} \hat{f}(\cdot; S_{\leq t})$ and note that $\gamma < 1$. $\square$

3.7 Proofs for Section 3.3: Main result

Proof of [Lemma 3.3.1](#): Martingale concentration inequalities

Lemma 3.3.1. *Let $W_1, \dots, W_T$ be a stochastic process adapted to the filtration $(\mathcal{F}_i)_{i \leq T}$. Suppose that $0 \leq W_t \leq 1$ almost surely. Then, for any $\delta, \lambda, \eta \in (0, 1)$, we have*

$$\Pr \left[\sum_{t=1}^T W_t \geq \frac{e^\lambda - 1}{\lambda} \sum_{t=1}^T \mathbb{E}[W_t | \mathcal{F}_{t-1}] + \frac{\log(1/\delta)}{\lambda} \right] \leq \delta, \quad (3.8)$$

and

$$\Pr \left[\sum_{t=1}^T \mathbb{E}[W_t | \mathcal{F}_{t-1}] \geq \frac{\eta e^\eta}{e^\eta - 1} \sum_{t=1}^T W_t + \frac{e^\eta \log(1/\delta)}{e^\eta - 1} \right] \leq \delta. \quad (3.9)$$

Proof. We first prove (3.8). Fix $\lambda \in (0, 1]$. We have

$$\mathbb{E} [e^{\lambda W_t} | \mathcal{F}_{t-1}] \leq \mathbb{E} [1 + (e^\lambda - 1)W_t | \mathcal{F}_{t-1}] \quad (3.18)$$

$$\begin{aligned} &= 1 + (e^\lambda - 1) \mathbb{E} [W_t | \mathcal{F}_{t-1}] \\ &\leq \exp((e^\lambda - 1) \mathbb{E}[W_t | \mathcal{F}_{t-1}]), \end{aligned} \quad (3.19)$$

where we used the facts that $e^{\lambda x} \leq 1 + (e^\lambda - 1)x$ for any $0 \leq x \leq 1$ in (3.18), and $1 + x \leq e^x$ for any x in (3.19). Set $\alpha = (e^\lambda - 1)$. We can now easily conclude that

$$\begin{aligned}
 \mathbb{E} \left[\prod_{t=1}^T e^{\lambda W_t - \alpha \mathbb{E}[W_t | \mathcal{F}_{t-1}]} \right] &= \mathbb{E} \left[\mathbb{E} \left[\prod_{t=1}^T e^{\lambda W_t - \alpha \mathbb{E}[W_t | \mathcal{F}_{t-1}]} \mid \mathcal{F}_{T-1} \right] \right] \\
 &= \mathbb{E} \left[\prod_{t=1}^{T-1} e^{\lambda W_t - \alpha \mathbb{E}[W_t | \mathcal{F}_{t-1}]} \cdot \mathbb{E} \left[e^{\lambda W_T - \alpha \mathbb{E}[W_T | \mathcal{F}_{T-1}]} \mid \mathcal{F}_{T-1} \right] \right] \\
 &= \mathbb{E} \left[\prod_{t=1}^{T-1} e^{\lambda W_t - \alpha \mathbb{E}[W_t | \mathcal{F}_{t-1}]} \cdot e^{-\alpha \mathbb{E}[W_T | \mathcal{F}_{T-1}]} \mathbb{E} \left[e^{\lambda W_T} \mid \mathcal{F}_{T-1} \right] \right] \\
 &\leq \mathbb{E} \left[\prod_{t=1}^{T-1} e^{\lambda W_t - \alpha \mathbb{E}[W_t | \mathcal{F}_{t-1}]} \right] \quad (\text{using (3.19)}) \\
 &\leq \mathbb{E} \left[e^{\lambda W_1 - \alpha \mathbb{E}[W_1 | \mathcal{F}_0]} \cdot 1 \right] \quad (\text{using induction}) \\
 &\leq 1.
 \end{aligned}$$

We can thus apply Markov's inequality on the random variable $\prod_{t=1}^T e^{\lambda W_t - \alpha \mathbb{E}[W_t | \mathcal{F}_{t-1}]}$ to conclude that

$$\Pr \left[\sum_{t=1}^T \lambda W_t - \alpha \mathbb{E}[W_t | \mathcal{F}_{t-1}] \geq \log(1/\delta) \right] \leq \delta.$$

(3.8) now follows after rearranging.

We now prove (3.9). Fix $\eta \in (0, 1]$. We have

$$\mathbb{E} \left[e^{-\eta W_t} \mid \mathcal{F}_{t-1} \right] \leq \mathbb{E} \left[1 - (1 - e^{-\eta}) W_t \mid \mathcal{F}_{t-1} \right] \quad (3.20)$$

$$\begin{aligned}
 &= 1 - (1 - e^{-\eta}) \mathbb{E} [W_t \mid \mathcal{F}_{t-1}] \\
 &\leq \exp(-(1 - e^{-\eta}) \mathbb{E}[W_t \mid \mathcal{F}_{t-1}]), \quad (3.21)
 \end{aligned}$$

where we used the facts that $e^{-\eta x} \leq 1 - (1 - e^{-\eta})x$ for any $0 \leq x \leq 1$ in (3.20), and $1 + x \leq e^x$

for any x in (3.21). Set $\beta = (1 - e^{-\eta})$. We can now conclude that

$$\begin{aligned}
 \mathbb{E} \left[\prod_{t=1}^T e^{-\eta W_t + \beta \mathbb{E}[W_t | \mathcal{F}_{t-1}]} \right] &= \mathbb{E} \left[\mathbb{E} \left[\prod_{t=1}^T e^{-\eta W_t + \beta \mathbb{E}[W_t | \mathcal{F}_{t-1}]} \mid \mathcal{F}_{T-1} \right] \right] \\
 &= \mathbb{E} \left[\prod_{t=1}^{T-1} e^{-\eta W_t + \beta \mathbb{E}[W_t | \mathcal{F}_{t-1}]} \cdot \mathbb{E} \left[e^{-\eta W_T + \beta \mathbb{E}[W_T | \mathcal{F}_{T-1}]} \mid \mathcal{F}_{T-1} \right] \right] \\
 &= \mathbb{E} \left[\prod_{t=1}^{T-1} e^{-\eta W_t + \beta \mathbb{E}[W_t | \mathcal{F}_{t-1}]} \cdot e^{\beta \mathbb{E}[W_T | \mathcal{F}_{T-1}]} \mathbb{E} \left[e^{-\eta W_T} \mid \mathcal{F}_{T-1} \right] \right] \\
 &\leq \mathbb{E} \left[\prod_{t=1}^{T-1} e^{-\eta W_t + \beta \mathbb{E}[W_t | \mathcal{F}_{t-1}]} \right] \quad \text{(using (3.21))} \\
 &\leq \mathbb{E} \left[e^{-\eta W_1 + \beta \mathbb{E}[W_1 | \mathcal{F}_0]} \cdot 1 \right] \quad \text{(using induction)} \\
 &\leq 1.
 \end{aligned}$$

We can thus apply Markov's inequality on the random variable $\prod_{t=1}^T e^{-\eta W_t + \beta \mathbb{E}[W_t | \mathcal{F}_{t-1}]}$ to conclude that

$$\Pr \left[-\sum_{t=1}^T \eta W_t + \beta \mathbb{E}[W_t | \mathcal{F}_{t-1}] \geq \log(1/\delta) \right] \leq \delta.$$

(3.9) now follows after rearranging. $\square$

3.8 Function Class for which our Adversarial Construction has Optimal Out-degree

The main theorem of this section is the following result which shows that there exist function classes and optimal orientations of its one-inclusion graphs, consistent with the orientation strategy of [42], with severely sub-optimal PAC performance. Furthermore, these orientations satisfy the optimal *fractional* out-degree constraints of [42].

We briefly overview *fractional* orientations of one-inclusion graphs before stating our main results. A *fractional orientation* of hypergraph $G = (V, E)$ is a function $\sigma_G : E \rightarrow \Delta(V)$ such that for any $e \in E$, $\text{supp}(\sigma_G(e)) \subset e$ where $\Delta(V)$ denotes the set of probability distributions over V and $\text{supp}(\cdot)$ denotes the support of a distribution. Given a hypergraph $G = (V, E)$ and fractional orientation σ_G , define the *out-degree* of a vertex $v \in V$ to be

$$\text{out}(v; \sigma_G) = \sum_{\substack{e \in E \\ v \in e}} \Pr_{V \sim \sigma_G(e)} \{V \neq v\}.$$

This definition generalizes the definition from Section 3.5 and recovers the previous definition if the orientations were restricted to be deterministic. The *max out-degree* for fractional

orientations are defined analogously and Algorithm 2 generalizes straightforwardly to the fractional setting by simply predicting according to a random hypothesis drawn from the orientation $\sigma_{\mathcal{X}_S \cup \{x\}}(e)$ in Line 4 of Algorithm 2 as opposed to the deterministic one determined by the integral orientations from Section 3.5. Finally, we note that stronger out-degree guarantees are possible for fractional orientations with the maximum out-degree of optimal orientations being equal to $\mu(G)$ in contrast to the weaker bound of $\lceil \mu(G) \rceil$ for integral orientations. Despite this strengthening, we still show that there exist function classes and *optimal* fractional orientations of their one-inclusion graphs (in terms of out-degree and hence, leave-one-out error) with extremely poor guarantees in the PAC setting. This is the content of the following theorem:

Theorem 3.8.1. *There are positive absolute constants C_1, C_2, c_1, c_2 and c_3 such that the following holds. For any $d \geq C_1$, $n \geq C_2 d$, and $\delta \in (c_1 d/n, c_2)$, there exists a hypothesis class $\mathcal{H}$ defined on a countable set $\mathcal{X}$ with VC dimension d , a valid one-inclusion graph algorithm whose out-degree is consistent with the optimal fractional orientation strategy of [42] whose output is a hypothesis $\hat{f}_{\text{OIG}}$ and a realizable distribution $P = P(n, \delta)$ over $\mathcal{X} \times \{0, 1\}$ satisfying:*

$$\Pr \left[\text{err}_P \left(\hat{f}_{\text{OIG}} \right) \geq c_3 \frac{d}{n\delta} \right] \geq \delta.$$

We choose the absolute constants so that $c_2 > c_1/C_2$, ensuring the interval $(c_1 d/n, c_2)$ is non-empty whenever $n \geq C_2 d$. In this section, we describe a function class and an orientation of its one-inclusion-graphs which are consistent with the max-flow based orientation strategy of Haussler, Littlestone, and Warmuth but nevertheless, suffers from poor performance in the PAC setting. Our proof strategy and construction are essentially identical to that of Theorem 3.4.3 and we include them here for the sake of completeness. The domain and the function class are parameterized by $d \in \mathbb{N}$ and are defined below:

$$\begin{aligned} \mathcal{X}_d &:= \mathcal{X}_1 \cup \mathcal{X}_2 \text{ where } \mathcal{X}_1 = \{0, 1, 2, \dots\} \text{ and } \mathcal{X}_2 = \{\alpha_1, \dots, \alpha_d\} \\ \mathcal{H}_d &:= \left\{ f : \forall x \in \mathcal{X}_2, f(x) = 0 \text{ and } \sum_{x \in \mathcal{X}_1} \mathbf{1}\{f(x) = 1\} \leq \frac{d}{4} \right\} \cup \{f : \forall x \in \mathcal{X}_1, f(x) = 1\}. \end{aligned}$$

We will use the notation $\mathcal{H}_d^1$ and $\mathcal{H}_d^2$ throughout to refer to the first and second subsets in the definition of $\mathcal{H}_d$, respectively. In the second subset, the values of f on $\mathcal{X}_2$ are unrestricted. In the remainder of this section we set $\mathcal{X} := \mathcal{X}_d$. For notational convenience, we treat quantities such as $d/2$ and $d/4$ as integers. We prove a simple result on the VC-dimension of $\mathcal{H}_d$.

Proposition 3.8.2. *The VC-dimension of $\mathcal{H}_d$ is d .*

Proof. Note that $\mathcal{H}_d$ shatters $\mathcal{X}_2$ and hence, its VC-dimension is at least d . To establish equality, let $\{x_1, \dots, x_{d+1}\} \subset \mathcal{X}$ be shattered by $\mathcal{H}_d$. Note that there exists x_i such that $x_i \in \mathcal{X}_1$. Since $\mathcal{H}_d$ shatters $\{x_j\}_{j=1}^{d+1}$, the class $\mathcal{H}' := \{f \in \mathcal{H}_d : f(x_i) = 0\}$ shatters

$\{x_j\}_{j=1}^{d+1} \setminus \{x_i\}$. However, note from the definition of $\mathcal{H}_d$ that

$$\mathcal{H}' = \left\{ f : f(x_i) = 0, \forall x \in \mathcal{X}_2, f(x) = 0 \text{ and } \sum_{x \in \mathcal{X}_1} \mathbf{1}\{f(x) = 1\} \leq d/4 \right\}$$

and consequently, the VC-dimension of $\mathcal{H}'$ is at most $d/4$, yielding a contradiction. $\square$

We now lower bound the edge density of the one inclusion graphs of $\mathcal{H}_d$ featuring a substantial number of points from $\mathcal{X}_2$.

Proposition 3.8.3. *Let $U = \{x_1, \dots, x_n\} \subset \mathcal{X}$ be such that $|U \cap \mathcal{X}_2| \geq d/2$. Then, we have:*

$$\mu(G(\mathcal{H}_d|_U)) \geq \frac{d}{4}.$$

Proof. The proof follows by considering the sub-graph

$$\{f \in \mathcal{H}_d|_U : \forall x \in U \setminus \mathcal{X}_2, f(x) = 1\},$$

which forms a complete cube over the elements $U \cap \mathcal{X}_2$ that has edge density at least $d/4$. $\square$

For any n , our hard distribution $P = P(n, d)$ will be defined as follows. In particular, although [Theorem 3.8.1](#) allows P to depend on δ , in the construction below P depends only on (n, d) ; δ only affects the choice of the orientation. The marginal over $\mathcal{X}$ is

$$P_{\mathcal{X}} = \frac{1}{2} \text{Unif}([2n]) + \frac{1}{2} \text{Unif}(\mathcal{X}_2)$$

and the labels are generated by $f^* = \mathbf{0}$. We will also only construct orientations for specific samples that are obtainable from i.i.d. samples from P . Concretely, we focus on the following set of realizations:

$$\mathcal{S} = \left\{ S : |\mathcal{X}_S \cap \mathcal{X}_1|, |\mathcal{X}_S \cap \mathcal{X}_2| \geq \frac{d}{2} \right\}.$$

For all other realizations $S \notin \mathcal{S}$ (and corresponding $(n+1)$ -point sets $\mathcal{X}_S \cup \{x\}$), we define the orientation rule σ arbitrarily (any optimal orientation). Next, we decompose the set $\mathcal{S}$ as follows:

$$\mathcal{S} := \bigcup_{\substack{Y \subseteq \mathcal{X}_2 \\ |Y| \geq d/2}} \mathcal{S}^Y \text{ where } \mathcal{S}^Y := \{S \in \mathcal{S} : \mathcal{X}_S \cap \mathcal{X}_2 = Y\}.$$

and further, decompose each $\mathcal{S}^Y$ as follows:

$$\mathcal{S}^Y := \bigcup_{k=d/2}^n \mathcal{S}_k^Y \text{ where } \mathcal{S}_k^Y := \{S \in \mathcal{S}^Y : |\mathcal{X}_S \cap \mathcal{X}_1| = k\}.$$

Now, fix one particular Y . Let $\mathcal{U}_k$ denote the set of all binary vectors of length $2n$ with exactly k entries equal to 1. We may associate each element of $\mathcal{S}_k^Y$ with an element $\mathcal{X}_S \in \mathcal{U}_k$ by

recording which elements of $[2n]$ appear in $\mathcal{X}_S \cap \mathcal{X}_1$. Similarly to the proof of [Theorem 3.4.3](#), we construct a bipartite graph with one set of vertices a subset of $\mathcal{U}_k$ (and hence, possible realizations of the subset of the training sample featuring elements in $\mathcal{X}_1$) and the other $\mathcal{U}_{k+1}$. The vertex sets are defined as follows:

$$\begin{aligned} V_1 &= \bigcup_{0 \leq i \leq 4\lceil \delta n \rceil} T_{(i)}, \\ V_2 &= \mathcal{U}_{k+1}. \end{aligned} \quad (\text{VERT-SETS-EXACT})$$

For the remainder of this subsection (with Y fixed), we identify any $u \in \mathcal{U}_k$ (or $u \in \mathcal{U}_{k+1}$) with its support $\text{supp}(u) \subseteq [2n]$, and we use the shorthand

$$\mathcal{H}_d|_u := \mathcal{H}_d|_{Y \cup \text{supp}(u)} \quad \text{and} \quad G(\mathcal{H}_d|_u) := G(\mathcal{H}_d|_{Y \cup \text{supp}(u)}).$$

Accordingly, set operations such as $u \setminus u'$ are interpreted as operations on supports. Denoting the bipartite graph by $G = (V_1, V_2, E)$ and defining the neighborhood function $N(v)$ analogously, we have as a consequence of [Lemma 3.4.7](#): Given $E \subseteq V_1 \times V_2$, define neighborhoods by

$$N(v) := \{v' \in V_2 : (v, v') \in E\} \quad \text{for } v \in V_1, \quad N(v') := \{v \in V_1 : (v, v') \in E\} \quad \text{for } v' \in V_2.$$

Moreover, for any $(v, v') \in E$ we write $v \prec v'$ to indicate that v' is obtained from v by flipping exactly one coordinate from 0 to 1.

Corollary 3.8.4. *There exists an edge set E such that the bipartite graph $G = (V_1, V_2, E)$ satisfies:*

$$\text{for all } v' \in V_2, |N(v')| \leq \frac{d}{4}, \quad \text{and} \quad \left| \left\{ v \in V_1 : |N(v)| \geq \frac{d}{32\delta} \right\} \right| \geq \frac{3}{4}|V_1|.$$

Moreover, every $(v, v') \in E$ satisfies $v \prec v'$.

As in [Section 3.4](#), we construct our orientations solely for the elements in $\mathcal{S}_{k+1}^Y$ corresponding to elements in V_2 . These orientations are constructed in [Algorithm 3](#) below. We emphasize (see the corresponding discussion in [Section 3.4](#)) that the one-inclusion prediction on a test point x depends on the one-inclusion graph induced by the set of distinct points $\mathcal{X}_S \cup \{x\}$ (and if the sample contains repetitions, the corresponding ordered-sample one-inclusion graph is a subgraph of this graph). Accordingly, the rule σ constructed below specifies an orientation $\sigma_{\mathcal{X}_S \cup \{x\}}$ for each such distinct-point set corresponding to an element of V_2 . On the subgraph $G(\mathcal{H}_d^1)$, we will use the orientation strategy that points every edge towards the vertex with lower hamming weight on the subgraph in our construction. This orientation strategy coincides with the *Closure algorithm* [\[68\]](#). Here and throughout, we write $f_0 := \mathbf{0}$ for the all-zero hypothesis.

