Bounds and Applications of Concentration of Measure in Fair Machine Learning and Data Science

by

Cyrus Cousins

Sc.M., Brown University, 2017

Sc.B., Tufts University, 2015

A dissertation submitted in partial fulfillment of the

requirements for the Degree of Doctor of Philosophy

in the Department of Computer Science at Brown University

Providence, Rhode Island

May 2021

This dissertation by Cyrus Cousins is accepted in its present form by
the Department of Computer Science as satisfying the dissertation requirement
for the degree of Doctor of Philosophy.

Date ______________ ________________________________
Eli Upfal, Director

Recommended to the Graduate Council

Date ______________ ________________________________
Michael Littman, Reader

Date ______________ ________________________________
Matteo Riondato, Reader

Approved by the Graduate Council

Date ______________ ________________________________
Andrew G. Campbell
Dean of the Graduate School

Curriculum Vitae

A native of Providence, RI, Cyrus showed an interest in computers and mathematics from an early age. He pursued his baccalaureate studies at Tufts University, and whilst there, he served as a teaching assistant in various courses, ranging from computational geometry to bioinformatics. He also took his first steps as a researcher, working with professors Donna Slonim, Norman Ramsey, and Mitchell Wand, on a variety of competition biology and inference problems. In 2015, Cyrus graduated, earning his Bachelor's degree in computer science, mathematics, and biology, and also earning highest honors on his senior thesis, on the subject of information-theoretic anomaly detection in biological systems. With an infectious thirst for knowledge, he immediately matriculated to Brown University to pursue his doctoral studies the next autumn. At Brown University, Cyrus served as a graduate teaching assistant for CS1810: *Computational Molecular Biology*, DATA2040: *Deep Learning and Special Topics in Data Science*, and CS1450: *Probability for Computing and Data Analysis*. He earned his master's degree in computer science from Brown University in 2017 en route to his Ph.D. In the summer of 2018, he took a position as research intern at Two Sigma labs, working with Matteo Riondato on statistical machine learning methods, and returned the following summer to work with Larry Rudolph on problems in multi-agent reinforcement learning. Throughout his time at Brown University, he published with myriad professors on a gamut of problems, ranging from empirical game theory to statistical learning theory, though a dedication to rigorous mathematical analysis of practically-relevant sampling problems echoes throughout his body of work. In 2021, Cyrus was awarded the Dean's Faculty Fellowship at Brown University.

Conference and Journal Publications

1. Cyrus Cousins and Matteo Riondato. Sharp uniform convergence bounds through empirical centralization. In *Advances in Neural Information Processing Systems*, 2020
2. Leonardo Pellegrina, Cyrus Cousins, Fabio Vandin, and Matteo Riondato. MCRapper: Monte-Carlo Rademacher averages for POSET families and approximate pattern mining. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2165–2174, 2020

3. Enrique Areyan Viqueira, Cyrus Cousins, and Amy Greenwald. Improved algorithms for learning equilibria in simulation-based games. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, pages 79–87, 2020
4. Enrique Areyan Viqueira, Cyrus Cousins, Yasser Mohammad, and Amy Greenwald. Empirical mechanism design: Designing mechanisms from data. In *Uncertainty in Artificial Intelligence*, pages 1094–1104. PMLR, 2020
5. Cyrus Cousins and Matteo Riondato. CaDET: interpretable parametric conditional density estimation with decision trees and forests. *Machine Learning*, 108(8-9):1613–1634, 2019
6. Enrique Areyan Viqueira, Cyrus Cousins, and Amy Greenwald. Learning simulation-based games from data. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, 2019
7. Cyrus Cousins and Eli Upfal. The k-nearest representatives classifier: A distance-based classifier with strong generalization bounds. In *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10. IEEE, 2017

Workshop Papers, Notable Preprints, and Extended Abstracts

1. Cyrus Cousins. An axiomatic theory of provably-fair welfare-centric machine learning. *arXiv preprint arXiv:2104.14504*, 2021
2. Enrique Areyan Viqueira, Cyrus Cousins, and Amy Greenwald. Learning competitive equilibria in noisy combinatorial markets. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, 2021
3. Cyrus Cousins, Shahrzad Haddadan, and Eli Upfal. Making mean-estimation more efficient using an MCMC trace variance approach: DynaMITE. *arXiv preprint arXiv:2011.11129*, 2020
4. Carsten Binnig, Benedetto Buratti, Yeounoh Chung, Cyrus Cousins, Tim Kraska, Zeyuan Shang, Eli Upfal, Robert Zeleznik, and Emanuel Zgraggen. Towards interactive curation & automatic tuning of ML pipelines. In *Proceedings of the Second Workshop on Data Management for End-To-End Machine Learning*, pages 1–4, 2018
5. Clayton Sanford, Cyrus Cousins, and Eli Upfal. Uniform convergence bounds for codec selection. *arXiv preprint arXiv:1812.07568*, 2018
6. Cyrus Cousins, Chirstopher M Pietras, and Donna K Slonim. Scalable FRaC variants: Anomaly detection for precision medicine. In *2017 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, pages 253–262. IEEE, 2017

Preface and Acknowledgements

I dedicate this thesis to my parents, my coauthors, my colleagues, and my mentors, without whom I would not have been able to complete this work. In particular, Shahrzad Haddadan, Amy Greenwald, Enrique Areyan Viqueira, Alessio Mazzetto, Clayton Sanford, Eli Upfal, Larry Rudolph, and Matteo Riondato have all worked tirelessly on many submissions, which have made their way into this thesis in one form or another. I would also like to thank Amir Dib, Kavosh Asadi, and Cristina Menghini for their insightful discussion and collaborative effort in exploring many topics, though our work has not made it into the thesis. Finally I wish to thank Eli Upfal, Shahrzad Haddadan, and Amir Dib a second time, for helping to connect my work with receptive audiences, for tolerating my flaws, and for always pushing me towards betterment.

Contents

The Prelude

In this thesis, I introduce novel concentration-of-measure bounds for the *supremum deviation*, several *variance concepts*, and a family of *game-theoretic welfare functions*. It is divided into three parts, the first focusing on statistical methods and machine learning, followed by fair machine learning, and concluding with non-i.i.d. statistical estimation tasks. In particular, I introduce *empirically centralized Rademacher averages* to probabilistically bound the deviation between the *empirical* and *true* means over a *family of functions*, with applications to *multiple comparisons problems* in statistical and scientific settings (smaller p-values and tighter confidence intervals), and to various supervised and unsupervised machine learning settings (reduced sample complexity and sharper generalization bounds). I then show applications of these bounds to various machine-learning, fair machine-learning, and data-science settings, with deep theoretical implications and impactful practical consequences. Parts I and II assume an *independently and identically distributed* (i.i.d.) setting, where we observe many statistically independent occurrences before drawing conclusions, but in closing, part III extends some of my methods and themes to study non-i.i.d. mean-estimation problems. Naturally, some conclusions are weaker in these relaxed settings, but I find that many of the same data-dependent and variance-sensitivity themes apply, and give practical algorithms for realistic problems where the i.i.d. assumption is prohibitive.

Special attention is made throughout this thesis to connect complicated bounds and ideas to simple and intuitive concepts (like the central limit theorem or linear regression), while explaining and rigorously justifying all finite-sample probabilistic guarantees, as well as assumptions and proof techniques. Consequently, at the heart of this thesis is the marriage of simple ideas about intuitive random processes to sophisticated techniques and bounds from the theory of concentration of measure and probabilistic methods, applied in complicated settings to a variety of learning and statistical estimation problems. Topics are chosen for both their theoretical relevance and how well they illustrate and connect such ideas, but also for their practical relevance and applicability to high-impact real-world problems. Note also that, except when sufficiently succinct and pedagogically pertinent, all proofs are relegated to the appendices.

Part I: Concentration of Measure and Uniform Convergence in Machine Learning Part I deals primarily with statistical estimation guarantees in standard machine-learning settings. Chapter 1 introduces the *empirically centralized Rademacher average*, which yields probabilistic data-dependent bounds on the *supremum deviation* (SD) of empirical means of functions in a family $\mathcal{F}$ from their expectations (i.e., $\sup_{f \in \mathcal{F}} \left| \hat{\mathbb{E}}[f] - \mathbb{E}[f] \right|$), with optimal dependence on the *supremum variance* and the function ranges. Such bounds are impactful in machine learning, as they bound the *generalization gap* between training and test error, and thus control overfitting and quantify the bias-complexity tradeoff (i.e., the tension at the heart of machine learning between *better fit* and *increased bias* as model complexity grows). More generally, such uniform convergence bounds have applications to *multiple comparisons problems* in statistical and scientific settings (smaller p-values and tighter confidence intervals), and to various supervised and unsupervised machine learning settings (reduced sample complexity and sharper generalization bounds). Empirical centralization yields *data-dependent* supremum deviation bounds that improve the dependence of *non-centralized* (standard) Rademacher averages on *raw variances* to *centralized variances*, thus matching (asymptotically) known lower-bounds for mean estimation. To compute the bounds in practice, I develop novel tightly-concentrated Monte-Carlo estimators for the empirical Rademacher average of the empirically-centralized family, and show novel concentration results for the empirical wimpy variance. My experimental evaluation shows that these bounds greatly outperform their non-centralized counterparts, and are extremely practical, even at small sample sizes.

A major issue, ubiquitous in supervised learning settings, is the cost of obtaining labeled training data. The methods of chapter 1 may be viewed as a mechanism to get more out of *limited labeled data*, i.e., by showing that far less of it is required than previously thought, but chapter 2 takes a complimentary and orthogonal approach. In chapter 2, I adopt a *transductive learning* setting, wherein we have a vast set of points, a tiny fraction of which are labeled, and wish to learn to predict the remaining labels. I then describe a transductive learning algorithm that uses these data, alongside additional knowledge in the form of *weak labelers*, which are arbitrarily inaccurate predictors for either the target task or some related task. In particular, I show that accurate learning is possible with a small labeled training set when (a subset of) the family of weak labelers is somewhat well-aligned with the target task. The strategy here is to estimate and (probabilistically) bound appropriately-chosen statistics of the weak labelers using the supervised data, and then consider a *feasible set* of *possible labelings* for the *unlabeled data* that respect these statistics. When sufficient labeled data are available to well-estimate the chosen family of statistics, and the feasible label space is of low diameter, learning the *minimax-optimal* classifier over this feasible set then yields strong generalization guarantees. In particular, if the weak-labeler statistics give sufficient information about the target task, and we have sufficient labeled data to accurately constrain their statistics, then all *feasible labelings* are reasonably accurate, and if sufficiently many *unlabeled data* are available, then we can learn a complicated model; these conditions neatly factor the labeled and unlabeled sample complexities, and describe conditions under which weak-labelers are sufficient to supplement a small labeled dataset.

Part II: Fairness with Population Means: Malfare, Welfare, and Fair PAC Learning In part II, I begin with an axiomatic justification to the *power mean* family of welfare functions, which summarize societal wellbeing, while making tradeoffs between the needs of the overall population and of marginalized groups (cf., *utilitarianism* versus *egalitarianism*). From the same axiomatization, starting with a measure of *discontentment* (loss) rather than *contentment* (utility), I derive the parallel concept of *malfare*. For linear welfare functions, malfare acts as the *negative welfare* of the *negative utility*, however *nonlinear* welfare functions (e.g., the egalitarian welfare, or the Nash social welfare) may be *undefined* for negative utilities, whereas nonlinear malfare functions may be applied directly to loss or risk values. This is crucial to fairness in ML, where we generally seek to minimize loss (rather than maximize utility), and we require nonlinear power-mean malfare functions to specify fairness tradeoffs. These arguments are strongly grounded in the economic theory of *cardinal welfare*, but from them I show statistical estimation and learning guarantees more characteristic of the computer science literature. In particular, malfare is a natural target in machine learning problems where we *minimize* (negatively connoted) loss, rather than *maximize* (positively connoted) utility.

As an application, I cast a *streaming-media codec-selection* problem as a *fairness-sensitive* learning problem, wherein we seek to *efficiently select* a *small set* of media-encoders that can mutually satisfy a userbase with *diverse preferences* (e.g., quality versus bandwidth consumption). Optimizing welfare objectives maximize the impact of each selected codec, whereas without considering fairness and welfare objectives, it's easy to optimize only for a target demographic, or under invalid assumptions on users, thus potentially discriminating against some groups of users. This is an important *accessibility issue*, as these types of considerations ensure that audiovisual streaming and telecommunications services effectively serve populations that are often sidelined by digital services, including those with limited internet access, as well as those with various audiovisual perception conditions. I explore various *welfare* and *Pareto* optimality concepts, and how the bias-complexity tradeoff manifests in multivariate settings and with fairness issues.

From a more theoretical angle, I also show *statistical estimation guarantees* for welfare and malfare, and from the *social planning problem*, develop a theory of fair machine learning, based on the *probably approximately correct* (PAC) learning framework, termed *fair-PAC* (FPAC) learning. An FPAC-learner is an algorithm that learns an ε-δ malfare-optimal model with bounded sample complexity, for *any data distribution*, and for *any* (axiomatically justified) malfare concept. We show broad conditions under which, with appropriate modifications, many standard PAC-learners may be converted to FPAC learners. This places FPAC learning on firm theoretical ground, as it yields statistical, and in some cases computational, efficiency guarantees for many well-studied machine-learning models. FPAC-learning is also practically relevant, as it democratizes fair machine learning, by providing concrete training algorithms and rigorous generalization guarantees for these models.

Part III: Sample-Efficient Mean-Estimation with Dependent Sequential Data Finally, part III extends the methods and themes of the previous parts, where I assume an *independently and identically distributed* (i.i.d.) sample, into more general *weakly dependent* settings. As in the *data-dependent* guarantees with *empirically centralized Rademacher averages*, the goal here is to get the *strongest guarantees* under the *weakest assumptions*: in particular, this means my algorithms must be sensitive to structure found *in the data*. While notions of *approximate independence* can be difficult to rigorously bound from data, I find that appropriate *variance concepts* are easy to bound, and are sufficient to obtain asymptotically near-optimal sample-complexity guarantees for *mean estimation* in two dependent settings. In particular, I first examine *block databases*, where the key assumption is that *contiguous blocks of records* can be accessed nearly as efficiently as *individual records*, and second, I examine *Markov chains*, where each step of the chain is a *memoryless* random variable (i.e., conditionally independent from its history given the previous step), and it is as easy to collect a *trace of dependent* (often correlated) samples as to collect a pair of near-independent samples.

In both cases, I give *data-dependent* algorithms for ε-δ mean-estimation that avoid *worst-case* sample-complexity behavior. In particular, by considering not just independent or near-independent samples (as guaranteed via structural assumptions), but instead all of the dependent samples (available at no extra cost), I find that often the *variance* of the mean estimate decreases; particularly so when the data are less dependent than indicated by *a priori* structural assumptions. This decrease in variance implies more rapid convergence of empirical mean estimates by various central limit theorems, though this work undertakes the significantly more challenging endeavor of showing commensurate improvements to *finite-sample guarantees* on convergence rates. Furthermore, as I do not assume *a priori* knowledge of the appropriate variance concepts, sufficient sample sizes are not known *a priori*, and thus I employ a *progressive sampling strategy* to avoid drawing too many samples. The details of variance-estimation and necessary multiple-comparisons corrections entailed are subtle, but intuitively, this acts as a "guess and check" method, wherein we optimistically select an initial small sample size, which I use to estimate variances and means, and increase it until a sufficiently large sample has been drawn to provide the desired guarantee. Surprisingly, despite being variance-oblivious, these strategies are asymptotically optimal, up to logarithmic factors, in both the block-database and Markov-chain settings. In both cases, *a priori* guarantees on the amount of dependence do appear in our bounds, but only transiently, and as the additive error is taken to 0, terms involving only variance, which is estimated entirely from the data, come to dominate. This calls into question the importance of difficult-to-bound quantities measuring the degree of dependence.

This Work, as a Whole Read separately, each part represents a significant advancement in its respective field; the first in *statistical learning theory*, with an emphasis on *algorithms* and *generalization guarantees* requiring less *labeled data* than previous methods; the second in the *axiomatic philosophy*, *practice*, and *statistical learning guarantees* of *fair machine learning*; and the third in showing that themes and bound forms from *sampling problems* in standard i.i.d. settings translate well into non i.i.d. settings (with some additional work). Special attention is paid to keep each part readable outside of the greater context of this work, however the reader will better appreciate thematic connections, applications, and technical synergy when they are considered as a whole.

To better appreciate the connections and common themes running through this thesis, the preface of each part contains a preview of the concentration inequalities derived within. In particular, I develop a leitmotif of analogues of the Hoeffding and Bennett inequalities for bounded random variables, in the i.i.d. mean-estimation, malfare estimation, and non-i.i.d. mean-estimation settings, laying bare the common threads between each part. In their original forms, both results characterize the rate (with exact constants) that i.i.d. sums of *bounded random variables* converges to its mean. The Bennett inequality yields sub-Poisson (Poisson-like) or sub-gamma tails, with asymptotic behavior dominated by the *variance* of the random variables, and the Hoeffding inequality yields similar sub-Gaussian tails, in terms of the *worst-case* variance bound ${}^{r^2}/_4$, where r is the range of the random variables in question. Where appropriate, I also discuss more sophisticated uniform-convergence bounds, and how they manifest in each setting. Such bounds are important in many settings, as they allow us to study the simultaneous convergence of not just one or a handful of functions to their means, but rather large and even infinite sets of functions, without relying on union bounds (a.k.a. the *Bonferroni correction*). In the first part, these bounds are straightforward, as we consider risk (mean loss) in machine-learning settings with i.i.d. training data. The second part generalizes these results to consider *malfare*, which is a *nonlinear function* of risk over multiple groups, but we still obtain similar range, variance, and Rademacher average dependent tail bounds in this setting. Finally, in part three, we relinquish the i.i.d. assumption, which complicates matters greatly, but ultimately we attain similar bounds, now with additional terms capturing the *degree of (approximate) independence*.

To further solidify the connections between the three main parts, I note that the methods of part I and part III are not mutually incompatible, opening the door to strong *uniform convergence bounds* in *non-i.i.d.* settings, and furthermore, the *fair learning setting* of part II obviously benefits from the statistical bounds of part I, but similar analysis is certainly possible in non-i.i.d. settings. Indeed, in a connected and interdependent world, it may be the case that practical fair systems need to consider the intricacies of non-i.i.d. learning, and ultimately the importance of philosophically grounded and statistically rigorous fair-learning systems, operating on real-world data with all the messy dependence structures that may entail, just may be exactly what is needed, not only to bring new deep and interesting problems to the computer science community, but also to solve problems of algorithmic and natural injustice and unfairness in the world at large.

Part I

Concentration of Measure and Uniform Convergence in Machine Learning

The highly theoretical work of this part is key to the success in the data science and fair machine learning domains. Furthermore, this work in *concentration of measure* and *uniform convergence* is key to understanding and assessing the quality of our solutions in the pattern mining, data science, and fairness settings, as the basic statistical estimation themes echo throughout in various forms. Throughout the work we maintain a running metaphor of *sub-Gaussian* and *sub-gamma* bounds, and describe how, in most cases we can't expect to beat the sub-Gaussian bound, which are *asymptotically equivalent* to sub-gamma bounds.

Tail Bounds and Concentration of Measure For some function f, we are interested in the convergence of the *empirical mean* $\hat{\mathbb{E}}[f]$ on m samples to its expectation $\mathbb{E}[f]$. The Central Limit Theorem (CLT) tells us that $\hat{\mathbb{E}}[f]$ is asymptotically Gaussian, thus by the Chernoff bound for the Gaussian distribution, we have

$$\lim_{m\to\infty} \mathbb{P}\left(\left|\mathbb{E}[f] - \hat{\mathbb{E}}[f]\right| \geq \sqrt{\frac{2\,\mathbb{V}[f]\ln\frac{2}{\delta}}{m}}\right) \leq \delta \enspace .$$

[Devroye et al., 2016] show matching finite-sample lower-bounds for *any mean estimator*, thus the best bounds we can hope to achieve should be on par with the above. This class of bound is generally called *sub-Gaussian*, with $\mathbb{V}[f]$ replaced by various *variance proxies*. If we restrict our attention to *bounded* f, i.e., $f \in \mathcal{X} \to [a, b]$, where $r \doteq b - a$, Hoeffding's inequality [Hoeffding, 1963] states

$$\mathbb{P}\left(\left|\mathbb{E}[f] - \hat{\mathbb{E}}[f]\right| \geq \underbrace{\sqrt{\frac{r^2\ln\frac{2}{\delta}}{2m}}}_{\text{Variance Term}}\right) \leq \delta \enspace . \tag{1}$$

Here the *sub-Gaussian variance proxy* is $\frac{r^2}{4}$, which by Popoviciu's inequality, is the *worst-case* variance of any range r random variable. Bennett's inequality [Bennett, 1962] tells us that

$$\mathbb{P}\left(\left|\mathbb{E}[f] - \hat{\mathbb{E}}[f]\right| \geq \underbrace{\frac{r\ln\frac{2}{\delta}}{3m}}_{\text{Scale Term}} + \underbrace{\sqrt{\frac{2\,\mathbb{V}[f]\ln\frac{2}{\delta}}{m}}}_{\text{Variance Term}}\right) \leq \delta \enspace , \tag{2}$$

which yields the desired *variance dependence*, at the cost of a fast-decaying *scale term.* As such, the bound is *sub-gamma* [see Boucheron et al., 2013, chapter 2], rather than sub-Gaussian. We may thus view the *scale term* as a *finite-sample correction* to a sub-Gaussian bound.

Our uniform convergence bounds then apply simultaneously to *large sets of functions*, which translate to various guarantees on generalization error, estimation quality, and fairness. The natural comparison is to

Hoeffding's inequality with a union bound over $\mathcal{F}$, which tells us that

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} \left| \mathbb{E}[f] - \hat{\mathbb{E}}[f] \right| \geq \sqrt{\frac{r \ln \frac{2|\mathcal{F}|}{\delta}}{2m}} \right) \leq \delta \enspace .$$

Despite its weaknesses, this result powers many sample complexity guarantees in computer science. The principal aim of chapter 1 is to show that in many applications, we can improve over Hoeffding's inequality, and may nearly match variance-sensitive lower-bounds or Bennett's inequality.

Semisupervised Learning with Weak Labelers In (semi)supervised learning settings, the need for sharp tail bounds and generalization inequalities is often motivated by a dearth of *labeled training data*. Empirically centralized Rademacher averages are a mechanism to get more out of *limited labeled data*, but in chapter 2, I investigate a *weakly semisupervised setting*, where a small number of training labels are augmented with the output of *weak labelers*, which are assumed to be loosely correlated with the target task. I then describe a transductive learning algorithm for this setting, with strong data-dependent learning guarantees when the family of weak labelers is somewhat well-aligned with the target task.

In particular, the strategy is to estimate and bound appropriately-chosen statistics of the weak labelers using *labeled data* and applying *concentration inequalities*, e.g., Hoeffding, Bennett, or (empirically centralized) Rademacher bounds. The algorithm then considers a *feasible set* $\mathbb{Y}$ of *possible labelings* for the *unlabeled data* that respect these statistics. When sufficient labeled data are available to sharply-bound the chosen family of statistics, and the feasible label space is of *low diameter*, learning the *minimax-optimal* classifier over this feasible set then yields strong generalization guarantees. In particular, if the weak-labeler statistics give sufficient information about the target task, and we have sufficient labeled data to constrain their statistics, then all *feasible labelings* are reasonably accurate. Furthermore, if sufficiently *unlabeled data* are available, then we can learn a *complicated model* (i.e., a model of high Rademacher complexity); these conditions neatly factor the labeled and unlabeled sample complexities, and describe conditions under which weak-labelers are sufficient to supplement a small labeled training set.

In this setting, generalization error bounds may be composed into three major terms, with some variations in their exact form, depending on technical details. The first is the *minimax empirical risk* term, representing both the fundamental difficulty of the (possibly unrealizable) learning task with the given hypothesis class, which may be improved by learning a more complicated model. Second, we have *uncertainty over the feasible set* $\mathbb{Y}$, which may quantified, e.g., as the *diameter* of $\mathbb{Y}$, and may be improved by constraining additional statistics, adding weak labelers, or adding more labeled or unlabeled points. Then, thirdly we have an *overfitting term*, which describes overfitting of a model to the minimax guarantee, which may be controlled *solely by adding more unlabeled training data.* To make this concrete: the first two terms are *independent* of the complexity of the model we are learning, so if we wish to, e.g., fit a sophisticated deep neural network to a small amount of *labeled training data*, we may do so, offsetting the model complexity only by increasing the size of the *unlabeled training set.* Surprising as this may be, the reader should consider that this term addresses only *overfitting*, and good performance also requires that the *diameter* of the feasible set remain small, which requires subtle conditions on the weak-labeler statistics: essentially it must be possible to reconstruct the true labels of the target task with reasonable accuracy from them. Thus the method does not perform the impossible alchemy of synthesizing an accurate model with generalization guarantees from thin air, but rather *provides a mechanism* by which to *utilize and quantify* the information provided by weak-labelers.

Chapter 1

Empirically Centralized Rademacher Averages

I introduce the use of *empirical centralization* to derive novel practical, probabilistic, sample-dependent bounds to the Supremum Deviation (SD) of empirical means of functions in a family from their expectations. My bounds have optimal dependence on the maximum (i.e., wimpy) variance and the function ranges, and the same dependence on the number of samples as existing SD bounds. To compute the bounds in practice, I develop novel tightly-concentrated Monte-Carlo estimators of the empirical Rademacher average of the empirically-centralized family, and we show novel concentration results for the empirical wimpy variance. Experimental evaluation shows that our bounds greatly outperform non-centralized bounds and are extremely practical even at small sample sizes. This chapter is adapted from Cousins and Riondato [2020].

1.1 Introduction

The *supremum deviation* of the empirical means of functions in a family $\mathcal{F} \subseteq \mathcal{X} \rightarrow [a, b] \subset \mathbb{R}$ from their expectations is a key object in the study of empirical processes [Pollard, 1984]. Formally, let $\mathcal{D}$ be a distribution on the domain $\mathcal{X}$ and $\boldsymbol{x} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\}$ be a collection of m independent samples from $\mathcal{D}$. The *Supremum Deviation (SD) of $\mathcal{F}$ on $\boldsymbol{x}$* is the quantity

$$\mathsf{SD}(\mathcal{F}, \boldsymbol{x}) \doteq \sup_{f \in \mathcal{F}} \left| \hat{\mathbb{E}}_{\boldsymbol{x}}[f] - \mathbb{E}_{\mathcal{D}}[f] \right| , \text{ where } \hat{\mathbb{E}}_{\boldsymbol{x}}[f] \doteq \frac{1}{m} \sum_{i=1}^{m} f(\boldsymbol{x}_i) \ .$$

The sample-dependent *Empirical Rademacher Average (ERA)* $\hat{\mathfrak{R}}_m(\mathcal{F}, \boldsymbol{x})$ *of* $\mathcal{F}$ *on* $\boldsymbol{x}$ and its expectation, the *Rademacher Average (RA)* $\mathfrak{R}_m(\mathcal{F}, \mathcal{D})$ *of* $\mathcal{F}$ [Koltchinskii, 2001, Bartlett and Mendelson, 2002], imply *bidirectional bounds* on the SD (see (4)). Let $\boldsymbol{\sigma}$ be a collection of m independent Rademacher variables (i.e., uniform on $\{-1, 1\}$). These two quantities are defined as

$$\hat{\mathfrak{R}}_m(\mathcal{F}, \boldsymbol{x}) \doteq \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{m} \sum_{i=1}^{m} \boldsymbol{\sigma}_i f(\boldsymbol{x}_i) \right| \right], \text{ and } \mathfrak{R}_m(\mathcal{F}, \mathcal{D}) \doteq \mathop{\mathbb{E}}_{\boldsymbol{x}}\left[\hat{\mathfrak{R}}_m(\mathcal{F}, \boldsymbol{x})\right] \ . \tag{3}$$

The RA controls the finite-sample *expected* SD as [Van der Vaart and Wellner, 1996]

$$\tfrac{1}{2}\mathfrak{R}_m(\mathcal{F}, \mathcal{D}) - \tfrac{1}{\sqrt{m}} \sup_{f \in \mathcal{F}} \|f\|_\infty \leq \mathop{\mathbb{E}}_{\boldsymbol{x}}[\mathsf{SD}(\mathcal{F}, \boldsymbol{x})] \leq 2\mathfrak{R}_m(\mathcal{F}, \mathcal{D}) \ . \tag{4}$$

Probabilistic deviation bounds can be obtained by studying the convergence properties of the SD, and sample-dependent versions use the ERA and its deviation from the RA (see also theorem 1.3.2). The dependence on the *maximum* $q \doteq \sup_{f \in \mathcal{F}} \|f\|_\infty$ of $\mathcal{F}$ makes the lower bound unsatisfactory, as this quantity can be very large. This downside is particularly evident at relatively small sample sizes, which are actually the most interesting in practice. As uniform convergence bounds are now used not "just" for the theoretical analysis of the performance of learning, but also to develop randomized approximation algorithms for many tasks [Riondato and Upfal, 2015, 2018, Pellegrina et al., 2019, Areyan Viqueira et al., 2019, Pellegrina et al., 2020], we believe it is extremely important to derive *practical* bounds to the SD that are optimized not just in terms of the number of samples, but also of other important parameters, such as the maximum and the *wimpy variance* (see (10)). In this work, we use various forms of *centralization* to develop such practical bounds. Define the *distributional centralization* $\mathrm{C}_{\mathcal{D}}(\mathcal{F})$ *w.r.t.* $\mathcal{D}$ as the family

$$\mathrm{C}_{\mathcal{D}}(\mathcal{F}) \doteq \{x \mapsto f(x) - \mathop{\mathbb{E}}_{\mathcal{D}}[f], f \in \mathcal{F}\} \ . \tag{5}$$

$\mathrm{C}_{\mathcal{D}}(\mathcal{F})$ contains one function g for each $f \in \mathcal{F}$, such that g is f shifted by its expectation w.r.t. $\mathcal{D}$, thus, $\mathbb{E}_{\mathcal{D}}[g] = 0$ for each $g \in \mathrm{C}_{\mathcal{D}}(\mathcal{F})$. The Rademacher Average $\mathfrak{R}_m(\mathrm{C}_{\mathcal{D}}(\mathcal{F}), \mathcal{D})$ of $\mathrm{C}_{\mathcal{D}}(\mathcal{F})$ *sharply* controls the finite-sample expected SD as [Boucheron et al., 2013, Lemma 11.4]

$$\tfrac{1}{2}\mathfrak{R}_m(\mathrm{C}_{\mathcal{D}}(\mathcal{F}), \mathcal{D}) \leq \mathop{\mathbb{E}}_{\boldsymbol{x}}[\mathsf{SD}(\mathcal{F}, \boldsymbol{x})] \leq 2\mathfrak{R}_m(\mathrm{C}_{\mathcal{D}}(\mathcal{F}), \mathcal{D}) \ . \tag{6}$$

Comparing (4) and (6), it is evident that the RA could be an *arbitrary large* multiplicative factor away from the expected SD, especially at small-sample regimes or when the maximum q of $\mathcal{F}$ is large. The RA of the distributional centralization instead is *always* at most a multiplicative factor *two* away in both directions. Distributional centralization is therefore already known to be beneficial in the *expected* case, but can this gain be generalized to the probabilistic case, possibly using only sample-dependent quantities?

Contributions. In this work we introduce the use of *empirical centralization* to derive *practical*, *probabilistic* bounds to the SD. Our bounds exhibit a better or no worse dependence on important parameters such as the wimpy variance, the range (see (8)), and the sample size m (see theorem 1.3.3). We also show that the dependence on the wimpy variance that we obtain is optimal (lemma 1.3.4 and corollary 1.3.5). We introduce a novel empirical counterpart to the RA of the distributional centralization which uses *empirical centralization* to bound the SD. We analyze the bias of this quantity (lemma 1.2.1) and derive its concentration properties (theorem 1.2.2) using tail bounds for *self-bounding functions* [Boucheron et al., 2000, 2009]. In order to obtain fully-sample-dependent bounds, we introduce a Monte-Carlo estimator with sharp deviation bounds (theorem 1.3.7), and we also develop novel tight bounds for the empirical wimpy variance (theorem 1.3.1), which we believe to be of independent interest (e.g., in *matrix concentration inequalities* for estimating *eigenvalues* of *covariance matrices*). The results of our experimental evaluation show the advantages of centralization: the computed bounds to the SD are much smaller than those computed without centralization, even at small sample sizes. All proofs are relegated to appendix A.1.

1.2 Empirical centralization

We define the *empirical centralization* $\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F})$ *of* $\mathcal{F}$ *w.r.t. the sample* $\boldsymbol{x} \in \mathcal{X}^m$ as

$$\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}) \doteq \{x \mapsto f(x) - \hat{\mathbb{E}}_{\boldsymbol{x}}[f], f \in \mathcal{F}\} \enspace .$$

This quantity is an empirical counterpart to the distributional centralization $\mathrm{C}_{\mathcal{D}}(\mathcal{F})$ of $\mathcal{F}$ (see (5)). The key quantity that we use to derive the sample-dependent probabilistic bounds to the SD (section 1.3) is the ERA of the empirical centralization of $\mathcal{F}$, i.e., the quantity

$$\hat{\mathfrak{R}}_m(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}), \boldsymbol{x}) \enspace .$$

This quantity is completely dependent on the realized $\boldsymbol{x}$, even more, in some sense, than a “standard” ERA (see (3)), because the considered family $\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F})$ is also a function of $\boldsymbol{x}$, i.e., it is *sample-dependent.* We now derive its important properties: bias and concentration.

Bias The expectation w.r.t. $\boldsymbol{x}$ of $\hat{\mathfrak{R}}_m(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}), \boldsymbol{x})$ is *not* the RA of the distributional centralization of $\mathcal{F}$ (i.e., $\mathfrak{R}_m(\mathrm{C}_{\mathcal{D}}(\mathcal{F}), \mathcal{D})$), but we now show that the bias decreases rapidly in m, i.e., $\mathfrak{R}_m(\mathrm{C}_{\mathcal{D}}(\mathcal{F}), \mathcal{D}) \in \boldsymbol{\Theta}(\mathbb{E}_{\boldsymbol{x}}[\hat{\mathfrak{R}}_m(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}), \boldsymbol{x})])$. For ease of notation, let

$$\mathrm{b}(m) \doteq \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\left|\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{\sigma}_i\right|\right] \left(\text{which is } \boldsymbol{\Theta}\left(\frac{1}{\sqrt{m}}\right)\right) \ . \tag{7}$$

Lemma 1.2.1. Suppose $m \geq 4$. Then

$$\frac{\mathbb{E}_{\boldsymbol{x}}\left[\hat{\mathfrak{R}}_m(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}), \boldsymbol{x})\right]}{1+2\mathrm{b}(m)} \leq \mathfrak{R}_m(\mathrm{C}_{\mathcal{D}}(\mathcal{F}), \mathcal{D}) \leq \frac{\mathbb{E}_{\boldsymbol{x}}\left[\hat{\mathfrak{R}}_m(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}), \boldsymbol{x})\right]}{1-2\mathrm{b}(m)} \ .$$

Concentration We now show that $\hat{\mathfrak{R}}_m(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}), \boldsymbol{x})$ is tightly concentrated around its expectation because it is a *self-bounding function* [Boucheron et al., 2000, 2009] (see also definition A.1.1 in the supplementary material). We call the *widest range of* $\mathcal{F}$ the quantity

$$r \doteq \sup_{f\in\mathcal{F}}\left(\max_{x\in\mathcal{X}} f(x) - \min_{y\in\mathcal{X}} f(y)\right) \quad (\leq b-a) \ . \tag{8}$$

It is possible that $r \ll b-a$, for example, when $\mathcal{F}$ contains a function f and a function $g = f + c$ for some $c \in \mathbb{R}$. The widest range of the empirical and distributional centralizations of $\mathcal{F}$ is the same as the widest range of $\mathcal{F}$.

Theorem 1.2.2. Suppose $m \geq 1$, and let $\chi \doteq 1 + 2\mathrm{b}(m)$. For any $\delta \in (0,1)$, with probability at least $1-\delta$ over the choice of $\boldsymbol{x}$, it holds that

$$\mathop{\mathbb{E}}_{\boldsymbol{x}}[\hat{\mathfrak{R}}_m(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}), \boldsymbol{x})] \leq \hat{\mathfrak{R}}_m(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}), \boldsymbol{x}) + \frac{2r\chi\ln\frac{1}{\delta}}{3m} + \sqrt{\left(\frac{r\chi\ln\frac{1}{\delta}}{\sqrt{3}m}\right)^2 + \frac{2r\chi(\hat{\mathfrak{R}}_m(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}), \boldsymbol{x}) + r\mathrm{b}(m))\ln\frac{1}{\delta}}{m}} \ . \tag{9}$$

The ERA of $\mathcal{F}$ is a self-bounding function [Boucheron et al., 2003, Sect. 5.1], but proving this fact for the ERA of the empirical centralization $\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F})$ of $\mathcal{F}$ is more challenging (see proof in the supplementary material), because the empirical centralization $\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F})$ itself depends on the sample $\boldsymbol{x}$. This result, together with lemma 1.2.1, enables us to use the ERA of the empirical centralization, and Monte-Carlo estimations of it, to derive practical sharp upper-bounds to the SD.

1.3 Uniform convergence bounds

We now introduce novel bounds to the SD using the ERA of the empirical centralization. Before doing so, we must introduce an important technical concept.

Wimpy variance The *raw* (i.e., non-centralized) *wimpy variance* $\mathrm{W}^{\mathrm{r}}(\mathcal{F})$ *of* $\mathcal{F}$ and the (centralized) *wimpy variance* $\mathrm{W}(\mathcal{F})$ *of* $\mathcal{F}$ are key quantities in the study of probabilistic tail bounds to the SD [Boucheron et al., 2013, Ch. 11]. They are defined as

$$\mathrm{W}^{\mathrm{r}}(\mathcal{F}) \doteq \sup_{f\in\mathcal{F}} \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\big(f(x)\big)^2\right],\ \text{and}\ \mathrm{W}(\mathcal{F}) \doteq \sup_{f\in\mathcal{F}} \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\left(f(x)-\mathop{\mathbb{E}}_{\mathcal{D}}[f]\right)^2\right] \ . \tag{10}$$

Naturally, the raw wimpy variance is always greater or equal to its centralized counterpart, and potentially much larger. A key identity that we use throughout this work is

$$\mathrm{W}(\mathcal{F}) = \mathrm{W}^{\mathrm{r}}(\mathrm{C}_{\mathcal{D}}(\mathcal{F})) = \mathrm{W}(\mathrm{C}_{\mathcal{D}}(\mathcal{F})) \ .$$

Empirical estimators on $\boldsymbol{x}$ for the raw wimpy variance and for the wimpy variance are

$$\widehat{\mathrm{W}}^{\mathrm{r}}_{\boldsymbol{x}}(\mathcal{F}) \doteq \sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^{m}\big(f(x_i)\big)^2,\ \text{and}\ \widehat{\mathrm{W}}_{\boldsymbol{x}}(\mathcal{F}) \doteq \sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^{m}\left(f(x_i)-\hat{\mathop{\mathbb{E}}_{\boldsymbol{x}}}[f]\right)^2 \ .$$

To compute the *sample-dependent* bounds to the SD that we introduce later in this section, we develop novel tail bounds to these estimators, which we believe to be of independent interest. Most prior work assumed known *a priori* bounds to the wimpy variances, but we show that they can be replaced by *empirical* bounds. Maurer and Pontil [2009] prove that the sample variance (i.e., when $\mathcal{F}$ is a singleton) is a *weakly self-bounding function* [McDiarmid and Reed, 2006]. Our result holds for general $\mathcal{F}$, and is stronger, as we show that the wimpy variance is a *(strongly) self-bounding function* [Boucheron et al., 2000, 2009] (see also definition A.1.1 in the supplementary material).

Theorem 1.3.1. Suppose $m \geq 2$. Let $\delta \in (0,1)$. With probability $\geq 1-\delta$ over the choice of $\boldsymbol{x}$,

$$\mathrm{W}(\mathcal{F}) \leq \tfrac{m}{m-1}\widehat{\mathrm{W}}_{\boldsymbol{x}}(\mathcal{F}) + \frac{r^2 \ln\frac{1}{\delta}}{m-1} + \sqrt{\left(\frac{r^2\ln\frac{1}{\delta}}{m-1}\right)^2 + \frac{2r^2\,\tfrac{m}{m-1}\widehat{\mathrm{W}}_{\boldsymbol{x}}(\mathcal{F})\ln\frac{1}{\delta}}{m-1}} \ . \tag{11}$$

Bounds to the SD Bousquet [2002, Thm. 2.3 (presented here for clarity in a slightly weaker form)] uses the wimpy variance to derive concentration bounds for the SD.

Theorem 1.3.2 (Bousquet, 2002, Thm. 2.3)**.** Let $\delta \in (0,1)$. With probability $\geq 1-\delta$ over the choice of $\boldsymbol{x}$,

$$\mathsf{SD}(\mathcal{F}, \boldsymbol{x}) \leq \mathop{\mathbb{E}}_{\boldsymbol{x}}\left[\mathsf{SD}(\mathcal{F}, \boldsymbol{x})\right] + \frac{2r \ln \frac{1}{\delta}}{3m} + \sqrt{\frac{2\left(\mathrm{W}(\mathcal{F}) + 4r\ \mathbb{E}_{\boldsymbol{x}}\left[\mathsf{SD}(\mathcal{F}, \boldsymbol{x})\right]\right) \ln \frac{1}{\delta}}{m}}\ . \tag{12}$$

By plugging the r.h.s. of the symmetrization inequalities (4) and (6) in the r.h.s. of (12), one can obtain bounds that depend on the RA of $\mathcal{F}$ or on the RA of the *distributional* centralization $\mathrm{C}_{\mathcal{D}}(\mathcal{F})$. Neither of these bounds are sample-dependent. Such a bound can be obtained, for example, by using the ERA of $\mathcal{F}$ and a tail bound (e.g., McDiarmid [1989]'s inequality or a tail bound for self-bounding functions [Boucheron et al., 2009]) on the deviation of the ERA from the RA. The following result states our sample-dependent bound to the SD using the *empirical centralization* $\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F})$ and tail bounds to the wimpy variance, obtained by combining lemma 1.2.1 and theorems 1.3.1 to 1.3.2.

Theorem 1.3.3. Assume $m \geq 4$, and let $\eta \in (0,1)$. Take ν to be the r.h.s. of (11) computed with $\delta = \eta/3$, so $\mathbb{P}(\mathrm{W}(\mathcal{F}) > \nu) \leq \eta/3$, and take λ to be the r.h.s. of (9) computed with $\delta = \eta/3$, so $\mathbb{P}(\mathbb{E}_{\boldsymbol{x}}[\hat{\mathfrak{R}}_m(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}), \boldsymbol{x})] > \lambda) \leq \eta/3$. With probability $\geq 1-\eta$ over the choice of $\boldsymbol{x}$, it holds that

$$\mathsf{SD}(\mathcal{F}, \boldsymbol{x}) \leq \frac{2\lambda}{1-2\mathrm{b}(m)} + \frac{2r \ln \frac{3}{\eta}}{3m} + \sqrt{\frac{2(\nu + 8r\lambda/(1-2\mathrm{b}(m))) \ln \frac{3}{\eta}}{m}}\ .$$

The r.h.s. is

$$\frac{2}{1-2\mathrm{b}(m)} \hat{\mathfrak{R}}_m\left(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}), \boldsymbol{x}\right) + \mathbf{O}\left(\frac{r \ln \frac{1}{\eta}}{m} + \sqrt{\frac{(\mathrm{W}(\mathcal{F}) + r\mathfrak{R}_m(\mathrm{C}_{\mathcal{D}}(\mathcal{F}), \mathcal{D}) + r^2/\sqrt{m}) \ln \frac{1}{\eta}}{m}}\right)\ .$$

Is there any advantage in using this bound, i.e., in using empirical centralization, rather than using a bound involving the ERA of $\mathcal{F}$? I.e., how does it compare to the standard bound

$$\mathsf{SD}(\mathcal{F}, \boldsymbol{x}) \leq 2\hat{\mathfrak{R}}_m\left(\mathcal{F}, \boldsymbol{x}\right) + \mathbf{O}\left(\frac{q \ln \frac{1}{\eta}}{m} + \sqrt{\frac{(\mathrm{W}(\mathcal{F}) + q\mathfrak{R}_m(\mathcal{F}, \mathcal{D}) + q\sqrt{\mathrm{W}^{\mathrm{r}}(\mathcal{F})}/\sqrt{m}) \ln \frac{1}{\eta}}{m}}\right)\ ?$$

We shall see that the $r^2/\sqrt{m}$ and $q\sqrt{\mathrm{W}^{\mathrm{r}}(\mathcal{F})}/\sqrt{m}$ terms are incomparable, though both appear only in transient $\mathbf{O}(m^{-3/4})$ terms, and the remaining differences all favor centralization. Most previous studies focused on the behavior of SD bounds as functions of the sample size m, but we believe that efficient SD bounds for practical applications (e.g., [Riondato and Upfal, 2018, 2015, Pellegrina et al., 2019, Areyan Viqueira et al., 2019, Pellegrina et al., 2020]), must improve the dependence also on the other parameters, the *wimpy variance* being the most important. Indeed, developing such bounds is the goal of this work.

First of all, we remark that the dependence on the wimpy variance shown in (12) cannot be improved: any bound to the SD of $\mathcal{F}$ must be $\mathbf{\Omega}\sqrt{\mathrm{W}(\mathcal{F})\ln\frac{1}{\delta}/m}$, as can be shown using minimax lower bounds and median-of-means bounds [Devroye et al., 2016, Lugosi and Mendelson, 2019]. The question is thus whether the *complexity terms*, i.e., $\hat{\mathfrak{R}}_m(\mathcal{F},\boldsymbol{x})$ and $\hat{\mathfrak{R}}_m(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}),\boldsymbol{x})$, can match this lower bound. Lemma 1.3.4 answers this question in the *negative* for $\hat{\mathfrak{R}}_m(\mathcal{F},\boldsymbol{x})$, and in the *positive* for $\hat{\mathfrak{R}}_m(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}),\boldsymbol{x})$: the ERA of $\mathcal{F}$ is controlled (in part) by the empirical *raw* wimpy variance, whereas the ERA of $\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F})$ has corresponding depence on the empirical (centralized) wimpy variance. As with ordinary function variances, the raw wimpy variance can be unboundedly larger than the (centralized) wimpy variance, e.g., in the *constant function family* $\mathcal{F} \doteq \{x \mapsto c\}$.

Lemma 1.3.4. For any $\boldsymbol{x} \in \mathcal{X}^m$, it holds

$$\hat{\mathfrak{R}}_m(\mathcal{F},\boldsymbol{x}) \geq \sqrt{\frac{\widehat{\mathrm{W}^{\mathrm{r}}_{\boldsymbol{x}}}(\mathcal{F})}{2m}} \text{ and } \hat{\mathfrak{R}}_m(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}),\boldsymbol{x}) \geq \sqrt{\frac{\widehat{\mathrm{W}_{\boldsymbol{x}}}(\mathcal{F})}{2m}} \enspace .$$

Furthermore, it holds

$$\lim_{m\to\infty}\sqrt{m}\mathfrak{R}_m(\mathcal{F},\mathcal{D}) \geq \sqrt{\tfrac{2}{\pi}\mathrm{W}^{\mathrm{r}}(\mathcal{F})} \text{ and } \lim_{m\to\infty}\sqrt{m}\mathfrak{R}_m(\mathrm{C}_{\mathcal{D}}(\mathcal{F}),\mathcal{D}) \geq \sqrt{\tfrac{2}{\pi}\mathrm{W}(\mathcal{F})} \enspace .$$

To make the result concrete, consider that as soon as $\mathcal{F}$ contains a function f and a "c-shifted" version of it $f + c$, for some $c \in \mathbb{R}^+$, then $\sup_{g\in\mathcal{F}}|\hat{\mathbb{E}}_{\boldsymbol{x}}[g]| \geq c/2$, thus $\widehat{\mathrm{W}^{\mathrm{r}}_{\boldsymbol{x}}}(\mathcal{F}) \geq c^2/4$, and from the above lemma, $\hat{\mathfrak{R}}_m(\mathcal{F},\boldsymbol{x}) \geq c/\sqrt{8m}$, but $\hat{\mathfrak{R}}_m(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}),\boldsymbol{x})$ does not suffer from this issue.

The significance of lemma 1.3.4 is that a dependence on the (*centralized*) wimpy variance *cannot* be obtained *without* empirical centralization. One must settle for dependence on the *raw* wimpy variance, which can be unboundendly larger than its centralized counterpart. The result also tells us that a dependence on the (centralized) wimpy variance *may* be attained with empirical centralization. We show next that such is indeed the case.

Optimal dependence on wimpy variance The quantity $\hat{\mathfrak{R}}_m(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}),\boldsymbol{x})$ is an ERA, thus it can be upper-bounded using Massart's finite-class lemma [Massart, 2000, lemma 5.2]. We now apply this celebrated result to bound the ERA under *empirical centralization* while including the *absolute value* (absent from some presentations) inside the supremum of the ERA.

Corollary 1.3.5. Assume that $\mathcal{F}$ is finite. Let $\mathcal{F}_\pm \doteq \mathcal{F} \cup \{-f, f \in \mathcal{F}\}$. It holds

$$\hat{\mathfrak{R}}_m(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}), \boldsymbol{x}) \le \sqrt{\frac{2\widehat{\mathrm{W}}_{\boldsymbol{x}}(\mathcal{F}) \ln|\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}_\pm)|}{m}} \ . \tag{13}$$

The use of $\mathcal{F}_\pm$ is needed[1] to handle the absolute value in our definition of the ERA (see (3)). Without empirical centralization, the dependence would be on the raw wimpy variance, which equals the squared ℓ_2 norm in "classic" presentations of Massart's lemma. Corollary 1.3.5 shows that empirical centralization enables *optimal* dependence on the *centralized* wimpy variance, which *cannot be obtained without empirical centralization*, as shown in lemma 1.3.4.

Monte-Carlo estimation The quantity $\hat{\mathfrak{R}}_m(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}), \boldsymbol{x})$ is an ERA, so it "suffers" from the usual issue of how to actually compute or bound it in order to bound the SD via theorem 1.3.3. While analytical methods (e.g., Massart's lemma) yield (generally loose) bounds, Monte-Carlo estimation with proper tail bounds gives better results in practice, and it was proposed almost concurrently with the introduction of the ERA [Bartlett and Mendelson, 2002].

Definition 1.3.6. Let $\boldsymbol{\sigma} \in (\pm 1)^{n \times m}$ be a matrix of i.i.d. Rademacher r.v.'s. The *Monte-Carlo ERA* $\hat{\mathfrak{R}}^n_m(\mathcal{F}, \boldsymbol{x}, \boldsymbol{\sigma})$ *of* $\mathcal{F}$ *on* $\boldsymbol{x}$ *w.r.t.* $\boldsymbol{\sigma}$ is the quantity

$$\hat{\mathfrak{R}}^n_m(\mathcal{F}, \boldsymbol{x}, \boldsymbol{\sigma}) \doteq \frac{1}{n} \sum_{i=1}^{n} \sup_{f \in \mathcal{F}} \left| \frac{1}{m} \sum_{j=1}^{m} \boldsymbol{\sigma}_{i,j} f(\boldsymbol{x}_i) \right| \ .$$

It clearly holds $\mathbb{E}_{\boldsymbol{\sigma}}[\hat{\mathfrak{R}}^n_m(\mathcal{F}, \boldsymbol{x}, \boldsymbol{\sigma})] = \hat{\mathfrak{R}}_m(\mathcal{F}, \boldsymbol{x})$. Bartlett and Mendelson [2002, Thm. 11] show that the MC-ERA with $n = 1$ is concentrated about the ERA as

$$\mathbb{P}_{\boldsymbol{\sigma}}\left(\left| \hat{\mathfrak{R}}_m(\mathcal{F}, \boldsymbol{x}) - \hat{\mathfrak{R}}^1_m(\mathcal{F}, \boldsymbol{x}, \boldsymbol{\sigma}) \right| \ge \varepsilon \right) \le 2 \exp\left(\frac{-2m\varepsilon^2}{q^2} \right) \ .$$

The r.h.s. can be used in theorem 1.3.3 inside the definition of λ (with the needed adjustment of the confidence parameter δ using a union bound), thus obtaining an upper bound to the SD using the MC-ERA. The leitmotif of this work is to obtain strong, practical, sample-dependent bounds to the SD, so we derive a novel tail bound to the MC-ERA (theorem 1.3.7) *for general* n, where the strong dependence on q^2 of the above bound is replaced by a much weaker dependence, primarily on $\mathrm{W}(\mathcal{F})$. This change is similar to how theorem 1.3.2 improves over textbook bounds to the SD that use McDiarmid's bounded difference inequality.

[1] Since $|\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}_\pm)| \le 2|\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F})|$, the bound can be reformulated as function of $|\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F})|$ only.

Our improved variance-sensitive bound uses a transportation-method inequality due to Samson [2007] to *upper bound* the *expectation* of suprema of empirical processes. This result is, to our knowledge, novel, and is worst-case asymptotically equivalent to the McDiarmid bounds, and improves over it when the wimpy variance is small. The bound uses the *empirical maximum* $\hat{q}_{\mathcal{F}}(\boldsymbol{x})$ *of* $\mathcal{F}$ *on* $\boldsymbol{x}$, defined as

$$\hat{q}_{\mathcal{F}}(\boldsymbol{x}) \doteq \sup_{f \in \mathcal{F}, x \in \boldsymbol{x}} |f(x)| \qquad (\leq q) \enspace .$$

Theorem 1.3.7. Let $\boldsymbol{\sigma} \in (\pm 1)^{n \times m}$ be a matrix of i.i.d. Rademacher r.v.'s. Let $\delta \in (0,1)$. With probability at least $1-\delta$ over the choice of $\boldsymbol{\sigma}$, it holds

$$\hat{\mathfrak{R}}_m(\mathcal{F}, \boldsymbol{x}) \leq \hat{\mathfrak{R}}^n_m(\mathcal{F}, \boldsymbol{x}, \boldsymbol{\sigma}) + \frac{2\hat{q}_{\mathcal{F}}(\boldsymbol{x}) \ln \frac{1}{\delta}}{3nm} + \sqrt{\frac{4\widehat{\mathrm{W}}^{\mathrm{r}}_{\boldsymbol{x}}(\mathcal{F}) \ln \frac{1}{\delta}}{nm}} \enspace . \tag{14}$$

Empirical centralization obtains a dependence on the empirical wimpy variance of $\mathcal{F}$, rather than on the raw (i.e., non-centralized) wimpy variance. This advantage propagates when using the MC-ERA of the empirical centralization to bound the SD of $\mathcal{F}$. The dependence on the empirical maximum changes from $\hat{q}_{\mathcal{F}}(\boldsymbol{x})$ to $\hat{q}_{\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F})}(\boldsymbol{x})$, which can be a large improvement (and $\hat{q}_{\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F})}(\boldsymbol{x}) < 2\hat{q}_{\mathcal{F}}(\boldsymbol{x})$ at most).

Corollary 1.3.8. Let $\boldsymbol{\sigma} \in (\pm 1)^{n \times m}$ be a matrix of i.i.d. Rademacher r.v.'s. Let $\delta \in (0,1)$. With probability at least $1-\delta$ over the choice of $\boldsymbol{\sigma}$, it holds

$$\hat{\mathfrak{R}}_m(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}), \boldsymbol{x}) \leq \hat{\mathfrak{R}}^n_m(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}), \boldsymbol{x}, \boldsymbol{\sigma}) + \frac{2\hat{q}_{\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F})}(\boldsymbol{x}) \ln \frac{1}{\delta}}{3nm} + \sqrt{\frac{4\widehat{\mathrm{W}}_{\boldsymbol{x}}(\mathcal{F}) \ln \frac{1}{\delta}}{nm}} \enspace .$$

Although $n = 1$ Monte-Carlo trials are sufficient to match the convergence rate of theorem 1.3.2, the Monte-Carlo estimation error term can still be a significant portion of the total SD bound. For practical usage, particularly with small sample sizes, or when extremely tight bounds are needed, more Monte-Carlo trials (i.e., larger n) rapidly reduce the Monte-Carlo estimation error, and this error is soon dominated by the tail bound terms of theorem 1.3.2.

Example: Batch panel of experts Consider now the *batch panel of experts* problem, where $\mathcal{F}$ is a *finite family* of *experts*, and the task is to select the (approximately) *most accurate* among them, given a sample of *labeled instances*. With the Monte-Carlo method, we may sharply bound the SD whenever evaluating the requisite suprema is computationally feasible, e.g., via *enumeration* of $\mathcal{F}$. Furthermore, we automatically benefit from *data-dependent* and *distribution-dependent* structure, e.g., highly correlated or anticorrelated experts, and *low wimpy variance* over uniformly accurate $\mathcal{F}$. This example immediately extends to *model selection* via *structural risk minimization* if, e.g., the experts are organized into *concentric groups* by some *a priori confidence* or *quality* estimate.

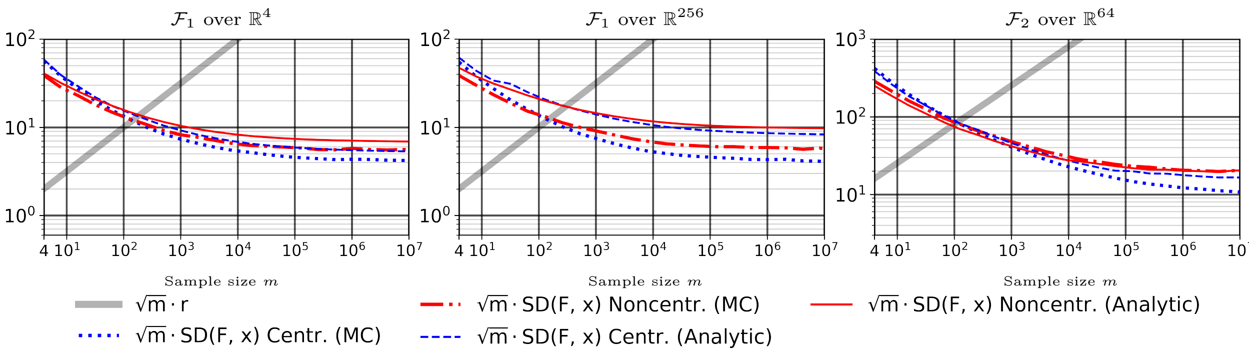


Figure 1.1: Comparison of SD bounds as functions of the sample size m. See the main text for an explanation of the results.

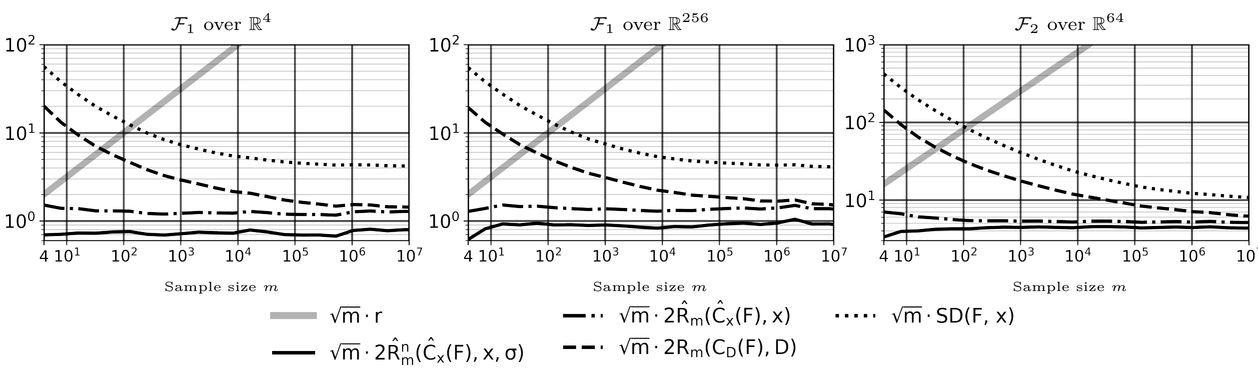


Figure 1.2: Upper bounds to complexity measures and SD as functions of the sample size m. See the main text for details.

1.4 Experimental evaluation

We performed experiments to evaluate the various bounds presented in the previous sections and compare the bounds to the SD using empirical centralization to those without centralization. The code is included in the supplementary material.

Function families We consider the function families $\mathcal{F}_p$, for any $p \geq 1$, containing all unit ℓ_p-norm-constrained linear functions in $\mathbb{R}^d$, i.e.,

$$\mathcal{F}_p \doteq \{\boldsymbol{x} \mapsto w \cdot \boldsymbol{x}, w \in \mathbb{R}^d \text{ s.t. } \|w\|_p \leq 1\} \ .$$

These families are of immediate interest in many machine learning settings, such as the analysis of support vector machines and neural networks, as both consist of Lipschitz loss and/or activation functions applied to one or more linear functions (see, e.g., [Bartlett and Mendelson, 2002] for analysis). Additionally, a bound on the SD of $\mathcal{F}_p$ over distribution $\mathcal{D}$ over $\mathbb{R}^d$ corresponds to the radius of the $\ell_{p/p-1}$ (Hölder dual norm) ball about $\hat{\mathbb{E}}[\boldsymbol{x}]$ in which $\mathbb{E}_{\mathcal{D}}[\boldsymbol{x}] \in \mathbb{R}^d$ falls. Such balls can be used to estimate *covariance matrices*, high-dimensional *sufficient statistics* in graphical models [Bradley and Guestrin, 2012], and to learn *equilibria* in simulation-based games [Areyan Viqueira et al., 2019, 2020].

Analytical bounds to the ERA of $\mathcal{F}_p$ on $\boldsymbol{x}$ and Monte-Carlo estimates of it (see definition 1.3.6) are relatively straightforward [Shalev-Shwartz and Ben-David, 2014, Lemmas 26.10, 26.11] (see lemma A.2.1 in the supplementary material). The following lemma extends these results to the empirical centralization.

Lemma 1.4.1. Let $\bar{\boldsymbol{x}} \doteq \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{x}_i \in \mathbb{R}^d$. For the ℓ_1 norm, it holds

$$\hat{\mathfrak{R}}_m(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}_1), \boldsymbol{x}) = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\left\|\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{\sigma}_i(\boldsymbol{x}_i - \bar{\boldsymbol{x}})\right\|_\infty\right] \leq \max_i \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}\|_\infty \sqrt{\frac{2\ln(2d)}{m}},$$

while for the ℓ_2 norm, it holds

$$\hat{\mathfrak{R}}_m(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}_2), \boldsymbol{x}) = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\left\|\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{\sigma}_i(\boldsymbol{x}_i - \bar{\boldsymbol{x}})\right\|_2\right] \leq \max_i \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}\|_2 \frac{1}{\sqrt{m}} \quad .$$

Similar bounds are possible for other values of p; e.g., by linearity, the case of $p = \infty$ is trivial. Note that in addition to computing MC-ERAs from $\ell_{p/p-1}$ dual norms, we may also compute (raw) empirical wimpy variances from *operator norms* of (raw) *covariance matrices* of $\boldsymbol{x}$. In particular, for $\mathcal{F}_1$, it is easy to show that the wimpy variance is simply the *largest variance* along any *standard basis vector*. Similarly, for $\mathcal{F}_2$, the wimpy variance is simply the *maximum variance* along any *unit vector*, i.e., the *spectral norm* of the covariance matrix.

Data generation and parameter values We generated the samples $\boldsymbol{x}$ for our experiments from random distributions over $\mathbb{R}^d$. The ERA of the family $\mathcal{F}_1$ is susceptible to the value of d (see lemma 1.4.1 and lemma A.2.1 in the supplementary material), so we use $d = 4$ and $d = 256$, while in the case of $\mathcal{F}_2$ the ERA is independent of d, so we use $d = 64$. Details of the distributions are in the supplementary material. Range-like quantities (i.e., q, $\hat{q}$, or r) can be computed from the data and/or known *a priori* bounds: $r = 1$ for our $\mathcal{F}_1$ experiments and $r = 8$ for the $\mathcal{F}_2$ case. (Raw) wimpy variances correspond to norms of the (raw) covariance matrices used for data generation (see the supplementary material for details). In all experiments, we used $\delta = 0.01$ and $n = 32$ (we comment on this choice below). The sample size m varied from 4 (the minimum possible, due to lemma 1.2.1) to 10^7.

A note on results visualization We present all of our results in plots with *log-log axes*, so that convergence rates are clearly visible as slopes, and constant factors as vertical offsets. The x-axis is the sample size m. Since we expect asymptotic convergence rates $\propto {}^{C}\!/\!\sqrt{m}$, where C depends on the (possibly raw) wimpy variance of $\mathcal{F}$, r, and δ, we plot all quantities *multiplied by* $\sqrt{m}$. This transformation allows to clearly visualize $\boldsymbol{\Theta}({}^{C}\!/\!\sqrt{m})$ behaviors as *straight horizontal lines*, and $\mathbf{o}({}^{C}\!/\!\sqrt{m})$ behaviors as (transient) downward slopes. For completeness, we show plots without the scaling by $\sqrt{m}$ in the supplementary material.

Results Figure 1.1 compares four bounds to the SD: using the Monte-Carlo estimate $\hat{\mathfrak{R}}^{n}_{m}(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}_p), \boldsymbol{x}, \boldsymbol{\sigma})$ for the ERA $\hat{\mathfrak{R}}_{m}(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}_p), \boldsymbol{x})$ of the empirical centralization of $\mathcal{F}_p$ on $\boldsymbol{x}$, using the Monte-Carlo estimate $\hat{\mathfrak{R}}^{n}_{m}(\mathcal{F}_p, \boldsymbol{x}, \boldsymbol{\sigma})$ for the ERA $\hat{\mathfrak{R}}_{m}(\mathcal{F}_p, \boldsymbol{x})$ of the non-centralized $\mathcal{F}_p$, using analytical bounds to $\hat{\mathfrak{R}}_{m}(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}_p), \boldsymbol{x})$ from lemma 1.4.1, and using analytical bounds to $\hat{\mathfrak{R}}_{m}(\mathcal{F}_p, \boldsymbol{x})$ from lemma A.2.1 (in the supplementary material). The thicker grey line is the quantity $\sqrt{m}r$; bounds above this line are vacuous.

At very small sample sizes (when all bounds are *vacuous*), the bounds obtained without centralization are sharper than the bounds with empirical centralization, due to the bias-correction of lemma 1.2.1 (see ξ in theorem 1.3.3) and the (fast-decaying) $\boldsymbol{\Theta}({}^{r}\!/\!{}_{m^{3/4}})$ term of theorem 1.2.2. Before $m \approx 200$, when bounds become non-vacuous, the advantages of empirical centralization become clear, and increase with the sample size. Recall that each bound is scaled by $\sqrt{m}$, thus all are *asymptotically horizontal*, as $\boldsymbol{\Theta}({}^{C}\!/\!\sqrt{m})$ terms eventually dominate the bound to the SD, where C varies greatly between bounds and methods. Thus without empirical centralization, obtaining the same bound to the SD would require a larger sample size m than with empirical centralization (this effect can be better observed in the non-$\sqrt{m}$-scaled plots in figure A.1 in the supplementary material.) The Monte-Carlo estimate, despite using only $n = 32$ Monte-Carlo trials, gives better bounds to the SD than an analytical approach.

In figure 1.2, we drill down on the SD bounds using the Monte-Carlo estimate $\hat{\mathfrak{R}}^{n}_{m}(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}_p), \boldsymbol{x}, \boldsymbol{\sigma})$ for the ERA $\hat{\mathfrak{R}}_{m}(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}_p), \boldsymbol{x})$ of the empirical centralization of $\mathcal{F}_p$ on $\boldsymbol{x}$, showing this quantity, together with the *upper bounds* to other intermediate quantities, that eventually lead to the SD bound: the ERA $\hat{\mathfrak{R}}_{m}(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}_p), \boldsymbol{x})$ (obtained by applying theorem 1.3.7 to the MC-ERA), the RA $\mathfrak{R}_{m}(\mathcal{F}_p, \mathcal{D})$ (obtained by applying theorem 1.2.2 and lemma 1.2.1 to the bound on the ERA), and SD (obtained by applying the r.h.s. of (6) and theorem 1.3.2 to the bound on the RA).

At small sample sizes, the fast-decaying terms dominate the bounds to the RA and SD, but, true to their nature, quickly become negligible: all bounds are asymptotically $\mathbf{\Theta}(C/\sqrt{m})$, where C, which in the plots in figure 1.2 appear as the vertical offset of each curve at high sample sizes, depends mostly on the wimpy variance of $\mathcal{F}$ and the range r. The bounds that decay as $\mathbf{\Theta}\sqrt{\mathrm{W}(\mathcal{F})/m}$ (i.e., the MC-ERA $\rightarrow$ ERA and RA $\rightarrow$ SD bounds) introduce constant factor terms, manifest as asymptotic vertical gaps, whereas the remaining bounds entirely vanish asymptotically. The gap from the MC-ERA to the ERA would disappear as the number n of Monte-Carlo trials (which we fixed at $n = 32$) increases.

The range and wimpy variances are approximately the same in both $\mathcal{F}_1$ experiments but the MC-ERA are much larger when $d = 256$ because here the RA is essentially the expected largest distance traveled over d random walks, which increases with d (see also lemma 1.4.1).