Algorithm 3 Constructing sub-optimal one-inclusion graphs

Input: Bipartite Graph $G = (V_1, V_2, E)$.

Output: Orientation rule σ for V_2 .

- 1: Set $\sigma = \{\}$.
 - 2: **for** $U \in V_2$ **do**
 - 3: Pick any optimal orientation σ_U of $G(\mathcal{H}_d|_U)$ matching the closure algorithm on $G(\mathcal{H}_d^1|_U)$.
 - 4: For every $\mathcal{X}_S \in N(U)$, let $i = i(U, \mathcal{X}_S)$ be the unique element of $U \setminus \mathcal{X}_S$ (well-defined since $\mathcal{X}_S \prec U$), and reorient $\sigma_U(\{f_0, f_i\}) = f_i$ where $f_i(\cdot) = \mathbf{1}\{\cdot \in U \setminus \mathcal{X}_S\}$.
 - 5: Add σ_U to σ .
 - 6: **end for**
 - 7: **return** σ .
-

Note that in Line 4, $U \setminus \mathcal{X}_S$ is a singleton $\{i\} \subseteq [2n] \subseteq \mathcal{X}_1$, so f_i is the indicator hypothesis of that point and in particular belongs to $\mathcal{H}_d^1|_U$. The subsequent lemma establishes that Algorithm 3 returns a valid orientation with optimal out-degree.

Lemma 3.8.5. *For any $U \in V_2$ and $G = (V_1, V_2, E)$ satisfying the conclusions of Corollary 3.8.4, the orientation σ_U returned by Algorithm 3 satisfies:*

$$\text{out}(\sigma_U) \leq \mu(G(\mathcal{H}_d|_U)).$$

Proof. First, we prove that Line 3 in Algorithm 3 has a valid orientation. First, note that $\mathcal{H}_d^1|_U$ and $\mathcal{H}_d^2|_U$ form disjoint connected components in $G(\mathcal{H}_d|_U)$ as any $f_1 \in \mathcal{H}_d^1|_U$ and $f_2 \in \mathcal{H}_d^2|_U$ differ in more than 1 element. Specifically, f_2 takes the value 1 on at least $d/2$ elements in $U \cap \mathcal{X}_1$ while f_1 evaluates to 1 on at most $d/4$ elements in the same set. Hence, the orientations for these two function classes may be constructed independently. We pick any optimal orientation for the vertices $\mathcal{H}_d^2|_U$. For the vertices $\mathcal{H}_d^1|_U$, observe that the closure orientation, which orients each edge towards the function with fewer 1, has out-degree at most $d/4$ from the definition of $\mathcal{H}_d^1$. Furthermore, from the guarantees of Corollary 3.8.4, the out-degree of f_0 is at most $d/4$ after the re-orientation step in Line 4 of Algorithm 3. The out-degrees of all the remaining vertices in $G(\mathcal{H}_d|_U)$ either stay the same or decrease by 1. Moreover, since $\mathcal{H}_d|_U$ is shorthand for $\mathcal{H}_d|_{Y \cup \text{supp}(U)}$ and $|Y| \geq d/2$, the condition of Proposition 3.8.3 applies to $Y \cup \text{supp}(U)$. Proposition 3.8.3 now establishes the optimality of the out-degree guarantees of σ_U as any vertex in $\mathcal{H}_d^1|_U$ has out-degree at most $d/4$ while we chose an optimal orientation for the vertices in $\mathcal{H}_d^2|_U$. $\square$

Now, similar to the proof of [Theorem 3.4.3](#), define $W_1^{k,Y}$ as

$$W_1^{k,Y} = \left\{ v \in V_1 : |N(v)| \geq \frac{d}{32\delta} \right\}.$$

The following lemma shows that any training sample S with $\mathcal{X}_S$ in $W_1^{k,Y}$ incurs large error.

Lemma 3.8.6. *For any training sample S with $\mathcal{X}_S \in W_1^{k,Y}$, the one-inclusion graph algorithm that predicts using the orientation rule σ defined by [Algorithm 3](#) on input G satisfying the conclusion of [Corollary 3.8.4](#) has*

$$\text{err}_P(\hat{f}_{\sigma,S}) \geq \frac{1}{128} \cdot \frac{d}{\delta n}.$$

Proof. By the construction of the orientation rule σ in [Algorithm 3](#), the algorithm predicts 1 on a test point $x \in \mathcal{X}_1$ (and hence makes a mistake since $f^* = \mathbf{0}$) if $\mathcal{X}_S \cup \{x\} \in N(\mathcal{X}_S)$. In particular, since $N(\mathcal{X}_S) \subseteq V_2$ consists of $(k+1)$ -subsets extending $\mathcal{X}_S$, this implies $x \notin \mathcal{X}_S$ (so the test point does not appear among the training points). We now have

$$\text{err}_P(\hat{f}_{\sigma,S}) \geq \frac{1}{2} \cdot \frac{|N(\mathcal{X}_S)|}{2n} \geq \frac{1}{128} \cdot \frac{d}{\delta n}.$$

□

We will now prove [Theorem 3.8.1](#). Throughout this proof, we use that $d \geq C_1$ and $n \geq C_2 d$ with C_1 and C_2 chosen sufficiently large, and that $\delta \in (c_1 d/n, c_2)$ with c_2 chosen sufficiently small (in particular $c_2 < 1/2$), whenever needed by the inequalities below. Recall that our distribution $P = P(n, d)$ is an equal mixture over two uniform distributions, the first over the elements $[2n]$ and the second over $\{\alpha_i\}_{i=1}^d$ and labels generated by $f^* = \mathbf{0}$.

Let Y denote $\mathcal{X}_S \cap \mathcal{X}_2$ and k denote $|\mathcal{X}_S \cap \mathcal{X}_1|$ which satisfy $|Y|, k \geq d/2$. Consider the family of one-inclusion prediction strategies constructed in [Algorithm 3](#) with the bipartite graph defined in [VERT-SETS-EXACT](#) and [Corollary 3.8.4](#). By [Lemma 3.8.6](#), any S with $\mathcal{X}_S \in W_1^{k,Y}$ incurs large error. Consequently, we have by symmetricity of P on the elements of $\mathcal{X}_1$:

$$\mathbf{Pr}_{S \sim P^n} \left[\mathcal{X}_S \in W_1^{k,Y} \mid Y, |\mathcal{X}_S \cap \mathcal{X}_1| = k \right] = \frac{|W_1^{k,Y}|}{|\mathcal{U}_k|} \geq \frac{3}{4} \cdot \frac{|V_1|}{|\mathcal{U}_k|} \geq 2\delta,$$

where the last step follows from a similar argument to that in [Section 3.4](#). We keep a factor of 2 here because later we multiply by $\mathbf{Pr}[|Y|, k \geq d/2]$, which we only lower bound by a constant < 1 ; since $\mathbf{Pr}[|Y|, k \geq d/2] \geq 8/10$, a conditional probability lower bound of 2δ ensures the final lower bound is at least δ . Finally, to conclude the proof, we require lower bounds on:

$$\mathbf{Pr} \left[|Y|, k \geq \frac{d}{2} \right].$$

We will do this via the union bound. Firstly, observe:

$$\begin{aligned}\mathbb{E}[|Y|] &= \sum_{i=1}^d \mathbb{E}[\mathbf{1}\{\alpha_i \in \mathcal{X}_S\}] = d \left(1 - \left(1 - \frac{1}{2d}\right)^n\right) \geq \frac{3d}{4} \quad (\text{for } n \geq C_2 d \text{ with } C_2 \text{ sufficiently large}) \\ \mathbb{V}\text{ar}[|Y|] &\leq \mathbb{E}[|Y|] \leq d,\end{aligned}$$

since the variables $\mathbf{1}\{\alpha_i \in \mathcal{X}_S\}$ are negatively correlated Bernoulli random variables. Hence, we have by Chebyshev's inequality (for $d \geq C_1$ with C_1 sufficiently large):

$$\Pr\left[|Y| \geq \frac{d}{2}\right] \geq \frac{9}{10}.$$

Similarly, for $n \geq C_2 d$ with C_2 sufficiently large, we have:

$$\begin{aligned}\mathbb{E}[k] &= \sum_{i=1}^{2n} \mathbb{E}[\mathbf{1}\{i \in \mathcal{X}_S\}] = 2n \left(1 - \left(1 - \frac{1}{4n}\right)^n\right) \geq 2n(1 - e^{-1/4}) \geq \frac{2n}{5} \geq d, \\ \mathbb{V}\text{ar}[k] &\leq \mathbb{E}[k].\end{aligned}$$

The inequality $\mathbb{V}\text{ar}[k] \leq \mathbb{E}[k]$ follows by the same negative correlation argument used above for $|Y|$, since the indicators $\mathbf{1}\{i \in \mathcal{X}_S\}$ for $i \in [2n]$ are also negatively correlated occupancy indicators under i.i.d. sampling. Chebyshev's inequality again yields for $d \geq C_1$ with C_1 sufficiently large:

$$\Pr\left[k \geq \frac{d}{2}\right] \geq \frac{9}{10}.$$

By putting the two results together, we have:

$$\Pr\left[|Y|, k \geq \frac{d}{2}\right] \geq \frac{8}{10}.$$

Therefore,

$$\begin{aligned}\Pr\left[\mathcal{X}_S \in \bigcup_{\substack{Y \subseteq \mathcal{X}_2, \\ n \geq k \geq d/2 \\ |Y| \geq d/2}} W_1^{k,Y}\right] &= \sum_{Y: |Y| \geq d/2} \sum_{n \geq k \geq d/2} \Pr\left[\mathcal{X}_S \in W_1^{k,Y} \mid Y, |\mathcal{X}_S \cap \mathcal{X}_1| = k\right] \Pr[Y, |\mathcal{X}_S \cap \mathcal{X}_1| = k] \\ &\geq 2\delta \sum_{Y: |Y| \geq d/2} \sum_{n \geq k \geq d/2} \Pr[Y, |\mathcal{X}_S \cap \mathcal{X}_1| = k] = 2\delta \Pr[|Y|, k \geq d/2] \geq \delta\end{aligned}$$

concluding the proof of [Theorem 3.8.1](#) where the optimality of the orientation strategy follows from [Lemma 3.8.5](#).

Chapter 4

Heavy-Tailed Mean Estimation under ℓ_p Norms

This chapter is based on joint work with Yeshwanth Cherapanamjeri, Mikael Møller Høgsgaard, Kasper Green Larsen, and Nikita Zhivotovskiy [5].

Mean estimation, or location estimation, is one of the foundational subroutines in machine learning and statistics. A modern line of work focuses on statistical estimators that are robust to heavy-tailed distributions. Starting with the systematic study pioneered by Catoni [26], one is interested in obtaining non-asymptotic *sub-Gaussian* rates for estimating the mean of random variables, usually by estimators such as the median of means, which appeared independently in [71, 47, 8], Catoni’s estimator [26], or trimmed mean estimators [60]. As we recall later, more involved versions of these estimators also work well in high dimensional regimes, where the primary focus has traditionally been on the Euclidean norm.

However, the Euclidean norm is arguably the most benign norm to work with. Indeed, given n independent vectors $X_1, \dots, X_n$ in $\mathbb{R}^d$ with common mean μ and identity covariance I_d , the sample mean is essentially optimal, up to universal constant factors:

$$\mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^n X_i - \mu \right\|_2 \leq \sqrt{\frac{d}{n}}, \quad (4.1)$$

which follows from Cauchy–Schwarz and does not require the existence of any moments above the second moment. Assuming only the existence of the covariance, and no higher moments, the main focus has therefore been on obtaining sub-Gaussian estimators, namely estimators with optimal sub-Gaussian type tails, where the sample mean is clearly the wrong object [66, 61, 27, 60].

When going beyond the Euclidean norm, while still keeping only the bounded covariance assumption, even the in expectation story of (4.1) becomes more delicate. In non-Hilbertian

geometries, the vanilla sample mean is no longer necessarily a good estimator.¹ Our key contribution is to show that the Euclidean picture is special even for the simple mean estimation problem in $\mathbb{R}^d$. For ℓ_p norms with $p > 2$, robustification can be necessary even for *in expectation* bounds of the form (4.1). This phenomenon does not appear in the regime $1 \leq p \leq 2$, where the sample mean remains optimal up to universal constants, but it becomes crucial once the geometry of the loss is sufficiently far from ℓ_2 .

Theorem 4.0.1 (Phase transition for location estimation in the ℓ_p norm, informal). *Under only the covariance assumption $\text{Cov}(X) = I_d$, the optimal expected error for estimating the mean in ℓ_p satisfies*

$$\inf_{\hat{\mu}} \sup_{\substack{\mu \in \mathbb{R}^d \\ P: \mathbb{E}_P X = \mu \\ \text{Cov}_P(X) = I_d}} \mathbb{E} \|\hat{\mu} - \mu\|_p \asymp \begin{cases} d^{\frac{1}{2\lceil n/2 \rceil}}, & \bar{p} > 2\lceil n/2 \rceil, \\ d^{\frac{1}{\bar{p}}} \sqrt{\frac{\bar{p}}{n}}, & \bar{p} \leq 2\lceil n/2 \rceil, \end{cases}$$

where

$$\bar{p} = \min\{p, \max\{\log d, 1\}\}$$

Thus the minimax risk undergoes a sharp phase transition. In the large sample regime, the rate has the familiar Gaussian form, corresponding to the Gaussian width of the dual ball inducing the norm, as we discuss below:

$$d^{1/\bar{p}} \sqrt{\frac{\bar{p}}{n}}.$$

In contrast, in the small sample or large p regime, the effective exponent is capped by the number of samples, and the optimal rate becomes

$$d^{\frac{1}{2\lceil n/2 \rceil}}.$$

We show that the estimator achieving this rate is a simple coordinate-wise median of means estimator, but with a nonstandard choice of the number of blocks. Recall that median of means first splits the sample into nearly equal blocks, computes the average inside each block, and then takes a median of these block averages. In the coordinate-wise version used here, if $\bar{X}^{(1)}, \dots, \bar{X}^{(m)}$ denote the nonempty block averages, the estimator is defined by

$$\hat{\mu}_j = \text{Median} \left(\{ \bar{X}_j^{(1)}, \dots, \bar{X}_j^{(m)} \} \right), \quad j = 1, \dots, d.$$

The number of blocks is dictated not by a confidence parameter δ , as in the usual high probability theory, but by the geometry of the loss, namely by the effective exponent $\bar{p}$. More precisely, the estimator uses

$$t(\bar{p}) = 2 \left\lceil \frac{\bar{p}}{2} \right\rceil - 1$$

¹As an example, for matrix valued observations the natural benchmark is often the operator norm. In this setting, heavy-tailed matrices are known to require additional truncation steps even for in expectation risk bounds, before one can ask for sub-Gaussian tails; see, for example, [2].

blocks, with the median taken over the nonempty blocks. When $1 \leq p < 2$, we have $t(\bar{p}) = 1$, so the estimator is just the ordinary sample mean. The same happens at $p = 2$, recovering exactly the Euclidean behavior in (4.1). When $p > 2$ and $\bar{p} > 2$, the estimator becomes a genuine coordinate-wise median of means estimator, with the number of blocks increasing with $\bar{p}$. Finally, in the extreme regime where $t(\bar{p}) > n$, every nonempty block contains a single observation, and the procedure reduces to the coordinate-wise median of the sample points. The technical difficulty is to show that these regimes match sharply and together give the precise minimax rate above.

Phase transition and the existing general norm results. Our *in expectation* analysis complements the existing high probability theory of robust mean estimation with respect to general norms. Sub-Gaussian mean estimators for general norms were studied, in different forms, in [59, 67, 32]. The original benchmark in [59] is distribution dependent and is centered around robustly reproducing the performance of the empirical mean, together with the correct one dimensional sub-Gaussian deviation term. Informally, in our ℓ_p setup this benchmark contains a term of the form

$$\mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^n (X_i - \mu) \right\|_p,$$

or its symmetrized analogue. This is the natural target if the goal is to obtain optimal high probability tails around the performance of the sample mean. However, for the problem considered in this paper, this benchmark can be too large: as follows from our results, for $p > 2$, the sample mean itself is already suboptimal even in expectation.

A related but different point is emphasized in [32]. If one takes Gaussian mean estimation as the benchmark, then the relevant complexity is the Gaussian width of the dual ball inducing the norm. In the case of the ℓ_p norm and identity covariance, this gives

$$\frac{\mathbb{E} \|G\|_p}{\sqrt{n}} \asymp d^{1/\bar{p}} \sqrt{\frac{\bar{p}}{n}},$$

where $G \sim N(0, I_d)$ and $\bar{p} = \min\{p, \max\{\log d, 1\}\}$. Thus, the second branch of our theorem,

$$d^{1/\bar{p}} \sqrt{\frac{\bar{p}}{n}},$$

coincides exactly with the Gaussian width benchmark. In this regime, our estimator achieves the same rate that the sample mean would have for Gaussian data; this may be called the sub-Gaussian regime. The genuinely new phenomenon provided by our analysis is the first branch,

$$d^{2\lceil \frac{1}{2} \rceil}.$$

This term does not appear in the Gaussian mean estimation setup. It reflects a finite sample heavy-tailed regime under only a covariance assumption. Morally, the Gaussian width rate

is correct once the number of samples is large enough relative to the effective exponent $\bar{p}$. When this condition fails, the minimax risk enters a different phase.

Phase transition and the Gaussian location estimation problem. The Gaussian width branch of Theorem 4.0.1 is also closely connected to the classical minimax theory of Gaussian location estimation. In the Gaussian location model, Anderson’s lemma implies an exact minimax identity for every bowl-shaped loss; see the original inequality of Anderson [12] and the modern presentation in [73, Section 28.6]. In particular, for $X_1, \dots, X_n \sim N(\mu, I_d)$, Anderson’s lemma implies

$$\inf_{\hat{\mu}} \sup_{\mu \in \mathbb{R}^d} \mathbb{E} \|\hat{\mu} - \mu\|_p = \frac{\mathbb{E} \|G\|_p}{\sqrt{n}} \asymp d^{1/\bar{p}} \sqrt{\frac{\bar{p}}{n}}, \quad G \sim N(0, I_d), \quad \bar{p} = p \wedge (\log d \vee 1).$$

Thus, as discussed above, the second branch of our theorem is exactly the classical Gaussian location obstruction.

Our result shows that this Gaussian answer is not the whole story under only a covariance assumption. For the much larger class of all distributions with covariance I_d , the Gaussian benchmark remains sharp only when

$$\bar{p} \leq 2 \left\lceil \frac{n}{2} \right\rceil.$$

When this condition fails, heavy-tailed distributions create an additional obstruction, and the minimax risk becomes $d^{2\lceil \frac{1}{n} \rceil}$. In this sense, Theorem 4.0.1 may be viewed as a heavy-tailed, covariance-only analogue of the classical Gaussian location minimax identity.

High probability implications and limitations of projection-based estimators. As of now, the full high probability minimax picture for ℓ_p norms remains unclear. In particular, the aforementioned general norm results do not directly provide the matching high probability behavior for ℓ_p norms under only the covariance assumption. Our theorem in expectation already shows that the sample mean can be suboptimal. On the negative side, we further show that natural projection-based median of means estimators, of the type used in the general norm theory [59, 67, 32], can also fail to achieve the correct ℓ_p behavior.

Informally, a projection-based median of means estimator works as follows. It searches for a point whose projection in every dual direction agrees with the median of the projected block means:

$$\hat{\mu}_{\text{proj}} \in \arg \min_x \sup_{\|v\|_q \leq 1} |\langle v, x \rangle - \text{Median}(\{\langle v, \bar{X}^{(1)} \rangle, \dots, \langle v, \bar{X}^{(m)} \rangle\})|, \quad \frac{1}{p} + \frac{1}{q} = 1.$$

This construction is natural as a generic norm-agnostic procedure, but our result shows that it can be too crude even for the classical family of ℓ_p norms once $p > 2$.

Theorem 4.0.2 (Projection-based median of means can be suboptimal at constant probability, informal). *Fix $p > 2$. There are universal constants $c, C > 0$ such that the following holds. If $n \leq c\sqrt{d}$ and $d \geq C$, then there exists a distribution P on $\mathbb{R}^d$ with mean zero and covariance I_d such that, for any admissible choice of the bucket size satisfying $n/t \geq 3$, the projection-based median of means estimator satisfies*

$$\Pr \left(\|\hat{\mu}_{\text{proj}}\|_p \gtrsim \frac{\sqrt{d}}{n^{1-1/p}} \right) \geq 0.9.$$

This should be compared with the coordinate-wise estimator analyzed above. In the Gaussian width regime, the latter achieves a much smaller rate. For example, for $p = \infty$ and $\log d \lesssim n \ll \sqrt{d}$, the coordinate-wise estimator has error of order, with constant probability,

$$\sqrt{\frac{\log d}{n}},$$

whereas the projection-based median of means lower bound is of order, again with constant probability,

$$\frac{\sqrt{d}}{n}.$$

Thus, in this regime the projection-based estimator pays a polynomial price in the dimension, while the coordinate-wise estimator has only logarithmic dependence. This shows that, for structured norms such as ℓ_p , one needs to go beyond projection-based estimators for general norms.

A high probability result for the ℓ_∞ norm. Although our main focus is on the results in expectation, as a complementary result we give a sharp high probability analysis for the ℓ_∞ norm. This result is the high probability analogue of the phase transition phenomenon above, and refines even the constant probability regimes implied by our analysis in expectation.

Theorem 4.0.3 (ℓ_∞ high probability phase transition, informal). *Under only the covariance assumption $\text{Cov}(X) = \sigma^2 I_d$, the optimal ℓ_∞ error at confidence level $1 - \delta$ is, up to universal constants,*

$$\begin{cases} \sigma \left(\frac{d}{\delta} \right)^{\frac{1}{2\lceil \frac{1}{2} \rceil}}, & n \lesssim \log(d/\delta), \\ \sigma \sqrt{\frac{\log(d/\delta)}{n}}, & n \gtrsim \log(d/\delta). \end{cases}$$

The upper bound is achieved by the coordinate-wise median of means estimator with $t = 2\lceil \log(d/\delta) \rceil$ blocks, and the lower bound holds for every estimator.

As before, the first regime is genuinely non-Gaussian: the error is polynomial in d/δ , although the exponent decreases with the sample size. Once n is of order $\log(d/\delta)$, the usual

sub-Gaussian rate takes over. Thus, even for the classical ℓ_∞ norm, the Gaussian width or sub-Gaussian benchmark captures only the large sample side of the theory, while under only a covariance assumption there is an additional finite sample heavy-tailed phase. This also highlights a limitation of projection-based estimators: in the regimes where the coordinate-wise estimator attains the sharp ℓ_∞ rate, projection-based median of means estimators can be polynomially suboptimal, as shown by Theorem 4.0.2.

On computation and coordinate-wise medians. A significant line of work on heavy tailed mean estimation in the Euclidean norm studies polynomial time estimators with sub-Gaussian guarantees under finite covariance assumptions; see, for example, [44, 28, 45, 33]. In the main regimes in expectation studied here, computation is essentially trivial: the optimal estimator is coordinate-wise and reduces to block averages and one dimensional medians. At the same time, coordinate-wise median of means is not universally optimal. In the high probability Euclidean problem, no single distribution independent choice of the number of buckets is optimal for all distributions with identity covariance. This is expected, and probably folklore, but we include it for completeness in Section 4.4. The complete high probability theory for general norms therefore remains substantially more delicate than the picture in expectation developed here.²

4.1 Preliminaries

We begin with some notation. We use $\log$ to denote the natural logarithm. Median for an even number of reals is defined as any point that is between the left and right median. To be concrete we pick the average of these quantities. For a positive integer a , we will often use the fact that $\lfloor \frac{a+1}{2} \rfloor$ is a short form for $\frac{a+1}{2}$ when a is odd and $\frac{a}{2}$ when a is even. For positive integer a we define $[a] = \{1, 2, \dots, a\}$. We use c, C for universal constants, and their value may change from line to line. We also set $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$.

We use I for identity matrix. For a distribution P over $\mathbb{R}^d$ with mean μ , we define the covariance matrix as $\text{Cov}_P(X) = \mathbb{E}_{X \sim P}[XX^\top] - \mu\mu^\top$.

4.2 In-expectation Results

In this section we prove sharp bounds for the mini-max risk of heavy-tailed isotropic mean estimation. The minimax risk of sample size n , dimension d and norm $\|\cdot\| : \mathbb{R}^d \rightarrow \mathbb{R}_{\geq 0}$ is

²Recent generic chaining based approaches are particularly relevant here; see [1, Theorem 5.1] and [17]. We note that these works do not provide a sharp high probability minimax theory for ℓ_p norms under the covariance only assumption.

defined as follows:

$$\mathfrak{R}(n, d, \|\cdot\|) = \inf_{\mathcal{A}: (\mathbb{R}^d)^n \rightarrow \mathbb{R}^d} \sup_{\substack{\mu \in \mathbb{R}^d \\ P: \\ \mathbb{E}_P[X] = \mu, \text{Cov}_P(X) = I}} \mathbb{E}_{X_1, \dots, X_n \sim P} [\|\mathcal{A}(X_1, \dots, X_n) - \mu\|]. \quad (4.2)$$

In words, this is the smallest expected error when estimating the mean, measured in the norm $\|\cdot\|$, that an estimator can guarantee based on n samples, uniformly over all d -dimensional distributions with identity covariance. For ℓ_p norms, we will simply write $\mathfrak{R}(n, d, p) = \mathfrak{R}(n, d, \|\cdot\|_p)$.

To state our main result, it will be useful to setup some notation that will be used throughout this section. We will make frequent use of the truncated version of $p \geq 1$ defined as

$$\bar{p} = p \wedge (\log d \vee 1). \quad (4.3)$$

The reason we truncate with $\log d$ is because for any p, q such that $\ln d \leq p \leq q$, we have that $e^{-1}\|\cdot\|_p \leq \|\cdot\|_q \leq \|\cdot\|_p$, i.e. the norms are equivalent up to a constant factor. Thus, the minimax risk with respect to ℓ_p and ℓ_q are equivalent up to a constant factor so it suffices to consider the range $p \in [1, \log d]$.³

We can now state our main result.

Theorem 4.2.1. *For any $n, d \in \mathbb{N}$ and $p \geq 1$ we have that*

$$\mathfrak{R}(n, p, d) \asymp \begin{cases} d^{\frac{1}{2\lceil n/2 \rceil}}, & \bar{p} > 2 \lceil \frac{n}{2} \rceil, \\ d^{\frac{1}{\bar{p}}} \sqrt{\frac{\bar{p}}{n}}, & \bar{p} \leq 2 \lceil \frac{n}{2} \rceil. \end{cases}$$

Equivalently, if we define $s = \bar{p} \wedge 2 \lceil \frac{n}{2} \rceil$, then

$$\mathfrak{R}(n, p, d) \asymp d^{\frac{1}{s}} \sqrt{\frac{s}{n}}.$$

Moreover, this rate is achieved by the coordinate-wise median-of-means estimator.

Remark 4.2.2. [Theorem 4.2.1](#) demonstrates that the expected minimax risk with respect to ℓ_p norms for $p \geq 1$ is governed by the truncated version of p defined in (4.3). Furthermore we see a phase transition exactly at $\bar{p} = 2 \lceil \frac{n}{2} \rceil$, where the small sample size regime $2 \lceil \frac{n}{2} \rceil < \bar{p}$ rate $d^{\frac{1}{2\lceil \frac{n}{2} \rceil}}$ changes to the large sample size regime $\bar{p} \leq 2 \lceil \frac{n}{2} \rceil$ rate $d^{\frac{1}{\bar{p}}} \sqrt{\frac{\bar{p}}{n}}$.

Upper bound

In this section we prove upper bounds on the expected error of the coordinate-wise median-of-means estimator ([Algorithm 4](#)) when used to estimate the mean with respect to ℓ_p norms

³The truncation with 1 is just to avoid technicalities when $d \leq 2$.

Algorithm 4 Coordinate-wise Median-of-Means**Require:** Number of buckets t , samples $X_1, \dots, X_n \in \mathbb{R}^d$

- 1: Initialize $\hat{\mu}^{(1)}, \dots, \hat{\mu}^{(t)} \leftarrow 0 \in \mathbb{R}^d$ ▷ bucket means
- 2: Initialize $c_1, \dots, c_t \leftarrow 0$ ▷ bucket counts
- 3: **for** $m = 1$ to n **do**
- 4: $i \leftarrow 1 + ((m - 1) \bmod t)$
- 5: $\hat{\mu}^{(i)} \leftarrow \hat{\mu}^{(i)} + X_m$
- 6: $c_i \leftarrow c_i + 1$
- 7: **end for**
- 8: **for** $i = 1$ to t **do**
- 9: **if** $c_i > 0$ **then**
- 10: $\hat{\mu}^{(i)} \leftarrow \hat{\mu}^{(i)} / c_i$
- 11: **end if**
- 12: **end for**
- 13: **return** $\hat{\mu} = \text{Median}(\{\hat{\mu}^{(i)} : c_i > 0\})$ ▷ coordinate-wise median of non-empty buckets

for $p \geq 1$. We will make frequent use of the function $t(\bar{p})$:

$$t(\bar{p}) = 2 \left\lceil \frac{\bar{p}}{2} \right\rceil - 1. \quad (4.4)$$

We will use $t(\bar{p})$ as the number of buckets in coordinate-wise median-of-means estimator [Algorithm 4](#) when estimating the mean with respect to the ℓ_p norm. Notice that in the case that $p \in [1, 2]$, the number of buckets is just 1 and [Algorithm 4](#) reduces to the usual sample mean that is optimal in-expectation for $p \in [1, 2]$.

We now state our upper bound on the performance of [Algorithm 4](#) for estimating the mean of a random vector in $\mathbb{R}^d$ under ℓ_p norm for any $p \geq 1$.

Theorem 4.2.3. *Let $X_1, \dots, X_n \in \mathbb{R}^d$ be i.i.d. random vectors with finite mean μ and covariance Σ . Fix $p \geq 1$. Given the number of buckets $t(\bar{p})$ and the sample $X_1, \dots, X_n$ as input, [Algorithm 4](#) returns an estimate $\hat{\mu}$ satisfying*

$$\mathbb{E} [\|\hat{\mu} - \mu\|_p] \leq \begin{cases} 2 \left(\sum_{j=1}^d \Sigma_{jj}^{\lceil \frac{n}{2} \rceil} \right)^{\frac{1}{2 \lceil \frac{n}{2} \rceil}}, & \bar{p} > 2 \lceil \frac{n}{2} \rceil, \\ 2 \frac{\left(\sum_{j=1}^d \Sigma_{jj}^{\bar{p}/2} \right)^{\frac{1}{\bar{p}}}}{\sqrt{\lceil n/t(\bar{p}) \rceil}}, & \bar{p} \leq 2 \lceil \frac{n}{2} \rceil. \end{cases}$$

Moreover, if we define $s = \bar{p} \wedge 2 \lceil \frac{n}{2} \rceil$ then

$$\mathbb{E} [\|\hat{\mu} - \mu\|_p] \leq 4 \left(\sum_{j=1}^d \Sigma_{jj}^{s/2} \right)^{\frac{1}{s}} \sqrt{\frac{s}{n}}.$$

To prove [Theorem 4.2.3](#), we will need the following moment bound on the deviation of the empirical median from the population mean which is a slight generalization of a result due to [\[53\]](#). For completeness, we include a proof in [Section 4.5](#).

Theorem 4.2.4. *Let $X_1, \dots, X_t$ be t independent random variables with a common mean μ and let $M = M(X_1, \dots, X_t)$ denote their median. For all $q \geq 0$ it holds that*

$$\mathbb{E} \left[|M - \mu|^q \right]^{\lceil \frac{t}{2} \rceil} \leq \left(\frac{t}{\lceil \frac{t}{2} \rceil} \right) \max_{i \in [t]} \mathbb{E} [|X_i - \mu|^q]^{\lceil \frac{t}{2} \rceil}.$$

Proof of Theorem 4.2.3. We consider two cases. First, let us consider the case that $\bar{p} \leq 2 \lceil \frac{n}{2} \rceil$. Notice that this condition is equivalent to the condition that $n \geq t(\bar{p})$, which will be more convenient to use in what follows. Indeed, we have that

$$t(\bar{p}) \leq n \iff \left\lceil \frac{\bar{p}}{2} \right\rceil \leq \left\lfloor \frac{n+1}{2} \right\rfloor \iff \left\lceil \frac{\bar{p}}{2} \right\rceil \leq \left\lceil \frac{n}{2} \right\rceil \iff \frac{\bar{p}}{2} \leq \left\lceil \frac{n}{2} \right\rceil \iff \bar{p} \leq 2 \left\lceil \frac{n}{2} \right\rceil. \quad (4.5)$$

Now, using the fact that $\|\cdot\|_p \leq \|\cdot\|_{\bar{p}}$ together with Jensen's inequality yields

$$\mathbb{E} [\|\hat{\mu} - \mu\|_p] \leq \mathbb{E} [\|\hat{\mu} - \mu\|_{\bar{p}}] \leq \mathbb{E} [\|\hat{\mu} - \mu\|_{\bar{p}}^{\frac{1}{\bar{p}}}]^{\frac{1}{\bar{p}}} = \left(\sum_{j=1}^d \mathbb{E} [(\bar{\mu}_j - \mu_j)^{\bar{p}}] \right)^{\frac{1}{\bar{p}}}. \quad (4.6)$$

It now suffices to bound the terms $\mathbb{E} [(\hat{\mu}_j - \mu_j)^{\bar{p}}]$. Notice that $\left\lceil \frac{t(\bar{p})}{2} \right\rceil = \left\lfloor \frac{t(\bar{p})+1}{2} \right\rfloor = \left\lceil \frac{\bar{p}}{2} \right\rceil \geq 1$, so there is some $q \in (1, 2]$ such that

$$q \left\lceil \frac{t(\bar{p})}{2} \right\rceil = \bar{p}.$$

Thus, we can invoke the moment bound of [Theorem 4.2.4](#) with $t = t(\bar{p})$ and $q \in (1, 2]$ as above together with Jensen's inequality to get

$$\begin{aligned} \mathbb{E} [(\hat{\mu}_j - \mu_j)^{\bar{p}}] &\leq \left(\frac{t(\bar{p})}{\left\lceil \frac{t(\bar{p})}{2} \right\rceil} \right) \max_{i \in [t(\bar{p})]} \mathbb{E} \left[|\hat{\mu}_j^{(i)} - \mu_j|^q \right]^{\left\lceil \frac{t(\bar{p})}{2} \right\rceil} \\ &\leq \left(\frac{t(\bar{p})}{\left\lceil \frac{t(\bar{p})}{2} \right\rceil} \right) \max_{i \in [t(\bar{p})]} \mathbb{E} \left[|\hat{\mu}_j^{(i)} - \mu_j|^2 \right]^{\frac{q}{2} \left\lceil \frac{t(\bar{p})}{2} \right\rceil} \\ &\leq \left(\frac{t(\bar{p})}{\left\lceil \frac{t(\bar{p})}{2} \right\rceil} \right) \left(\frac{\Sigma_{jj}}{\left\lfloor \frac{n}{t(\bar{p})} \right\rfloor} \right)^{\frac{q}{2} \left\lceil \frac{t(\bar{p})}{2} \right\rceil} = \left(\frac{t(\bar{p})}{\left\lfloor \frac{t(\bar{p})+1}{2} \right\rfloor} \right) \left(\frac{\Sigma_{jj}}{\left\lfloor \frac{n}{t(\bar{p})} \right\rfloor} \right)^{\frac{\bar{p}}{2}}, \end{aligned}$$

where the last inequality follows from the fact that each bucket gets at least $\lfloor n/t(\bar{p}) \rfloor$ samples and that the variance of the j th coordinate is Σ_{jj} . Plugging the above back into (4.6) yields

$$\mathbb{E} [\|\hat{\mu} - \mu\|_p] \leq \left(\frac{t(\bar{p})}{\left\lceil \frac{t(\bar{p})}{2} \right\rceil} \right)^{\frac{1}{\bar{p}}} \frac{\left(\sum_{j=1}^d (\Sigma_{jj})^{\frac{\bar{p}}{2}} \right)^{\frac{1}{\bar{p}}}}{\sqrt{\lfloor n/t(\bar{p}) \rfloor}} \leq 2 \frac{\left(\sum_{j=1}^d (\Sigma_{jj})^{\frac{\bar{p}}{2}} \right)^{\frac{1}{\bar{p}}}}{\sqrt{\lfloor n/t(\bar{p}) \rfloor}},$$

where the final inequality follows from the simple bound

$$\binom{t(\bar{p})}{\lceil \frac{t(\bar{p})}{2} \rceil} = \binom{2\lceil \bar{p}/2 \rceil - 1}{\lceil \bar{p}/2 \rceil} = \binom{2\lceil \bar{p}/2 \rceil - 1}{\lceil \bar{p}/2 \rceil - 1} \leq \sum_{k=0}^{\lceil \bar{p}/2 \rceil - 1} \binom{2\lceil \bar{p}/2 \rceil - 1}{k} = 2^{2\lceil \frac{\bar{p}}{2} \rceil - 2} \leq 2^{\bar{p}}.$$