In conclusion, the results confirm the advantages of empirical centralization to obtain tighter bounds to the SD with optimal dependence on the wimpy variance, while still maintaining the same behavior in terms of the number of samples as bounds not using centralization.

1.5 Conclusions

We develop practical, sharp, sample-dependent probabilistic bounds to the SD through *empirical centralization*, together with novel results on the concentration of the wimpy variance and of Monte-Carlo estimates of the ERA. Our bounds exhibit optimal dependence on the wimpy variance and the same dependence on the number of samples as bounds not using centralization. The results of our experimental evaluation show that the advantage is significant even at small sample sizes, and remains so as the sample size grows. In future work, we will explore the important relationship between centralization and localization [Koltchinskii, 2006, Giné and Koltchinskii, 2006].

Chapter 2

Adversarial Learning with Weak Supervision

In this work, I develop a rigorous approach for using a set of possibly correlated weak supervision sources in order to solve a multi class classification task, when only a very small set of labeled data is available. The algorithm optimally learns a model that has minimum expected risk among all the feasible solutions for a set of unlabeled data, where the feasibility of a labeling is computed through constraints defined by estimated statistics of the weak supervision sources. I provide theoretical guarantees for this approach that depend on the information provided by the weak supervision sources. Notably, this method does not require the weak supervision sources to have the same labeling space as the multi-class task. Finally, I demonstrate the effectiveness of the approach with experiments on various image classification tasks. This chapter is based on joint work with Alessio Mazzetto, Dylan Sam, Stephen Bach, and Eli Upfal.

2.1 Introduction

In the last decade, deep neural networks have been applied to accurately solve a wide range of classification tasks in different domains, but the supervised learning of these models requires a considerable amount of labeled data. An alternative strategy is to learn from *weak supervision*, i.e., sources of labels that are noisy or heuristic. Examples include hand-written rules [Ratner et al., 2017, Wu et al., 2018, Safranchik et al., 2020] and classifiers trained for related tasks [Varma et al., 2017, Bach et al., 2019, Chen et al., 2019]. Even if these sources of information are noisy, results show that they can lead to high-quality models, particularly when the outputs from many sources are combined.

A key technical challenge in such work is how to combine multiple sources of weak supervision, since they might conflict with each other. We assume access to little to no ground truth data. Instead, much prior work on aggregating noisy labels [Dawid and Skene, 1979, Zhang et al., 2016, Gao and Zhou, 2013, Karger et al., 2014, Ghosh et al., 2011, Dalvi et al., 2013, Ratner et al., 2016, 2019] assumes that the sources make independent errors, which is a very strong assumption. Some recent works [Bach et al., 2017, Varma et al., 2019] attempt to learn more sophisticated distributions but still rely on parametric assumptions that make conditional independence assumptions. Independence assumptions in models of weak supervision sources are hard to verify and limiting in practice. Many useful weak supervision sources, particularly ones learned from related datasets, can be arbitrarily correlated, as there are systematic differences between the target classification task and the mildly related tasks used to learn them. For example, if all the labelers are fine-tuned from the same pretrained model, they are likely to inherit some of the same biases.

Recently, two pieces of work have addressed the problem of combining weak labelers without distribution assumptions by taking an adversarial approach. Adversarial label learning (ALL) [Arachie and Huang, 2019] first formulated the problem as a minimax optimization, learning a model that minimizes risk using the worst-case assignment to the unknown ground truth labels. ALL is restricted to binary classification, it does not solve the minimax optimization problem optimally, and provides no theoretical guarantees for the models it learns. A later work, performance guaranteed majority vote (PGMV) [Mazzetto et al., 2021], takes an alternative approach for the binary classification setting. It uses a small amount of labeled data and a large amount of unlabeled data to empirically estimate properties of the labelers and define a constrained optimization over the assignments to unknown labels with a worst-case error bound. However, this approach is also limited to binary classification because it exploits the fact that when two labelers disagree, one must be correct. Further, it requires performing the worst case optimization over each subset of labelers considered. The resulting combinatorial blowup limits the number of labelers that can be considered simultaneously.

In this work, we address the limitations of ALL and PGMV by providing a computationally efficient framework for multi class weak supervision with performance guarantees. Similar to ALL, we formulate the search for ground truth as a search over a set of ground truth labels that satisfy statistical constraints on the weak supervision sources. Unlike ALL, we optimally solve the optimization problem, and we show a novel analysis that factors in the information provided by the weak supervision sources with respect to the target classification, geometrically represented as the ℓ_1 diameter of the region of feasible labelings, to evaluate the quality of the learnt model.

Contributions. We introduce a novel method to use the information provided by a set of arbitrarily correlated weak supervision sources in order to learn a classifier for a target classification task of interest. Inspired by previous work, we use a little amount of labeled data to compute statistics on the weak supervision sources, and we formulate an optimization problem to find the prediction model that achieves the lowest risk among all the labelings of an unlabeled dataset that agree with those statistics.

Our main contributions are the following:

1. We develop the first rigorous method for learning multi class classification from weak supervision sources (Section 4).
2. We provide theoretical analysis of our method, proving approximation guarantees on the quality of our solution, and complexity bounds for the algorithm's run-time (Section 4).
3. We evaluate the quality of the solution provided by our method using a geometrical quantity that represents the aggregate information provided by the weak supervision sources with respect to the target classification task (Section 4.2).
4. While the presentation of our method is general, we demonstrate the applicability of our approach through two practical instances of prediction model and loss function: convex combination of the weak supervision sources and multinomial logistic regression (Section 4.1)
5. We show how to extend our method for the case when the weak supervision sources do not have the same labeling space as the multi-classification task. This is useful, for example, when learning with attributes. In many weak supervision tasks we have information that partially constrains the correct label, such as knowing that if a labeler detects stripes on an animal, it narrows down the possible animal types (Section 4.3).
6. We conduct extensive experiments demonstrating the effectiveness of our novel approach for multi class classification tasks. We also show that the performance of our approach compares favorably with the best published adversarial weak learning algorithms for binary classification, ALL [Arachie and Huang, 2019] and PGMV [Mazzetto et al., 2021].

2.2 Related Work

The problem of learning from multiple, possibly conflicting weak labelers with little to no ground truth data has received considerable attention recently [Ratner et al., 2016, Bach et al., 2017, Ratner et al., 2017, Varma et al., 2019, Arachie and Huang, 2019, Mazzetto et al., 2021]. This setting is distinct from much work on ensemble learning [Zhang and Ma, 2012], such as boosting [Schapire, 1990, Freund, 1995], where abundant labeled examples are used to learn to combine ensemble members. Other ensemble methods, such as bagging [Breiman, 1996] take an unweighted vote of ensemble members but rely on the assumption that each member is trained on labeled data sampled from the target distribution. Unlike these methods, in weak supervision, the goal is to use other statistical properties of the labelers, such as their agreements and disagreements, to learn to combine them. In this way, the labelers can potentially be improved without increasing the need for labeled training data.

This work has its roots in crowdsourcing, where the "labelers" are people with different, unknown levels of reliability. Dawid and Skene's seminal work [Dawid and Skene, 1979] showed how the accuracy of each labeler could be estimated with expectation maximization by assuming a naive Bayes distribution over the labelers' votes and the latent ground truth. Since then, much work has provided theoretically guaranteed algorithms for learning under these assumptions [Zhang et al., 2016, Gao and Zhou, 2013, Karger et al., 2014, Ghosh et al., 2011, Dalvi et al., 2013]. When the labelers are humans working without coordination, the independence assumption is a reasonable one.

Recently, frameworks for weakly supervised machine learning like Snorkel [Ratner et al., 2016, Bach et al., 2017, Ratner et al., 2017] have used and extended these learning techniques to the setting in which the labelers are programmed rules, weak classifiers, or other heuristics. As described in the introduction, learned and programmed labelers can have heavily correlated errors because of common elements in the heuristics they use. This potential problem has motivated attempts to relax the independence assumption. One line of work [Bach et al., 2017, Varma et al., 2019] has tried to learn more sophisticated parametric models of the labelers, but they are still limited by how correct their assumptions are, which are hard to verify in practice. In this work, we therefore focus on methods for learning from weak supervision without assumptions on the distribution of labeler outputs and ground truth.

2.3 Preliminaries

We denote scalar and generic items as lowercase symbols, vectors as lowercase bold symbols, and matrices are bold uppercase symbols. The i-th column of a matrix $\boldsymbol{A}$ as denoted by the corresponding lowercase symbol $\boldsymbol{a}_i$, i.e., $\boldsymbol{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n]$. Due to space constraints, the proofs are deferred to the Appendix.

In multi class learning, we have a domain $\mathcal{X}$ and a classifier function h that maps each $x \in \mathcal{X}$ to one of k possible labels (classes). Since we will work later with distributions over the k classes, it is convenient to represent label i, $1 \leq i \leq k$ as a k dimension vector $\boldsymbol{e}_i$ with all the components set to 0 except for the i-th component that is set to 1. Thus, $h : \mathcal{X} \to \mathcal{Y} = \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_k\}$. A classifier (e.g., softmax layer of a neural network) may output a probability distribution vector $\boldsymbol{y} \in \mathbb{R}^k_{\geq 0}$ over the k classes, where y_c is the probability that the item belongs to class c, and $\sum_c y_c = 1$. We call $\mathcal{Y}^\diamond \supset \mathcal{Y}$ the set of all possible probability vectors. A *loss function* $\ell : \mathcal{Y}^\diamond \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ quantifies the error of the classifier's output $h(x)$ with respect to the true label $\boldsymbol{y}$. Let $p_{\mathcal{X}\mathcal{Y}}$ be the probability distribution over $\mathcal{X} \times \mathcal{Y}$. Given a classifier h, its *risk* is defined as

$$\mathrm{R}(h) \doteq \mathop{\mathbb{E}}_{x \sim p_{\mathcal{X}\mathcal{Y}}} \ell(h(x), \boldsymbol{y})$$

In standard supervised learning, we are given *labeled samples* from $p_{\mathcal{X}\mathcal{Y}}$, and we find a classifier with low risk among a set of classifiers $\mathcal{H}$, which is also called a *hypothesis class*. The amount of labeled data required to guarantee that we can find (or *train*) such a classifier is referred to *sample complexity*, and it is proportional to a measure of richness of $\mathcal{H}$. For many classification tasks of interest, there could be low availability of labeled data, and this is a critical problem for a wide range of domains, where the most successful hypothesis classes have high dimension (e.g., convolutional neural networks for images).

In this work we assume access to m_L i.i.d. labeled samples $(\tilde{x}_1, \tilde{\boldsymbol{y}}_1), \ldots, (\tilde{x}_{m_L}, \tilde{\boldsymbol{y}}_{m_L})$ from $p_{\mathcal{X}\mathcal{Y}}$, where the amount m_L is insufficient for the direct supervised learning of $\mathcal{H}$. To circumvent the lack of sufficient training data, we assume access to a set of weak labelers (classifiers) $\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_n$, also called weak supervision sources. These labelers are *weak* in that they can be inaccurate with respect to the target classification task. For example, the weak labelers could be trained for classification tasks that are only mildly related to the target classification task. For example, we could train a labeler to detect stripes on zebras and horses, and then attempt to use it to label data as either tigers or lions. Moreover, we add no further assumptions on the properties of those classifiers, and their output could be arbitrarily correlated. We also assume access to m unlabeled data points $X = \{x_1, \ldots, x_m\}$ sampled independently from the marginal distribution $p_{\mathcal{X}}$. Our method uses the weak supervision sources $\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_n$ to constraint the space of the possible labels that can be given to those unlabeled data points. We use the limited labeled data to compute statistics about the weak classifier, and then consider possible labeling of the unlabeled data X that satisfy those statistics.

As an example, suppose that we use the m_L labeled data points to compute as a statistic the *empirical risk* of each weak supervision source, i.e., we take

$$\hat{\mu}_i \doteq \frac{1}{m} \sum_{j=1}^{m} \ell(\varphi_i(\tilde{x}_j), \tilde{\boldsymbol{y}}_j) \enspace ,$$

for $i \in 1, \ldots, n$. In section 2.4 we will use related statistics in order to prove strong theoretical guarantees. If we were to assign a labeling to the unlabeled data points X, a reasonable approach would be to find a labeling such that the empirical risk of the weak supervision source i computed with respect of those labels is equal to $\hat{\mu}_i$. However, this is a computationally hard problem, as we have to assign a discrete label (from $\mathcal{Y}$) to each item, and each label affects the empirical risk of all the weak supervision sources. Moreover, there is no guarantee that we can find such a labeling for the unlabeled data, and it is unclear which labeling to choose in case there are multiple solutions.

To address the computational issues with a discrete selection of the labels, we assign a probability vector from $\mathcal{Y}^\diamond$ to each unlabeled data point. In other words, for each unlabeled item x_j, we assign a probability vector $\boldsymbol{y}$, where the c-th component of that vector represents the probability that item x_j belongs to class c. Given a classifier h, we compute the loss of the classifier on item x with respect to the probability vector $\boldsymbol{y} \in \mathcal{Y}^\diamond$ as *expected loss*. Abusing notation, let $\boldsymbol{e} \sim \boldsymbol{y}$ denote that $\boldsymbol{e} = \boldsymbol{e}_c \in \mathcal{Y}$ with probability y_c, and define

$$\ell^\diamond(h(x), \boldsymbol{y}) \doteq \mathop{\mathbb{E}}_{\boldsymbol{e}\sim\boldsymbol{y}} \ell(h(x), \boldsymbol{e}) = \sum_{c=1}^{k} y_c \cdot \ell(h(x), \boldsymbol{e}_c) \ . \tag{15}$$

We observe that this definition of loss generalizes the one computed with respect to a discrete labeling, in fact for each $\boldsymbol{e} \in \mathcal{Y}$, we have that $\ell(h(x), \boldsymbol{e}) = \ell^\diamond(h(x), \boldsymbol{e})$. The loss (15) is computationally easier to work with, as it is *linear* with respect to the labeling $\boldsymbol{y}$. Let $\boldsymbol{Y} \in \mathbb{R}^{k\times m}$ be a matrix that describes a possible labeling of the unlabeled data points; in particular the jth column of the matrix $\boldsymbol{Y}$ is $\boldsymbol{y}_j \in \mathcal{Y}^\diamond$, and it denotes the probability vector of the labeling of the item x_j. The empirical risk of a classifier h on the unlabeled data X with labeling $\boldsymbol{Y}$ is then defined as

$$\hat{\mathrm{R}}(h; X, \boldsymbol{Y}) \doteq \frac{1}{m} \sum_{j=1}^{m} \ell^\diamond(h(x_j), \boldsymbol{y}_j) \ .$$

Finding a labeling $\boldsymbol{Y}$ for which $\hat{\mathrm{R}}(h; X, \boldsymbol{Y}) = \hat{\mu}_i$ for $i \in 1, \dots, n$ is equivalent to the computationally easy task of solving a linear system with $\mathbf{O}(n + m)$ constraints (the n constraints on the empirical risk equality and m constraints on probability vectors summing to 1) and $\mathbf{O}(mk)$ variables. However, there still could be multiple solutions to such a linear system. The core idea of the method presented in section 2.4 is to find a model that has the lowest risk across all those possible solutions.

2.4 Adversial Labeling Method

Let $\mathcal{H} = \{\boldsymbol{h_\theta} : \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d\}$ be the hypothesis class that we will use to find the classifier for the classification task of interest, where each classifier $h \in \mathcal{H}$ is parametrized by a vector of weights $\boldsymbol{\theta}$.

Let $\boldsymbol{Y}^*$ be the (unknown) true labelings of the unlabeled data X. For each weak supervision source i, we use the labeled data to compute a interval $\triangle_i$ such that with high probability we have that $\hat{\mathrm{R}}(\boldsymbol{\varphi}_i(x); X, \boldsymbol{Y}^*) \in \triangle_i$ for $i \in 1, \ldots, n$. This is a crucial property that we will need to show our theoretical bound (theorem 2.4.8), and one may refer to lemma 2.4.1 to see how to build such intervals.

Let $\mathbb{Y}^\diamond$ be the set of all possible labeling matrices $\boldsymbol{Y}$ such that the empirical risk of $\boldsymbol{\varphi}_i$ computed with respect to the labeling $\boldsymbol{Y}$ of the unlabeled data X belongs to the corresponding interval $\triangle_i$ for each weak supervision source. Formally, the set $\mathbb{Y}^\diamond$ is defined as

$$\mathbb{Y}^\diamond \doteq \left\{ \boldsymbol{Y} \in \mathbb{R}^{k \times m} \;\middle|\; \begin{array}{ccc} \boldsymbol{y}_j & \in \mathcal{Y}^\diamond & \forall j \in 1, \ldots, m \\ \hat{\mathrm{R}}(\varphi_i; X, \boldsymbol{Y}) & \in \triangle_i & \forall i \in 1, \ldots, n \end{array} \right\} .$$

We will refer to $\mathbb{Y}^\diamond$ as the set of feasible labelings. The next lemma shows how to build the intervals $\triangle_i$ to guarantee that with high probability the true labeling $\boldsymbol{Y}^*$ is feasible.

Lemma 2.4.1 (Two-Sample Bounded Difference Interval Statistics)**.** Suppose that the codomain of the loss function ℓ is contained in the interval $[0, B]$. Let $\hat{\mu}_1, \ldots, \hat{\mu}_n$ be the empirical risks of $\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_n$ computed with respect to the m_L labeled samples. Fix a value $\delta \in (0, 1)$ and let γ be defined as

$$\gamma \doteq B \sqrt{\frac{(m_L + m) \ln \frac{2n}{\delta}}{2 m_L m}} .$$

If we set $\triangle_i = [\mu_i - \gamma, \mu_i + \gamma]$, then with probability at least $1 - \delta$ it holds that $\boldsymbol{Y}^* \in \mathbb{Y}^\diamond$.

We want to find the classifier that achieves the lowest empirical risk among the feasible labelings of the unlabeled data points. That is, we choose the classifier $\boldsymbol{h}_{\hat{\boldsymbol{\theta}}} \in \mathcal{H}$, where $\hat{\boldsymbol{\theta}}$ is the solution of the following minimax problem

$$\hat{\boldsymbol{\theta}} \doteq \operatorname*{argmin}_{\boldsymbol{\theta} \in \Theta} \max_{\boldsymbol{Y} \in \mathbb{Y}^\diamond} \hat{\mathrm{R}}(\boldsymbol{h}_{\boldsymbol{\theta}}; X, \boldsymbol{Y}) \ . \tag{16}$$

The optimization problem above has some nice properties. The set $\mathbb{Y}^\diamond$ is specified by linear constraints in $\boldsymbol{Y}$. Moreover, the objective of the minimax (16) problem is also linear in $\boldsymbol{Y}$. Hence, it is easy to see that for a given $\boldsymbol{\theta} \in \Theta$, it is possible to solve the maximization problem

$$f(\boldsymbol{\theta}) \doteq \max_{\boldsymbol{Y} \in \mathbb{Y}^\diamond} \hat{\mathrm{R}}(\boldsymbol{h}_{\boldsymbol{\theta}}; X, \boldsymbol{Y}) \tag{17}$$

through a *linear program* with $\mathbf{O}(mk)$ variables and $\mathbf{O}(m + n)$ linear constraints.

In order to solve the minimax problem (16), we will introduce a few assumptions on the loss function and the model choice $\mathcal{H}$, which are satisfied by many classic machine learning settings. In particular, we would like the function $f(\boldsymbol{\theta})$ to be convex, so that we can solve the minimization problem $\min_{\boldsymbol{\theta}\in\Theta} f(\boldsymbol{\theta})$. Even if $f(\boldsymbol{\theta})$ is convex, we may not be able to apply a gradient-based optimization method as $f(\boldsymbol{\theta})$ involves a maximization, hence it is not differentiable everywhere. To solve this issue, we will use the concept of subgradient, which is a generalization of the concept of gradient. This will require the loss function to be Lipschitz. A function $g : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ is said to be L-Lipschitz if for any $\boldsymbol{a}, \boldsymbol{a'} \in \mathbb{R}^{d_1}$, it holds that $\left\|g(\boldsymbol{a}) - g(\boldsymbol{a'})\right\|_2 \leq L\left\|\boldsymbol{a} - \boldsymbol{a'}\right\|_2$.

Definition 2.4.2 (The Subgradient)**.** Let $\mathcal{A} \subseteq \mathbb{R}^b$ be the domain of a function g. A vector $\boldsymbol{v} \in \mathbb{R}^b$ is a subgradient for a function g at $\boldsymbol{x} \in \mathcal{A}$ if for any $\boldsymbol{y} \in \mathcal{A}$ we have that

$$g(\boldsymbol{a'}) - g(\boldsymbol{a}) \geq \boldsymbol{v}^T \cdot (\boldsymbol{a'} - \boldsymbol{a}) \ .$$

For each $\boldsymbol{a} \in \mathcal{A}$, we define

$$\partial g(\boldsymbol{a}) \doteq \{\boldsymbol{v} : \boldsymbol{v} \text{ is a subgradient of } g \text{ at } \boldsymbol{a}\} \ .$$

If a function is differentiable in a point, then its subgradient with respect to that point is unique, and it is equal to the gradient.

Fact 2.4.3. If a function g is convex, then there exists at least one subgradient for each point of its domain.

The following intermediate result that immediately follows from the definition of $\ell^{\diamond}$ will prove useful throughout this discussion.

Lemma 2.4.4 (Lipschitz Constants under Convex Combination)**.** Let $\ell(\boldsymbol{h}_{\boldsymbol{\theta}}(x), \boldsymbol{e})$ be convex and L-Lipschitz with respect to $\boldsymbol{\theta}$ for any $(x, \boldsymbol{e}) \in \mathcal{X} \times \mathcal{Y}$. Then, for any probability vector $\boldsymbol{y} \in \mathcal{Y}^{\diamond}$, the function $\ell^{\diamond}(\boldsymbol{h}_{\boldsymbol{\theta}}(x), \boldsymbol{y})$ is also convex and L-Lipschitz with respect to $\boldsymbol{\theta}$.

The next Lemma shows that under some conditions, it is possible to compute the subgradient of the function f.

Lemma 2.4.5 (The Subgradient at Nondifferentiable Points)**.** Fix a value $\boldsymbol{\theta} \in \Theta$, let $\boldsymbol{Y'} \doteq \operatorname{argmax}_{\boldsymbol{Y}\in\mathbb{Y}^{\diamond}} \hat{\mathrm{R}}(\boldsymbol{h}_{\boldsymbol{\theta}}; X, \boldsymbol{Y})$, and assume that $\ell(\boldsymbol{h}_{\boldsymbol{\theta}}(x), \boldsymbol{e})$ is convex with respect to $\boldsymbol{\theta}$ for any $x \in \mathcal{X}$ and $\boldsymbol{e} \in \mathcal{Y}$. Then

$$\emptyset \neq \partial \mathrm{R}(\boldsymbol{h}_{\boldsymbol{\theta}}; X, \boldsymbol{Y'}) \subseteq \partial f(\boldsymbol{\theta}) \ .$$

A subgradient-based optimization approach [Shor, 2012] is similar to gradient descent, however at each

Algorithm 1 Adversarial Subgradient Algorithm

Input: Number of iterations T, step size h, $\mathcal{H}$, X, $\boldsymbol{\varphi}_1, \dots, \boldsymbol{\varphi}_n$

Output: Approximate solution $\tilde{\boldsymbol{\theta}}$ of (16) (See theorem 2.4.6)

$\tilde{\boldsymbol{\theta}}^{(0)} = \boldsymbol{\theta}^{(0)} \leftarrow$ arbitrary point $\boldsymbol{\theta} \in \Theta$

for $t \in 1, \dots, T$ **do**

$\quad \boldsymbol{Y}' \leftarrow \mathrm{argmax}_{\boldsymbol{Y} \in \mathbb{Y}} \hat{\mathrm{R}}(\boldsymbol{h}_{\boldsymbol{\theta}^{(t-1)}}, \boldsymbol{Y})$

$\quad \boldsymbol{v} \leftarrow$ arbitrary vector from $\partial \hat{\mathrm{R}}(\boldsymbol{h}_{\boldsymbol{\theta}}, \boldsymbol{Y}')$

$\quad \boldsymbol{\theta}^{(t)} \leftarrow \mathrm{Proj}_{\Theta}(\boldsymbol{\theta}^{(t-1)} - h\boldsymbol{v})$ $\qquad$ ($\mathrm{Proj}_{\Theta}(\cdot)$ denotes projection onto Θ)

$\quad \tilde{\boldsymbol{\theta}}^{(t)} \leftarrow \mathrm{argmin}\{f(\tilde{\boldsymbol{\theta}}^{(t-1)}), f(\boldsymbol{\theta}^{(t)})\}$

end for

Return $\tilde{\boldsymbol{\theta}}^{(T)}$

iteration we use a subgradient instead of the gradient, and we memorize the best solution found among all the iterations.

The subgradient-based optimization algorithm used to solve the optimization problem (16) is presented in algorithm 1.

As observed before, $\boldsymbol{Y}'$ as defined in the algorithm can be computed by solving a linear program. The projection step depends on the set of parameters Θ. While this is not a requirement for our approach, if the loss function $\ell(\boldsymbol{h}_{\boldsymbol{\theta}}(x), \boldsymbol{y})$, is differentiable with respect to $\boldsymbol{\theta}$, then we can compute the gradient of the empirical risk instead of a subgradient.

Theorem 2.4.6 (Optimality Guarantees of the Subgradient Method)**.** Suppose that for any $(x, \boldsymbol{y}) \in \mathcal{X} \times \mathcal{Y}$, $\ell(\boldsymbol{h}_{\boldsymbol{\theta}}(x), \boldsymbol{y})$ is L-Lipschitz continuous and convex with respect to $\boldsymbol{\theta}$. Let *step size* $h > 0$, and *iteration count* $T \in \mathbb{N}$, and $\tilde{\boldsymbol{\theta}}$ as returned by algorithm 1. Then, we have that

$$f(\tilde{\boldsymbol{\theta}}) - f(\hat{\boldsymbol{\theta}}) \leq \frac{\varnothing^2(\Theta) + L^2 h^2 T}{2hT} \enspace ,$$

where $\varnothing(\cdot)$ is computed with respect to the ℓ_2-norm, i.e., $\varnothing^2(\Theta) \doteq \max_{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta} \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2^2$, and $\hat{\boldsymbol{\theta}}$ is defined as in (16).

Corollary 2.4.7 (Convergence Rates of the Adversarial Subgradient Algorithm)**.** Suppose that the same assumptions of theorem 2.4.6 hold. For any $\varepsilon > 0$, then if $h = \varepsilon / L^2$ and $T \geq \frac{L^2 \varnothing^2(\Theta)}{\varepsilon^2}$, we have that

$$f(\tilde{\boldsymbol{\theta}}) - f(\hat{\boldsymbol{\theta}}) \leq \varepsilon \enspace .$$

Therefore, we can compute a solution within ε-additive error of (16) in $\mathbf{O}(\frac{L^2 \varnothing^2(\Theta)}{\varepsilon^2})$ steps of the subgradient algorithm.

2.4.1 Risk Guarantees and Statistical Learning

In this section, we develop a bound on the true risk of the classifier $\boldsymbol{h}_{\hat{\boldsymbol{\theta}}}$ that is a solution of the optimization problem (16). The bounds are expressed in function of the Rademacher complexity of the function family $\mathcal{L} \doteq \{\ell^\diamond \circ h : h \in \mathcal{H}\}$ that describes the loss of each function $h \in \mathcal{H}$, the risk minimizer $\boldsymbol{\theta}^* = \operatorname{argmin}_{\boldsymbol{\theta}\in\Theta} \mathrm{R}(h_{\boldsymbol{\theta}})$, and the normalized ℓ_1 diameter $D_{\mathbb{Y}^\diamond}$ of the feasible set of solutions $\mathbb{Y}^\diamond$.

$$D_{\mathbb{Y}^\diamond} \doteq \sup_{\boldsymbol{Y}', \boldsymbol{Y}'' \in \mathbb{Y}^\diamond} \frac{1}{m} \sum_{j=1}^{m} \left\| \boldsymbol{y}'_j - \boldsymbol{y}''_j \right\|_1 \quad . \tag{18}$$

The quantity $D_{\mathbb{Y}^\diamond}$ characterizes the information given by the classifiers $\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_n$ on the classification task. In particular, a weak supervision source provides useful information on the classification task of interest only if it reduces the size of the feasible set, and it provably improves the performance of our algorithm if it decreases the average diameter $D_{\mathbb{Y}^\diamond}$.

Given a function family $\mathcal{L}$, we define the empirical Rademacher average [see Mitzenmacher and Upfal, 2017] of the unlabeled items X and a possible labeling $\boldsymbol{Y}$ of those items as

$$\hat{\mathfrak{R}}(\mathcal{L}; X, \boldsymbol{Y}) \doteq \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{\ell^\diamond \circ h \in \mathcal{L}} \frac{1}{m} \sum_{i=1}^{m} \boldsymbol{\sigma}_i \ell^\diamond(h(x_i), \boldsymbol{y}_i) \right] \quad ,$$

where $\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_m$ are independent random variables from the Rademacher distribution, i.e., $\mathbb{P}(\boldsymbol{\sigma}_i = 1) = \mathbb{P}(\boldsymbol{\sigma}_i = -1) = \frac{1}{2}$. Intuitively, this quantity measures the *capacity* of $\mathcal{H}$ to *overfit*, and under mild conditions, approaches 0 as sample size m tends to infinity, in which case overfitting becomes impossible.

Theorem 2.4.8 (Optimality of Adversarial Solutions)**.** Let $\boldsymbol{h}_{\hat{\boldsymbol{\theta}}}$ be the solution of (16). Let $\boldsymbol{\theta}^* \in \operatorname{argmin}_{\boldsymbol{\theta}\in\Theta} \mathrm{R}(\boldsymbol{h}_{\boldsymbol{\theta}})$. Suppose that the codomain of the loss function ℓ is contained in the interval $[0, B]$. Let $\boldsymbol{Y}^*$ be the true (unknown) labeling of the unlabeled data X, and assume that $\boldsymbol{Y}^* \in \mathbb{Y}^\diamond$. Then, with probability $1 - \delta$ it holds that

$$\mathrm{R}(\boldsymbol{h}_{\hat{\boldsymbol{\theta}}}) \leq \mathrm{R}(\boldsymbol{h}_{\boldsymbol{\theta}^*}) + B D_{\mathbb{Y}^\diamond} + \sup_{\boldsymbol{Y} \in \mathbb{Y}^\diamond} 4\hat{\mathfrak{R}}(\mathcal{L}; X, \boldsymbol{Y}) + \mathbf{O}\left(B \sqrt{\frac{\ln \frac{1}{\delta}}{m}} \right) \quad . \tag{19}$$

The bound of theorem 2.4.8 requires the computation of the diameter of the set $\mathbb{Y}^\diamond$ with respect to the

ℓ_1-norm. This set is a polytope $\mathbb{Y}^\diamond \subseteq \mathbb{R}^{km}$, though due to the *equality constraints* for each unit k-simplex on each of the m unlabeled points, it may be embedded and expressed as a standard *H-polytope* in $\mathbb{R}^{(k-1)m}$, as it is specified by *linear inequality constraints* on the possible labelings. While this bound provides insight on how the information provided by the weak supervision sources affect the quality of our solution, computing the diameter is hard. Brieden et al. [2001] show that computing the ℓ_1 diameter is NP-hard, as is computing even a multiplicative $\sqrt{(k-1)m}$ approximation. Furthermore, particularly in the case of $k = 2$, it seems straightforward to construct problem instances with a great degree of control over the constraints, so there is no reason to think our problem is in general easier than the general case.

2.4.2 The Centerpoint Strategy: Bounds without Adversarial Learning

Although we have shown a polynomial-time algorithm for (convex) minimax optimization, it is worth noting that solving a linear program at each subgradient step adds significant overhead to training time, and this strategy also precludes using our method with out-of-the-box optimizers and machine learning algorithms. We also note that the minimax gap is necessarily small when the feasible set $\mathbb{Y}^\diamond$ is small, as captured by the *diameter term* of theorem 2.4.8. Taking this intuition into account, we know that if the feasible set is small, then there is little difference between a *minimax-optimal* model, and a model trained to optimize *any point* in the *feasible set*.

In this subsection, we adopt a methodology based on this idea, where a point in the feasible set, termed a *centerpoint*, is selected, a model is optimised for that (fixed) labelling, using any black-box optimizer, and generalization bounds are obtained for said model. We derive these bounds as corollaries of a minimax bound, so they are not as statistically efficient, however they may be computationally more efficient, and are clearly more modular, and thus this method is easier to incorporate into existing systems. Finally, while any point in the feasible set may be used as a centerpoint, we know that particular choices may yield desirable properties and intuitive interpretations, so we discuss such choices. The next theorem quantifies the performance of the centerpoint method.

Theorem 2.4.9 (Generalization Bounds for Feasibility-Constrained Learning)**.** Fix $\boldsymbol{\theta} \in \Theta$, and let $\boldsymbol{Y} \in \mathbb{Y}^\diamond$ be a feasible labeling. Suppose that the codomain of the loss function ℓ is contained in the interval $[0, B]$. Let $\boldsymbol{Y}^*$ be the true (unknown) labeling of the unlabeled data X, and assume that $\boldsymbol{Y}^* \in \mathbb{Y}^\diamond$. Then, with probability $1-\delta$, it holds that

$$\mathrm{R}(\boldsymbol{h_\theta}) \leq \sup_{\boldsymbol{Y}'\in\mathbb{Y}^\diamond} \hat{\mathrm{R}}(\boldsymbol{h_\theta};(X,\boldsymbol{Y}')) + 2\sup_{\boldsymbol{Y}'\in\mathbb{Y}^\diamond} \hat{\boldsymbol{\mathfrak{R}}}_m(\mathcal{L};(X,\boldsymbol{Y}')) + 3B\sqrt{\frac{\ln\frac{1}{\delta}}{2m}} \tag{20}$$

$$\leq \hat{\mathrm{R}}(\boldsymbol{h_\theta};(X,\boldsymbol{Y})) + \frac{B}{2}\sup_{\boldsymbol{Y}'\in\mathbb{Y}^\diamond}\left\|\boldsymbol{Y}-\boldsymbol{Y}'\right\|_{1,1} + 2\sup_{\boldsymbol{Y}'\in\mathbb{Y}^\diamond} \hat{\boldsymbol{\mathfrak{R}}}_m(\mathcal{L};(X,\boldsymbol{Y}')) + 3B\sqrt{\frac{\ln\frac{1}{\delta}}{2m}} \tag{21}$$

$$\leq \hat{\mathrm{R}}(\boldsymbol{h_\theta};(X,\boldsymbol{Y})) + BD_{\mathbb{Y}^\diamond} + 2\sup_{\boldsymbol{Y}'\in\mathbb{Y}^\diamond} \hat{\boldsymbol{\mathfrak{R}}}_m(\mathcal{L};(X,\boldsymbol{Y}')) + 3B\sqrt{\frac{\ln\frac{1}{\delta}}{2m}} \ . \tag{22}$$

Theorem 2.4.9 provides generalization bounds with respect to a fixed model $\boldsymbol{\theta}\in\Theta$. The inequality (20) is a standard minimax bound, and it can be observed that the right-hand side is minimized by choosing $\boldsymbol{\theta}$ equal to the solution of (16). Naturally, the question arises, how should $\boldsymbol{\theta}$ be selected to optimize the other forms? In particular, we seek reasonable choices of $\boldsymbol{Y}$ such that $\boldsymbol{\theta}$ trained to optimize $\boldsymbol{Y}$ perform well. For this purpose, we will ignore the $\hat{\mathrm{R}}(\boldsymbol{h_\theta};(X,\boldsymbol{Y}))$ term, as this can't be computed without knowing $\boldsymbol{Y}$, and also the $2\sup_{\boldsymbol{Y}'\in\mathbb{Y}^\diamond}\hat{\boldsymbol{\mathfrak{R}}}_m(\mathcal{L};(X,\boldsymbol{Y}'))$ term, as this term can generally be bounded independently of $\boldsymbol{Y}$.

The next bound (21) shows that the generalization bound is minimized by picking the point $\boldsymbol{Y}$ that minimizes the *radius term* $\sup_{\boldsymbol{Y}'\in\mathbb{Y}^\diamond}\left\|\boldsymbol{Y}-\boldsymbol{Y}'\right\|_{1,1}$. Such a minimizer is known as an ℓ_1 *outer Chebyshev center*, and for this reason is a natural choice of centerpoint. Unfortunately, Brieden et al. [2001] show that computing the ℓ_1 outer Chebyshev center of a polytope is as hard as computing the diameter ℓ_1, making this centerpoint choice difficult to work with. In a related problem, Eldar et al. [2008] obtain promising results with a relaxation of the ℓ_2 outer Chebyshev center, and we are hopeful that similar techniques may lead to an effective and easily computed centerpoint choice in this setting.

Another intuitive choice is the *arithmetic mean* (a.k.a. the *centroid* or *barycenter*) of the feasible region $\mathbb{Y}^\diamond$. This is rather elegant, as the hypothesis $h_{\boldsymbol{\theta}}$ that minimizes the empirical risk *on average* over $\mathbb{Y}^\diamond$ also minimizes empirical risk of the *average labeling* in $\mathbb{Y}^\diamond$, i.e., we have

$$\boldsymbol{Y} \doteq \min_{\boldsymbol{\theta}\in\Theta}\frac{1}{Z}\int_{\boldsymbol{y}\in\mathbb{Y}^\diamond}\hat{\mathrm{R}}(h_{\boldsymbol{\theta}};X,Y)\;\mathrm{d}Y = \min_{\boldsymbol{\theta}\in\Theta}\hat{\mathrm{R}}\left(h_{\boldsymbol{\theta}};X,\frac{1}{Z}\int_{\mathbb{Y}^\diamond}Y\,\mathrm{d}Y\right) \ ,$$

for normalizing constant Z. This gives us a nice interpretation as an *average-case optimal solution* over the uncertainty over feasible labelings. Similarly, we can also consider only the *integer solutions* (i.e., integrate and normalize w.r.t. $\mathbb{Y}$ instead of $\mathbb{Y}^\diamond$). This is an instance of the *vertex centroid problem*, which Elbassioni and Tiwary [2008] show to be #P hard, however they also show it can be approximated efficiently via sampling

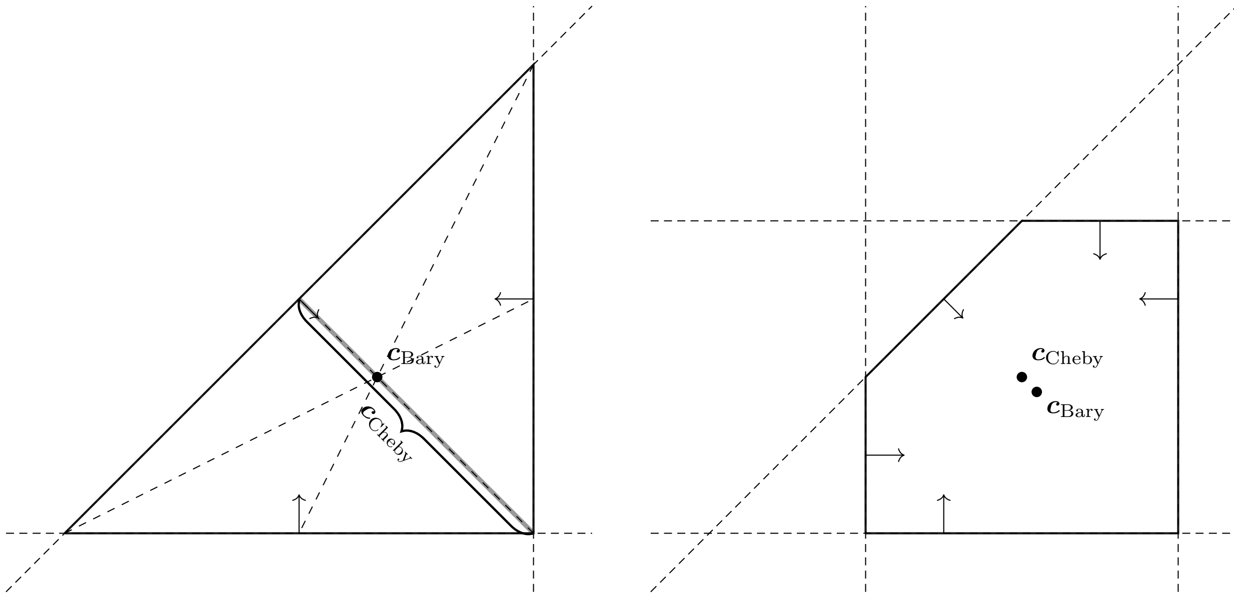


Figure 2.1: Two polytopes in $\mathbb{R}^2$, with their ℓ_1 outer Chebyshev centers and barycenters. The barycenter is always uniquely defined, but in the right isosceles triangle, the Chebyshev center is not unique, but rather consists of the entire perpendicular bisector of the hypotenuse.

methods. The ℓ_1 outer Chebyshev center and barycenter are contrasted graphically on example polytopes in figure 2.1.

2.4.3 Applications

In order to feature the generality of our framework, we show two examples of different instantiations of the optimization problem (16) for different choices of loss function and prediction models for which we can apply corollary 2.4.7.

Convex combination of the weak supervision sources. Let $\Theta = \{\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n) \in \mathbb{R}_+^n : \sum_{i=1}^n \theta_i = 1\}$. Our prediction model is a convex combination of the output of the weak classifiers $\boldsymbol{\varphi}, \ldots, \boldsymbol{\varphi}_n$. In particular, given $\boldsymbol{\theta} \in \Theta$, the classifier $\boldsymbol{h_\theta}$ is defined as $\boldsymbol{h_\theta}(x) = \sum_{i=1}^n \theta_i \boldsymbol{\varphi}_i(x)$ for any $x \in \mathcal{X}$. It is easy to see that $\varnothing(\Theta) \leq \sqrt{2}$. Given an arbitrary vector $\boldsymbol{v} \in \mathbb{R}^n$, the projection step to Θ can be done efficiently by using for example the algorithm in Wang and Carreira-Perpinán [2013].

Let ℓ be the Brier loss, defined as for any $(x, \boldsymbol{e}) \in \mathcal{X} \times \mathcal{Y}$ is defined as:

$$
\begin{aligned}
\ell(\boldsymbol{h_\theta}(x), \boldsymbol{e}) &\doteq \sum_{c=1}^{k} \left(\boldsymbol{h_\theta}(x)_c - \boldsymbol{e}_c\right)^2 \\
&= ||\boldsymbol{h_\theta}(x)||_2^2 - 2\boldsymbol{h_\theta}(x)^T \cdot \boldsymbol{e} + 1 \ \ .
\end{aligned}
$$

It is easy to see that the function $\ell(\boldsymbol{h}_{\boldsymbol{\theta}}(x), \boldsymbol{e})$ is convex and differentiable with respect to $\boldsymbol{\theta}$. Moreover, the loss function has codomain $[0, 2]$.

Lemma 2.4.10. The loss $\ell(\boldsymbol{h}_{\boldsymbol{\theta}}(x), \boldsymbol{e})$ of a prediction model $\boldsymbol{h}_{\boldsymbol{\theta}}$ defined as in this subsection is $2\sqrt{n}$-Lipschitz continuous with respect to $\boldsymbol{\theta}$.

softMax (multinomial logistic legression). Suppose that each item is a vector in $\mathbb{R}^b$, i.e., $\mathcal{X} \subseteq \mathbb{R}^b$, and assume that $||\boldsymbol{x}||_2 \leq B_x$ for any $x \in \mathcal{X}$. Let $\Theta \doteq \{\boldsymbol{\theta} = (\boldsymbol{w}_1^T \dots \boldsymbol{w}_k^T) \in \mathbb{R}^{k \times b} : \boldsymbol{w}_c \in \mathbb{R}^b \wedge ||\boldsymbol{w}_c||_2 \leq B_w$ for $c \in 1, \dots, k\}$. That is, $\boldsymbol{\theta}$ is the concatenation of k vectors with bounded norm. If we are given $\boldsymbol{\theta}$ Observe that with this definition of Θ, we have that $\varnothing(\Theta) \leq \sqrt{2k} B_w$. Given a vector $\boldsymbol{\theta} = (\boldsymbol{w}_1^T \dots \boldsymbol{w}_k^T)$, the projection step to Θ is simply $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{w}}_1^T \dots \tilde{\boldsymbol{w}}_k^T)$, where $\tilde{\boldsymbol{w}}_c = \boldsymbol{w}_c / \min(B_w/||\boldsymbol{w}_c||_2, 1)$ for $c \in 1, \dots, k$.

Given $\boldsymbol{\theta} = (\boldsymbol{w}_1^T \dots \boldsymbol{w}_k^T) \in \Theta$, and $\boldsymbol{x} \in \mathcal{X}$, we define

$$\boldsymbol{h}_{\boldsymbol{\theta}}(\boldsymbol{x}) \doteq \left(\frac{\exp(\boldsymbol{w}_1^T \cdot \boldsymbol{x})}{\sum_{c=1}^k \exp(\boldsymbol{w}_c^T \cdot \boldsymbol{x})}, \dots, \frac{\exp(\boldsymbol{w}_k^T \cdot \boldsymbol{x})}{\sum_{c=1}^k \exp(\boldsymbol{w}_c^T \cdot \boldsymbol{x})} \right)^T \quad .$$

This classifier is a particular instantiation of softMax combined with a linear model. For a vector $\boldsymbol{v} = (v_1, \dots, v_d)^T$, define $\ln \boldsymbol{v} \doteq (\ln v_1, \dots, \ln v_d)^T$. Given $(x, \boldsymbol{e}) \in \mathcal{X} \times \mathcal{Y}$, we define the cross-entropy loss ℓ of the prediction model $\boldsymbol{h}_{\boldsymbol{\theta}}$ as

$$\ell(\boldsymbol{h}_{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{e}) \doteq -\boldsymbol{e}^T \cdot \ln(\boldsymbol{h}_{\boldsymbol{\theta}}(\boldsymbol{x})) \quad .$$

This combination of prediction model and loss function is also known as multinomial logistic regression. It is easy to see that the loss function is differentiable with respect to $\boldsymbol{\theta}$, and it is a known result that $\ell(\boldsymbol{h}_{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{e})$ is convex with respect to $\boldsymbol{\theta}$ for any $(\boldsymbol{x}, \boldsymbol{e}) \in \mathcal{X} \times \mathcal{Y}$ [Böhning, 1992]. Moreover, the loss function is bounded, and Lipschitz continuous, as shown by the next two lemmas.

Lemma 2.4.11. For any $(x, \boldsymbol{e}) \in \mathcal{X} \times \mathcal{Y}$, and $\theta \in \Theta$, we have that $\ell(\boldsymbol{h}_{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{e}) \in [0, B_w B_x + \ln k]$.

Lemma 2.4.12. For any $(\boldsymbol{x}, \boldsymbol{e})$, the loss $\ell(\boldsymbol{h}_{\boldsymbol{\theta}}(x), \boldsymbol{e})$ of a prediction model $\boldsymbol{h}_{\boldsymbol{\theta}}$ defined as in this subsection is (kB_x)-Lipschitz continuous with respect to $\boldsymbol{\theta}$.

2.4.4 Constraining the feasible set

The presentation of section 2.4 implicitly assumes an alignment between the output classes of the weak supervision sources $\boldsymbol{\varphi}_1, \dots, \boldsymbol{\varphi}_n$ and the target classification task. In fact, as seen in lemma 2.4.1, we compute the intervals $\triangle_i$ based on the empirical risk of the weak supervision sources using labeled data of the target classification task. However, for many application of interests, the weak supervision sources could output to a different codomain, potentially with a unequal number of classes. As an example, suppose that we would like to distinguish between images of $\{\text{cat}, \text{dog}, \text{rabbit}, \text{bear}\}$. A binary classifier that tells us if the animal represented in an image has a tail or not still provides a useful clue with respect to the target classification task, and we would like to use that information.

In this section, we will show how to constrain the feasible set of labelings $\mathbb{Y}^\diamond$ in a more general setting, where the weak supervision source $\boldsymbol{\varphi}_i$ is a classifier that maps elements from the domain $\mathcal{X}$ to soft labels over k_i classes, i.e., $\boldsymbol{\varphi}_i : \mathcal{X} \to \mathcal{Y}^\diamond_{k_i}$, where $\mathcal{Y}^\diamond_{k_i} = \{\boldsymbol{v} \in \mathbb{R}^{k_i}_{\geq 0} : \sum_c v_c = 1\}$.

Consider the weak supervision source $\boldsymbol{\varphi}_i$. For each $c \in 1, \ldots, k$ and $\tilde{c} \in 1, \ldots, k_i$, we use the m_L labeled data $(\tilde{x}_1, \tilde{\boldsymbol{y}}_1), \ldots, (\tilde{x}_{m_L}, \tilde{\boldsymbol{y}}_{m_L})$ to compute the statistic

$$\hat{\mu}_{i,c,\tilde{c}}(\tilde{X}, \tilde{\boldsymbol{Y}}) \doteq \frac{1}{|\tilde{X}|} \sum_{j=1}^{|\tilde{X}|} y_{j,c} [\boldsymbol{\varphi}_i(x_j)]_{\tilde{c}} \ .$$

It is clear that the function $\hat{\mu}_{i,c,\tilde{c}}(\tilde{X}, \tilde{\boldsymbol{Y}})$ is linear in $\tilde{\boldsymbol{Y}}$. For each weak supervision source $\boldsymbol{\varphi}_i$, true class $c \in 1, \ldots, k$, and weak supervision source's output class $\tilde{c} \in 1, \ldots, k_i$, based on the value $\hat{\mu}_{i,c,\tilde{c}}(\tilde{X}, \tilde{Y})$, we compute an interval $\triangle_{i,c,\hat{c}}$, defined as

$$\triangle_{i,c,\hat{c}} \doteq [\hat{\mu}_{i,c,\tilde{c}}(\tilde{X}, \tilde{Y}) - \gamma, \hat{\mu}_{i,c,\tilde{c}}(\tilde{X}, \tilde{Y}) + \gamma] \ .$$

where the value γ is specified in lemma 2.4.13.

Given a labeling $\boldsymbol{Y}$ of the unlabeled dataset X, we say that $\boldsymbol{Y}$ is a feasible solution if for each i, c and $\tilde{c}$, it holds that

$$\hat{\mu}_{i,c,\tilde{c}}(X, \boldsymbol{Y}) \in \triangle_{i,c,\hat{c}} \ . \tag{23}$$

That is, the set of all the feasible solutions $\mathbb{Y}^\diamond$ is defined as

$$\mathbb{Y}^\diamond \doteq \left\{ \boldsymbol{Y} \in \mathbb{R}^{k \times m} \,\middle|\, \begin{array}{ccc} \boldsymbol{y}_j & \in \mathcal{Y}^\diamond & \forall j \in 1, \ldots, m \\ \hat{\mu}_{i,c,\tilde{c}} & \in \triangle_{i,c,\tilde{c}} & \forall i \in 1, \ldots, n, \forall c, \tilde{c} \end{array} \right\} \ .$$

Notice that the constraints specified in $\mathbb{Y}^\diamond$ are still linear in $\boldsymbol{Y}$, therefore we can still compute the value $f(\boldsymbol{\theta})$ (as in (17)) by solving a linear program, and all discussion of empirical-risk based constraints still applies.

In order to be able to give the theoretical bound of theorem 2.4.8, we need to guarantee that the true labeling

$\boldsymbol{Y}^*$ of the unlabeled data X is feasible. This is possible by choosing a suitable value γ when defining the intervals $\triangle_{i,c,\tilde{c}}$, in the following result, which closely resembles lemma 2.4.1.

Lemma 2.4.13 (Confusion Matrix Interval Risk Bounds)**.** Suppose failure probability $\delta \in (0, 1)$. For every $i \in 1, \ldots, n$, $c \in 1, \ldots, k$, and $\tilde{c} = 1, \ldots, k_i$ let $\triangle_{i,c,\tilde{c}}$ be computed as in (23), and let $K \doteq k \sum_{i=1}^{n} k_i$. Now, if we use the value

$$\gamma \doteq \sqrt{\frac{(m_L + m) \ln \frac{2nK}{\delta}}{2 m_L m}}$$

as the radius of these intervals, then with probability at least $1 - \delta$ it holds that $\boldsymbol{Y}^* \in \mathbb{Y}^{\diamond}$.

2.5 Experiments

We demonstrate the applicability and performance of our method on image multi class classification tasks derived from the DomainNet [Peng et al., 2019] dataset. We also provide experiments on image binary classification tasks derived from the Animals with Attributes 2 [Xian et al., 2018] dataset in order to compare our methods with additional baselines.

DomainNet contains images from 6 different domains $P = \{$clipart, infograph, painting, quickdraw, real, sketch$\}$ and features 345 different classes. Animals with Attributes 2 contains natural images of 50 types of animals. Associated with the dataset is a list of 85 attributes for each animal class, which we use to create weak supervision sources. Animals with Attributes 2 is divided into 40 "seen" classes and 10 "unseen" classes, where the seen classes can be used to train attribute classifiers without leaking information about the unseen classes.

We refer to our algorithms by using the acronyms **AMCL-CC** and **AMCL-LR**, where AMCL stands for **A**dversial **M**ulti **C**lass **L**earning. AMCL-CC is an implementation of our method that uses a convex combination of the weak supervision sources as the prediction model, whereas AMCL-LR uses multinomial logistic regression (see Section 4.1). For every image, we compute the output of a pretrained ResNet-18 and use it as inputs for AMCL-LR.

2.5.1 Setup

From DomainNet, we select $k = 5$ random classes from the 25 classes with the largest number of data. Then, for each domain $p \in P$, we learn a multi class classifier $\boldsymbol{\varphi}_p$ for those k classes in domain p. The classifier $\boldsymbol{\varphi}_p$ is trained by fine-tuning a pretrained ResNet-18 network [He et al., 2016] using 60% of the labeled data for that domain. For each domain p, we consider the classifiers trained in domains other than p as weak

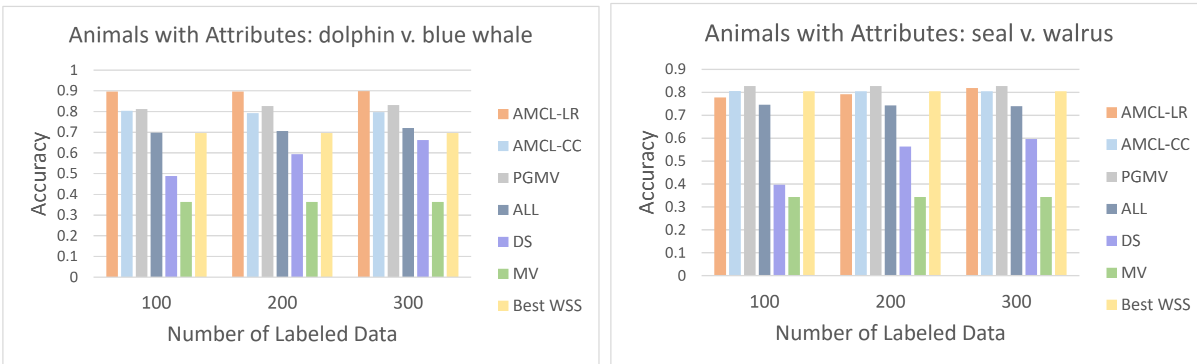


Figure 2.2: Experimental results on Animals with Attributes for the binary classification tasks of dolphin v. blue whale (left) and seal v. walrus (right) as we vary the amount of labeled data. Each method uses 560 unlabeled data for dolphin v. blue whale and 602 unlabeled data for seal v. walrus.

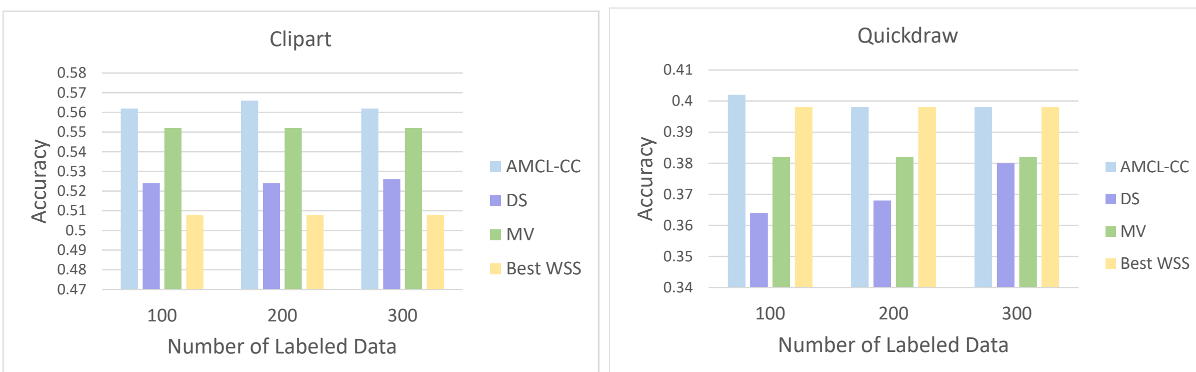


Figure 2.3: Experimental results on Domain Net for the clipart and quickdraw domains as we vary the amount of labeled data. Each method uses 500 unlabeled data. Results are listed for the 5 classes of {sea turtle, vase, whale, bird, violin }.

supervision sources, i.e., the classifiers $\{\varphi_q\}$ for $q \in P \setminus \{p\}$. We remark that these weak supervision sources never have access to samples from domain p.

From Animals With Attributes, we create binary classification tasks by selecting pairs of unseen classes. Following Mazzetto et al. [2021], we create weak supervision sources by using the seen classes to train classifiers for the attributes that distinguish them. Similarly to Domain Net, these classifiers are learned by fine-tuning a pretrained ResNet-18 network using labeled data from the seen classes. In order to focus on the most challenging tasks, where the weak supervision sources are not highly accurate, we select the 4 class pairs among the unseen classes with the lowest majority vote accuracy.

We remark that all algorithms that require unlabeled data are evaluated in a transductive setting: the unlabeled data used by the algorithms is also used to evaluate the final learnt prediction models.

2.5.2 Baselines and algorithms

Following the example of [Mazzetto et al., 2021], we compare our method with the following baselines and algorithms:

Best Weak Supervision Source (Best WSS): We report the accuracy of the best weak supervision source.

Majority Vote (MV): We consider a simple approach to combining the weak supervision sources: we average their output and select the most voted class. This approach requires no learning, but is suboptimal when weak supervision sources' errors are not independent and/or have unequal levels of accuracy.

Semi-Supervised Dawid-Skene Estimator (DS): We also consider a semi-supervised extension to the standard crowdsourcing algorithm [Dawid and Skene, 1979] that finds the optimal aggregation of the outputs of independent weak supervision sources. The Dawid-Skene estimator is also the default aggregation method for the Snorkel system [Ratner et al., 2017]. Here, we use a semi-supervised version of this algorithm, for a fair comparison with our work. We simply optimize the marginal likelihood of the weak supervision sources' outputs for unlabeled data and the joint likelihood with the label when it is observed.

Adversarial Label Learning (ALL): This algorithm [Arachie and Huang, 2019] learns a prediction model that has the highest expected accuracy with respect to an adversarial labeling of an unlabeled dataset, where this labeling must satisfy error constraints on the weak supervision sources. This approach shares similarities with our method; however, it fails to provide theoretical guarantees on the learning of the prediction model. For a fair comparison to our method, we use logistic regression as the prediction model, and use the same features given to AMCL-LR.

Performance-guaranteed Majority Vote (PGMV): This method finds a subset of weak supervision sources whose majority vote achieves high accuracy with respect to the worst-case distribution of the output the weak supervision sources. Again, this worst-case distribution is constrained by using statistics computed on the weak supervision sources (individual error rates and pairwise differences).

Due to the limitations of PGMV and ALL, we can run those algorithms only for binary classification tasks.

2.5.3 Results

Animals With Attributes (binary classification): In figure 2.2, we report the results on the Animals With Attributes dataset for two binary classification tasks.

In the binary setting, our methods match or outperform the state-of-the-art methods PGMV and ALL over all amounts of labeled data. We note that even though AMCL-LR and ALL use the same inputs and train the same prediction model, our method achieves overall higher accuracies in addition to providing theoretical guarantees on the learning of the prediction model.

Domain Net (multi-class classification): In figure 2.3, we report the accuracies of the different algorithms on the Domain Net dataset for the clipart and quickdraw domains. We remark that ALL and PGMV cannot be used in this setting, as they are restricted to binary classification.

In the multi class setting, our methods again match or outperform the baselines over all amounts of labeled data. We note that in the quickdraw domain, the weak supervision sources are overall very inaccurate, and it is difficult to recover useful information from them. However, differently from the baselines DS and MV, AMCL-CC can still recover and improve upon the best weak supervision source.

Again, as noted by the Best WSS column, the weak supervision sources are quite inaccurate in this dataset. Therefore, we do not report the results for the AMCL-LR algorithm as the weak supervision sources do not constrain the feasible set of solutions enough for our method to learn a more complex model such as multinomial logistic regression.

Due to space constraints, additional plots and experimental details for both datasets are reported in the Appendix.

2.6 Conclusion

We develop the first rigorous method that can use information provided by arbitrarily correlated weak supervision sources in order to learn a prediction model for a multi class classification task. In many practical settings, our method can optimally learn the model that achieves the smallest risk with respect to an adversarial feasible labeling of a unlabeled dataset, and we provide theoretical guarantees on the quality of the learnt model based on a measure of the information provided by the weak supervision sources. Finally, we provide experiments that illustrate the applicability of our approach and its advantages over existing methods.

Part II

Fairness with Population Means: Malfare, Welfare, and Fair PAC Learning

The rise of machine learning in industry and government has brought about a commensurate rise in algorithmic bias and fairness issues, with increasing recognition by academics that such issues must be explicitly considered in practical fairness-sensitive machine learning settings. It is now well-known that historic data reflect historic biases and injustices, and that, due to availability bias, machine learning models are frequently trained on data from prominent, privileged, or majority groups while minorities remain understudied and underdocumented, and these issues have well-studied effects on the efficacy and (un)fairness of many contemporary machine learning systems. While part I addresses this issue only implicitly, by tailoring machine learning methods and generalization guarantees to the *small sample* setting (and thus democratizing such methods to operate not just on well-studied large groups, but also on smaller populations), here I take a more direct approach to confronting algorithmic bias and fairness issues in machine learning.

For context, traditional constraint-based notions of algorithmic fairness have risen to prominence in machine learning, with the potential to correct for some forms of data and/or algorithmic bias by ensuring *demographic-parity* concepts, e.g., *equality of opportunity*, *equality of outcome*, and other such desiderata. Fairness by demographic-parity constraint has several prominent flaws: most notably, many parity constraints are *mutually unsatisfiable*, and by nature, there is inherent tension between *minimizing loss* and *constraining for fairness*, where additional *tolerance parameters* (which quantify the degree to which fairness constraints impact a solution) must be specified to strike a balance between *accuracy* and *fairness*. Perhaps in response to these issues, some recent work has trended toward welfare-based fairness-concepts, wherein both *accuracy* and *fairness* are encoded in a *welfare function* defined on a *group of subpopulations*. Welfare metrics quantify *overall wellbeing* across such populations, and welfare-based objectives and constraints incentivize *fair machine learning methods* to produce satisfactory solutions that consider the diverse needs of multiple groups. Unfortunately, welfare based approaches require a notion of (positive) per-group utility, and I argue that this is not natural to many machine learning tasks, which instead seek to *minimize* some *negatively connoted* loss value.

Introducing Malfare In this part, I derive an equivalently-axiomatically justified alternative to welfare, termed malfare, and propose its use as a cardinal objective in *fair machine learning tasks*. In particular, malfare measures overall *societal harm* (rather than wellbeing), and *malfare minimization* naturally generalizes *risk minimization* to (nonlinearly) consider model performance across multiple protected groups, and is thus ideal for tasks like ensuring *facial recognition* or *speech recognition* systems are *accurate* and *accessible* across a diverse population. I then cast fair machine learning as a direct *malfare minimization* problem, where a group's dissatisfaction is simply their risk (expected loss). Surprisingly, except in two trivial cases, the axioms of cardinal welfare (malfare) dictate that this is not equivalent to simply defining utility as negative loss. The framework is established in chapter 3, and it is instantiated in a practical learning setting in chapter 4.

Building upon these concepts, in chapter 5 I define *fair-PAC learning*, where a fair PAC-learner is an algorithm that learns an ε-δ malfare-optimal model with bounded sample complexity, for *any data distribution*, and for *any* (axiomatically justified) malfare concept. We show broad conditions under which, with appropriate modifications, many standard PAC-learners may be converted to fair-PAC learners. This places fair-PAC learning on firm theoretical ground, as it yields *statistical efficiency* guarantees for many well-studied machine-learning models, and is also practically relevant, as it democratizes fair machine learning by providing concrete training algorithms and rigorous generalization guarantees for these models. In particular, I show via a constructive polynomial reduction that *realizable fair PAC-learning* reduces to *realizable PAC-learning*, and furthermore, I show, non-constructively, that for learning problems where PAC-learnability implies uniform convergence, it is equivalent to fair-PAC-learnability. In addition to *statistical guarantees*, I also show *computational guarantees*: in particular, I show that when training is possible via *convex optimization* or efficient-enumeration of an approximate cover of the hypothesis space, then ε-δ training malfare-optimal models, as with risk-optimal models, requires polynomial time.

A Preview of Technical Results As for statistical guarantees in malfare estimation, corollary 3.4.2 shows that for any fair $\boldsymbol{w}$-weighted malfare function $\mathrm{M}(\cdot; \boldsymbol{w})$ (precisely defined in chapter 3), empirical malfare is tightly concentrated about true malfare. In particular, assume g groups, any *loss function* $\ell : \mathcal{X} \to [0, r]$, risk values $\mathcal{S} \in [0, r]^g$ s.t. $\mathcal{S}_i = \mathbb{E}_{\mathcal{D}_i}[\ell]$, samples $\boldsymbol{x}_i \sim \mathcal{D}_i^m$ and empirical risk values $\hat{\mathcal{S}}_i = \frac{1}{m} \sum_{j=1}^{m} \ell(\boldsymbol{x}_{i,j})$. Then with probability at least $1 - \delta$ over choice of $\boldsymbol{x}$, we have

$$\left| \mathrm{M}_p(\mathcal{S}; \boldsymbol{w}) - \mathrm{M}_p(\hat{\mathcal{S}}; \boldsymbol{w}) \right| \leq r \sqrt{\frac{\ln \frac{2g}{\delta}}{2m}} \ . \tag{24}$$

Malfare estimation is more challenging than risk estimation, as the empirical malfare is in general a *biased estimator* of true malfare. Still, this result is clearly analogous to Hoeffding's inequality for the risk, suffering only a $\sqrt{\log g}$ dependence on the number of groups g. Alternatively, again with probability at least $1 - \delta$ over choice of $\boldsymbol{x}$, we have a (variance-dependent) Bennett's inequality analogue of

$$\left| \mathrm{M}_p(\mathcal{S}; \boldsymbol{w}) - \mathrm{M}_p(\hat{\mathcal{S}}; \boldsymbol{w}) \right| \leq \frac{r \ln \frac{2g}{\delta}}{3m} + \max_{i \in 1, \dots, g} \sqrt{\frac{2\, \mathbb{V}_{\mathcal{D}_i}[\ell] \ln \frac{2g}{\delta}}{m}} \ . \tag{25}$$

These results should be contrasted with inequalities (1) and (2) for mean-estimation, as they generalize said results to malfare estimation. It is a straightforward matter to extend them to depend on *empirical variances*, and I further generalize to give *uniform convergence guarantees* over *function families* in chapter 4.

Chapter 3

Quantifying Population Sentiment with Welfare and Malfare

In this chapter, I address an inherent difficulty in welfare-theoretic fair machine learning by proposing an equivalently-axiomatically justified alternative and studying the resulting computational and statistical learning questions. Welfare metrics quantify *overall wellbeing* across a population consisting of one or more groups, and welfare-based objectives and constraints have recently been proposed to incentivize *fair machine learning methods* to produce satisfactory solutions that consider the diverse (often mutually incompatible) needs of multiple groups. Unfortunately, many machine-learning problems are more naturally cast as *loss minimization*, rather than *utility maximization* tasks, which complicates direct application of welfare-centric methods to fair machine learning tasks. I thus define a complementary measure to *welfare*, termed *malfare*, measuring overall *societal harm* (rather than wellbeing), with axiomatic justification via the standard axioms of cardinal welfare.

With this definition in hand, I then cast fair machine learning as a direct *malfare minimization* problem, where a group's malfare is their risk (expected loss). Surprisingly, except in two trivial cases, the axioms of cardinal welfare (malfare) dictate that this is not equivalent to simply defining utility as negative loss. Previous welfare based methods have generally defined welfare in an *ad-hoc* manner, and I show that, in particular, seemingly intuitive definitions of welfare in terms of loss lead to absurdities or non-trivial and uninterpretable hyperparameters. Furthermore, I show that malfare exhibits rapidly-converging finite-sample concentration-of-measure guarantees, and thus may be estimated with great accuracy from a small sample. This chapter is adapted from the first half of Cousins [2021].