Now consider the case that $\bar{p} > 2\lceil \frac{n}{2} \rceil$, which is equivalent to $n < t(\bar{p})$ by (4.5). In this case, every non-empty bucket has a single sample and the estimator is the coordinate-wise median. Since $n < t(\bar{p}) = 2\lceil \bar{p}/2 \rceil - 1$, a simple case analysis on the parity of n shows that $p > 2\lceil \frac{n}{2} \rceil =: r(n)$. Thus, using the fact that $\|\cdot\|_p \leq \|\cdot\|_{2\lceil \frac{n}{2} \rceil}$ and Jensen's inequality yields

$$\mathbb{E} [\|\hat{\mu} - \mu\|_p] \leq \mathbb{E} [\|\hat{\mu} - \mu\|_{2\lceil \frac{n}{2} \rceil}] \leq \mathbb{E} [\|\hat{\mu} - \mu\|_{r(n)}^{r(n)}]^{\frac{1}{r(n)}} = \left(\sum_{j=1}^d \mathbb{E} [(\bar{\mu}_j - \mu_j)^{r(n)}] \right)^{\frac{1}{r(n)}}. \quad (4.7)$$

So, it suffices to bound the terms $\mathbb{E} [(\hat{\mu}_j - \mu_j)^{r(n)}]$. As in the previous case, we can invoke [Theorem 4.2.4](#) with $q = 2$ and $t = n$ to get the bound

$$\mathbb{E} [(\hat{\mu}_j - \mu_j)^{r(n)}] = \mathbb{E} [(\hat{\mu}_j - \mu_j)^{2\lceil \frac{n}{2} \rceil}] \leq \binom{n}{\lceil \frac{n}{2} \rceil} \Sigma_{jj}^{\lceil \frac{n}{2} \rceil}.$$

Plugging this into the bound (4.7) and using a similar approximation for the binomial term as before yields the final bound

$$\mathbb{E} [\|\hat{\mu} - \mu\|_p] \leq 2 \left(\sum_{j=1}^d (\Sigma_{jj})^{\lceil \frac{n}{2} \rceil} \right)^{\frac{1}{2\lceil \frac{n}{2} \rceil}},$$

as desired.

We now derive the uniform bound based on $s = \bar{p} \wedge 2\lceil \frac{n}{2} \rceil$. Notice that in the case that $n \geq t(\bar{p})$ we have $s = \bar{p}$ and we have the simple bound

$$\mathbb{E} [\|\hat{\mu} - \mu\|_p] \leq 2 \frac{\left(\sum_{j=1}^d (\Sigma_{jj})^{\frac{\bar{p}}{2}} \right)^{\frac{1}{\bar{p}}}}{\sqrt{\lceil n/t(\bar{p}) \rceil}} = 2 \frac{\left(\sum_{j=1}^d (\Sigma_{jj})^{\frac{s}{2}} \right)^{\frac{1}{s}}}{\sqrt{\lceil n/t(\bar{p}) \rceil}} \leq 4 \left(\sum_{j=1}^d (\Sigma_{jj})^{\frac{s}{2}} \right)^{\frac{1}{s}} \sqrt{\frac{s}{n}},$$

where the final inequality follows from the facts that $n \geq t(\bar{p})$ implies $\lfloor \frac{n}{t(\bar{p})} \rfloor \geq \frac{n}{2t(\bar{p})}$ and $t(\bar{p}) = 2\lceil \frac{\bar{p}}{2} \rceil - 1 \leq 2\bar{p} = 2s$ and so $\frac{1}{\sqrt{\lceil n/t(\bar{p}) \rceil}} \leq \sqrt{\frac{2t(\bar{p})}{n}} \leq 2\sqrt{\frac{s}{n}}$. Now in the case that $n < t(\bar{p})$ we have $s = 2\lceil \frac{n}{2} \rceil = r(n)$ and hence

$$\mathbb{E} [\|\hat{\mu} - \mu\|_p] \leq 2 \left(\sum_{j=1}^d (\Sigma_{jj})^{\lceil \frac{n}{2} \rceil} \right)^{\frac{1}{2\lceil \frac{n}{2} \rceil}} = 2 \left(\sum_{j=1}^d (\Sigma_{jj})^{\frac{s}{2}} \right)^{\frac{1}{s}} \leq 2 \left(\sum_{j=1}^d (\Sigma_{jj})^{\frac{s}{2}} \right)^{\frac{1}{s}} \sqrt{\frac{s}{n}},$$

where the final inequality uses the fact that $s = 2\lceil \frac{n}{2} \rceil \geq n$. This completes the proof. $\square$

Lower bounds

In this section we prove that the rate achieved in [Theorem 4.2.3](#) is optimal. The lower bound has two components, proved as two separate lemmas, corresponding to the two regimes discussed in the introduction. The first obstruction is already present in the Gaussian location model and gives the Gaussian-width branch. The second obstruction is genuinely non-Gaussian and is witnessed by heavy-tailed distributions.

We begin with the Gaussian lower bound. It matches [Theorem 4.2.3](#) up to constants in the regime $\bar{p} \leq 2 \lceil \frac{n}{2} \rceil$.

Lemma 4.2.5 (Gaussian-width lower bound). *There exists a universal constant $c > 0$ such that for any $d, n \in \mathbb{N}$, any $p \in [1, \infty]$, and any estimator $\mathcal{A} : (\mathbb{R}^d)^n \rightarrow \mathbb{R}^d$,*

$$\sup_{\mu \in \mathbb{R}^d} \mathbb{E}_{X_1, \dots, X_n \sim \mathcal{N}(\mu, I)} [\|\mathcal{A}(X_1, \dots, X_n) - \mu\|_p] \geq c d^{\frac{1}{\bar{p}}} \sqrt{\frac{\bar{p}}{n}}.$$

Consequently, the same lower bound holds for the minimax risk $\mathfrak{R}(n, d, p)$.

Proof. Let $P_\mu = \mathcal{N}(\mu, I)$ and let $G \sim \mathcal{N}(0, I)$. Since the Gaussian family $\{P_\mu : \mu \in \mathbb{R}^d\}$ is contained in the class of distributions with covariance I , it suffices to prove the lower bound over this subfamily. By [\[73, Theorem 28.7\]](#), which is a consequence of Anderson's lemma [\[12\]](#), applied with $\rho(x) = \|x\|_p$, the minimax risk in the Gaussian location model is

$$\inf_{\hat{\mu}} \sup_{\mu \in \mathbb{R}^d} \mathbb{E}_{X_1, \dots, X_n \sim P_\mu} [\|\hat{\mu}(X_1, \dots, X_n) - \mu\|_p] = \mathbb{E} \left\| \frac{G}{\sqrt{n}} \right\|_p = \frac{\mathbb{E} \|G\|_p}{\sqrt{n}}.$$

Indeed, the function $\rho(x) = \|x\|_p$ is bowl-shaped and lower-semicontinuous, since its sublevel sets are symmetric and convex. Therefore every estimator $\mathcal{A}$ satisfies

$$\sup_{\mu \in \mathbb{R}^d} \mathbb{E}_{X_1, \dots, X_n \sim P_\mu} [\|\mathcal{A}(X_1, \dots, X_n) - \mu\|_p] \geq \frac{\mathbb{E} \|G\|_p}{\sqrt{n}}. \quad (4.8)$$

It remains to lower bound the Gaussian width term. If q is conjugate to p , with the conventions $q = \infty$ when $p = 1$ and $q = 1$ when $p = \infty$, then

$$\mathbb{E} \|G\|_p = \mathbb{E} \left[\sup_{\|v\|_q \leq 1} \langle G, v \rangle \right].$$

This is the Gaussian width of the ℓ_q unit ball. The standard Gaussian norm estimate gives

$$\mathbb{E} \|G\|_p \asymp d^{\frac{1}{\bar{p}}} \sqrt{\bar{p}}, \quad \bar{p} = p \wedge (\log d \vee 1).$$

Plugging this estimate into [\(4.8\)](#) yields the claim. □

We now prove the second lower bound. Unlike [Lemma 4.2.5](#), this obstruction is not present in the Gaussian location model. It is a genuinely heavy-tailed obstruction and gives the small-sample branch of the minimax rate.

Since $\|x\|_p \geq \|x\|_\infty$ for all $x \in \mathbb{R}^d$ and all $p \in [1, \infty]$, it suffices to prove a lower bound in the ℓ_∞ norm.

Lemma 4.2.6 (Heavy-tailed small-sample lower bound). *There exists a universal constant $c > 0$ satisfying the following. For any $d, n \in \mathbb{N}$ with $\log d \geq 2 \lceil \frac{n}{2} \rceil$ and any estimator $\mathcal{A} : (\mathbb{R}^d)^n \rightarrow \mathbb{R}^d$, there exists a distribution P with mean $\mu \in \mathbb{R}^d$ and covariance I such that*

$$\mathbb{E}_{X_1, \dots, X_n \sim P} [\|\mathcal{A}(X_1, \dots, X_n) - \mu\|_\infty] \geq c d^{\frac{1}{2 \lceil \frac{n}{2} \rceil}}.$$

Consequently, in the regime $\log d \geq 2 \lceil \frac{n}{2} \rceil$, the same lower bound holds for the minimax risk $\mathfrak{R}(n, d, p)$ for every $p \in [1, \infty]$.

Proof. We use the high-probability ℓ_∞ lower bound proved in [Theorem 4.3.3](#). Apply [Theorem 4.3.3](#) with $\sigma = 1$ and $\delta_0 = 1/4$. Then for any estimator $\mathcal{A}$, there exists a distribution P on $\mathbb{R}^d$ with mean μ and covariance I such that

$$\Pr_{X_1, \dots, X_n \sim P} \left[\|\mathcal{A}(X_1, \dots, X_n) - \mu\|_\infty \geq \frac{1}{4} \left(1 - 2d^{-\frac{1}{\lceil \frac{n}{2} \rceil}} \right) (4d)^{\frac{1}{2 \lceil \frac{n}{2} \rceil}} \right] > \frac{1}{4}.$$

Here we used that, for $\delta_0 = 1/4$, the auxiliary parameter appearing in [Theorem 4.3.3](#) is

$$\left(\frac{4\delta_0}{d} \right)^{\frac{1}{\lceil \frac{n}{2} \rceil}} = d^{-\frac{1}{\lceil \frac{n}{2} \rceil}}.$$

By the assumption $\log d \geq 2 \lceil \frac{n}{2} \rceil$, we have

$$d^{-\frac{1}{\lceil \frac{n}{2} \rceil}} \leq e^{-2}.$$

Therefore,

$$\frac{1}{4} \left(1 - 2d^{-\frac{1}{\lceil \frac{n}{2} \rceil}} \right) (4d)^{\frac{1}{2 \lceil \frac{n}{2} \rceil}} \geq \frac{1 - 2e^{-2}}{4} d^{\frac{1}{2 \lceil \frac{n}{2} \rceil}}.$$

Using the elementary inequality $\mathbb{E} Z \geq a \Pr[Z \geq a]$ for nonnegative random variables Z , we conclude that

$$\mathbb{E}_{X_1, \dots, X_n \sim P} [\|\mathcal{A}(X_1, \dots, X_n) - \mu\|_\infty] \geq \frac{1 - 2e^{-2}}{16} d^{\frac{1}{2 \lceil \frac{n}{2} \rceil}}.$$

This proves the first claim with, for instance,

$$c = \frac{1 - 2e^{-2}}{16}.$$

The minimax-risk consequence follows from the fact that $\|x\|_p \geq \|x\|_\infty$ for every $p \in [1, \infty]$. □

4.3 Tight deviation bounds for the ℓ_∞ norm

High Probability Bounds

In this section we prove high probability bounds for the performance of the coordinate wise median of means estimator for estimating the mean of a random vector in $\mathbb{R}^d$ under the ℓ_∞ norm. Towards this end, we begin with the one dimensional case.

Theorem 4.3.1. *Let $X_1, \dots, X_n \in \mathbb{R}$ be i.i.d. random variables with finite mean μ and variance σ^2 . Fix a confidence parameter $\delta \in (0, 1)$. Given the number of buckets $t = 2\lceil \log(1/\delta) \rceil$ and the sample $X_1, \dots, X_n$ as input, [Algorithm 4](#) returns an estimate $\hat{\mu}$ satisfying*

$$|\hat{\mu} - \mu| \leq \begin{cases} 2\sigma \left(\frac{1}{\delta}\right)^{\frac{1}{2\lceil \frac{n}{2} \rceil}}, & n < 2\lceil \log\left(\frac{1}{\delta}\right) \rceil, \\ 4\sigma \sqrt{\frac{e}{n} \log\left(\frac{1}{\delta}\right)}, & n \geq 2\lceil \log\left(\frac{1}{\delta}\right) \rceil, \end{cases} \quad (4.9)$$

with probability at least $1 - \delta$ over the randomness of $X_1, \dots, X_n$.

Proof. We split the analysis into two very different cases. First, let us consider the case that the sample size n is smaller than the number of buckets t . In this case, [Algorithm 4](#) returns the median of the n sample points. Using Markov's inequality together with the moment bound of [Theorem 4.2.4](#) with $q = 2$ and $t = n$ implies that our output $\hat{\mu}$ satisfies

$$\Pr \left[|\hat{\mu} - \mu| > 2\sigma \left(\frac{1}{\delta}\right)^{\frac{1}{2\lceil \frac{n}{2} \rceil}} \right] \leq \frac{\delta \mathbb{E} \left[|\hat{\mu} - \mu|^{2\lceil \frac{n}{2} \rceil} \right]}{(2\sigma)^{2\lceil \frac{n}{2} \rceil}} \leq \frac{\delta \binom{n}{\lceil \frac{n}{2} \rceil}}{2^{2\lceil \frac{n}{2} \rceil}} < \frac{\delta 2^n}{2^{2\lceil \frac{n}{2} \rceil}} \leq \delta,$$

where the second-to-last inequality uses a simple upper bound on centered binomial coefficients.

Now consider the (well studied) case that n is at least t . In this case, [Algorithm 4](#) returns the median of the t bucket means. Each bucket contains at least $\lfloor \frac{n}{t} \rfloor$ samples, so the variance of each bucket mean is at most $\frac{\sigma^2}{\lfloor n/t \rfloor}$. Thus, Markov's inequality applied together with moment bound of [Theorem 4.2.4](#) tells us that our output $\hat{\mu} = \text{Median}(\hat{\mu}^{(1)}, \dots, \hat{\mu}^{(t)})$ satisfies

$$\Pr \left[|\hat{\mu} - \mu| > 2\sigma \sqrt{e \left(\frac{1}{\lfloor n/t \rfloor}\right)} \right] \leq \left(\frac{\lfloor n/t \rfloor}{4e\sigma^2}\right)^{\lceil \frac{t}{2} \rceil} \binom{t}{\lceil \frac{t}{2} \rceil} \max_{i \in [t]} \mathbb{E} \left[(\hat{\mu}^{(i)} - \mu)^2 \right]^{\lceil \frac{t}{2} \rceil} \leq e^{-\lceil \frac{t}{2} \rceil} \leq \delta.$$

The above implies the second bound in (4.9) since $\lfloor \frac{n}{t} \rfloor \geq \frac{n}{4\log(1/\delta)}$. This completes the proof. $\square$

Applying the above MoM estimator to each coordinate with failure parameter δ/d together with a union bound gives us the following corollary for estimation under the ℓ_∞ norm.

Corollary 4.3.2. *Let $X_1, \dots, X_n \in \mathbb{R}^d$ be i.i.d. random vectors with finite mean μ and covariance $\sigma^2 I$. Fix a confidence parameter $\delta \in (0, 1)$. Given the number of buckets $t = 2\lceil \log(d/\delta) \rceil$ and the sample $X_1, \dots, X_n$ as input, [Algorithm 4](#) returns an estimate $\hat{\mu}$ satisfying*

$$\|\hat{\mu} - \mu\|_\infty \leq \begin{cases} 2\sigma \left(\frac{d}{\delta}\right)^{\frac{1}{2\lceil \frac{n}{2} \rceil}}, & n < 2\lceil \log\left(\frac{d}{\delta}\right) \rceil, \\ 4\sigma \sqrt{\frac{e}{n} \log\left(\frac{d}{\delta}\right)}, & n \geq 2\lceil \log\left(\frac{d}{\delta}\right) \rceil, \end{cases} \quad (4.10)$$

with probability at least $1 - \delta$ over the randomness of $X_1, \dots, X_n$.

Lower bound in ℓ_∞ norm with high probability

In this sub-section we show a lower bound for any estimator in ℓ_∞ norm with high probability. Formally we show the following theorem.

Theorem 4.3.3. *For any $\delta \in (0, 1/2)$, any $\sigma > 0$, any $n, d \in \mathbb{N}$ and any deterministic estimator $\mathcal{A} : (\mathbb{R}^d)^n \rightarrow \mathbb{R}^d$, there exists a distribution P on $\mathbb{R}^d$ with mean μ and covariance matrix $\sigma^2 I$ such that*

$$\Pr_{X_1, \dots, X_n \sim P} \left[\|\mathcal{A}(X_1, \dots, X_n) - \mu\|_\infty \geq \frac{\sigma(1-2p)}{4} \left(\frac{d}{\delta}\right)^{\frac{1}{2\lceil \frac{n}{2} \rceil}} \right] > \delta,$$

where $p = \left(\frac{4\delta}{d}\right)^{\frac{1}{2\lceil \frac{n}{2} \rceil}}$.

Proof. If $p \geq 1/2$, then $1 - 2p \leq 0$, so the lower bound in the theorem is non-positive. In this case the claim is trivial since the norm is always non-negative and the claim follows for any distribution P with covariance $\sigma^2 I$. Thus, we assume that $p < 1/2$ throughout. Define

$$\ell = \frac{\sigma}{\sqrt{p(1-p)}}.$$

We consider two distributions P_0 and P_1 on $\{0, \ell\}$ given by

$$\Pr_{Y \sim P_0} [Y = \ell] = p, \quad \Pr_{Y \sim P_1} [Y = \ell] = 1 - p.$$

Equivalently, under P_0 the point ℓ occurs with probability p , while under P_1 it occurs with probability $1 - p$. Write $\mu_0 = \mathbb{E}_{Y \sim P_0} [Y] = p\ell$ and $\mu_1 = \mathbb{E}_{Y \sim P_1} [Y] = (1 - p)\ell$ for the mean of P_0 and P_1 respectively. It is immediate that both distributions have variance $p(1-p)\ell^2 = \sigma^2$. Moreover, we define the separation of the means to be

$$\Delta = \mu_1 - \mu_0 = (1 - 2p)\ell,$$

and since $p < 1/2$, we have $\Delta > 0$. We now construct the hard distribution. For each $b = (b_1, \dots, b_d) \in \{0, 1\}^d$, let P_b be the product distribution on $\mathbb{R}^d$ whose j th coordinate

has distribution P_{b_j} , independently over $j = 1, \dots, d$. Notice that P_b has mean μ_b and covariance matrix $\sigma^2 I$ where $\mu_b = (\mu_{b_1}, \dots, \mu_{b_d})$. We will show that for any deterministic estimator $\mathcal{A}$, there exists *some* choice of $b \in \{0, 1\}^d$ witnessing the lower bound. Given samples $X_1, \dots, X_n \in \mathbb{R}^d$, define a decoder $\hat{b} = \hat{b}(X_1, \dots, X_n) \in \{0, 1\}^d$ coordinate-wise by

$$\hat{b}_j \in \arg \min_{u \in \{0, 1\}} |\mathcal{A}(X_1, \dots, X_n)_j - \mu_u|,$$

with ties broken arbitrarily. Now, if $\|\mathcal{A}(X_1, \dots, X_n) - \mu_b\|_\infty < \frac{\Delta}{2}$, then every coordinate of $\mathcal{A}(X_1, \dots, X_n)$ is strictly closer to the correct mean than to the incorrect mean, hence

$$\|\mathcal{A}(X_1, \dots, X_n) - \mu_b\|_\infty < \frac{\Delta}{2} \implies \hat{b} = b. \quad (4.11)$$

Let U be the uniform distribution over $\{0, 1\}^d$ and let $B = (B_1, \dots, B_d) \sim U$. Conditioned on $B = b$, let $X_1, \dots, X_n$ be i.i.d. samples from P_b . For each coordinate j , define the coordinate samples

$$Z_j = (X_{1,j}, X_{2,j}, \dots, X_{n,j}) \in \mathcal{Z},$$

where $\mathcal{Z} := \{0, \ell\}^n$. Conditioned on $B = b$, the random vectors $Z_1, \dots, Z_d$ are independent and Z_j follows $P_{b_j}^n$. Finally define $\eta = \frac{1}{2} \sum_{z \in \{0, \ell\}^n} \min\{P_0^n(z), P_1^n(z)\}$, which, by restricting the sum to those z satisfying $|\{i : z_i = \ell\}| = \lfloor n/2 \rfloor$, can be lower bounded as follows:

$$\eta \geq \frac{1}{2} \binom{n}{\lfloor \frac{n}{2} \rfloor} \min\{p^{\lceil \frac{n}{2} \rceil} (1-p)^{\lfloor \frac{n}{2} \rfloor}, (1-p)^{\lceil \frac{n}{2} \rceil} p^{\lfloor \frac{n}{2} \rfloor}\} \geq \frac{1}{2} \binom{n}{\lfloor \frac{n}{2} \rfloor} p^{\lceil \frac{n}{2} \rceil} 2^{-\lfloor \frac{n}{2} \rfloor} \geq \frac{1}{2} p^{\lceil \frac{n}{2} \rceil} = \frac{2\delta}{d}, \quad (4.12)$$

where the second inequality uses the fact that $p < 1/2$ twice, first to say the first term in the minimum is no larger than the second and then lower bound $1-p$ by $1/2$, the final inequality uses the well known inequality $\binom{n}{r} \geq (\frac{n}{r})^r$ for $n \geq r \geq 1$ to conclude that $2^{-\lfloor \frac{n}{2} \rfloor} \binom{n}{\lfloor \frac{n}{2} \rfloor} \geq 2^{-\lfloor \frac{n}{2} \rfloor} (\frac{n}{\lfloor \frac{n}{2} \rfloor})^{\lfloor \frac{n}{2} \rfloor} \geq 1$,⁴ and the equality follows from the definition of p . We now claim that *every* decoder $\phi : (\{0, \ell\}^n)^d \rightarrow \{0, 1\}^d$ satisfies

$$\Pr_{B \sim U, X_1, \dots, X_n \sim P_B} [\phi(Z) = B] \leq (1 - \eta)^d.$$

⁴If $n = 1$, then the claimed lower bound trivially holds since $2^{-\lfloor \frac{n}{2} \rfloor} = 1$ and $\binom{n}{\lfloor \frac{n}{2} \rfloor} = \binom{1}{0} = 1$ by convention.

Indeed, writing $Z = (Z_1, \dots, Z_d)$ and summing over the sample space we have that

$$\begin{aligned}
 \Pr_{B \sim U, X_1, \dots, X_n \sim P_B} [\phi(Z) = B] &= 2^{-d} \sum_{b \in \{0,1\}^d} \sum_{z \in \mathcal{Z}^d} \mathbf{1}\{\phi(z) = b\} \prod_{j=1}^d P_{b_j}^n(z_j) \\
 &= 2^{-d} \sum_{z \in \mathcal{Z}^d} \prod_{j=1}^d P_{\phi(z)_j}^n(z_j) \\
 &\leq 2^{-d} \sum_{z \in \mathcal{Z}^d} \prod_{j=1}^d \max_{u \in \{0,1\}} P_u^n(z_j) \\
 &= 2^{-d} \sum_{z_1 \in \mathcal{Z}} \cdots \sum_{z_d \in \mathcal{Z}} \prod_{j=1}^d \max_{u \in \{0,1\}} P_u^n(z_j) \\
 &= \left(\frac{1}{2} \sum_{w \in \mathcal{Z}} \max_{u \in \{0,1\}} P_u^n(w) \right)^d.
 \end{aligned}$$

Now notice that

$$\frac{1}{2} \sum_{w \in \mathcal{Z}} \max\{P_0^n(w), P_1^n(w)\} = \frac{1}{2} \sum_{w \in \mathcal{Z}} (P_0^n(w) + P_1^n(w) - \min\{P_0^n(w), P_1^n(w)\}) = 1 - \eta,$$

so,

$$\Pr_{B \sim U, X_1, \dots, X_n \sim P_B} [\phi(Z) = B] \leq (1 - \eta)^d.$$

The above upper bound certainly holds for our decoder $\hat{b}$, so an averaging argument implies the existence of *some* $b^* \in \{0, 1\}^d$ such that

$$\Pr_{X_1, \dots, X_n \sim P_{b^*}} [\hat{b}(Z) = b^*] \leq (1 - \eta)^d \leq \left(1 - \frac{2\delta}{d}\right)^d \leq e^{-2\delta} < 1 - \delta,$$

where we used the lower bound on η from (4.12) and the inequality $e^{-2\delta} \leq 1 - \delta$ that holds for $\delta \in (0, \frac{1}{2})$. Now, using the above with (4.11), we conclude that

$$\Pr_{X_1, \dots, X_n \sim P_{b^*}} \left[\|\mathcal{A}(X_1, \dots, X_n) - \mu_{b^*}\|_\infty \geq \frac{\Delta}{2} \right] \geq \Pr_{X_1, \dots, X_n \sim P_{b^*}} [\hat{b}(Z) \neq b^*] > \delta.$$

All that remains is to lower bound $\Delta/2$. We have

$$\begin{aligned}
 \frac{\Delta}{2} &= \frac{\sigma(1-2p)}{2\sqrt{p(1-p)}} \\
 &\geq \frac{\sigma(1-2p)}{2} \left(\frac{d}{4\delta} \right)^{\frac{1}{2\lceil \frac{n}{2} \rceil}} \\
 &\geq \frac{\sigma(1-2p)}{4} \left(\frac{d}{\delta} \right)^{\frac{1}{2\lceil \frac{n}{2} \rceil}}.
 \end{aligned}$$

where we used the facts $p = \left(\frac{d}{4\delta}\right)^{\frac{1}{\lceil \frac{n}{2} \rceil}}$, $\sqrt{\frac{1}{1-p}} \geq 1$ and $4^{-\frac{1}{2\lceil \frac{n}{2} \rceil}} \geq \frac{1}{2}$. This completes the proof. $\square$

4.4 Lower bounds

Sub-optimality of the coordinate-wise median-means estimator for ℓ_2 norm

In this section, we show that [Algorithm 4](#) cannot be optimal for the ℓ_2 norm. The optimal rate for heavy-tailed mean estimation for isotropic data in this setting is known to be of order

$$\sqrt{\frac{d}{n}} + \sqrt{\frac{1}{n} \log \left(\frac{1}{\delta} \right)}. \quad (4.13)$$

We will show that there are two different isotropic distributions P_1 and P_2 that require [Algorithm 4](#) to use two very different choices for the number of buckets t for a wide range of n , d and confidence δ . This demonstrates the coordinate-wise median-of-means algorithm cannot be optimal in general.

Concretely, our first result shows that there is an isotropic distribution P_1 over $\mathbb{R}^d$ such that the number of buckets must satisfy $t = O\left(\max\left\{1, \frac{\log(1/\delta)}{d}\right\}\right)$ in order to achieve the optimal rate.

Theorem 4.4.1. *There is a universal constants $C > 0$ such that the following holds. For any $\delta \in (0, 1/3)$ and $n, d, t \in \mathbb{N}$ with $t \leq n$, there exists a distribution P_1 over $\mathbb{R}^d$ with mean 0 and covariance I such that if [Algorithm 4](#) is given the number of buckets t and the sample $X_1, \dots, X_n \sim P_1$ as input, it returns an estimate $\hat{\mu}$ satisfying*

$$\Pr_{X_1, \dots, X_n \sim P_1} \left[\|\hat{\mu}\|_2 \geq C \sqrt{\frac{dt}{n}} \right] > \delta.$$

Our second result shows that there is a distribution P_2 such that the required number of bucket must be at least $\Omega\left(\frac{\log(d/\delta)}{\log \log(d/\delta)}\right)$ to have any hope of achieving the optimal rate (4.13) for a wide range of the choice of parameter d and δ .

Theorem 4.4.2. *There are a universal constant $c, C > 0$ such that the following holds. For any $\delta \in (0, 1/e)$ and $d \in \mathbb{N}$, there exists a distribution P_2 with mean 0 and covariance I such that for any sample size n satisfying $\log n \leq 4 \log(ed) \log(e/\delta)$, it holds that if [Algorithm 4](#) is run with a number of buckets $t \leq \frac{\log(d/2\delta)}{c(\log \log ed + \log \log(e/\delta))}$ then its output $\hat{\mu}$ satisfies*

$$\Pr_{X_1, \dots, X_n \sim P_2} \left[\|\hat{\mu}\|_2 \geq C \sqrt{\frac{t \log d \log(1/\delta)}{n}} \right] > \delta.$$

Let us now consider a concrete setting of the parameters that makes the sub-optimality claim transparent. Consider the regime where $d \gg 1$, the sample size satisfies $n \leq \exp(cd \log d)$, and the confidence parameter is set to $\delta = \exp(-d)$. In this regime, [Theorem 4.4.1](#) tells us that we must use $t = O(1)$ buckets to achieve the optimal rate. On the other, [Theorem 4.4.2](#) tells us that whenever $t \leq c \frac{d}{\log d}$, [Algorithm 4](#) suffers error *at least*

$$C \sqrt{\frac{d \log d}{n}},$$

which is asymptotically larger than the optimal rate (4.13) in this regime. Hence, in order to optimally estimate the mean of P_2 , it is necessary to set the number of buckets to be larger than $c \frac{d}{\log d}$ which is asymptotically larger than the constant bucket regime demanded by P_1 .

Proof of Theorem 4.4.1. Set $p = 1 - \frac{t}{8n}$ and $\ell = \frac{1}{\sqrt{p(1-p)}}$ and define random variable $Z \sim Q$ by

$$\Pr_{Z \sim Q}[Z = (1-p)\ell] = p, \quad \Pr_{Z \sim Q}[Z = -p\ell] = 1-p.$$

Notice that $\mathbb{E}_{Z \sim Q}[Z] = 0$ and Z has variance 1. Let P_1 be the product distribution on $\mathbb{R}^d$ whose coordinates are independent copies of Z . Clearly P_1 has mean 0 and covariance I .

Let $X_1, \dots, X_n$ be an i.i.d. sample from P_1 and run [Algorithm 4](#) using t buckets. Denote the bucket sizes by $m_1, \dots, m_t$, so that

$$m_j \leq \left\lceil \frac{n}{t} \right\rceil \leq \frac{n}{t} + 1.$$

Fix a coordinate $k \in [d]$ and a bucket $j \in [t]$. The empirical mean of the k th coordinate in bucket j equals $(1-p)\ell$ whenever all samples in that bucket have k th coordinate equal to $(1-p)\ell$. This event has probability

$$p^{m_j} = \left(1 - \frac{t}{8n}\right)^{m_j} \geq 1 - \frac{m_j t}{8n} \geq 1 - \left(\frac{n}{t} + 1\right) \frac{t}{8n} \geq \frac{3}{4},$$

where we used the inequality $(1-x)^n \geq 1-nx$ which holds for $x \geq -1$ and $n \geq 1$, and the fact that $t \leq n$.

For the fixed coordinate $k \in [d]$, let Y_k be the number of buckets whose coordinate- k empirical mean is equal to $(1-p)\ell$. Then Y_k is a sum of independent Bernoulli random variables with

$$\mathbb{E} Y_k \geq \frac{3t}{4}.$$

We can apply Markov's inequality to the random variable $t - Y_k$ to get

$$\Pr_{X_1, \dots, X_n \sim P_1} \left[Y_k > \frac{t}{2} \right] = 1 - \Pr_{X_1, \dots, X_n \sim P_1} \left[t - Y_k \geq \frac{t}{2} \right] \geq 1 - \frac{\frac{t}{4}}{\frac{t}{2}} = \frac{1}{2}.$$

On the event $\{Y_k > t/2\}$, strictly more than half of the bucket means in coordinate k are equal to $(1-p)\ell$, and therefore the coordinate-wise median M_k is $(1-p)\ell$. Now define

$$W = \sum_{k=1}^d \mathbf{1}\{M_k = (1-p)\ell\}.$$

Since the coordinates of P_1 are independent, the events $\{M_k = (1-p)\ell\}$ are independent over $k \in [d]$. Moreover since,

$$\mathbb{E} W \geq \frac{d}{2},$$

we can invoke Markov's inequality on the random variable $d - W$ to conclude that

$$\Pr_{X_1, \dots, X_n \sim P_1} \left[W \geq \frac{d}{4} \right] = 1 - \Pr_{X_1, \dots, X_n \sim P_1} \left[d - W \geq \frac{3d}{4} \right] \geq 1 - \frac{\frac{d}{2}}{\frac{3d}{4}} = \frac{1}{3}.$$

On this event, we have that

$$\|\hat{\mu}\|_2 \geq \sqrt{\frac{d}{4}}(1-p)\ell = \sqrt{\frac{dt}{4(8n-t)}} \geq C\sqrt{\frac{dt}{n}},$$

for some universal constant $C > 0$. By setting $c = \frac{1}{3}$ it holds for every $\delta \in (0, c)$ that

$$\Pr_{X_1, \dots, X_n \sim P_1} \left[\|\hat{\mu}\|_2 \geq C\sqrt{\frac{dt}{n}} \right] > \delta,$$

as claimed. This completes the proof. $\square$

Proof of Theorem 4.4.2. Define $K = \log(ed)$, $L = \log(e/\delta)$, and $\varepsilon = \frac{1}{2KL}$. Let Y be the random variable with density

$$f_Y(y) = \frac{2+\varepsilon}{2}|y|^{-3-\varepsilon}\mathbf{1}\{|y| \geq 1\}.$$

Then $\mathbb{E} Y = 0$ and $\mathbb{E} Y^2 = \frac{2+\varepsilon}{\varepsilon} =: \sigma^2$. Define $X = Y/\sigma$, so that X has mean 0 and variance 1. Moreover, for every $u \geq 1/\sigma$,

$$\Pr[X \geq u] = \frac{1}{2}(\sigma u)^{-2-\varepsilon}, \quad \Pr[|X| \geq u] = (\sigma u)^{-2-\varepsilon}. \quad (4.14)$$

Let P_2 be the product distribution on $\mathbb{R}^d$ whose coordinates are independent copies of X . Clearly P_2 has mean 0 and covariance I .