3.1 Introduction

It is now well-understood that contemporary ML systems for tasks like facial recognition [Buolamwini and Gebru, 2018, Cook et al., 2019, Cavazos et al., 2020], medical settings [Mac Namee et al., 2002, Ashraf et al., 2018], and many others exhibit differential accuracy across gender, race, and other protected-group membership. This immediately yields *accessibility issues* to users of such systems, and can lead to *direct discrimination*, e.g., facial recognition in policing yielding disproportionate false-arrest rates, and ML in medical technology yielding disproportionate health outcomes, exacerbating existing structural and societal inequalities impacting many minority groups. In welfare-centric ML methods, both *accuracy* and *fairness* are encoded in a single *welfare function* defined on a *group of subpopulations*. Welfare is then directly optimized [Rolf et al., 2020] or constrained [Speicher et al., 2018, Heidari et al., 2018] to promote *fair learning* across *all groups*. This addresses differential performance and bias issues across groups by ensuring that (1), each group is *seen* and *considered* during training, and (2), an outcome is incentivized that is desirable overall, ideally according to some mutually-agreed-upon welfare function. Unfortunately, welfare based metrics require a notion of (positive) utility, and we argue that this is not natural to many machine learning tasks, where we instead *minimize* some *negatively connoted* loss value (e.g., in decision-theory or with proper scoring rules). We thus define a complementary measure to *welfare*, termed *malfare*, measuring societal harm (rather than wellbeing). In particular, malfare arises naturally when one applies the standard *axioms of cardinal welfare* (with appropriate modifications) to *loss* rather than *utility*. With this framework, we then cast fair machine learning as a direct *malfare minimization* problem, where a group's malfare is their risk (expected loss).

Perhaps surprisingly, defining and minimizing a *malfare function* is *not equivalent* to defining and maximizing some *welfare function*, while taking utility to be negative loss (except in the trivial cases of egalitarian and utilitarian malfare). This is essentially because nearly every function satisfying the standard axioms of cardinal welfare requires *nonnegative* inputs, and it is not in general possible to contort a loss function into a utility function while satisfying this requirement. For example, while minimizing the 0-1 loss, which simply counts the number of mistakes a classifier makes, is isomorphic to maximizing the 1-0 gain, which counts number of correct classifications, minimizing some *malfare function* defined on 0-1 loss over groups *is not* in general equivalent to maximizing any *welfare function* defined on 1-0 gain. More strikingly, for problems like minimizing square error (i.e., in regression), it is in general not even possible to define a complementary nonnegative gain function without changing the optimal solution, *even for a single group*.

We briefly summarize our contributions as follows.

1. We derive in section 3.3 the *malfare concept*, extending welfare to measure *negatively-connoted* attributes,

and show that *malfare-minimization* naturally generalizes *risk-minimization* to produce *fairness-sensitive* machine-learning objectives that consider multiple protected groups.

2. We show in section 3.4 that in many cases, while empirical estimates of welfare and malfare are *statistically biased*, they may be *sharply estimated* using standard finite-sample concentration-of-measure bounds.

3.2 Related Work

Constraint-based notions of algorithmic fairness have risen to prominence in machine learning, with the potential to ensure demographic-parity (e.g., equality of opportunity, equality of outcome, or equalized odds) thus correcting for some forms of data or algorithmic bias. While noble in intent and intuitive by design, fairness by demographic-parity constraints has several prominent flaws: most notably, several popular parity constraints are mutually unsatisfiable [Kleinberg et al., 2017], and their constraint-based formulation inherently puts *accuracy* and *fairness* at odds, where additional *tolerance parameters* are required to strike a balance between the two. Furthermore, recent works [Hu and Chen, 2020, Kasy and Abebe, 2021] have shown that *welfare* and even *disadvantaged group utility* can decrease even as fairness constraints are tightened, calling into question whether demographic parity constraints are even beneficial to those they purport to aid. Without deeper axiomatic underpinnings, it is unclear why one should choose one notion of fairness over another, and it is unsatisfying that we do not have a way of comparing and selecting between them without appealing to informal arguments.

Perhaps in response to these issues, some recent work has trended toward welfare-based fairness-concepts [Hu and Chen, 2020, Rolf et al., 2020], wherein both *accuracy* and *fairness* are encoded in a *welfare function* defined on a *group of subpopulations*. Welfare is then directly optimized [Rolf et al., 2020] or constrained [Speicher et al., 2018, Heidari et al., 2018] to promote *fair learning* across *all groups* Perhaps the most similar to our work is a method of [Hu and Chen, 2020], wherein they *directly maximize* empirical welfare over linear (halfspace) classifiers; however as with other previous works, an appropriate utility function must be selected, which we avoid by instead using malfare. We argue that *empirical welfare maximization* is an effective strategy when an appropriate and natural measure of *utility* is available, but in machine learning contexts like this, there is no "correct" or clearly neutral way to convert loss to utility. Our strategy avoids this issue by working directly in terms of malfare and loss.

The most poignant contrast to existing work we can make is to the *Seldonian learner* [Thomas et al., 2019] framework, which can be thought of as extending PAC-learning to learning problems with both *constraints* and *arbitrary nonlinear objectives*. We argue that this generality is harmful to the utility of the concept as a mathematical or practical object, as nearly any ML problem can be posed as a constrained nonlinear optimization task. The utility in fair PAC learning is that it is sophisticated enough to handle fairness issues, with a particular axiomatically justified objective, but remains simple enough to study as a mathematical object; in particular reductions between various PAC and FPAC learnable classes are of great value in understanding FPAC learning, which would not be possible in a more general framework.

3.3 Quantifying Population-Level Sentiment

A generic *population mean* function $\mathrm{M}(\mathcal{S}; \boldsymbol{w})$ quantifies some *sentiment value* $\mathcal{S}$, across a population Ω weighted by $\boldsymbol{w}$. In particular, $\mathcal{S} : \Omega \to \mathbb{R}_{0+}$ describes the *values* over which we take the mean, and $\boldsymbol{w}$, a probability measure over Ω, describes their *weights*. When $\mathcal{S}$ measures a *desirable quantity*, generally termed *utility*, the population mean is a measure of *cardinal welfare* [Moulin, 2004], and thus quantifies overall *well-being*. We also consider the inverse-notion of *ill-being*, termed *malfare*, in terms of an *undesirable* $\mathcal{S}$, generally *loss* or *risk*, which naturally extends the concept. We show an equivalent *axiomatic justification* for malfare, and argue that its use is more natural in many situations, particularly when considering or optimizing *loss functions*.

Definition 3.3.1 (Population Means: Welfare and Malfare)**.** A *population mean* function $\mathrm{M}(\mathcal{S}; \boldsymbol{w})$ measures the *overall sentiment* of population Ω, measured by *sentiment function* $\mathcal{S} : \Omega \to \mathbb{R}_{0+}$, weighted by *probability measure* $\boldsymbol{w}$ over Ω. If $\mathcal{S}$ denotes a desirable quantity (e.g., utility), we call $\mathrm{M}(\mathcal{S}; \boldsymbol{w})$ a *welfare function*, written $\mathrm{W}(\mathcal{S}; \boldsymbol{w})$, and inversely, if it is undesirable (e.g., disutility, loss, or risk), we call $\mathrm{M}(\mathcal{S}; \boldsymbol{w})$ a *malfare function*, written $\mathrm{M\!\!\!\!\backslash}(\mathcal{S}; \boldsymbol{w})$.

For now, think of the term *population mean* as signifying that an entire population, with diverse and subjective desiderata, is considered and summarized, rather than a single objective viewpoint. As we introduce axioms and show consequent properties, the appropriateness of the term shall become more apparent. Note that we use the term *sentiment* to refer to $\mathcal{S}$ with neutral connotation, but when discussing welfare or malfare, we often refer to $\mathcal{S}$ as *utility* or *risk*, respectively, as in these cases, $\mathcal{S}$ describes a well-understood pre-existing concept. Coarsely speaking, the three notions are identical, all being functions of the form $(\Omega \to \mathbb{R}_{0+}) \times \textsc{Measure}(\Omega, 1) \to \mathbb{R}_{0+}$, however we shall see that in order to promote fairness, the desirable axioms of malfare and welfare functions differ slightly. The notation reflects this; $\mathrm{M}(\mathcal{S}; \boldsymbol{w})$ is an M for *mean*, whereas $\mathrm{W}(\mathcal{S}; \boldsymbol{w})$ is a W for *welfare*, and $\mathrm{M\!\!\!\!\backslash}(\mathcal{S}; \boldsymbol{w})$ is an $\mathrm{M\!\!\!\!\backslash}$ (*inverted* W), to emphasize its inverted nature.

Often we are interested in *unweighted* population means, given sentiments as a *vector* $\mathcal{S} \in \mathbb{R}^{g}_{0+}$. Unweighted means may be defined in terms of weighted means, as

$$\mathrm{M}(\mathcal{S}) \doteq \mathrm{M}\left(i \mapsto \mathcal{S}_i; \{i \mapsto \tfrac{1}{g}\}\right) \enspace ,$$

abusing notation to concisely express the uniform measure. Indeed, it may seem antithetical to fairness to allow for weights in malfare and welfare definitions; consider however that weights can represent *differential population sizes*, and thus ensure that the welfare or malfare of *weight-preserving decompositions* of groups into subgroups with equal risk or utility remains constant.

Example 3.3.2 (Utilitarian Welfare)**.** Suppose individuals reside in some space $\mathcal{X}$, where *distributions* $\mathcal{D}_{1:g}$ over domain $\mathcal{X}$ describe the distribution over individuals in *each group*. Suppose also *utility function* $\mathrm{U}(x) : \mathcal{X} \to \mathbb{R}_{0+}$, describing the *level of satisfaction* of an individual, w.r.t., e.g., some *situation*, *allocation*, or *classifier*. We now take the *sentiment function* to be the *mean utility* (per-group), i.e.,

$$\mathcal{S}(\omega_i) \doteq \mathop{\mathbb{E}}_{x \sim \mathcal{D}_i}[\mathrm{U}(x)] = \mathop{\mathbb{E}}_{\mathcal{D}_i}[\mathrm{U}] \enspace .$$

Now, given a *weights vector* $\boldsymbol{w}$, describing the *relative frequencies* of membership in each of the g groups, we define the *utilitarian welfare* as

$$\mathrm{W}_1(\mathcal{S}; \boldsymbol{w}) \doteq \sum_{i=1}^{g} \boldsymbol{w}(\omega_i)\mathcal{S}(\omega_i) = \mathop{\mathbb{E}}_{\omega \sim \boldsymbol{w}}[\mathcal{S}(\omega)] = \mathop{\mathbb{E}}_{\boldsymbol{w}}[\mathcal{S}] \enspace .$$

Of course, in statistical, sampling, and machine learning contexts, $\mathcal{D}_{1:g}$ and $\boldsymbol{w}$ may be unknown, so we now discuss an *empirical analogue*. Section 3.4 is then devoted to showing how and when empirical population means well-approximate their true counterparts.

Example 3.3.3 (Empirical Utilitarian Welfare)**.** Now suppose $\mathcal{D}_{1:g}$ are unknown, but instead, we are given a *sample* $\boldsymbol{x}_{1:g,1:m} \in \mathcal{X}^{g \times m}$, where $\boldsymbol{x}_{i,1:m} \sim \mathcal{D}_i$. We define an *empirical analogue* of the utilitarian welfare as in example 3.3.2, instead taking

$$\hat{\mathcal{S}}(\omega_i) \doteq \mathop{\hat{\mathbb{E}}}_{x \in \boldsymbol{x}_i}[\mathrm{U}(x)] \quad \& \quad \hat{\mathrm{W}}_1(\hat{\mathcal{S}}, \boldsymbol{w}) \doteq \mathop{\mathbb{E}}_{\boldsymbol{w}}[\hat{\mathcal{S}}] \enspace .$$

Similarly, if $\boldsymbol{w}$ is unknown, but we may sample from some $\mathcal{D}$ over $\Omega \times \mathcal{X}$, we can use *empirical frequencies* $\hat{\boldsymbol{w}}$ in place of *true frequencies* $\boldsymbol{w}$, and define $\hat{\mathcal{S}}(\omega_i)$ as *conditional averages* over the subsample associated with group i.

3.3.1 Axioms of Cardinal Welfare and Malfare

Definition 3.3.4 (Axioms of Cardinal Welfare and Malfare)**.** We define the *population-mean axioms* for population-mean function $\mathrm{M}(\mathcal{S};\boldsymbol{w})$ below. For each item, assume (if necessary) that the axiom applies $\forall \mathcal{S}, \mathcal{S}' \in \Omega \to \mathbb{R}_{0+}$, scalars $\alpha, \beta \in \mathbb{R}_{0+}$, and probability measures $\boldsymbol{w}$ over Ω.

1. (Strict) Monotonicity: $\forall \boldsymbol{\varepsilon} : \Omega \to \mathbb{R}_{0+}$ s.t. $\int_{\boldsymbol{w}} \boldsymbol{\varepsilon}(\omega) \,\mathrm{d}(\omega) > 0$: $\mathrm{M}(\mathcal{S};\boldsymbol{w}) < \mathrm{M}(\mathcal{S} + \boldsymbol{\varepsilon};\boldsymbol{w})$.
2. Symmetry: $\forall$ permutations π over Ω: $\mathrm{M}(\mathcal{S};\boldsymbol{w}) = \mathrm{M}(\pi(\mathcal{S});\pi(\boldsymbol{w}))$.
3. Continuity: $\{\mathcal{S}' \mid \mathrm{M}(\mathcal{S}';\boldsymbol{w}) \le \mathrm{M}(\mathcal{S};\boldsymbol{w})\}$ and $\{\mathcal{S}' \mid \mathrm{M}(\mathcal{S}';\boldsymbol{w}) \ge \mathrm{M}(\mathcal{S};\boldsymbol{w})\}$ are closed sets.
4. Independence of unconcerned agents: Suppose subpopulation $\Omega' \subseteq \Omega$. Then $$\mathrm{M}\left(\begin{cases} p \in \Omega' : \alpha \\ p \notin \Omega' : \mathcal{S}(p) \end{cases};\boldsymbol{w}\right) \le \mathrm{M}\left(\begin{cases} p \in \Omega' : \alpha \\ p \notin \Omega' : \mathcal{S}'(p) \end{cases};\boldsymbol{w}\right) \Rightarrow \mathrm{M}\left(\begin{cases} p \in \Omega' : \beta \\ p \notin \Omega' : \mathcal{S}(p) \end{cases};\boldsymbol{w}\right) \le \mathrm{M}\left(\begin{cases} p \in \Omega' : \beta \\ p \notin \Omega' : \mathcal{S}'(p) \end{cases};\boldsymbol{w}\right) .$$
5. Independence of common scale: $\mathrm{M}(\mathcal{S};\boldsymbol{w}) \le \mathrm{M}(\mathcal{S}';\boldsymbol{w}) \Rightarrow \mathrm{M}(\alpha\mathcal{S};\boldsymbol{w}) \le \mathrm{M}(\alpha\mathcal{S}';\boldsymbol{w})$.
6. Multiplicative linearity: $\mathrm{M}(\alpha\mathcal{S};\boldsymbol{w}) = \alpha\mathrm{M}(\mathcal{S};\boldsymbol{w})$.
7. Unit scale: $\mathrm{M}(\mathbf{1};\boldsymbol{w}) = \mathrm{M}(1,\dots,1;\boldsymbol{w}) = 1$.
8. Pigou-Dalton transfer principle: Suppose $\mu = \mathbb{E}_{\boldsymbol{w}}[\mathcal{S}] = \mathbb{E}_{\boldsymbol{w}}[\mathcal{S}']$, and $\forall p \in \Omega : \left|\mu - \mathcal{S}'(p)\right| \le \left|\mu - \mathcal{S}(p)\right|$. Then $\mathrm{W}(\mathcal{S}';\boldsymbol{w}) \ge \mathrm{W}(\mathcal{S};\boldsymbol{w})$.
9. Anti-Pigou-Dalton transfer principle: Suppose as in 8, and conclude $\text{M}\!\!\text{M}(\mathcal{S}';\boldsymbol{w}) \le \text{M}\!\!\text{M}(\mathcal{S};\boldsymbol{w})$.

We take a moment to comment on each of these axioms, to preview their purpose and assure the reader of their necessity. Axioms 1-5 are the standard *axioms of cardinal welfarism* (1-4 are discussed by Sen [1977], Roberts [1980], and 5 by Debreu [1959], Gorman [1968]). Together, they imply (via the Debreu-Gorman theorem) that any population-mean can be decomposed as a *monotonic function* of a *sum* (over groups) of *log* or *power* functions. Axiom 6 is a natural and useful property, and ensures that *dimensional analysis* on mean functions is possible: in particular, the *units* of mean functions match those of sentiment. Note that axiom 6 implies axiom 4, and it is thus a simple strengthening of a traditional cardinal welfare axiom. This axiom also ensures that *units* of population means preserve the *units* of $\mathcal{S}$, making dimensional analysis greatly more convenient; we will also see that it is essential to show convenient *statistical* and *learnability* properties. Axiom 7 furthers this theme, as it ensures that not only do *units* of means match those of $\mathcal{S}$, but scale does as well (making comparisons like $\mathcal{S}_i$ is *above* the population-welfare meaningful), and also enabling comparison *across populations* (in the sense that comparing *averages* is more meaningful than *sums*). Finally, axiom 8 (the *Pigou-Dalton transfer principle* [see Pigou, 1912, Dalton, 1920]) is also standard in cardinal welfare theory as it ensures fairness, in the sense that welfare is higher when utility values are more uniform, i.e., incentivizing *equitable redistribution of "wealth"* in welfare. Its antithesis, axiom 9, encourages the opposite; in the context of welfare, this perversely incentivizes an expansion of inequality, but for malfare, which we generally wish to *minimize*, the opposite occurs, thus this axiom characterizes *fairness* in the context of *malfare*.

Axioms 6-4 are novel to this work, and are key in strengthening the Debreu-Gorman theorem to ensure that all welfare and malfare functions are *power means* in the sequel. Axiom 9 is also novel, as it is necessary to flip the inequality of axiom 8 when the sense of the population mean is inverted from welfare to malfare; in particular, the semantic meaning shifts from requiring that "redistribution of utility is desirable" to "redistribution of disutility is not undesirable."

For context, we present an additional axiom; that of *additive separability*. For simplicity, we present it only in the unweighted discrete case, as there is some subtlety to an equivalent measure-theoretic formulation, and we derive no benefit from assuming this axiom, as it is largely incompatible with our assumptions, and presented only for comparison purposes.

Definition 3.3.5 (Additive Separability)**.** Population-mean $\mathrm{M}(\mathcal{S}_{1:g})$ is *additively separable* if there exist functions $f_{1:g}$ s.t. each $f_i \in \mathbb{R}_{0+} \to \mathbb{R}_{0+}$, and $\mathrm{M}(\mathcal{S}_{1:g})$ may be decomposed as

$$\mathrm{M}(\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_g) = \sum_{i=1}^{g} f_i(\mathcal{S}_i) \enspace .$$

While seemingly quite important to early welfare theorists (e.g., the Debreu-Gorman theorem is generally presented in additively-separable form), and it gives rise to some convenient interpretability and computational properties, we argue that these are far-outstripped by those stemming from the *multiplicative linearity* and *unit scale* axioms, with no real difference in generality.[1] Furthermore, unlike these and other standard cardinal welfare axioms, additive separability seems a bit-heavy handed, assuming something very specific that is supposedly convenient for the economist, with little justification as to why and how it serves as a fundamental property of *cardinal welfare* itself. These properties are discussed in section 3.3.3, and directly compared with the additively-separable form in section 3.3.4.

[1]In this sense, the *additive welfare functionals* are somewhat like the *natural sufficient statistics* of the exponential family, in that, while additivity is sometimes convenient, it comes only with the sacrifice of other desiderata.

3.3.2 The Power Mean

We now define the *p-power mean* $\mathrm{M}_p(\cdot)$, for any $p \in \mathbb{R} \cup \pm\infty$, which we shall use to quantify both malfare and welfare. Power means arise often when analyzing population means obeying the various axioms of definition 3.3.4, and as we shall see in theorem 3.3.7, are a particularly important class of population means.

Definition 3.3.6 (Power-Mean Welfare and Malfare)**.** Suppose $p \in \mathbb{R} \cup \pm\infty$. We first define the *unweighted power mean* of sentiment vector $\mathcal{S} \in \mathbb{R}^g_{0+}$ as

$$\mathrm{M}_p(\mathcal{S}) \doteq \begin{cases} p \in \mathbb{R} \setminus \{0\} & \sqrt[p]{\frac{1}{g}\sum_{i=1}^{g} \mathcal{S}_i^p} \\ p = -\infty & \min_{i \in 1,\dots,g} \mathcal{S}_i \\ p = 0 & \sqrt[g]{\prod_{i=1}^{g} \mathcal{S}_i} = \exp\left(\frac{1}{g}\sum_{i=1}^{g} \ln(\mathcal{S}_i)\right) \\ p = \infty & \max_{i \in 1,\dots,g} \mathcal{S}_i \end{cases} .$$

We now define the *weighted power mean*, given *sentiment value function* $\mathcal{S} : \Omega \to \mathbb{R}_{0+}$ and probability measure $\boldsymbol{w}$ over Ω, as

$$\mathrm{M}_p(\mathcal{S}; \boldsymbol{w}) \doteq \begin{cases} p \in \mathbb{R} \setminus \{0\} & \sqrt[p]{\int_{\boldsymbol{w}} \mathcal{S}^p(\omega)\,\mathrm{d}(\omega)} = \sqrt[p]{\mathop{\mathbb{E}}_{\omega\sim\boldsymbol{w}}[\mathcal{S}^p(\omega)]} \\ p = -\infty & \min_{\omega \in \mathrm{Support}(\boldsymbol{w})} \mathcal{S}(\omega) \\ p = 0 & \exp\left(\int_{\boldsymbol{w}} \ln \mathcal{S}(\omega)\,\mathrm{d}(\omega)\right) = \exp\left(\mathop{\mathbb{E}}_{\omega\sim\boldsymbol{w}}[\ln \mathcal{S}(\omega)]\right) \\ p = \infty & \max_{\omega \in \mathrm{Support}(\boldsymbol{w})} \mathcal{S}(\omega) \end{cases} .$$

In both the weighted and unweighted cases, $p \in \{-\infty, 0, \infty\}$ resolve as their (unique) limits, and for all $p \in \mathbb{R}$, note that *power means* are special cases of the (weighted) *generalized mean*, defined for strictly monotonic f as $\mathrm{M}_f(\mathcal{S}; \boldsymbol{w}) \doteq f^{-1}\big(\mathbb{E}_{\omega\sim\boldsymbol{w}}[\ln \mathcal{S}(\omega)]\big)$.

Theorem 3.3.7 (Properties of the Power-Mean)**.** Suppose $\mathcal{S}, \boldsymbol{\varepsilon}$ are sentiment values $\Omega \to \mathbb{R}_{0+}$, and $\boldsymbol{w}$ is a probability measure over Ω. Then

1. Monotonicity: $\mathrm{M}_p(\mathcal{S}; \boldsymbol{w})$ is weakly-monotonically-increasing in p, and strictly if $\mathcal{S}$ attains distinct $a, b \in \mathbb{R}$ with nonnegligible probability.
2. Subadditivity: $\forall p \geq 1 :\ \mathrm{M}_p(\mathcal{S} + \boldsymbol{\varepsilon}; \boldsymbol{w}) \leq \mathrm{M}_p(\mathcal{S}; \boldsymbol{w}) + \mathrm{M}_p(\boldsymbol{\varepsilon}; \boldsymbol{w})$.

3. Contraction: $\forall p \geq 1 : \left|\mathrm{M}_p(\mathcal{S}; \boldsymbol{w}) - \mathrm{M}_p(\mathcal{S}'; \boldsymbol{w})\right| \leq \mathrm{M}_p(\left|\mathcal{S} - \mathcal{S}'\right|; \boldsymbol{w}) \leq \left\|\mathcal{S} - \mathcal{S}'\right\|_\infty$.

4. Curvature: $\mathrm{M}_p(\mathcal{S}; \boldsymbol{w})$ is concave in $\mathcal{S}$ for $p \in [-\infty, 1]$ and convex for $p \in [1, \infty]$.

3.3.3 Properties of Welfare and Malfare Functions

We now show that the axioms of definition 3.3.4 are sufficient to characterize many properties of welfare and malfare.

Theorem 3.3.8 (Population Mean Properties)**.** Suppose population-mean function $\mathrm{M}(\mathcal{S}; \boldsymbol{w})$. If $\mathrm{M}(\cdot; \cdot)$ satisfies (subsets of) the population-mean axioms (see definition 3.3.4), we have that $\mathrm{M}(\cdot; \cdot)$ exhibits the following properties. For each, assume arbitrary sentiment-value function $\mathcal{S} : \Omega \to \mathbb{R}_{0+}$ and weights measure $\boldsymbol{w}$ over Ω. The following then hold.

1. *Identity*: Axioms 6 & 7 imply $\mathrm{M}(\omega \mapsto \alpha; \boldsymbol{w}) = \alpha$.

2. Axioms 1-5 imply $\exists p \in \mathbb{R}$, strictly-monotonically-increasing continuous $F : \mathbb{R} \to \mathbb{R}_{0+}$ s.t.

$$\mathrm{M}(\mathcal{S}; \boldsymbol{w}) = F\left(\int_{\boldsymbol{w}} f_p(\mathcal{S}(\omega)) \, \mathrm{d}(\omega)\right) = F\left(\mathop{\mathbb{E}}_{\omega \sim \boldsymbol{w}}\left[f_p(\mathcal{S}(\omega))\right]\right) \enspace , \enspace \text{with} \begin{cases} p = 0 & f_0(x) \doteq \ln(x) \\ p \neq 0 & f_p(x) \doteq \mathrm{sgn}(p) x^p \end{cases} .$$

3. Axioms 1-7 imply $F(x) = f_p^{-1}(x)$, thus $\mathrm{M}(\mathcal{S}; \boldsymbol{w}) = \mathrm{M}_p(\mathcal{S}; \boldsymbol{w})$.

4. Axioms 1-5 and 8 imply $p \in (-\infty, 1]$.

5. Axioms 1-5 and 9 imply $p \in [1, \infty)$.

Taken together, the items of theorem 3.3.8 tell us that the mild conditions of axioms 1-4 (generally assumed for welfare), along with *multiplicative linearity*, imply that welfare and utility, or malfare and loss, are measured in the *same units* (e.g., *nats* or *bits* for *cross-entropy loss*, square-$\mathcal{Y}$-units for *square error*, or *dollars* for *income utility*), and power-mean malfare is effectively the only reasonable choice of welfare or malfare function. Even without multiplicative linearity, axioms 1-5 imply population mean functions are still *monotonic transformations* of power-mean. Furthermore, the entirely milquetoast *unit scale* axiom implies that sentiment values and population means have the same *scale*, making comparisons like "the risk of group i is above (or below) the population malfare" meaningful. Finally, we also have that $p \in [-\infty, 1)$ incentivizes redistribution of utility from better-off groups to worse-off groups, and similarly $p \in [1, \infty)$ incentivizes redistribution of harm[2] from worse-off groups to better-off groups.

[2]Note that, mathematically speaking, it is entirely *valid* to quantify welfare with $p > 1$ or malfare with $p < 1$, and indeed such characterizations may arise in the analysis of unfair systems; however we generally advocate against intentionally creating such unfair systems.

3.3.4 A Comparison with the Additively Separable Form

In particular, assuming axioms 1-5, all population means are *monotonic transformations* of the power mean, thus whether we assume *additive separability*, and get the additively-separable form

$$\mathrm{M}(\mathcal{S}) = c\sum_{i=1}^{g} f_p(\mathcal{S}_i) = cg\mathrm{M}_p^p(\mathcal{S}) \ ,$$

for $p \in \mathbb{R} \setminus \{0\}$ (where usually $c = 1$ is taken as the canonical form), or we assume axioms 6 & 7, and get the power-mean, there is no real loss of descriptiveness, as all such population-means remain *isomorphic* under the binary comparison operator ($\leq$). With additive separability, it is straightforward to compute the welfare of a population from the welfares of *subpopulations*, but it is still computationally trivial to do this with power means. Furthermore, the limiting cases of $p \in \pm\infty$ become undefined in the additively separable form, and we lose monotonicity in p (see theorem 3.3.7), both of which are remedied with power means.

With additive separability, welfare summarizes population sentiment by intuitively generalizing the idea of *summation*. In contrast, in our setting, we prefer to think of welfare as a generalized *average*. This yields desirable statistical estimation and learnability properties (shown in the sequel), but is also useful in and of itself, as it allows us to, for instance compare individual sentiment values to welfare, as both the *units* and *scale* match.

Another reason to prefer the power-mean over the additively separable form is the potential for direct comparisons between *group sentiments*, population means of *subgroups*, and *overall population means*. In particular, the *dimensional analysis* properties of the *power mean* are convenient, as these comparisons agree in *units* (due to axiom 6) and *scale* (due to axiom 7). No such dimensional analysis is possible with the additively separable form, as, e.g., if $\mathcal{S}$ is measured in dollars, then $\mathrm{W}_2^2(\mathcal{S};\boldsymbol{w})$ is measured in *square dollars* (a rather unintuitive unit).

3.3.5 Relating Power Means and Inequality Indices

We now discuss and define *relative inequality indices* $\mathrm{I}(\mathcal{S};\boldsymbol{w})$, which have been employed in the literature [Sen et al., 1997] to construct *welfare functions* of the form

$$\mathrm{W}(\mathcal{S};\boldsymbol{w}) = \mathrm{W}_1(\mathcal{S};\boldsymbol{w})\big(1 - \mathrm{I}(\mathcal{S};\boldsymbol{w})\big) \ .$$

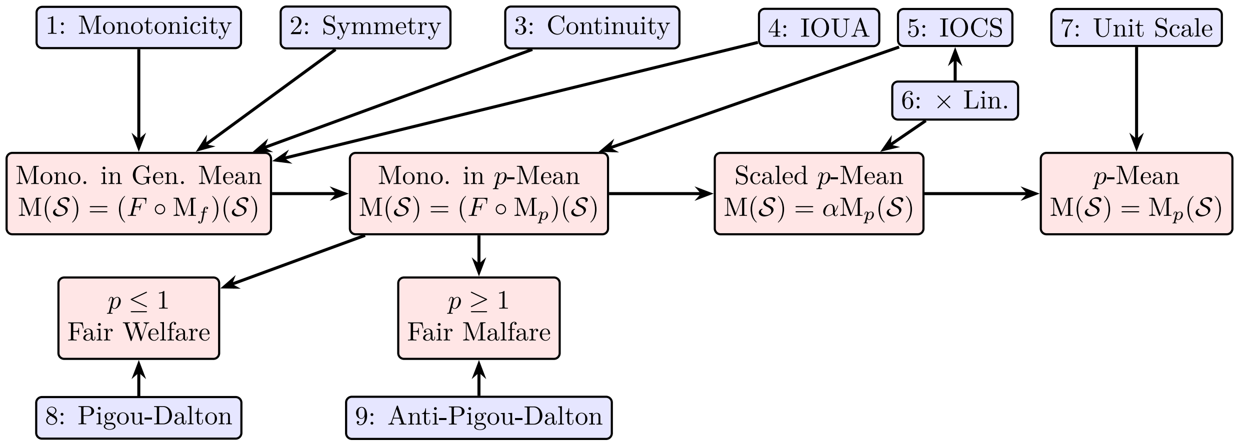


Figure 3.1: Relationships between population-mean axioms and properties. Assumptions and axioms shown in pastel blue, and properties shown in pastel red. Equivalent properties hold for weighted population mean functions.

This characterization intuitively starts with the *utilitarian welfare*, which measures *overall satisfaction* and then *downweights* based on how unfairly distributed utility is amongst the population. The "relative" in relative inequality indices connotes the fact that they are restricted to domain $[0, 1]$, thus the welfare metric matches the utilitarian under no inequality, and is 0 under maximal inequality.

We show that a large class of such functions are actually power means, which both gives them axiomatic justification, and shows prior support in the literature for the power mean. In particular, we first consider the *Atkinson index* relative inequality measure family [Atkinson et al., 1970].

Definition 3.3.9 (Atkinson Index)**.** For all $\varepsilon \in \mathbb{R}$, we define the Atkinson index as

$$\mathrm{Atk}_\varepsilon(\mathcal{S}; \boldsymbol{w}) \doteq 1 - \frac{\mathrm{M}_{1-\varepsilon}(\mathcal{S}; \boldsymbol{w})}{\mathrm{M}_1(\mathcal{S}; \boldsymbol{w})} \enspace .$$

Note that often the Atkinson index is restricted to $\varepsilon \in [0, 1]$; outside this range, it may exceed 1. Furthermore, the Atkinson index is generally stated without weights, and in a mathematically equivalent form, in which the resemblance to the power mean is less obvious, but for our purposes the above form is clearer. From it, we immediately have the following lemma.

Lemma 3.3.10 (Relating Atkinson's Indices and Power Means)**.** Suppose some $\varepsilon \in \mathbb{R}$, and take $p = 1 - \varepsilon$. It then holds that

$$\mathrm{M}_p(\mathcal{S}; \boldsymbol{w}) = \mathrm{M}_1(\mathcal{S}; \boldsymbol{w})(1 - \mathrm{Atk}_\varepsilon(\mathcal{S}; \boldsymbol{w})) \enspace .$$

Proof. This is a direct consequence of definition 3.3.9, noting $p = 1 - \varepsilon \Leftrightarrow \varepsilon = 1 - p$. $\square$

This is not particularly surprising in light of the welfare-centric derivation of [Atkinson et al., 1970], but nonetheless it yields a valuable alternative way to think about power means and inequality-weighted welfare functions. In particular, it gives a direct *axiomatic justification* of the welfare function $\mathrm{W}(\mathcal{S};\boldsymbol{w}) = \mathrm{W}_1(\mathcal{S};\boldsymbol{w})(1 - \mathrm{Atk}_\varepsilon(\mathcal{S};\boldsymbol{w}))$ (see theorem 3.3.8), and also gives an alternative intuitive interpretation of power-mean welfare (as inequality-weighted utilitarian welfare).

Note that similar properties may be shown for the isomorphic inequality measures of *generalized entropy indices* and *Theil indices* [Theil, 1967], though in this context their forms are generally less pleasing.

3.4 Statistical Estimation of Malfare Values

We first illustrate the ease with which p-power means can be estimated, in contrast to the standard additive welfare formulations. Perhaps surprisingly, we find that empirical welfare and malfare are *biased estimators*, yet they admit much sharper finite-sample tail bounds than the additive welfare formulations, which are unbiased.

Lemma 3.4.1 (Statistical Estimation)**.** Suppose probability distribution $\mathcal{D}$, population-mean M obeying *monotonicity*, sentiment value function $\mathcal{S}$ such that, given functions f_ω, we have $\mathcal{S}(\omega) = \mathbb{E}_{x\sim\mathcal{D}}[f_\omega(x)]$, sample $\boldsymbol{x} \sim \mathcal{D}^m$, and *empirical sentiment value estimate* $\hat{\mathcal{S}} \doteq \hat{\mathbb{E}}_{x\in\boldsymbol{x}}[f_\omega(x)]$. If it holds with probability at least $1-\delta$ that $\forall\omega:\ \mathcal{S}'(\omega) - \boldsymbol{\varepsilon}(\omega) \leq \mathcal{S}(\omega) \leq \mathcal{S}'(\omega) + \boldsymbol{\varepsilon}(\omega)$, then with said probability, we have

$$\mathrm{M}_p(\mathbf{0} \vee (\hat{\mathcal{S}} - \boldsymbol{\varepsilon});\boldsymbol{w}) \leq \mathrm{M}_p(\hat{\mathcal{S}};\boldsymbol{w}) \leq \mathrm{M}_p(\hat{\mathcal{S}} + \boldsymbol{\varepsilon};\boldsymbol{w}) \enspace ,$$

where $\boldsymbol{a} \vee \boldsymbol{b}$ denotes the (elementwise) minimum.

Proof. This result follows from the assumption, and the *monotonicity* axiom (i.e., adding or subtracting $\boldsymbol{\varepsilon}$ can not decrease or increase the power mean, respectively). The minimum with 0 on the LHS is required simply because by definition, sentiment values are nonnegative, and M_p is in general undefined with negative inputs. □

We now reify this result, applying the well-known Hoeffding [1963] and Bennett [1962] bounds to show concentration and derive an explicit form for $\boldsymbol{\varepsilon}$.

Corollary 3.4.2 (Statistical Estimation with Hoeffding and Bennett Bounds)**.** Suppose fair power-mean malfare $\text{M}(\cdot;\cdot)$ (i.e., $p \geq 1$), *loss function* $\ell : \mathcal{X} \to [0,r]$, $\mathcal{S} \in [0,r]^g$ s.t. $\mathcal{S}_i = \mathbb{E}_{\mathcal{D}_i}[\ell]$, samples $\boldsymbol{x}_i \sim \mathcal{D}_i^m$, and take $\hat{\mathcal{S}}_i \doteq \frac{1}{m}\sum_{j=1}^{m} \ell(\boldsymbol{x}_{i,j})$. Then with probability at least $1-\delta$ over choice of $\boldsymbol{x}$,

$$\left|\mathrm{M}_p(\mathcal{S};\boldsymbol{w}) - \mathrm{M}_p(\hat{\mathcal{S}};\boldsymbol{w})\right| \leq r\sqrt{\frac{\ln\frac{2g}{\delta}}{2m}} \enspace .$$

Alternatively, again with probability at least $1-\delta$ over choice of $\boldsymbol{x}$, we have

$$\left|\mathrm{M}_p(\mathcal{S};\boldsymbol{w}) - \mathrm{M}_p(\hat{\mathcal{S}};\boldsymbol{w})\right| \leq \frac{r\ln\frac{2g}{\delta}}{3m} + \max_{i\in 1,\dots,g}\sqrt{\frac{2\,\mathbb{V}_{\mathcal{D}_i}[\ell]\ln\frac{2g}{\delta}}{m}} \enspace .$$

As corollary 3.4.2 follows directly from lemma 3.4.1, with Hoeffding and Bennett inequalities applied to derive $\boldsymbol{\varepsilon}$ bounds, similar results are immediately possible with arbitrary concentration inequalities. In particular, similar *data-dependent* bounds may be shown, e.g., with *empirical Bennett bounds*, removing dependence on *a priori known variance*. In particular, we may apply theorem 1.3.3 to bound $\boldsymbol{\varepsilon}$, either elementwise (over groups) with a *union bound*, or jointly, if the loss functions for each group may be combined into a single function family.

Furthermore, note that while these bounds may be used for *evaluating* the welfare or malfare of a *particular* classifier or mechanism (through $\mathcal{S}$ and $\hat{\mathcal{S}}$), they immediately extend to *learning* over a *finite family* via the union bound. Much like in standard statistical learning theory, the exponential tail bounds of corollary 3.4.2 allow the family to grow *exponentially*, at *linear* cost to sample complexity. Also as in standard uniform convergence analysis (generally discussed in the context of *empirical risk minimization*), we can easily handle infinite hypothesis classes and obtain much sharper bounds by considering *uniform convergence bounds* over the family, e.g., with Rademacher averages and appropriate concentration-of-measure bounds.

In learning contexts, minimizing the *malfare* among all groups immediately generalizes minimizing *risk* of a single group. These statistical estimation bounds immediately imply that the *empirical malfare-optimal* solution is a reasonable proxy for the true malfare-optimal solution, as we now formalize.

Definition 3.4.3 (The Empirical Malfare Minimization (EMM) Principle)**.** Suppose hypothesis class $\mathcal{H} \subseteq \mathcal{X} \to \mathcal{Y}$, training samples $\boldsymbol{z}_{1:g}$ drawn from distributions $\mathcal{D}_{1:g}$ over $\mathcal{X}\times\mathcal{Y}$, loss function $\ell : \mathcal{Y}\times\mathcal{Y} \to \mathbb{R}_{0+}$, malfare function M, group weights $\boldsymbol{w}$. The *empirical malfare minimizer* is then defined as

$$\hat{h} \doteq \operatorname*{argmin}_{h\in\mathcal{H}} \mathrm{M}\left(i \mapsto \hat{\mathrm{R}}(h;\ell,\boldsymbol{z}_i);\boldsymbol{w}\right) \enspace ,$$

and the EMM principle states that $\hat{h}$ is a reasonable proxy for the true malfare minimizer

$$h^* \doteq \operatorname*{argmin}_{h^* \in \mathcal{H}} \mathrm{M}\left(i \mapsto \mathrm{R}(h^*; \ell, \mathcal{D}_i); \boldsymbol{w}\right) \enspace .$$

In all cases, these bounds provide further justification for using the *power-mean* over the *additively separable* form, as it is substantially harder to achieve comparable bounds on $\mathrm{M}_p^p(\mathcal{S}; \boldsymbol{w})$, due to the increased difficulty of controlling the *range* and *Lipschitz constant* of these quantities. Of course, as power-mean bounds imply bounds on the additively-separable form (and vice-versa), we recommend working with power means, and then converting back to the additively separable form (if so desired).

Chapter 4

Ewo$k$s: an Algorithm for Fair Codec Selection

In this chapter, I frame the problem of selecting *sets of streaming media codecs* as a *supervised learning task*, which I address with the *Empirical Welfare-Optimal k codec Selection* (EWOkS) algorithm. I consider multivariate optimality concepts (over, e.g., *audio quality* versus *bandwidth consumption*), such as *Pareto optimality*, *minimax-optimality*, and various *welfare* concepts. Welfare solutions perform well *in aggregate* over *sets* or *distributions* of users with *varied preferences*, quantifying and providing fine-grained control over *algorithmic fairness* and *bias* issues. I prove that EWOkS-solutions are approximately welfare-optimal over the underlying media distribution via rigorous statistical learning guarantees, and show empirically that EWOkS solutions are fairer (as measured by welfare) than single-objective solutions. This chapter is based on joint work with Clayton Sanford and Eli Upfal.

4.1 Introduction

Streaming video now accounts for an estimated 58% of all internet traffic [Sandvine, 2018], thus there is strong user demand and economic incentive to optimize stream quality and resource utilization. Most streamed audio and video data is *compressed* with a *codec* (coder-decoder), with complicated, subjective, multivariate (often conflicting) objectives; for example *resource utilization metrics*, (e.g., *compression ratio*, *encode/decode* CPU time), *fidelity / distortion (quality) metrics*, and *fault tolerance*, are all important, to varying degrees for various users, applications, and media types. Furthermore, while codecs are generally developed and tested on particular use cases (e.g., lossless audio, low-bandwidth speech data, various qualities of music data), and relative performance on each attribute varies depending on the media being encoded.

While many codecs purport efficacy for various use cases, little effort has been put into rigorous multivariate comparative statistical analysis. In practice, sets of streaming media codecs are generally selected *ad-hoc* by domain experts to target particular bitrates or quality thresholds, which may be suboptimal for users with different preferences, or over different media distributions. For example, low-quality `mp3` encodings may be used for a low-bitrate music stream, and lossless `FLAC` encodings for a high-quality stream, but for a speech stream, there exist specialized codecs that will achieve higher quality than a general-purpose codec at a given compression ratio. Furthermore, relative codec performance across media is inconsistent, and usually selecting the best encoding among a set of codecs outperforms any individual codec *on average*.

The multivariate aspect of the codec-selection problem induces both deep theoretical quandries and impactful real-world fairness and algorithmic bias issues. New technologies often fail to adequately serve global users, resulting in products with varying degrees of success across markets, harming both potential users and developers. We model the preferences of a user with a *loss function* ℓ mapping *codec attributes* to *user dissatisfaction*, and model *user diversity* by considering a *set* or *distribution* $\mathcal{L}$ over loss functions.

We address all of these issues with a novel *fair media streaming strategy*, employing a *codec set* $\boldsymbol{c}$, s.t. when a user with loss function ℓ requests media x, we return the encoding $c(x)$ for $c \in \boldsymbol{c}$ that minimizes $(\ell \circ c)(x)$, thus optimizing for *data-dependent user preference*. We then propose the *Empirical Welfare-Optimal k codec Selection* (Ewo*k*s) algorithm, which *learns* a data-dependent set of k codecs $\boldsymbol{c}$ that is *welfare-optimal* over $\mathcal{L}$. Data-dependence is key, as Ewo*k*s considers each codec in an unbiased way, allowing empirical performance to dictate whether it is selected, which automates the process of selecting domain-appropriate codecs, eliminating bias due to distribution-shift.

While various *welfare concepts* characterize different *notions of fairness*, Ewo*k*s is not tied to any particular welfare function. Rather, we show how to optimize and bound generalization error across a *broad class of welfare concepts*, where both *utilitarian* and *egalitarian* welfare arise as special cases.

It may be computationally intractable to optimize welfare over large datasets, and one may wish to approximate with a *sample* [see e.g. Riondato and Upfal, 2015, 2016, 2018], and in dynamic settings, new media are continuously uploaded, and thus are unavailable during training. We show that Ewo*k*s solutions are not only *empirically optimal*, but also *approximately optimal* over the *underlying media distribution* of the training data, thus mitigating overfitting concerns, by posing codec selection as a *statistical learning problem* and showing data-dependent uniform-convergence bounds.

We show novel finite-sample data-dependent welfare-generalization bounds with Monte-Carlo Rademacher averages [Koltchinskii, 2001, Bartlett and Mendelson, 2002] and modern entropy-method concentration inequalities [Boucheron et al., 2003, 2013], which outperform standard Martingale-based Rademacher bounds [Mitzenmacher and Upfal, 2017], particularly in the small-sample setting. Our bounds are sensitive to the *empirical variances* of codec performances, and can substantially improve upon the generalization guarantees of variance-insensitive methods. Such welfare-generalization guarantees are particularly important, as unlike with risk and loss functions, *empirical welfare* is *not* an unbiased estimator of *welfare*.

We now summarize our contributions:

1. We introduce k *codec selection*, reframing traditionally manual codec selection as a *supervised learning problem*, and propose the EWOkS algorithm it.
2. We control for *algorithmic bias* and *fairness* issues by learning *distribution-dependent* models and robustly optimizing over a broad class of *welfare concepts*.
3. We show welfare generalization bounds through *uniform convergence* theory and novel sharp *variance-sensitive* tail bounds.
4. We validate experimentally that EWOkS robustly controls for overfitting and unfairness due to optimizing for specific user models or data distributions.

4.2 The Ewo$k$s Algorithm

In the codec selection framework, we assume media exist in some space $\mathcal{X}$, and a family $\mathcal{C}$ of *codecs*, which act as functions from $\mathcal{X}$ onto a compressed $\mathcal{X}'$ and back, which we abstract to functions from $\mathcal{X}$ onto the space of *attribute values* $\mathbb{R}^a_{0+}$. Attributes represent various *objectively measurable* properties such as *quality* and *resource utilization* metrics. For example, `mp3`, with fixed encoder hyperparameters, is a codec, and the quality / distortion metrics and compression ratio are *attributes* of an encoding. We generally assume monotonically minimizing each objective is desirable (i.e., decreasing some attribute while others remain constant is never bad), though often this may be relaxed.

We also assume a family $\mathcal{L} \subseteq \mathbb{R}^a_{0+} \to \mathbb{R}_{0+}$ of *loss functions*, such that each loss $\ell \in \mathcal{L}$, codec $c \in \mathcal{C}$, media sample $x \in \mathcal{X}$, $(\ell \circ c)(x)$ quantifies the *subjective dissatisfaction* of ℓ on encoding $c(x)$. Therefore $\mathcal{L}$ represents the *space of user preferences* for various options and tradeoffs, for instance for video *framerate*, *quality*, *resolution*, and *compression ratio*. In other words, the *family $\mathcal{L}$ objectively captures* the *subjective desiderata* of users, each represented by some $\ell \in \mathcal{L}$. While each $\ell \in \mathcal{L}$ is an objective metric, by considering all of $\mathcal{L}$

simultaneously, we remain sensitive to the *subjective nature* of the problem.

Given a set $\boldsymbol{c}$ of k codecs, when a user with some loss function ℓ requests media $x \in \mathcal{X}$, we serve the request by selecting the $c \in \boldsymbol{c}$ that *minimizes* $(\ell \circ c)(x)$, and returning $(c, c(x))$. Consequently, it's generally sufficient to select a codec set such that *on average* (over the media distribution), *at least one* codec is satisfactory *for each user*. We quantify the aversion of a user with loss function ℓ to codec set $\boldsymbol{c}$ over media distribution $\mathcal{D}$ with the *risk*

$$\mathrm{R}(\boldsymbol{c}; \ell, \mathcal{D}) \doteq \mathop{\mathbb{E}}_{x \sim \mathcal{D}}\left[\min_{c \in \boldsymbol{c}} (\ell \circ c)(x)\right] \ . \tag{26}$$

The brief explanation above, and the optimality concepts discussed in the sequel, are sufficient to understand and analyze EWOkS.

4.2.1 Welfare-Optimal k codec Selection

Given a large codec family $\mathcal{C}$, the computation and storage costs of serving each query with the optimal $c \in \mathcal{C}$ become prohibitive. Depending on the data distribution $\mathcal{D}$ and user needs (represented by $\mathcal{L}$), many codecs may contribute little marginal benefit, hence a subset of $\mathcal{C}$ is often sufficient. For a single loss function ℓ, ℓ-optimality encourages selecting a *diverse codec set* $\boldsymbol{c}$, as there is little marginal benefit to adding a high-performing codec c to $\boldsymbol{c}$ if its performance is highly correlated with some $c' \in \boldsymbol{c}$ (as may occur with *similar encoding algorithms* or *parameters*). Furthermore, in general, EWOkS considers a *loss family* $\mathcal{L}$ *in aggregate*, and welfare-objectives incentivize selecting a *diverse codec set* that performs well for all $\ell \in \mathcal{L}$. In short, our goal is to select some $\boldsymbol{c} \subseteq \mathcal{C}$ from among the set of size-k *subsets* of $\mathcal{C}$, henceforth written $\binom{\mathcal{C}}{k}$, that performs well in aggregate over $\mathcal{L}$ w.r.t. $\mathcal{D}$.

With a specific user or objective use-case, represented by loss function ℓ, *risk* is the natural way to quantify performance. We then define the *risk minimizer* $\boldsymbol{c}^*_\ell$ as

$$\boldsymbol{c}^*_\ell \doteq \mathop{\mathrm{argmin}}_{\boldsymbol{c} \in \binom{\mathcal{C}}{k}} \mathrm{R}(\boldsymbol{c}; \ell, \mathcal{D}) \ . \tag{27}$$

With a *family* or *distribution* $\mathcal{L}$ over loss functions, we face tradeoffs as to the degree to which each $\ell \in \mathcal{L}$ may be satisfied. One approach to algorithmic fairness and robustness against arbitrary use-cases is to select a set of codecs $\boldsymbol{c}$ that optimizes worst-case performance (minimax-optimality) over $\mathcal{L}$. This corresponds to the *egalitarian-welfare optimal* solution concept[1] over $\mathcal{L}$. Welfare concepts capture various aspects of multivariate

[1]Although welfare is usually defined in terms of (desirable) *utility* $u(\cdot)$, we instead consider (undesirable) *loss* $\ell(\cdot)$, and thus in some sense seek to *minimize* a notion of *antiwelfare* over loss. Our methods can be reposed to work in either direction, e.g., by defining utility $u(\boldsymbol{a}) = 1 - \ell(\boldsymbol{a})$.

optimality in multi-agent systems [Gibbons, 1992], and egalitarian welfare ensures fairness, because it always incentivizes selecting $\boldsymbol{c}$ to benefit the *least-satisfied* users. Alternatively, *utilitarian welfare* considers the *mean loss* across a *distribution* over users, which is fair in a different sense. Utilitarian welfare will not ignore the desires of large groups in favor of improving small worst-case groups, though in ML systems it may create positive feedback loops, magnifying preexisting inequity. Rather than argue for a particular welfare concept, a deep philosophical choice which depends on many extrinsic factors, we show that our statistical methods are sufficient to control for generalization error *within a broad class* of welfare concepts.

In general, take *welfare concept* $\mathrm{M}(\boldsymbol{c}; \mathcal{L}, \mathcal{D})$ to be a function that takes a *codec set* $\boldsymbol{c}$, a *distribution over loss functions* $\mathcal{L}$, and a *media distribution* $\mathcal{D}$ over $\mathcal{X}$, and yields a real-valued measure of overall well-being. In this section, we show all results in full generality, but for intuition, recall the *power mean* definition 3.3.6. We also refer to both *welfare* and *malfare* as *welfare*, as whether we wish to minimize or maximize depends on whether codec attributes are positive or negative (e.g., *positive quality* or *negative distortion*). Recall that $\mathrm{M}_p(\dots)$ is *monotonic* in p, 1 & ∞ correspond to *utilitarian* and *egalitarian* welfare, respectively, and intermediate p are of interest as *compromises thereof* (e.g., the $p = 2$ *quadratic welfare* lies between *egalitarian* and *utilitarian* welfare, and weighs high-risk users more heavily than low-risk groups). We require only that $\mathrm{M}(\cdot)$ is λ-*Lipschitz continuous* in the ℓ_∞-norm of $\ell \mapsto \mathbb{E}_{x \sim \mathcal{D}}[\min_{c \in \boldsymbol{c}}(\ell \circ c)(x)]$ (i.e., an ε-perturbation to *all* risk values changes $\mathrm{M}(\cdot)$ by $\leq \lambda\varepsilon$). We also assume that $\mathrm{M}(\cdot)$ is *monotonic* in the risk of each user, making *Pareto-optimality* a meaningful optimality concept.

Given *loss function family* or *distribution* $\mathcal{L}$, media distribution $\mathcal{D}$ over $\mathcal{X}$, and welfare concept $\mathrm{M}(\cdot)$, we now quantify the performance of some k codec selection $\boldsymbol{c} \in \binom{\mathcal{C}}{k}$ as its *welfare* w.r.t. $\mathrm{M}(\cdot)$. The M-optimal codec set $\boldsymbol{c}^*_{\mathrm{M}}$ is then defined as

$$\boldsymbol{c}^*_{\mathrm{M}} \doteq \operatorname*{argmin}_{\boldsymbol{c} \in \binom{\mathcal{C}}{k}} \mathrm{M}(\boldsymbol{c}; \mathcal{L}, \mathcal{D}) \enspace . \tag{28}$$

4.2.2 Empirical Welfare Optimization

We now consider the task of *approximating* the risk-optimal $\boldsymbol{c}^*_\ell$ or welfare-optimal $\boldsymbol{c}^*_{\mathrm{M}}$, given only a *sample* $\boldsymbol{x} \sim \mathcal{D}^m$. We define the *empirical risk* $\hat{\mathrm{R}}(\boldsymbol{c}; \ell, \boldsymbol{x})$ as the *sample average loss*

$$\hat{\mathrm{R}}(\boldsymbol{c}; \ell, \boldsymbol{x}) \doteq \hat{\mathbb{E}}_{x \in \boldsymbol{x}} \left[\min_{c \in \boldsymbol{c}} (\ell \circ c)(x) \right] \enspace , \tag{29}$$

and similarly, define the *empirical risk minimizer* $\hat{\boldsymbol{c}}_\ell$ as

$$\hat{\boldsymbol{c}}_\ell \doteq \operatorname*{argmin}_{\boldsymbol{c} \in \binom{\mathcal{C}}{k}} \hat{\mathrm{R}}(\boldsymbol{c}; \ell, \boldsymbol{x}) \enspace , \tag{30}$$

corresponding to eqs. 26 and 27, respectively.

Similarly, for welfare $\mathrm{M}(\cdot)$, we take the *empirical welfare* $\hat{\mathrm{M}}(\boldsymbol{c};\mathcal{L},\boldsymbol{x})$ to be welfare computed with *empirical risk* $\hat{\mathrm{R}}(\boldsymbol{c};\ell,\boldsymbol{x})$ instead of *risk* $\mathrm{R}(\boldsymbol{c};\ell,\mathcal{D})$, and define the *empirical welfare minimizer*, by analogy with equation 28, as

$$\hat{\boldsymbol{c}}_{\mathrm{M}} \doteq \operatorname*{argmin}_{\boldsymbol{c}\in\binom{\mathcal{C}}{k}} \hat{\mathrm{M}}(\boldsymbol{c};\mathcal{L},\boldsymbol{x}) \quad . \tag{31}$$

The *Empirical Welfare-Optimal k codec Selection* (EWO*k*S) algorithm computes $\hat{\boldsymbol{c}}_{\mathrm{M}}$ as a proxy for $\boldsymbol{c}^*_{\mathrm{M}}$. We assume the combinatorial optimization aspect of this task is tractable via enumeration, or by exploiting its inherent *submodularity* (diminishing returns in adding codecs), as usually the computational bottleneck is encoding each $\boldsymbol{x}_i$ with each $c \in \mathcal{C}$. Although EWO*k*S is straightforward, the difficulty arises in determining the *minimum sample size* that guarantees that the welfare (over $\mathcal{D}$) of $\hat{\boldsymbol{c}}_{\mathrm{M}}$ approaches that of $\boldsymbol{c}^*_{\mathrm{M}}$.

In the statistical parlance, the *empirical risk* is an *unbiased estimator* of *risk*, i.e., $\mathbb{E}[\hat{\mathrm{R}}(\boldsymbol{c};\ell,\boldsymbol{x})] = \mathrm{R}(\boldsymbol{c};\ell,\mathcal{D})$, and $\hat{\boldsymbol{c}}_\ell$ is an *M-estimator* [Huber, 1964]. The *Empirical Risk Minimization* (ERM) paradigm [Vapnik, 1992] states that $\hat{\boldsymbol{c}}_\ell$ is a reasonable proxy for $\boldsymbol{c}^*_\ell$, and much of *statistical learning theory* quantifies the finite-sample *selection bias* (overfitting) of $\hat{\boldsymbol{c}}_\ell$. Similarly, with insufficient data, $\hat{\boldsymbol{c}}_{\mathrm{M}}$ can overfit, in which case the *estimation error* $\mathrm{M}(\hat{\boldsymbol{c}}_{\mathrm{M}};\mathcal{L},\mathcal{D}) - \mathrm{M}(\boldsymbol{c}^*_{\mathrm{M}};\mathcal{L},\mathcal{D}) \gg 0$, and $\hat{\boldsymbol{c}}_{\mathrm{M}}$ is a poor proxy for $\boldsymbol{c}^*_{\mathrm{M}}$. In general, the *empirical welfare* is a *biased estimator* of *welfare* (i.e., $\mathbb{E}[\hat{\mathrm{M}}(\boldsymbol{c};\mathcal{L},\boldsymbol{x})] \neq \mathrm{M}(\boldsymbol{c},\mathcal{L},\mathcal{D})$), however in section 4.2.3 we show tail bounds on the welfare-estimation error, thus controlling for overfitting and providing methods to guarantee that EWO*k*S solutions are statistically significant.

Note that k is a fixed hyperparameter of the EWO*k*S algorithm, and should be selected to balance the tradeoff as k increases between the *cost* of *managing more codecs* and *increased overfitting* against empirical-welfare improvement. As model complexity increases in k, optimizing k to balance the bias-variance tradeoff is statistically-efficient with *structural risk minimization* [Koltchinskii, 2001].

4.2.3 Generalization Analysis

We now show exponential tail bounds on the *welfare optimality-gap* $\mathrm{M}(\hat{\boldsymbol{c}}_{\mathrm{M}};\mathcal{L},\mathcal{D}) - \mathrm{M}(\boldsymbol{c}^*_{\mathrm{M}};\mathcal{L},\mathcal{D})$. We bound the *supremum deviation* over all *risk values* with novel *variance-sensitive tail bounds* which we then use to bound the welfare optimality-gap. Here it is convenient to discuss a generic function family, so given codec family $\boldsymbol{c} \subseteq \mathcal{X} \to \mathbb{R}^a_{0+}$, loss family $\mathcal{L} \subseteq \mathbb{R}^a_{0+} \to \mathbb{R}_{0+}$, and $k \in \mathbb{N}$, we define

$$\mathcal{F} \doteq \min\circ\mathcal{L}\circ\binom{\mathcal{C}}{k} = \left\{ x \mapsto \min_{c\in\boldsymbol{c}} (\ell\circ c)(x) \,\middle|\, \boldsymbol{c}\in\binom{\mathcal{C}}{k},\ \ell\in\mathcal{L} \right\} \ .$$

$\mathcal{F}$ captures the k codec selection-specific details of the problem, thus we seek to show *uniform convergence* over $\mathcal{F}$, which we will use to bound the gap between *empirical* and *expected* welfare of Ewo$k$s solutions.

Definition 4.2.1 (Rademacher Averages)**.** Suppose *distribution* $\mathcal{D}$ over $\mathcal{X}$, *sample* $\boldsymbol{x} \in \mathcal{X}^m$, n *Monte-Carlo trials*, and Rademacher matrix $\boldsymbol{\sigma} \in (\pm 1)^{n \times m}$. We define

1. for fixed $\boldsymbol{\sigma}$ and $\boldsymbol{x}$, the *n-Monte-Carlo Empirical Rademacher Average* (n-MCERA):

$$\hat{\mathfrak{R}}^n_m(\mathcal{F}, \boldsymbol{x}, \boldsymbol{\sigma}) \doteq \frac{1}{n}\sum_{j=1}^{n} \sup_{f \in \mathcal{F}} \left| \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\sigma}_{j,i} f(\boldsymbol{x}_i) \right| \;\; ;$$

2. averaging over $\boldsymbol{\sigma} \sim$ i.i.d. Rademacher (uniform on ± 1), the *Empirical Rademacher Average* (ERA):

$$\hat{\mathfrak{R}}_m(\mathcal{F}, \boldsymbol{x}) \doteq \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\hat{\mathfrak{R}}^1_m(\mathcal{F}, \boldsymbol{x}, \boldsymbol{\sigma})\right] \;\; ; \;\; \&$$

3. averaging over $\boldsymbol{x} \sim \mathcal{D}^m$, the *Rademacher Avg.* (RA):

$$\mathfrak{R}_m(\mathcal{F}, \mathcal{D}) \doteq \mathop{\mathbb{E}}_{\boldsymbol{x}}\left[\hat{\mathfrak{R}}_m(\mathcal{F}, \boldsymbol{x})\right] \;\; .$$

The n-MCERA is a *Monte-Carlo* estimate (w.r.t. $\boldsymbol{\sigma}$) of the ERA, and the ERA is a *sample estimate* (w.r.t. $\boldsymbol{x}$) of the RA. Recall that through symmetrization, we have

$$\mathop{\mathbb{E}}_{\boldsymbol{x}}\Big[\underbrace{\sup_{f \in \mathcal{F}} \left| \mathop{\mathbb{E}}_{\mathcal{D}}[f] - \mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}}[f] \right|}_{\text{Supremum Deviation}} \Big] \leq 2\mathfrak{R}_m(\mathcal{F}, \mathcal{D}) \;\; , \tag{32}$$

where the *supremum deviation* (SD) *uniformly bounds* the gap between *empirical* and *expected* values of each $f \in \mathcal{F}$ *simultaneously.*

While RAs control the *expected* SD, and boundedness is sufficient to obtain sharp tail bounds on the convergence of n-MCERA and ERA to the RA, sharp bounds on the SD require also that the *variance* of each $f \in \mathcal{F}$ is controlled. We define the *wimpy variance* (see Boucheron et al. [2013, ch. 11]), and our novel estimate, the *empirical wimpy variance* as

$$v \doteq \sup_{f \in \mathcal{F}} \mathop{\mathbb{V}}_{\mathcal{D}}[f] \;\; \& \;\; \hat{v} \doteq \sup_{f \in \mathcal{F}} \frac{1}{m}\sum_{i=1}^{m}\left(f(\boldsymbol{x}_i) - \mathop{\hat{\mathbb{E}}}_{\mathcal{D}}[f]\right)^2 \;\; ,$$

respectively. We now show bounds that quantify the relationships between the SD, n-MCERA, v, and $\hat{v}$ in the following novel theorem, which enjoys the asymptotic improvements of *variance-sensitive bounds* [Bousquet, 2002] without assuming *a priori* variance knowledge. In particular, our bound excels in the small-sample setting for bounded functions of uniformly low variance, as occurs in the codec selection problem.

Theorem 4.2.2 (Variance-Sensitive Bounds)**.** Suppose *distribution* $\mathcal{D}$ over $\mathcal{X}$, *training sample* $\boldsymbol{x} \sim \mathcal{D}^m$, *Monte-Carlo trial count* n, i.i.d. Rademacher matrix $\boldsymbol{\sigma} \in (\pm 1)^{n \times m}$, codec family $\mathcal{C}$, loss family $\mathcal{L} \subseteq \mathbb{R}^a_{0+} \to [-r, r]$, λ-Lipschitz *welfare* $\mathrm{M}(\cdot)$, and codec selection size k. $\forall \delta \in (0,1)$, let $\mathcal{F} \doteq \min \circ \mathcal{L} \circ \binom{\mathcal{C}}{k}$, take $\hat{v}$ to be the *empirical wimpy variance* of $\mathcal{F}$ over $\boldsymbol{x}$ and $\hat{v}^{\mathsf{raw}}$ to be the *empirical raw wimpy variance*, and take

1. $\varepsilon_{\mathrm{ERA}} \doteq \frac{4r\frac{4}{\delta}}{3nm} + \sqrt{\frac{4\hat{v}^{\mathsf{raw}} \ln \frac{4}{\delta}}{nm}}$;
2. $\varepsilon_{\mathrm{RA}} \doteq \frac{2r \ln \frac{4}{\delta}}{m} + \sqrt{\frac{2r(\hat{\mathfrak{R}}^n_m(\mathcal{F}, \boldsymbol{x}, \boldsymbol{\sigma}) + \varepsilon_{\mathrm{ERA}}) \ln \frac{4}{\delta}}{m}}$;
3. $\widetilde{\mathfrak{R}} \doteq \hat{\mathfrak{R}}^n_m(\mathcal{F}, \boldsymbol{x}, \boldsymbol{\sigma}) + \varepsilon_{\mathrm{ERA}} + \varepsilon_{\mathrm{RA}}$;
4. $\tilde{v} \doteq \frac{m}{m-1}\hat{v} + \frac{4r^2 \ln \frac{4}{\delta}}{m} + \sqrt{\frac{2r^2 \hat{v} \ln \frac{4}{\delta}}{m-1}}$; &
5. $\varepsilon_{\mathrm{SD}} \doteq \frac{r \ln \frac{4}{\delta}}{3m} + \sqrt{\frac{2(\tilde{v} + 4r\widetilde{\mathfrak{R}}) \ln \frac{4}{\delta}}{m}}$.

It then holds with with pr. $\geq 1 - \delta$ over $\boldsymbol{x}, \boldsymbol{\sigma}$ that

1. $\left| \mathrm{M}(\hat{\boldsymbol{c}}_{\mathrm{M}}, \mathcal{L}, \mathcal{D}) - \hat{\mathrm{M}}(\hat{\boldsymbol{c}}_{\mathrm{M}}, \mathcal{L}, \boldsymbol{x}) \right| \leq 2\lambda\widetilde{\mathfrak{R}} + \lambda\varepsilon_{\mathrm{SD}}$ &
2. $\mathrm{M}(\hat{\boldsymbol{c}}_{\mathrm{M}}, \mathcal{L}, \mathcal{D}) - \mathrm{M}(\boldsymbol{c}^*_{\mathrm{M}}, \mathcal{L}, \mathcal{D}) \leq 4\lambda\widetilde{\mathfrak{R}} + 2\lambda\varepsilon_{\mathrm{SD}}$.

The details of the proof are rather cumbersome, but the key idea is that we take a union bound over 4 probabilistic upper-tail-bounds; in particular, w.h.p., $v \leq \tilde{v}$ (1 tail bound), and $\mathfrak{R}_m(\mathcal{F}, \mathcal{D}) \leq \widetilde{\mathfrak{R}}$ (2 tail bounds for n-MCERA $\to$ ERA $\to$ RA). The fourth tail bound is on the SD of $\mathcal{F}$, which by Lipschitz properties yields bounds on welfare-gaps. Note that if we assume $\lambda = 1$, $4\widetilde{\mathfrak{R}} + 2\varepsilon_{\mathrm{SD}}$ is

$$4\hat{\mathfrak{R}}^n_m(\mathcal{F}, \boldsymbol{x}, \boldsymbol{\sigma}) + \boldsymbol{\Theta}\left(\frac{r \ln\frac{1}{\delta}}{m} + \sqrt{\frac{(v + r\mathfrak{R}_m(\mathcal{F}, \mathcal{D})) \ln\frac{1}{\delta}}{m}} \right) . \tag{33}$$

With worst-case variance $v = {r^2}/{4}$, we recover the slow-decaying $\boldsymbol{\Theta}\sqrt{r^2 \ln\frac{1}{\delta}/m}$ McDiarmid terms of standard methods, however generally we get asymptotic *mixed-rate convergence*, where the $\boldsymbol{\Theta}(r \ln\frac{1}{\delta}/m)$ term depends on the scale r, but *decays quickly* in m, and the $\boldsymbol{\Theta}\sqrt{(v + r\mathfrak{R}_m(\mathcal{F}, \mathcal{D})) \ln\frac{1}{\delta}/m}$ term depends on $\sqrt{v}$ instead of r, but *decays slowly* in m. Note that the latter term simplifies to $\boldsymbol{\Theta}\sqrt{v \ln\frac{1}{\delta}/m}$ when we consider the limiting behavior as $m \to \infty$ and assume $\mathfrak{R}_m(\mathcal{F}, \mathcal{D})$ tends to 0, thus the essence of the bound is *fast-decay* as r/m plus *slow decay* as $\sqrt{v/m}$.

This dichotomy is key to understanding the bound's strong performance in the small sample setting: we see initial *rapid decay* while the r term is dominant, followed by *slow decay* as the $\sqrt{v}$ term comes to dominate.

Where the transition occurs is primarily dictated by the relative values of r and v, but for sufficiently small v, we see fast-decaying tail bounds and excellent performance in the small-sample regime (see section 4.3.3).