We first analyze a fixed coordinate $k \in [d]$. Let $m_1, \dots, m_t$ be the bucket sizes, and write

$$\Delta = \sqrt{\frac{KL}{\lceil n/t \rceil}}.$$

For bucket $i \in [t]$, let $\bar{X}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} X_{i,j}$ and define the event

$$E_i = \left\{ \bar{X}_i \geq \sqrt{\frac{KL}{m_i}} \right\}.$$

We will prove that for every $i \in [t]$ it holds that

$$\Pr[E_i] \geq \frac{c_0}{(KL)^{5/2}},$$

for a universal constant $c_0 > 0$. Fix a bucket $i \in [t]$ and set $s = \sqrt{m_i KL}$. Since $s \geq 1/\sigma$, we can apply (4.14). For each $r \in [m_i]$, consider the event

$$F_r = \left\{ X_{i,r} \geq s, \quad |X_{i,j}| < s \quad \forall j \neq r, \quad \sum_{j \neq r} X_{i,j} \geq 0 \right\}.$$

Conditioned on $|X_{i,j}| < s$ for all $j \neq r$, the probability that $\sum_{j \neq r} X_{i,j} \geq 0$ is at least $1/2$ by symmetry. Therefore

$$\Pr[F_r] \geq \frac{1}{2} \Pr[X \geq s] (1 - \Pr[|X| \geq s])^{m_i-1} \geq \frac{1}{2} \Pr[X \geq s] (1 - \Pr[|X| \geq s])^{m_i}.$$

Notice that the events $F_1, \dots, F_{m_i}$ are disjoint and each F_r implies E_i . Thus,

$$\begin{aligned} \Pr[E_i] &\geq \sum_{r=1}^{m_i} \Pr[F_r] \\ &\geq \frac{m_i}{2} \Pr[X \geq s] (1 - \Pr[|X| \geq s])^{m_i} \\ &\geq \frac{m_i}{2} \Pr[X \geq s] (1 - m_i \Pr[|X| \geq s]) \\ &= \frac{m_i}{4} (\sigma s)^{-2-\varepsilon} (1 - m_i (\sigma s)^{-2-\varepsilon}) \\ &= \frac{m_i}{4} (\sigma s)^{-2-\varepsilon} \left(1 - \frac{1}{\sigma^{2+\varepsilon} m_i^{\varepsilon/2} (KL)^{1+\varepsilon/2}} \right) \\ &\geq \frac{3m_i}{16} (\sigma s)^{-2-\varepsilon} \\ &= \frac{3}{16 \sigma^{2+\varepsilon} m_i^{\varepsilon/2} (KL)^{1+\varepsilon/2}}. \end{aligned}$$

In the above, we used in the first inequality the inequality $(1+x)^r \geq 1+rx$ that is valid for $x \geq -1$ and $r \geq 1$ and in the second inequality that $\sigma = \sqrt{(2+\varepsilon)/\varepsilon} \geq \sqrt{4}$ by $\varepsilon = \frac{1}{2KL} \leq 1/2$. Since $\varepsilon = \frac{1}{2KL}$, we have $\log m_i \leq \log n \leq 4KL = 2/\varepsilon$, so $m_i^{\varepsilon/2} \leq e$ and

$\sigma = \sqrt{(2+\varepsilon)/\varepsilon} = \sqrt{4KL+1} \leq \sqrt{5KL}$ implying $\sigma^{2+\varepsilon} \leq (5KL)^{1+\varepsilon/2}$. Finally, using the fact that $\varepsilon \leq 1/2$ yields

$$\mathbf{Pr}[E_i] \geq \frac{3}{16e\sigma^{2+\varepsilon}(KL)^{1+\varepsilon/2}} \geq \frac{3}{16e(5KL)^{1+\varepsilon/2}(KL)^{1+\varepsilon/2}} \geq \frac{c_0}{(KL)^{5/2}},$$

for a universal constant $c_0 > 0$. This proves the one-bucket lower bound.

Let $\hat{\mu}_k$ denote the k -th coordinate of the output $\hat{\mu}$. Notice that events $E_1, \dots, E_t$ are independent and if all of them occur, every bucket mean is at least Δ , since $\sqrt{KL/m_i} \geq \Delta$. Thus

$$\mathbf{Pr}_{X_1, \dots, X_n \sim P_2}[\hat{\mu}_k \geq \Delta] \geq \left(\frac{c_0}{(KL)^{5/2}} \right)^t,$$

and since

$$\log \left(\frac{(KL)^{5/2}}{c_0} \right) \leq c(\log \log ed + \log \log(e/\delta))$$

for some universal constant $c > 0$, we have that for

$$t \leq \frac{\log(d/2\delta)}{c(\log \log ed + \log \log(e/\delta))}$$

it holds that

$$\left(\frac{c_0}{(KL)^{5/2}} \right)^t \geq \frac{2\delta}{d}.$$

We now pass to d dimensions. Since P_2 is a product distribution, the coordinate-wise median-of-means estimator outputs are independent across coordinates. Therefore

$$\mathbf{Pr}[\|\hat{\mu}\|_2 \geq \Delta] \geq \mathbf{Pr}[\exists k \in [d] : \hat{\mu}_k \geq \Delta] \geq 1 - \left(1 - \frac{2\delta}{d}\right)^d \geq 1 - e^{-2\delta} > \delta,$$

where the last inequality holds for $\delta \in (0, 1/e)$. Finally, since $t \leq n$, we have that

$$\Delta = \sqrt{\frac{KL}{\lceil n/t \rceil}} \geq C \sqrt{\frac{t \log d \log(1/\delta)}{n}},$$

for an appropriate choice of $C > 0$. Combining this with the above gives

$$\mathbf{Pr}_{X_1, \dots, X_n \sim P_2} \left[\|\hat{\mu}\|_2 \geq C \sqrt{\frac{t \log d \log(1/\delta)}{n}} \right] > \delta,$$

as claimed. This proves the theorem. $\square$

Sub-optimality of the projection-based median-of-means estimator

In this section, we establish that projection-based Median-of-Means estimators are provably sub-optimal for ℓ_p norms when $p > 2$. Concretely, when $n \ll d$, we show that they incur a polynomial dependence on dimension in the achieved error bound. Let $p > 2$ and $q < 2$ denote the order of the corresponding dual ball. The projection-based median is now defined as follows for a pointset $Y = \{y_i\}_{i=1}^n$ with $y_i \in \mathbb{R}^d$:

$$\text{Med}_{\text{proj}}(Y) = \arg \min_x \max_{v: \|v\|_q=1} |\langle v, x \rangle - \text{Med}(\{\langle v, y_i \rangle\}_{i=1}^n)|. \quad (\text{PROJ-MED})$$

The theorem is formally stated below with the proof deferred to [Section 4.5](#).

Theorem 4.4.3. *Let $p > 2$. There exist absolute constants $C, c > 0$ such that the following hold. Let $d, n \in \mathbb{N}$ be such that $n \leq c\sqrt{d}$ and $d > C$. Then, there exists a distribution D with mean 0 and covariance I such that the projection-based median-of-means estimator with any bucket size m satisfying $n/m \geq 3$, $\hat{\mu}_{\text{Proj}}$, satisfies:*

$$\Pr_{\mathbf{X} \sim D^{\otimes n}} \left(\|\hat{\mu}_{\text{Proj}}(\mathbf{X})\| \geq \frac{\sqrt{d}}{2n^{1-1/p}} \right) \geq 0.9.$$

Remark 4.4.4. Firstly, observe that the lower bound in [Theorem 4.4.3](#) directly translates to an in-expectation lower bound for the projection-based median-of-means estimator. Secondly, observe that [Theorem 4.2.3](#) when $\Sigma = I$ (and $n = \omega(\log(d))$) yields the in-expectation bound:

$$\mathbb{E} [\|\hat{\mu} - \mu\|_p] \leq 4d^{1/p} \sqrt{\frac{\log d}{n}}.$$

Comparing this bound with [Theorem 4.4.3](#), we see that the projection-based median-of-means estimator is strictly worse when $n = \tilde{o}(d)$. In particular, for $p = \infty$, [Theorem 4.4.3](#) incurs a *polynomial* dependence on d while [Theorem 4.2.3](#) is *logarithmic*.

4.5 Omitted Proofs

Lemma 4.5.1. *Let $f_1, \dots, f_t : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ be t , measurable, non-increasing and integrable functions. Then*

$$\int_0^\infty \prod_{i=1}^t f_i(x^{1/t}) dx \leq \prod_{i=1}^t \int_0^\infty f_i(x) dx$$

Proof. We have that

$$\begin{aligned}
 \prod_{i=1}^t \int_0^\infty f_i(x) dx &= \prod_{i=1}^t \int_0^\infty f_i(x_i) dx_i \\
 &= \int_0^\infty \dots \int_0^\infty \prod_{i=1}^t f_i(x_i) dx_1 \dots dx_t \\
 &= \sum_{\substack{x_{i_1} < \dots < x_{i_t} \\ \{i_1, \dots, i_t\} = [t]}} \int_0^\infty \int_0^{x_{i_t}} \dots \int_0^{x_{i_2}} \prod_{i=1}^t f_i(x_i) dx_{i_1} \dots dx_{i_{t-1}} dx_{i_t} \\
 &\geq \sum_{\substack{x_{i_1} < \dots < x_{i_t} \\ \{i_1, \dots, i_t\} = [t]}} \int_0^\infty \prod_{i=1}^t f_i(x_{i_t}) \int_0^{x_{i_t}} \dots \int_0^{x_{i_2}} 1 dx_{i_1} \dots dx_{i_{t-1}} dx_{i_t} \\
 &= t! \int_0^\infty \prod_{i=1}^t f_i(x) \frac{x^{t-1}}{(t-1)!} dx \\
 &= \int_0^\infty \prod_{i=1}^t f_i(x^{1/t}) dx,
 \end{aligned}$$

where the third equality follows from decomposing the integration domain $\mathbb{R}_+^t$ into the $t!$ disjoint regions corresponding to the possible orderings $x_{i_1} < \dots < x_{i_t}$ of the coordinates. The inequality follows from the fact that each f_i is non-increasing, so on any region where $x_{i_1} < \dots < x_{i_t}$ we have $f_i(x_i) \geq f_i(x_{i_t})$ for every $i \in [t]$. The second last equality follows from the fact that there are $t!$ such orderings, together with an explicit evaluation of the inner integrals. The final equality follows from a change of variables. $\square$

Theorem 4.2.4. *Let $X_1, \dots, X_t$ be t independent random variables with a common mean μ and let $M = M(X_1, \dots, X_t)$ denote their median. For all $q \geq 0$ it holds that*

$$\mathbb{E} \left[|M - \mu|^{q \lceil \frac{t}{2} \rceil} \right] \leq \binom{t}{\lceil \frac{t}{2} \rceil} \max_{i \in [t]} \mathbb{E} [|X_i - \mu|^q]^{\lceil \frac{t}{2} \rceil}.$$

Proof. The case $q = 0$ follows from the left hand side of the inequality to be less than or equal to the right hand side of the inequality. Furthermore, if the right hand side is infinite then the inequality also holds. Thus, we can assume that $q > 0$ and that the right hand side is finite, meaning all the expectations $\mathbb{E} [|X_i - \mu|^q]$ are finite. By calculations we have that

$$\mathbb{E} \left[|M - \mu|^{q \lceil \frac{t}{2} \rceil} \right] = \int_0^\infty \mathbf{Pr} \left[|M - \mu|^{q \lceil \frac{t}{2} \rceil} \geq x \right] dx = \int_0^\infty \mathbf{Pr} \left[|M - \mu|^q \geq x^{\frac{1}{\lceil \frac{t}{2} \rceil}} \right] dx,$$

where we have used that the expectation of a positive random variable is the integral of its right tail. Now for the median M of $X_1, \dots, X_t$ to be x away from the mean μ it must

be the case that there are $\lceil \frac{t}{2} \rceil$ X_i 's such that $|X_i - \mu| \geq x$. By this, a union bound and independence of X_i 's we then get that:

$$\begin{aligned} \mathbb{E} \left[|M - \mu|^q \lceil \frac{t}{2} \rceil \right] &\leq \int_0^\infty \mathbf{Pr} \left[\exists I \in \binom{[t]}{\lceil \frac{t}{2} \rceil} : \forall i \in I \text{ such that } |X_i - \mu|^q \geq x^{\lceil \frac{1}{\lceil \frac{t}{2} \rceil}} \right] dx \\ &\leq \sum_{I \in \binom{[t]}{\lceil \frac{t}{2} \rceil}} \int_0^\infty \mathbf{Pr} \left[\forall i \in I : |X_i - \mu|^q \geq x^{\lceil \frac{1}{\lceil \frac{t}{2} \rceil}} \right] dx \\ &= \sum_{I \in \binom{[t]}{\lceil \frac{t}{2} \rceil}} \int_0^\infty \prod_{i \in I} \mathbf{Pr} \left[|X_i - \mu|^q \geq x^{\lceil \frac{1}{\lceil \frac{t}{2} \rceil}} \right] dx. \end{aligned}$$

Now using that for each i , the function $\mathbf{Pr} [|X_i - \mu|^q \geq x]$ is non-increasing in x , integrable by considering the case where the expectation is finite for all X_i , and measurable by X_i being a random variable, we can use [Lemma 4.5.1](#) to get that

$$\begin{aligned} \mathbb{E} \left[|M - \mu|^q \lceil \frac{t}{2} \rceil \right] &\leq \sum_{I \in \binom{[t]}{\lceil \frac{t}{2} \rceil}} \prod_{i \in I} \int_0^\infty \mathbf{Pr} [|X_i - \mu|^q \geq x] dx \\ &= \sum_{I \in \binom{[t]}{\lceil \frac{t}{2} \rceil}} \prod_{i \in I} \mathbb{E} [|X_i - \mu|^q] \\ &\leq \binom{t}{\lceil \frac{t}{2} \rceil} \max_{i \in [t]} \mathbb{E} [|X_i - \mu|^q]^{\lceil \frac{t}{2} \rceil}. \end{aligned}$$

where the first inequality follows from [Lemma 4.5.1](#), the first equality follows from the fact that the expectation of a positive random variable is the integral of its right tail, and the last inequality follows from taking max under the product. $\square$

Proof of [Theorem 4.4.3](#)

Here, we establish [Theorem 4.4.3](#) which establishes the sub-optimality of the projection based median-of-means estimator for *any* choice of bucketsize. [Theorem 4.4.3](#) is restated below for convenience.

Theorem 4.4.3. *Let $p > 2$. There exist absolute constants $C, c > 0$ such that the following hold. Let $d, n \in \mathbb{N}$ be such that $n \leq c\sqrt{d}$ and $d > C$. Then, there exists a distribution D with mean 0 and covariance I such that the projection-based median-of-means estimator with any bucket size m satisfying $n/m \geq 3$, $\hat{\mu}_{\text{Proj}}$, satisfies:*

$$\mathbf{Pr}_{\mathbf{X} \sim D^{\otimes n}} \left(\|\hat{\mu}_{\text{Proj}}(\mathbf{X})\| \geq \frac{\sqrt{d}}{2n^{1-1/p}} \right) \geq 0.9.$$

We will show that the median-of-means estimator with the projection-based median defined above is sub-optimal for any choice of number of buckets, t , equivalently bucket size $m = n/t$, for the distribution D , corresponding to the uniform distribution over the signed and scaled basis vectors. Formally:

$$\forall i \in [d], s \in \pm 1 : \Pr_{X \sim D} (X = s\sqrt{d}e_i) = \frac{1}{2d}.$$

where e_i denote the standard basis vectors in $\mathbb{R}^d$. Note that D has mean 0 and covariance I .

Now, let $\mathbf{X} = \{X_i\}_{i=1}^n$ be n iid samples from from D . Furthermore, from our definition of D , $X_i = s_i\sqrt{d}e_{c(i)}$ with $s_i = \pm 1$ and $c(i) \in [d]$. We first prove a simple lemma establishing that the $c(i)$ are unique with high-probability:

Lemma 4.5.2. *Under the preconditions of [Theorem 4.4.3](#), we have:*

$$\forall i, j \in [n], i \neq j : c(i) \neq c(j)$$

with probability at least 0.9.

Proof. We have:

$$\begin{aligned} \Pr(\forall i, j \in [n], i \neq j : c(i) \neq c(j)) &= \prod_{i=1}^n \left(1 - \frac{(i-1)}{d}\right) \geq \prod_{i=1}^n \left(1 - \frac{n}{d}\right) \\ &= \left(1 - \frac{n}{d}\right)^n \geq \exp\left(-\frac{2n}{d} \cdot n\right) \geq 0.9 \end{aligned}$$

as $n \leq c\sqrt{d}$. □

Now, we condition on the uniqueness of $c(i)$. Define the bucketed means Z_i as follows for $i \in [t]$ with $m := n/t$:

$$Z_i = \frac{1}{m} \sum_{j=(i-1)m+1}^{im} X_j$$

with the collection denoted by $\mathbf{Z}$. From [PROJ-MED](#), the projection-based median-of-means estimate is the minimizer of the following function:

$$F(x) := \max_{v: \|v\|_q=1} |\langle v, x \rangle - \text{Med}(\{\langle v, Z_i \rangle\}_{i=1}^t)|$$

We now make some observations:

Observation 4.5.3. F is a convex function of x .

Proof. This follows from F being defined as the maximum of a set of convex functions. □

Next, F satisfies the following symmetries:

Observation 4.5.4. Defining:

$$\tilde{X}_i = \sqrt{d}e_{c(i)}, S \in \mathbb{R}^{d \times d} : S_{i,j} := \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j, \forall \ell, c(\ell) \neq i \\ 1 & \text{if } i = j, \exists \ell, c(\ell) = i \text{ and } s_\ell = 1 \\ -1 & \text{otherwise} \end{cases}, \text{ and } \tilde{Z}_i := \frac{1}{m} \sum_{j=(i-1)m+1}^{im} \tilde{X}_j.$$

where uniqueness of $c(i)$ from [Lemma 4.5.2](#) ensures that S is well-defined. Finally, defining $\tilde{F}$ analogously to F , we observe the following:

$$\forall x : F(x) = \tilde{F}(Sx).$$

Consequently,

$$\arg \min_x F(x) = S \arg \min_x \tilde{F}(x).$$

Proof. From the definition of S we have

$$S\tilde{X}_i = S\sqrt{d}e_{c(i)} = S_{c(i),c(i)}\sqrt{d}e_{c(i)} = s_i\sqrt{d}e_{c(i)} = X_i$$

which implies

$$\forall v, \|v\|_q = 1, i \in [t] : \langle v, Z_i \rangle = \langle Sv, \tilde{Z}_i \rangle.$$

and hence

$$F(x) = \max_{v: \|v\|_q=1} |\langle v, x \rangle - \text{Med}(\{\langle v, Z_i \rangle\}_{i=1}^t)| = \max_{v: \|v\|_q=1} |\langle Sv, Sx \rangle - \text{Med}\left(\left\{\langle Sv, \tilde{Z}_i \rangle\right\}_{i=1}^t\right)| = \tilde{F}(Sx)$$

where the last equality follows from the fact that S is a diagonal matrix with entries in $\{\pm 1\}$, so Sv ranges over all vectors with ℓ_q norm at most 1 as v does. $\square$

On the event of [Lemma 4.5.2](#), define the bucket supports

$$C_r := \{c((r-1)m+1), \dots, c(rm)\}, \quad r \in [t].$$

Then the sets $C_1, \dots, C_t$ are disjoint, $|C_r| = m$ for all $r \in [t]$, and

$$\tilde{Z}_r = \frac{\sqrt{d}}{m} \sum_{j \in C_r} e_j.$$

Hence,

$$\tilde{F}(x) = \max_{v: \|v\|_q=1} \left| \langle v, x \rangle - \text{Med} \left(\left\{ \frac{\sqrt{d}}{m} \sum_{j \in C_r} v_j \right\}_{r=1}^t \right) \right|.$$

Lemma 4.5.5. *Fix $r \in [t]$, and let P be any permutation matrix that permutes the coordinates in C_r and fixes all coordinates outside C_r . Then*

$$\forall x \in \mathbb{R}^d : \tilde{F}(Px) = \tilde{F}(x).$$

Proof. Since

$$\tilde{Z}_\ell = \frac{\sqrt{d}}{m} \sum_{j \in C_\ell} e_j$$

for every $\ell \in [t]$, permuting coordinates inside C_r leaves every $\tilde{Z}_\ell$ unchanged:

$$\forall \ell \in [t] : P\tilde{Z}_\ell = \tilde{Z}_\ell.$$

Therefore,

$$\begin{aligned} \tilde{F}(Px) &= \max_{v: \|v\|_q=1} \left| \langle v, Px \rangle - \text{Med} \left(\left\{ \langle v, \tilde{Z}_\ell \rangle \right\}_{\ell=1}^t \right) \right| \\ &= \max_{v: \|v\|_q=1} \left| \langle P^\top v, x \rangle - \text{Med} \left(\left\{ \langle P^\top v, \tilde{Z}_\ell \rangle \right\}_{\ell=1}^t \right) \right|. \end{aligned}$$

Since P is a permutation matrix, $\|P^\top v\|_q = \|v\|_q$. Hence, after the change of variables $u = P^\top v$,

$$\tilde{F}(Px) = \max_{u: \|u\|_q=1} \left| \langle u, x \rangle - \text{Med} \left(\left\{ \langle u, \tilde{Z}_\ell \rangle \right\}_{\ell=1}^t \right) \right| = \tilde{F}(x).$$

□

Corollary 4.5.6. *There exists a minimizer x^* of $\tilde{F}$ such that for every $r \in [t]$, all coordinates of x^* in C_r are equal.*

Proof. Let $\mathcal{G}_{\text{in}}$ denote the finite group of permutation matrices that permute coordinates inside each C_r and fix all coordinates outside $\cup_{r=1}^t C_r$. For any minimizer x of $\tilde{F}$, define its symmetrization

$$\bar{x} := \frac{1}{|\mathcal{G}_{\text{in}}|} \sum_{P \in \mathcal{G}_{\text{in}}} Px.$$

By convexity of $\tilde{F}$ from [Observation 4.5.3](#) and [Lemma 4.5.5](#),

$$\tilde{F}(\bar{x}) \leq \frac{1}{|\mathcal{G}_{\text{in}}|} \sum_{P \in \mathcal{G}_{\text{in}}} \tilde{F}(Px) = \frac{1}{|\mathcal{G}_{\text{in}}|} \sum_{P \in \mathcal{G}_{\text{in}}} \tilde{F}(x) = \tilde{F}(x).$$

Since x is already a minimizer, $\bar{x}$ is also a minimizer.

By construction, $\bar{x}$ is invariant under every permutation inside each C_r . Hence,

$$\forall r \in [t], \forall i, j \in C_r : \bar{x}_i = \bar{x}_j.$$

□

Lemma 4.5.7. *Let σ be any permutation of $[t]$. For each $r \in [t]$, choose any bijection*

$$\phi_r : C_r \rightarrow C_{\sigma(r)}.$$

Let $P_{\sigma,\phi}$ be the permutation matrix that sends each coordinate $j \in C_r$ to $\phi_r(j)$, and fixes all coordinates outside $\cup_{r=1}^t C_r$. Then

$$\forall x \in \mathbb{R}^d : \tilde{F}(P_{\sigma,\phi}x) = \tilde{F}(x).$$

Proof. By construction,

$$P_{\sigma,\phi}\tilde{Z}_r = \frac{\sqrt{d}}{m} \sum_{j \in C_r} P_{\sigma,\phi}e_j = \frac{\sqrt{d}}{m} \sum_{k \in C_{\sigma(r)}} e_k = \tilde{Z}_{\sigma(r)}.$$

Therefore,

$$\begin{aligned} \tilde{F}(P_{\sigma,\phi}x) &= \max_{v: \|v\|_q=1} \left| \langle v, P_{\sigma,\phi}x \rangle - \text{Med} \left(\left\{ \langle v, \tilde{Z}_r \rangle \right\}_{r=1}^t \right) \right| \\ &= \max_{v: \|v\|_q=1} \left| \langle P_{\sigma,\phi}^\top v, x \rangle - \text{Med} \left(\left\{ \langle P_{\sigma,\phi}^\top v, \tilde{Z}_{\sigma(r)} \rangle \right\}_{r=1}^t \right) \right|. \end{aligned}$$

Since σ only reorders the multiset of bucket projections, the median is unchanged:

$$\text{Med} \left(\left\{ \langle P_{\sigma,\phi}^\top v, \tilde{Z}_{\sigma(r)} \rangle \right\}_{r=1}^t \right) = \text{Med} \left(\left\{ \langle P_{\sigma,\phi}^\top v, \tilde{Z}_r \rangle \right\}_{r=1}^t \right).$$

Since $P_{\sigma,\phi}$ is a permutation matrix $\|P_{\sigma,\phi}^\top v\|_q = \|v\|_q$. Hence, after the change of variables $u = P_{\sigma,\phi}^\top v$,

$$\tilde{F}(P_{\sigma,\phi}x) = \max_{u: \|u\|_q=1} \left| \langle u, x \rangle - \text{Med} \left(\left\{ \langle u, \tilde{Z}_r \rangle \right\}_{r=1}^t \right) \right| = \tilde{F}(x).$$

□

Corollary 4.5.8. *There exists a minimizer x^* of $\tilde{F}$ and a scalar $a \in \mathbb{R}$ such that*

$$\forall j \in \bigcup_{r=1}^t C_r : x_j^* = a.$$

Proof. Start from the minimizer given by [Corollary 4.5.6](#), so that

$$\forall r \in [t], \forall j \in C_r : x_j = a_r$$

for some scalars $a_1, \dots, a_t$. Let $\mathcal{G}_{\text{buck}}$ denote the finite set of all permutations $P_{\sigma,\phi}$ from [Lemma 4.5.7](#). Define

$$\bar{x} := \frac{1}{|\mathcal{G}_{\text{buck}}|} \sum_{P \in \mathcal{G}_{\text{buck}}} Px.$$

By convexity of $\tilde{F}$ from [Observation 4.5.3](#) and [Lemma 4.5.7](#),

$$\tilde{F}(\bar{x}) \leq \frac{1}{|\mathcal{G}_{\text{buck}}|} \sum_{P \in \mathcal{G}_{\text{buck}}} \tilde{F}(Px) = \tilde{F}(x).$$

Hence $\bar{x}$ is also a minimizer. The group $\mathcal{G}_{\text{buck}}$ acts transitively on the active coordinates $\cup_{r=1}^t C_r$: any active coordinate can be mapped to any other active coordinate by first permuting buckets and then using a bijection inside the target bucket. Therefore, $\bar{x}$ must take the same value on all active coordinates, i.e.,

$$\exists a \in \mathbb{R} \text{ such that } \forall j \in \bigcup_{r=1}^t C_r : \bar{x}_j = a.$$

□

Lemma 4.5.9. *There exists a minimizer x^* of $\tilde{F}$ such that*

$$\forall j \notin \bigcup_{r=1}^t C_r : x_j^* = 0.$$

Consequently, there exists a minimizer of the form

$$x^* = a \sum_{r=1}^t \sum_{j \in C_r} e_j$$

for some scalar $a \in \mathbb{R}$.

Proof. Fix $j \notin \cup_{r=1}^t C_r$, and let D_j be the diagonal matrix defined by

$$(D_j)_{k,k} = \begin{cases} -1 & \text{if } k = j, \\ 1 & \text{otherwise.} \end{cases}$$

Since every $\tilde{Z}_r$ is supported on C_r , we have

$$\forall r \in [t] : D_j \tilde{Z}_r = \tilde{Z}_r.$$

Hence, arguing exactly as in [Lemma 4.5.5](#),

$$\forall x \in \mathbb{R}^d : \tilde{F}(D_j x) = \tilde{F}(x).$$

Let x be any minimizer of $\tilde{F}$, and define

$$\bar{x} := \frac{x + D_j x}{2}.$$

By convexity of $\tilde{F}$ from [Observation 4.5.3](#),

$$\tilde{F}(\bar{x}) \leq \frac{\tilde{F}(x) + \tilde{F}(D_j x)}{2} = \tilde{F}(x).$$

Thus $\bar{x}$ is also a minimizer. Moreover, $\bar{x}$ agrees with x on every coordinate except j , and

$$\bar{x}_j = 0.$$

Repeating this argument for every $j \notin \cup_{r=1}^t C_r$ yields a minimizer supported on $\cup_{r=1}^t C_r$.

Combining this with [Corollary 4.5.8](#) proves that there exists a minimizer of the form

$$x^* = a \sum_{r=1}^t \sum_{j \in C_r} e_j = a \sum_{i=1}^n e_{c(i)}.$$

□

At this point, by [Lemma 4.5.9](#), it remains only to optimize over the scalar multiplying the active coordinates. Define

$$\rho := \left\lfloor \frac{t}{2} \right\rfloor + 1, \quad \eta := t - \rho.$$

Let $y_{(1)} \geq \dots \geq y_{(t)}$ denote the non-increasing rearrangement of a multiset of scalars $\mathbf{Y} = \{y_i\}_{i=1}^t$. We use the convention

$$\text{Med}(\mathbf{Y}) = \begin{cases} y_{(k)} & \text{if } t = 2k - 1, \\ \frac{y_{(k)} + y_{(k+1)}}{2} & \text{if } t = 2k. \end{cases}$$

For $a \in \mathbb{R}$, define

$$x^{(a)} := a \sum_{i=1}^n e_{c(i)} = a \sum_{r=1}^t \sum_{j \in C_r} e_j.$$

Lemma 4.5.10. *For every $a \in \mathbb{R}$,*

$$\tilde{F}(x^{(a)}) = m^{1/p} \max_{u \in \mathbb{R}^t: \|u\|_q \leq 1} \left| a \sum_{r=1}^t u_r - \frac{\sqrt{d}}{m} \text{Med}(\{u_r\}_{r=1}^t) \right|.$$

Proof. For $v \in \mathbb{R}^d$, define

$$u_r := m^{-1/p} \sum_{j \in C_r} v_j, \quad r \in [t].$$

By Hölder,

$$|u_r| = m^{-1/p} \left| \sum_{j \in C_r} v_j \right| \leq \|v_{C_r}\|_q.$$

Hence,

$$\sum_{r=1}^t |u_r|^q \leq \sum_{r=1}^t \|v_{C_r}\|_q^q \leq \|v\|_q^q.$$

In particular, every v with $\|v\|_q \leq 1$ produces a vector u with $\|u\|_q \leq 1$. Moreover,

$$\sum_{j \in C_r} v_j = m^{1/p} u_r \quad \text{for all } r \in [t].$$

Therefore,

$$\begin{aligned} \tilde{F}(x^{(a)}) &= \max_{v: \|v\|_q=1} \left| a \sum_{r=1}^t \sum_{j \in C_r} v_j - \frac{\sqrt{d}}{m} \text{Med} \left(\left\{ \sum_{j \in C_r} v_j \right\}_{r=1}^t \right) \right| \\ &= m^{1/p} \max_{v: \|v\|_q=1} \left| a \sum_{r=1}^t u_r - \frac{\sqrt{d}}{m} \text{Med}(\{u_r\}_{r=1}^t) \right| \\ &\leq m^{1/p} \max_{u: \|u\|_q \leq 1} \left| a \sum_{r=1}^t u_r - \frac{\sqrt{d}}{m} \text{Med}(\{u_r\}_{r=1}^t) \right|. \end{aligned}$$

For the reverse inequality, fix any $u \in \mathbb{R}^t$ with $\|u\|_q \leq 1$, and define $v \in \mathbb{R}^d$ by setting

$$\forall r \in [t], \forall j \in C_r : v_j := m^{-1+1/p} u_r,$$

and $v_j := 0$ outside $\cup_{r=1}^t C_r$. Then

$$\sum_{j \in C_r} v_j = m^{1/p} u_r \quad \text{for all } r \in [t],$$

and

$$\|v\|_q^q = \sum_{r=1}^t \sum_{j \in C_r} m^{-q+q/p} |u_r|^q = \sum_{r=1}^t m^{(1-q)+q/p} |u_r|^q = \sum_{r=1}^t |u_r|^q = \|u\|_q^q \leq 1,$$

since $(1-q) = -q/p$ as $1/p + 1/q = 1$. Plugging v into the definition of $\tilde{F}$ now yields the reverse inequality, concluding the proof of the lemma. $\square$

Lemma 4.5.11. *For every $a \in \mathbb{R}$,*

$$\tilde{F}(x^{(a)}) = m^{1/p} \left(\rho \left| \frac{\sqrt{d}}{\rho m} - a \right|^p + \eta |a|^p \right)^{1/p}.$$

Proof. Let

$$\lambda := \frac{\sqrt{d}}{m}.$$

By Lemma 4.5.10, it suffices to compute

$$\max_{\|u\|_q \leq 1} \left| a \sum_{r=1}^t u_r - \lambda \text{Med}(\{u_r\}_{r=1}^t) \right|.$$

Since $\text{Med}(-u) = -\text{Med}(u)$, this maximum equals

$$\max_{\|u\|_q \leq 1} \left\{ \lambda \text{Med}(\{u_r\}_{r=1}^t) - a \sum_{r=1}^t u_r \right\}.$$

Let

$$u_{(1)} \geq u_{(2)} \geq \cdots \geq u_{(t)}$$

be the non-increasing rearrangement of u . Now, observe that:

$$\text{Med}(u) \leq \frac{1}{\rho} \sum_{r=1}^{\rho} u_{(r)}.$$

Therefore,

$$\lambda \text{Med}(u) - a \sum_{r=1}^t u_r \leq \frac{\lambda}{\rho} \sum_{r=1}^{\rho} u_{(r)} - a \sum_{r=1}^t u_{(r)} = \left(\frac{\lambda}{\rho} - a \right) \sum_{r=1}^{\rho} u_{(r)} - a \sum_{r=\rho+1}^t u_{(r)}.$$

By Hölder,

$$\lambda \text{Med}(u) - a \sum_{r=1}^t u_r \leq \left(\rho \left| \frac{\lambda}{\rho} - a \right|^p + \eta |a|^p \right)^{1/p} \|u\|_q.$$

Since $\|u\|_q \leq 1$, this establishes an upper bound.

For the lower bound, set

$$A := \frac{\lambda}{\rho} - a, \quad B := -a, \quad \alpha := (\rho |A|^p + \eta |B|^p)^{-1/q}, \text{ and } J(z) := \text{sgn}(z) |z|^{p-1}.$$

Now, define $u \in \mathbb{R}^t$ by

$$u_1 = \cdots = u_{\rho} = \alpha J(A), \quad u_{\rho+1} = \cdots = u_t = \alpha J(B).$$

Since $A - B > 0$, and J is increasing, this vector is non-increasing. Also, observe

$$\|u\|_q^q = \alpha^q (\rho |A|^p + \eta |B|^p) = 1 \text{ and } \text{Med}(u) = \alpha J(A).$$

Hence

$$\begin{aligned} \lambda \text{Med}(u) - a \sum_{r=1}^t u_r &= \lambda \alpha J(A) - a (\rho \alpha J(A) + \eta \alpha J(B)) = \alpha (\rho A J(A) + \eta B J(B)) \\ &= \alpha (\rho |A|^p + \eta |B|^p) = (\rho |A|^p + \eta |B|^p)^{1/p}. \end{aligned}$$

matching the upper bound and proving the lemma. $\square$

Corollary 4.5.12. *The optimal scalar is*

$$a^* = \frac{\sqrt{d}}{\rho m} \cdot \frac{1}{1 + (\eta/\rho)^{1/(p-1)}}.$$

Consequently, a minimizer of $\tilde{F}$ is

$$x^* = a^* \sum_{i=1}^n e_{c(i)},$$

and hence, by [Observation 4.5.4](#), a minimizer of F is

$$\hat{\mu}_{\text{proj}}(\mathbf{X}) = Sx^* = a^* \sum_{i=1}^n s_i e_{c(i)}.$$

Proof. Observing that $\tilde{F}(x^{(a)})$ is convex in a and computing its derivative, the corollary follows. $\square$

Corollary 4.5.13. *The resulting estimator has*

$$\|\hat{\mu}_{\text{proj}}(\mathbf{X})\|_p = \|x^*\|_p = a^* n^{1/p} \geq \frac{\sqrt{d}}{2n^{1-1/p}}.$$

Proof. Since x^* has exactly n nonzero coordinates, each of magnitude a^* ,

$$\|x^*\|_p = a^* n^{1/p}.$$

The corollary follows from [Corollary 4.5.12](#). $\square$

Proof of [Theorem 4.4.3](#). The theorem follows from [Corollary 4.5.13](#) and [Lemma 4.5.2](#). $\square$

Bibliography

- [1] Pedro Abdalla and Shahar Mendelson. “Covariance estimation with direction dependence accuracy”. In: *Probability Theory and Related Fields* 194.1–2 (2026), pp. 371–409. DOI: [10.1007/s00440-025-01376-7](https://doi.org/10.1007/s00440-025-01376-7). arXiv: [2402.08288](https://arxiv.org/abs/2402.08288) [math.ST].
- [2] Pedro Abdalla and Nikita Zhivotovskiy. “Covariance estimation: Optimal dimension-free guarantees for adversarial corruption and heavy tails”. In: *Journal of the European Mathematical Society* 28.4 (2026), pp. 1809–1847. DOI: [10.4171/JEMS/1505](https://doi.org/10.4171/JEMS/1505). URL: <https://doi.org/10.4171/JEMS/1505>.
- [3] Ishaq Aden-Ali et al. “Majority-of-Three: The Simplest Optimal Learner?” In: *Proceedings of Thirty Seventh Conference on Learning Theory*. Ed. by Shipra Agrawal and Aaron Roth. Vol. 247. Proceedings of Machine Learning Research. PMLR, June 2024, pp. 22–45. URL: <https://proceedings.mlr.press/v247/aden-ali24a.html>.
- [4] Ishaq Aden-Ali et al. “Optimal PAC Bounds without Uniform Convergence”. In: *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE Computer Society, 2023, pp. 1203–1223.
- [5] Ishaq Aden-Ali et al. “Phase Transitions in Heavy-Tailed Mean Estimation under ℓ_p Norms”. Preprint. 2026.
- [6] Ishaq Aden-Ali et al. “The One-Inclusion Graph Algorithm is not Always Optimal”. In: *The Thirty Sixth Annual Conference on Learning Theory, COLT 2023*. Vol. 195. Proceedings of Machine Learning Research. PMLR, 2023, pp. 72–88. URL: <https://proceedings.mlr.press/v195/aden-ali23a.html>.
- [7] Mark Aizerman, Emmanuil Braverman, and Lev Rozonoer. “Theoretical foundations of the potential function method in pattern recognition learning”. English. In: *Automation and Remote Control* 25 (1965), pp. 821–837.
- [8] Noga Alon, Yossi Matias, and Mario Szegedy. “The space complexity of approximating the frequency moments”. In: *Journal of Computer and System Sciences* 58.1 (1999), pp. 137–147. DOI: [10.1006/jcss.1997.1545](https://doi.org/10.1006/jcss.1997.1545). URL: <https://doi.org/10.1006/jcss.1997.1545>.
- [9] Noga Alon and Michael Tarsi. “Colorings and orientations of graphs”. In: *Combinatorica* 12.2 (1992), pp. 125–134.