4.2.4 Linear Loss Families

For large $\mathcal{C}$ and k, it can become computationally intractable to compute the wimpy variance and n-MCERA, and in general they are rather opaque, due to the supremum over codec combinations and the minimum over selected codecs. We now show simple *compositional bounds* for *linear loss families* mapping *attribute vectors* $\boldsymbol{a} \in \mathbb{R}^a_{0+}$ onto $\mathbb{R}_{0+}$, taking

$$\mathcal{L}_p \doteq \{\, \ell(\boldsymbol{a}) \doteq \boldsymbol{w} \cdot \boldsymbol{a} \mid \boldsymbol{w} \in \mathbb{R}^a_{0+} \text{ s.t. } \|\boldsymbol{w}\|_p \leq 1 \,\} \tag{34}$$

for $p \geq 1$. To control the *range* of $\mathcal{L}_p$, we assume also a *bounded domain* $\mathcal{X}$, in particular we assume bounded q-norm for $q \doteq \frac{p-1}{p}$, or $q = 1$ for $p = \infty$, thus $\|\cdot\|_p$ and $\|\cdot\|_q$ are *dual norms* (i.e., by Hölder's inequality, $\big|\boldsymbol{w} \cdot c(x)\big| \leq \|\boldsymbol{w}\|_p \|c(x)\|_q$). We now show that both the variance and ERA of the (nonlinear) k codec selection problem can be bounded with their (linear) counterparts in 1 codec selection.

Theorem 4.2.3 (Bounds for Linear Loss Families)**.** Suppose p, q, s.t. $|p^{-1}| + |q^{-1}| = 1$, *codec family* $\mathcal{C}$, $\mathcal{L}_p$ as in (34), and assume $\sup\limits_{c \in \mathcal{C}, x \in \mathcal{X}} \|c(x)\|_q \leq s$. For all $k \in 1, \dots |\mathcal{C}|$, take $\mathcal{F}_k \doteq \min \circ \mathcal{L}_p \circ \binom{\mathcal{C}}{k}$, and $\hat{v}_k$ to be the *empirical wimpy variance* of $\mathcal{F}_k$ on *sample* $\boldsymbol{x}$. Taking $\hat{\mathbb{C}}_{\boldsymbol{x}}[c] \in \mathbb{R}^{a \times a}$ to denote the *empirical covariance matrix* of the attributes of codec c over $\boldsymbol{x}$, and $\|\cdot\|_{p \to q}$ to be the ℓ_p-ℓ_q *operator norm*, it holds that

$$\hat{v}_1 = \sup_{c \in \mathcal{C}} \sup_{\ell \in \mathcal{L}} \hat{\mathbb{V}}_{\boldsymbol{x}}[\ell \circ c] \leq \sup_{c \in \mathcal{C}} \big\| \hat{\mathbb{C}}_{\boldsymbol{x}}[c] \big\|_{q \to p} \ , \ \&$$

$$\hat{\mathfrak{R}}^n_m(\mathcal{F}_1, \boldsymbol{x}, \boldsymbol{\sigma}) \leq \frac{1}{n} \sum_{j=1}^{n} \sup_{c \in \mathcal{C}} \left\| \frac{1}{m} \sum_{i=1}^{m} \boldsymbol{\sigma}_{j,i} c(\boldsymbol{x}_i) \right\|_q \ .$$

Furthermore, for any $a, b \in \mathbb{N}$ s.t. $k = a + b$, we have

$$\hat{v}_k \leq \hat{v}_a + \hat{v}_b \leq k \hat{v}_1 \ , \ \&$$

$$\hat{\mathfrak{R}}_m(\mathcal{F}_k, \boldsymbol{x}) \leq \hat{\mathfrak{R}}_m(\mathcal{F}_a, \boldsymbol{x}) + \hat{\mathfrak{R}}_m(\mathcal{F}_b, \boldsymbol{x}) \leq k \hat{\mathfrak{R}}_m(\mathcal{F}_1, \boldsymbol{x}) \ .$$

Here we see that linearity of $\mathcal{L}_p$ is convenient, though it *does not* in general convert the EWO*k*S problem to a linear problem, due to the *minimum* over selected codecs. However, $\mathcal{F}_1$ is completely linear, thus for $k = 1$ it is easy to compute Rademacher averages and variances. With sufficient computational resources, $\hat{\mathfrak{R}}_m(\mathcal{F}_k, \boldsymbol{x})$ and $\hat{v}_k$ can be computed as precisely as required, and theorem 4.2.3 allows us to trade *statistical efficiency* for *computational efficiency* by truncating the computation at some selection size $j < k$, and then loosely bounding $\hat{\mathfrak{R}}_m(\mathcal{F}_k, \boldsymbol{x})$ and $\hat{v}_k$ in terms of $\hat{\mathfrak{R}}_m(\mathcal{F}_j, \boldsymbol{x})$ and $\hat{v}_j$, with particularly efficient computation for $j = 1$. We now illustrate the use of theorem 4.2.3 with a simple corollary for the ℓ_2 linear loss family.

Corollary 4.2.4 (Bounds for Linear Loss Families)**.** Suppose as in theorem 4.2.3, and also $p = q = 2$, and all attribute values are contained by the unit sphere. Then

$$\hat{v}_k \leq k \sup_{c \in \mathcal{C}} \left\| \hat{\mathbb{C}}_{\boldsymbol{x}}[c] \right\|_{2 \to 2} \quad ,$$

where $\|\cdot\|_{2\to 2}$ is the *spectral norm* (largest eigenvalue). Additionally, with pr. $\geq 1 - \delta$ over $\boldsymbol{x}, \boldsymbol{\sigma}$, it holds that

$$\begin{aligned} \hat{\mathfrak{R}}_m(\mathcal{F}_k, \boldsymbol{x}) &\leq k \left(\hat{\mathfrak{R}}^n_m(\mathcal{F}_1, \boldsymbol{x}, \boldsymbol{\sigma}) + \sqrt{\tfrac{\ln \frac{1}{\delta}}{2nm}} \right) \\ &\leq k \left(\tfrac{1}{n} \sum_{i=1}^{n} \sup_{c \in \mathcal{C}} \left\| \tfrac{1}{m} \sum_{j=1}^{m} \boldsymbol{\sigma}_{i,j} c(\boldsymbol{x}_j) \right\|_2 + \sqrt{\tfrac{\ln \frac{1}{\delta}}{2nm}} \right) \quad . \end{aligned}$$

Consequently, theorem 4.2.2 holds with

$$\begin{aligned} \varepsilon_{\mathrm{RA}} &\leq \frac{\ln \frac{4}{\delta}}{3m} + \sqrt{\frac{2k(\hat{\mathfrak{R}}^n_m(\mathcal{F}_1, \boldsymbol{x}, \boldsymbol{\sigma}) + \varepsilon_{\mathrm{ERA}}) \ln \frac{4}{\delta}}{m}}; \\ \tilde{\mathfrak{R}} &\leq k\big(\hat{\mathfrak{R}}^n_m(\mathcal{F}_1, \boldsymbol{x}, \boldsymbol{\sigma}) + \varepsilon_{\mathrm{ERA}}\big) + \varepsilon_{\mathrm{RA}}; \ \& \\ \tilde{v} &\leq \tfrac{m}{m-1} k \hat{v}_1 + \frac{2 \ln \frac{4}{\delta}}{m} + \sqrt{\frac{2k\hat{v}_1 \ln \frac{4}{\delta}}{m-1}} \quad . \end{aligned}$$

Here we see that, for linear loss family $\mathcal{L}_2$, we can bound the wimpy variance by computing the *maximum variance* along any *unit vector* (much like in PCA or SVD). Similarly, bounding the ERA is straightforward as well, as each trial of the n-MCERA essentially corresponds to measuring the maximum (over codecs) *distance traveled* (ℓ_2 norm) in *attribute space* over a *random walk* (directions dictated by each $\boldsymbol{\sigma}_{i,j}$). Note that together, corollary D.3.6 & theorem 4.2.2 imply that for 1-Lipschitz welfare $\mathrm{M}(\cdot)$, $r \in \boldsymbol{\Theta}(1)$, with pr. $\geq 1 - \delta$, we have $\mathrm{M}(\hat{\boldsymbol{c}}_{\mathrm{M}}, \mathcal{L}, \mathcal{D}) - \mathrm{M}(\boldsymbol{c}^*_{\mathrm{M}}, \mathcal{L}, \mathcal{D}) \leq$

$$4k\hat{\mathfrak{R}}^n_m(\mathcal{F}_1, \boldsymbol{x}, \boldsymbol{\sigma}) + \boldsymbol{\Theta}\left(\frac{\ln \frac{1}{\delta}}{m} + \sqrt{\frac{k(v_1 + \mathfrak{R}_m(\mathcal{F}_1, \mathcal{D})) \ln \frac{1}{\delta}}{m}} \right) . \tag{35}$$

Type	Codec #	Parameter	Distortion
VBR	$\texttt{c}_0$-$\texttt{c}_9$	0, 1, ..., 9	Low-Med
ABR	$\texttt{c}_{10}$-$\texttt{c}_{13}$	320, 256, 128, 64	Low-Med
VBR	$\texttt{c}_{14}$-$\texttt{c}_{16}$	9.9, 9.99, 9.999	Med-High
ABR	$\texttt{c}_{17}$-$\texttt{c}_{19}$	32, 16, 8	Med-High
WAV	$\texttt{c}_{20}$	—	None

Table 4.1: We employ *Variable Bit Rate* (VBR) `mp3`, parameterized by *quality*, *Average Bit Rate* (ABR) `mp3`, parameterized by *bitrate*, and uncompressed `wav` codecs.

4.3 Experimental Evaluation of Ewo$k$s

We study various notions of multivariate optimality on four datasets, spanning speech and several music genres. The *Debussy* dataset consists of the complete orchestral works of Claude Debussy, conducted by Yan Pascal Tortelier; the *Explosions* dataset is the 2000–2013 discography of progressive rock band *Explosions in the Sky*; the *Zeppelin* dataset is the complete discography of hard rock band *Led Zeppelin*; and the *LibriVox* dataset contains public-domain LibriVox audiobooks, all cut into short nonoverlapping audio samples. In some experiments, we pool all music datasets or all four datasets to evaluate EWOkS on *heterogeneous data*. Table 4.1 describes the codecs $\texttt{c}_0, \texttt{c}_1, \ldots, \texttt{c}_{20}$ used in all experiments.

We quantify distortion in the audio domain with the *Perceptual Evaluation of Audio Quality* (PEAQ) [Thiede et al., 2000] metric, which is an *objective* measure of how well one audio sample approximates another, based on psychoacoustic principles, that aims to measure the degree of human-perceived difference. In all experiments, we consider two attributes, the first being PEAQ distortion, normalized to $[0, 1]$, and the second being the *compression ratio*, also in $[0, 1]$.

4.3.1 Data-Dependence and Pareto Optimality

A codec set $\boldsymbol{c} \in \binom{\mathcal{C}}{k}$ is said to be *Pareto-optimal* among $\binom{\mathcal{C}}{k}$ if there exists a *linear loss function* (nonnegative linear combination of attributes; see equation 34) $\ell \in \mathcal{L}_2$ for which the risk (equation 26) of $\boldsymbol{c}$ is minimal among $\binom{\mathcal{C}}{k}$, i.e.,

$$\mathrm{Pareto}\big(\tbinom{\mathcal{C}}{k}, \mathcal{D}\big) \doteq \bigcup_{\ell \in \mathcal{L}_2} \operatorname*{argmin}_{\boldsymbol{c} \in \binom{\mathcal{C}}{k}} \mathrm{R}(\boldsymbol{c}; \ell, \mathcal{D}) \ . \tag{36}$$

For $k = 1$, the *Pareto-optimal frontier* of $\mathcal{C}$ in $\mathbb{R}^d_{0+}$ is then the *lower convex-hull*[2] $\mathrm{LCH}(\cdot)$ of the set of *mean attribute values* of all Pareto-optimal codecs, i.e.,

$$\mathrm{PFrnt}(\mathcal{C}, \mathcal{D}) \doteq \mathrm{LCH}\big\{ \mathop{\mathbb{E}}_{x \sim \mathcal{D}}[c(x)] \ \big| \ c \in \mathcal{C} \big\} \ . \tag{37}$$

[2]NB the *convex hull* $\mathrm{CH}(\boldsymbol{a})$ of pointset $\boldsymbol{a} \subseteq \mathbb{R}^a$ is the *convex closure* of $\boldsymbol{a}$, and the *lower convex hull* is $\mathrm{LCH}(\boldsymbol{x}) \doteq \{\boldsymbol{a}' \in \mathrm{CH}(\boldsymbol{a}) \mid \forall \boldsymbol{\varepsilon} \succ \mathbf{0} \colon \boldsymbol{a}' - \boldsymbol{\varepsilon} \notin \mathrm{CH}(\boldsymbol{a})\}$.

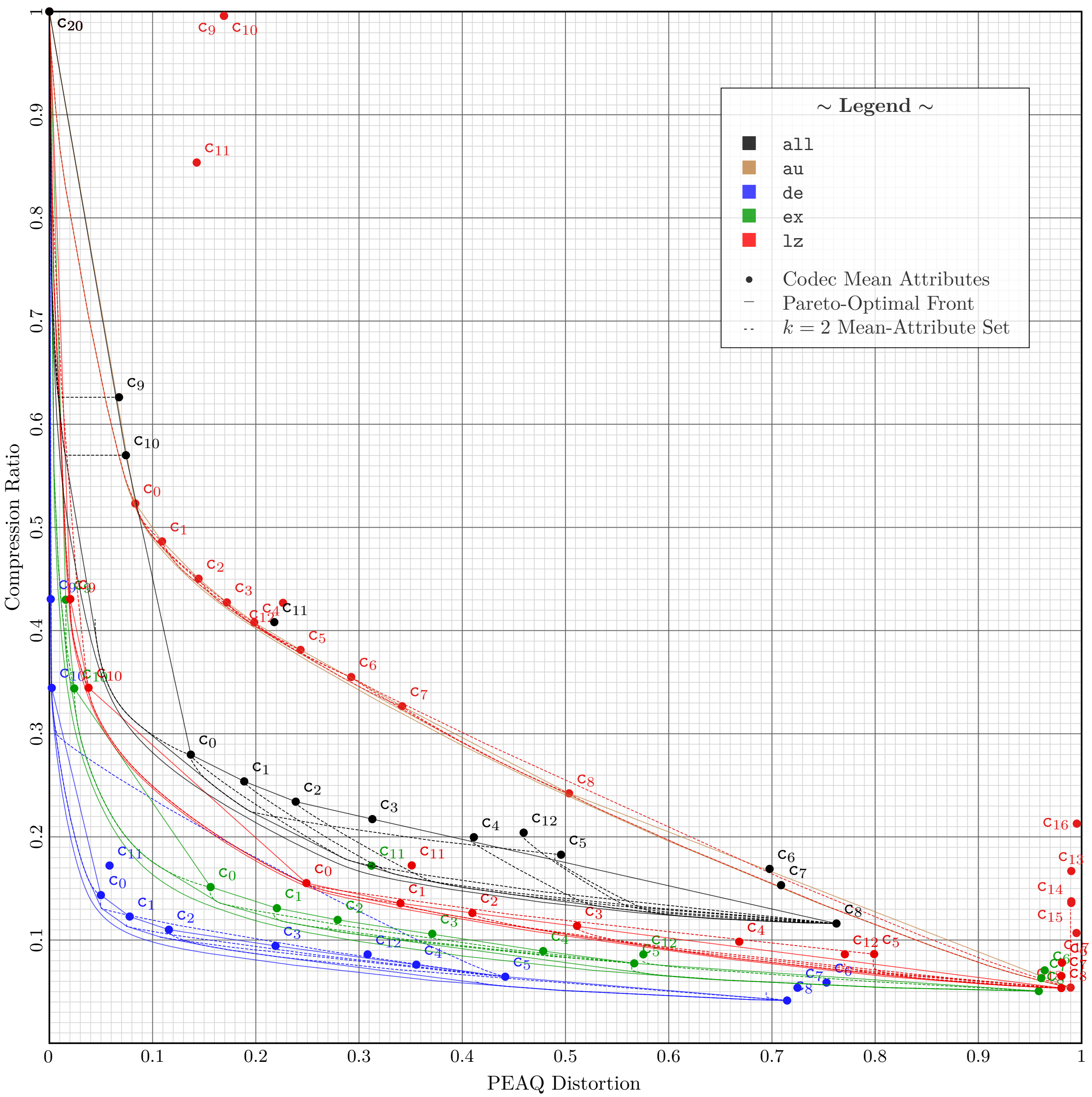


Figure 4.1: Pareto-optimal frontiers for all datasets. Mean distortion (x axis) and compression ratio (y axis) shown for each codec $\bullet$, with Pareto-optimal frontiers for codec families of sizes $k \in \{1, 2, |\mathcal{C}|\}$, color-coded by dataset. Curves for each k are unambiguous, since k *increases* from top-right to bottom-left. Mean-attribute sets for Pareto-optimal $k = 2$ pairs are shown in the inset as dotted curves.

This notion is easily visualized (see figure 4.1), and it concisely represents the set of *mean attribute values* (and thus risk values) of $\boldsymbol{c}^*_\ell \in \mathrm{CH}(\mathcal{C})$ for any $\ell \in \mathcal{L}_2$.

To generalize the *Pareto-optimal frontier* to $k > 1$, first observe that the *mean attribute values* of a k codec set depend on $\ell \in \mathcal{L}_2$, as EWOkS takes a *minimum* over k codecs. Consequently, each $\boldsymbol{c} \in \binom{\mathcal{C}}{k}$ is assoociated with a *mean-attribute set*, and we take the Pareto-optimal frontier of $\binom{\mathcal{C}}{k}$ to be the *lower convex hull* of the *union* of the mean-attribute sets of each $\boldsymbol{c} \in \binom{\mathcal{C}}{k}$. Algebraically, $\mathrm{PFrnt}(\binom{\mathcal{C}}{k}, \mathcal{D})$ is the LCH of

$$\bigcup_{\boldsymbol{c} \in \binom{\mathcal{C}}{k}} \left\{ \mathop{\mathbb{E}}_{x \sim \mathcal{D}} \left[\mathop{\mathrm{argmin}}_{c(x) \in \boldsymbol{c}(x)} \ell(c(x)) \right] \middle| \ell \in \mathcal{L}_2 \right\} \ . \tag{38}$$

Observe that equation 38 generalizes equation 37, as for $k = 1$ and $\mathcal{C} = \{c\}$, we have $\mathrm{argmin}_{c(x) \in \boldsymbol{c}(x)} \ell(c(x)) = c(x)$. Note that the *subadditive nonlinearity* of the argmin (i.e., the improvement from selecting the *best* among k codecs) improves $\mathrm{PFrnt}(\binom{\mathcal{C}}{k}, \mathcal{D})$ *monotonically* in k. Again the Pareto-optimal frontier concisely visualizes the possible *mean attribute values* (and thus *risks*) of the optimal $\boldsymbol{c} \in \binom{\mathcal{C}}{k}$ for any $\ell \in \mathcal{L}_2$.

Note that, like the welfare concepts of section 4.2.1, Pareto-optimality is a notion of multivariate optimality. As with welfare, we may define the *empirical Pareto-optimal set* and the *empirical Pareto-optimal front*, w.r.t. sample $\boldsymbol{x}$, by replacing $\mathbb{E}_{\mathcal{D}}[\cdot]$ with $\hat{\mathbb{E}}_{\boldsymbol{x}}[\cdot]$. Furthermore, by its inherently linear nature, it is trivial to adapt the uniform convergence guarantees of theorem 4.2.3 to bound the gap between empirical and true Pareto-optimaltity concepts.

In figure 4.1, we plot the empirical mean PEAQ distortion and compression ratios of codecs, and empirical Pareto-optimal frontiers of codec sets of sizes $k \in \{1, 2, |\mathcal{C}|\}$ on the music, speech, and mixed datasets. Immediately we see that the Pareto-optimal frontiers differ between datasets, illustrating the importance of data-dependent codec selection. We see that on the speech and music datasets, the $k = 1$ (top-right most) Pareto-optimal frontier usually performs similarly to the $k = |\mathcal{C}|$, which indicates little variability in codec performance over these distributions, thus a static codec for each objective performs reasonably well. However, for the *mixed dataset*, we see a massive improvement in moving from $k = 1$ to $k = 2$ codecs, and modest diminishing returns for $k > 2$. This makes sense, as for fixed objectives, different codecs are often optimal for the LibriVox and music datasets, thus combining them is optimal in the mixed dataset.

4.3.2 Fairness and Welfare-Optimality

The above Pareto-optimality experiments indicate that *for any fixed objective*, there exists a *small codec set* that is near-optimal (i.e., $\mathrm{PFrnt}(\binom{\mathcal{C}}{2}, \mathcal{D}) \approx \mathrm{PFrnt}(\binom{\mathcal{C}}{|\mathcal{C}|}, \mathcal{D})$) However, we see in the inset that *individual pairs* of codecs (dotted lines) only achieve Pareto-optimality for *small ranges* of $\mathcal{L}$, thus $k > 2$ codecs may

Table 4.2: Experiments optimizing $\ell_{1/2}$ loss, and $\mathrm{M}_1(\cdot)$, $\mathrm{M}_\infty(\cdot)$, and $\mathrm{M}_2(\cdot)$ welfares on `au`. **a)** *Objective values*, and **b)** *regret* for each row-objective, when optimizing $\hat{\boldsymbol{c}}_C$ for each column-objective o_C.

a) Objective Values:

k	$\ell_{1/2}$	M_1	M_∞	M_2
1	.297	.297	.341	.309
2	.294	.217	.328	.234
3	.293	.208	.327	.229
$\lvert\mathcal{C}\rvert$	.293	.202	.326	.223

b) $k = 3$ Regret Matrix:

	$\ell_{1/2}$	M_1	M_∞	M_2
$\ell_{1/2}$	**0**	.0098	.0097	.0015
M_1	.1343	**0**	.1221	.0189
M_∞	.0811	.0374	**0**	.0103
M_2	.0607	.0035	.0454	**0**

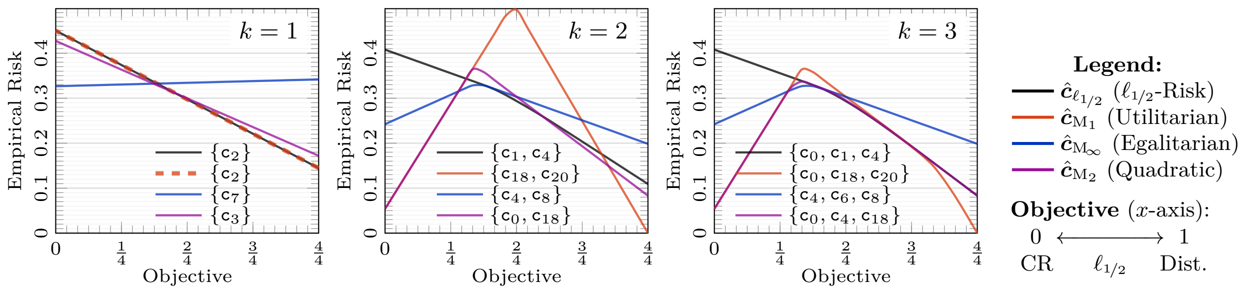


Figure 4.2: Empirical risk (y-axis) plotted against linear objective family $\mathcal{L}_1$ (x-axis), with *distortion weight* $\boldsymbol{w}_1 = x$ and *compression ratio weight* $\boldsymbol{w}_2 = 1 - x$, on `au`, for $k \in \{1, 2, 3\}$. Empirically-optimal codec sets $\hat{\boldsymbol{c}}_o \in \binom{\mathcal{C}}{k}$ for each objective o are also given.

be required to well-optimize many loss functions *simultaneously*; consequently, we expect various *welfare concepts* to continue improving beyond $k = 2$. In table 4.2 & figure 4.2, we study the effect of increasing k on the *welfare* of various EWOkS-optimal solution concepts, which is sensitive to performance *across all of* $\mathcal{L}$.

In this experiment, we consider the 2-dimensional ℓ_1-*linear loss family* $\mathcal{L}_1 = \{\, \boldsymbol{a} \mapsto \boldsymbol{w} \cdot \boldsymbol{a} \,:\, \boldsymbol{w}_1 \in [0, 1], \boldsymbol{w}_2 = 1 - \boldsymbol{w}_1 \}$ (see eq 34), where $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ represent the penalty-weight a user places on *distortion* and *compression ratio*, respectively. We optimize the neutral single linear objective (see eq 34) $\ell_{1/2}(\boldsymbol{a}) \doteq (1/2, 1/2) \cdot \boldsymbol{a}$, *utilitarian welfare* $\mathrm{M}_1(\cdot)$ on the loss distribution over $\mathcal{L}_1$ with density $\propto |\boldsymbol{w}_1 - 1/2|^2$ (i.e., the user distribution is biased towards the extremes), as well as *egalitarian welfare* $\mathrm{M}_\infty(\cdot)$ over $\mathcal{L}_1$ and *quadratic welfare* $\mathrm{M}_2(\cdot)$ with uniform density over $\mathcal{L}_1$.

Table 4.2a shows the raw empirical objective values of each optimality concept, for $k \in \{1, 2, 3, |\mathcal{C}|\}$, and table 4.2b shows the *regret* of (row) objective o_R on $\hat{\boldsymbol{c}}_C$ optimized for each (column) objective o_C, which is the amount by which objective o_R would *improve* if $\hat{\boldsymbol{c}}_C$ were optimized for o_R instead of o_C.

To better understand the tradeoffs made by EWOkS under various objectives, for each objective o, figure 4.2 plots the *empirical risk* (see equation 29) of each o-optimal k codec selection $\hat{\boldsymbol{c}}_o$ w.r.t. each $\ell \in \mathcal{L}_1$ as a function of *distortion weight* $\boldsymbol{w}_1$. In particular, each plot corresponds to a *selection size* k, and each line represents the *empirical risk* w.r.t. the linear loss function with $\boldsymbol{w}_1 = x$, i.e., from left to right we interpolate from *pure distortion* to *pure compression ratio* losses.

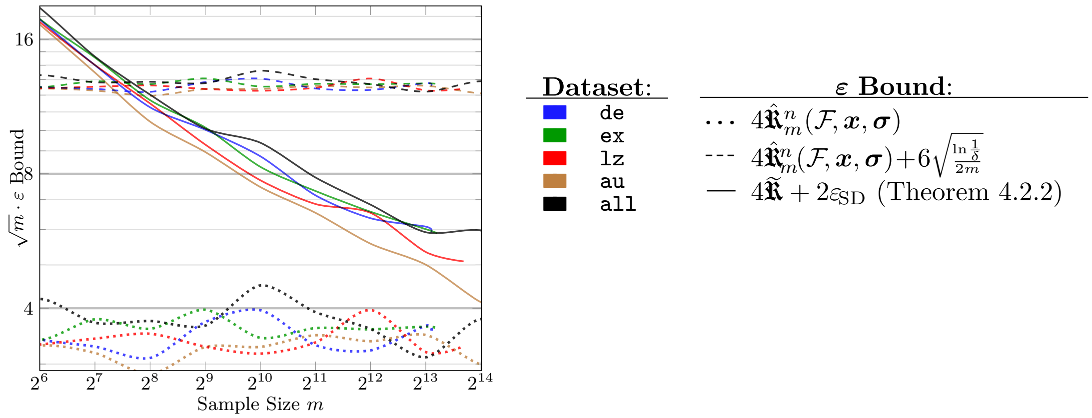


Figure 4.3: Log-log plots of Ewo*k*s $\sqrt{m}$-scaled $\mathrm{M}(\cdot)$-optimality gap bound $\sqrt{m}\varepsilon$ (x-axis), versus sample size m (y-axis), in expectation with $4\hat{\mathfrak{R}}^n_m(\mathcal{F}, \boldsymbol{x}, \boldsymbol{\sigma})$, and w.h.p. with McDiarmid and theorem 4.2.2.

From the *objective values* matrix in table 4.2, we see that the single-objective $\ell_{1/2}$ experiences rapidly diminishing returns as k increases, improving by only 0.004 from $k = 1$ to $k = |\mathcal{C}|$, whereas the welfare objectives gradually improve with increasing k. This is as expected, because a sufficiently diverse set of codecs must be selected so as to perform reasonably well *across all* $\ell \in \mathcal{L}$, rather than *for some fixed* ℓ.

This experiment also clearly illustrates the *fairness impact* of objective choice, as we see that *improving utilitarian welfare* $\mathrm{M}_1(\cdot)$ can *decrease egalitarian welfare* $\mathrm{M}_\infty(\cdot)$, as occurs for $k = 2$, when $\hat{\boldsymbol{c}}_{\mathrm{M}_1} = \{\mathtt{c}_{18}, \mathtt{c}_{20}\}$, which are the second-lowest bitrate and uncompressed codecs, respectively, resulting in high regret (0.1343) for $\mathrm{M}_\infty(\cdot)$ (see table 4.2a). This is unsurprising, as it is easy to produce low distortion *or* low compression ratio encodings; the difficulty lies in optimizing both *simultaneously*. The regret matrix (table 4.2b) also shows fairness tradeoffs; here we see that the quadratic-welfare-optimal $\hat{\boldsymbol{c}}_{\mathrm{M}_2}$ is much better on-average than $\hat{\boldsymbol{c}}_{\mathrm{M}_\infty}$ (utilitarian regret 0.0189 versus 0.1121), despite worst-case performance only 0.01 worse than $\hat{\boldsymbol{c}}_{\mathrm{M}_\infty}$. This confirms that while optimizing the *worst-case* $\mathrm{M}_\infty(\cdot)$ is at-odds with optimizing the *average-case* $\mathrm{M}_1(\cdot)$, the quadratic $\mathrm{M}_2(\cdot)$ is a reasonable compromise.

4.3.3 Uniform Convergence Bounds

We now show that our uniform convergence bounds yield valuable conclusions as to the approximate optimality of Ewo*k*s solutions, and outperform McDiarmid bounds on real data. In figure 4.3, we plot Ewo*k*s optimality-gap bounds against sample size m, assuming $\mathrm{M}(\cdot)$ is 1-Lipschitz, and using selection-size $k = 3$, loss family $\mathcal{L}_1$ (see (34)), $n = 100$ Monte-Carlo trials, and tail bounds with failure probability $\delta = 0.01$. Here all bounds are scaled by $\sqrt{m}$ to visualize the *deviation* of *rates* from $\boldsymbol{\Theta}\sqrt{1/m}$, which appears flat under $\sqrt{m}$-scaling.

We immediately see that our variance-sensitive bounds (theorem 4.2.2) are significantly sharper than the

McDiarmid bounds, yielding valuable conclusions after only $m \approx 1000$ samples. The *mixed convergence rates* (see theorem 4.2.2) of these bounds are visible as initial *fast-decay* (*negative slopes*), straightening out as $m \to \infty$ and the asymptotic $\boldsymbol{\Theta}\sqrt{v/m}$ slow rate is reached. In contrast, the McDiarmid bounds, and the n-MCERAs themselves, all exhibit slow $\boldsymbol{\Theta}\sqrt{1/m}$ rates (even slopes). In particular, the poor performance of the McDiarmid bounds occurs because they are always dominated by their concentration term $6\sqrt{\ln\frac{1}{\delta}/2m}$, visible as a large parallel gap between them and the n-MCERAs.

We see also from their relative order that `au` is the easiest to learn, and `all` the most difficult. This is unsurprising; in music and mixed data, we expect more *variability* (thus *variance*), and also more benefit (overfitting) with $k > 1$ (thus higher $\widetilde{\mathfrak{R}}$).

4.4 Discussion

We propose a novel media streaming strategy, where a small set of codecs $\boldsymbol{c}$ is selected such that the set includes a "good" solution for each $\ell \in \mathcal{L}$ over $\mathcal{D}$. We develop the *Empirical Welfare-Optimal k codec Selection* (EWO*k*S) algorithm, which *learns* a data-dependent set of k codecs $\boldsymbol{c}$ that is *welfare-optimal* over $\mathcal{L}$. The *selection size k* balances between the cost of *compressing / storing* multiple encodings of media, and the improvement to user experience and fairness.

Our experiments show the importance of considering both the objective *and* the data-distribution. We see that while k codec selection yields modest improvement for *single-objectives* (section 4.3.1), particularly for heterogenous (mixed speech / music) data, it greatly improves *welfare-objectives* over *loss families* (section 4.3.2). Our experiments confirm that bias-issues exist in codec selection, and exhibit several surprising phenomena (e.g., optimizing a single neutral loss function can yield bias issues, learning a more sophisticated [higher k] model for one welfare-concept can harm other welfare concepts), further motivating rigorous analysis and provable methods for controlling algorithmic bias.

We rigorously analyze the generalization error of EWO*k*S (section 4.2.3), applying *Rademacher averages* with a novel variance-sensitive bound for welfare-generalization. We handle arbitrary Lipschitz welfare functions, whereas previous applications gave weaker bounds limited to simple cases, such as *utilitarian welfare* in *auctions* [Hoy et al., 2017]. Since general welfare-concepts depend nonlinearly on many loss values *simultaneously*, uniform convergence is an ideal tool to analyze them. Our experiments (section 4.3.3) also show our bounds significant improvement traditional methods in the small-sample domain.

Our data-driven codec-selection approach follows trends in machine learning and databases. We replace *fixed codecs* with *learned codecs*, in much the same way that autoML systems replace fixed models with learned pipelines [Hutter et al., 2011], and database systems replace static indices with learned indices [Kraska et al., 2018]. As this trend continues, and more critical services integrate learned components, the importance of fairness continues to grow; we are confident that our welfare analysis methods for provable fairness can be applied beyond the domain of codec selection.

Chapter 5

Fair Probably Approximately Correct Learning

Building upon the concept of malfare of chapter 3, I define *fair-PAC learning*, where a fair PAC-learner is an algorithm that learns an ε-δ malfare-optimal model with bounded sample complexity, for *any data distribution*, and for *any* (axiomatically justified) malfare concept, i.e., power-mean malfare with $p \in [1, \infty)$. This definition extends Valiant's [Valiant, 1984] classic PAC-learning formalization of machine learning, and I show that, with appropriate modifications, many (standard) PAC-learners may be converted to fair-PAC learners, and argue that fair-PAC-learners are intuitive and easy to use, as one must only select an axiomatically justified malfare concept (encoding their desired fairness concept), hypothesis class, and confidence guarantees, and then receives a provably ε-δ optimal model. I show broad conditions under which, with appropriate modifications, many standard PAC-learners may be converted to fair-PAC learners. I also argue that fair-PAC-learners are intuitive and easy to use, as one must only select a malfare concept (encoding their desired fairness concept), hypothesis class, and confidence guarantees, and then one receives a provably (probabilistically) near-optimal model.

This is in contrast to similar frameworks, e.g., the *Seldonian learner* of [Thomas et al., 2019], where (1), the arbitrarily statistically infeasible task of constrained optimization must be addressed, and (2), learning guarantees are in terms of less-interpretable *constrained objectives* (e.g., specified objectives, parity constraints, and tolerance values), rather than singular cardinal *malfare objectives*, which directly encode an interpretable single-parameter fairness concept that ranges between *average case* ($p = 1$) and *worst case* ($p = \infty$) performance across groups. This places fair-PAC learning on firm theoretical ground, as it yields statistical, and in some cases computational, efficiency guarantees for many well-studied machine-learning models, and is also practically relevant, as it democratizes fair machine learning by providing concrete training algorithms and rigorous generalization guarantees for these models.

In this framework, both PAC and FPAC learning are parameterized by a *learning task* (model space and loss function), and I explore the rich hierarchy of learnability under variations of these concepts. In particular, I show that for many loss functions, PAC and FPAC learning are *statistically equivalent* (i.e., PAC-learnability implies FPAC-learnability) in section 5.2. Furthermore, in section 5.3 I show that standard *convexity conditions* sufficient for PAC-learnability are also sufficient for FPAC-learnability, as are certain *coverability conditions* that ensure the hypothesis space is sufficiently small so as to limit overfitting, and may be efficiently approximately searched. I explore the basic relationships between various learnability classes, but many open questions remain, and I hope that future work will further characterize these practically interesting and theoretically deep problems. This chapter is adapted from the latter half of Cousins [2021].

5.1 Statistical and Computational Efficiency Learning Guarantees

In this section, we define a formal notion of fair-learnability, termed *fair PAC-learning*, where a loss function and hypothesis class are fair PAC-learnable essentially if any distribution can be approximately learned from a polynomially-sized sample (w.h.p.). We then construct various fair-PAC learners, and relate the concept to standard PAC learning [Valiant, 1984], with the understanding that this allows the vast breadth of research of PAC-learning algorithms, and quite saliently, necessary and sufficient conditions, to be applied to fair-PAC learning. In particular, we show a hierarchy of fair-learnability via generic statistical and computational learning theoretic bounds and reductions.

Hypothesis Classes and Sequences We now define *hypothesis class sequences*, required to discuss nontrivial *computational complexity* in learning. This definition is adapted from [Shalev-Shwartz and Ben-David, 2014, Def. 8.1], which treats only *binary classification.*

Definition 5.1.1 (Hypothesis Class Sequence)**.** A *hypothesis class sequence* $\mathcal{H} = (\mathcal{H}_1, \mathcal{H}_2, \dots)$ is a *nested sequence* of *hypothesis classes* mapping $\mathcal{X} \to \mathcal{Y}$. In other words, $\mathcal{H}_1 \subseteq \mathcal{H}_2 \subseteq \dots$.

Usually, $\mathcal{H}_i$ is easily derived from $\mathcal{H}_j$. For instance, *linear classifiers* naturally form a sequence of families using their *dimension*:

$$\mathcal{H}_d \doteq \left\{ \boldsymbol{x} \mapsto \operatorname{sgn}\big(\boldsymbol{x} \cdot (\boldsymbol{w} \circ \boldsymbol{0})\big) \;\middle|\; \boldsymbol{w} \in \mathbb{R}^d \right\} \enspace .$$

Here each $\mathcal{H}_i$ is defined over domain $\mathbb{R}^\infty$, but it is more natural to discuss them as objects over $\mathbb{R}^d$. Similarly, unit-scale *univariate polynomial regression* naturally decomposes as

$$\mathcal{H}_d \doteq \left\{ x \mapsto (x, x^2, \ldots, x^d) \cdot \boldsymbol{w} \,\middle|\, \boldsymbol{w} \in [-1, 1]^d \right\} \ .$$

The *hypothesis class sequence* concept is necessary, as it allows us to distinguish statistically easy problems, like learning hyperplanes in finite-dimensional space, from the statistically challenging problem of learning hyperplanes in $\mathbb{R}^\infty$. It is also important to studying computational-hardness of statistically-easy learning problems, as this essentially boils down to selecting a hypothesis class sequence such that learning *time complexity* grows *superpolynomially* (in d).

For context, we first present a generalized notion of PAC-learnability, which we then generalize to fair-PAC-learnability. Standard presentations consider only classification under 0-1 loss, but following the generalized learning setting of Vapnik [2013], some authors consider generalized notions for other learning problems (see, e.g., Shalev-Shwartz and Ben-David [2014, Definition 3.4]).

Definition 5.1.2 (PAC-Learnability)**.** Suppose *nested hypothesis class sequence* $\mathcal{H}_1 \subseteq \mathcal{H}_2 \subseteq \ldots$, all over $\mathcal{X} \to \mathcal{Y}$, and *loss function* $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{0+}$. We say $\mathcal{H}$ is *PAC-learnable* w.r.t. ℓ if $\exists$ a (randomized) algorithm $\mathcal{A}$, such that $\forall$:

1. sequence index d;
2. instance distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$;
3. additive approximation error $\varepsilon > 0$; and
4. failure probability $\delta \in (0, 1)$;

it holds that $\mathcal{A}$ can identify a hypothesis $\hat{h} \in \mathcal{H}$, i.e., $\hat{h} \leftarrow \mathcal{A}(\mathcal{D}, \varepsilon, \delta, d)$, such that

1. there exists some *sample complexity* function $\mathrm{m}(\varepsilon, \delta, d) : \big(\mathbb{R}_+ \times (0, 1) \times \mathbb{N}\big) \to \mathbb{N}$ s.t. $\mathcal{A}(\mathcal{D}, \varepsilon, \delta, d)$ consumes no more than $\mathrm{m}(\varepsilon, \delta, d)$ samples from $\mathcal{D}$ (i.e., has finite sample complexity); and
2. with probability at least $1 - \delta$ (over randomness of $\mathcal{A}$), $\hat{h}$ obeys

$$\mathrm{R}(\hat{h}; \ell, \mathcal{D}) \leq \inf_{h^* \in \mathcal{H}} \mathrm{R}(h^*; \ell, \mathcal{D}) + \varepsilon \ .$$

The class of such learning problems is denoted PAC, thus we write $(\mathcal{H}, \ell) \in \mathrm{PAC}$ to denote PAC-learnability.

Furthermore, if for all d, the space of $\mathcal{D}$ is restricted such that

$$\exists h \in \mathcal{H}_d \text{ s.t. } \mathrm{R}(h; \ell, \mathcal{D}) = 0 \ ,$$

then $(\mathcal{H}, \ell)$ is *realizable-PAC-learnable*, written $(\mathcal{H}, \ell) \in \mathrm{PAC}^{\mathrm{Rlz}}$.

Observation 5.1.3 (On Realizable Learning)**.** Our definition of realizability appears to differ from the standard form, in which $\mathcal{D}$ is a distribution over only $\mathcal{X}$, and y is simply computed as $h^*(x)$, for some $h^* \in \mathcal{H}$. We instead constrain $\mathcal{D}$ such that there exists a 0-risk $h^* \in \mathcal{H}$, which is equivalent for any loss function ℓ such that $\ell(y, \hat{y}) = 0 \Leftrightarrow y = \hat{y}$, e.g., 0-1 classification loss, or absolute or square regression loss. With our definition, it is much clearer that realizable learning is a special case of agnostic learning, and furthermore, we handle a much broader class of problems, for which there may be some amount of *noise*, or wherein a ground truth may not even exist.

For instance, in a *recommender system*, y may represent the *set of items* that x will like, and h may predict a singleton set, and thus we take $\ell(y, \{\hat{y}\}) = \mathbb{1}_y(\hat{y})$. There is no ground-truth here, but rather we seek a *compatible solution* that recommends appropriate items to everyone. Similarly, in high-dimensional classification, we often treat classifier output as a *ranked list* of predictions, and the *top-k loss* is taken to be $\ell(y, \hat{y}) = \mathbb{1}_{\hat{y}_{1:k}}(y)$. Again, no ground truth exists, but 0-risk learning is still possible if there is not "too much" ambiguity (e.g., foxes and dogs can be confused, as long as they are separated from horses and zebras).

We now generalize this concept to *fair-PAC-learnability.* In particular, we replace the *univariate risk-minimization* task with a *multivariate malfare-minimization* task. Following the theory of section 3.3.3, we do not commit to any particular objective, but instead require that a fair-PAC-learner is able to maximize *any* fair malfare function satisfying the standard axioms. Furthermore, here problem instances grow not just in problem complexity d, but also in the number of groups g.

Definition 5.1.4 (Fair PAC (FPAC)-Learnability)**.** Suppose *nested hypothesis class sequence* $\mathcal{H}_1 \subseteq \mathcal{H}_2 \subseteq \cdots \subseteq \mathcal{X} \to \mathcal{Y}$, and *loss function* $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{0+}$. We say $\mathcal{H}$ is *fair PAC-learnable* w.r.t. *group count* g and *loss function* ℓ if $\exists$ a (randomized) algorithm $\mathcal{A}$, such that $\forall$:

1. sequence index d;
2. group count g;
3. instance distribution $\mathcal{D}_{1:g}$ over $(\mathcal{X} \times \mathcal{Y})^g$;
4. group weights measure $\boldsymbol{w}$ over $\{1, \ldots, g\}$;
5. malfare concept $\text{M}(\cdot; \cdot)$ satisfying axioms 1-7 and 9;

6. additive approximation error $\varepsilon > 0$; and

7. failure probability $\delta \in (0, 1)$;

it holds that $\mathcal{A}$ can identify a hypothesis $\hat{h} \in \mathcal{H}$, i.e., $\hat{h} \leftarrow \mathcal{A}(\mathcal{D}_{1:g}, \boldsymbol{w}, \mathrm{M}, \varepsilon, \delta, d)$, such that

1. there exists some *sample complexity* function $\mathrm{m}(\varepsilon, \delta, d, g) : \left(\mathbb{R}_+ \times (0,1) \times \mathbb{N} \times \mathbb{N}\right) \to \mathbb{N}$ s.t. $\mathcal{A}(\mathcal{D}_{1:g}, \boldsymbol{w}, \mathrm{M}, \varepsilon, \delta, d)$ consumes no more than $\mathrm{m}(\varepsilon, \delta, d, g)$ samples (finite sample complexity); and

2. with probability at least $1 - \delta$ (over randomness of $\mathcal{A}$), $\hat{h}$ obeys

$$\mathrm{M}\left(i \mapsto \mathrm{R}(\hat{h}; \ell, \mathcal{D}_i); \boldsymbol{w}\right) \leq \operatorname*{argmin}_{h^* \in \mathcal{H}} \mathrm{M}\left(i \mapsto \mathrm{R}(h^*; \ell, \mathcal{D}_i); \boldsymbol{w}\right) + \varepsilon \ .$$

The class of such fair-learning problems is denoted FPAC, thus we write $(\mathcal{H}, \ell) \in \mathrm{FPAC}$ to denote fair-PAC-learnability.

Finally, if for all d, the space of $\mathcal{D}$ is restricted such that

$$\exists h \in \mathcal{H}_d \text{ s.t. } \max_{i \in 1, \ldots, g} \mathrm{R}(h; \ell, \mathcal{D}_i) = 0 \ ,$$

then $(\mathcal{H}, \ell)$ is *realizable-fair-PAC-learnable*, written $(\mathcal{H}, \ell) \in \mathrm{FPAC}^{\mathrm{Rlz}}$.

We now observe that a few special cases are familiar learning problems, though we argue that all cases are of interest, and simply represent different ideals of fairness which may be situationally appropriate.

Observation 5.1.5 (Malfare Functions and Special Cases)**.** By assumption, $\mathrm{M}(\cdot; \cdot)$ must be $\mathrm{M}_p(\cdot; \cdot)$ for $p \in [1, \infty)$. Taking $g = 1$ implies $\boldsymbol{w} = (1)$, and $\mathrm{M}_p(\mathcal{S}; \boldsymbol{w}) = \mathcal{S}_1$, thus reducing the problem to standard PAC-learning. Similarly, taking $p = 1$ converts the problem to *weighted loss minimization* (weights determined by $\boldsymbol{w}$), and $p = \infty$ yields a *minimax optimization problem*, where the max is over groups, commonly encountered in adversarial and robust learning settings.

An aside on the flexibility of fair-PAC-learnability Note that the generalized definition of (fair) PAC-learnability is sufficiently broad so as to include many *supervised*, *semi-supervised*, and *unsupervised* learning problems. While this is not immediately apparent, consider that, for instance, *k-means clustering* can be expressed as a *learning problem* where the task is to identify a set of k vectors in $\mathbb{R}^d$ where the hypothesis class maps a vector $\boldsymbol{x}$ on to the nearest cluster center, and the loss function is the square distance. This is a surprisingly natural fairness issue when cast as a *resource allocation problem*; if, for instance each cluster center represents a cellphone tower, then we seek to place towers to serve *all groups*, and to avoid serving one or more groups particularly well at the expense of others.

On Computational Efficiency Some authors consider not just the *statistical* but also the *computational* performance of learning, generally requiring that $\mathcal{A}$ have *polynomial* time complexity (and thus implicitly sample complexity). In other words, they require that $\mathcal{A}(\mathcal{D}, \varepsilon, \delta, d)$ terminates in $\mathrm{m}(\varepsilon, \delta, d) \in \mathrm{Poly}(\frac{1}{\varepsilon}, \frac{1}{\delta}, d)$ steps. The main focus of this manuscript is sample complexity, but for completeness, we note that a similar concept of *polynomial-time fair-PAC learnability* is equally interesting, where here we assume $\mathcal{A}(\mathcal{D}_{1:g}, \boldsymbol{w}, \mathrm{M}, \varepsilon, \delta, d)$ may be computed by a Turing machine (with access to *sampling* and *entropy* oracles) in $\mathrm{m}(\varepsilon, \delta, d, g) \in \mathrm{Poly}(\frac{1}{\varepsilon}, \frac{1}{\delta}, d, g)$ steps, and thus implicitly the same bound on sample complexity. We denote these concepts $\mathrm{PAC}^{\mathrm{Agn}}_{\mathrm{Poly}}$, $\mathrm{PAC}^{\mathrm{Rlz}}_{\mathrm{Poly}}$, $\mathrm{FPAC}^{\mathrm{Agn}}_{\mathrm{Poly}}$, and $\mathrm{FPAC}^{\mathrm{Rlz}}_{\mathrm{Poly}}$.

Some trivial reductions We first observe (immediately from definitions 5.1.2 and 5.1.4) that PAC-learning is a special case of fair-PAC-learning. In particular, taking $g = 1$ implies $\mathrm{M}_p(\mathcal{S}) = \mathrm{M}_1(\mathcal{S}) = \mathcal{S}_1$, thus *malfare-minimization* coincides with *risk minimization*. The more interesting question, which we seek to answer in the remainder of this document, is *when* and *whether* the *converse* holds. Furthermore, when possible, we would like to show practical, sample-and-compute-efficient *constructive reductions*.

Realizability We first show that in the *realizable case*, *PAC-learnability* implies *fair PAC-learnability*. In particular, we employ a simple and practical *constructive* polynomial-time reduction. More efficient reductions are possible for particular values of p, g. Our reduction simply takes a sufficiently number of samples from the *uniform mixture distribution* over all g groups, and PAC-learns on this distribution. As the reduction is constructive, (and efficiency-preserving) this gives us generic algorithms for (efficient) FPAC-learning in terms of algorithms for PAC-learning.

Theorem 5.1.6 (Realizable Reductions)**.** Suppose loss function ℓ and hypothesis class $\mathcal{H}$. Then

1. $(\mathcal{H}, \ell) \in \mathrm{PAC}^{\mathrm{Rlz}} \Rightarrow (\mathcal{H}, \ell) \in \mathrm{FPAC}^{\mathrm{Rlz}}$; and
2. $(\mathcal{H}, \ell) \in \mathrm{PAC}^{\mathrm{Rlz}}_{\mathrm{Poly}} \Rightarrow (\mathcal{H}, \ell) \in \mathrm{FPAC}^{\mathrm{Rlz}}_{\mathrm{Poly}}$.

We construct a(n) (efficient) FPAC-learner for $(\mathcal{H}, \ell)$ by noting that there exists some $\mathcal{A}'$ with sample-complexity $\mathrm{m}_{\mathcal{A}'}(\varepsilon, \delta, d)$ and time complexity $\mathrm{t}_{\mathcal{A}'}(\varepsilon, \delta, d)$ to PAC-learn $(\mathcal{H}, \ell)$, and taking $\mathcal{A}(\mathcal{D}_{1:g}, \boldsymbol{w}, \mathrm{M}, \varepsilon, \delta, d) \doteq \mathcal{A}'(\mathrm{mix}(\mathcal{D}_{1:g}), \frac{\varepsilon}{g}, \delta, d)$. Then $\mathcal{A}$ FPAC-learns $(\mathcal{H}, \ell)$, with sample-complexity $\mathrm{m}_{\mathcal{A}}(\varepsilon, \delta, d, g) = \mathrm{m}_{\mathcal{A}'}(\frac{\varepsilon}{g}, \delta, d)$, and time-complexity $\mathrm{t}_{\mathcal{A}}(\varepsilon, \delta, d, g) = \mathrm{t}_{\mathcal{A}'}(\frac{\varepsilon}{g}, \delta, d)$.

Proof. We first show the sample complexity result. Suppose $\hat{h} \leftarrow \mathcal{A}'(p, \boldsymbol{w}, \mathcal{D}_{1:g}, \varepsilon, \delta, d)$. Then, with probability at least $1 - \delta$ (by the guarantee of $\mathcal{A}$), we have

$$\begin{aligned} \mathrm{M}_p(i \mapsto \mathrm{R}(h; \ell, \mathcal{D}_i)), \boldsymbol{w}) &\leq \mathrm{M}_\infty\big(i \mapsto \mathrm{R}(h; \ell, \mathcal{D}_i), i \mapsto \tfrac{1}{g}\big) \\ &\leq g\mathrm{M}_1\big(i \mapsto \mathrm{R}(h; \ell, \mathcal{D}_i), i \mapsto \tfrac{1}{g}\big) \\ &= g\mathrm{R}\big(h; \ell, \mathrm{mix}(\mathcal{D}_{1:g})\big) \leq g\tfrac{\varepsilon}{g} = \varepsilon \ . \end{aligned}$$

We thus may conclude that $(\mathcal{H}, \ell)$ is efficiently *realizable-PAC-learnable* by $\mathcal{A}'$, with *sample complexity* $\mathrm{m}_{\mathcal{A}}(\varepsilon, \delta, d, g) = \mathrm{m}_{\mathcal{A}'}(\frac{\varepsilon}{g}, \delta, d)$. Similarly, if $\mathcal{A}$ has polynomial runtime, then so too does $\mathcal{A}'$, thus we may also conclude efficiency. $\square$

While mathematically correct, if somewhat trivial, unfortunately, this argument does not extend to the agnostic case. This is essentially because in general it is not possible to simultaneously satisfy all groups, and even with reweighting, the malfare of ERM solutions are highly unstable in the weights. To see this, suppose $\mathcal{Y} \doteq \{a, b\}$, group A always wants a, and group B always wants b, with symmetric preferences, and we wish to optimize egalitarian malfare. For any reweighting, the utility-optimal solution is always to produce all a or all b, except with equal weights, when all solutions are equally good. Thus for no reweighting do all reweighted-risk solutions even approximate the egalitarian-optimal solution (which is evenly split between a and b). We thus conclude that simple constructive reductions using PAC-learners as subroutines is likely to solve the general problem.

In addition to the argument being inextensible to the agnostic case, we note that, philosophically speaking, realizable FPAC learning is rather uninteresting, essentially because in a world where all parties may be satisfied completely, the obvious solution is to do so (and this solution is in fact an equilibrium). Thus unfairness and bias issues logically only arise in a world of *conflict* (e.g., in zero-sum settings, or under limited resources constraints, which foster *competition* between groups). Thus while the realizable case is straightforward and solvable, we argue that in practice, the agnostic learning setting is far more relevant.

5.2 Characterizing Fair Statistical Learnability with FPAC-Learners

We first consider only questions of *statistical learning.* In other words, we ignore computation for now, and show only that *there exist* fair-PAC-learning algorithms of unbounded runtime. In particular, we show a generalization of the *fundamental theorem of statistical learning* to fair learning problems. The aforementioned result relates *uniform convergence* and *PAC-learnability*, and is generally stated for binary classification only. We define a natural generalization of uniform convergence to arbitrary learning problems within our framework, and then show conditions under which a generalized fundamental theorem of (fair) statistical learning holds. In particular, we show that, neglecting computational concerns, PAC-learnability and fair-PAC learnability are equivalent for learning problems with a particular *no-free-lunch* guarantee.

We first define a *generalized notion* of *uniform convergence* as defined by [Shalev-Shwartz and Ben-David, 2014]. In particular, our definition applies to *any bounded loss function*,[1] thus greatly generalizes the standard notion for binary classification.

Definition 5.2.1 (Uniform Convergence)**.** Suppose $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, r] \subseteq \mathbb{R}$ and hypothesis class $\mathcal{H} \subseteq \mathcal{X} \to \mathcal{Y}$. We say $(\mathcal{H}, \ell) \in \mathrm{UC}$ if

$$\lim_{m \to \infty} \sup_{\mathcal{D} \text{ over } \mathcal{X} \times \mathcal{Y}} \mathop{\mathbb{E}}_{\boldsymbol{z} \sim \mathcal{D}^m} \left[\sup_{h \in \mathcal{H}} \left| \hat{\mathrm{R}}(h; \ell, \boldsymbol{z}) - \mathrm{R}(h; \ell, \mathcal{D}) \right| \right] = 0 \enspace .$$

Observation 5.2.2. We stress that this definition is both uniform over ℓ composed with the *hypothesis class* $\mathcal{H}$ and uniform over *all possible distributions* $\mathcal{D}$. The classical definition of *uniform convergence in probability* applies to a singular $\mathcal{D}$, however it is standard in PAC-learning and VC theory to assume uniformity over $\mathcal{D}$, so we adopt this convention.

Standard uniform convergence definitions consider only the convergence of *empirical frequencies* of events to their *true frequencies*, whereas we generalize to consider uniform convergence of the *empirical means* of functions to their *expected values*.

In discussing uniform convergence, it is often necessary to consider not the loss function or hypothesis class *in isolation*, but their *composition*, defined as

$$\forall h \in \mathcal{H}: \; (\ell \circ h)(x, y) \doteq \ell(y, h(x)) \quad \& \quad \ell \circ \mathcal{H} \doteq \{ \ell \circ h \,|\, h \in \mathcal{H} \} \enspace .$$

It is also helpful to consider the *sample complexity* of ε-δ uniform-convergence, where we take

$$\mathrm{m}_{\mathrm{UC}}(\ell \circ \mathcal{H}, \varepsilon, \delta) \doteq \operatorname{argmin} \left\{ m \;\middle|\; \sup_{\mathcal{D} \text{ over } \mathcal{X} \times \mathcal{Y}} \mathbb{P} \left(\sup_{h \in \mathcal{H}} \left| \mathop{\mathbb{E}}_{\mathcal{D}}[\ell \circ h] - \mathop{\hat{\mathbb{E}}}_{\boldsymbol{z} \sim \mathcal{D}^m} [\ell \circ h] \right| > \varepsilon \right) \leq \delta \right\} \enspace ,$$

i.e., the minimum sufficient sample size to ensure ε-δ uniform-convergence over the *loss family* $\ell \circ \mathcal{H}$.

It is in general true that *uniform convergence* implies *PAC-learnability*; this is well-known for binary

[1]Boundedness should not be strictly necessary for learnability even uniform convergence, but vastly simplifies all aspects of the analysis. In many cases, it can be relaxed to moment-conditions, such as *sub-Gaussian* or *sub-exponential* assumptions.

classification, but we show the generalized result for completeness. The converse is true for some learning problems, but not for others, which we shall use as a powerful tool to characterize when PAC-learnability implies fair-PAC-learnability. We first state the fundamental theorem of statistical learning (for classification) [Shalev-Shwartz and Ben-David, 2014, theorems 6.2, 29.3].

Theorem 5.2.3 (Fundamental Theorem of Statistical Learning (Classification))**.** Suppose ℓ is the 0-1 loss for k-class classification, where $k < \infty$. Then the following are equivalent.

1. $\forall d \in \mathbb{N}$: $\mathcal{H}_d$ has finite Natarajan-dimension (= VC dimension for $k = 2$ classes).
2. $\forall d \in \mathbb{N}$: $(\ell, \mathcal{H}_d)$ has the uniform convergence property.
3. Any ERM rule is a successful agnostic-PAC learner for $\mathcal{H}$.
4. $\mathcal{H}$ is agnostic-PAC learnable.
5. Any ERM rule is a successful realizable-PAC learner for $\mathcal{H}$.
6. $\mathcal{H}$ is realizable-PAC learnable.

It is somewhat subtle to generalize this result to arbitrary learning problems. In particular, there are PAC-learnable problems for which uniform convergence *does not hold*. However, Alon et al. [1997] show similar results for various regression problems, with the (scale-sensitive) γ*-fat-shattering dimension* playing the role of the Vapnik-Chervonenkis or Natarajan dimension in classification.

Theorem 5.2.4 (Fundamental Theorem of Fair Statistical Learning)**.** Suppose ℓ such that $\forall \mathcal{H} : (\mathcal{H}, \ell) \in \mathrm{PAC}^{\mathrm{Rlz}} \Rightarrow (\mathcal{H}, \ell) \in \mathrm{UC}$. Then, for any hypothesis class sequence $\mathcal{H}$, the following are equivalent:

1. $\forall d \in \mathbb{N}$: $(\ell, \mathcal{H}_d)$ has the (generalized) uniform convergence property.
2. Any EMM rule is a successful agnostic-fair-PAC learner for $(\ell, \mathcal{H})$.
3. $(\ell, \mathcal{H})$ is agnostic-fair-PAC learnable.
4. Any EMM rule is a successful realizable-fair-PAC learner for $(\ell, \mathcal{H})$.
5. $(\ell, \mathcal{H})$ is realizable-fair-PAC learnable.

Proof. First note that $1 \Rightarrow 2$ is a rather straightforward consequence of the definition of uniform convergence and the contraction property of fair malfare functions (theorem 3.3.7 item 3). In particular, take $m \doteq \mathrm{m_{UC}}(\ell \circ \mathcal{H}_d, \frac{\varepsilon}{2}, \frac{\delta}{g})$. By union bound, this implies that with probability at least $1 - \delta$, taking samples $\boldsymbol{z}_{1:g,1:m} \sim \mathcal{D}_1^m \times \cdots \times \mathcal{D}_g^m$, we have

$$\forall i \in \{1, \ldots, g\} : \sup_{h \in \mathcal{H}_d} \left| \mathrm{R}(h; \ell, \mathcal{D}_i) - \hat{\mathrm{R}}(h; \ell, \boldsymbol{z}_i) \right| \leq \frac{\varepsilon}{2} \ .$$

Consequently, as $\mathrm{M}(\cdot;\boldsymbol{w})$ is 1-$\|\cdot\|_\infty$-$|\cdot|$ in risk (see lemma 3.4.1), it holds with probability at least $1-\delta$ that

$$\forall h \in \mathcal{H}_d : \left| \mathrm{M}\big(i \mapsto \hat{\mathrm{R}}(h;\ell,\boldsymbol{z}_i);\boldsymbol{w}\big) - \mathrm{M}\big(i \mapsto \mathrm{R}(h);\ell,\mathcal{D}_i);\boldsymbol{w}\big) \right| \leq \frac{\varepsilon}{2} \enspace .$$

Now, for EMM-optimal $\hat{h}$, and malfare-optimal h^*, we apply this result twice to get

$$\begin{aligned}
\mathrm{M}\big(i \mapsto \mathrm{R}(\hat{h};\ell,\mathcal{D}_i);\boldsymbol{w}\big) &\leq \mathrm{M}\big(i \mapsto \hat{\mathrm{R}}(\hat{h};\ell,\boldsymbol{z}_i);\boldsymbol{w}\big) + \tfrac{\varepsilon}{2} \\
&\leq \mathrm{M}\big(i \mapsto \hat{\mathrm{R}}(h^*;\ell,\boldsymbol{z}_i);\boldsymbol{w}\big) + \tfrac{\varepsilon}{2} \\
&\leq \mathrm{M}\big(i \mapsto \mathrm{R}(h^*;\ell,\mathcal{D}_i);\boldsymbol{w}\big) + \varepsilon \enspace .
\end{aligned}$$

Therefore, under uniform convergence, the EMM algorithm agnostic fair-PAC learns $(\mathcal{H},\ell)$ with finite sample complexity $\mathrm{m}_{\mathcal{A}}(\varepsilon,\delta,d,g) = g \cdot \mathrm{m}_{\mathrm{UC}}(\ell \circ \mathcal{H}_d, \frac{\varepsilon}{2}, \frac{\delta}{g})$, completing $1 \Rightarrow 2$.

Now, observe that $2 \Rightarrow 3$ and $4 \Rightarrow 5$ are almost tautological: the existence of (agnostic / realizable) fair-PAC learning algorithms imply (agnostic / realizable) fair-PAC learnability.

Now, $2 \Rightarrow 4$ and $3 \Rightarrow 5$ hold, as realizable learning is a special case of agnostic learning.

As 1 implies 2-4, which in turn each imply 5, it remains only to show that $5 \Rightarrow 1$, i.e., if $\mathcal{H}$ is realizable fair PAC learnable, then $\mathcal{H}$ has the uniform convergence property. In general, the question is rather subtle, but here the assumption "suppose ℓ such that $(\mathcal{H},\ell) \in \mathrm{PAC}^{\mathrm{Rlz}} \Rightarrow (\mathcal{H},\ell) \in \mathrm{UC}$" does most of the work. In particular, as PAC-learning is a special case of FPAC-learning, we have

$$(\mathcal{H},\ell) \in \mathrm{FPAC}^{\mathrm{Agn}} \Rightarrow (\mathcal{H},\ell) \in \mathrm{PAC}^{\mathrm{Agn}} \enspace ,$$

then applying the assumption yields $(\mathcal{H},\ell) \in \mathrm{UC}$. □

The reductions and equivalences that compose this result are graphically depicted in figure 5.1.

Observation 5.2.5 (The Gap Between Uniform Convergence and (fair) PAC-Learnability)**.** Note that the assumption "suppose ℓ such that $(\mathcal{H},\ell) \in \mathrm{PAC}^{\mathrm{Rlz}} \Rightarrow (\mathcal{H},\ell) \in \mathrm{UC}$" does not in general hold. In many cases of interest, it is known to hold, e.g., finite-class classification under 0-1 loss, and bounded regression under square and absolute loss.

In general, verifying this condition is a rather subtle task, and until more general techniques are developed, must be repeated for each learning problem (loss function). Regardless, this lies beyond the scope of this

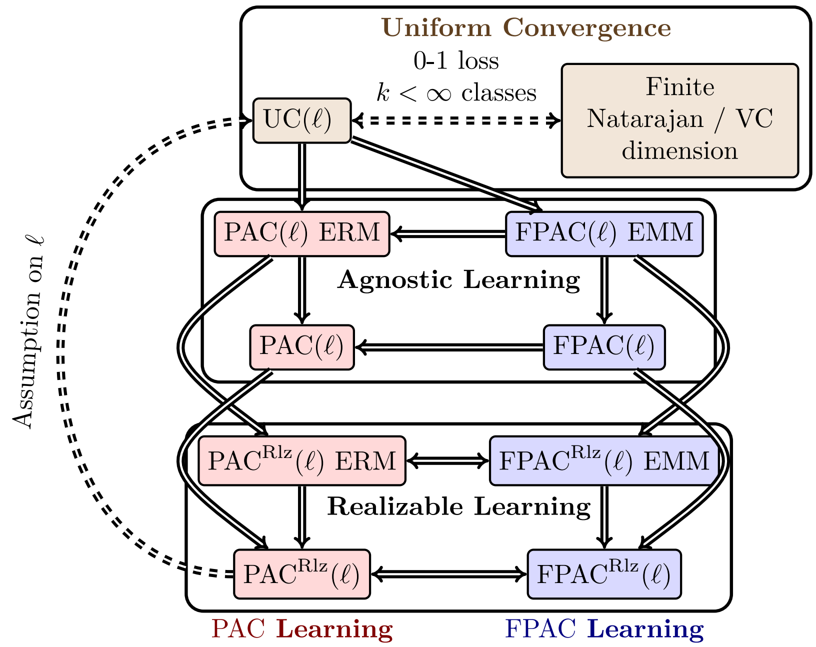


Figure 5.1: Implications between membership in PAC and fair-PAC classes. In particular, for arbitrary fixed ℓ, implication denotes *implication of membership* of some $\mathcal{H}$ (i.e., containment); see theorems 5.2.4 and 5.1.6. Dashed implication arrows hold conditionally on ℓ. Note that when the assumption on ℓ (see theorem 5.2.4) holds, the hierarchy collapses, and in general, under realizability, some classes are known to coincide.

manuscript, as we have reduced the question of *fair*-PAC learnability to one of (generic) PAC-learnability.

5.3 Characterizing Computational Learnability with Efficient FPAC Learners

In this section, we consider the more granular question of whether FPAC learning is computationally harder than PAC learning. In other words, where previously we showed conditions under which PAC = FPAC, here we focus on the subset of models with polynomial time training efficiency guarantees, i.e., $\mathrm{PAC}_{\mathrm{Poly}} = \mathrm{FPAC}_{\mathrm{Poly}}$. Theorem 5.1.6 has already characterized the computational complexity of *realizable* FPAC-learning, so we now focus on the agnostic case. In the *agnostic case*, we show neither a generic reduction or non-constructive proof that $\mathrm{PAC}_{\mathrm{Poly}} = \mathrm{FPAC}_{\mathrm{Poly}}$, nor do we show a counterexample; rather we leave this question for future work. We do, however, show that under conditions commonly leveraged as sufficient for efficient PAC-learning, so too is efficient FPAC-learning possible. In particular, section 5.3.1 provides an efficient constructive reduction (i.e., an algorithm) for efficient FPAC-learning under standard *convex optimization* settings, and section 5.3.2 shows the same when a small cover of $\mathcal{H}$ exists and may be efficiently enumerated. The results of this section are summarized graphically in figure 5.2.