- [10] Noga Alon et al. “A theory of PAC learnability of partial concept classes”. In: *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2022, pp. 658–671.
- [11] Noga Alon et al. “Scale-sensitive dimensions, uniform convergence, and learnability”. In: *Journal of the ACM* 44.4 (1997), pp. 615–631.
- [12] T. W. Anderson. “The integral of a symmetric unimodal function over a symmetric convex set and some probability inequalities”. In: *Proceedings of the American Mathematical Society* 6.2 (1955), pp. 170–176.
- [13] Martin Anthony and Peter Bartlett. *Neural Network Learning: Theoretical Foundations*. Vol. 9. Cambridge University Press, 1999.
- [14] Peter Auer. “Learning nested differences in the presence of malicious noise”. In: *Theoretical Computer Science* 185.1 (1997), pp. 159–175.
- [15] Peter Auer and Ronald Ortner. “A new PAC bound for intersection-closed concept classes”. In: *Machine Learning* 66.2 (2007), pp. 151–163.
- [16] Maria-Florina Balcan and Phil Long. “Active and passive learning of linear separators under log-concave distributions”. In: *Conference on Learning Theory*. 2013, pp. 288–316.
- [17] Daniel Bartl and Shahar Mendelson. “Uniform mean estimation via generic chaining”. In: *Advances in Mathematics* 493, 110918 (2026). DOI: [10.1016/j.aim.2026.110918](https://doi.org/10.1016/j.aim.2026.110918). arXiv: [2502.15116](https://arxiv.org/abs/2502.15116) [math.PR].
- [18] Peter L Bartlett and Philip M Long. “Prediction, learning, uniform convergence, and scale-sensitive dimensions”. In: *Journal of Computer and System Sciences* 56.2 (1998), pp. 174–190.
- [19] Shai Ben-David et al. “Characterizations of Learnability for Classes of $\{0, \dots, n\}$ -Valued Functions”. In: *Journal of Computer and System Sciences* 50.1 (1995), pp. 74–86.
- [20] Anselm Blumer and Nick Littlestone. “Learning faster than promised by the Vapnik-Chervonenkis dimension”. In: *Discrete Applied Mathematics* 24.1-3 (1989), pp. 47–53.
- [21] Anselm Blumer et al. “Learnability and the Vapnik-Chervonenkis Dimension”. In: *Journal of the ACM* 36.4 (1989), pp. 929–965.
- [22] Olivier Bousquet et al. “Proper learning, Helly number, and an optimal SVM bound”. In: *Conference on Learning Theory*. PMLR. 2020, pp. 582–609.
- [23] Leo Breiman. “Bagging Predictors”. In: *Mach. Learn.* 24.2 (Aug. 1996), pp. 123–140. ISSN: 0885-6125. DOI: [10.1023/A:1018054314350](https://doi.org/10.1023/A:1018054314350). URL: <https://doi.org/10.1023/A:1018054314350>.
- [24] Nataly Brukhim et al. “A Characterization of Multiclass Learnability”. In: *arXiv preprint arXiv:2203.01550* (2022).

- [25] Nader H Bshouty, Yi Li, and Philip M Long. “Using the doubling dimension to analyze the generalization of learning algorithms”. In: *Journal of Computer and System Sciences* 75.6 (2009), pp. 323–335.
- [26] Olivier Catoni. “Challenging the Empirical Mean and Empirical Variance: A Deviation Study”. In: *Annales de l’Institut Henri Poincaré, Probabilités et Statistiques*. Vol. 48. 4. 2012, pp. 1148–1185.
- [27] Olivier Catoni and Ilaria Giulini. *Dimension-free PAC-Bayesian bounds for matrices, vectors, and linear least squares regression*. 2017. DOI: [10.48550/arXiv.1712.02747](https://doi.org/10.48550/arXiv.1712.02747). arXiv: [1712.02747](https://arxiv.org/abs/1712.02747) [math.ST]. URL: <https://arxiv.org/abs/1712.02747>.
- [28] Yeshwanth Cherapanamjeri, Nicolas Flammarion, and Peter L. Bartlett. “Fast Mean Estimation with Sub-Gaussian Rates”. In: *Proceedings of the Thirty-Second Conference on Learning Theory*. Vol. 99. Proceedings of Machine Learning Research. PMLR, 2019, pp. 786–806. URL: <https://proceedings.mlr.press/v99/cherapanamjeri19b.html>.
- [29] Amit Daniely and Shai Shalev-Shwartz. “Optimal learners for multiclass problems”. In: *Conference on Learning Theory*. PMLR. 2014, pp. 287–316.
- [30] Amit Daniely et al. “Multiclass learnability and the ERM principle.” In: *J. Mach. Learn. Res.* 16.1 (2015), pp. 2377–2404.
- [31] Malte Darnstädt. “The optimal PAC bound for intersection-closed concept classes”. In: *Information Processing Letters* 115.4 (2015), pp. 458–461.
- [32] Jules Depersin and Guillaume Lecué. “Optimal Robust Mean and Location Estimation via Convex Programs with Respect to Any Pseudo-Norms”. In: *Probability Theory and Related Fields* 183.3 (2022), pp. 1005–1056.
- [33] Ilias Diakonikolas, Daniel M. Kane, and Ankit Pensia. “Outlier Robust Mean Estimation with Subgaussian Rates via Stability”. In: *Advances in Neural Information Processing Systems*. Vol. 33. 2020, pp. 1830–1840. URL: <https://proceedings.neurips.cc/paper/2020/hash/13ec9935e17e00bed6ec8f06230e33a9-Abstract.html>.
- [34] Andrzej Ehrenfeucht et al. “A General Lower Bound on the Number of Examples Needed for Learning”. In: *Inf. Comput.* 82.3 (1989), pp. 247–261.
- [35] Sally Floyd and Manfred Warmuth. “Sample compression, learnability, and the Vapnik-Chervonenkis dimension”. In: *Machine learning* 21.3 (1995), pp. 269–304.
- [36] Steve Hanneke. “The Optimal Sample Complexity of PAC Learning”. In: *Journal of Machine Learning Research* 17.38 (2016), pp. 1–15.
- [37] Steve Hanneke. “Theoretical Foundations of Active Learning”. Doctoral thesis. Carnegie-Mellon University, Machine Learning Department, 2009.
- [38] Steve Hanneke and Aryeh Kontorovich. “Stable sample compression schemes: New applications and an optimal SVM margin bound”. In: *Algorithmic Learning Theory*. PMLR. 2021, pp. 697–721.

- [39] Nicholas Harvey et al. “Tight analyses for non-smooth stochastic gradient descent”. In: *Conference on Learning Theory*. 2019, pp. 1579–1613.
- [40] David Haussler. “Sphere packing numbers for subsets of the Boolean n -cube with bounded Vapnik-Chervonenkis dimension”. In: *Journal of Combinatorial Theory, Series A* 69.2 (1995), pp. 217–232.
- [41] David Haussler, Nick Littlestone, and Manfred Warmuth. “Predicting $\{0, 1\}$ -functions on randomly drawn points (Extended Abstract)”. In: *29th Annual Symposium on Foundations of Computer Science (FOCS)*. 1988, pp. 100–109.
- [42] David Haussler, Nick Littlestone, and Manfred K. Warmuth. “Predicting 0-1 Functions on Randomly Drawn Points”. In: *Information and Computation* 115.2 (1994), pp. 248–292.
- [43] David Haussler and Philip M Long. “A generalization of Sauer’s lemma”. In: *Journal of Combinatorial Theory, Series A* 71.2 (1995), pp. 219–240.
- [44] Samuel B. Hopkins. “Mean estimation with sub-Gaussian rates in polynomial time”. In: *The Annals of Statistics* 48.2 (2020), pp. 1193–1213. DOI: [10.1214/19-AOS1843](https://doi.org/10.1214/19-AOS1843). URL: <https://doi.org/10.1214/19-AOS1843>.
- [45] Samuel B. Hopkins, Jerry Li, and Fred Zhang. “Robust and Heavy-Tailed Mean Estimation Made Simple, via Regret Minimization”. In: *Advances in Neural Information Processing Systems*. Vol. 33. 2020, pp. 11902–11912. URL: <https://proceedings.neurips.cc/paper/2020/hash/8a1276c25f5efe85f0fc4020fbf5b4f8-Abstract.html>.
- [46] Svante Janson. “Tail bounds for sums of geometric and exponential variables”. In: *Statistics and Probability Letters* 135 (2018), pp. 1–6.
- [47] Mark R. Jerrum, Leslie G. Valiant, and Vijay V. Vazirani. “Random generation of combinatorial structures from a uniform distribution”. In: *Theoretical Computer Science* 43.2–3 (1986), pp. 169–188. DOI: [10.1016/0304-3975\(86\)90174-X](https://doi.org/10.1016/0304-3975(86)90174-X). URL: [https://doi.org/10.1016/0304-3975\(86\)90174-X](https://doi.org/10.1016/0304-3975(86)90174-X).
- [48] Alkis Kalavasis, Grigoris Velegkas, and Amin Karbasi. “Multiclass Learnability Beyond the PAC Framework: Universal Rates and Partial Concept Classes”. In: *arXiv preprint arXiv:2210.02297* (2022).
- [49] Michael J Kearns and Robert E Schapire. “Efficient distribution-free learning of probabilistic concepts”. In: *Journal of Computer and System Sciences* 48.3 (1994), pp. 464–497.
- [50] Pieter Kleer and Hans Simon. “Primal and dual combinatorial dimensions”. In: *Discrete Applied Mathematics* 327 (2023), pp. 185–196.
- [51] Wojciech Kotłowski, Wouter M Koolen, and Alan Malek. “Random permutation online isotonic regression”. In: *Advances in Neural Information Processing Systems* 30 (2017).

- [52] Kasper Green Larsen. “Bagging is an Optimal PAC Learner”. In: *Proceedings of Thirty Sixth Conference on Learning Theory*. Vol. 195. Proceedings of Machine Learning Research. 2023, pp. 450–468.
- [53] Kasper Green Larsen, Rasmus Pagh, and Jakub Tetek. “CountSketches, Feature Hashing and the Median of Three”. In: *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*. Ed. by Marina Meila and Tong Zhang. Proceedings of Machine Learning Research. PMLR, 2021, pp. 6011–6020. URL: <http://proceedings.mlr.press/v139/larsen21a.html>.
- [54] Yi Li, Philip M. Long, and Aravind Srinivasan. “The One-Inclusion Graph Algorithm is Near-Optimal for the Prediction Model of Learning”. In: *IEEE Transactions on Information Theory* 47.3 (2001), pp. 1257–1261.
- [55] Nick Littlestone. “From On-Line to Batch Learning”. In: *Proceedings of the Second Annual Workshop on Computational Learning Theory, COLT 1989, July 31 - August 2, 1989*. Ed. by Ronald L. Rivest, David Haussler, and Manfred K. Warmuth. Morgan Kaufmann, 1989, pp. 269–284.
- [56] Nick Littlestone. “Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm”. In: *Machine learning* 2 (1988), pp. 285–318.
- [57] Philip M Long. “An upper bound on the sample complexity of PAC-learning halfspaces with respect to the uniform distribution”. In: *Information Processing Letters* 87.5 (2003), pp. 229–234.
- [58] Philip M Long. “On agnostic learning with $\{0, \star, 1\}$ -valued and real-valued hypotheses”. In: *Conference on Computational Learning Theory*. 2001, pp. 289–302.
- [59] Gábor Lugosi and Shahar Mendelson. “Near-Optimal Mean Estimators with Respect to General Norms”. In: *Probability Theory and Related Fields* 175.3–4 (2019), pp. 957–973.
- [60] Gábor Lugosi and Shahar Mendelson. “Robust multivariate mean estimation: The optimality of trimmed mean”. In: *The Annals of Statistics* 49.1 (2021), pp. 393–410. DOI: [10.1214/20-AOS1961](https://doi.org/10.1214/20-AOS1961). URL: <https://doi.org/10.1214/20-AOS1961>.
- [61] Gábor Lugosi and Shahar Mendelson. “Sub-Gaussian Estimators of the Mean of a Random Vector”. In: *The Annals of Statistics* 47.2 (2019), pp. 783–794.
- [62] Alexander Lunts and Victor Brailovsky. “Evaluation of attributes obtained in statistical decision rules”. In: *Engineering Cybernetics* 3 (1967), pp. 98–109.
- [63] Tudor Manole and Aaditya Ramdas. “Martingale Methods for Sequential Estimation of Convex Functionals and Divergences”. In: *IEEE Transactions on Information Theory* (2023).
- [64] Pascal Massart and Élodie Nédélec. “Risk bounds for statistical learning”. In: *The Annals of Statistics* 34.5 (2006).

- [65] Shahar Mendelson and Roman Vershynin. “Entropy and the combinatorial dimension”. In: *Inventiones mathematicae* 152.1 (2003), pp. 37–55.
- [66] Stanislav Minsker. “Geometric Median and Robust Estimation in Banach Spaces”. In: *Bernoulli* 21.4 (2015), pp. 2308–2335.
- [67] Stanislav Minsker. “Uniform bounds for robust mean estimators”. In: *Stochastic Processes and their Applications* 190 (2025), p. 104724. DOI: [10.1016/j.spa.2025.104724](https://doi.org/10.1016/j.spa.2025.104724). URL: <https://doi.org/10.1016/j.spa.2025.104724>.
- [68] B. K. Natarajan. “On learning Boolean functions”. In: *Proceedings of the Nineteenth Annual ACM Symposium on Theory of Computing*. STOC ’87. New York, New York, USA: Association for Computing Machinery, 1987, pp. 296–304. ISBN: 0897912217.
- [69] Balas K Natarajan. “On learning sets and functions”. In: *Machine Learning* 4 (1989), pp. 67–97.
- [70] Balas K Natarajan and Prasad Tadepalli. “Two new frameworks for learning”. In: *Machine Learning Proceedings 1988*. Elsevier, 1988, pp. 402–415.
- [71] A. S. Nemirovsky and D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience Series in Discrete Mathematics. Chichester: John Wiley & Sons, 1983. ISBN: 9780471103455.
- [72] David Pollard. “Limit theorems for empirical processes”. In: *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 57.2 (1981), pp. 181–195. ISSN: 0044-3719. DOI: [10.1007/BF00535488](https://doi.org/10.1007/BF00535488). URL: <https://doi.org/10.1007/BF00535488>.
- [73] Yury Polyanskiy and Yihong Wu. *Information Theory: From Coding to Learning*. Cambridge University Press, 2025.
- [74] Benjamin Rubinstein, Peter Bartlett, and Hyam Rubinstein. “Shifting: One-inclusion mistake bounds and sample compression”. In: *Journal of Computer and System Sciences* 75.1 (2009), pp. 37–59.
- [75] Mark Rudelson and Roman Vershynin. “Combinatorics of random processes and sections of convex bodies”. In: *Annals of Mathematics* (2006), pp. 603–648.
- [76] Robert E. Schapire. *The Design and Analysis of Efficient Learning Algorithms*. ACM Doctoral Dissertation Awards. The MIT Press, 1992, p. 231. ISBN: 9780262193252.
- [77] Hans U. Simon. “An Almost Optimal PAC Algorithm”. In: *Proceedings of the 28th Annual Conference on Learning Theory*. Vol. 40. Proceedings of Machine Learning Research. 2015, pp. 1552–1563.
- [78] Hans U Simon and Balázs Szörényi. “One-inclusion hypergraph density revisited”. In: *Information processing letters* 110.8-9 (2010), pp. 341–344.
- [79] Michel Talagrand. “The Glivenko-Cantelli problem, ten years later”. In: *Journal of Theoretical probability* 9 (1996), pp. 371–384.

- [80] Michel Talagrand. “Vapnik–Chervonenkis type conditions and uniform Donsker classes of functions”. In: *The Annals of Probability* 31.3 (2003), pp. 1565–1582.
- [81] Leslie G. Valiant. “A Theory of the Learnable”. In: *Communications of the ACM* 27.11 (1984), pp. 1134–1142.
- [82] Aad Van der Vaart and Jon Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.
- [83] Vladimir N. Vapnik and Alexey Y. Chervonenkis. “On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities”. In: *Theory of Probability and its Applications* 16.2 (1971), pp. 264–280.
- [84] Vladimir Vapnik and Alexey Chervonenkis. “A class of algorithms for pattern recognition learning”. In: *Avtomatika i Telemekhanika* 25.6 (1964), pp. 937–945.
- [85] Vladimir Vapnik and Alexey Chervonenkis. “Algorithms with complete memory and recurrent algorithms in the problem of learning pattern recognition”. In: *Avtomatika i Telemekhanika* (4 1968), pp. 95–106.
- [86] Vladimir Vapnik and Alexey Chervonenkis. “Necessary and sufficient conditions for the uniform convergence of means to their expectations”. In: *Theory of Probability & Its Applications* 26.3 (1982), pp. 532–553.
- [87] Vladimir Vapnik and Alexey Chervonenkis. *Theory of Pattern Recognition*. Moscow: Nauka, 1974.
- [88] Vladimir Vapnik and Alexey Chervonenkis. “Theory of uniform convergence of frequencies of appearance of attributes to their probabilities and problems of defining optimal solution by empiric data”. In: *Avtomatika i Telemekhanika* 2 (1971), pp. 42–53.
- [89] Rom Varshamov and Grigory Tenengolts. “Code Correcting Single Asymmetric Errors”. In: *Avtomatika i Telemekhanika* 26.2 (1965), pp. 288–292.
- [90] Manfred K Warmuth. “The optimal PAC algorithm”. In: *International Conference on Computational Learning Theory*. Springer. 2004, pp. 641–642.
- [91] Changlong Wu et al. “Expected Worst Case Regret via Stochastic Sequential Covering”. In: *arXiv preprint arXiv:2209.04417* (2022).
- [92] Yuhong Yang and Andrew Barron. “Information-theoretic determination of minimax rates of convergence”. In: *Annals of Statistics* (1999), pp. 1564–1599.
- [93] Nikita Zhivotovskiy. “Optimal learning via local entropies and sample compression”. In: *Conference on Learning Theory*. PMLR. 2017, pp. 2023–2065.