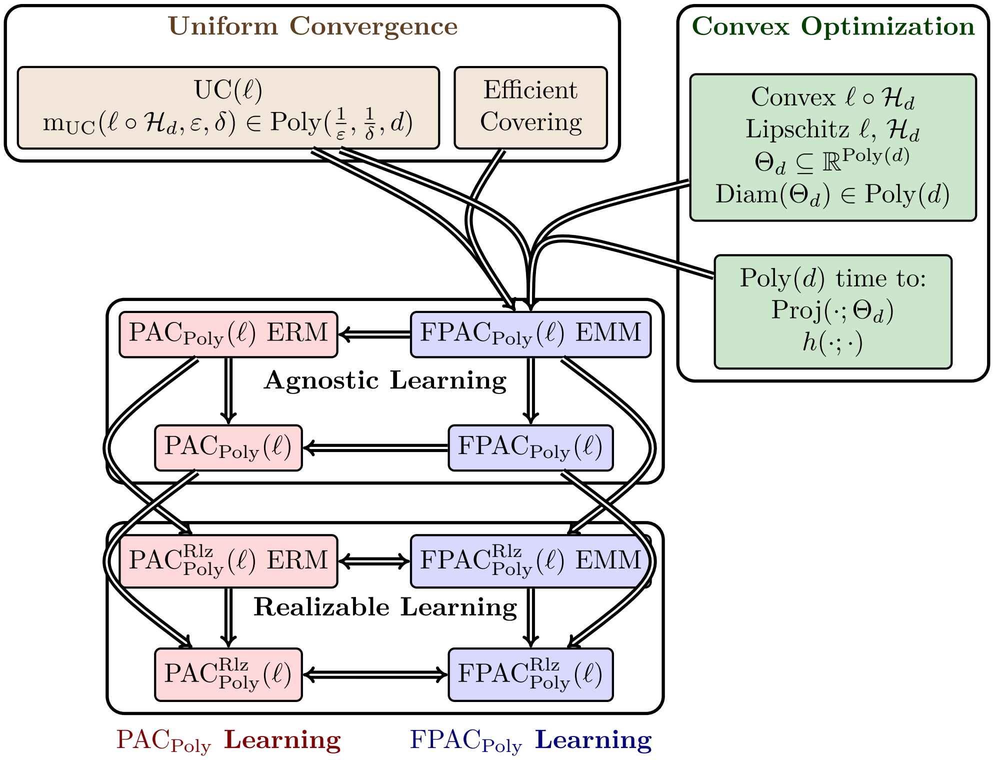


Figure 5.2: Implications between membership in various poly-time PAC and fair-PAC classes. In particular, for arbitrary but fixed ℓ, implication denotes *implication of membership* of some $\mathcal{H}$ (i.e., containment). See theorems 5.3.1 and 5.3.2.

In both the convex optimization and efficient enumeration settings, the proofs take the same general form: we show that ε-*approximate* EMM on m total samples is computationally efficient (in $\mathrm{Poly}(m, \varepsilon, d)$ time), and then argue that so long as *sample complexity* of FPAC-learning via EMM is polynomial, then we may construct an FPAC learner using ε-approximate EMM with polynomial *time complexity*. In particular, the proofs simply account for *optimization* and *sampling* error, and in both cases construct efficient FPAC-learners.

5.3.1 Efficient FPAC Learning with Convex Optimization

Here we show concretely and constructively the existence of fair-PAC-learners under standard convex optimization assumptions via the *subgradient method* [Shor, 2012], with constants fully derived. Sharper analyses are of course possible, and potential improvements are discussed subsequently, but our result is immediately practical and can be applied verbatim to problems like *generalized linear models* [Nelder and Wedderburn, 1972] and many *kernel methods* with little analytical effort.

Theorem 5.3.1 (Efficient FPAC Learning via Convex Optimization)**.** Suppose each hypothesis space $\mathcal{H}_d \in \mathcal{H}$ is indexed by $\Theta_d \subseteq \mathbb{R}^{\mathrm{Poly}(d)}$, i.e., $\mathcal{H}_d = \{h(\cdot;\theta) \mid \theta \in \Theta_d\}$, s.t. (Euclidean) $\mathrm{Diam}(\Theta_d) \in \mathrm{Poly}(d)$, and

Algorithm 2 Approximate Empirical Malfare Minimization via the Subgradient Method

procedure $\mathcal{A}_{\mathrm{PSG}}(\ell, \mathcal{H}, \theta_0, \mathrm{m}_{\mathrm{UC}}(\cdot,\cdot), \mathcal{D}_{1:g}, \boldsymbol{w}, \mathrm{M}(\cdot;\cdot), \varepsilon, \delta)$

  Input: λ_ℓ-Lipschitz loss function ℓ, $\lambda_{\mathcal{H}}$-Lipschitz hypothesis class $\mathcal{H}$ with parameter space Θ, initial guess $\theta_0 \in \Theta$, uniform-convergence sample-complexity bound $\mathrm{m}_{\mathrm{UC}}(\cdot,\cdot)$, group distributions $\mathcal{D}_{1:g}$, group weights $\boldsymbol{w}$, malfare function $\mathrm{M}(\cdot;\cdot)$, solution optimality guarantee ε-δ.

  Output: ε-δ-$\mathrm{M}(\cdot;\cdot)$-optimal $h \in \mathcal{H}$ (under the conditions of theorem 5.3.1).

  $\mathrm{m}_{\mathcal{A}} \leftarrow \mathrm{m}_{\mathrm{UC}}(\frac{\varepsilon}{3}, \frac{\delta}{g})$ ▷ Determine sufficient sample size

  $\boldsymbol{z}_{1:g,1:\mathrm{m}_{\mathcal{A}}} \sim \mathcal{D}_1^{\mathrm{m}_{\mathcal{A}}} \times \cdots \times \mathcal{D}_g^{\mathrm{m}_{\mathcal{A}}}$ ▷ Draw training sample for each group

  $n \leftarrow \left\lceil \left(\frac{3\,\mathrm{Diam}(\Theta)\lambda_\ell\lambda_{\mathcal{H}}}{\varepsilon} \right)^2 \right\rceil$ ▷ Iteration count

  $\alpha \leftarrow \frac{\mathrm{Diam}(\Theta)}{\lambda_\ell\lambda_{\mathcal{H}}\sqrt{n}}$ ▷ Learning rate ($\approx \frac{\varepsilon}{3\lambda_\ell^2\lambda_{\mathcal{H}}^2}$)

  $f(\theta) : \Theta \mapsto \mathbb{R}_{0+} \doteq \mathrm{M}\big(i \mapsto \hat{\mathrm{R}}(h(\cdot;\theta); \ell, \boldsymbol{z}_i); \boldsymbol{w}\big)$ ▷ Define empirical malfare objective

  $\hat{\theta} \leftarrow \textsc{ProjectedSubgradient}(f, \Theta, \theta_0, n, \alpha)$ ▷ Run PSG algorithm on empirical malfare

  return $h(\cdot;\hat{\theta})$ ▷ Return ε-δ optimal model

end procedure

$\forall x \in \mathcal{X}, \theta \in \Theta_d$, $h(x;\theta)$ can be evaluated in $\mathrm{Poly}(d)$ time, and $\tilde{\theta} \in \mathbb{R}^{\mathrm{Poly}(d)}$ can be Euclidean-projected onto Θ_d in $\mathrm{Poly}(d)$ time. Suppose also ℓ such that $\forall x \in \mathcal{X}, y \in \mathcal{Y}:\ \theta \mapsto \ell(y, h(x;\theta))$ is a *convex function*, and suppose Lipschitz constants $\lambda_\ell, \lambda_{\mathcal{H}} \in \mathrm{Poly}(d)$ and some norm $\|\cdot\|_{\mathcal{Y}}$ over $\mathcal{Y}$ s.t. ℓ is λ_ℓ-$\|\cdot\|_{\mathcal{Y}}$-$|\cdot|$-Lipschitz in $\hat{y}$, i.e.,

$$\forall y, \hat{y}, \hat{y}' \in \mathcal{Y}:\ \big|\ell(y,\hat{y}) - \ell(y,\hat{y}')\big| \leq \lambda_\ell \big\|\hat{y} - \hat{y}'\big\|_{\mathcal{Y}} \enspace ,$$

and also that each $\mathcal{H}_d$ is $\lambda_{\mathcal{H}}$-$\|\cdot\|_2$-$\|\cdot\|_{\mathcal{Y}}$-Lipschitz in θ, i.e.,

$$\forall x \in \mathcal{X}, \theta, \theta' \in \Theta_d:\ \big\|h(x;\theta) - h(x;\theta')\big\|_{\mathcal{Y}} \leq \lambda_{\mathcal{H}} \big\|\theta - \theta'\big\|_2 \enspace .$$

Finally, assume $\ell \circ \mathcal{H}_d$ exhibits ε-δ *uniform convergence* with sample complexity $\mathrm{m}_{\mathrm{UC}}(\varepsilon, \delta, d) \in \mathrm{Poly}(\frac{1}{\varepsilon}, \frac{1}{\delta}, d)$. It then holds that, for arbitrary initial guess $\theta_0 \in \Theta_d$, for any group distributions $\mathcal{D}_{1:g}$, group weights $\boldsymbol{w}$, and fair malfare function $\mathrm{M}(\cdot;\cdot)$, the algorithm (see algorithm 2)

$$\mathcal{A}(\mathcal{D}_{1:g}, \boldsymbol{w}, \mathrm{M}(\cdot;\cdot), \varepsilon, \delta, d) \doteq \mathcal{A}_{\mathrm{PSG}}\big(\ell, \mathcal{H}_d, \theta_0, \mathrm{m}_{\mathrm{UC}}(\cdot,\cdot,d), \mathcal{D}_{1:g}, \boldsymbol{w}, \mathrm{M}(\cdot;\cdot), \varepsilon, \delta\big)$$

fair-PAC-learns $(\mathcal{H}, \ell)$ with sample complexity $\mathrm{m}(\varepsilon, \delta, d, g) = g \cdot \mathrm{m}_{\mathrm{UC}}(\frac{\varepsilon}{3}, \frac{\delta}{g}, d)$, and (training) time-complexity $\in \mathrm{Poly}(\frac{1}{\varepsilon}, \frac{1}{\delta}, d, g)$, thus $(\mathcal{H}, \ell) \in \mathrm{FPAC}^{\mathrm{Agn}}_{\mathrm{Poly}}$.

It is of course possible to show similar guarantees under relaxed conditions. and with sharper sample complexity and time complexity bounds. The above result merely characterizes a simple and standard convex optimization setting under which standard convex optimization risk minimization guarantees readily translate to malfare minimization.

5.3.2 Uniform Convergence and Efficient Covering

As we have seen in section 5.2, uniform convergence implies, and is often equivalent to, (fair) PAC-learnability. However, these results all consider only *statistical learning*, and to analyze *computational learning* questions, we introduce a strengthening of uniform convergence that considers *computation*. We now show sufficient conditions for polynomial-time fair PAC-learnability via *covering numbers*, which we use both to show uniform convergence and to construct an efficient training algorithm.

In particular, we show that if a *polynomially-large cover* of each $\ell \circ \mathcal{H}_d$ exists and can be efficiently enumerated, then $(\mathcal{H}, \ell) \in \mathrm{FPAC}_{\mathrm{Poly}}$. In what follows, an ℓ_2-γ-empirical-cover of loss family $(\ell \circ \mathcal{H}_d) \subseteq \mathcal{X} \mapsto \mathbb{R}_{0+}$ on a sample $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^m$ is any $\mathcal{H}_{d,\gamma}$ such that

$$\forall h \in \mathcal{H}_d : \min_{h_\gamma \in \mathcal{H}_{d,\gamma}} \sqrt{\frac{1}{m} \sum_{i=1}^{m} \big((\ell \circ h)(\boldsymbol{z}_i) - (\ell \circ h_\gamma)(\boldsymbol{z}_i) \big)^2} \leq \gamma \ .$$

We take $\mathcal{C}(\ell \circ \mathcal{H}_d, \boldsymbol{z}, \gamma)$ to denote such a cover, and $\mathcal{C}^*(\ell \circ \mathcal{H}_d, \boldsymbol{z}, \gamma)$ to denote such a cover *of minimum cardinality.* Finally, we define the *uniform covering numbers*

$$\mathcal{N}(\ell \circ \mathcal{H}_d, m, \gamma) \doteq \sup_{\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^m} \left| \mathcal{C}^*(\ell \circ \mathcal{H}_d, \boldsymbol{z}, \gamma) \right| \quad \& \quad \mathcal{N}(\ell \circ \mathcal{H}_d, \gamma) \doteq \sup_{m \in \mathbb{N}} \mathcal{N}(\ell \circ \mathcal{H}_d, m, \gamma) \ .$$

This concept is crucial to both our *uniform convergence* and *optimization efficiency* guarantees. In particular, our construction ensures that $\mathcal{N}(\ell \circ \mathcal{H}_d, m, \gamma)$ is sufficiently small so as to ensure *polynomial training time* on a *polynomially-large* training sample is sufficient to FPAC-learn $(\ell, \mathcal{H})$.

With this exposition complete, the FPAC-learning algorithm we present is simply EMM on a cover $\hat{\mathcal{C}}(\ell \circ \mathcal{H}_d, \boldsymbol{z}, \gamma)$ that is not superpolynomially larger than $\mathcal{N}(\ell \circ \mathcal{H}_d, m, \gamma)$, which our construction ensures exists and may be efficiently enumerated, and upon which we guarantee that a polynomially-large training sample is sufficient. We now state the result, noting that full derivation with exact constants appears in the appendix.

Theorem 5.3.2 (Efficient FPAC-Learning by Covering)**.** Suppose loss function ℓ of bounded codomain (i.e., $\|\ell\|_\infty$ is bounded), and hypothesis class sequence $\mathcal{H}$, s.t. $\forall m, d \in \mathbb{N}$, $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^m$, there exist

1. a γ-ℓ_2 cover $\mathcal{C}^*(\ell \circ \mathcal{H}_d, \boldsymbol{z}, \gamma)$, where $\left|\mathcal{C}^*(\ell \circ \mathcal{H}_d, \boldsymbol{z}, \gamma)\right| \leq \mathcal{N}(\ell \circ \mathcal{H}_d, \gamma) \in \mathrm{Poly}(\frac{1}{\gamma}, d)$; and
2. an algorithm to enumerate a γ-ℓ_2 cover $\hat{\mathcal{C}}(\ell \circ \mathcal{H}_d, \boldsymbol{z}, \gamma)$ of size $\mathrm{Poly}\, \mathcal{N}(\ell \circ \mathcal{H}_d, m, \gamma)$ in $\mathrm{Poly}(m, \frac{1}{\gamma}, d)$ time.

Then $(\ell, \mathcal{H}) \in \mathrm{FPAC}_{\mathrm{Poly}}$, where (1) is sufficient to show that $(\ell, \mathcal{H})$ is fair-PAC learnable with polynomial *sample complexity*, and (2) is required only to show polynomial *training time complexity.*

This immediately implies that fixed $\mathcal{H}$ that are *finite*, or of bounded *pseudodimension* or γ*-fat-shattering dimension*[2] are fair-PAC-learnable. For instance, this includes *classifiers* such as *all possible languages* of *Boolean formulae* over (constant) d variables, or *halfspaces* (i.e., linear hard classifiers $\mathcal{H} \doteq \{\boldsymbol{x} \mapsto \operatorname{sgn}(\boldsymbol{x} \cdot \boldsymbol{w}) \mid \boldsymbol{w} \in \mathbb{R}^d\}$), as well as GLM, subject to regularity constraints to appropriately control the loss function. However, it is perhaps not as powerful as it appears; it applies to *fixed hypothesis classes*, thus each of the above linear models over $\mathbb{R}^d$ is efficiently fair-PAC learnable, but it says nothing about their performance as $d \to \infty$.

This is essentially because the *statistical analysis* to show polynomial *sample complexity* requires only that $\ln \mathcal{N}(\ell \circ \mathcal{H}_d, \gamma) \in \operatorname{Poly}(\frac{1}{\gamma}, d)$, whereas our *training algorithm* must actually *enumerate* an empirical cover, which yields the (exponentially) stronger requirement that $\mathcal{N}(\ell \circ \mathcal{H}_d, m, \gamma) \in \operatorname{Poly}(\frac{1}{\gamma}, d)$ for polynomial *time complexity*. Indeed, we see that while the covering numbers we assume imply uniform convergence with *sample complexity* polynomial in d, when covering numbers grow exponentially in d, then our algorithm yields only *exponential* time complexity in d. Consequently, the result only implies polynomial-time algorithms w.r.t. sequences that grow slowly in complexity; e.g., sequences of linear classifiers that grow only *logarithmically* in dimension $\mathcal{H}_d \doteq \{\boldsymbol{x} \mapsto \operatorname{sgn}(\boldsymbol{x} \cdot \boldsymbol{w}) \mid \boldsymbol{w} \in \mathbb{R}^{\lfloor \ln d \rfloor}\}$.

Note also that theorem 5.3.2 leverages *covering arguments* in both their *statistical* and *computational* capacities. Statistical bounds based on covering are generally well-regarded, particularly when strong analytical bounds on covering numbers are available, although sharper results are possible (e.g., through the *entropy integral* or *majorizing measures*). Furthermore, while we do construct a *polynomial time* training algorithm, in many cases, specific optimization methods (e.g., stochastic gradient descent or Newton's method) exist to perform EMM *more efficiently* and with *higher accuracy*. Worse yet, *efficient enumerability* of a cover may be non-trivial in some cases; while most covering arguments in the wild are either constructive, or compositional to the point where each component can easily be constructed, it may hold for some problems that computing or enumerating a cover is computationally prohibitive.

On Compositionality and Coverability Conditions The covers and covering numbers discussed above are of course properties of each $\ell \circ \mathcal{H}_d$, rather than ℓ and each $\mathcal{H}_d$ individually. This creates proof obligation for each loss function of interest, in contrast to theorem 5.3.1, wherein only Lipschitz continuity of ℓ is assumed, and the remaining analysis is on $\mathcal{H}$. Fortunately, in many cases it is still possible to analyze covers of each $\mathcal{H}_d$ in isolation, and then draw conclusions across a broad family of ℓ composed with each $\mathcal{H}_d$. In particular, via standard properties of covering numbers, if ℓ is Lipschitz-continuous w.r.t. some pseudonorm

[2]The reader is invited to consult [Anthony and Bartlett, 2009] for an encyclopedic overview of various combinatorial dimensions, associated covering-number and shattering-coefficient concepts, and their applications to statistical learning theory.

$\|\cdot\|_{\mathcal{Y}}$ over $\mathcal{Y}$, and γ-ℓ_2 covering numbers of each $\mathcal{H}_d$ w.r.t. $\|\cdot\|_{\mathcal{Y}}$ are well-behaved, it can be shown that the conditions of theorem 5.3.2 are met. This is useful as, for example, regression losses like *square error*, *absolute error*, and *Huber loss* are all Lipschitz continuous on bounded domains, and thus analysis on each $\mathcal{H}_d$ alone is sufficient to apply theorem 5.3.2 with each such loss function.

5.4 Conclusion

The central thrust of this paper introduces malfare minimization as a fair learning task, and shows relationships between the statistical and computational issues of malfare and risk minimization. In particular, we argue that our method is more in line with welfare-centric machine learning theory than demographic-parity theory, however malfare is better aligned to address machine learning tasks cast as loss minimization problems than is welfare. As such, the first half of this manuscript is dedicated to deriving and motivating malfare minimization, while the latter studies the problem from statistical and computational learning theoretic perspectives.

5.4.1 Contrasting Malfare and Welfare

With our framework now fully laid out and initial results presented, we now contrast our malfare-minimization framework with traditional welfare-maximization approaches in greater detail. We do not claim that malfare is a better or more useful concept than welfare; but rather we argue only that it is *significantly different* (with surprising non-equivalence results between power-mean welfare and malfare functions), stands on an equal axiomatic footing, and it stands to reason that the right tool (malfare) should be used for the right job (fair loss minimization). With this said, we acknowledge that some learning tasks, e.g., bandit problems and reinforcement learning tasks, are more naturally phrased as *maximizing* utility or (discounted) reward. However, with a few exceptions, e.g., the *spherical scoring rule* from decision theory, most supervised learning problems are naturally cast as minimizing nonnegative *loss functions* (arguably via cross-entropy or KL-divergence minimization through maximum-likelihood, either as explicitly intended [Nelder and Wedderburn, 1972], or *ex-post-facto* through subsequent analysis [Cousins and Riondato, 2019]).

We are highly interested in exploring a parallel theory of fair welfare optimization, however some key malfare properties do not hold for welfare. In particular, fair welfare functions $\mathrm{W}_p(\cdot;\cdot)$ for $p \in [0,1)$ are not Lipschitz continuous; for example, the *Nash social welfare* (a.k.a. *unweighted geometric welfare*) $\mathrm{W}_0(\mathcal{S}; \omega \mapsto \frac{1}{g}) = \sqrt[g]{\prod_{i=1}^{g} \mathcal{S}_i}$ is unstable to perturbations of each $\mathcal{S}_i$ around 0, which causes difficulty in both the *statistical* and *computational* aspects of learning. Leveraging this fact, it is trivial to construct welfare-maximization learning problems for which sample complexity is unbounded, making straightforward translation of our FPAC framework into a welfare setting rather vacuous, except in contrived, trivial, or degenerate cases. Essentially, this is because while corollary 3.4.2 holds for both welfare and malfare, it does not imply *uniform sample-complexity* bounds, whereas, such bounds are trivial for fair malfare (see lemma 3.4.1), due to the *contraction property* (theorem 3.3.7 item 3). It thus seems such a theory would need either to either impose additional assumptions to avoid non-Lipschitz behavior (e.g., artificially limit the permitted range of p), or otherwise provide weaker (non-uniform) learning guarantees.

5.4.2 FPAC Learning: Contributions and Open Questions

After motivating the malfare-minimization machine learning setting, we introduce fair PAC-learning to study the statistical and computational difficulty of malfare minimization. As a generalization of PAC-learning, known hardness results (e.g., lower-bounds on computational and sample complexity of loss minimization) immediately apply, thus, coarsely speaking, the interesting question is whether, for some tasks, malfare minimization is harder than risk minimization. Theorem 5.2.4 answers this question in the negative for *sample complexity*, under appropriate conditions on the loss function, as does theorem 5.1.6 under realizability. The remaining cases are left open, though section 5.3 at least shows that many conditions sufficient for PAC-learnability are also sufficient for FPAC-learnability.

We are optimistic that our FPAC-learning definitions will motivate the community to further pursue the deep connections between various PAC and FPAC learning settings, as well as promote cross-pollination between computational learning theory and fair machine learning research. We believe that deeper inquiry into these questions will lead to both a better understanding of what is and is not FPAC-learnable, as well as more practical and efficient reductions and FPAC-learning algorithms.

Part III

Sample-Efficient Mean-Estimation with Dependent Sequential Data

In this part, I examine sampling and data science problems where we do not assume our setting can be reduced to an independently and identically distributed (i.i.d.) statistical estimation task. As the analysis of non-i.i.d. settings requires far more sophisticated techniques than i.i.d. settings, I focus on the simpler statistical problem of *mean estimation*, rather than uniform convergence bounds.[3] I am confident that our methods can be generalized to handle more sophisticated concentration-of-measure techniques, such as Rademacher averages, though with significantly more analytical complexity than in the i.i.d. case.

In particular, I consider two settings where the data are somehow temporally or sequentially dependent (e.g., *correlated*), and we still wish to recover guarantees similar to those in the i.i.d. case. While the examples I give are for *sampling* and *data science* problems, there is no reason that similar methods can not be applied to machine learning tasks and other application domains.

As in the machine learning settings of parts I and II, where data dependent bounds are desirable to leverage the structure of the data-distribution, without unreasonably assuming *a priori* knowledge of properties like variances, I wish to show similar guarantees (w.h.p.) while assuming as little as possible *a priori*, and providing efficient algorithms, with sample complexity bounds comparable to those attainable assuming additional information were available. This is very important in practical sampling problems, as the reason we are sampling in the first place is generally that we know very little about a distribution, and it is infeasible to compute such properties exactly, so assuming known properties greatly hinders the practical usability of such methods.

Two Settings of Non-I.I.D. Mean Estimation I now describe our two settings, while aggressively focusing on their similarities, perhaps to the point of extreme reductionism. Note that despite their intuitive similarities, and the great similarity between the algorithms and sample complexity guarantees they admit, the subtle technical differences between these results are quite substantial. In particular, results in the setting are essentially self-contained, depending primarily on known bounds in the i.i.d. setting, while the second has deep roots in recent developments in the theory of concentration inequalities for Markov chains.

I first consider an important setting in databases and data analysis, where data are stored in such a way that *random access* is expensive, but access to contiguous blocks of data is cheap, e.g., in *sequentially accessible* media. This includes many physical media, such as *magnetic tapes*, *LaserDiscs* (and other optical media), *spinning magnetic disk drives*, and *wax cylinders*, but also more generally, in modern digital systems, including in blocks stored across multiple *files*, *drives*, or *servers* across a network. In each of these settings, there is a

[3]Do note that elementary machine learning guarantees for finite families, early VC theoretic bounds, and to some extent covering number analysis methods are all based on simple mean-estimation, making these results applicable in these settings as well.

simple and efficient access pattern for the data, however standard sampling methods require *random access*, which may be orders of magnitude slower.

I give an algorithm that is able to leverage this database structure by sampling large (but dependent) blocks, at cost comparable to drawing the same quantity of independent samples. In the worst case, samples within blocks are *highly correlated*, and the method is no better than if one sample were drawn from each block. However, often, in particular when samples within each block are actually independent, the sample complexity of my method nearly matches that of methods that assume each sample in each block were drawn independently.

In the second setting, I adopt a widely-used theoretical model of dependent processes, namely, I assume we wish to sample to estimate the expectation of a *function* f over the *stationary distribution* of a *Markov chain* $\mathcal{M}$. Here I continue the theme of *approximate independence*, where now rather than sampling blocks with independence between them and dependence within them, we draw a long chain of samples, where nearby samples may be dependent, but the chain is guaranteed to *mix*, in the sense that sufficiently-distant samples are *nearly independent*. Precise definitions of *mixing* and *mixing time*, as well as the closely-related *relaxation time*, are given in chapter 7, but for now it suffices to understand that there exists some T such that samples at least T steps apart in the chain behave near-independently for the purposes of mean-estimation.

Preview of the Method At a high level, both settings are the same, in the sense that we have an *a priori* guarantee of *approximate independence*, i.e., between blocks, or across large gaps in Markov chains. In fact, both have a parameter T describing the degree of *known independence*, i.e., the *size of each block*, or the *relaxation time* (which, roughly describes the length between samples in the chain for them to behave *near-independently* for *mean estimation*). They also may exploit structure in the potentially-dependent space between these near-independent milestones, which yields stronger guarantees when more independence exists in the data (distribution) than would be implied by the *a priori* guarantees.

Upon closer examination, one may note that in the block sampling setting, we have *complete independence* across spans of T records, whereas in the Markov chain setting, we have only *approximate independence*. Fortunately, through modern concentration inequalities, this only results in small constant-factor changes to tail bounds and sample complexity, thus despite the significantly more challenging *partial independence* setting of Markov chains, I recover the same category of guarantee, with comparable constants in all bounds. It is also worth noting that in both cases, the guarantees and bounds are shown by analyzing novel *variance concepts*. In particular, in block sampling I use the *inter-block variance*, defined as the variance *over blocks* of the *average of* f (rather than the variance of f *over all records*), and in the Markov chain setting, I use the *inter-trace variance*, defined as the *stationary variance* over *length-T traces* of $\mathcal{M}$ of the *average of f* (rather than the standard *stationary variance*, i.e., the variance of f over the *stationary distribution* of $\mathcal{M}$). In both cases, I show that when more independence exists than an *a priori* assumption on T indicates, these variance concepts are commensurately smaller. Such independence structure is sufficient but not necessary for our methods to yield improvements, and this analysis leads to intuitive and simple conditions under which the method is guaranteed to outperform classical alternatives.

As for the actual mechanism of these methods, both work by taking averages over all available samples (regardless of their potential dependence). This doesn't impact expectation, via linearity of expectation, and can only reduce the variance of the estimate. Because they don't require *a priori* variance bounds, the algorithms employ *progressive sampling*, meaning they operate on an iteratively-doubling sample of blocks or traces, and at each iteration, upper-bound the appropriate variance-concept, and terminate when the sample size is sufficiently small given said variance bound to ensure that the user-supplied ε-δ mean-estimation guarantee is met. Due to this dynamic sampling schedule, these algorithms dynamically adapt their sample-consumption to structure found within the data, and consequently, the dominant terms of their sample complexities match known lower-bounds for (variance-aware) mean-estimation to within logarithmic factors.

The Form of the Bounds I now briefly preview the sample complexity guarantees derived in this part. Let T denote the *block size* or a *relaxation time* bound, r the range of f, and v_T the inter-block or inter-trace variance of f. Then both algorithms consume no more than

$$\mathbf{O}\left(\log\left(\frac{\log {}^{r}\!/\!_{\varepsilon}}{\delta}\right)\left(\frac{Tr}{\varepsilon}+\frac{Tv_T}{\varepsilon^2}\right)\right) \tag{39}$$

records or Markov-chain steps, with probability at least $1-\delta$. First note that the $\log\log {}^{r}\!/\!_{\varepsilon}$ term is incurred by correcting for *multiple comparisons* (early stopping) made over iterations of *progressive sampling.* Ignoring this term, if we then consider that this bounds the number of *dependent samples*, and we have a factor-T fewer of *near-independent* samples, the result then matches *Bennett's inequality*, with v_T playing the role of v.

In the worst case, $v_T \in \mathbf{\Theta}(v_1)$, where v_1 is the variance or stationary variance of f, and we match i.i.d. guarantees that take a single record from each length-T block or sequence. However, when more independence exists in the data, $v_T \ll v_1$, and we get a more interesting sample complexity bound. Note that we assume no *a priori* knowledge of v_T, v_1, or their relationship, yet nevertheless, these quantities appear in sample complexity bounds.

In particular, in the block sampling setting, if we assume $v_T \leq \frac{\alpha v_1}{T}$, which occurs with $\alpha = 1$ assuming perfect independence within blocks, then the number of *blocks* consumed by the block sampling algorithm, via equation 39 is

$$\mathbf{O}\left(\log\left(\frac{\log {r}/{\varepsilon}}{\delta}\right)\left(\frac{r}{\varepsilon} + \frac{\alpha v_1}{T\varepsilon^2}\right)\right) \ . \tag{40}$$

Similarly, in the Markov-chain setting, I show that $Tv_T \in \mathbf{O}(\tau_{\mathrm{rel}} v_1)$, where τ_{rel} is the *relaxation time* of chain $\mathcal{M}$, thus equation 39 yields a sample complexity bound of

$$\mathbf{O}\left(\log\left(\frac{\log {r}/{\varepsilon}}{\delta}\right)\left(\frac{Tr}{\varepsilon} + \frac{\tau_{\mathrm{rel}} v_1}{\varepsilon^2}\right)\right) \ . \tag{41}$$

The fascinating insight here is that, although *estimating the relaxation time* τ_{rel} is generally impossible without strong assumptions on $\mathcal{M}$, τ_{rel} still appears in the sample complexity bound. In particular, the loose bound T appears in equation 41, but is asymptotically dominated by a term involving τ_{rel} as $\varepsilon \to 0$. Again these bounds are presented in loose simplified forms to emphasize their similarity; in the subsequent two chapters, sharper forms, non-asymptotic guarantees, and corresponding insights are derived.

Chapter 6

Mean Estimation in Block Databases

I formulate and analyze *BlockSample*, a suite of sampling algorithms that use the entire information in each sampled block to estimate the mean values of arbitrary functions of record values. BlockSample is tailored to modern data storage where the cost of retrieving a random record is essentially the same as the cost of reading a block of records. The algorithms *progressively sample* blocks of records, rather than individual records, and estimate the variance of the mean functions between blocks. By applying a specially tailored version of Bernstein's large deviation bound, the algorithm adaptively computes the required number of blocks to yield provable confidence intervals (or ε-δ-approximations) of function means. Thus, the sampling cost of BlockSample adapts to the *amount of variance* and *degree of independence* within each block, despite assuming no *a priori* information about either. In experiments on real data, a $40\times$ speedup over i.i.d. sampling was achieved for a block size of 128 records. This chapter is based on joint work with Larry Rudolph and Eli Upfal.

6.1 Preliminaries

We assume that our database $\boldsymbol{D}$ is organized into M *fixed-size blocks*, each containing exactly B *records* in space $\mathcal{X}$, thus $\boldsymbol{D}$ may be thought of as a matrix in $\mathcal{X}^{M\times B}$, where $D_{i,j}$ is the j record in block i. Records may contain several *fields*, e.g., an `Employee` record $x \in \mathcal{X}$ may be $(\texttt{Name} \in \textsc{String}, \texttt{ID} \in \mathbb{N}, \texttt{Salary} \in \mathbb{R})$. Let $\mathcal{D}$ be uniform distribution over the $M \times B$ records of $\boldsymbol{D}$. Given a function $f : \mathcal{X} \to \mathbb{R}$, or a function family $\mathcal{F} \subseteq \mathcal{X} \to \mathbb{R}$, where $\mathcal{F} = \{f_1, \ldots, f_n\}$, we want to estimate the expected value, i.e., the scalar $\mathbb{E}_{\mathcal{D}}[f]$ or the vector $\mathbb{E}_{\mathcal{D}}[\mathcal{F}] = (\mathbb{E}_{\mathcal{D}}[f_1], \ldots, \mathbb{E}_{\mathcal{D}}[f_n])$.

Our algorithms do not have access to the distribution $\mathcal{D}$ over the individual records. Instead the algorithms sample from a uniform distribution over the M blocks. To estimate $\mathbb{E}_{\mathcal{D}}[f]$ the algorithm samples *blocks* $\boldsymbol{x}_{r,1:B} = \boldsymbol{D}_{i,1:B}$, for $i \sim \mathcal{U}(1:M)$, and computes $\hat{\mathbb{E}}_{\boldsymbol{x}_{r,1:B}}[f] = \frac{1}{B}\sum_{j=1}^{B} f(\boldsymbol{x}_{r,j})$, i.e., the mean value of the block's B records, for each block. Since by linearity,

$$\mathop{\mathbb{E}}_{\mathcal{D}}[f] = \mathop{\mathbb{E}}_{i\sim\mathcal{U}(1:M)}\Big[\mathop{\hat{\mathbb{E}}}_{\boldsymbol{D}_{i,1:B}}[f]\Big] \enspace .$$

we can estimate $\mathbb{E}_{\mathcal{D}}[f]$ by $\frac{1}{m_B}\sum_{i=1}^{m_B} \hat{\mathbb{E}}_{\boldsymbol{x}_{i,1:B}}[f]$ for m_B blocks of size B, chosen uniformly at random.

We are interested in the minimum number of block samples needed to obtain an (ε,δ)-estimate of $\mathbb{E}_{\mathcal{D}}[f]$, i.e.,

$$m_B^*(\varepsilon,\delta) = \inf_{m_B^*\in\mathbb{N}} : \sup_{m_B > m_B^*} \mathbb{P}\left(\left|\frac{1}{m_B}\sum_{i=1}^{m_B} \mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}_{i,1:B}}[f] - \mathop{\mathbb{E}}_{\mathcal{D}}[f]\right| \geq \varepsilon\right) \leq \delta \enspace .$$

Thus, our algorithms and analysis focus on the random variable $\hat{\mathbb{E}}_{\boldsymbol{x}_{i,1:B}}[f]$ for block i chosen uniformly at random. The mean and range of $\hat{\mathbb{E}}_{\boldsymbol{x}_{i,1:B}}[f]$ is equal to that of f, its variance, the *inter-block-variance* $V_B \doteq \mathbb{V}_{i\sim\mathcal{U}(1:M)}[\hat{\mathbb{E}}_{\boldsymbol{D}_{i,1:B}}[f]]$, is typically smaller than $v \doteq \mathbb{V}_{\mathcal{D}}[f]$, the variance of f.

To show this, we apply the *law of total variance* to decompose the *total variance* $v = \mathbb{V}_{\mathcal{D}}[f]$ into *intra-block* variance $V_B^{\odot}$ and *inter-block* variance V_B as

$$\begin{aligned}
&\mathop{\mathbb{V}}_{\mathcal{D}}[f] = \frac{1}{MB}\sum_{i=1}^{M}\sum_{j=1}^{B}\Big(f(\boldsymbol{D}_{i,j}) - \mathop{\hat{\mathbb{E}}}_{\boldsymbol{D}_{1:M,1:B}}[f]\Big)^2 = \\
&\underbrace{\frac{1}{M}\sum_{i=1}^{M}\frac{1}{B}\sum_{j=1}^{B}\Big(f(\boldsymbol{D}_{i,j}) - \mathop{\hat{\mathbb{E}}}_{\boldsymbol{D}_{i,1:B}}[f]\Big)^2}_{\text{Intra-Block Variance } V_B^{\odot}} + \underbrace{\frac{1}{M}\sum_{i=1}^{M}\Big(\mathop{\hat{\mathbb{E}}}_{\boldsymbol{D}_{i,1:B}}[f] - \mathop{\hat{\mathbb{E}}}_{\boldsymbol{D}_{1:M,1:B}}[f]\Big)^2}_{\text{Inter-Block Variance } V_B} \enspace .
\end{aligned}$$

Here we are primarily interested in the inter-block variance V_B, and its impact on the required block-sample size. The above shows us that $0 \leq V_B \leq v$; the minimal $V_B = 0$ is obtained when all blocks are identical, while the maximum value $V_B = v$ is obtained when the records within each block are identical, but different between blocks. A third case of interest is $V_B = v/B$ which corresponds to a uniform random partition of record to blocks. In general, V_B corresponds to the dependence between the allocations of items to blocks and the value of f. For example, if the data represents some sequential process such as time series, positive sequential dependence leads to higher V_B, while negative sequential dependence gives smaller V_B. Note that $B = 1$ corresponds to standard uniform random sample of individual items, with $V_1^{\odot} = 0$, thus $V_1 = v$.

In addition to *variance* our algorithms are often concerned with the *range* of the function f. In particular, we assume knowledge of an interval r such that $\forall x \in \mathcal{X} : f(x) \in r$, and we denote the lower and upper limits of

r as $\lfloor r \rfloor$ and $\lceil r \rceil$, and the *length* as $\lfloor r \rceil = \lceil r \rceil - \lfloor r \rfloor$.

6.1.1 Variance and Concentration Inequalities

We saw in the above discussion that averaging the value of a function over all records in a randomly chosen block gives an unbiased estimate of $\mathbb{E}_{\mathcal{D}}[f]$ with possibly smaller variance than the value of the function on one random record. To fully utilise the possible reduction in variance we need:

1. A tight bound on the minimum sample size $m^*_B(\varepsilon, \delta)$ as a function of the sample variance; and

2. Adapting the bound in (1) to use the empirical variance instead of the true variance of the sample.

In the asymptotic regime, the *Central Limit Theorem* (CLT) gives an optimal solution to (1) and the t-distribution properties resolve (2). In the finite sample regime these tasks are harder and do not have optimal solution. Nevertheless, rigorous finite-sample guarantees are needed in algorithmic applications because asymptotic bounds for samples from common distributions such has heavy-tailed distributions can be extremely inaccurate on finite samples.

The standard CS tool for *finite-sample* bounds is *Hoeffding's inequality* [Hoeffding, 1963], which generalizes Chernoff's inequality (see, e.g., [Mitzenmacher and Upfal, 2017]) to provide *sub-Gaussian* guarantees that depend on the range of the random variable.

Theorem 6.1.1 (Hoeffding's Inequality)**.** Suppose distribution $\mathcal{D}$ over $r \subseteq \mathbb{R}$ and mean μ, and sample $\boldsymbol{x} \sim \mathcal{D}^m$. We then have

$$\begin{gathered} \mathop{\mathbb{P}}_{\boldsymbol{x}_{1:m}}\left(\left|\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{x}_i - \mu\right| \geq \sqrt{\frac{2\frac{1}{4}\lfloor r \rceil^2 \ln(\frac{2}{\delta})}{m}}\right) \leq \delta \ \& \\ m^*(\varepsilon, \delta, \lfloor r \rceil) \leq \frac{2\frac{1}{4}\lfloor r \rceil^2 \ln(\frac{2}{\delta})}{\varepsilon^2} \ . \end{gathered} \tag{42}$$

Hoeffding's inequality is tight to within constant factors when variance is maximal (i.e., $v = \frac{1}{4}\lfloor r \rceil^2$), but is quite loose as $v \to 0$.

In our work, we seek bounds that merge the *simple assumptions* and *rigorous finite-sample guarantees* of Hoeffding's inequality [Hoeffding, 1963] with the *sharpness* and *variance-dependence* of the asymptotic CLT-Gaussian bounds. To this end, we build on *Bennett's (1962) inequality* (which improves an earlier result of Sergei Bernstein Bernstein [1924] under identical conditions), by replacing dependence on *known a priori variance* with *empirical variance.*

Theorem 6.1.2 (Bennett's inequality (sub-gamma form))**.** Suppose $\mathcal{D}$ a distribution over $r \subseteq \mathbb{R}$, where $\lfloor r \rceil$

and variance v. Then

$$\mathop{\mathbb{P}}_{\boldsymbol{x}_{1:m}}\left(\left|\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{x}_i - \mu\right| \geq \underbrace{\frac{\lfloor r\rceil \ln(\frac{2}{\delta})}{3m}}_{\text{Scale}} + \underbrace{\sqrt{\frac{2v\ln(\frac{2}{\delta})}{m}}}_{\text{Scale}}\right) \leq \delta \;\; \& \\ m^*(\varepsilon, \delta, \lfloor r\rceil, v) \leq \underbrace{\frac{2\lfloor r\rceil \ln(\frac{2}{\delta})}{3\varepsilon}}_{\text{Scale}} + \underbrace{\frac{2v\ln(\frac{2}{\delta})}{\varepsilon^2}}_{\text{Variance}} \;\; . \tag{43}$$

This bound has *sub-gamma form*, i.e., resembles the Chernoff bound of a *gamma random variable* (see [Boucheron et al., 2013]). The form of ε is a convenient lens through which to view the result; we decompose the bound into a fast-decaying (in m) $\lfloor r\rceil$-dependent *scale term* and a slow-decaying (in m) v-dependent *variance-term.* If we ignore the scale term, Bennett's inequality matches the asymptotic CLT-Gaussian bound, *despite holding non-asymptotically*, and improves upon Hoeffding's inequality, by replacing the worst-case $\frac{1}{4}\lfloor r\rceil^2$ sub-Gaussian variance-proxy with a sub-gamma variance-proxy of v. The fast-decaying $\lfloor r\rceil$-term is usually quite small, and it acts as a finite-sample correction for an asymptotic sub-Gaussian bound.

Our analysis relies of two large deviation bounds. We apply a sub-gamma bound to the error in estimating the variance by the empirical variance, and then use that estimate of the variance to compute a bound on the required sample size using a Bennett's type bound. Together, this results can be viewed as a sub-gamma finite sample approximation for the t-statistic.

6.1.2 Prior Work

Sampling techniques and their analysis have been studied extensively, mostly in the asymptotic regime. Here we are interested in finite-sample analysis, as required for algorithmic analysis. We are particularly focused on large deviation bounds that are sensitive to the sample variance, and on tight substituting the variance with empirical one. Similar bounds exist in the literature [Audibert et al., 2007, Mnih et al., 2008, Audibert et al., 2009, Maurer and Pontil, 2009], however our analysis is sharper, and we show a complementary lower-tail bound, which we required for high-probability sample-complexity guarantees.

With *known variance*, it is easy to determine a sufficient sample size m to satisfy an ε-δ error guarantee, but with *empirical variance*, we can't determine if a sample-size is sufficient until we've already drawn it, necessitating a sort of *guess and check* approach. We address this issue with *progressive sampling*, which has been used in data-science settings where sufficient sample sizes are not known *a priori*, e.g., with Rademacher averages in *pattern mining* problems, [Pietracaprina et al., 2010, Riondato and Upfal, 2015] and with *variance* in *relative confidence interval estimation* [Mnih et al., 2008]. The basic idea is that we pick a *sampling* *schedule* (*a priori*) that tells us how many samples to *progressively draw* at each timestep, and we follow the schedule until our ε-δ guarantee is met, after which we can stop sampling.

6.2 Algorithms

In this section, we present our mean-query sampling algorithms for block data, alongside intuitive explanation, and rigorous finite-sample correctness and efficiency guarantees. We first consider in section 6.2.1 the simple case of *single-function mean estimation* with *additive guarantees*, emphasizing simplicity and intuitive presentation over sampling efficiency and flexibility. We then present an optimized algorithm for *multi-function mean estimation* in section 6.2.2. Our first algorithm assumes that the function f being estimated has range $[0, 1]$, but we lift this restriction in the generalized algorithm. We close by sketching how these algorithms may be modified to support *arbitrary error guarantees* and address other related tasks.

6.2.1 A Simple Block Sampling Algorithm

We first present a simple technique for single-function block-mean estimation, pseudocode given as algorithm 3. Algorithm 3 takes as input *query parameters*, which are a database $\boldsymbol{D}$ of M blocks of size B and a function $f : \mathcal{X} \to [0, 1]$ upon which to perform mean estimation, and the *confidence parameters* ε and δ, which determine the width and confidence of the resulting additive-ε guarantee. In each iteration, algorithm 3 draws a geometrically-increasing number of blocks to sample uniformly from the database. It then averages f over all the records in each block, and then applies an empirical-variance sensitive variant of Bennett's inequality, repeating until it produces a 2ε-diameter CI.

Intuitively, algorithm 3 enforces its ε-δ additive guarantees by establishing a *fixed schedule* of geometrically increasing sample-sizes and bound-applications that are all valid simultaneously (via union bound). Essentially, algorithm 3 initially draws a small sample, after which it may terminate immediately, if the *empirical variance* (and thus w.h.p. the true variance) is ≈ 0, via application of theorem 6.1.2; otherwise, it samples additional blocks to double the total sample size at each iteration, until *either* the Bennett confidence interval satisfies the requested ε-δ guarantee, *or* the *maximum iteration count* N is reached, after which a Hoeffding confidence interval implies the guarantee. In both cases, the bounds depend on $\ln(N)$ to account for the multiple-comparisons made by the algorithm in iterations before termination (i.e., the possibility that a small sample is drawn with *empirical variance* $\hat{v} \ll v$, leading to early termination). Formally, we characterize the guarantees of algorithm 3 below, with proof given in the appendix.

Algorithm 3 Simple Block Sampling

procedure UBMS($\boldsymbol{D}$, f, ε, δ) $\to \hat{\boldsymbol{\mu}}$
  Input: *Database* $\boldsymbol{D}$ of M blocks of size B, *function* $f : \mathcal{X} \to [0, 1]$, *confidence interval width* ε, and *failure probability* $\delta \in (0, 1)$
  Output: Additive ε-δ approximation $\hat{\boldsymbol{\mu}}$ of $\mathop{\mathbb{E}}_{\mathcal{D}}[f]$
  $N \leftarrow \left\lceil \log_2\left(\frac{3}{28\varepsilon}\right) \right\rceil$ ▷ Maximum iteration index
  $\alpha \leftarrow \frac{14 \ln(\frac{3N}{\delta})}{3\varepsilon}$ ▷ Minimal Sufficient Sample Size
  $\beta \leftarrow 2$ ▷ Geometric Schedule Ratio
  $\boldsymbol{U}_0 \leftarrow 0; \boldsymbol{V}_0 \leftarrow 0$ ▷ Avg & avg^2 accumulators
  $\boldsymbol{m}_0 \leftarrow 0$ ▷ Sample count begins at 0
  for $t \in 1, 2, \ldots, N$ **do**
    $\boldsymbol{m}_t \leftarrow \lceil \alpha \beta^t \rceil$ ▷ Geometric sampling schedule
    $\boldsymbol{m}'_t \leftarrow \boldsymbol{m}_t - \boldsymbol{m}_{t-1}$ ▷ Marginal sample size
    $\mathcal{I} \sim \mathcal{U}^{\boldsymbol{m}'_t}(1, \ldots, M)$ ▷ Draw uniform block indices
    $\boldsymbol{x}_{t,\cdot,\cdot} \leftarrow \textsc{Read}(\boldsymbol{D}, \mathcal{I})$ ▷ Read of $\boldsymbol{x}_{t,\cdot,\cdot} \in \mathcal{X}^{\boldsymbol{m}'_t \times B}$
    $\boldsymbol{U}_t \leftarrow \boldsymbol{U}_{t-1} + \sum_{j=1}^{\boldsymbol{m}'_t} \hat{\mathop{\mathbb{E}}}_{\boldsymbol{x}_{t,j,1:B}}[f]$ ▷ Sum block avgs
    $\hat{\boldsymbol{\mu}}_t \leftarrow \frac{1}{\boldsymbol{m}_t} \boldsymbol{U}_t$ ▷ Grand mean
    $\boldsymbol{V}_t \leftarrow \boldsymbol{V}_{t-1} + \sum_{j=1}^{\boldsymbol{m}'_t} \hat{\mathop{\mathbb{E}}}^2_{\boldsymbol{x}_{t,j,1:B}}[f]$ ▷ Sum square block avgs
    $\hat{\boldsymbol{v}}_t \leftarrow \frac{\boldsymbol{m}_t}{\boldsymbol{m}_t - 1} \left(\frac{1}{\boldsymbol{m}_t} \boldsymbol{V}_t - \hat{\boldsymbol{\mu}}_t^2 \right)$ ▷ Inter-block variance
    if $(t = N) \vee \left(\frac{7 \ln(\frac{3N}{\delta})}{3\boldsymbol{m}_t} + \sqrt{\frac{2 \ln(\frac{3N}{\delta}) \hat{\boldsymbol{v}}_t}{\boldsymbol{m}_t}} \leq \varepsilon \right)$ **then**
      return $\hat{\boldsymbol{\mu}}_t$
    end if
  end for
end procedure

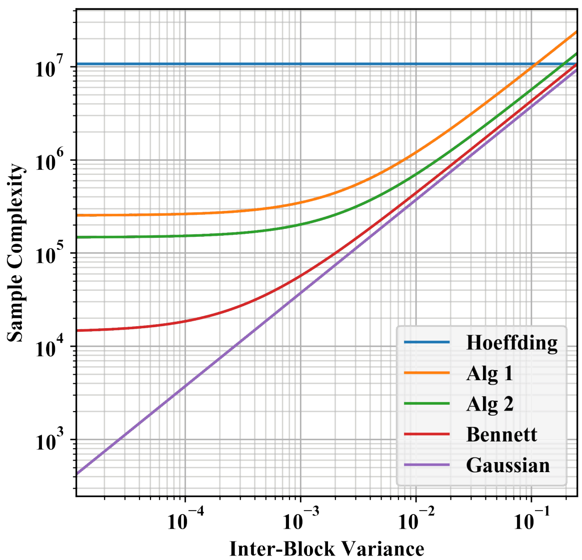


Figure 6.1: Sample complexity (y-axis) of algorithm 3 (theorem 6.2.1.1), and algorithm 4 (theorem 6.2.3) with optimized β, compared to Hoeffding, Bennett, and asymptotic Gaussian bounds, as a function of variance (x-axis). Note that variance is not known *a priori* to algorithm 3, but is used in the other bounds. Here $r = 1$, $\varepsilon = 10^{-3}$ and $\delta = 10^{-9}$.

Theorem 6.2.1 (Algorithm 3 Guarantees)**.** Suppose $\varepsilon > 0$ and $\delta \in (0, 1)$, then with respect to the distribution $\mathcal{M}$ of sampling blocks uniformly at random,

1. $\mathbb{P}\left(\left|\hat{\mu} - \mathbb{E}_{\mathcal{D}}[f]\right| \leq \varepsilon\right) \geq 1 - \delta$; &
2. If $\delta \leq 0.01$, then $\mathbb{P}\left(\hat{m} \leq 2 + \frac{4\ln(\frac{3N}{\delta})}{\varepsilon}\left(\frac{v}{\varepsilon} + \frac{8}{3}\right)\right) \leq 1 - \delta$.

Item 1 simply tells us that algorithm 3 works "as advertised," and produces a confidence interval of the user-requested diameter (tolerance) with the user-supplied certainty (failure probability). Item 2 is rather more informative; it tells us that, at least with high probability, the *sample consumption* of algorithm 3 is small, and *mostly* dependent on the *inter-block variance* of f. It is interesting to note that although this variance is *not known a priori*, it still characterizes the sample consumption of algorithm 3.

In figure 6.1, we contrast the high-probability sample-complexity bound of algorithm 3 with the corresponding sample complexities of Bennett and asymptotic Gaussian bounds. We see that, although the latter two methods are not usable in practice, as they require *known variance*, algorithm 3 is nearly as sample-efficient, despite only using *empirical variance*. Note also that the figure is not entirely fair, as it compares a rather loose *high probability bound* on algorithm 3 sample consumption to the exact required sample-size of the Bennett and Hoeffding bounds.

Algorithm 4 Block Mean Sampling

procedure BMS($\boldsymbol{D}$, $(\mathcal{F}, \boldsymbol{r}_{1:n}, \boldsymbol{\mu}_{0,1:n}, \boldsymbol{v}_{0,1:n})$, (ε, δ), $\beta) \to \boldsymbol{\mu}_{1:n}$

  Input: *Database* $\boldsymbol{D}$ of M size B blocks, *function family* $\mathcal{F} \in \prod_{i=1}^{n}(\mathcal{X} \to \boldsymbol{r}_i)$, (*range*, *mean*, *variance*) interval bounds $(\boldsymbol{r}_i, \boldsymbol{\mu}_{0,i}, \boldsymbol{v}_{0,i})$ for each $f_i \in \mathcal{F}$, target CI (*radius*, *error probability*) bounds (ε, δ), geometric ratio β

  Output: *Confidence intervals* $\boldsymbol{\mu}_{i\in 1:n}$ of each $\mathop{\mathbb{E}}_{\mathcal{D}}[f_i]$

  ▷ Compute max timestep N and min sufficient size $\lceil \alpha \rfloor$

  $N \leftarrow 1 \vee \left\lceil \log_\beta \left(\dfrac{\max_{i\in 1:n} 4(\lceil \boldsymbol{v}_i \rceil + \frac{7\lfloor \boldsymbol{r}_i \rceil}{3}\varepsilon) \wedge \lfloor \boldsymbol{r}_i \rceil^2}{\max_{i\in 1:n} 4(\lfloor \boldsymbol{v}_i \rfloor + \frac{7\lfloor \boldsymbol{r}_i \rceil}{3}\varepsilon) \wedge \lfloor \boldsymbol{r}_i \rceil^2} \right) \right\rceil$

  $\alpha \leftarrow \frac{1}{2\varepsilon^2} \ln(\frac{3nN}{\delta}) \big(\max_{i\in 1:n} 4(\lfloor \boldsymbol{v}_i \rfloor + \frac{7\lfloor \boldsymbol{r}_i \rceil}{3}\varepsilon) \wedge \lfloor \boldsymbol{r}_i \rceil^2 \big)$

  $\boldsymbol{U}_{0,1:n} \leftarrow \boldsymbol{0}$; $\boldsymbol{V}_{0,1:n} \leftarrow \boldsymbol{0}$ ▷ Avg & avg^2 accumulators

  $\boldsymbol{m}_0 \leftarrow 0$ ▷ Initialize sample count

  $T \leftarrow 3nN$ ▷ Compute total # of tail bounds

  for $t \in 1, 2, N$ **do** ▷ Main loop

    $\boldsymbol{m}_t \leftarrow \lceil \alpha\beta^t \rceil$ ▷ Geometric sampling schedule (cumulative)

    $\boldsymbol{m}'_t \leftarrow \boldsymbol{m}_t - \boldsymbol{m}_{t-1}$ ▷ # of blocks to process this round

    $\mathcal{I} \sim \mathcal{U}^{\boldsymbol{m}'_t}(1, \ldots, M)$ ▷ Draw $\boldsymbol{m}'_t$ uniform block indices

    $\boldsymbol{x}_{t,\cdot,\cdot} \leftarrow \textsc{Read}(\boldsymbol{D}, \mathcal{I})$ ▷ Read of $\boldsymbol{x}_{t,\cdot,\cdot} \in \mathcal{X}^{\boldsymbol{m}'_t \times B}$

    for $f_i \in \mathcal{F}$ **do**

      $\boldsymbol{U}_{t,i} \leftarrow \boldsymbol{U}_{t-1,i} + \sum_{j=1}^{\boldsymbol{m}'_t} \hat{\mathop{\mathbb{E}}}_{\boldsymbol{x}_{t,j,\cdot}}[f_i]$ ▷ Sum of block avgs

      $\hat{\boldsymbol{\mu}}_{t,i} \leftarrow \frac{1}{\boldsymbol{m}_t} \boldsymbol{U}_{t,i}$ ▷ Grand mean

      $\boldsymbol{V}_{t,i} \leftarrow \boldsymbol{V}_{t-1,i} + \sum_{j=1}^{\boldsymbol{m}'_t} \hat{\mathop{\mathbb{E}}}_{\boldsymbol{x}_{t,j,\cdot}}{}^2[f_i]$ ▷ Sum of square block means

      $\hat{\boldsymbol{v}}_{t,i} \leftarrow \frac{\boldsymbol{m}_t}{\boldsymbol{m}_t - 1} \left(\frac{1}{\boldsymbol{m}_t} \boldsymbol{V}_{t,i} - \hat{\boldsymbol{\mu}}_{t,i}^2 \right)$ ▷ Inter-block variance

      $\varepsilon_v \leftarrow \dfrac{2\lfloor \boldsymbol{r}_i \rceil^2 \ln(\frac{T}{\delta})}{3\boldsymbol{m}_t} + \sqrt{\left(\dfrac{\lfloor \boldsymbol{r}_i \rceil^2 \ln(\frac{T}{\delta})}{\sqrt{3}\boldsymbol{m}_t} \right)^2 + \dfrac{2\lfloor \boldsymbol{r}_i \rceil^2 \ln(\frac{T}{\delta}) (\hat{v} + \frac{\lfloor \boldsymbol{r}_i \rceil^2 \boldsymbol{m}_t}{4(\boldsymbol{m}_t - 1)^2})}{\boldsymbol{m}_t}}$

      $\boldsymbol{v}_{t,i} \leftarrow \boldsymbol{v}_{t-1,i} \cap \big[0,\ \hat{\boldsymbol{v}}_{t,i} + \varepsilon_v\big]$ ▷ Refine variance bound

      $\varepsilon_- \leftarrow \min\left(\frac{(\lceil \boldsymbol{r}_i \rceil - \lfloor \boldsymbol{\mu}_{t-1,i} \rfloor) \ln(\frac{T}{\delta})}{3\boldsymbol{m}_t} + \sqrt{\frac{2\lceil \boldsymbol{v}_{t,i} \rceil \ln(\frac{T}{\delta})}{\boldsymbol{m}_t}},\ \lfloor \boldsymbol{r}_i \rceil \sqrt{\frac{\ln(\frac{T}{\delta})}{2\boldsymbol{m}_t}} \right)$

      $\varepsilon_+ \leftarrow \min\left(\frac{(\lceil \boldsymbol{\mu}_{t-1,i} \rceil - \lfloor \boldsymbol{r}_i \rfloor) \ln(\frac{T}{\delta})}{3\boldsymbol{m}_t} + \sqrt{\frac{2\lceil \boldsymbol{v}_{t,i} \rceil \ln(\frac{T}{\delta})}{\boldsymbol{m}_t}},\ \lfloor \boldsymbol{r}_i \rceil \sqrt{\frac{\ln(\frac{T}{\delta})}{2\boldsymbol{m}_t}} \right)$

      $\boldsymbol{\mu}_{t,i} \leftarrow \boldsymbol{\mu}_{t-1,i} \cap \big[\hat{\boldsymbol{\mu}}_{t,i} - \varepsilon_-,\ \hat{\boldsymbol{\mu}}_{t,i} + \varepsilon_+\big]$ ▷ Refine mean

    end for

    if $\max_{i\in 1:n} \lfloor \boldsymbol{\mu}_{t,i} \rceil \leq 2\varepsilon$ **then** ▷ Check additive ε-error

      return $\boldsymbol{\mu}_{t,1:n}$

    end if

  end for

end procedure

6.2.2 A Refined Block Sampling Algorithm

We now present algorithm 4, which supports compound queries (multiple functions), can use *a priori* variance and expectation knowledge, and is generally more sample-efficient than algorithm 3. Briefly put, the sample complexity improvements derive from using tighter concentration inequalities, propagating bounds between timesteps, though the overall structure is largely unchanged.

Here we walk through algorithm 4, in particular emphasizing where and how it differs from algorithm 3. The first major difference is in its input; rather than taking a single function f, algorithm 4 operates on a *function family* $\mathcal{F}$. Additionally, algorithm 4 relaxes the *unit range* assumption of algorithm 3, and thus needs the *range* $\boldsymbol{r}_i$ of each $f_i \in \mathcal{F}$, and also takes (possibly vacuous) *a priori mean* and *variance* bounds $\boldsymbol{\mu}_{0,\cdot}$ and $\boldsymbol{v}_{0,\cdot}$ of each $f_i \in \mathcal{F}$. The algorithm also adds a β parameter, which determines the *geometric ratio* between successive sample sizes (discussed further in section 6.2.3), whereas algorithm 3 simply fixes $\beta = 2$.

Another major difference is in the *sampling schedule* of algorithm 4. Because the algorithm must operate on a *family* of functions, each with their own range, and possibly variance bounds, determining the *range* of sample sizes to draw, and thus the *number* N of doubling geometric samples to draw, is more involved. We use a union bound over a variance and 2-tailed mean bound for each function at each timestep, for a total of $3nN$ bounds. Because algorithm 4 can not terminate until *each* $f_i \in \mathcal{F}$ has been ε-additively estimated, the smallest and largest sample sizes are both worst-case over $\mathcal{F}$, and depend on the known range $\boldsymbol{r}_i$, which increases both, and *variance bounds* $\boldsymbol{v}_i$, where *lower bounds* can *increase* minimal sample sizes, and *upper-bounds* can *decrease* maximal sample sizes. The details are rather complicated (see proof of theorem 6.2.2, but intuitively, a sampling schedule is selected using *a priori* knowledge to avoid statistical inefficiency by ignoring infeasible sample sizes. However, if we assume $\mathcal{F} = \{f_1\}$ & $\boldsymbol{r}_1 = [0, 1]$, (as in algorithm 3), the schedule of algorithm 4 is more efficient, using fewer iterations, ($N \leq \lceil \log_2(3/28\varepsilon) \rceil$) and a finer geometric sample size grid ($\beta \leq 2$).

Each step of the main loop progressively samples blocks (line 14), and for each $f \in \mathcal{F}$ computes the empirical variances and means of block-averages (lines 17 & 19). The algorithm then refines its existing bounds on variances and means (lines 21-24), which synthesizes the information gained this iteration with the *a priori* bounds and previous iterations. Algorithm 4 may require fewer samples than algorithm 3, as it also considers upper and lower bounds from previous timesteps, which due to random fluctuations in empirical means and variances, may be sharper than the current bound.

While the exact form of the tail bounds in algorithm 4 is rather subtle, they are defined and discussed further in section 6.2.2, and they always improve upon the simple empirical Bennett inequality (line 18) of algorithm 3. The last step of each iteration is then to check the additive ε-error *termination condition*, which is satisfied with certainty by the end of iteration N.

Concentration Inequalities Algorithm 3 uses a generic sub-gamma empirical Bennett bound, whereas algorithm 4 uses an asymmetric refined bound with weaker dependence on *function ranges*. This is significant in the block-sampling setting, as we generally expect very small inter-block variances ($V_B \ll V_1 \leq \frac{1}{4}\lfloor r\rceil^2$), in which case the fast-decaying $\mathbf{\Theta}(\lfloor r\rceil \ln(\frac{1}{\delta})/m)$ *scale term* can dominate the slow-decaying $\mathbf{\Theta}\sqrt{\ln(\frac{1}{\delta})v/m}$ *variance term*, thus moving closer to purely sub-Gaussian tails greatly reduces sample complexity.

Algorithm 4 bounds the *variance* v in terms of the unbiased empirical variance $\hat{v}$ with probability at least $1-\delta$ as

$$v \leq \hat{v} + \underbrace{\frac{2\lfloor r\rceil \ln(\frac{1}{\delta})}{3m}}_{\textsc{Scale}} + \underbrace{\sqrt{\left(\frac{\lfloor r\rceil \ln(\frac{1}{\delta})}{\sqrt{3}m}\right)^2 + \frac{2\lfloor r\rceil^2 \ln(\frac{1}{\delta})(\hat{v}+\frac{m}{4(m-1)^2})}{m}}}_{\textsc{Variance}} , \tag{44}$$

(used on line 20) and similarly, our efficiency guarantees require a bound in the opposite direction, namely, with probability at least $1-\delta$, we have

$$\hat{v} \leq v + \underbrace{\frac{\lfloor r\rceil \ln(\frac{1}{\delta})}{3m}}_{\textsc{Scale}} + \underbrace{\sqrt{\frac{2\lfloor r\rceil^2 \ln(\frac{1}{\delta})(v + \frac{m}{4(m-1)^2})}{m}}}_{\textsc{Variance}} , \tag{45}$$

both discussed further in the appendix. Additional we apply Bennett's inequality more carefully, noting that the upper and lower tails can use scale terms $\frac{\lceil \boldsymbol{r}_i\rceil - \mathbb{E}[f_i]}{3}$ and $\frac{\mathbb{E}[f_i]-\lfloor \boldsymbol{r}_i\rfloor}{3}$, respectively (see lines 22 & 23), which gives asymmetric tail bounds up to a factor-2 sharper than the scale term $\frac{\lfloor \boldsymbol{r}_i\rceil}{3}$ used by algorithm 3.

6.2.3 Setting Parameters

Schedule Selection With the introduction of the β parameter, algorithm 4 allows the user to tune the geometric sampling schedule to tradeoff between a large number of closely-spaced sample sizes (and thus a large union bound), or a small number of distant sample sizes (and thus potentially overshooting the sufficient sample size).

We note that while algorithm 3 simply fixes $\beta = 2$, as this intuitively simple choice leads to a straightforward *doubling* of sample size at every step, this choice is often quite sub-optimal, by up to a 2-factor (hence the $\approx$ 2-factor gap between UBMS and Bennett's inequality in figure 6.1). Decreasing β can remove of this 2-factor, though only to a point, as taking β too close to 1 increases N, which counteracts this improvement.

The other difficulty of using algorithm 4 is in setting the *range*, *mean*, and *variance* parameters $\boldsymbol{r}$, $\boldsymbol{\mu}$, and $\boldsymbol{v}$. First note that $f : \mathcal{X} \to r$ implies $\mathbb{E}_{\mathcal{D}}[f] \in r$ and $\mathbb{V}_{\mathcal{D}}[f] \leq \frac{1}{4}\lfloor r\rceil^2$, thus bounding $\boldsymbol{r}$ is *sufficient* to invoke algorithm 4, and sharper bounds on $\boldsymbol{\mu}$ & $\boldsymbol{v}$ are only used to improve performance.

Often f has bounded range (e.g., indicator functions, percentages, etc.), but in general, we need to assume the database and its indices maintain enough information to compute appropriate bounds. In particular, we assume they maintain a few elementary statistics on each field. If f is defined on a *single field* (similar methods apply for multivariate f) with range $\boldsymbol{r}_{\mathcal{X}}$, then

$$r \subseteq \left[\min_{x \in \boldsymbol{r}_x} f(x), \max_{x \in \boldsymbol{r}_x} f(x)\right] \enspace ,$$

which simplifies to $[f(\lfloor \boldsymbol{r}_x \rfloor), f(\lceil \boldsymbol{r}_x \rceil)]$ for increasing f.

Expectation and variance bounds are trickier, and depend on exactly which statistics we maintain and the properties of f. To illustrate the principle, we consider the case of *contraction functions*, i.e., $f : \mathbb{R} \to \mathbb{R}$ s.t. $\forall x \in \mathbb{R} : \left|\frac{\mathrm{d}}{\mathrm{d}x}\right| \leq 1$. If f is a contraction, it holds that $V_B \leq \mathbb{V}_{\mathcal{D}}[f] \leq \mathbb{V}[\mathcal{D}]$, thus this is another case where field-statistics maintained by the database yield *a priori* knowledge exploitable by algorithm 4. Similar properties requiring various assumptions on f and database statistics can be shown to bound $\boldsymbol{v}$ and $\boldsymbol{\mu}$, which in all cases can be given to algorithm 4 to improve sampling efficiency.

Algorithm 4 Guarantees The correctness and efficiency guarantees of algorithm 4 are much like in algorithm 3, although they must depend on $|\mathcal{F}|$ to account for multiple comparisons, and are also sensitive to *a priori* range and variance knowledge.

Theorem 6.2.2 (Algorithm 4 Correctness Guarantees)**.** Suppose $\varepsilon > 0$ and $\delta \in (0, 1)$, and take $\hat{\mu} \leftarrow \mathrm{BMS}(\boldsymbol{D}, (\mathcal{F}, \boldsymbol{r}, \boldsymbol{\mu}, \boldsymbol{v}), (\varepsilon, \delta), \beta)$. Let $n = |\mathcal{F}|$ and take N as computed on line 5. Then with probability at least $1 - \delta$ over randomness of BMS, it holds that

$$\forall i \in 1, \ldots, n : \left(\mathop{\mathbb{E}}_{\mathcal{D}}[f_i] \in \hat{\boldsymbol{\mu}} \wedge \lfloor \hat{\boldsymbol{\mu}}_i \rceil \leq 2\varepsilon\right) \enspace ,$$

and therefore $\max_{i \in 1, \ldots, n} \left|\mathop{\mathbb{E}}_{\mathcal{D}}[f_i] - \frac{\lceil \boldsymbol{\mu}_i \rceil - \lfloor \boldsymbol{\mu}_i \rfloor}{2}\right| \leq \varepsilon$.

Note that algorithm 4 produces *interval estimates*, which are more informative, not necessarily centered on the empirical mean, and often sharper than requested, particularly in multi-function estimation, when some function means are more easily estimated than others, and this information may be useful to the user downstream. The first statement in theorem 6.2.2 is a guarantee for these interval estimates, and the second converts their centers to *point estimates*, and more closely resembles theorem 6.2.1.1.

Theorem 6.2.3. Suppose $\varepsilon > 0$ and $\delta \leq 0.01n$, and take

$$\hat{\mu} \leftarrow \textsc{BlockMeanSampling}(\boldsymbol{D}, (\mathcal{F}, \boldsymbol{r}, \boldsymbol{\mu}, \boldsymbol{v}), (\varepsilon, \delta), \beta) \enspace .$$

Let $n = |\mathcal{F}|$ and take N as computed on line 5. Then, taking $V_{\max}$ and $r_{\max}$ to be the largest inter-block variance and range among $\mathcal{F}$, with probability at least $1 - \frac{\delta}{3N}$ over randomness of $\textsc{BlockMeanSampling}$, it holds that

$$\hat{m}(\varepsilon, \delta, v) \leq \left\lceil \frac{2\beta \ln(\frac{3nN}{\delta})}{\varepsilon} \left(\frac{V_{\max}}{\varepsilon} + \frac{8 r_{\max}}{3} \right) \right\rceil \enspace .$$

Proof. To see this result, note that together, eqs. 44 & 45 and Bennett's inequality are sufficient to show that if algorithm 4 reaches a sample of size at least

$$\frac{2 \ln(\frac{3nN}{\delta})}{\varepsilon} \left(\frac{V_{\max}}{\varepsilon} + \frac{8 r_{\max}}{3} \right) \enspace ,$$

then with probability at least $1 - \frac{\delta}{3N}$, the *empirical variance* of each f_i will be small enough to satisfy the termination condition. This exact sample size is not necessarily considered by algorithm 4, but because of its geometric structure, we can guarantee that a sample size a factor $\leq \beta$ in excess of the above (rounded up) is present in the sampling schedule, and termination w.h.p. will occur there. Note that here equation 44 is required, as we need to show that the *empirical variance* is not much greater than the *true variance*, whereas in the correctness proof, we only cared about the reverse bound. Note also that the statement holds with probability independent of N, as it only considers the largest single timestep where the sample size is no greater than the stated bound: terminating earlier only decreases $\hat{m}$, and we show that terminating later occurs with probability $\leq \frac{\delta}{3N}$. The $\delta \leq 0.01n$ requirement is necessary only to get the $\frac{8}{3}$ constant from eqs. 44 & 45; see appendix for details. $\square$

This result mirrors that of theorem 6.2.1.1, although a more careful analysis gives a higher-probability guarantee, removes a nuisance constant term, and is now sensitive to β and the other parameters of algorithm 4. This bound with optimal selection of β is plotted against the corresponding result for theorem 6.2.1 and various tail bounds in figure 6.1.

Comparative Sample Complexity Analysis In some sense, the strongest possible way to show that algorithm 4 is near optimal is to compare its sample complexity to that of the *minimax-optimal mean estimation algorithm* under identical conditions. In particular, assuming variance v and range $\geq \sqrt{4v}$, suppose that $\mathcal{A}$ is a *non-adaptive* (fixed-sample size) ε-δ single-function mean-estimation algorithm, and $m^*(\varepsilon, \delta, v)$

its sample complexity. [Devroye et al., 2016] show that under mild-conditions on δ,

$$m^*(\varepsilon, \delta, v) \geq \frac{\ln(\frac{1}{\delta})v}{\varepsilon^2} \ \ ,$$

i.e., the *minimax optimal mean estimator* requires at least this many samples. The best-known upper-bounds nearly match up to constant factors [Lugosi and Mendelson, 2019], but induce additional dependence on other parameters. In our case, we have *additive dependence* on *range*, and thus can't get a purely competitive ratio, but the following theorem gives an affine-characterization of our bound, which is nearly multiplicative when $r_{\max}/\varepsilon$ is small.

Theorem 6.2.4. Suppose as in theorem 6.2.3, and also $\ln(\frac{1}{\delta})v/\varepsilon^2 > 5$ & $\delta < \min(\frac{1}{2}, 0.01n, 2^{4\varepsilon^2/v})$. With probability at least $1-\delta$, it then holds that

$$\hat{m}(\varepsilon, \delta, v) \leq 2\beta\left(1 + \frac{\ln(3nN)}{\ln(\frac{1}{\delta})}\right)m^*(\varepsilon, \delta, v) + \beta\frac{16\ln(\frac{3nN}{\delta})r_{\max}}{3\varepsilon} \ \ .$$

Proof. This result is a consequence of theorem 6.2.1 item 2 and the sample-complexity lower-bound of [Devroye et al., 2016]. Note that the sub-gamma variance term is purely multiplicative term, but the sub-gamma scale term is additive, as the lower-bound has no range-dependence. The assumptions of δ are required by the sample-complexity lower-bound and theorem 6.2.3. □

This result is significant, as it tells us that despite its relative simplicity, no amount of cleverness can yield a general algorithm that performs significantly better than algorithm 4.

6.3 Experiments

This section compares the behavior of the two algorithms for various mean queries on a very large industrial dataset containing public disk block traces on block accesses in a data-center[1]. There is no need to compare the results of the sample to the full dataset, rather what is important is to see how the results of the i.i.d. sample compare to the block samples. The experiments vary block sizes $B \in \{1, \ldots, 128\}$, with 1 being i.i.d. sampling, and 128 being blocks of 128 records. We require more samples for higher confidence (small δ) and tighter intervals (small ε), and similarly, when inter-block variance V_B (which generally decreases with B) is large. Here we evaluate performance for various queries, parameter settings, and block sizes.

In each experiment, we estimate means to within target confidence interval diameter 2ε. The first two

[1] The Two Sigma data object trace set will be made generally available soon; it has been used in several other publications.

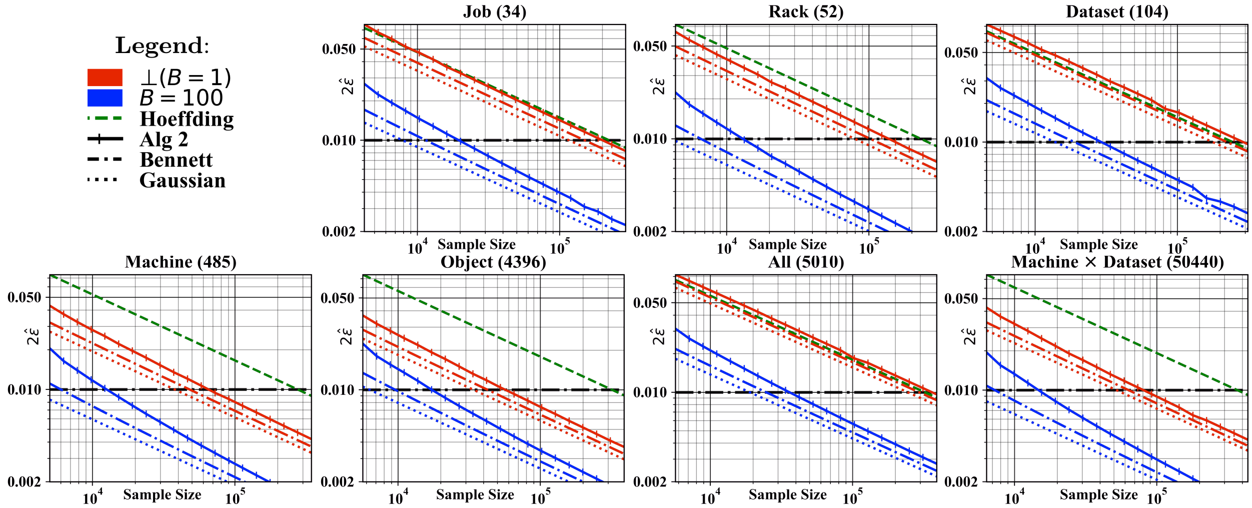


Figure 6.2: Confidence interval width $2\hat{\varepsilon}$ vs. sample size for frequency estimation tasks on the *Two Sigma disk block traces* dataset, with base-10 log-log axes. In each experiment, target $\varepsilon = \frac{1}{100}$ and $\delta = \frac{1}{1000}$, and the column field and $|\mathcal{F}|$ are given in the title. *Independent sampling* shown in red, and *block sampling* with block size $B = 100$ in blue. We show algorithm 4 (defined only at specific sample sizes, denoted by ı marks) comparing to Hoeffding, Bennett, and (asymptotic) Gaussian tail bounds. Termination occurs when a bound crosses the $\varepsilon = \frac{1}{100}$ line, so lower curves are better.

experiments show how various bounds behave *as a function of sample size* m, and the final experiment examines how *sample size* $\hat{m}$ changes as a function of *block size* B. Our primary goal in these experiments is to show that the sample complexity of block sampling is much less than that of independent sampling, which we see by comparing performance across different block sizes.

The Hoeffding bound is our baseline as a non-asymptotic sufficient sample size is computed *a priori* given only range information, thus is invariant under *block size* and *variance*. The Bennett and asymptotic Gaussian bounds, on the other hand, depend on the (unknown) true variance V_B of the data, so are not usable in practice. They are presented only for comparison purposes. A secondary goal of these experiments is to show that our algorithms, despite having less *a priori* information, are still competitive with the Bennett and Gaussian bounds.

All plots are presented on log-log axes, so as to distinguish many similar curves across a broad range of sample sizes and confidence interval diameters. This is quite helpful as it visually contrasts the *linear* $\boldsymbol{\Theta}(m^{-1/2})$ *convergence rates* of the sub-Gaussian Hoeffding and Gaussian bounds to the *curved mixed-rate* sub-Gamma Bennett and algorithm 4 bounds.

6.3.1 Frequency Estimation Tasks

We first address several *frequency estimation tasks* on each categorical fields of the dataset (`Job`, `Rack`, `Dataset`, `Machine`, and `Object` IDs), as well as the joint task of frequency estimation on *all five categorical* *fields*, and on categorical estimation of (`Machine`, `Dataset`) tuples. In each task, we require a single indicator function for *each category* of the relevant (combinations of) *categorical field(s)*, the expectation of which is the *category frequency*, thus all $f \in \mathcal{F}$ have range $[0, 1]$, and variance $\leq {}^1\!/_4$. Each query is performed with algorithm 4 using $\beta = 1.25$, and results are presented in figure 6.2.

The queries are arranged by increasing *function family size* $|\mathcal{F}|$, thus Hoeffding sample complexity is monotonic in $\mathcal{F}$. However, the remaining bounds, which are *variance-sensitive*, often attain *improved sample complexity* for larger $\mathcal{F}$, as *maximum variance* more strongly impacts performance than *function family size.*

Except for the `Job`, `Dataset`, and `All` experiments, which each contain a single high-variance function, we see that the variance-dependent bounds outperform Hoeffding's inequality, even for independent sampling. Furthermore, using $B = 10$ or $B = 100$ reduces the variance substantially, and algorithm 4 beats Hoeffding's inequality in all datasets.

In all experiments, we see sample consumption decrease as a function of block size, empirically validating the thesis that block-information yields superior convergence rates over independent sampling. Indeed we generally see sample complexity behave roughly $\propto \frac{1}{B}$, though due to *dependence within blocks*, the constant-of-proportionality is ≥ 1. Nevertheless, the improvement with increasing block size is still quite substantial. This also illustrates why standard i.i.d. bounds can't be applied to the mB samples at each timestep directly, and so our variance-based approach is required.

Also of note is the clear curvature of the *Bennett* and algorithm 4 bounds in log-log space. Both the *Hoeffding* and *Gaussian* plots are linear (slope $-{}^1\!/_2$), as confidence interval radius ε obeys $\varepsilon \propto c_{\max}/\sqrt{m}$ and $\varepsilon \propto \sqrt{V_{\max}/m}$, respectively, whereas the Bennett and empirical Bennett (algorithm 4) inequalities yield curves with initial $c_{\max}/m$-rate decay (slope -1), followed by subsequent $\sqrt{V_{\max}/m}$-rate decay (slope $-{}^1\!/_2$). This curvature is more visible when $V_{\max}$ is very close to 0 (e.g., when $B = 100$), as with the chosen parameters, the algorithm 4 generally starts with a large enough sample size that the $\sqrt{V_{\max}/m}$ term dominates. This is significant in block sampling, as for a sufficiently large block size (thus small variance), we can expect to remain in the fast-decaying (slope -1) regime for longer, and thus quickly reach a sufficient sample size.

A final point of interest is that algorithm 4 bounds exhibit slight deviations from the ideal sub-Gaussian and sub-Gamma curves of the other bounds, which generally appear as dips in the algorithm 4 curves. These deviations are most visible in the `Job` and `Dataset` field experiments, and are mainly due to using empirical variance instead of variance. Correcting for these fluctuations is also the reason why the algorithm 4 bounds are necessarily looser than the Bennett bounds.

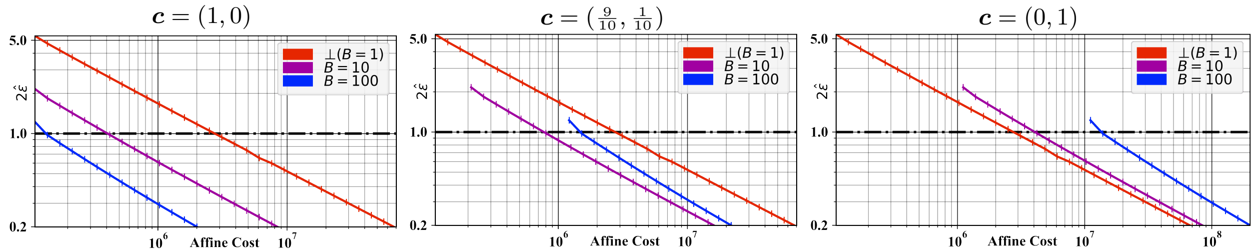


Figure 6.3: Sigmoid read size experiments, with log-log axes and various affine cost models. Confidence interval width $2\hat{\varepsilon}$ is plotted as a function of *affine cost* $\boldsymbol{c}_0 m + \boldsymbol{c}_1 Bm$ for $B \in \{1, 10, 100\}$. The left plot only pays per-block, and the right plot pays per record.

6.3.2 Sigmoid Functions

To understand how the algorithms behave for particular functions this subsection determines averages of *sigmoid functions* of `read sizes` (in mb) on the *Two Sigma block traces* dataset. We consider 15 sigmoid functions (see appendix), each a contraction with range $[-10^3, 10^3]$, and want simultaneous $\varepsilon = 0.5$, $\delta = 10^{-6}$ confidence intervals on each. Unlike the frequency experiments, these functions have large range, so are expected to have much higher variance. Additionally, we assume *a priori* knowledge of the *variance* v and *range* $\boldsymbol{r}$ of the `read size` field. Because each sigmoid is a monotonic contraction function, we use v and $\boldsymbol{r}$ to provide algorithm 4 with *a priori* $\boldsymbol{r}_i$ and $\boldsymbol{v}_{0,i}$ bounds (see section 6.2.3), to improve sampling efficiency.

We again compare Hoeffding, Bennett, and Gaussian bounds to algorithm 4 with $\beta = 1.25$, but now plot, figure 6.3, confidence interval width against various measures of *cost*. In particular, we consider *affine costs* of *block size*, measured as $m\boldsymbol{c}_1 + mB\boldsymbol{c}_2$ where $\boldsymbol{c}_1$ is the cost of reading and processing each *block*, and $\boldsymbol{c}_2$ is the cost of reading and processing each *record*. Intuitively, $\boldsymbol{c}_1$ represents the networking and communication costs of *accessing* a block, and $\boldsymbol{c}_2$ represents the cost of *processing* each element of a block.

In the left plot, where $\boldsymbol{c} = (1, 0)$ (pay-per-block), we see similar trends as in the frequency estimation tasks; again algorithm 4 performance is quite near that of the Gaussian bound, and increasing block size greatly improves all variance-sensitive bounds. In the middle plot, where $\boldsymbol{c} = (\frac{9}{10}, \frac{1}{10})$, we see great improvement moving from $B = 1$ to $B = 10$, but diminishing returns thereafter, as large blocks, with their high-cost and low variance, can't compete with a larger number of low-cost small blocks. In the right plot, we have $\boldsymbol{c} = (0, 1)$ (pay-per-sample), where independent sampling outperforms the larger block sizes. This is unsurprising, as here dependent and independent data have the same cost, and due to autocorrelation of nearby read-sizes, block-averages have higher variance than averages of B independent samples (thus Gaussian performance is worse), and Bennett and algorithm 4 performance is significantly worse, as they also pay the price of having

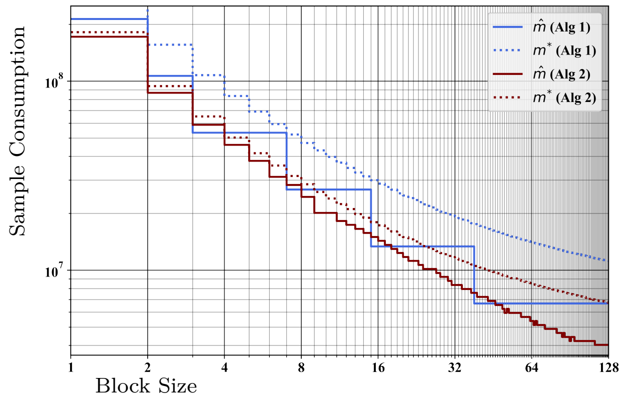


Figure 6.4: Sample Complexity versus Block Size on Mean Read Offset Task. For all $B \in 1, \ldots, 100$, we plot empirical sample consumption $\hat{m}$, and sample consumption bound m^* for both algorithm 3 and algorithm 4 with $\beta = 1.05$.

fewer independent samples in their sub-gamma scale terms.

These experiments nicely illustrate the limitations of block sampling; the improvement one gets depends entirely upon how one measures, and how one measures should depend on a specific application or architecture. In the convenient model where one pays only for blocks, one can generally expect large improvement moving from independent sampling to block sampling. With a hybrid affine cost, which may be reasonable for balancing IO and computation costs the $\boldsymbol{c} = (\frac{9}{10}, \frac{1}{10})$ plot is illustrative, perhaps with even greater gains, when compute costs are much less than $1/9$ of IO costs. Finally, the takeaway of the $\boldsymbol{c} = (0, 1)$ is almost tautological: if there is no cost benefit to operating on blocks, then we generally can't expect to benefit from block sampling.

6.3.3 Block Size and Sample Complexity

In this experiment, we estimate the *mean read offset*, truncated to 4GiB (i.e., reads above 4GiB were considered to be 4GiB), in the block-traces dataset. Like the sigmoid experiments, this is a continuous task, but here we provide no *a priori* variance bounds, and only the range bound $r = [0, 2^{32}]$B. Our goal is to see how efficient block-sampling schedules are, and how much they benefit from large block sizes.

In section 6.3.3, we plot ($\varepsilon = 0.1$,$\delta = 10^{-6}$) *sample consumption* $\hat{m}$, as well as the sample complexity bounds m^* (from theorems 6.2.1 and 6.2.3) for algorithm 3 (scaled as necessary to establish unit range), and algorithm 4 with $\beta = 1.05$. With $\beta = 1.05$, taking $B = 100$, we see that block-sampling yields a greater than 40-fold improvement in sample complexity over independent sampling. We also see that sample consumption decreases roughly hyperbolically as a function of block size (as under independence), which explains the strong performance of block sampling on this dataset.

These experiments are also quite revealing as to the differences between algorithms 3 and 4 and the importance of the geometric schedule constant β. In algorithm 3, where effectively $\beta = 2$, we see large and infrequent *discrete jumps* in sample consumption, due to the *coarse sampling schedule*. We can see how inefficient this is by comparing it to the much finer $\beta \approx 1$ sampling schedule of algorithm 4, where a smaller sample size is almost always sufficient.

6.4 Conclusion

Motivated by the fact that in modern data storage the cost of accessing a random record is essentially the same as the cost of reading a block of records, we present a rigorous technique for taking full advantage of all the record in the sampled block instead of just a random one, as done in standard sampling methods. The algorithms are progressive, continually fetching random blocks until the stopping conditions are met. This is well suited to progressively refined results when a coarse-grain result is shown almost immediately and then improved over time.

While we focused here on mean-estimation algorithms, the concept of block sampling is far more general. Our methods extend to any estimation task with a *termination condition* that can be progressively estimated at each step. In particular, with trivial modifications, our can estimate *relative confidence intervals* and *conditional means*. Similarly, in the multi-function case, (e.g., estimating averages over multiple fields), the uniform bound (FWER) over $\mathcal{F}$ is rather stringent, and it is a simple matter to instead bound the ℓ_2-distance between the point-estimate-means and true means, or provide a *false discovery rate* (FDR) guarantee, both of which are less sensitive to bad behavior of small subsets of $\mathcal{F}$.

The disaggregation of storage from compute, especially in the cloud, means that there no single model for the cost of accessing data. Large datasets are distributed across multiple storage servers, which is appropriate for parallel processing over large data sets. Parallel or distributed processing often involves higher compute costs, non-uniform startup times, and delays due to stragglers. As a result, it is sub-second response time for simple operations are difficult to achieve without block sampling.

Chapter 7

Mean Estimation and the Markov-Chain Monte-Carlo Method

I introduce a novel method for MCMC-mean estimation, whose sample complexity depends on a novel measure: the *inter-trace variance*, which characterizes the sample complexity of MCMC-mean estimation better than stationary variance (used widely in classical mean estimators). For a range R function f with stationary variance v_π, standard *variance-agnostic* MCMC mean-estimators run the chain for $\tilde{\mathcal{O}}(\frac{TR^2}{\varepsilon^2})$ steps, when given as input an (often loose) upper-bound T on the *relaxation time* τ_{rel} ($\leq$ mixing time τ_{mix}), and when an upper-bound V on v_π is known, $\tilde{\mathcal{O}}\big(\frac{TR}{\varepsilon} + \frac{TV}{\varepsilon^2}\big)$ steps suffice.

For an arbitrary trace length τ, I define the *inter-trace variance* $\mathrm{trv}^{(\tau)}$, and using it I introduce the DYNAmic MCMC Inter-Trace variance Estimation (DYNAMITE) algorithm for mean-estimation. DYNAMITE estimates w.h.p. the mean to within ε additive error in $\tilde{\mathcal{O}}\big(\frac{TR}{\varepsilon} + \frac{\tau_{\mathrm{rel}} \cdot \mathrm{trv}^{(\tau_{\mathrm{rel}})}}{\varepsilon^2}\big)$ steps, without *a priori* variance bounds (w.r.t. f) on the *stationary variance* v_π, or the *trace variance* $\mathrm{trv}^{(\tau_{\mathrm{rel}})}$. When ε is small, the dominating complexity term is $\tau_{\mathrm{rel}} \cdot \mathrm{trv}^{(\tau_{\mathrm{rel}})}$, thus the complexity of DYNAMITE principally depends on the *a priori unknown* τ_{rel} and $\mathrm{trv}^{(\tau_{\mathrm{rel}})}$. I show that often $\mathrm{trv}^{(T)} = o(v_\pi)$, and furthermore, it always holds that $\mathrm{trv}^{(\tau_{\mathrm{rel}})} \leq 2v_\pi$, thus the worst-case complexity of DYNAMITE is $\tilde{\mathcal{O}}(\frac{TR}{\varepsilon} + \frac{\tau_{\mathrm{rel}} \cdot v_\pi}{\varepsilon^2})$, improving the dependence of classical methods on the loose bounds T and V. This chapter is adapted from Cousins, Haddadan, and Upfal [2020].

7.1 Introduction and Related Work

Assume we have a bounded, real-valued function $f : \Omega \to [a, b]$, and a distribution π defined on the domain Ω, and our goal is to estimate the average of this function, $\mu \doteq \mathbb{E}_\pi[f(x)]$. Having enough samples from π, one can estimate μ by taking the average of such samples, usually referred to as the *empirical mean* $\hat{\mu}$. The accuracy of this estimate depends on the variance of f and the number of samples. Markov chains are often used to *approximately sample* from π efficiently Aldous et al. [1997], Brightwell and Winkler [1991], Wilson [2004], Randall [2006], Levin and Peres [2017], which can be used to construct a FPRAS for many #P-hard problems which aim to find the size of Ω.

Classical *Markov Chain Monte Carlo* (MCMC) approaches work by analytically determining a sufficient quantity m of approximately-independent samples to estimate μ, and then generating such samples by running a Markov chain. One approach is to run m copies of a Markov chain $\mathcal{M}$ converging to a stationary distribution π for T steps, where T is chosen to exceed the *mixing time* τ_{mix}, which ensures each sample is approximately π-distributed. Another approach is to run a *single instance* of a Markov chain for a larger number of steps, and estimate μ from these samples (rather than restarting the chain every T steps to collect a single sample). In the latter approach, m is taken proportional to an upper-bound T on the *relaxation time* τ_{rel} ($\leq \tau_{\text{mix}}$); both τ_{mix} and τ_{rel} are formally defined in section 7.2. Unfortunately, because MCMC is generally used with distributions too complicated to analyze precisely, we lack sharp bounds on $v_\pi \doteq \mathbb{V}_\pi[f]$, and furthermore, it is often difficult to sharply bound τ_{rel} (or τ_{mix}) analytically Guruswami [2016], Levin and Peres [2017], Lovász and Winkler [1998].

A series of papers analyze the sample complexity of such approaches Lezaud [1998], Miasojedow [2012], Paulin [2015], Jiang et al. [2018], Mitzenmacher and Upfal [2017], Chung et al. [2012], Fan et al. [2018], Leon and Perron [2004], Peskun [1973], finding it to be $\tilde{\mathcal{O}}\left(T({}^{V}\!/\!_{\varepsilon^2} + {}^{R}\!/\!_{\varepsilon})\right)$ for variance-aware bounds, where $V \geq v_\pi$, $r \doteq b - a$, and $\tilde{\mathcal{O}}$ hides log factors. For variance-agnostic bounds, which assume no prior knowledge of v_π, the bound becomes $\tilde{\mathcal{O}}({}^{TR^2}\!/\!_{\varepsilon^2})$ (see theorems 7.2.5 and 7.2.6 for more details). The drawbacks of these approaches are twofold: (1) among these algorithms, the more-efficient variance-aware methods need an *a priori* bound $V \geq v_\pi$, and (2) the *complexity* of these methods depend on the upper-bounds they receive on *relaxation time* and *the stationary variance* as input, and are thus inefficient when given loose upper-bounds.

Mean estimation is an interesting problem, well-studied in theoretical computer science, and has also borne fruit as a tool for solving other difficult theoretical computer science problems. For example, Jerrum, Valiant, and Vazirani reduced estimating the *size* of a *self-reducible set* to solving a series of *mean-estimation* tasks, and ongoing work in the area seeks to reduce the complexity of such reductions Stefankovic et al. [2007], Huber [2015], David G. Harris [2007], Kolmogorov [2018]. Furthermore, there has been an emerging line of research to analyze *sublinear algorithms* having query access to input graphs using random walks Chierichetti and Haddadan [2018], Ben-Hamou et al. [2018], Bera and Seshadhri [2020], for example to estimate the mean of a real-valued function on vertices Chierichetti and Haddadan [2018]. In addition, mean estimation using MCMC has enormous applications in other fields of science such as *computational biology* Vandin et al. [2016, 2012a,b, 2011], Enright et al. [2002], Azad et al. [2018], *chemistry* Gillespie [2007], *statistical physics* Del Moral and Miclo [2007], Mignani and Rodolfo [2001], Diaconis [2009], Navarro et al. [2016], and various subfields of *computer science* e.g. *statistical learning*, *image processing*, etc Fan et al. [2013], Salatino et al. [2019], Yuan and Kendziorski [2006], Tu and Zhu [2002].

In this work, we show that the *inter-trace variance* $v_\cdot$ better characterizes MCMC sample-complexity than does the *stationary variance* v_π (we formally define v_τ for arbitrary τ in section 7.2.1). In particular, we present and analyze the DynaMITE algorithm, which begins by running a Markov chain for $T \geq \tau_{\mathrm{rel}}$ steps, and *averaging* f along this trace. We call any τ consecutive steps of a chain a τ*-trace* (dropping τ when it is clear from context). Using a *progressive sampling schedule*, DynaMITE repeatedly doubles the chain length, taking more and more samples $f^1_{\mathrm{avg}}, f^2_{\mathrm{avg}}, \ldots$, each f^i_{avg} being the average of the ith T-trace. At each iteration, DynaMITE approximates the inter-trace variance v_T by empirically estimating $\mathbb{V}[f_{\mathrm{avg}}]$, and stops when the termination condition is satisfied.

We show that w.h.p., the sample consumption of DynaMITE is $\tilde{\mathcal{O}}\big(\frac{Tr}{\varepsilon} + \frac{T\cdot v_T}{\varepsilon^2}\big) = \tilde{\mathcal{O}}\big(\frac{Tr}{\varepsilon} + \frac{\tau_{\mathrm{rel}}\cdot v_{\tau\mathrm{rel}}}{\varepsilon^2}\big)$, where $v_{\tau\mathrm{rel}}$ is the trace variance of a τ_{rel}-trace. Note that this improves the sample complexity of traditional variance-agnostic approaches, and also dominates variance-aware approaches when ε is small, *despite assuming no a priori knowledge of* v_T. Moreover, the complexity of DynaMITE is primarily dependent on the *a priori unknown* relaxation time and inter-trace variance, rather than their loose upper bounds, which limits the efficiency of standard approaches. Note that while T appears in DynaMITE and its sample complexity bound, its contribution is negligible for small ε, and a loose T bound is compensated for by earlier termination.

Our goal in this paper is to demonstrate the significance of using the *inter trace variance* v_τ instead of the stationary variance v_π. We develop techniques to theoretically guarantee $v_{\tau_{\mathrm{rel}}} = o(v_\pi)$ in specific scenarios, and then prove this phenomenon occurs in example 7.2.2 (see lemma 7.2.3).

Road map The paper is organized as follows: Section 7.2 contains all the preliminaries, important definitions, and classical results that we use in our proofs. In section 7.3, we introduce our algorithm, and provide a proof sketch for its correctness and efficiency. Finally, proofs are presented in full detail in appendix F.1.

Our Contributions:

1. We introduce the *inter-trace variance* $v_\cdot$, and show that $v_{\tau\mathrm{rel}}$ characterizes the sample complexity of MCMC-mean estimators better than v_π. It always holds that $v_{\tau\mathrm{rel}} \leq 2v_\pi$, furthermore, we identify conditions under which $v_{\tau\mathrm{rel}} = o(v_\pi)$. We conjecture that this is common, and show a technique for bounding $v_{\tau\mathrm{rel}}$, when a simpler *projection chain* approximates f's image on MCMC-traces.

2. Leveraging $v_\cdot$, we design DynaMITE. Under the standard MCMC assumptions of known relaxation time bound $T \geq \tau_{\mathrm{rel}}$ and bounded range r, DynaMITE requires $\tilde{\mathcal{O}}\big(\frac{Tr}{\varepsilon} + \frac{\tau_{\mathrm{rel}} \cdot v_{\tau\mathrm{rel}}}{\varepsilon^2}\big)$ samples, *without assuming* any *a priori* knowledge of τ_{rel}, $v_{\tau\mathrm{rel}}$ or v_T. For small ε, note that while DynaMITE takes a loose upper bound on τ_{rel} as input, its complexity depends on τ_{rel} and $v_{\tau\mathrm{rel}}$; this is particularly impactful, as practical MCMC efficacy is often limited by loose bounds on τ_{rel}, and sharply upper bounding $v_{\tau\mathrm{rel}}$ is seldom an easy task.

3. DynaMITE's sample complexity is always better than classic MCMC mean-estimators, since it reduces the dependency on loose upper bounds T and V. Furthermore, if $v_T = o(v_\pi)$, then DynaMITE is asymptotically better than classic algorithms. We note that complexity of DynaMITE depends on $v_{\tau\mathrm{rel}}$ oblivious to any prior knowledge of $v_{\tau\mathrm{rel}}$.

7.2 Preliminaries

In this section, we first introduce notation that we use throughout the text and relevant theorems from the literature. Throughout this work, we take $f : \mathcal{X} \to [a, b]$ to be a function, and π a probability distribution on $\mathcal{X}$. We define the *mean* and *stationary variance* of f as $\mu \doteq \mathbb{E}_{x\sim\pi}[f] \doteq \int_{\mathcal{X}} f(x)\,\mathrm{d}\pi(x)$ and $v_\pi \doteq \mathbb{V}_\pi(f) = \mathbb{E}_{x\sim\pi}[(f(x) - \mu)^2]$, respectively. An *MCMC mean-estimator* estimates μ by collecting samples running a Markov chain.

7.2.1 The Inter-Trace Variance

MCMC Terminology A Markov chain $\mathcal{M}$ on sample space $\mathcal{X}$ is a sequence of random variables $X_1, X_2, \ldots$ that is *memoryless*. The evolution of $\mathcal{M}$ is characterized by its *transition matrix*; here we abuse notation and also use $\mathcal{M}$ to represent its transition matrix, i.e., for any $x, y \in \mathcal{X}$, $\mathcal{M}(x, y) \doteq \mathbb{P}(X_i = y | X_{i-1} = x)$, and thus $\mathbb{P}(X_{i+k} = y | X_i = x) = \mathcal{M}^k(x, y)$, where $\mathcal{M}^k$ is the standard matrix powering notation. When ν is a probability distribution on $\mathcal{X}$, by $\mathcal{M}(\nu)$ we mean the distribution of X_i when $X_{i-1} \sim \nu$. The distribution of the chain after k steps of $\mathcal{M}$ is denoted by $X_{i+k} = \mathcal{M}^k(X_i)$. For arbitrary τ, we call a sequence of random variables generated by $\mathcal{M}$ a *τ-trace* of Markov chain $\mathcal{M}$. We use *vector notation* to denote traces of Markov chains, i.e., $\boldsymbol{X}_{1:\tau} : X_1, X_2, \ldots X_\tau$ is a τ-trace of $\mathcal{M}$ (we drop the subscript $1:\tau$ when the length is clear from context). We assume all the Markov chains we discuss are ergodic and lazy[1] (see Levin and Peres [2017] if not familiar with definitions), and denote the *stationary distribution* by π. For ergodic chains, and w.r.t. a precision parameter ε, the *mixing time*, denoted by $\tau_{\text{mix}}(\varepsilon)$, is defined as the minimum τ satisfying $\text{TVD}(\mathcal{M}^\tau(\nu), \pi) \leq \varepsilon$ for any arbitrary starting distribution ν, and we take $\tau_{\text{mix}} \doteq \tau_{\text{mix}}(1/4)$. It is known that $\tau_{\text{mix}}(\varepsilon) \leq \tau_{\text{mix}} \log(\varepsilon^{-1})$. We denote the *second largest absolute eigenvalue* $\mathcal{M}$ by λ, and the *relaxation time* by τ_{rel}. Note that $\tau_{\text{rel}} \doteq (1-\lambda)^{-1}$ and it is closely related to τ_{mix}, as $\big(\tau_{\text{rel}}(\mathcal{M}) - 1\big) \ln(2) \leq \tau_{\text{mix}}(\mathcal{M}) \leq \tau_{\text{rel}}(\mathcal{M})\ \ln\big(2/\sqrt{\pi_{\min}}\big)$, see Levin and Peres [2017]. We use T to denote a known upper-bound on the relaxation time.

We are interested in designing algorithms to estimate μ using samples from π collected by running $\mathcal{M}$. Assume $\boldsymbol{X} : X_1, X_2, \ldots, X_\tau$ is an τ-trace of $\mathcal{M}$, we define the *empirical mean* $\hat{\mu}(\boldsymbol{X}) \doteq \frac{1}{\tau}\sum_{i=1}^{\tau} f(X_i)$. Note that $\hat{\mu}$ is an *unbiased estimator* for μ, i.e., $\mathbb{E}[\hat{\mu}(\boldsymbol{X})] = \mu$. Since $\boldsymbol{X}$ are not i.i.d. samples, we can not use the standard Bessel's correction as an unbiased variance estimator. Therefore, we devise a novel MCMC *unbiased estimator for variance*, which is based on running two independent Markov chains in parallel, and averaging the square difference of f between them (see definition 7.3.4).

We now introduce the *inter-trace variance* $v_\cdot$, which is the backbone of our proposed algorithm DynaMITE. Letting T be an upper-bound on τ_{rel}, DynaMITE uses progressive sampling to estimate v_T, thus it requires no *a priori* knowledge of v_T, $v_{\tau_{\text{rel}}}$ or v_π.

Inter-Trace Variance Let τ be an arbitrary length. On the set of all τ-traces, we define the probability distribution $\pi^{(\tau)}$ as $\pi^{(\tau)}(\boldsymbol{X}_{1:\tau}) \doteq \pi(X_1)\prod_{i=1}^{\tau-1}\mathcal{M}(X_i, X_{i+1})$. For an arbitrary F defined on Ω^τ, by $\mathbb{E}_{\boldsymbol{X}\sim\pi^{(\tau)}}[F(\boldsymbol{X})]$, we mean the expectation of F on any τ-trace of the Markov chain when $X_1 \sim \pi$, and other consecutive X_is follow the transition of the Markov chain. For arbitrary τ and $\boldsymbol{X}_{1:\tau}$, we define $f_{\text{avg}}(\boldsymbol{X}_{1:\tau}) \doteq \frac{1}{\tau}\sum_{i=1}^{\tau} f(X_i)$. Note that by linearity of expectation, $\mathbb{E}_{\boldsymbol{X}\sim\pi^{(\tau)}}[f_{\text{avg}}(\boldsymbol{X})] = \mu$. We denote the variance of f_{avg} by v_τ, i.e., $v_\tau \doteq \mathbb{E}_{\boldsymbol{X}\sim\pi^{(\tau)}}[(f_{\text{avg}}(\boldsymbol{X}) - \mu)^2]$. Note that under independence, we have $v_\tau = \frac{v_\pi}{\tau}$, and for mixing processes (see Thm. 3.1 of Paulin [2015]), we get

$$\frac{1}{\tau} v_\pi \leq v_\tau \leq \frac{2\tau_{\text{rel}}}{\tau} v_\pi \quad . \tag{46}$$

Receiving as input T, an upper-bound on relation time, DynaMITE uses averages along T-traces of $\mathcal{M}$ and it estimates v_T. We show that DynaMITE's sample complexity is dominated by $v_{\tau\text{rel}}$, thus when $v_{\tau\text{rel}}$ is significantly smaller than the stationary variance (i.e., $v_{\tau\text{rel}} = o(v_\pi)$), DynaMITE outperforms classical

[1]We explicitly note when reversibility is needed.

methods. We show this occurs, when f partitions the state space of $\mathcal{M}$ in such a way that the *projection chain* (defined below in definition 7.2.1) has a much smaller relaxation time ($v_{\tau\text{rel}} = o(v_\pi)$).

Definition 7.2.1 (Projection chain Levin and Peres [2017])**.** Having an equivalence relationship $\simeq$ on $\mathcal{X}$ and classes $\tilde{\mathcal{X}} = \{[x]; x \in \mathcal{X}\}$ such that if $x \simeq x'$, then $\mathcal{M}(x, [y]) = \mathcal{M}(x', [y])$, we call the Markov chain $\tilde{\mathcal{M}}$ with state space $\tilde{\mathcal{X}}$ transition probabilities $\tilde{\mathcal{M}}([x], [y]) = \mathcal{M}(x, [y])$ a *projection chain* (see section 2.3 of Levin and Peres [2017] for full discussion).

In example 7.2.2, we introduce a class of functions having equal *means* and *stationary variances*, but projecting differently on traces of a fixed chain. Thus we can prove that while they all have equal v_π, under appropriate parameterization, $v_{\tau\text{rel}}$ can take any arbitrary value in the range of equation 46.

Example 7.2.2. Consider the Markov chain $\mathcal{C}$, known as the cycle, defined on $[n] = \{1, 2, \ldots, n\}$ having transition probabilities: $\mathcal{C}(i, i) = 1/2$, $\mathcal{C}(i, i+1) = 1/4$ and $\mathcal{C}(i, i-1) = 1/4$, where $i-1$ and $i+1$ are taken mod n. Clearly the stationary distribution on this chain is uniform, and it is well known that the mixing time and relaxation time are both $\Theta(n^2)$.

The following class of functions defined on $[n]$ all satisfy $\mathbb{E}_\pi(f_i) = 1/2$ and $v_\pi(f_i) = 1/4$. However, as shown in lemma 7.2.3, each has a different value for $v_{\tau\text{rel}}$, covering the entire spectrum in equation 46:

For $1 \le i \le \frac{n}{2}$, let $f_i : [n] \to \{0, 1\}$ be $f_i(x) = 0$ if and only if $x \bmod 2i < i$, so f_i's image on the cycle is consecutive length-i runs of 0s and 1s (see figure 7.1). Note that in the two extreme cases, we have, (1) $f_1 : [n] \to \{0, 1\}, f_1(x) = 0$ if and only if $x \bmod 2 = 0$, and (2) $f_{n/2} : [n] \to \{0, 1\}, f_{n/2}(x) = 0$ if and only if $x \le \frac{n}{2}$.

On equivalence classes of mod 2, we find the corresponding projection chain, denoted by $\mathcal{C}_{f_1}$, which has transition probabilities: $\mathcal{C}_{f_1}(0, 1) = \mathcal{C}_{f_1}(1, 0) = 1/4$ and $\mathcal{C}_{f_1}(0, 0) = \mathcal{C}_{f_1}(1, 1) = 1/2$. Similarly, for arbitrary i, we denote the projection chain on equivalence classes of mod $2i$ by $\mathcal{C}_{f_i}$. Clearly the trace variance of f_i on $\mathcal{C}$ and on $\mathcal{C}_{f_i}$ are the same. Figure 7.1 illustrates these projections.

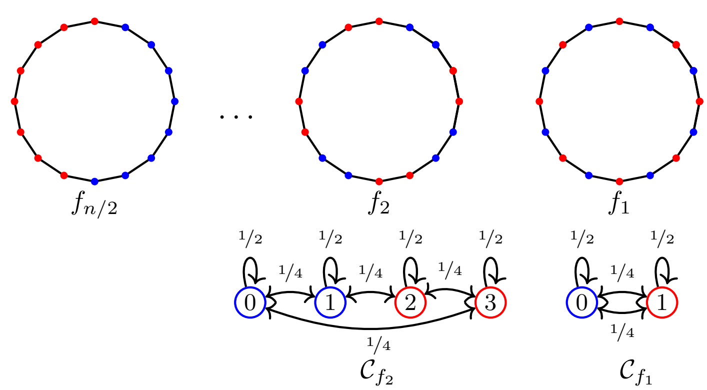


Figure 7.1: The image of f_is on a cycle of length $n = 16$, and corresponding *projection chains* with the same trace variance.

We now present the following lemma with a proof sketch. The complete proof is presented in the appendix.

Lemma 7.2.3. For $1 \leq i \leq \frac{n}{2}$, let f_i be defined as above, for arbitrary τ we have $v_\tau(f_i) = \Theta({}^{i^2}\!/\tau)$. In particular, for $\tau = \Theta(\tau_{\mathrm{rel}}(\mathcal{M})) = \Theta(n^2)$, we have $v_\tau(f_{n/2}) = \Theta(1)$, and $v_\tau(f_1) = \Theta({}^1\!/n^2)$.

Proof sketch. We first prove that $v_\tau(f_i) \leq \Theta({}^{i^2}\!/\tau)$. Note that $x \mapsto x \bmod 2i$ partitions the set $[n]$ into $2i$ partitions, and the relaxation time of the projection chain $\mathcal{C}_{f_i}$ is $\Theta(i^2)$. Since the trace variance of f_i on $\mathcal{C}$ and $\mathcal{C}_{f_i}$ are the same, by applying inequality 46, we get $v_\tau(f_i) \leq \Theta({}^{i^2}\!/\tau)$.

We now proceed to prove that $v_\tau(f_i) \geq \Theta({}^{i^2}\!/\tau)$. Let $S_0 = \{i/3+1, i/3+2, \ldots, 2i/3\} \subseteq [n]$. In the appendix by an application of Lévy's inequality and Hoeffding's inequality, we show that the trace variance of any trace of length at most i^2 conditioned starting at S_0 is $\Theta(1)$. Similarly, we define S_k to be $S_k \doteq \{k(2i)+(i/3+1), k(2i)+(i/3+2), \ldots, k(2i)+2i/3\}$, and using a similar argument we can show that any walk starting at $S = \bigcup_{k=0}^{n/2i} S_k$ has trace variance at least constant. Since the stationary probability of S is $\frac{1}{3}$, we conclude $v_\tau \geq \Theta(1)$. See Figure 7.2 for visualization of f_i on $\mathcal{C}$, and the region S. $\square$

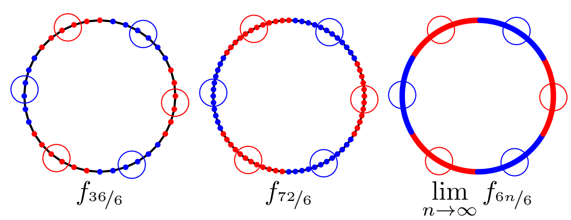


Figure 7.2: Image of $f_{n/6}$ on a cycle for various values of n, zero values are colored in red, and one values in blue. The trace variance for trace length than $(n/52)^2$, conditioned starting at $S = \bigcup_{k=1}^{6} S_k$ (circled regions), is constant.

7.2.2 Classic MCMC-Mean Estimators

In contrast to our proposed algorithm DynaMITE, which increases the sample size dynamically, the classic methods are *static*; they receive a fixed m as input and run the chain for m steps to generate the trace $\boldsymbol{X}_{1:m} : X_1, X_2, \ldots, X_m$. These algorithms then returns the empirical mean $\hat{\mu}$ as the output.

In the following theorems $\boldsymbol{X}_{1:m} : X_1, X_2, \ldots, X_m$, is a length m stationary trace of $\mathcal{M}$, (i.e., $\boldsymbol{X}_{1:m} \sim \pi^{(T)}$), with mixing time τ_{mix}, relaxation time τ_{rel}, and second largest eigenvalue λ. Based on each of these theorems we obtain the sample complexity (sufficient m) of a variance-aware or variance-agnostic algorithm. Remember that f is a bounded function with range $[a,b]$, we define $r \doteq b - a$ as the range of f. We use the following definition for *sample complexity*.

Definition 7.2.4 (Sample complexity)**.** Assume algorithm A runs a Markov chain for m steps and generates a trace $X_1, X_2, \ldots, X_m$. Algorithm A uses this trace to find an estimation for μ, namely $\hat{\mu}$. If there exists m_A such that for any $m \geq m_A$, we have $\mathbb{P}(|\mu - \hat{\mu}| \geq \varepsilon) \leq \delta$, we call m_A the *sample complexity* of algorithm A.

Theorem 7.2.5 (Hoeffding-Type Bounds for Mixing Processes, (see Thm. 2.1 of [Fan et al., 2018])). For any $\delta \in (0,1)$, we have

$$\mathbb{P}\left(|\hat{\mu} - \mu| \geq \sqrt{\frac{2(1+\lambda)(\frac{r^2}{4})\ln(\frac{2}{\delta})}{(1-\lambda)m}}\right) \leq \delta \ . \tag{47}$$

This implies sample complexity

$$m_H(\lambda, r, \varepsilon, \delta) = \frac{1+\lambda}{1-\lambda}\ln(\tfrac{2}{\delta})\frac{r^2}{2\varepsilon^2} \in \Theta\Big(\tau_{\mathrm{rel}}\ln(\tfrac{1}{\delta})\frac{r^2}{\varepsilon^2}\Big) \ .$$

Largely due to their simplicity and convenience, Hoeffding-type bounds enjoy ubiquity in computer science. Unfortunately, they are generally much looser than variance-aware bounds, as the dependence on variance v is replaced by *variance proxy* $(\frac{r}{2})^2$ (maximal assuming range r). While it is not generally possible to remove dependence on r entirely, Bernstein-type inequalities greatly reduce this dependence, often yielding *asymptotic improvement* to sample complexity Audibert et al. [2007], Mnih et al. [2008], Audibert et al. [2009], Maurer and Pontil [2009].

Theorem 7.2.6 (Bernstein-Type Bound for Mixing Process [Jiang et al., 2018, Thm. 1.2]). [2] For any $\delta \in (0,1)$, we have

$$\mathbb{P}\left(|\hat{\mu} - \mu| \geq \frac{10r\ln(\frac{2}{\delta})}{(1-\lambda)m} + \sqrt{\frac{2(1+\lambda)v_\pi\ln(\frac{2}{\delta})}{(1-\lambda)m}}\right) \leq \delta \ . \tag{48}$$

Sample complexity of the static variance-aware algorithm: This implies sample complexity

$$m_B(\lambda, r, v, \varepsilon, \delta) = \frac{2}{1-\lambda}\ln(\tfrac{2}{\delta})\Big(\frac{5r}{\varepsilon} + \frac{(1+\lambda)v_\pi}{\varepsilon^2}\Big) \in \Theta\Big(\tau_{\mathrm{rel}}\ln(\tfrac{1}{\delta})\Big(\frac{r}{\varepsilon} + \frac{v_\pi}{\varepsilon^2}\Big)\Big) \ .$$

7.3 DynaMITE

In this section we present the DynaMITE: DYNAmic Mcmc Inter-Trace variance Estimation method. We show that its sample complexity is dependent on the *a priori unknown* trace variance v_T. As discussed previously, in various scenarios, v_T is significantly smaller than the stationary variance v_π.

Our algorithm consists of three subroutines: (1) McmcPro employs *progressive sampling*, beginning with a small sample obtained by running the chain, and we *progressively* increase the sample size until a *stopping condition* is met, namely the mean is ε-δ well-estimated. In each round of McmcPro, we bound v_T using

[2] Note that thms. 7.2.5 and 7.2.6 assume *stationarity*, i.e., they consider traces $\boldsymbol{X}_{1:m} : X_1, X_2, \ldots X_m$ assuming $X_1 \sim \pi$. This can be done using what is known as a *warm start*. For more details see subsection F.1.1 in the appendix.

an unbiased estimator of it, which we obtain from running two *independent copies* of $\mathcal{M}$ in parallel (see definition 7.3.4). In order to find a suitable stopping conditioning guaranteeing that the empirical mean is sufficiently accurate, we employ the Bernstein bound (theorem 7.2.6). (2) Our main procedure DYNAMITE runs the above progressive schedule on a *trace chain* (defined in definition 7.3.3), using the *trace averaging technique* explained in section 7.2. (3) Finally, in procedure WARMSTARTDYNAMITE, we begin at an arbitrary (nonstationary) state, but run it chain for at least the *uniform mixing time* τ_{unif} steps before starting sampling, and apply an appropriate *nonstationarity correction.* Note that this is performed only once, and does not significantly impact sample complexity, which thus depends on τ_{rel} rather than τ_{unif}.

The following theorems, proved in appendix F.1, guarantee correctness and efficiency of DYNAMITE.

Theorem 7.3.1 (Correctness of DYNAMITE)**.** Consider a Markov chain $\mathcal{M}$ over Ω, and assume its second absolute eigenvalue is less than Λ. Assume we have function $f : \Omega \to [a, b]$ with $r \doteq b - a$, and we want to estimate *true mean*, $\mu \doteq \mathbb{E}_{\pi}[f]$ with precision $\varepsilon > 0$, and confidence $(1 - \delta) \in (0, 1)$.

If we start at stationarity, i.e., $(x_0, x_1) \sim \pi(\mathcal{M} \otimes \mathcal{M})$, we take *mean estimate* $\hat{\mu}$ as either

1. $\hat{\mu} \leftarrow \text{MCMCPRO}((x_0, x_1), \mathcal{M}, \Lambda, f, \varepsilon, \delta)$; or
2. $\hat{\mu} \leftarrow \text{DYNAMITE}((x_0, x_1), \mathcal{M}, \Lambda, f, \varepsilon, \delta)$ for lazy $\mathcal{M}$.

More generally, for a nonstationary start from any $\omega \in \mathrm{Support}(\pi(\mathcal{M}))$, given minimum supported stationary probability at least $\pi_{\min}$, we may take

3. $\hat{\mu} \leftarrow \text{WARMSTARTDYNAMITE}(\omega, \mathcal{M}, \Lambda, \pi_{\min}, f, \varepsilon, \delta)$ for lazy, reversible $\mathcal{M}$.

In all cases, with probability at least $1 - \delta$, we have $|\hat{\mu} - \mu| \leq \varepsilon$.

Theorem 7.3.2 (Efficiency of DYNAMITE)**.** Suppose as in theorem 7.3.1. With probability at least $1 - \delta$, it holds that total sample consumption $\hat{m}$ of DYNAMITE obeys

$$\hat{m} \in \mathcal{O}\left(\log\left(\frac{\log(r/\varepsilon)}{\delta}\right)\left(\frac{r}{(1-\Lambda)\varepsilon} + \frac{\tau_{\mathrm{rel}} v_{\tau\mathrm{rel}}}{\varepsilon^2}\right)\right)$$

Note on input parameters and their interplay with efficiency The improvement of DYNAMITE over standard methods discussed in section 7.2 is threefold: (1) First, in MCMCPRO, we develop a *dynamic sampling schedule*, which is called in line 24 of DYNAMITE. Note that MCMCPRO also works if we simply run it on the original chain. In this simpler version, the sample complexity depends on the *a priori unknown* v_{π}. Thus we get an improvement over standard methods, whose complexity depend on loose upper bound V. (2) In DYNAMITE, we employ the trace averaging technique to reduce the sample complexity from Tv_{π} to

Algorithm 5 MCMCPRO, DYNAMITE, and WARMSTARTDYNAMITE routines

procedure MCMCPRO((X_0, X_1), $\mathcal{M}$, Λ, f, ε, δ) $\mapsto \hat{\mu}$

  Input: Initial state $(X_0, X_1) \in \mathcal{X}^2$ of tensor product chain, *Markov chain* $\mathcal{M}$ over $\mathcal{X}$, *second absolute eigenvalue upper-bound* Λ, *function* $f : \mathcal{X} \to [a, b]$ with $r \doteq b - a$, *confidence interval radius* ε, and *failure probability* $\delta \in (0, 1)$.

  Output: Additive ε, δ approximation $\hat{\mu}$ of $\mu = \mathbb{E}_\pi[f]$.

  $N \leftarrow \max\Big(1, \lfloor \log_2\big(\frac{r}{2\varepsilon}\big) \rfloor\Big)$; $\alpha \leftarrow \dfrac{(1+\Lambda) r \ln \frac{3N}{\delta}}{(1-\Lambda)\varepsilon}$; $m_0 \leftarrow 0$ ▷ Initialize *sampling schedule*

  for $i \in 1, 2, \ldots, N$ **do**

    $m_i \leftarrow \lceil \alpha 2^i \rceil$ ▷ Total sample count at iteration i

    for $j \in (m_{i-1} + 1), \ldots, m_i$ **do**

      $(X_{j,1}, X_{j,2}) \sim (\mathcal{M} \otimes \mathcal{M})(X_{j-1,1}, X_{j-1,2})$ ▷ Run both chains up to step m_i

    end for

    $\hat{\boldsymbol{\mu}}_i \leftarrow \frac{1}{2m_i} \sum_{j=1}^{m_i} (f \oplus f)(X_{j,0}, X_{j,1})$ ▷ Empirical mean

    $\hat{\boldsymbol{v}}_i \leftarrow \frac{1}{2m_i} \sum_{j=1}^{m_i} (f \ominus f)^2(X_{j,0}, X_{j,1})$ ▷ Empirical variance

    $\boldsymbol{u}_i \leftarrow \hat{\boldsymbol{v}}_i + \dfrac{(11+\sqrt{21})(1+ {}^{\Lambda}\!/\!\sqrt{21}) r^2 \ln \frac{3N}{\delta}}{(1-\Lambda) m_i} + \sqrt{\dfrac{(1+\Lambda) r^2 \hat{\boldsymbol{v}}_i \ln \frac{3N}{\delta}}{(1-\Lambda) m_i}}$ ▷ Variance upper bound

    $\hat{\boldsymbol{\varepsilon}}_i \leftarrow \dfrac{10 r \ln \frac{3N}{\delta}}{(1-\Lambda) m_i} + \sqrt{\dfrac{(1+\Lambda) \boldsymbol{u}_i \ln \frac{3N}{\delta}}{(1-\Lambda) m_i}}$ ▷ Apply Bernstein bound

    if $(i = N) \vee (\hat{\boldsymbol{\varepsilon}}_i \leq \varepsilon)$ **then** ▷ Terminate if accuracy guarantee is met

      return $\hat{\boldsymbol{\mu}}_i$

    end if

  end for

end procedure

---

procedure DYNAMITE((x_0, x_1), $\mathcal{M}$, Λ, f, ε, δ) $\mapsto \hat{\mu}$

  Input: Initial state $(x_0, x_1) \in \mathcal{X} \times \mathcal{X}$, *lazy Markov chain* $\mathcal{M}$ over $\mathcal{X}$, *second absolute eigenvalue upper-bound* Λ, *function* $f : \mathcal{X} \to [a, b]$, *confidence interval radius* ε, and *failure probability* $\delta \in (0, 1)$.

  Output: Additive ε, δ approximation $\hat{\mu}$ of $\mu = \mathbb{E}_\pi[f]$.

  $T \leftarrow \lceil \frac{1+\Lambda}{1-\Lambda} \ln \sqrt{2} \rceil$ ▷ Select T s.t. $\tau_{\mathrm{rel}}(\mathcal{M}^{(T)}) \leq 2$

  $(X_0, X_1) \leftarrow \Big((\underbrace{\omega_1, \omega_2, \omega_3, \ldots, \omega_{T-1}}_{\text{ARBITRARY } \omega_{1:T-1} \in \Omega^{T-1}}, x_0), (\underbrace{\omega_1, \omega_2, \omega_3, \ldots, \omega_{T-1}}_{\text{ARBITRARY } \omega_{1:T-1} \in \Omega^{T-1}}, x_1) \Big)$ ▷ Initialize trace chain state

  return MCMCPRO$\big((X_0, X_1), \mathcal{M}^{(T)}, \Lambda^T, f_{\mathrm{avg}}, \varepsilon, \delta\big)$ ▷ Run MCMCPRO on trace chain

end procedure

---

procedure WARMSTARTDYNAMITE(ω_0, $\mathcal{M}$, Λ, $\pi_{\min}$, f, ε, δ) $\mapsto \hat{\mu}$

  Input: Initial state $\omega_0 \in \mathrm{Support}(\pi(\mathcal{M}))$, lazy reversible *Markov chain* $\mathcal{M}$ over $\mathcal{X}$, *second absolute eigenvalue upper-bound* Λ, *minimum probability lower-bound* $\pi_{\min}$, *function* $f : \mathcal{X} \to [a, b]$, *confidence interval radius* ε, and *failure probability* $\delta \in (0, 1)$.

  Output: Additive ε, δ approximation $\hat{\mu}$ of $\mu = \mathbb{E}_\pi[f]$.

  $\tau_{\mathrm{unif}} \leftarrow \left\lceil \dfrac{\ln \frac{1}{\pi_{\min}}}{\ln \frac{1}{\Lambda}} \right\rceil$ ▷ *Uniform mixing time* bound

  $(x_0, x_1) \sim (\mathcal{M} \otimes \mathcal{M})^{\tau_{\mathrm{unif}}}(\omega_0, \omega_0)$ ▷ Run product chain until *uniformly mixed*

  return DYNAMITE$\big((x_0, x_1), \mathcal{M}, \Lambda, f, \varepsilon, \frac{\delta}{4}\big)$ ▷ Run DYNAMITE with nonstationarity correction

end procedure

Tv_T, thus, if $v_T = o(v_\pi)$ we obtain asymptotic improvement over standard methods (again we require no *a priori* knowledge of v_T). (3) Since it always holds $Tv_T \leq 2\tau_{\mathrm{rel}} v_{\tau\mathrm{rel}}$, DynaMITE's complexity depends on $\tau_{\mathrm{rel}} v_{\tau\mathrm{rel}}$, i.e., the *a priori unknown* τ_{rel} and $v_{\tau\mathrm{rel}}$. Thus, even in the worst case when $v_{\tau\mathrm{rel}} \approx v_\pi$, our algorithm outperforms static MCMC-mean estimators whose complexity is TV (see theorem 7.2.6 in section 7.2).

Starting at Nonstationarity Often we can't assume even a single (perfectly) stationary sample may be efficiently drawn, and realistically can only start with an *arbitrary* $\omega \in \mathrm{Support}(\pi)$. WarmStartDynaMITE uses a standard *warm start technique* (widely used in MCMC algorithms), where the chain is run from an arbitrary point for its *uniform mixing time*, and then DynaMITE is run, applying a standard *nonstationarity correction* (see appendix F.1.1) to account for the nonstationary start. Note that this correction applies to both theorems 7.2.5 and 7.2.6 of section 7.2.

Prelude to a Proof We show theorem 3.1 in several steps. First, given $\mathcal{M}$, we define a *trace chain* $\mathcal{M}^{(T)}$, whose state space is all possible traces of length T (definition 7.3.3), and we show that for $T \geq \tau_{\mathrm{mix}}(\mathcal{M})$, mixing and relaxation times of $\mathcal{M}^{(T)}$ are at most constant (lemma F.1.5). Second, in definition 7.3.4 we introduce a novel MCMC *unbiased variance estimator*, which runs two copies of $\mathcal{M}$ in parallel, denoted $\mathcal{M} \otimes \mathcal{M}$, and averages the square difference of f in them, i.e., we note that $v_\pi = \frac{1}{2}\,\mathbb{E}_{\pi(\mathcal{M}\otimes\mathcal{M})}[(f \ominus f)^2]$. We prove the correctness of our variance estimator in lemma 7.3.5, and in line 11 of DynaMITE we employ it to estimate v_T. Third, we use variance-sensitive concentration bounds for mixing processes to bound the *true mean* μ in terms of the *empirical mean* $\hat{\mu}$ and *empirical variance* $\hat{v}$. Note that since we apply the variance estimation algorithm to the trace chain $\mathcal{M}^{(T)}$, $\hat{v}$ is an estimator for v_T. Since our algorithm depends on the *empirical variance*, we don't know *a priori* how many samples will be required to estimate the mean, so we run with a *doubling sampling schedule* until the desired guarantee is met.

We now define the T*-trace Markov chain* of $\mathcal{M}$, which we denote $\mathcal{M}^{(T)}$.[3] We use the trace chain knowing (for proof, see lemma F.1.5) that $\mathcal{M}^{(T)}$ has constant mixing and relaxation time for $T \geq \tau_{\mathrm{rel}}$.

Definition 7.3.3 (Trace chain)**.** For a Markov chain $\mathcal{M}$ on state space Ω, we define $\mathcal{M}^{(T)}$ on state space Ω^T as follows: given $\boldsymbol{a} = (a_1, a_2 \dots a_T)$ and $\boldsymbol{b} = (b_1, b_2, \dots b_T)$ in Ω^T, the probability of going from $\boldsymbol{a}$ to $\boldsymbol{b}$ is $\mathcal{M}^{(T)}(\boldsymbol{a}, \boldsymbol{b}) \doteq \mathcal{M}(a_T, b_1) \prod_{i=1}^{T-1} \mathcal{M}(b_i, b_{i+1})$.

The following definition provides us with an *unbiased MCMC estimator for variance*.

Definition 7.3.4 (Unbiased MCMC variance estimator)**.** Suppose function $f : \mathcal{X} \mapsto \mathbb{R}$, chain $\mathcal{M}$ over $\mathcal{X}$, and assume $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are two length m independent traces of $\mathcal{M}$. The following definition for $\hat{v}$ satisfies $\mathbb{E}[\hat{v}] = v_\pi$, thus $\hat{v}$ is an unbiased estimator for v_π.

[3] We use parentheses in the exponent to preserve standard matrix powering notation.

$$\hat{v} \doteq \frac{1}{2m}\sum_{i=1}^{m}(f(X_{1,i}) - f(X_{2,i}))^2 \enspace .$$

In other words, we can think $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ as the *two factors* of a single trace $\boldsymbol{X}$ of the *tensor product chain* $\mathcal{M} \otimes \mathcal{M}$, which steps forward by simultaneously running both factor chains for one step. Given a sample $\boldsymbol{X} \sim (\mathcal{M} \otimes \mathcal{M})^{(m)}$, this variance estimate $\hat{v}$ can then be represented as the *empirical mean* of half the *square difference function*, i.e., the function $\frac{1}{2}(f \ominus f)^2$.

Note that applying the above estimator to $\mathcal{M}$ yields an unbiased estimate of v_π, but applying it to $\mathcal{M}^{(T)}$ yields an unbiased estimator of v_T, thus it is sufficient to estimate both the stationary and trace variances. The following lemma, proved in the appendix, bounds the error of this estimator.

Lemma 7.3.5 (Trace Variance Estimation)**.** Let $\hat{v}$ be defined as in definition 7.3.4, it then holds that, for any $\delta \in (0, 1)$, we have

$$\mathbb{P}\left(v_\pi \geq \hat{v} + \frac{(11+\sqrt{21})(1+\frac{\lambda}{\sqrt{21}})r^2 \ln\frac{1}{\delta}}{(1-\lambda)m} + \sqrt{\frac{(1+\lambda)r^2\hat{v}\ln\frac{1}{\delta}}{(1-\lambda)m}}\right) \leq \delta \enspace .$$

Sampling Schedule and Proof Sketch We first note that correctness of DynaMITE and WarmStartDynaMITE follow from that of McmcPro, as they simply apply McmcPro to a trace chain and/or apply a *nonstationarity correction*.

The key to McmcPro is an *a priori* fixed *sampling schedule*, which determines the sizes of *progressively larger samples*. To ensure correctness, a sequence of tail bounds must all hold simultaneously w.h.p. by union bound. The main difficulty is that the schedule length N and the probability concentration bounds are codependent. A shorter schedule is more *statistically efficient*, but can *overshoot* the sufficient sample size, and the opposite holds for longer schedules. We explain how to resolve this cyclic dependence in the next paragraph.

Over a run of DynaMITE, we take (up to) $3N$ probability concentration bounds (N bounds for variance, line 12, and N bounds each for upper and lower mean bounds, line 13). We first establish the *worst-case* Hoeffding sample complexity m_N (theorem 7.2.5) and *best-case* Bernstein sample complexity α (theorem 7.2.6) by taking $v = 0$,

$$m_I \doteq m_H(\Lambda, r, \varepsilon, \tfrac{2\delta}{3N}) = \frac{(1+\Lambda)r^2\ln\frac{3N}{\delta}}{2(1-\Lambda)\varepsilon^2} \enspace , \enspace \& \enspace \alpha \doteq \frac{(1+\Lambda)r\ln\frac{3N}{\delta}}{(1-\Lambda)\varepsilon} \approx m_B(\Lambda, r, 0, \varepsilon, \tfrac{2\delta}{3N}) \enspace .$$

We know from these bounds that once a sample of size m_N is drawn, the desired guarantee has been met, and similarly, before a sample of size $\approx \alpha$ is drawn, the desired Bernstein bound *can not* be met. We now select the minimal N such that each sample size m_i obeys $m_i \leq 2m_{i-1}$. Observe that in the ratio $\frac{m_N}{\alpha}$, all dependence on N is divided out ($\ln \frac{3N}{\delta}$ terms cancel). We initially run the chain for m_1 to be $\lceil 2\alpha \rceil$, and in iteration i, we run up to $\lceil 2\alpha^i \rceil$ steps, thus we conclude $N \doteq \lfloor \log_2(\frac{r}{2\varepsilon}) \rfloor \geq \lceil \log_2(\frac{r}{2\varepsilon}) \rceil - 1$ (doubling) iterations are sufficient. These computations are repeated verbatim by McmcPro (lines 4 & 6) to compute the sampling schedule.

Conclusion and Discussion

Part I introduces new statistical techniques for variance-sensitive uniform convergence bounds and generalization guarantees in machine learning. The centralization strategy of chapter 1 achieves asymptotically-optimal bound convergence rates, and our Monte-Carlo estimation procedure yields sharp bounds with both centralized (theorem 1.3.7) and non-centralized (corollary 1.3.8) Rademacher averages. This has broad implications in machine learning, data science, and statistical settings, and can also be used as a component statistical method, mutatis mutandis, in all subsequent chapters. Chapter 2 then continues this theme of efficient use of data by introducing an algorithm that augments a small amount of labeled data with the output of *weak labelers*, which are assumed to be *machine learning models* associated with *correlated tasks*. I show both *computational guarantees* on efficient learnability and statistical guarantees on generalization bounds involving both the number of labeled and unlabeled samples.

Overall, the work of part I may be interpreted as novel methods to reduce the burden of acquiring large amounts of labeled data to train sophisticated machine learning models. Indeed, with the ever-expanding available computational power available (i.e., as characterized by Moore's law, and more recently with learning on GPUs, FPGAs, and specialized hardware) we have trained more complicated models (in particular, deeper neural networks), but datasets have had to grow commensurately to support training such models without overfitting. However, as computation becomes cheaper, the economic costs of collecting larger datasets do not always scale similarly, leaving learning from small datasets a problem that is both practically important and theoretically challenging. In fact, these methods are particularly impactful, as poor performance and analysis in the small sample setting can be a fairness issue, in the sense that by nature of their size and visibility, more data may be available on majority groups than minority or understudied groups, which contributes to marginalized groups being often poorly-served by machine learning systems.

Following the pure statistical methods of part I, in part II, I define *malfare* parallel to *welfare* (definition 3.3.1), and show that subject to several intuitive and basic axioms (section 3.3.1), all welfare and malfare functions are *power means* (definition 3.3.6, see theorem 3.3.8). I then argue that fair ML should seek to *minimize malfare*, as this naturally generalizes *loss minimization* to multiple groups. I then combine the uniform convergence guarantees of part I with the fairness setting of part II in chapter 4, where I introduce the *fair codec selection problem*. Finally, chapter 5 defines a concept of *fair-PAC-learnability* in terms of malfare minimization, and characterizes both necessary and sufficient conditions for various flavors of fair learnability. Following Valiant's introduction of classical PAC-learning in order to rigorously characterize machine learning problems in the lexicon of computer scientists, I define malfare minimization as the target of fair-PAC learning, with axiomatically justified fairness characteristics, the precise details of which also lead to interesting theoretical computer science and statistical estimation questions. In particular, I show a hierarchy of PAC-learning and fair-PAC-learning settings, and am excited to see additional connections drawn in future work.

Finally, in part III I leave the i.i.d. setting behind, with two dependent *mean-estimation* settings of great practical and theoretical interest. Chapter 6 describes a common real-world problem in databases, where a user wants to estimate the *mean of a function*, without waiting for the database to iterate over all records. Motivated by the real-world performance characteristics of modern database systems, I show that, while in the worst case, sampling entire blocks may not improve over sampling individual records, my algorithm is able to detect independence within blocks automatically, and adapts accordingly. Chapter 7 adapts this idea of *algorithmically adapting* to independence detected within the data, and applies it to sample from Markov chains. Such settings are both of great interest to the theoretical computer science community, as increasingly attention is placed on *randomized algorithms*, and also to practitioners in diverse fields, from *significance testing* in the sciences to *machine learning* and *artificial intelligence*, wherein such methods are vital to efficiently reason under partial knowledge in an uncertain world.

I have additionally published works in several related areas that have not made it into this thesis, namely *anomaly detection* [Cousins et al., 2017], *automated machine learning* [Binnig et al., 2015, 2018], *empirical game theory* and *mechanism design* [Viqueira et al., 2019, 2020, 2021], and *statistical data science* [Pellegrina et al., 2020]. Of special note are two somewhat more theoretical papers in pure machine learning. In Cousins and Upfal [2017], I show generalization bounds for a distance-based classifier, including what is to my knowledge the only known *exact (closed form) expression* for the *empirical Rademacher average* of a nontrivial classification model. In Cousins and Riondato [2019], I theoretically unify *classification trees* with the *entropy* impurity criterion and *regression trees* with the *square-error* impurity criterion, as well as an infinite family of parametric conditional-density estimation trees, under the *cross-entropy* information criterion, and show all to be instances of the *minimax-entropy principle*, with consequent efficient training heuristics.

The theme running through all of my work is the importance placed on rigorous guarantees for domain-appropriate properties in various probabilistic processes, most notably in *machine learning* (generalization bounds), *sampling* (mean estimation guarantees), and *data science* (frequent itemsets, equilibria in empirical game theory, etc.) settings. In particular, almost all of my papers rigorously show *finite-sample probabilistic guarantees* in some probabilistic setting, and overall my work illustrates that while such guarantees may induce significant proof complexity, often they asymptotically match or nearly match known lower bounds for sample complexity and approximate central-limit-theorem bounds. I argue that this additional analytical overhead is a price worth paying, as finite-sample guarantees are a necessary component in understanding the behavior and failure modes of machine learning and data science systems. As we see machine learning increasingly employed in our day-to-day lives, and witness the catastrophic consequences of failures of such systems, the importance of rigorous analysis of these methods to avoid negative outcomes becomes increasingly apparent.

Bibliography

D. Aldous, L. Lovasz, and P. Winkler. Mixing times for uniformly ergodic Markov chains. *Stochastic Processes and their Applications*, 71:165–185, 1997.

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

Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations*. Cambridge University Press, 2009.

Chidubem Arachie and Bert Huang. Adversarial label learning. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

Enrique Areyan Viqueira, Amy Greenwald, Cyrus Cousins, and Eli Upfal. Learning simulation-based games from data. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1778–1780. International Foundation for Autonomous Agents and Multiagent Systems, 2019.

Enrique Areyan Viqueira, Cyrus Cousins, and Amy Greenwald. Improved algorithms for learning equilibria in simulation-based games. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, pages 79–87, 2020.

Ahmed Ashraf, Shehroz Khan, Nikhil Bhagwat, Mallar Chakravarty, and Babak Taati. Learning to unlearn: Building immunity to dataset bias in medical imaging studies. *arXiv preprint arXiv:1812.01716*, 2018.

Anthony B Atkinson et al. On the measurement of inequality. *Journal of Economic Theory*, 2(3):244–263, 1970.

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Tuning bandit algorithms in stochastic environments. In *International conference on algorithmic learning theory*, pages 150–165. Springer, 2007.

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009.

Ariful Azad, Georgios Pavlopoulos, Christos Ouzounis, Nikos Kyrpides, and Aydin Buluç. HipMCL: a high-performance parallel implementation of the Markov clustering algorithm for large-scale networks. *Nucleic Acids Research*, 46: 1–11, 01 2018. doi: 10.1093/nar/gkx1313.

Stephen H. Bach, Bryan He, Alexander Ratner, and Christopher Ré. Learning the structure of generative models without labeled data. In *International Conference on Machine Learning (ICML)*, 2017.

Stephen H. Bach, Daniel Rodriguez, Yintao Liu, Chong Luo, Haidong Shao, Cassandra Xia, Souvik Sen, Alexander Ratner, Braden Hancock, Houman Alborzi, Rahul Kuchhal, Christopher Ré, and Rob Malkin. Snorkel DryBell: A case study in deploying weak supervision at industrial scale. 2019.

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

Anna Ben-Hamou, Roberto I. Oliveira, and Yuval Peres. Estimating graph parameters via random walks with restarts. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, page 1702–1714, USA, 2018. Society for Industrial and Applied Mathematics. ISBN 9781611975031.

George Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57(297):33–45, 1962.

Suman K. Bera and C. Seshadhri. How to count triangles, without seeing the whole graph. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, page 306–316, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379984. doi: 10.1145/3394486.3403073. URL `https://doi.org/10.1145/3394486.3403073`.

Sergei Bernstein. On a modification of Chebyshev's inequality and of the error formula of Laplace. *Annals of the Science Institute SAV. Ukraine, Sect. Math*, 1(4):38–49, 1924.

D.P. Bertsekas. *Convex Optimization Algorithms*. Athena Scientific, 2015.

Carsten Binnig, Fuat Basık, Benedetto Buratti, Ugur Cetintemel, Yeounoh Chung, Andrew Crotty, Cyrus Cousins, Dylan Ebert, Philipp Eichmann, Alex Galakatos, et al. Towards interactive data exploration. In *Real-Time Business Intelligence and Analytics*, pages 177–190. Springer, 2015.

Carsten Binnig, Benedetto Buratti, Yeounoh Chung, Cyrus Cousins, Tim Kraska, Zeyuan Shang, Eli Upfal, Robert Zeleznik, and Emanuel Zgraggen. Towards interactive curation & automatic tuning of ML pipelines. In *Proceedings of the Second Workshop on Data Management for End-To-End Machine Learning*, pages 1–4, 2018.

Dankmar Böhning. Multinomial logistic regression algorithm. *Annals of the institute of Statistical Mathematics*, 44(1): 197–200, 1992.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. A sharp concentration inequality with applications. *Random Structures & Algorithms*, 16(3):277–292, 2000.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. Concentration inequalities using the entropy method. *The Annals of Probability*, 31(3):1583–1614, 2003.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. On concentration of self-bounding functions. *Electronic Journal of Probability*, 14:1884–1899, 2009.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Olivier Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematique*, 334(6):495–500, 2002.

Joseph Bradley and Carlos Guestrin. Sample complexity of composite likelihood. In *Artificial Intelligence and Statistics*, pages 136–160, 2012.

Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

Andreas Brieden, Peter Gritzmann, Ravindran Kannan, Victor Klee, László Lovász, and Miklós Simonovits. Deterministic and randomized polynomial-time approximation of radii. *Mathematika*, 48(1-2):63–105, 2001.

G. Brightwell and P. Winkler. Counting linear extensions. *Order*, 8:225–242, 1991.

Peter S Bullen. *Handbook of means and their inequalities*, volume 560. Springer Science & Business Media, 2013.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

Jacqueline G Cavazos, P Jonathon Phillips, Carlos D Castillo, and Alice J O'Toole. Accuracy comparison across face recognition algorithms: Where are we on measuring race bias? *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2020.

Vincent S Chen, Paroma Varma, Ranjay Krishna, Michael Bernstein, Christopher Ré, and Li Fei-Fei. Scene graph prediction with limited labels. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Flavio Chierichetti and Shahrzad Haddadan. On the complexity of sampling vertices uniformly from a graph. In *45th International Colloquium on Automata, Languages, and Programming (ICALP 2018)*, volume 107 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 149:1–149:13. Schloß Dagstuhl–Leibniz-Zentrum für Informatik, 2018. ISBN 978-3-95977-076-7.

Kai-Min Chung, Henry Lam, Zhenming Liu, and Michael Mitzenmacher. Chernoff-Hoeffding bounds for Markov chains: Generalized and simplified. *arXiv preprint arXiv:1201.0559*, 2012.

Cynthia M Cook, John J Howard, Yevgeniy B Sirotin, Jerry L Tipton, and Arun R Vemury. Demographic effects in facial recognition and their dependence on image acquisition: An evaluation of eleven commercial systems. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 1(1):32–41, 2019.

Cyrus Cousins. An axiomatic theory of provably-fair welfare-centric machine learning. *arXiv preprint arXiv:2104.14504*, 2021.

Cyrus Cousins and Matteo Riondato. CaDET: interpretable parametric conditional density estimation with decision trees and forests. *Machine Learning*, 108(8-9):1613–1634, 2019.

Cyrus Cousins and Matteo Riondato. Sharp uniform convergence bounds through empirical centralization. In *Advances in Neural Information Processing Systems*, 2020.

Cyrus Cousins and Eli Upfal. The k-nearest representatives classifier: A distance-based classifier with strong generalization bounds. In *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10. IEEE, 2017.

Cyrus Cousins, Chirstopher M Pietras, and Donna K Slonim. Scalable FRaC variants: Anomaly detection for precision medicine. In *2017 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, pages 253–262. IEEE, 2017.

Cyrus Cousins, Shahrzad Haddadan, and Eli Upfal. Making mean-estimation more efficient using an MCMC trace variance approach: DynaMITE. *arXiv preprint arXiv:2011.11129*, 2020.

Hugh Dalton. The measurement of the inequality of incomes. *The Economic Journal*, 30(119):348–361, 1920.

Nilesh Dalvi, Anirban Dasgupta, Ravi Kumar, and Vibhor Rastogi. Aggregating crowdsourced binary ratings. In *Proceedings of the 22nd international conference on World Wide Web*, pages 285–294, 2013.

Vladimir Kolmogorov David G. Harris. Parameter estimation for Gibbs distributions. 2007. URL `https://arxiv.org/abs/2007.10824`.

A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society C*, 28(1):20–28, 1979.

Gerard Debreu. Topological methods in cardinal utility theory. Technical report, Cowles Foundation for Research in Economics, Yale University, 1959.

Pierre Del Moral and L. Miclo. Branching and interacting particle systems approximations of Feynman-Kac formulae with applications to non-linear filtering. *Sem. Probab. Stras.*, 34:1–145, 05 2007. doi: 10.1007/BFb0103798.

Luc Devroye, Matthieu Lerasle, Gábor Lugosi, and Roberto I. Oliveira. Sub-Gaussian mean estimators. *The Annals of Statistics*, 44(6):2695–2725, 2016.

Persi Diaconis. The Markov chain Monte Carlo revolution. *Bulletin of the American Mathematical Society*, 46:179–205, 04 2009. doi: 10.1090/S0273-0979-08-01238-X.

Khaled Elbassioni and Hans Raj Tiwary. On computing the vertex centroid of a polyhedron. *arXiv preprint arXiv:0806.3456*, 2008.

Yonina C Eldar, Amir Beck, and Marc Teboulle. A minimax Chebyshev estimator for bounded error estimation. *IEEE Transactions on Signal Processing*, 56(4):1388–1397, 2008.

Anton Enright, S Dongen, and C.A. Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic acids research*, 30:1575–84, 05 2002. doi: 10.1093/nar/30.7.1575.

J. Erfurt, C. R. Helmrich, S. Bosse, H. Schwarz, D. Marpe, and T. Wiegand. A study of the perceptually weighted peak signal-to-noise ratio (WPSNR) for image compression. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 2339–2343, Sep. 2019. doi: 10.1109/ICIP.2019.8803307.

Jianqing Fan, Yuan Liao, and Martina Mincheva. Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 75, 09 2013. doi: 10.1111/rssb.12016.

Jianqing Fan, Bai Jiang, and Qiang Sun. Hoeffding's lemma for Markov chains and its applications to statistical learning. *arXiv preprint arXiv:1802.00211*, 2018.

Yoav Freund. Boosting a weak learning algorithm by majority. *Information and Computation*, 121(2):256–285, 1995.

Bolin Gao and Lacra Pavel. On the properties of the softmax function with application in game theory and reinforcement learning, 2018.

Chao Gao and Dengyong Zhou. Minimax optimal convergence rates for estimating ground truth from crowdsourced labels. *CoRR*, abs/1207.0016, 2013.

Arpita Ghosh, Satyen Kale, and Preston McAfee. Who moderates the moderators? Crowdsourcing abuse detection in user-generated content. In *Proceedings of the 12th ACM conference on Electronic commerce*, pages 167–176, 2011.

Robert Gibbons. *Game theory for applied economists*. Princeton University Press, 1992.

Daniel T. Gillespie. Stochastic simulation of chemical kinetics. *Annual Review of Physical Chemistry*, 58(1):35–55, 2007. doi: 10.1146/annurev.physchem.58.032806.104637.

Evarist Giné and Vladimir Koltchinskii. Concentration inequalities and asymptotic results for ratio type empirical processes. *The Annals of Probability*, 34(3):1143–1216, 2006.

Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing*. Addison-Wesley Longman Publishing Co., Inc., USA, 2nd edition, 1992. ISBN 0201508036.

William M Gorman. The structure of utility functions. *The Review of Economic Studies*, 35(4):367–390, 1968.

V. Guruswami. Rapidly mixing Markov chains: A comparison of techniques (a survey). *ArXiv*, abs/1603.01512, 2016.

Uffe Haagerup. The best constants in the Khintchine inequality. *Studia Mathematica*, 70(3):231–283, 1982.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Robert Hegemann, Alexander Leidinger, and Rogério Brito. LAME MP3 encoder, 2017. URL `http://lame.sourceforge.net/`.

Hoda Heidari, Claudio Ferrari, Krishna Gummadi, and Andreas Krause. Fairness behind a veil of ignorance: A welfare analysis for automated decision making. In *Advances in Neural Information Processing Systems*, pages 1265–1276, 2018.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

Darrell Hoy, Denis Nekipelov, and Vasilis Syrgkanis. Welfare guarantees from data. In *Advances in Neural Information Processing Systems*, pages 3768–3777, 2017.

Lily Hu and Yiling Chen. Fair classification and social welfare. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 535–545, 2020.

Mark Huber. Approximation algorithms for the normalizing constant of Gibbs distributions. *The Annals of Applied Probability*, 25(2):974–985, 2015. ISSN 10505164. URL `http://www.jstor.org/stable/24519939`.

Peter J Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

Frank Hutter, Holger H Hoos, and Kevin Leyton-Brown. Sequential model-based optimization for general algorithm configuration. In *International Conference on Learning and Intelligent Optimization*, pages 507–523. Springer, 2011.

Bai Jiang, Qiang Sun, and Jianqing Fan. Bernstein's inequality for general Markov chains. *arXiv preprint arXiv:1805.10721*, 2018.

David R. Karger, Sewoong Oh, and Devavrat Shah. Budget-optimal task allocation for reliable crowdsourcing systems. *Operations Research*, 62(1):1–24, 2014.

Maximilian Kasy and Rediet Abebe. Fairness, equality, and power in algorithmic decision-making. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 576–586, 2021.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, volume 67, page 43. Schloß Dagstuhl–Leibniz-Zentrum für Informatik, 2017.

V. Kolmogorov. A faster approximation algorithm for the Gibbs partition function. *ArXiv*, abs/1608.04223, 2018.

Vladimir Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902–1914, 2001.

Vladimir Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006.

Tim Kraska, Alex Beutel, Ed H Chi, Jeffrey Dean, and Neoklis Polyzotis. The case for learned index structures. In *Proceedings of the 2018 International Conference on Management of Data*, pages 489–504. ACM, 2018.

Carlos Leon and François Perron. Optimal Hoeffding bounds for discrete reversible Markov chains. *The Annals of Applied Probability*, 14, 05 2004. doi: 10.1214/105051604000000170.

David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Society, 2017.

Pascal Lezaud. Chernoff-type bound for finite Markov chains. *The Annals of Applied Probability*, 8, 08 1998. doi: 10.1214/aoap/1028903453.

László Lovász and Peter Winkler. Mixing times. *AMS DIMACS Series*, 41:189–204, 1998.

Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019.

Brian Mac Namee, Padraig Cunningham, Stephen Byrne, and Owen I Corrigan. The problem of bias in training data in regression problems in medical decision support. *Artificial intelligence in medicine*, 24(1):51–70, 2002.

Pascal Massart. Some applications of concentration inequalities to statistics. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, volume 9, pages 245–303, 2000.

Andreas Maurer. Concentration inequalities for functions of independent variables. *Random Structures & Algorithms*, 29(2):121–138, 2006.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

A. Mazzetto, D. Sam, A. Park, E. Upfal, and S. H. Bach. Semi-supervised aggregation of dependent weak supervision sources with performance guarantees. In *Artificial Intelligence and Statistics (AISTATS)*, 2021.

Colin McDiarmid. On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188, 1989.

Colin McDiarmid and Bruce Reed. Concentration for self-bounding functions and an inequality of Talagrand. *Random Structures & Algorithms*, 29(4):549–557, 2006.

Hugh McGuire. LibriVox: Free public domain audiobooks. URL `https://librivox.org/`.

Blazej Miasojedow. Hoeffding's inequalities for geometrically ergodic Markov chains on general state space. *Statistics & Probability Letters*, 87, 01 2012. doi: 10.1016/j.spl.2014.01.013.

Stefania Mignani and Rosa Rodolfo. Markov chain Monte Carlo in statistical mechanics: The problem of accuracy. *Technometrics*, 43:347–55, 2001.

Michael Mitzenmacher and Eli Upfal. *Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis*. Cambridge university press, second edition, 2017.

Volodymyr Mnih, Csaba Szepesvári, and Jean-Yves Audibert. Empirical Bernstein stopping. In *Proceedings of the 25th international conference on Machine learning*, pages 672–679. ACM, 2008.

Hervé Moulin. *Fair division and collective welfare*. MIT Press, 2004.

Cristóbal A. Navarro, Wei Huang, and Youjin Deng. Adaptive multi-GPU exchange Monte Carlo for the 3D random field Ising model. *Computer Physics Communications*, 205:48–60, 2016.

John Ashworth Nelder and Robert WM Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384, 1972.

Luca Oneto, Alessandro Ghio, Davide Anguita, and Sandro Ridella. An improved analysis of the Rademacher data-dependent bound using its self bounding property. *Neural Networks*, 44:107–111, 2013.

Daniel Paulin. Concentration inequalities for Markov chains by Marton couplings and spectral methods. *Electronic Journal of Probabability*, 20, 2015.

Leonardo Pellegrina, Matteo Riondato, and Fabio Vandin. SPuManTE: Significant pattern mining with unconditional testing. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1528–1538. ACM, 2019.

Leonardo Pellegrina, Cyrus Cousins, Fabio Vandin, and Matteo Riondato. MCRapper: Monte-Carlo Rademacher averages for POSET families and approximate pattern mining. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2165–2174, 2020.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1406–1415, 2019.

P. H. Peskun. Optimum Monte-Carlo sampling using Markov chain. *Biometrika*, 60(3):607–612, 1973. URL `JSTOR,www.jstor.org/stable/2335011`.

Andrea Pietracaprina, Matteo Riondato, Eli Upfal, and Fabio Vandin. Mining top-k frequent itemsets through progressive sampling. *Data Mining and Knowledge Discovery*, 21(2):310–326, 2010.

Arthur Cecil Pigou. *Wealth and welfare*. Macmillan and Company, limited, 1912.

David Pollard. *Convergence of Stochastic Processes*. Springer, New York, 1984.

Dana Randall. Rapidly mixing Markov chains with applications in computer science and physics. 8(2):30–41, March 2006. ISSN 1521-9615. doi: 10.1109/MCSE.2006.30. URL `https://doi.org/10.1109/MCSE.2006.30`.

Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: Rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment*, 11(3):269–282, 2017.

Alexander J Ratner, Christopher M De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. Data programming: Creating large training sets, quickly. In *Neural Information Processing Systems (NeurIPS)*, 2016.

Alexander J Ratner, Braden Hancock, Jared Dunnmon, Frederic Sala, Shreyash Pandey, and Christopher Ré. Training complex models with multi-task weak supervision. In *AAAI*, 2019.

Matteo Riondato and Eli Upfal. Mining frequent itemsets through progressive sampling with Rademacher averages. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1005–1014. ACM, 2015.

Matteo Riondato and Eli Upfal. ABRA: Approximating betweenness centrality in static and dynamic graphs with Rademacher averages. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1145–1154. ACM, 2016.

Matteo Riondato and Eli Upfal. ABRA: Approximating betweenness centrality in static and dynamic graphs with Rademacher averages. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 12(5):61, 2018.

Kevin WS Roberts. Interpersonal comparability and social choice theory. *The Review of Economic Studies*, pages 421–439, 1980.

Esther Rolf, Max Simchowitz, Sarah Dean, Lydia T Liu, Daniel Björkegren, Moritz Hardt, and Joshua Blumenstock. Balancing competing objectives with noisy data: Score-based classifiers for welfare-aware machine learning. *arXiv preprint arXiv:2003.06740*, 2020.

Esteban Safranchik, Shiying Luo, and Stephen H. Bach. Weakly supervised sequence tagging from noisy rules. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2020.

M. Salatino, J. Austermann, J. A. Beall, S. Choi, K. T. Crowley, S. Duff, S. W. Henderson, G. Hilton, S. Ho, J. Hubmayr, Y. Li, M. D. Niemack, S. M. Simon, S. T. Staggs, and E. J. Wollack. Machine learning, Markov chain Monte Carlo, and optimal algorithms to characterize the AdvACT kilopixel transition-edge sensor arrays. *IEEE Transactions on Applied Superconductivity*, 29(5):1–5, 2019.

Paul-Marie Samson. Infimum-convolution description of concentration properties of product probability measures, with applications. In *Annales de l'IHP Probabilités et statistiques*, volume 43, pages 321–338, 2007.

Sandvine. The global internet phenomena report October 2018. `https://www.sandvine.com/hubfs/downloads/`

phenomena/2018-phenomena-report.pdf, October 2018.

Clayton Sanford, Cyrus Cousins, and Eli Upfal. Uniform convergence bounds for codec selection. *arXiv preprint arXiv:1812.07568*, 2018.

Robert E Schapire. The strength of weak learnability. *Machine Learning*, 5(2):197–227, 1990.

Amartya Sen. On weights and measures: Informational constraints in social welfare analysis. *Econometrica: Journal of the Econometric Society*, pages 1539–1572, 1977.

Amartya Sen, Master Amartya Sen, Sen Amartya, James E Foster, James E Foster, et al. *On economic inequality*. Oxford university press, 1997.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

Naum Zuselevich Shor. *Minimization methods for non-differentiable functions*, volume 3. Springer Science & Business Media, 2012.

Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. A unified approach to quantifying algorithmic unfairness: Measuring individual & group unfairness via inequality indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2239–2248, 2018.

D. Stefankovic, S. Vempala, and E. Vigoda. Adaptive simulated annealing: A near-optimal connection between sampling and counting. In *48th Annual IEEE Symposium on Foundations of Computer Science*, pages 183–193, 2007. doi: 10.1109/FOCS.2007.67.

Henri Theil. Economics and information theory. Technical report, Econometric Institute, Netherlands School of Economics, 1967.

Thilo Thiede, William C. Treurniet, Roland Bitto, Christian Schmidmer, Thomas Sporer, John G. Beerends, Catherine Colomes, Michael Keyhl, Gerhard Stoll, Karlheinz Brandenburg, and Bernhard Feiten. PEAQ—the ITU standard for objective measurement of perceived audio quality. *Journal of the Audio Engineering Society*, 48(1):3–29, 2000.

Philip S Thomas, Bruno Castro da Silva, Andrew G Barto, Stephen Giguere, Yuriy Brun, and Emma Brunskill. Preventing undesirable behavior of intelligent machines. *Science*, 366(6468):999–1004, 2019.

Z. Tu and Song Zhu. Image segmentation by data-driven Markov chain Monte Carlo. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24:657–673, 06 2002. doi: 10.1109/34.1000239.

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

AW Van der Vaart and JA Wellner. *Weak Convergence and Empirical Processes*. Springer, New York, 1996.

Fabio Vandin, Eli Upfal, and Ben Raphael. Algorithms for detecting significantly mutated pathways in cancer. *Journal of computational biology : a journal of computational molecular cell biology*, 18:507–22, 03 2011. doi: 10.1089/cmb.2010.0265.

Fabio Vandin, Patrick Clay, Eli Upfal, and Benjamin J Raphael. Discovery of mutated subnetworks associated with clinical data in cancer. In *Biocomputing 2012*, pages 55–66. World Scientific, 2012a.

Fabio Vandin, Eli Upfal, and Benjamin J Raphael. De novo discovery of mutated driver pathways in cancer. *Genome research*, 22(2):375–385, 2012b.

Fabio Vandin, Benjamin J Raphael, and Eli Upfal. On the sample complexity of cancer pathways identification. *Journal of Computational Biology*, 23(1):30–41, 2016.

Vladimir Vapnik. Principles of risk minimization for learning theory. In *Advances in neural information processing systems*, pages 831–838, 1992.

Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

Paroma Varma, Bryan He, Payal Bajaj, Imon Banerjee, Nishith Khandwala, Daniel L Rubin, and Christopher Ré. Inferring generative model structure with static analysis. *Advances in neural information processing systems*, 30: 239, 2017.

Paroma Varma, Frederic Sala, Ann He, Alexander Ratner, and Christopher Ré. Learning dependency structures for weak supervision models. In *International Conference on Machine Learning (ICML)*, 2019.

Enrique Areyan Viqueira, Cyrus Cousins, and Amy Greenwald. Learning simulation-based games from data. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, 2019.

Enrique Areyan Viqueira, Cyrus Cousins, Yasser Mohammad, and Amy Greenwald. Empirical mechanism design: Designing mechanisms from data. In *Uncertainty in Artificial Intelligence*, pages 1094–1104. PMLR, 2020.

Enrique Areyan Viqueira, Cyrus Cousins, and Amy Greenwald. Learning competitive equilibria in noisy combinatorial markets. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, 2021.

Weiran Wang and Miguel A Carreira-Perpinán. Projection onto the probability simplex: An efficient algorithm with a simple proof, and an application. *arXiv preprint arXiv:1309.1541*, 2013.

David Bruce Wilson. Mixing times of lozenge tiling and card shuffling Markov chains. *The Annals of Applied Probability*, (1):274–325, 2004.

Sen Wu, Luke Hsiao, Xiao Cheng, Braden Hancock, Theodoros Rekatsinas, Philip Levis, and Christopher Ré. Fonduer: Knowledge base construction from richly formatted data. In *International Conference on Management of Data*, 2018.

Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2018.

Ming Yuan and Christina Kendziorski. Hidden Markov models for microarray time course data in multiple biological conditions. *Journal of the American Statistical Association*, 101:1323–1332, 02 2006. doi: 10.1198/016214505000000394.

Cha Zhang and Yunqian Ma. *Ensemble Machine Learning: Methods and Applications*. Springer, 2012.

Yuchen Zhang, Xi Chen, Dengyong Zhou, and Michael I Jordan. Spectral methods meet EM: A provably optimal algorithm for crowdsourcing. *The Journal of Machine Learning Research*, 17(1):3537–3580, 2016.

Appendices and Supplementary Material

Appendix A

Supplementary Material for Chapter 1

A.1 Proofs

Lemma 1.2.1. Suppose $m \geq 4$. Then

$$\frac{\mathbb{E}_{\boldsymbol{x}}\left[\hat{\mathfrak{R}}_m(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}), \boldsymbol{x})\right]}{1+2\mathrm{b}(m)} \leq \mathfrak{R}_m(\mathrm{C}_{\mathcal{D}}(\mathcal{F}), \mathcal{D}) \leq \frac{\mathbb{E}_{\boldsymbol{x}}\left[\hat{\mathfrak{R}}_m(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}), \boldsymbol{x})\right]}{1-2\mathrm{b}(m)} \ .$$

Proof. We first show the rightmost inequality. Starting from the definition of the RA of the distributional centralization, and then subtracting and adding $\hat{\mathbb{E}}_{\boldsymbol{x}}[f]$, it holds

$$\mathfrak{R}_m(\mathrm{C}_{\mathcal{D}}(\mathcal{F}), \mathcal{D}) = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}, \boldsymbol{x}}\left[\sup_{f \in \mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\sigma}_i\Big(\big(f(\boldsymbol{x}_i) - \hat{\mathop{\mathbb{E}}_{\boldsymbol{x}}}[f]\big) + \big(\hat{\mathop{\mathbb{E}}_{\boldsymbol{x}}}[f] - \mathop{\mathbb{E}}_{\mathcal{D}}[f]\big)\Big)\right|\right] \ .$$

The subadditivity of the supremum and of the absolute value, and the linearity of the expectation allow us to split the r.h.s. into two summands and obtain

$$\mathfrak{R}_m(\mathrm{C}_{\mathcal{D}}(\mathcal{F}), \mathcal{D}) \leq \mathop{\mathbb{E}}_{\boldsymbol{\sigma}, \boldsymbol{x}}\left[\sup_{f \in \mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\sigma}_i\big(f(\boldsymbol{x}_i) - \hat{\mathop{\mathbb{E}}_{\boldsymbol{x}}}[f]\big)\right|\right] + \mathop{\mathbb{E}}_{\boldsymbol{\sigma}, \boldsymbol{x}}\left[\sup_{f \in \mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\sigma}_i\big(\hat{\mathop{\mathbb{E}}_{\boldsymbol{x}}}[f] - \mathop{\mathbb{E}}_{\mathcal{D}}[f]\big)\right|\right] \ .$$

Both terms on the r.h.s. can be seen as expectations w.r.t. $\boldsymbol{x}$ of the ERAs on $\boldsymbol{x}$ of two sample-dependent families: the empirical centralization of $\mathcal{F}$, and the family

$$\mathcal{K}_{\boldsymbol{x}} \doteq \{y \mapsto \hat{\mathop{\mathbb{E}}_{\boldsymbol{x}}}[f] - \mathop{\mathbb{E}}_{\mathcal{D}}[f], f \in \mathcal{F}\} \ .$$

Each function in $\mathcal{K}_{\boldsymbol{x}}$ is *constant*. Thus, we can write

$$\mathfrak{R}_m(\mathrm{C}_{\mathcal{D}}(\mathcal{F}),\mathcal{D}) \leq \mathop{\mathbb{E}}_{\boldsymbol{x}}\Big[\hat{\mathfrak{R}}_m(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}),\boldsymbol{x})\Big] + \mathop{\mathbb{E}}_{\boldsymbol{x}}\Big[\hat{\mathfrak{R}}_m(\mathcal{K}_{\boldsymbol{x}},\boldsymbol{x})\Big] \ . \tag{49}$$

Using (7) and the linearity of expectation we have that, for each $\boldsymbol{x} \in \mathcal{X}^m$, it holds

$$\hat{\mathfrak{R}}_m(\mathcal{K}_{\boldsymbol{x}},\boldsymbol{x}) = \sup_{f\in\mathcal{F}} |\hat{\mathop{\mathbb{E}}_{\boldsymbol{x}}}[f] - \mathop{\mathbb{E}}_{\mathcal{D}}[f]|\mathrm{b}(m) = \mathsf{SD}(\mathcal{F},\boldsymbol{x})\mathrm{b}(m) = \mathsf{SD}(\mathrm{C}_{\mathcal{D}}(\mathcal{F}),\boldsymbol{x})\mathrm{b}(m), \tag{50}$$

where in the last step we use the fact that the SD is invariant to shifting of functions. Continuing from (49) and using (50) and the rightmost inequality of (6), we obtain

$$\mathfrak{R}_m(\mathrm{C}_{\mathcal{D}}(\mathcal{F}),\mathcal{D}) \leq \mathop{\mathbb{E}}_{\boldsymbol{x}}\Big[\hat{\mathfrak{R}}_m(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}),\boldsymbol{x})\Big] + 2\mathfrak{R}_m(\mathrm{C}_{\mathcal{D}}(\mathcal{F}),\mathcal{D})\mathrm{b}(m) \ .$$

The hypothesis $m \geq 4$ implies $1 - 2\mathrm{b}(m) > 0$ (see (7)), so we can rewrite the above as

$$\mathfrak{R}_m(\mathrm{C}_{\mathcal{D}}(\mathcal{F}),\mathcal{D}) \leq \frac{1}{1-2\mathrm{b}(m)} \mathop{\mathbb{E}}_{\boldsymbol{x}}\Big[\hat{\mathfrak{R}}_m(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}),\boldsymbol{x})\Big],$$

which completes the proof of the upper bound.

We next show the lower bound. Starting from the definition of $\hat{\mathfrak{R}}_m(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}),\boldsymbol{x})$ and subtracting and adding $\mathbb{E}_{\mathcal{D}}[f]$, it holds

$$\mathop{\mathbb{E}}_{\boldsymbol{x}}\big[\hat{\mathfrak{R}}_m(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}),\boldsymbol{x})\big] = \mathop{\mathbb{E}}_{\boldsymbol{\sigma},\boldsymbol{x}}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{\sigma}_i\Big(\big(f(\boldsymbol{x}_i) - \mathop{\mathbb{E}}_{\mathcal{D}}[f]\big) + \big(\mathop{\mathbb{E}}_{\mathcal{D}}[f] - \hat{\mathop{\mathbb{E}}_{\boldsymbol{x}}}[f]\big)\Big)\right|\right] \ .$$

The subadditivity of the supremum and of the absolute value, and the linearity of the expectation allow us to split the r.h.s. into two summands and obtain

$$\begin{aligned}\mathop{\mathbb{E}}_{\boldsymbol{x}}\big[\hat{\mathfrak{R}}_m(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}),\boldsymbol{x})\big] \leq &\mathop{\mathbb{E}}_{\boldsymbol{\sigma},\boldsymbol{x}}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{\sigma}_i\big(f(\boldsymbol{x}_i) - \mathop{\mathbb{E}}_{\mathcal{D}}[f]\big)\right|\right] \\ &+ \mathop{\mathbb{E}}_{\boldsymbol{\sigma},\boldsymbol{x}}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{\sigma}_i\big(\mathop{\mathbb{E}}_{\mathcal{D}}[f] - \hat{\mathop{\mathbb{E}}_{\boldsymbol{x}}}[f]\big)\right|\right] \ .\end{aligned} \tag{51}$$

The first term on the r.h.s. is the RA of the *distributional* centralization of $\mathcal{F}$, i.e., it is $\mathfrak{R}_m(\mathrm{C}_{\mathcal{D}}(\mathcal{F}),\mathcal{D})$. The second term is the expectation w.r.t. $\boldsymbol{x}$ of the ERA on $\boldsymbol{x}$ of the family

$$\mathcal{Z}_{\boldsymbol{x}} \doteq \{x \mapsto \mathop{\mathbb{E}}_{\mathcal{D}}[f] - \hat{\mathop{\mathbb{E}}_{\boldsymbol{x}}}[f], f \in \mathcal{F}\} \ .$$

Each function in $\mathcal{Z}_{\boldsymbol{x}}$ is *constant*. Proceeding in exactly the same way as we did for the family $\mathcal{K}_{\boldsymbol{x}}$ in the proof of the upper bound, we can write

$$\hat{\mathfrak{R}}_m(\mathcal{Z}_{\boldsymbol{x}}, \boldsymbol{x}) = \mathsf{SD}(\mathrm{C}_{\mathcal{D}}(\mathcal{F}), \boldsymbol{x})\mathrm{b}(m) \ . \tag{52}$$

Continuing from (51) and using (52) and the rightmost inequality of (6), we obtain

$$\mathop{\mathbb{E}}_{\boldsymbol{x}}\big[\hat{\mathfrak{R}}_m(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}), \boldsymbol{x})\big] \leq \mathfrak{R}_m(\mathrm{C}_{\mathcal{D}}(\mathcal{F}), \mathcal{D}) + 2\mathfrak{R}_m(\mathrm{C}_{\mathcal{D}}(\mathcal{F}), \mathcal{D})\mathrm{b}(m) \leq (1 + 2\mathrm{b}(m))\mathfrak{R}_m(\mathrm{C}_{\mathcal{D}}(\mathcal{F}), \mathcal{D}),$$

and our proof is complete. ☐

Definition A.1.1. A function $\mathrm{Z} \in \mathcal{X}^m \to \mathbb{R}$ is *(α,β)-self-bounding with scale* γ, for some $\alpha > 0$, $\beta \geq 0$, $\gamma \geq 0$ if for each $j = 1, \ldots, m$, there exists a function $\mathrm{Z}_j \in \mathcal{X}^m \to \mathbb{R}$ such that, for any $\boldsymbol{x} \in \mathcal{X}^m$ it holds that

1. $\mathrm{Z}_j(\boldsymbol{x})$ does not depend on the j-th component $\boldsymbol{x}_j$ of $\boldsymbol{x}$; and
2. it holds $\mathrm{Z}_j(\boldsymbol{x}) \leq \mathrm{Z}(\boldsymbol{x}) \leq \mathrm{Z}_j(\boldsymbol{x}) + \gamma$;

Additionally, the functions Z_j, $j = 1, \ldots, m$, must be such that, for any $\boldsymbol{x} \in \mathcal{X}^m$, it holds $\sum_{j=1}^{m}\Big(\mathrm{Z}(\boldsymbol{x}) - \mathrm{Z}_j(\boldsymbol{x})\Big) \leq \alpha\mathrm{Z}(\boldsymbol{x}) + \beta$.

Theorem A.1.2. Let Z be a function from $\mathcal{X}^m$ to $\mathbb{R}$ that is (α, β)-self-bounding with scale γ, for $\alpha \geq 1/3$. Let $\delta \in (0, 1)$ and let $\boldsymbol{x}$ be a collection of m i.i.d. samples from $\mathcal{X}$. With probability at least $1 - \delta$ over the choice of $\boldsymbol{x}$, it holds

$$\mathop{\mathbb{E}}_{\boldsymbol{x}}\left[\mathrm{Z}(\boldsymbol{x})\right] \leq \mathrm{Z}(\boldsymbol{x}) + \alpha\gamma \ln\frac{1}{\delta} + \sqrt{\left(\alpha\gamma \ln\frac{1}{\delta}\right)^2 + 2\gamma(\alpha\mathrm{Z}(\boldsymbol{x}) + \beta)\ln\frac{1}{\delta}} \ . \tag{53}$$

Additionally, when $\alpha = 1$, we may improve the constants to

$$\mathop{\mathbb{E}}_{\boldsymbol{x}}\left[\mathrm{Z}(\boldsymbol{x})\right] \leq \mathrm{Z}(\boldsymbol{x}) + \frac{2}{3}\gamma \ln\frac{1}{\delta} + \sqrt{\left(\frac{1}{\sqrt{3}}\gamma \ln\frac{1}{\delta}\right)^2 + 2\gamma(\mathrm{Z}(\boldsymbol{x}) + \beta)\ln\frac{1}{\delta}} \ . \tag{54}$$

Proof. In both cases, we will assume WLOG $\gamma = 1$. The results then hold by linearity, noting that if $\mathrm{Z}(\cdot)$ is α-β self-bounding, with scale γ, then $\frac{1}{\gamma}\mathrm{Z}(\cdot)$ is α-β/γ self-bounding, with scale 1; the general case thus follows by dividing out γ, obtaining a bound, and then multiplying through by γ.

We first show equation 53. Assume scale $\gamma = 1$. It is known that for $\gamma = 1$, we have for all $\alpha \geq \frac{1}{3}$, as described in [Boucheron et al., 2009, Thm. 1], which improves the earlier bounds of [Maurer, 2006]

$$\mathbb{P}\left(\mathrm{Z}(\boldsymbol{x}) \leq \mathop{\mathbb{E}}_{\boldsymbol{x}}[\mathrm{Z}(\boldsymbol{x})] - \varepsilon\right) \leq \exp\left(\frac{-\varepsilon^2}{2(\alpha\,\mathbb{E}_{\boldsymbol{x}}[\mathrm{Z}(\boldsymbol{x})] + \beta)}\right) \ . \tag{55}$$

Now, taking δ equal to the RHS of (55), and solving for ε, this implies that with probability at least $1 - \delta$, we have

$$\mathrm{Z}(\boldsymbol{x}) + \frac{\beta}{\alpha} \geq \mathop{\mathbb{E}}_{\boldsymbol{x}}\left[\mathrm{Z}(\boldsymbol{x})\right] + \frac{\beta}{\alpha} - \sqrt{2(\alpha \mathop{\mathbb{E}}_{\boldsymbol{x}}\left[\mathrm{Z}(\boldsymbol{x})\right] + \beta)\ln\frac{1}{\delta}} \ .$$

Note that this is a quadratic inequality in $\sqrt{\mathbb{E}_{\boldsymbol{x}}\left[\mathrm{Z}(\boldsymbol{x})\right] + \frac{\beta}{\alpha}}$, solving for which (via the quadratic formula) yields nondegenerate solution

$$\mathop{\mathbb{E}}_{\boldsymbol{x}}\left[\mathrm{Z}(\boldsymbol{x})\right] \leq \mathrm{Z}(\boldsymbol{x}) + \alpha\ln\frac{1}{\delta} + \sqrt{\left(\alpha\ln\frac{1}{\delta}\right)^2 + 2\alpha(\mathop{\mathbb{E}}_{\boldsymbol{x}}\left[\mathrm{Z}(\boldsymbol{x})\right] + \beta)\ln\frac{1}{\delta}} \ .$$

Finally, in the general case, with γ-scaling, we have

$$\mathop{\mathbb{E}}_{\boldsymbol{x}}\left[\mathrm{Z}(\boldsymbol{x})\right] \leq \mathrm{Z}(\boldsymbol{x}) + \gamma\alpha\ln\frac{1}{\delta} + \sqrt{\left(\gamma\alpha\ln\frac{1}{\delta}\right)^2 + 2\gamma\alpha(\mathop{\mathbb{E}}_{\boldsymbol{x}}\left[\mathrm{Z}(\boldsymbol{x})\right] + \beta)\ln\frac{1}{\delta}} \ .$$

We now show equation 54 (i.e., assume $\alpha = 1$). Again assume $\gamma = 1$. This result follows via identical logic to the above, this time using the *sub-gamma* form (see Boucheron et al. [2013, Ch. 2.1], section 2.1) of the stronger *sub-Poisson* 1-β self-bounding function inequality [Boucheron et al., 2000, Thm. 1].

In particular, here we have that with probability at least $1 - \delta$,

$$\mathrm{Z}(\boldsymbol{x}) \geq \mathop{\mathbb{E}}_{\boldsymbol{x}}\left[\mathrm{Z}(\boldsymbol{x})\right] + \frac{1}{3}\ln\frac{1}{\delta} - \sqrt{2(\mathop{\mathbb{E}}_{\boldsymbol{x}}\left[\mathrm{Z}(\boldsymbol{x})\right] + \beta)\ln\frac{1}{\delta}} \ ,$$

which by the quadratic formula, yields

$$\mathop{\mathbb{E}}_{\boldsymbol{x}}\left[\mathrm{Z}(\boldsymbol{x})\right] \leq \mathrm{Z}(\boldsymbol{x}) + \frac{2}{3}\ln\frac{1}{\delta} + \sqrt{\left(\frac{\gamma}{\sqrt{3}}\ln\frac{1}{\delta}\right)^2 + 2(\mathop{\mathbb{E}}_{\boldsymbol{x}}\left[\mathrm{Z}(\boldsymbol{x})\right] + \beta)\ln\frac{1}{\delta}} \ .$$

The general result then follows via γ-scaling.

$\square$

Theorem 1.2.2. Suppose $m \geq 1$, and let $\chi \doteq 1 + 2\mathrm{b}(m)$. For any $\delta \in (0,1)$, with probability at least $1-\delta$ over the choice of $\boldsymbol{x}$, it holds that

$$\mathop{\mathbb{E}}_{\boldsymbol{x}}[\hat{\mathfrak{R}}_m(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}),\boldsymbol{x})] \leq \hat{\mathfrak{R}}_m(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}),\boldsymbol{x}) + \frac{2r\chi\ln\frac{1}{\delta}}{3m} + \sqrt{\left(\frac{r\chi\ln\frac{1}{\delta}}{\sqrt{3}m}\right)^2 + \frac{2r\chi(\hat{\mathfrak{R}}_m(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}),\boldsymbol{x}) + r\mathrm{b}(m))\ln\frac{1}{\delta}}{m}} \ . \tag{9}$$

Proof. This proof proceeds by showing that $\hat{\mathfrak{R}}_m(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}),\boldsymbol{x})$ is a $(1, r\mathrm{b}(m))$-self-bounding function with scale $r\chi/m$, then applying (54) from theorem A.1.2. First note that the result trivially holds for $m=1$, as the empirically centralized ERA will always be 0, thus we assume $m \geq 2$ henceforth.

For any $\boldsymbol{x} \in \mathcal{X}^m$, let

$$\mathrm{Y}(\boldsymbol{x}) \doteq \hat{\mathfrak{R}}_m(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}),\boldsymbol{x}),$$

and let $\boldsymbol{x}_{\setminus j}$ (resp. $\boldsymbol{\sigma}_{\setminus j}$) denote the $m-1$-dimensional vector of all but the j-th element of $\boldsymbol{x}$ (resp. $\boldsymbol{\sigma}$). Define

$$\mathrm{Y}_j(\boldsymbol{x}) \doteq \frac{m-1}{m}\hat{\mathfrak{R}}_{m-1}\left(\hat{\mathrm{C}}_{\boldsymbol{x}_{\setminus j}}(\mathcal{F}),\boldsymbol{x}_{\setminus j}\right) = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\left|\sum_{i=1,i\neq j}^{m}\boldsymbol{\sigma}_i\left(f(\boldsymbol{x}_i) - \mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}_{\setminus j}}[f]\right)\right|\right] \ .$$

We define these functions for convenience of notation. They will be handy when we later introduce the functions Z and Z_j, $j = 1,\ldots,m$ that we want to show to be self-bounding.

We now show that $\mathrm{Y}_j(\boldsymbol{x}) \leq \mathrm{Y}(\boldsymbol{x}) + r/m\,\mathrm{b}(m)$. Starting from the definition of $\mathrm{Y}_j(\boldsymbol{x})$ and adding and subtracting $(f(x_j) - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f])/2m$ to the argument of the supremum, it holds

$$\mathrm{Y}_j(\boldsymbol{x}) = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}_{\setminus j}}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\left|\left(\sum_{\substack{i=1\\ i\neq j}}^{m}\boldsymbol{\sigma}_i\left(f(\boldsymbol{x}_i) - \mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}_{\setminus j}}[f]\right)\right) + \frac{1}{2}\big(f(\boldsymbol{x}_j) - \mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}_{\setminus j}}[f]\big) - \frac{1}{2}\big(f(\boldsymbol{x}_j) - \mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}_{\setminus j}}[f]\big)\right|\right] \ .$$

Doubling and halving the sum in the argument of the expectation, and leveraging the subadditivity of the supremum and of the absolute value, we obtain

$$\mathrm{Y}_j(\boldsymbol{x}) \leq \mathop{\mathbb{E}}_{\boldsymbol{\sigma}_{\setminus j}}\left[\begin{array}{rl} & \frac{1}{2}\left(\sup_{f\in\mathcal{F}}\frac{1}{m}\left|\sum_{\substack{i=1\\ i\neq j}}^{m}\boldsymbol{\sigma}_i\Big(f(\boldsymbol{x}_i) - \mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}_{\setminus j}}[f]\Big) + \Big(f(\boldsymbol{x}_j) - \mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}_{\setminus j}}[f]\Big)\right|\right) \\ + & \frac{1}{2}\left(\sup_{f\in\mathcal{F}}\frac{1}{m}\left|\sum_{\substack{i=1\\ i\neq j}}^{m}\boldsymbol{\sigma}_i\Big(f(\boldsymbol{x}_i) - \mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}_{\setminus j}}[f]\Big) - \Big(f(\boldsymbol{x}_j) - \mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}_{\setminus j}}[f]\Big)\right|\right)\end{array}\right] \ .$$

The two-term sum forming the argument of the outermost expectation is the expectation *w.r.t. only* $\boldsymbol{\sigma}_j$ (i.e., *conditioned* on $\boldsymbol{\sigma}_{\setminus j}$) of the quantity

$$\sup_{f\in\mathcal{F}} \frac{1}{m}\left|\sum_{i=1}^{m} \boldsymbol{\sigma}_i \left(f(\boldsymbol{x}_i) - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f]\right)\right| \enspace .$$

Thus, using the law of total expectation, we can write

$$\mathrm{Y}_j(\boldsymbol{x}) \leq \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\left|\sum_{i=1}^{m} \boldsymbol{\sigma}_i \left(f(\boldsymbol{x}_i) - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f]\right)\right|\right] \enspace .$$

By subtracting and adding $\hat{\mathbb{E}}_{\boldsymbol{x}}[f]$ to each term of the sum, and using the subadditivity of the supremum and of the absolute value, and the linearity of the expectation, we obtain

$$\mathrm{Y}_j(\boldsymbol{x}) \leq \underbrace{\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\left|\sum_{i=1}^{m} \boldsymbol{\sigma}_i \left(f(\boldsymbol{x}_i) - \hat{\mathbb{E}}_{\boldsymbol{x}}[f]\right)\right|\right]}_{=\mathrm{Y}(\boldsymbol{x})} + \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\left|\sum_{i=1}^{m} \boldsymbol{\sigma}_i \left(\hat{\mathbb{E}}_{\boldsymbol{x}}[f] - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f]\right)\right|\right] \enspace . \tag{56}$$

The first term on the r.h.s. is $\mathrm{Y}(\boldsymbol{x})$. The second term is the ERA of the sample-dependent family

$$\mathcal{W}_{\boldsymbol{x}} \doteq \left\{ y \mapsto \frac{1}{m}(f(\boldsymbol{x}_j) - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f]), f \in \mathcal{F} \right\} \enspace .$$

Each function in $\mathcal{W}_{\boldsymbol{x}}$ is *constant*. Using (7) and the linearity of expectation, like we did in the proof of lemma 1.2.1 for the family $\mathcal{K}_{\boldsymbol{x}}$ (see (50)), it holds

$$\hat{\mathfrak{R}}_m(\mathcal{W}_{\boldsymbol{x}}, \boldsymbol{x}) = \frac{1}{m}\sup_{f\in\mathcal{F}}|f(\boldsymbol{x}_j) - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f]|\mathrm{b}(m) \leq \frac{r}{m}\mathrm{b}(m) \enspace .$$

Thus, continuing from (56) by incorporating the above fact, it holds

$$\mathrm{Y}_j(\boldsymbol{x}) \leq \mathrm{Y}(\boldsymbol{x}) + \frac{r}{m}\mathrm{b}(m) \enspace . \tag{57}$$

We now show that $\mathrm{Y}_j(\boldsymbol{x}) \geq \mathrm{Y}(\boldsymbol{x}) - (1+\mathrm{b}(m))r/m$. Starting from the definition of Y_j and adding and removing

$$\frac{1}{m}\left(\boldsymbol{\sigma}_j\big(f(\boldsymbol{x}_j) - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f]\big)\right)$$

to the argument of the supremum, it holds

$$\mathrm{Y}_j(\boldsymbol{x}) = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\left|\left(\sum_{\substack{i=1\\ i\neq j}}^{m} \boldsymbol{\sigma}_i \left(f(\boldsymbol{x}_i) - \mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}_{\setminus j}}[f]\right)\right) + \boldsymbol{\sigma}_j\big(f(\boldsymbol{x}_j) - \mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}_{\setminus j}}[f]\big) - \boldsymbol{\sigma}_j\big(f(\boldsymbol{x}_j) - \mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}_{\setminus j}}[f]\big)\right|\right] \ .$$

Then, from the triangle inequality and the fact that

$$\sup_{f\in\mathcal{F}}\left|\boldsymbol{\sigma}_j\big(f(\boldsymbol{x}_j) - \mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}_{\setminus j}}[f]\big)\right| \leq r,$$

we obtain

$$\mathrm{Y}_j(\boldsymbol{x}) \geq \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\left|\sum_{i=1}^{m} \boldsymbol{\sigma}_i \left(f(\boldsymbol{x}_i) - \mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}_{\setminus j}}[f]\right)\right|\right] - \frac{r}{m} \ .$$

From here, we add and subtract $\boldsymbol{\sigma}_i \hat{\mathbb{E}}_{\boldsymbol{x}}[f]$ to each term of the sum, and then use the triangle inequality, the subadditivity of the supremum, and the linearity of expectation, to obtain

$$\mathrm{Y}_j(\boldsymbol{x}) \geq \underbrace{\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\left|\sum_{i=1}^{m} \boldsymbol{\sigma}_i \left(f(\boldsymbol{x}_i) - \mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}}[f]\right)\right|\right]}_{=\mathrm{Y}(\boldsymbol{x})} - \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\left|\sum_{i=1}^{m} \boldsymbol{\sigma}_i \left(\mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}}[f] - \mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}_{\setminus j}}[f]\right)\right|\right] - \frac{r}{m} \ .$$

The second term on the r.h.s. is again the ERA of a family of constant functions, each of them taking value at most r/m. Thus using (7), it follows that

$$\mathrm{Y}_j(\boldsymbol{x}) \geq \mathrm{Y}(\boldsymbol{x}) - (1 + \mathrm{b}(m))\frac{r}{m} \ .$$

Combining the above and (57), we obtain

$$\mathrm{Y}(\boldsymbol{x}) - (1 + \mathrm{b}(m))\frac{r}{m} \leq \mathrm{Y}_j(\boldsymbol{x}) \leq \mathrm{Y}(\boldsymbol{x}) + \frac{r}{m}\mathrm{b}(m) \ . \tag{58}$$

We now show that

$$\sum_{j=1}^{m}\big(\mathrm{Y}(\boldsymbol{x}) - \mathrm{Y}_j(\boldsymbol{x})\big) \leq \mathrm{Y}(\boldsymbol{x}) \ . \tag{59}$$

Starting from the definition of the Y_j functions, and using the linearity of expectation and the subadditivity of the supremum

$$\begin{aligned}\sum_{j=1}^{m} \mathrm{Y}_j(\boldsymbol{x}) &= \sum_{j=1}^{m} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\left|\sum_{i=1, i\neq j}^{m} \boldsymbol{\sigma}_i\left(f(\boldsymbol{x}_i) - \mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}_{\setminus j}}[f]\right)\right|\right] \\ &\geq \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\left|\sum_{j=1}^{m}\sum_{i=1, i\neq j}^{m} \boldsymbol{\sigma}_i\left(f(\boldsymbol{x}_i) - \mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}_{\setminus j}}[f]\right) .\right|\right] \ .\end{aligned}$$

We rearrange the terms in the double sums, and use the linearity of expectation to obtain

$$\begin{aligned}\sum_{j=1}^{m} \mathrm{Y}_j(\boldsymbol{x}) &\geq \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\left|(m-1)\sum_{i=1}^{m} \boldsymbol{\sigma}_i\left(f(\boldsymbol{x}_i) - \mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}}[f]\right)\right|\right] \\ &\geq (m-1)\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\left|\sum_{i=1}^{m} \boldsymbol{\sigma}_i\left(f(\boldsymbol{x}_i) - \mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}}[f]\right)\right|\right] ,\end{aligned}$$

which completes our proof of (59), as the last expectation is $\mathrm{Y}(\boldsymbol{x})$.

Define now the functions

$$\mathrm{Z}(\boldsymbol{x}) \doteq \mathrm{Y}(\boldsymbol{x}) \text{ and } \mathrm{Z}_j(\boldsymbol{x}) \doteq \mathrm{Y}_j(\boldsymbol{x}) - \frac{r}{m}\mathrm{b}(m) \text{ for each } j = 1, \ldots, m \ .$$

The value of $\mathrm{Z}_j(\boldsymbol{x})$ clearly does not dependent on the j-th component of $\boldsymbol{x}$. Also, from (58) it follows that

$$\mathrm{Z}_j(\boldsymbol{x}) \leq \mathrm{Z}(\boldsymbol{x}) \leq \mathrm{Z}_j(\boldsymbol{x}) + (1 + 2\mathrm{b}(m))\frac{r}{m} \text{ for each } j = 1, \ldots, m \ .$$

A consequence of (59) is finally that

$$\sum_{j=1}^{m} \big(\mathrm{Z}(\boldsymbol{x}) - \mathrm{Z}_j(\boldsymbol{x})\big) \leq \mathrm{Z}(\boldsymbol{x}) + r\mathrm{b}(m) \ .$$

Thus Z, i.e., $\hat{\mathfrak{R}}_m(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}), \boldsymbol{x})$, is a $(1, r\mathrm{b}(m))$-self-bounding function with scale $(1 + 2\mathrm{b}(m))r/m$. An application of (54) from theorem A.1.2 completes the proof. ☐

Before proving theorem 1.3.1, we need the following lemma.

Lemma A.1.3. It holds

$$\mathrm{W}(\mathcal{F}) \leq \frac{m}{m-1}\mathop{\mathbb{E}}_{\boldsymbol{x}}[\widehat{\mathrm{W}}_{\boldsymbol{x}}(\mathcal{F})] \ .$$

Proof. Using Bessel's correction, we can rewrite the definition of wimpy variance to use the empirical expectation as

$$\mathrm{W}(\mathcal{F}) = \sup_{f\in\mathcal{F}} \mathop{\mathbb{E}}_{\boldsymbol{x}}\left[\frac{1}{m}\sum_{i=1}^{m}\left(f(\boldsymbol{x}_i) - \mathop{\mathbb{E}}_{\mathcal{D}}[f]\right)^2\right] = \sup_{f\in\mathcal{F}} \mathop{\mathbb{E}}_{\boldsymbol{x}}\left[\frac{1}{m-1}\sum_{i=1}^{m}\left(f(\boldsymbol{x}_i) - \mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}}[f]\right)^2\right] \ .$$

An application of Jensen's inequality gives

$$\mathrm{W}(\mathcal{F}) \le \mathop{\mathbb{E}}_{\boldsymbol{x}}\Bigg[\underbrace{\sup_{f\in\mathcal{F}} \frac{1}{m-1}\sum_{i=1}^{m}\left(f(\boldsymbol{x}_i) - \mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}}[f]\right)^2}_{=\frac{m}{m-1}\widehat{\mathrm{W}}_{\boldsymbol{x}}(\mathcal{F})}\Bigg] \ .$$

□

Theorem 1.3.1. Suppose $m \ge 2$. Let $\delta \in (0,1)$. With probability $\ge 1-\delta$ over the choice of $\boldsymbol{x}$,

$$\mathrm{W}(\mathcal{F}) \le \tfrac{m}{m-1}\widehat{\mathrm{W}}_{\boldsymbol{x}}(\mathcal{F}) + \frac{r^2 \ln\frac{1}{\delta}}{m-1} + \sqrt{\left(\frac{r^2\ln\frac{1}{\delta}}{m-1}\right)^2 + \frac{2r^2\,\frac{m}{m-1}\widehat{\mathrm{W}}_{\boldsymbol{x}}(\mathcal{F})\ln\frac{1}{\delta}}{m-1}} \ . \tag{11}$$

Proof. This proof proceeds by showing that $\widehat{\mathrm{W}}_{\boldsymbol{x}}(\mathcal{F})$ is a $({}^{m}\!/_{m-1}, 0)$-self-bounding with scale ${}^{r^2}\!/_{m}$, then applying lemma A.1.3, and finally (53) from theorem A.1.2.

Let $\boldsymbol{x}_{\setminus j}$ denote the vector $\boldsymbol{x}$ with the j-th component removed, as we defined it also in the proof for theorem 1.2.2. Let $\hat{\mathrm{V}}_{\boldsymbol{x}}[f]$ denote the (unbiased) sample variance of f over $\boldsymbol{x}$, i.e.,

$$\hat{\mathrm{V}}_{\boldsymbol{x}}[f] \doteq \frac{1}{m-1}\sum_{i=1}^{m}\left(f(\boldsymbol{x}_i) - \mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}}[f]\right)^2 \ .$$

Define

$$\mathrm{Z}(\boldsymbol{x}) \doteq \frac{m}{m-1}\widehat{\mathrm{W}}_{\boldsymbol{x}}(\mathcal{F}) = \sup_{f\in\mathcal{F}} \hat{\mathrm{V}}_{\boldsymbol{x}}[f] = \sup_{f\in\mathcal{F}} \frac{1}{m-1}\sum_{i=1}^{m}\left(f(\boldsymbol{x}_i) - \mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}}[f]\right)^2$$

and

$$\mathrm{Z}_j(\boldsymbol{x}) \doteq \sup_{f\in\mathcal{F}} \frac{1}{m-1}\sum_{i=1, i\neq j}^{m}\left(f(\boldsymbol{x}_i) - \mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}_{\setminus j}}[f]\right)^2 \ . \tag{60}$$

We first show that

$$\mathrm{Z}_j(\boldsymbol{x}) = \sup_{f\in\mathcal{F}}\left[\hat{\mathrm{V}}_{\boldsymbol{x}}[f] - \frac{1}{m}\left(f(x_j) - \mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}_{\setminus j}}[f]\right)^2\right], \tag{61}$$

as this form comes in handy many times. Starting from the definition of Z_j in (60), we add and subtract $\frac{1}{m-1}\left(f(x_j) - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f]\right)^2$ to the argument of the supremum, and then add and subtract $\hat{\mathbb{E}}_{\boldsymbol{x}}[f]$ to the argument of the sum, to obtain:

$$\begin{aligned}
\mathrm{Z}_j(\boldsymbol{x}) &= \sup_{f\in\mathcal{F}} \frac{1}{m-1}\left[\left(\sum_{i=1}^{m}\big(f(\boldsymbol{x}_i) - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f]\big)^2\right) - \big(f(\boldsymbol{x}_j) - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f]\big)^2\right] \\
&= \sup_{f\in\mathcal{F}} \frac{1}{m-1}\left[\left(\sum_{i=1}^{m}\Big(\big(f(\boldsymbol{x}_i) - \hat{\mathbb{E}}_{\boldsymbol{x}}[f]\big) + \big(\hat{\mathbb{E}}_{\boldsymbol{x}}[f] - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f]\big)\Big)^2\right) - \big(f(\boldsymbol{x}_j) - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f]\big)^2\right]
\end{aligned}$$

By expressing the square in the argument of the sum, separating the three resulting terms in three distinct sums (associative property of the sum), and noticing that one of these sum is $\sum_{i=1}^{m}(f(\boldsymbol{x}_i) - \hat{\mathbb{E}}_{\boldsymbol{x}}[f]) = 0$, and another has argument $(\hat{\mathbb{E}}_{\boldsymbol{x}}[f] - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f])^2$ independent from i, we obtain

$$\mathrm{Z}_j(\boldsymbol{x}) = \sup_{f\in\mathcal{F}} \frac{1}{m-1}\left[\underbrace{\left(\sum_{i=1}^{m}\big(f(\boldsymbol{x}_i) - \hat{\mathbb{E}}_{\boldsymbol{x}}[f]\big)^2\right)}_{=(m-1)\hat{\mathrm{V}}_{\boldsymbol{x}}[f]} + m\big(\hat{\mathbb{E}}_{\boldsymbol{x}}[f] - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f]\big)^2 - \big(f(\boldsymbol{x}_j) - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f]\big)^2\right] .$$

It holds $\hat{\mathbb{E}}_{\boldsymbol{x}}[f] = \frac{1}{m}f(\boldsymbol{x}_j) + \frac{m-1}{m}\hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f]$, so we have

$$\mathrm{Z}_j(\boldsymbol{x}) = \sup_{f\in\mathcal{F}} \frac{1}{m-1}\left[(m-1)\hat{\mathrm{V}}_{\boldsymbol{x}}[f] + m\left(\frac{1}{m}f(\boldsymbol{x}_j) - \frac{1}{m}\hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f]\right)^2 - \big(f(\boldsymbol{x}_j) - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f]\big)^2\right] .$$

The identity in (68) then follows through simple algebraic steps.

We want to show that Z is a $({}^{m}/_{m-1}, 0)$-self-bounding function with scale ${}^{r^2}/_{m}$ (see definition A.1.1). By definition of Z_j in (60), the value of $\mathrm{Z}_j(\boldsymbol{x})$ does not depend on the j-th component of $\boldsymbol{x}$, as required by the first point in definition A.1.1.

We now show that, for any $j = 1, \dots, m$, it holds,

$$\mathrm{Z}_j(\boldsymbol{x}) \leq \mathrm{Z}(\boldsymbol{x}) \leq \mathrm{Z}_j(\boldsymbol{x}) + \frac{r^2}{m} \text{ for any } \boldsymbol{x} \in \mathcal{X}^m, \tag{62}$$

as required by the second point in definition A.1.1. The leftmost inequality follows from the definitions of Z and Z_j. To show the rightmost inequality, we start from (68), and use the subadditivity of the supremum to obtain

$$\mathrm{Z}_j(\boldsymbol{x}) \geq \left[\underbrace{\left(\sup_{f\in\mathcal{F}} \hat{\mathrm{V}}_{\boldsymbol{x}}[f]\right)}_{=\mathrm{Z}(\boldsymbol{x})} - \left(\sup_{f\in\mathcal{F}} \frac{1}{m}\big(f(\boldsymbol{x}_j) - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f]\big)^2\right)\right] .$$

The rightmost supremum is always smaller than ${}^{r^2}/_{m}$ because $|f(\boldsymbol{x}_j) - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f]| \leq r$, thus we have obtained the rightmost inequality in (62).

We now show that, for any $\boldsymbol{x} \in \mathcal{X}^m$, it holds

$$\sum_{i=1}^{m} \big(\mathrm{Z}(\boldsymbol{x}) - \mathrm{Z}_j(\boldsymbol{x})\big) \leq \frac{m}{m-1}\mathrm{Z}(\boldsymbol{x}),$$

as in the last requirement of definition A.1.1. Starting again from (68) and using the subadditivity of the supremum, it holds

$$\sum_{j=1}^{m} \mathrm{Z}_j(\boldsymbol{x}) = \sum_{j=1}^{m} \sup_{f \in \mathcal{F}} \left[\hat{\mathrm{V}}_{\boldsymbol{x}}[f] - \frac{1}{m}\big(f(\boldsymbol{x}_j) - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f]\big)^2\right] \geq \sup_{f \in \mathcal{F}} \sum_{j=1}^{m} \left[\hat{\mathrm{V}}_{\boldsymbol{x}}[f] - \frac{1}{m}\big(f(\boldsymbol{x}_j) - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f]\big)^2\right] \ .$$

By simple algebra we then get

$$\sum_{j=1}^{m} \mathrm{Z}_j(\boldsymbol{x}) \geq \sup_{f \in \mathcal{F}} \left[m\hat{\mathrm{V}}_{\boldsymbol{x}}[f] - \frac{1}{m}\sum_{j=1}^{m}\big(f(\boldsymbol{x}_j) - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f]\big)^2\right] \ .$$

From here, we use the fact that

$$\hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f] = \frac{1}{m-1}\big(m\,\hat{\mathbb{E}}_{\boldsymbol{x}}[f] - f(\boldsymbol{x}_j)\big),$$

to get

$$\sum_{j=1}^{m} \mathrm{Z}_j(\boldsymbol{x}) \geq \sup_{f \in \mathcal{F}} \left[m\hat{\mathrm{V}}_{\boldsymbol{x}}[f] - \frac{1}{m}\sum_{j=1}^{m}\left(\frac{m}{m-1}f(\boldsymbol{x}_j) - \frac{m}{m-1}\hat{\mathbb{E}}_{\boldsymbol{x}}[f]\right)^2\right] \ .$$

Now by simplifying some terms on the r.h.s., we obtain

$$\sum_{j=1}^{m} \mathrm{Z}_j(\boldsymbol{x}) \geq \sup_{f \in \mathcal{F}} \left[m\hat{\mathrm{V}}_{\boldsymbol{x}}[f] - \frac{m}{(m-1)}\underbrace{\frac{1}{m-1}\sum_{j=1}^{m}\left(f(\boldsymbol{x}_j) - \hat{\mathbb{E}}_{\boldsymbol{x}}[f]\right)^2}_{=\hat{\mathrm{V}}_{\boldsymbol{x}}[f]}\right] \ .$$

Collecting terms and using the original definition of Z results in

$$\sum_{j=1}^{m} \mathrm{Z}_j(\boldsymbol{x}) \geq \left(m - \frac{m}{m-1}\right)\mathrm{Z}(\boldsymbol{x}) \ .$$

Thus,

$$\sum_{j=1}^{m} \big(\mathrm{Z}(\boldsymbol{x}) - \mathrm{Z}_j(\boldsymbol{x})\big) \leq m\mathrm{Z}(\boldsymbol{x}) - \left(m - \frac{m}{m-1}\right)\mathrm{Z}(\boldsymbol{x}) \leq \frac{m}{m-1}\mathrm{Z}(\boldsymbol{x}),$$

which concludes our proof that Z, is $({}^{m}/_{m-1}, 0)$-self-bounding with scale ${}^{r^2}/_{m}$.

We now use the above fact to prove the thesis. A consequence of lemma A.1.3 is

$$\mathop{\mathbb{P}}_{\boldsymbol{x}}\left(\widehat{\mathrm{W}}_{\boldsymbol{x}}(\mathcal{F}) \leq \mathrm{W}(\mathcal{F}) - \varepsilon\right) \leq \mathop{\mathbb{P}}_{\boldsymbol{x}}\left(\widehat{\mathrm{W}}_{\boldsymbol{x}}(\mathcal{F}) \leq \frac{m}{m-1} \mathop{\mathbb{E}}_{\boldsymbol{x}}[\widehat{\mathrm{W}}_{\boldsymbol{x}}(\mathcal{F})] - \varepsilon\right) .$$

From here, we use the definition

$$\mathrm{Z}(\boldsymbol{x}) = \frac{m}{m-1}\widehat{\mathrm{W}}_{\boldsymbol{x}}(\mathcal{F})$$

and apply (53) from theorem A.1.2 to obtain the thesis. ☐

The constants in this bound are somewhat sub-optimal, as there is a significant gap between the best-known (sub-Poisson) tails for $(1,0)$-self-bounding and the best-known (sub-gamma) tails for $(1+\varepsilon,0)$-self-bounding functions. We hope that future work leads to refined analysis of tail bounds for $(\alpha,0)$-self-bounding functions that decay gracefully as α exceeds 1.

Lemma 1.3.4. For any $\boldsymbol{x} \in \mathcal{X}^m$, it holds

$$\hat{\mathfrak{R}}_m(\mathcal{F},\boldsymbol{x}) \geq \sqrt{\frac{\widehat{\mathrm{W}}^{\mathrm{r}}_{\boldsymbol{x}}(\mathcal{F})}{2m}} \text{ and } \hat{\mathfrak{R}}_m(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}),\boldsymbol{x}) \geq \sqrt{\frac{\widehat{\mathrm{W}}_{\boldsymbol{x}}(\mathcal{F})}{2m}} .$$

Furthermore, it holds

$$\lim_{m\to\infty} \sqrt{m}\mathfrak{R}_m(\mathcal{F},\mathcal{D}) \geq \sqrt{\tfrac{2}{\pi}\mathrm{W}^{\mathrm{r}}(\mathcal{F})} \text{ and } \lim_{m\to\infty} \sqrt{m}\mathfrak{R}_m(\mathrm{C}_{\mathcal{D}}(\mathcal{F}),\mathcal{D}) \geq \sqrt{\tfrac{2}{\pi}\mathrm{W}(\mathcal{F})} .$$

Proof. From the subadditivity of the supremum, it holds that

$$\hat{\mathfrak{R}}_m(\mathcal{F},\boldsymbol{x}) \geq \sup_{f\in\mathcal{F}} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\left|\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{\sigma}_i f(\boldsymbol{x}_i)\right|\right] .$$

An application of Khintchine's inequality [Haagerup, 1982] gives

$$\hat{\mathfrak{R}}_m(\mathcal{F},\boldsymbol{x}) \geq \sup_{f\in\mathcal{F}} \frac{1}{\sqrt{2}}\sqrt{\frac{\|f(\boldsymbol{x})\|_2^2}{m^2}},$$

where $f(\boldsymbol{x})$ denotes the m-dimensional vector of values of f on $\boldsymbol{x}$. The proof of the leftmost inequality in the thesis ends by noting that

$$\widehat{\mathrm{W}}^{\mathrm{r}}_{\boldsymbol{x}}(\mathcal{F}) = \frac{\|f(\boldsymbol{x})\|_2^2}{m} .$$

The rightmost inequality is then a corollary, using the identity $\widehat{\mathrm{W}}^{\mathrm{r}}_{\boldsymbol{x}}(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F})) = \widehat{\mathrm{W}}_{\boldsymbol{x}}(\mathcal{F})$.

The asymptotic lower bounds follow by replacing the Khintchine's inequality step with an application of the central limit theorem. $\square$

Before proving theorem 1.3.7 we need to introduce an important technical result. For any $u \in \mathbb{R}$, let $\mathrm{h}(u) \doteq (1+u)\ln(1+u) - u$, and let $(u)_+ \doteq \max(0, u)$.

Theorem A.1.4 (Samson's bound, [Boucheron et al., 2013, Thm. 12.11])**.** Let $\mathcal{Q}_1, \ldots, \mathcal{Q}_m$ be possibly different probability distributions over a domain $\mathcal{Y}$. Let $\mathcal{G} \subseteq \mathcal{X} \to [-1, 1]$. Furthermore, assume that for each $g \in \mathcal{G}$ and $i \in \{1, \ldots, m\}$, it holds $\mathbb{E}_{\mathcal{Q}_i}[g] = 0$. Now, for any $\boldsymbol{y} \in \mathcal{Y}^m$, let

$$\mathrm{Z}(\boldsymbol{y}) \doteq \sup_{g \in \mathcal{G}} \sum_{i=1}^{m} g(y_i) \text{ and } S^2 \doteq \mathop{\mathbb{E}}_{\boldsymbol{y}}\left[\sup_{g \in \mathcal{F}} \sum_{i=1}^{m} \mathop{\mathbb{E}}_{y_i' \sim \mathcal{Q}_i}\left[\left(\left(g(y_i) - g(y_i')\right)_+\right)^2\right]\right] .$$

Let $\boldsymbol{y} \in \mathcal{Y}^m$, with each $y_i \sim \mathcal{Q}_i$, independently (but not necessarily identically, since the distributions may be different). It holds[1]

$$\mathop{\mathbb{P}}_{\boldsymbol{y}}\left(\mathrm{Z}(\boldsymbol{y}) \leq \mathop{\mathbb{E}}_{\mathcal{Q}_{1:m}}[\mathrm{Z}] - \varepsilon\right) \leq \exp\left(-\frac{S^2}{4}\mathrm{h}\left(\frac{2\varepsilon}{S^2}\right)\right) . \tag{63}$$

Theorem 1.3.7. Let $\boldsymbol{\sigma} \in (\pm 1)^{n \times m}$ be a matrix of i.i.d. Rademacher r.v.'s. Let $\delta \in (0, 1)$. With probability at least $1 - \delta$ over the choice of $\boldsymbol{\sigma}$, it holds

$$\hat{\mathfrak{R}}_m(\mathcal{F}, \boldsymbol{x}) \leq \hat{\mathfrak{R}}^n_m(\mathcal{F}, \boldsymbol{x}, \boldsymbol{\sigma}) + \frac{2\hat{q}_{\mathcal{F}}(\boldsymbol{x}) \ln \frac{1}{\delta}}{3nm} + \sqrt{\frac{4\widehat{\mathrm{W}}^{\mathrm{r}}_{\boldsymbol{x}}(\mathcal{F}) \ln \frac{1}{\delta}}{nm}} . \tag{14}$$

Proof. Without loss of generality, we assume that $\hat{q}_{\mathcal{F}}(\boldsymbol{x}) = 1$. The general case then follows via scaling.

Let

$$\mathrm{Z}(\boldsymbol{\sigma}) \doteq nm\hat{\mathfrak{R}}^n_m(\mathcal{F}, \boldsymbol{x}, \boldsymbol{\sigma}) = \sum_{j=1}^{n} \sup_{f \in \mathcal{F}} \left|\sum_{i=1}^{m} \boldsymbol{\sigma}_{j,i} f(\boldsymbol{x}_i)\right| .$$

It holds $\mathbb{E}_{\boldsymbol{\sigma}}[\mathrm{Z}] = nm\hat{\mathfrak{R}}_m(\mathcal{F}, \boldsymbol{x})$.

We first show that we can apply Samson's bound (theorem A.1.4) to Z, i.e., to the scaled MC-ERA. Consider the function family $\mathcal{F}_\pm$ introduced in corollary 1.3.5, and consider the n-times Cartesian product of $\mathcal{F}_\pm$ with itself

$$(\mathcal{F}_\pm)^n = \underbrace{\mathcal{F}_\pm \times \cdots \times \mathcal{F}_\pm}_{n \text{ times}} .$$

We use $\boldsymbol{f} = (f_1, \ldots, f_n)$ to denote an element of $(\mathcal{F}_\pm)^n$. Now, define the family

$$\mathcal{G} \doteq \{g(\boldsymbol{\sigma}_{j,i}) \doteq \boldsymbol{\sigma}_{j,i} f_j(\boldsymbol{x}_i), \boldsymbol{f} \in (\mathcal{F}_\pm)^n\} .$$

[1] To be precise, this is an immediate consequence of the statement of [Boucheron et al., 2013, Thm. 2.11], through an application of the Chernoff method to the moment generating function given therein.

The functions in $\mathcal{G}$ have domain $\mathcal{Y} = \{-1, 1\}$ and values in $[-1, 1]$. It holds

$$\mathrm{Z}(\boldsymbol{\sigma}) = \sup_{\boldsymbol{f} \in (\mathcal{F}_{\pm})^n} \sum_{j=1}^{n} \sum_{i=1}^{m} \boldsymbol{\sigma}_{j,i} f_j(\boldsymbol{x}_i) = \sup_{g \in \mathcal{G}} \sum_{(j,i) \in \{1,\ldots n\} \times \{1,\ldots,m\}} g(\boldsymbol{\sigma}_{j,i}) \ . \tag{64}$$

Thus Z has the form required by theorem A.1.4.

Let $\boldsymbol{\sigma'}$ denote a second $n \times m$ i.i.d. Rademacher matrix (like $\boldsymbol{\sigma}$), and define

$$\begin{aligned} S^2 &\doteq \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{\boldsymbol{f} \in (\mathcal{F}_{\pm})^n} \sum_{j=1}^{n} \sum_{i=1}^{m} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}'_{j,i}} \left[\left(\left(\boldsymbol{\sigma}_{j,i} f_j(\boldsymbol{x}_i) - \boldsymbol{\sigma}'_{j,i} f_j(\boldsymbol{x}_i) \right)_+ \right)^2 \right] \right] \\ &= n \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{f \in \mathcal{F}_{\pm}} \sum_{i=1}^{m} 2 \left(\left(\boldsymbol{\sigma}_{1,i} f(\boldsymbol{x}_i) \right)_+ \right)^2 \right] \ . \end{aligned}$$

It holds

$$S^2 \leq 2nm \widehat{\mathrm{W}}^{\mathrm{r}}_{\boldsymbol{x}}(\mathcal{F}) \ . \tag{65}$$

For each $g \in \mathcal{G}$, $g(\boldsymbol{\sigma}_{j,i})$ and $g(\boldsymbol{\sigma}_{j',i'})$ are *independent*, though not necessarily *identically distributed*, for $(j,i) \neq (j',i')$, due to the dependence of $g(\boldsymbol{\sigma}_{j,i})$ on indices (j,i). It also holds, for each $g \in \mathcal{G}$, and indices (j,i), that $\mathbb{E}_{\boldsymbol{\sigma}_{i,j}}[g(\boldsymbol{\sigma}_{i,j})] = 0$, simply due to multiplication by symmetric (Rademacher) r.v.'s.

Thus, we can use Samson's bound (theorem A.1.4) on $\mathcal{G}$, Z, and S^2, although it is generally more convenient to work with $\mathcal{F}$ and $(\mathcal{F}_{\pm})^n$.

We now show the thesis. Fix $\varepsilon \in (0,1)$. It follows from Samson's bound that

$$\mathop{\mathbb{P}}_{\boldsymbol{\sigma}} \left(\hat{\mathfrak{R}}_m(\mathcal{F}, \boldsymbol{x}) \geq \hat{\mathfrak{R}}^n_m(\mathcal{F}, \boldsymbol{x}, \boldsymbol{\sigma}) + \varepsilon \right) = \mathop{\mathbb{P}}_{\boldsymbol{\sigma}} \left(\mathbb{E}[\mathrm{Z}] \geq \mathrm{Z}(\boldsymbol{\sigma}) + nm\varepsilon \right) \leq \exp\left(-\frac{S^2}{4} \mathrm{h}\left(\frac{2nm\varepsilon}{S^2} \right) \right) \ .$$

The function

$$g(x) \doteq x \mathrm{h}\left(\frac{2nm\varepsilon}{x} \right)$$

is monotonically decreasing in its argument. Thus, using (65) gives

$$\mathop{\mathbb{P}}_{\boldsymbol{\sigma}} \left(\hat{\mathfrak{R}}_m(\mathcal{F}, \boldsymbol{x}) \geq \hat{\mathfrak{R}}^n_m(\mathcal{F}, \boldsymbol{x}, \boldsymbol{\sigma}) + \varepsilon \right) \leq \exp\left(\frac{-nm \widehat{\mathrm{W}}^{\mathrm{r}}_{\boldsymbol{x}}(\mathcal{F})}{2} \mathrm{h}\left(\frac{\varepsilon}{\widehat{\mathrm{W}}^{\mathrm{r}}_{\boldsymbol{x}}(\mathcal{F})} \right) \right) \ .$$

Now, for $u > -1/2$, define the function

$$\mathrm{h}_1(u) \doteq 1 + u - \sqrt{1 + 2u} \ .$$

Using the fact (see Boucheron et al. [2013, Ch. 2.4]) that

$$\mathrm{h}(u) \geq 9\mathrm{h}_1\left(\frac{u}{3}\right) \text{ for every } u \in (-1, +\infty),$$

we obtain

$$\mathop{\mathbb{P}}_{\boldsymbol{\sigma}}\left(\hat{\mathfrak{R}}_m(\mathcal{F}, \boldsymbol{x}) \geq \hat{\mathfrak{R}}^n_m(\mathcal{F}, \boldsymbol{x}, \boldsymbol{\sigma}) + \varepsilon\right) \leq \exp\left(-\frac{9}{2}nm\widehat{\mathrm{W}}^{\mathrm{r}}_{\boldsymbol{x}}(\mathcal{F})\mathrm{h}_1\left(\frac{\varepsilon}{3\widehat{\mathrm{W}}^{\mathrm{r}}_{\boldsymbol{x}}(\mathcal{F})}\right)\right) .$$

The result for $\hat{q}_{\mathcal{F}}(\boldsymbol{x}) = 1$ is obtained by imposing that the r.h.s. be at most δ and solving for ε using standard sub-gamma inequalities. The general case then follows via linear scaling. □

This bound is quite comparable to Bousquet's bound on the SD (see theorem 1.3.2). The variance factors $\widehat{\mathrm{W}}^{\mathrm{r}}_{\boldsymbol{x}}(\mathcal{F})$ and $\widehat{\mathrm{W}}_{\boldsymbol{x}}(\mathcal{F})$ are convenient, as they depend only on sample variances, rather than true variances and expected supremum deviations.

Even if Samson's inequality introduces additional 2-factors on both the range and variance w.r.t. theorem 1.3.2, both are divided by MC-trial count n, so for $n \geq 2$ trials, the Monte-Carlo error terms become negligible.

A.2 Details on the Experimental Evaluation

As mentioned in the main text, lemma 1.4.1 is a consequence of [Shalev-Shwartz and Ben-David, 2014, Lemmas 26.11, 26.10], reported here for completeness.[2]

Lemma A.2.1 (Shalev-Shwartz and Ben-David, 2014, Lemmas 26.11, 26.10)**.** It holds

$$\hat{\mathfrak{R}}_m(\mathcal{F}_1, \boldsymbol{x}) = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\left\|\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{\sigma}_i\boldsymbol{x}_i\right\|_\infty\right] \leq \max_i \|\boldsymbol{x}_i\|_\infty\sqrt{\frac{2\ln(2d)}{m}},$$

and

$$\hat{\mathfrak{R}}_m(\mathcal{F}_2, \boldsymbol{x}) = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\left\|\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{\sigma}_i\boldsymbol{x}_i\right\|_2\right] \leq \max_i \|\boldsymbol{x}_i\|_2\frac{1}{\sqrt{m}} .$$

We now show the centralized variants.

[2]The identities in the lemma are not reported in the original, but can be easily obtained through a slightly more refined proof than the one presented in the original. See the proof of lemma 1.4.1 for intuition.

Lemma 1.4.1. Let $\bar{\boldsymbol{x}} \doteq \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{x}_i \in \mathbb{R}^d$. For the ℓ_1 norm, it holds

$$\hat{\mathfrak{R}}_m(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}_1), \boldsymbol{x}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\left\|\frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\sigma}_i(\boldsymbol{x}_i - \bar{\boldsymbol{x}})\right\|_\infty\right] \leq \max_i \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}\|_\infty \sqrt{\frac{2\ln(2d)}{m}},$$

while for the ℓ_2 norm, it holds

$$\hat{\mathfrak{R}}_m(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}_2), \boldsymbol{x}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\left\|\frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\sigma}_i(\boldsymbol{x}_i - \bar{\boldsymbol{x}})\right\|_2\right] \leq \max_i \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}\|_2 \frac{1}{\sqrt{m}} \ .$$

Proof. We show the ℓ_2 case in detail; the reasoning for the ℓ_1 case is essentially the same (see details at the end of the proof). The definition of $\hat{\mathfrak{R}}_m(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}_2), \boldsymbol{x})$ is

$$\hat{\mathfrak{R}}_m(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}_2), \boldsymbol{x}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{w:\|w\|_2 \leq 1}\left|\frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\sigma}_i\big(w \cdot \boldsymbol{x}_i - \hat{\mathbb{E}}_{\boldsymbol{x}}[w]\big)\right|\right],$$

where

$$\hat{\mathbb{E}}_{\boldsymbol{x}}[w] = \frac{1}{m}\sum_{i=1}^{m}(w \cdot \boldsymbol{x}_i) = w \cdot \bar{\boldsymbol{x}} \ .$$

Using linearity, we then get

$$\hat{\mathfrak{R}}_m(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}_2), \boldsymbol{x}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{w:\|w\|_2 \leq 1}\left|w \cdot \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\sigma}_i(\boldsymbol{x}_i - \bar{\boldsymbol{x}})\right|\right] \ .$$

Now, for ease of notation, let $u \doteq \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\sigma}_i(\boldsymbol{x}_i - \bar{\boldsymbol{x}})$. The supremum is realized when

$$w = \frac{u}{\|u\|_2},$$

because in this case the vector w has the same direction as u, and the largest possible norm $\|w\|_2 = 1$. Since the two vectors w and u are collinear, the Cauchy-Schwarz inequality holds with equality, and we have

$$w \cdot u = \|w\|_2\|u\|_2 = \|u\|_2 = \left\|\frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\sigma}_i(\boldsymbol{x}_i - \bar{\boldsymbol{x}})\right\|_2 \ .$$

We thus obtain

$$\hat{\mathfrak{R}}_m(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}_2), \boldsymbol{x}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\left\|\frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\sigma}_i(\boldsymbol{x}_i - \bar{\boldsymbol{x}})\right\|_2\right] \ .$$

From here, we can proceed as in the second part of the proof of [Shalev-Shwartz and Ben-David, 2014, Lemma 26.10] to obtain the thesis.

By similar reasoning (now with Hölder's inequality in place of the Cauchy-Schwarz inequality, and following the proof of Shalev-Shwartz and Ben-David [2014, Lemma 26.11]), we get that

$$\hat{\mathfrak{R}}_m(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}_1), \boldsymbol{x}) = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\left\|\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{\sigma}_i(\boldsymbol{x}_i - \bar{\boldsymbol{x}})\right\|_\infty\right] \leq \max_i \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}\|_\infty \sqrt{\frac{2\ln(2d)}{m}} \enspace .$$

□

A.2.1 Data Generation

Our data distributions for both the ℓ_1 and ℓ_2 constrained linear family experiments are both randomized and parameterized by dimension d. Rademacher averages and wimpy variances depend on the randomization and d, and ranges may be bounded *a priori* in terms of d.

ℓ_1 **Datasets** In our ℓ_1 experiments, each $\boldsymbol{x}_j$ is independently Beta-distributed, thus $\boldsymbol{x} \sim \mathrm{B}(\boldsymbol{\alpha}_1, \boldsymbol{\beta}_1) \times \cdots \times \mathrm{B}(\boldsymbol{\alpha}_d, \boldsymbol{\beta}_d)$. The parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are themselves randomized, in particular, we sample $\boldsymbol{\alpha}_j$ and $\boldsymbol{\beta}_j$ from $\sqrt{\chi^2_j}$, where χ^2_k is the χ^2 distribution with k degrees of freedom. In these datasets, $r = q = 1$.

ℓ_2 **Datasets** In our ℓ_2 experiments, we generate random *mean vector* $\boldsymbol{\mu} \in \mathbb{R}^d$ and *covariance matrix* $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$, then sample $\boldsymbol{x}' \sim \mathcal{N}(\mu, \boldsymbol{\Sigma})$, and finally obtain sample $\boldsymbol{x}$ by projecting $\boldsymbol{x}'$ to the nonnegative hyperquadrant of the radius $\sqrt{d}$ ℓ_2 sphere; i.e.,

$$\boldsymbol{x} = \mathop{\mathrm{argmin}}_{\boldsymbol{x}\in\mathbb{R}^d : \|\boldsymbol{x}\|_2 \leq \sqrt{d} \wedge \mathbf{0} \preceq \boldsymbol{x}} \|\boldsymbol{x} - \boldsymbol{x}'\|_2 \enspace .$$

Taking I_d to be the identity matrix, we sample $\boldsymbol{\mu} \sim \mathcal{N}(\mathbf{1}, \mathrm{I}_d)$, and taking $\boldsymbol{a} \sim \mathcal{U}(0,1)^{d\times d}$, we let $\boldsymbol{\Sigma} \doteq \frac{\boldsymbol{a}\boldsymbol{a}^\top}{d} + \mathrm{I}_d$. In these datasets, $r = q = \sqrt{d}$.

A.2.2 Supplementary Plots

Figure A.1 shows the same results as figure 1.1 (in the main text), but without the scaling of the quantities by $\sqrt{m}$. Similarly, figure A.2 shows the same results as figure 1.2, sans scaling by $\sqrt{m}$. Additionally, both plots also include a *McDiarmid term* $3r\sqrt{\ln \frac{1}{\eta}/2m}$, representing the *additive error* incurred bounding the SD in terms of $\hat{\mathfrak{R}}^1_m(\mathcal{F}, \boldsymbol{x}, \boldsymbol{\sigma})$. We stress that this term *does not* include the MC-ERA itsef, and thus is just

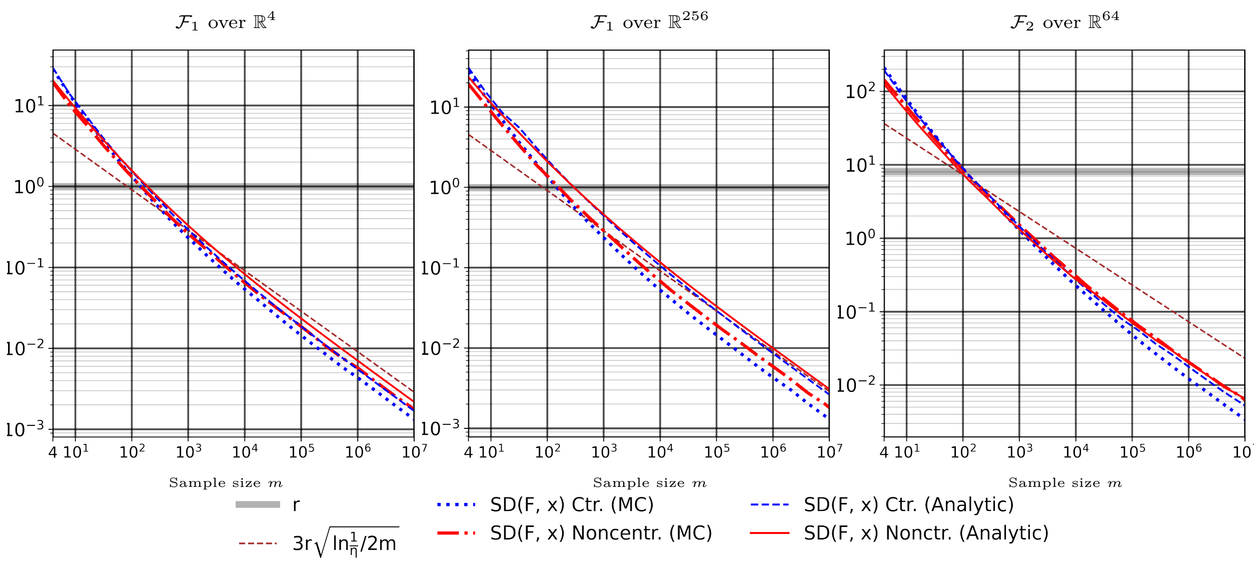


Figure A.1: Comparison of SD bounds as functions of the sample size m. See the main text for details.

one summand of the total McDiarmid SD bound. Nevertheless, the McDiarmid term alone asymptotically exceeds *all other bounds* in all experiments, except for the (loose) noncentralized analytical bound of $\mathcal{F}_1$ over $\mathbb{R}^{256}$. This sharply illustrates the benefit of *variance-sensitive* bounds over (range-only) McDiarmid bounds.

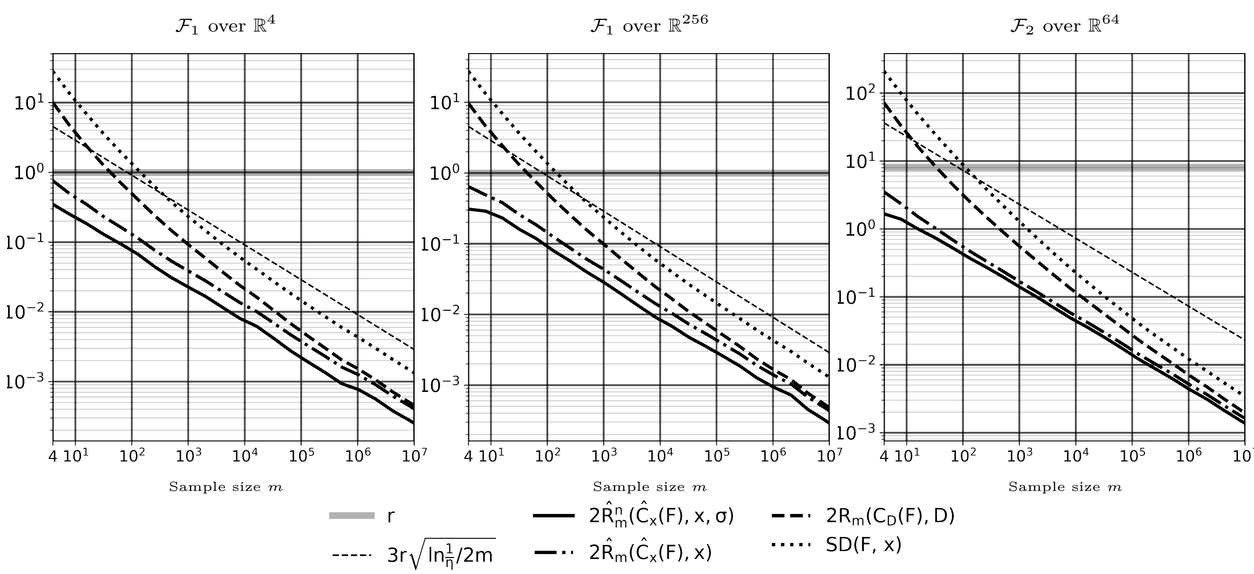


Figure A.2: Comparison of SD bounds as functions of the sample size m. See the main text for details.

Appendix B

Supplementary Material for Chapter 2

B.1 Deferred proofs

Proof of Lemma 2.4.1. For the sake of the proof, assume that we have two labeled set of samples of size m and m_L from $p_{\mathcal{X}\mathcal{Y}}$, call them respectively S and S_L. The set S represents our unlabeled sample, and the set S_L represents the labeled sample. For any $\delta \in (0,1)$, we would like to find a $\gamma > 0$ such that with probability $1-\delta$, for all $i \in 1,\dots,n$ (simultaneously), we have

$$\left| \frac{1}{m} \sum_{(x,\boldsymbol{y}) \in S} \ell(\boldsymbol{\varphi}_i(x), \boldsymbol{y}) - \frac{1}{m_L} \sum_{(x,\boldsymbol{y}) \in S_L} \ell(\boldsymbol{\varphi}_i(x), \boldsymbol{y}) \right| \leq \gamma \ . \tag{66}$$

The sample S represents the unlabeled data $x_1,\dots,x_m$ we have access to. In fact, $\frac{1}{m}\sum_{(x,\boldsymbol{y})\in S} \ell(\boldsymbol{\varphi}_i(x), \boldsymbol{y}) = \hat{\mathrm{R}}(\boldsymbol{\varphi}_i; X, \boldsymbol{Y}^*)$. The inequality (66) implies that for the true labeling of the unlabeled data $x_1,\dots,x_m$, for all $i \in 1,\dots,n$, it holds that:

$$\hat{\mathrm{R}}(\boldsymbol{\varphi}_i; X, \boldsymbol{Y}^*) \in [\mu_i - \gamma, \mu_i + \gamma]$$

where $\mu_i = \frac{1}{m_L}\sum_{(x,\boldsymbol{y})\in S_L} \ell(\boldsymbol{\varphi}_i(x), \boldsymbol{y})$ is the empirical mean computed from the labeled sample S_L.

By using Hoeffding's inequality, we have that for a fixed i, it holds that:

$$\begin{aligned}
&\mathop{\mathbb{P}}_{S,S_L}\left(\left|\frac{1}{m}\sum_{(x,\boldsymbol{y})\in S}\ell(\boldsymbol{\varphi}_i(x),\boldsymbol{y})-\frac{1}{m_L}\sum_{(x,\boldsymbol{y})\in S_L}\ell(\boldsymbol{\varphi}_i(x),\boldsymbol{y})\right|>\gamma\right)\\
&\leq 2\exp\left(\frac{-2\gamma^2}{\sum_{j=1}^{m}(\frac{B}{m})^2+\sum_{j=1}^{m_L}(\frac{B}{m_L})^2}\right)\\
&=2\exp\left(\frac{-2m_L m\gamma^2}{B^2(m+m_L)}\right)=\frac{\delta}{n}
\end{aligned}\tag{67}$$

By taking a union bound and solving (67) with respect to γ, the statement follows.

Proof of Lemma 2.4.5. By invoking lemma 2.4.4, it is easy to see that the function $\mathrm{R}(\boldsymbol{h}_{\boldsymbol{\theta}}; X, \boldsymbol{Y'})$ is a convex combination of convex functions with respect to $\boldsymbol{\theta}$, hence it is also convex in $\boldsymbol{\theta}$. Let $\boldsymbol{v} \in \partial\mathrm{R}(\boldsymbol{h}_{\boldsymbol{\theta}}; X, \boldsymbol{Y'})$. By Fact 3, such a vector always exists. Then, we have that for any $\boldsymbol{\theta'} \in \Theta$, it holds that:

$$\mathrm{R}(\boldsymbol{h}_{\boldsymbol{\theta'}}, \boldsymbol{Y'}) - \mathrm{R}(\boldsymbol{h}_{\boldsymbol{\theta}}, \boldsymbol{Y'}) \geq v^T(\boldsymbol{\theta'} - \boldsymbol{\theta})$$

As $f(\boldsymbol{\theta'}) \geq \mathrm{R}(\boldsymbol{h}_{\boldsymbol{\theta'}}, \boldsymbol{Y'})$, we have that:

$$f(\boldsymbol{\theta'}) - f(\boldsymbol{\theta}) \geq \boldsymbol{v}^T(\boldsymbol{\theta'} - \boldsymbol{\theta})$$

which implies that $\boldsymbol{v}$ is a subgradient of f at $\boldsymbol{\theta}$.

Proof of Theorem 2.4.6 We need to show that $f(\boldsymbol{\theta})$ is convex and L-Lipschitz continuous with respect to $\boldsymbol{\theta}$ to apply the standard convergence result for constant step size subgradient optimization [Bertsekas, 2015]

To show that $f(\boldsymbol{\theta})$ is convex it is straightforward to see that $\hat{\mathrm{R}}(\boldsymbol{h}_{\boldsymbol{\theta}}, \boldsymbol{Y})$ is convex in $\boldsymbol{\theta}$ as it is the convex combination of convex functions in $\boldsymbol{\theta}$. For any $\lambda \in [0, 1]$, we have that:

$$\begin{aligned}
f(\lambda\boldsymbol{\theta}_1 + (1-\lambda)\boldsymbol{\theta}_2) &= \max_{\boldsymbol{Y}\in\mathbb{Y}} L_c(h_{\lambda\boldsymbol{\theta}_1+(1-\lambda)\boldsymbol{\theta}_2}, \boldsymbol{Y})\\
&\leq \max_{\boldsymbol{Y}\in\mathbb{Y}} \left[\lambda L_c(\boldsymbol{h}_{\boldsymbol{\theta}_1}, \boldsymbol{Y} + (1-\lambda)L_c(\boldsymbol{h}_{\boldsymbol{\theta}_2}, \boldsymbol{Y})\right]\\
&\leq \lambda \max_{\boldsymbol{Y}\in\mathbb{Y}} L_c(\boldsymbol{h}_{\boldsymbol{\theta}_1}, \boldsymbol{Y}) + (1-\lambda)\max_{\boldsymbol{Y}\in\mathbb{Y}} L_c(\boldsymbol{h}_{\boldsymbol{\theta}_2}, \boldsymbol{Y})\\
&= \lambda f(\boldsymbol{\theta}_1) + (1-\lambda) f(\boldsymbol{\theta}_2)
\end{aligned}$$

Also, $f(\boldsymbol{\theta})$ is L-Lipschitz in $\boldsymbol{\theta}$. In fact, it is straightforward to see that $\hat{\mathrm{R}}(\boldsymbol{h}_{\boldsymbol{\theta}}; X, \boldsymbol{Y})$ is also L-Lipschitz continuous in $\boldsymbol{\theta}$. For any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$, we have that:

$$\begin{aligned}\left|f(\boldsymbol{\theta}_1) - f(\boldsymbol{\theta}_2)\right| &\leq \max_{\boldsymbol{Y}\in\mathbb{Y}}\left|\hat{\mathrm{R}}(\boldsymbol{h}_{\boldsymbol{\theta}_1}; X, \boldsymbol{Y}) - \hat{\mathrm{R}}(\boldsymbol{h}_{\boldsymbol{\theta}_2}; X, \boldsymbol{Y})\right| \\ &\leq L\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2\end{aligned}$$

The subgradient of $f(\boldsymbol{\theta})$ in $\boldsymbol{\theta}$ is computed by using lemma 2.4.5.

Proof of Lemma 2.4.10. For each $i \in 1, \ldots, n$, we have that:

$$\frac{\partial}{\partial \theta_i}\ell(h_{\boldsymbol{\theta}}(x), \boldsymbol{e}) = 2\left(\boldsymbol{\varphi}_i(x)^T \cdot \boldsymbol{h}_{\boldsymbol{\theta}}(x) - \boldsymbol{\varphi}_i(x)^T \cdot \boldsymbol{e}\right)$$

Therefore, we can bound the norm of the gradient of ℓ as:

$$\begin{aligned}\left\|\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{h}_{\boldsymbol{\theta}}(x), \boldsymbol{e})\right\|_2 &= 2\sqrt{\sum_{i=1}^{n}\left(\boldsymbol{\varphi}_i(x)^T \cdot (\boldsymbol{h}_{\boldsymbol{\theta}}(x) - \boldsymbol{e})\right)^2} && \textsc{See Above} \\ &\leq 2\sqrt{\sum_{i=1}^{n}(1)^2} && \begin{array}{r}\left\|\boldsymbol{\varphi}_i(x)^T\right\|_1 = 1 \\ \left\|\boldsymbol{h}_{\boldsymbol{\theta}}(x) - \boldsymbol{e}\right\|_\infty \leq 1 \\ \textsc{Hölder's Inequality}\end{array} \\ &= 2\sqrt{n}\ .\end{aligned}$$

This implies that the function $\ell(\boldsymbol{h}_{\boldsymbol{\theta}}(x), \boldsymbol{e})$ is $2\sqrt{n}$-Lipschitz continuous with respect to $\boldsymbol{\theta}$.

Proof of Lemma 2.4.11. Without loss of generality, suppose that $e_i = 1$. We have that:

$$\ell(\boldsymbol{h}_{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{e}) = -\ln\left(\frac{\exp(\boldsymbol{w}_i^T \cdot \boldsymbol{x})}{\sum_{c=1}^{k}\exp(\boldsymbol{w}_c^T \cdot \boldsymbol{x})}\right)$$

It is easy to see that $\ell(\boldsymbol{h}_{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{e}) \geq 0$. By using Cauchy-Schwarz inequality, we have that:

$$\begin{aligned}\ell(\boldsymbol{h}_{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{e}) &= -\ln\left(\frac{\exp(\boldsymbol{w}_i^T \cdot \boldsymbol{x})}{\sum_{c=1}^{k}\exp(\boldsymbol{w}_c^T \cdot \boldsymbol{x})}\right) \\ &\leq -\ln\left(\frac{\exp(-B_w B_x)}{k\exp(B_w B_x)}\right) \\ &\leq 2B_w B_x + \ln k\end{aligned}$$

Proof of Lemma 2.4.12. For a fixed $(\boldsymbol{x}, \boldsymbol{e}) \in \mathcal{X} \times \mathcal{Y}$, consider the function $\omega(\boldsymbol{p}) : \mathbb{R}^k \to \mathcal{Y}^\diamond$ defined as

$$\omega(\boldsymbol{p}) \doteq -\sum_{c=1}^{k} e_c \cdot \ln(p_c) \ ,$$

and let

$$\boldsymbol{h}(\boldsymbol{\theta}) \doteq \left(\frac{\exp(\boldsymbol{w}_1^T \cdot \boldsymbol{x})}{\sum_{c=1}^k \exp(\boldsymbol{w}_c^T \cdot \boldsymbol{x})}, \dots, \frac{\exp(\boldsymbol{w}_k^T \cdot \boldsymbol{x})}{\sum_{c=1}^k \exp(\boldsymbol{w}_c^T \cdot \boldsymbol{x})} \right)^T \quad ,$$

where $\boldsymbol{\theta} = (\boldsymbol{w_1} \dots \boldsymbol{w_k})^T$, and observe that $\ell(\boldsymbol{h}_\theta(x), \boldsymbol{e}) = \omega \circ \boldsymbol{h}(\boldsymbol{\theta})$.

It is well known that ℓ is $L_\omega L_{\boldsymbol{h}}$-Lipschitz continuous with respect to $\boldsymbol{\theta}$, where L_ω and $L_{\boldsymbol{h}}$ are the Lipschitz constants respectively of ω and $\boldsymbol{h}$. It is also a known result that $L_\omega \leq 1$ (see for example Proposition 4 of Gao and Pavel [2018]).

We now want to compute $L_{\boldsymbol{h}}$. We will use the fact that $\max_{\boldsymbol{\theta} \in \Theta} ||J_{\boldsymbol{h}}(\boldsymbol{\theta})||_F \leq L_{\boldsymbol{h}}$, where $J_{\boldsymbol{h}}$ denotes the Jacobian matrix of $\boldsymbol{h}$ and $|| \cdot ||_F$ denotes the Frobenius norm.

For ease of notation, let $\boldsymbol{h}(\theta) = \boldsymbol{p} = (p_1, \dots, p_k)^T$. We have that for each $i \in 1, \dots, k$, it holds that

$$\begin{aligned} \frac{\partial [\boldsymbol{h}(\boldsymbol{\theta})]_i}{\partial \boldsymbol{w}_j} &= p_i p_j \boldsymbol{x} \qquad \text{for } j \neq i \\ \frac{\partial [\boldsymbol{h}(\boldsymbol{\theta})]_i}{\partial \boldsymbol{w}_i} &= (p_i - p_i^2) \boldsymbol{x} \end{aligned}$$

Therefore, we can bound the square of the Frobenius norm of the Jacobian matrix of $\boldsymbol{h}$ with:

$$\begin{aligned} ||J_{\boldsymbol{h}}(\boldsymbol{\theta})||_F^2 &= \sum_{i,j} \left|\left| \frac{\partial [\boldsymbol{h}(\boldsymbol{\theta})]_i}{\partial \boldsymbol{w}_j} \right|\right|_2^2 \\ &\leq ||\boldsymbol{x}||_2^2 \left(\sum_i [p_i(1-p_i)]^2 + \sum_{i \neq j} [p_i p_j]^2 \right) \\ &\leq ||x||_2^2 (k + k^2/2) \leq ||kx||_2^2 \end{aligned}$$

We can conclude that $\boldsymbol{h}$ is kB_x-Lipschitz continuous, and the statement follows.

Proof of Theorem 2.4.8 From Chapter 14 of [Mitzenmacher and Upfal, 2017], we know that:

$$\mathrm{R}(\boldsymbol{h}_{\hat{\boldsymbol{\theta}}}) \leq \hat{\mathrm{R}}(\boldsymbol{h}_{\hat{\boldsymbol{\theta}}}; X, \boldsymbol{Y}^*) + 2\hat{\mathfrak{R}}(\mathcal{L}; X, \boldsymbol{Y}^*) + \mathbf{O}\left(B \sqrt{\frac{\ln \frac{1}{\delta}}{m}} \right)$$

By definition of $f(\cdot)$, it holds that $\hat{\mathrm{R}}(\boldsymbol{h}_{\hat{\boldsymbol{\theta}}}; X, \boldsymbol{Y}^*) \leq f(\hat{\boldsymbol{\theta}})$. As $\hat{\boldsymbol{\theta}}$ is the optimal solution of (16), we have that $f(\hat{\boldsymbol{\theta}}) \leq f(\boldsymbol{\theta^*})$. Let $\boldsymbol{Y'} \doteq \mathrm{argmax}_{\boldsymbol{Y} \in \mathbb{Y}^\diamond} \hat{\mathrm{R}}(\boldsymbol{h}_{\boldsymbol{\theta}^*}; X, \boldsymbol{Y})$. It holds that:

$$
\begin{aligned}
f(\boldsymbol{\theta^*}) &= \hat{\mathrm{R}}(\boldsymbol{h}_{\boldsymbol{\theta}^*}; X, \boldsymbol{Y'}) \\
&= \hat{\mathrm{R}}(\boldsymbol{h}_{\boldsymbol{\theta}^*}; X, \boldsymbol{Y'}) + \hat{\mathrm{R}}(\boldsymbol{h}_{\boldsymbol{\theta}^*}; X, \boldsymbol{Y}^*) - \hat{\mathrm{R}}(\boldsymbol{h}_{\boldsymbol{\theta}^*}; X, \boldsymbol{Y}^*) \\
&= \hat{\mathrm{R}}(\boldsymbol{h}_{\boldsymbol{\theta}^*}; X, \boldsymbol{Y}^*) + \left| \hat{\mathrm{R}}(\boldsymbol{h}_{\boldsymbol{\theta}^*}; X, \boldsymbol{Y'}) - \hat{\mathrm{R}}(\boldsymbol{h}_{\boldsymbol{\theta}^*}; X, \boldsymbol{Y}^*) \right|
\end{aligned}
$$

By using the fact that ℓ is bounded, and the definition of diameter $D_{\mathbb{Y}^\diamond}$, we have that:

$$
\begin{aligned}
\left| \hat{\mathrm{R}}(\boldsymbol{h}_{\boldsymbol{\theta}^*}; X, Y') - \hat{\mathrm{R}}(\boldsymbol{h}_{\boldsymbol{\theta}^*}; X, Y^*) \right| &= \left| \frac{1}{m} \sum_{j=1}^{m} \sum_{c=1}^{k} \ell(\boldsymbol{h}(x_j), \boldsymbol{e}_c)(\boldsymbol{y}'_{jc} - \boldsymbol{y}^*_{jc}) \right| \\
&\leq B \frac{1}{m} \sum_{j=1}^{m} \sum_{c=1}^{k} \left| \boldsymbol{y}'_{jc} - \boldsymbol{y}^*_{jc} \right| \leq B D_{\mathbb{Y}^\diamond}
\end{aligned}
$$

To wrap it up, it results that

$$
\begin{aligned}
\mathrm{R}(\boldsymbol{h}_{\hat{\boldsymbol{\theta}}}) \leq & \hat{\mathrm{R}}(\boldsymbol{h}_{\boldsymbol{\theta}^*}; X, \boldsymbol{Y}^*) + B D_{\mathbb{Y}^\diamond} + 2\hat{\mathfrak{R}}(\mathcal{L}; X, \boldsymbol{Y}^*) + \mathbf{O}\left(B \sqrt{\frac{\ln \frac{1}{\delta}}{m}} \right) \\
& \leq \mathrm{R}(\boldsymbol{h}_{\boldsymbol{\theta}^*}) + B D_{\mathbb{Y}^\diamond} + 4\hat{\mathfrak{R}}(\mathcal{L}; X, \boldsymbol{Y}^*) + \mathbf{O}\left(B \sqrt{\frac{\ln \frac{1}{\delta}}{m}} \right) \\
& \leq \mathrm{R}(\boldsymbol{h}_{\boldsymbol{\theta}^*}) + B D_{\mathbb{Y}^\diamond} + \sup_{Y \in \boldsymbol{Y}^\diamond} 4\hat{\mathfrak{R}}(\mathcal{L}; X, \boldsymbol{Y}) + \mathbf{O}\left(B \sqrt{\frac{\ln \frac{1}{\delta}}{m}} \right) \quad .
\end{aligned}
$$

Proof of Theorem 2.4.9 This result is essentially a corollary of theorem 2.4.8. After bounding the minimax-risk, the next lines follow via elementary geometric inequalities and range assumptions.

Proof of Lemma 2.4.13 The proof is along the same lines of the proof of lemma 2.4.1, but we take a union bound with respect to all the nK intervals $\triangle_{i,c,\hat{c}}$ for $i \in 1, \ldots, n$, $c \in 1, \ldots, k$, and $\hat{c} = 1, \ldots, k_i$. Moreover, as for any $j \in 1, \ldots, m$, we have that $y_{j,c}[\boldsymbol{\varphi}_i(x_j)]_{\hat{c}} \leq 1$, we take $B = 1$ during the proof (as in lemma 2.4.1).

B.2 Additional Experimental Details

We provide further information specifying the experimental setup used to generate our figures.

B.2.1 Weak Supervision Sources

We first build the weak supervision sources on our two datasets as follows.

Animals with Attributes Each class is annotated with a binary vector of attributes. For each attribute, we train a binary classifier by finetuning a ResNet-18 using labeled data from the seen classes. When we consider a classification task between two unseen classes, we use as weak supervision sources the classifiers for the attributes which are different between the animals of these two unseen classes. We report the results of the 4 binary classification tasks which have the lowest majority vote accuracy. We chose these particular tasks to demonstrate the abilities of our methods on the tasks that have the least accurate and most highly correlated weak supervision sources.

DomainNet We sample 5 classes among the 25 classes of DomainNet with the largest number of data. For each domain, we use 60% of the available data for those classes to fine tune a pretrained ResNet-18 network. We perform this procedure on two disjoint samples of test classes to illustrate our results on two distinct multi class classification tasks.

In our experiments, we use the pretrained ResNet-18 from PyTorch. We finetune this ResNet-18 network following the approach described in [He et al., 2016], using cross-entropy loss.

B.2.2 Algorithm Hyperparameters

The subgradient method (algorithm 1) used to train AMCL-CC and AMCL-LR uses the following hyperparameters:

AMCL-CC We set $\delta = 0.1$ and build the constraints as in lemma 2.4.1. We use $\varepsilon = 0.1$ and define the step size h and the number of iterations T as in corollary 2.4.7.

AMCL-LR In this case, the loss function is bounded as in lemma 2.4.11. Since this value could be potentially very large, which in turn it would result in large intervals and number of iterations, we use the value $B = 1$ in the experiments. We set δ to 0.1 and build the constraints as in lemma 2.4.1. We do not bound the set of weights Θ: in the experiments, the norm of the weights of the multinomial logistic regression model has never diverged. We run the subgradient algorithm for $1/\varepsilon^2$ iterations with step size ε, with $\varepsilon = 0.1$

B.3 Additional Figures

B.3.1 Animals with Attributes

We provide the remaining figures for our experiments on the Animals with Attributes dataset. The last two binary classification tasks are *bat* versus *rat* and *horse* versus *giraffe*.

From figure B.1, we note that our methods show similar results as the figures displayed in the main body of the paper. AMCL-LR matches or outperforms all other methods on both tasks, over all ranges labeled data. AMCL-CC is within a few accuracy points of the other baselines and AMCL-LR on these tasks.

B.3.2 DomainNet

We provide the remaining figures for our experiments on the DomainNet dataset. We provide histograms when using the other 4 domains as the target task and also provide histograms for results on another of the samples of 5 classes. The first sample of classes as mentioned in the main body of the paper is {sea turtle, vase, whale, bird, violin }. The second sample is {tornado, trombone, submarine, feather, zebra }.

From figures B.2–B.6, we note that in most domains our methods perform better than or match all other approaches, namely in both samples of *clipart*, *quickdraw*, *painting*, and the second sample of *sketch*. Our methods achieve slightly lower accuracy than the best performing baseline on the *real* domain and on the second sample of the *infograph* domain, although they are not beaten by a single baseline in all of these tasks. We believe that the combination of our theoretical guarantees and that our methods achieve similar or sometimes better empirical performance captures the benefits of AMCL.

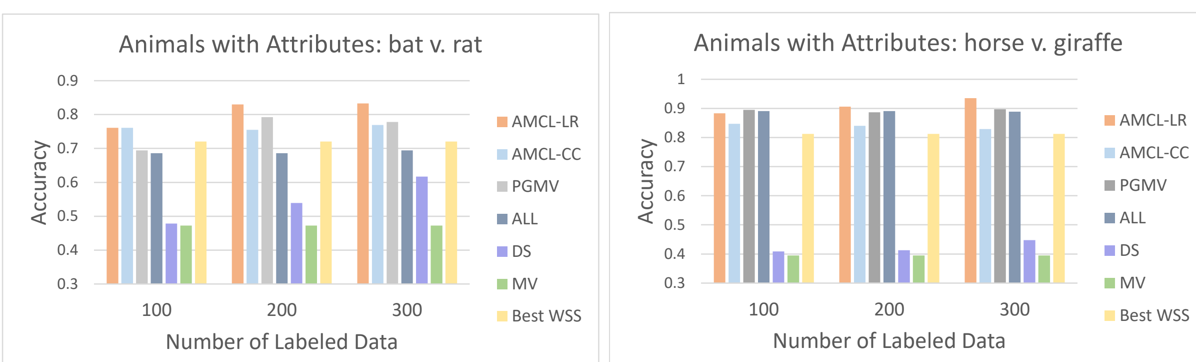


Figure B.1: Experimental results on the Animals with Attributes dataset for the binary classification tasks of *bat* versus *rat* (left) and *horse* versus *giraffe* (right) as we vary the amount of labeled data. Each method uses 347 unlabeled data points for bat versus rat and 1424 unlabeled data points for horse versus giraffe.

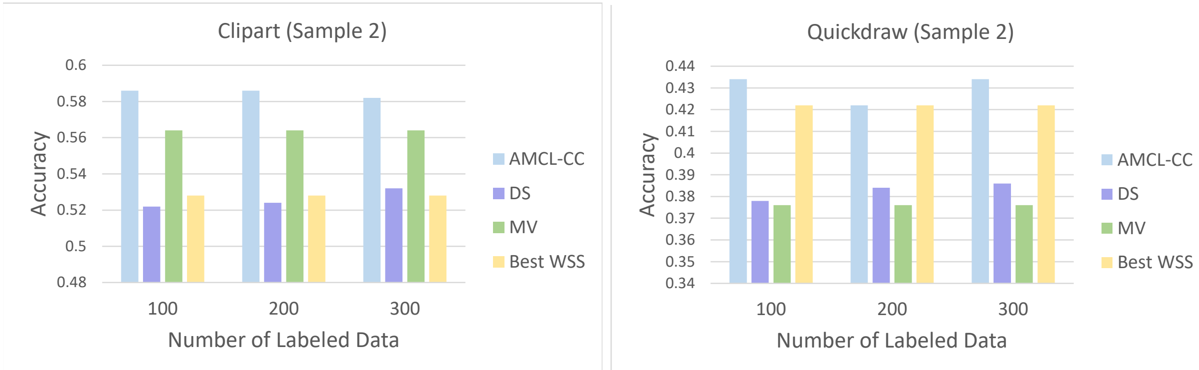


Figure B.2: Experimental results on the second sample of Domain Net for the *clipart* and *quickdraw* domains as we vary the amount of labeled data. Each method uses 500 unlabeled data points.

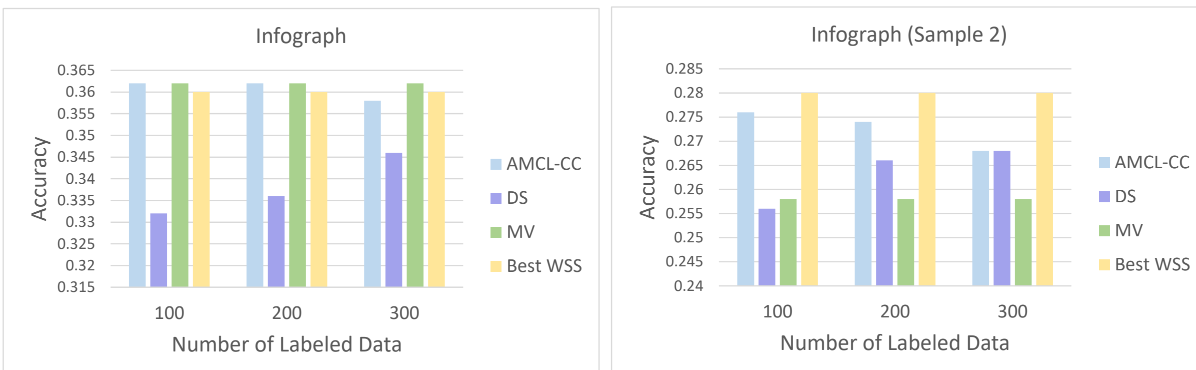


Figure B.3: Experimental results on both samples of Domain Net for the *infograph* domain as we vary the amount of labeled data. Each method uses 500 unlabeled data points.

B.4 Experiments on Synthetic Data

We run synthetic experiments to show that our method is robust with respect to the addition of correlated weak supervision sources. (Similar discussion has been done in Arachie and Huang [2019]).

We consider a multi class classification task over 5 classes, and 25 weak supervision sources $\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_{25}$. In this classification task, each item of the domain $\mathcal{X}$ has an unique true label. Given an item $x \in \mathcal{X}$, for $i \in 1, \ldots, 10$, the weak supervision source φ_i returns the correct label with probability $\frac{1}{2}$, and a random label with probability $\frac{1}{2}$. The output of the weak supervision source φ_i is independent to the output of the weak supervision sources $\boldsymbol{\varphi}_j$ for $j \in \{1, \ldots, 10\} \setminus \{i\}$. Therefore, the weak supervision source φ_i is correct with probability $\frac{1}{2}(1 + \frac{1}{k})$. For $i \in 11, \ldots, 25$, the weak supervision sources φ_i outputs the same result than the weak supervision source $\boldsymbol{\varphi}_1$. Note that the weak supervision sources $\boldsymbol{\varphi}_{11}, \ldots, \boldsymbol{\varphi}_{25}$ do not provide any additional information with respect to the target classification task, as they add redundant constraints to the set of feasible labelings $\mathbb{Y}^{\diamond}$. The majority vote of the weak supervision sources $\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_{25}$ is highly affected

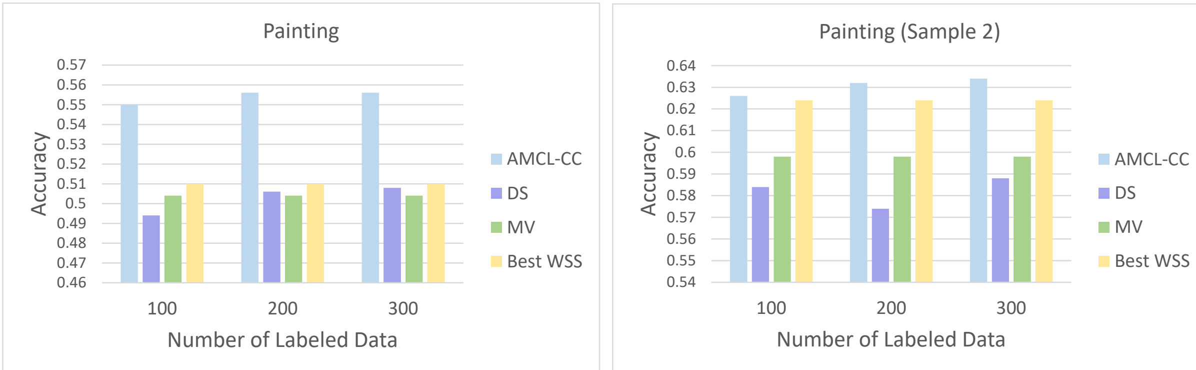


Figure B.4: Experimental results on both samples of Domain Net for the *painting* domain as we vary the amount of labeled data. Each method uses 500 unlabeled data points.

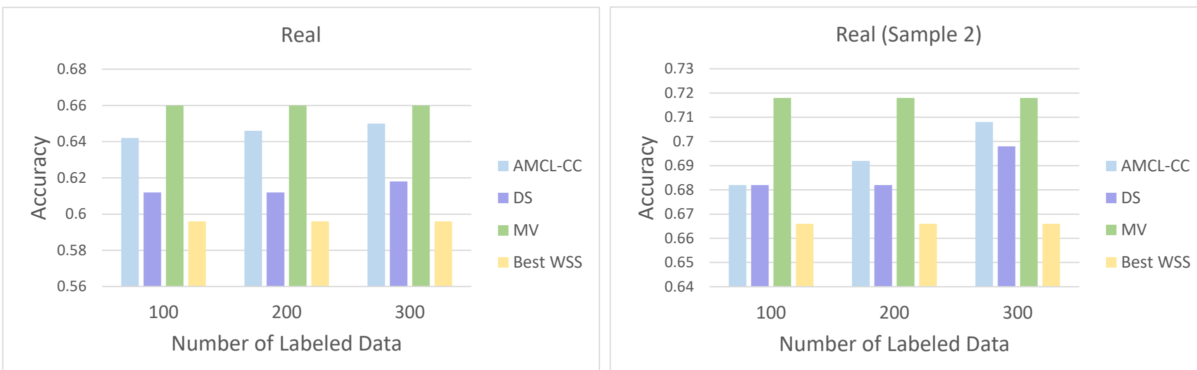
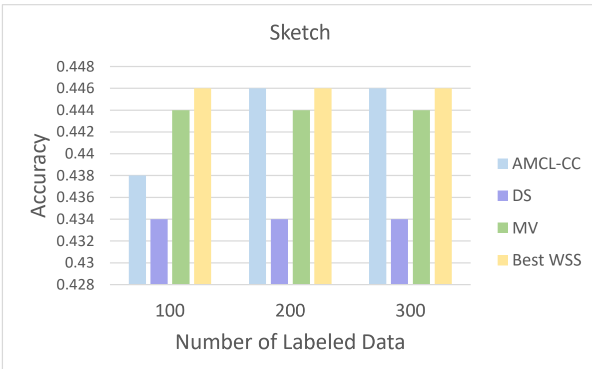


Figure B.5: Experimental results on both samples of Domain Net for the *real* domain as we vary the amount of labeled data. Each method uses 500 unlabeled data points.

by these dependencies, and it is very likely to provide the same answer as φ_1, which is only $\frac{1}{2}(1+\frac{1}{k})$ accurate on average. On the other hand, the majority vote of the weak supervision sources $\varphi_1, \ldots, \varphi_{10}$ would improve upon the individual accuracy of the weak supervision sources, as their output is independent.

We use 500 unlabeled examples, run experiments varying the amount of labeled data, and show that our method AMCL-CC is robust against those dependencies. For the sake of these experiments, as we want to use very small amount of labeled data, we set $\gamma = 0$ when building the constraints for $\mathbb{Y}^\diamond$ as in lemma 2.4.1. The experimental results are reported in appendix B.4. The table shows that AMCL-CC is robust with respect to dependencies among weak supervision sources, whereas majority vote is greatly affected by them. In fact, in this case the majority vote does not improve upon the individual accuracy of the weak supervision sources, which is on average $\frac{1}{2}(1+\frac{1}{k}) = \frac{3}{5}$.



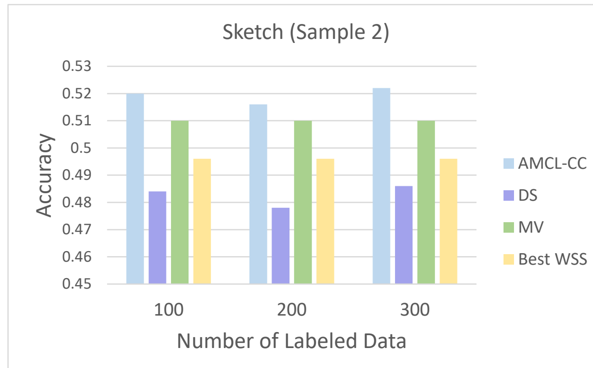


Figure B.6: Experimental results on both samples of Domain Net for the *sketch* domain as we vary the amount of labeled data. Each method uses 500 unlabeled data points.

Table B.1: We report the experimental results on the synthetic dataset. We report the accuracy obtained by our method AMCL-CC and the majority vote, when varying the amount of labeled examples (we report the average accuracy over 3 distinct runs).

Labeled Examples	AMCL-CC	Majority Vote
100	0.902	0.595
50	0.828	0.602
25	0.819	0.598

Appendix C

Supplementary Material for Chapter 3

C.1 Welfare and Malfare

We now show theorem 3.3.7.

Theorem 3.3.7 (Properties of the Power-Mean)**.** Suppose $\mathcal{S}, \boldsymbol{\varepsilon}$ are sentiment values $\Omega \to \mathbb{R}_{0+}$, and $\boldsymbol{w}$ is a probability measure over Ω. Then

1. Monotonicity: $\mathrm{M}_p(\mathcal{S}; \boldsymbol{w})$ is weakly-monotonically-increasing in p, and strictly if $\mathcal{S}$ attains distinct $a, b \in \mathbb{R}$ with nonnegligible probability.
2. Subadditivity: $\forall p \geq 1 : \mathrm{M}_p(\mathcal{S} + \boldsymbol{\varepsilon}; \boldsymbol{w}) \leq \mathrm{M}_p(\mathcal{S}; \boldsymbol{w}) + \mathrm{M}_p(\boldsymbol{\varepsilon}; \boldsymbol{w})$.
3. Contraction: $\forall p \geq 1 : \left|\mathrm{M}_p(\mathcal{S}; \boldsymbol{w}) - \mathrm{M}_p(\mathcal{S}'; \boldsymbol{w})\right| \leq \mathrm{M}_p(\left|\mathcal{S} - \mathcal{S}'\right|; \boldsymbol{w}) \leq \left\|\mathcal{S} - \mathcal{S}'\right\|_\infty$.
4. Curvature: $\mathrm{M}_p(\mathcal{S}; \boldsymbol{w})$ is concave in $\mathcal{S}$ for $p \in [-\infty, 1]$ and convex for $p \in [1, \infty]$.

Proof. We omit proof of item 1, as this is a standard property of power-means (generally termed the *power mean inequality* [Bullen, 2013, chapter 3]).

We first show item 2. By the triangle inequality (for $p \geq 1$), we have

$$\mathrm{M}_p(\mathcal{S} + \boldsymbol{\varepsilon}; \boldsymbol{w}) \leq \mathrm{M}_p(\mathcal{S}; \boldsymbol{w}) + \mathrm{M}_p(\boldsymbol{\varepsilon}; \boldsymbol{w}) \ .$$

We now show item 3 First take $\boldsymbol{\varepsilon} \doteq \mathcal{S} - \mathcal{S}'$. Now consider

$$
\begin{aligned}
\mathrm{M}_p(\mathcal{S}; \boldsymbol{w}) &= \mathrm{M}_p(\mathcal{S}' + \boldsymbol{\varepsilon}; \boldsymbol{w}) && \textsc{Definition of } \boldsymbol{\varepsilon} \\
&\leq \mathrm{M}_p(\mathcal{S}' + \boldsymbol{\varepsilon}_+; \boldsymbol{w}) && \textsc{Monotonicity} \\
&\leq \mathrm{M}_p(\mathcal{S}'; \boldsymbol{w}) + \mathrm{M}_p(\boldsymbol{\varepsilon}_+; \boldsymbol{w}) && \textsc{item } 2 \\
&\leq \mathrm{M}_p(\mathcal{S}'; \boldsymbol{w}) + \mathrm{M}_p(\left|\mathcal{S} - \mathcal{S}'\right|; \boldsymbol{w}) \ . && \textsc{Monotonicity}
\end{aligned}
$$

By symmetry, we have $\mathrm{M}_p(\mathcal{S}', \boldsymbol{w}) \leq \mathrm{M}_p(\mathcal{S}, \boldsymbol{w}) + \mathrm{M}_p(\left|\mathcal{S} - \mathcal{S}'\right|; \boldsymbol{w})$, which implies the result.

We now show item 4. First note the special cases of $p \in \pm\infty$ follow by convexity of the maximum ($p = \infty$) and concavity of the minimum ($p = -\infty$).

Now, note that for $p \geq 1$, by concavity of $\sqrt[p]{\cdot}$, Jensen's inequality gives us

$$
\mathrm{M}_1(\mathcal{S}; \boldsymbol{w}) = \mathop{\mathbb{E}}_{\omega \sim \boldsymbol{w}}[\mathcal{S}(\omega)] = \underbrace{\mathop{\mathbb{E}}_{\omega \sim \boldsymbol{w}}[\sqrt[p]{\mathcal{S}^p(\omega)}] \leq \sqrt[p]{\mathop{\mathbb{E}}_{\omega \sim \boldsymbol{w}}[\mathcal{S}^p(\omega)]}}_{\textsc{Definition of Convexity}} = \mathrm{M}_p(\mathcal{S}; \boldsymbol{w}) \ ,
$$

i.e., convexity, and similarly, for $p \leq 1$, $p \neq 0$, we have by convexity of $\sqrt[p]{\cdot}$, we have

$$
\mathrm{M}_1(\mathcal{S}; \boldsymbol{w}) = \mathop{\mathbb{E}}_{\omega \sim \boldsymbol{w}}[\mathcal{S}(\omega)] = \underbrace{\mathop{\mathbb{E}}_{\omega \sim \boldsymbol{w}}[\sqrt[p]{\mathcal{S}^p(\omega)}] \geq \sqrt[p]{\mathop{\mathbb{E}}_{\omega \sim \boldsymbol{w}}[\mathcal{S}^p(\omega)]}}_{\textsc{Definition of Concavity}} = \mathrm{M}_p(\mathcal{S}; \boldsymbol{w}) \ .
$$

Similar reasoning, now by convexity of $\ln(\cdot)$, shows the case of $p = 0$. □

We now show theorem 3.3.8.

Theorem 3.3.8 (Population Mean Properties)**.** Suppose population-mean function $\mathrm{M}(\mathcal{S}; \boldsymbol{w})$. If $\mathrm{M}(\cdot;\cdot)$ satisfies (subsets of) the population-mean axioms (see definition 3.3.4), we have that $\mathrm{M}(\cdot;\cdot)$ exhibits the following properties. For each, assume arbitrary sentiment-value function $\mathcal{S} : \Omega \to \mathbb{R}_{0+}$ and weights measure $\boldsymbol{w}$ over Ω. The following then hold.

1. *Identity*: Axioms 6 & 7 imply $\mathrm{M}(\omega \mapsto \alpha; \boldsymbol{w}) = \alpha$.
2. Axioms 1-5 imply $\exists p \in \mathbb{R}$, strictly-monotonically-increasing continuous $F : \mathbb{R} \to \mathbb{R}_{0+}$ s.t.

$$
\mathrm{M}(\mathcal{S}; \boldsymbol{w}) = F\left(\int_{\boldsymbol{w}} f_p(\mathcal{S}(\omega)) \, \mathrm{d}(\omega)\right) = F\left(\mathop{\mathbb{E}}_{\omega \sim \boldsymbol{w}}\left[f_p(\mathcal{S}(\omega))\right]\right) \ , \ \text{with} \begin{cases} p = 0 & f_0(x) \doteq \ln(x) \\ p \neq 0 & f_p(x) \doteq \mathrm{sgn}(p) x^p \end{cases} .
$$

3. Axioms 1-7 imply $F(x) = f_p^{-1}(x)$, thus $\mathrm{M}(\mathcal{S}; \boldsymbol{w}) = \mathrm{M}_p(\mathcal{S}; \boldsymbol{w})$.

4. Axioms 1-5 and 8 imply $p \in (-\infty, 1]$.

5. Axioms 1-5 and 9 imply $p \in [1, \infty)$.

Proof. Item 1 is an immediate consequence of axioms 6 & 7 (multiplicative linearity and unit scale).

We now note that item 2 is the celebrated Debreu-Gorman theorem [Debreu, 1959, Gorman, 1968], extended by continuity and measurability of $\mathcal{S}$ to the weighted case.

We now show item 3. This result is essentially a corollary of item 2, hence the dependence on axioms 1-4. Suppose $\mathcal{S}(\cdot) = 1$. By item 1, for all $p \neq 0$, we have

$$\alpha = \alpha \mathrm{M}(\mathcal{S}; \boldsymbol{w}) = \mathrm{M}(\alpha\mathcal{S}; \boldsymbol{w}) = F\left(\mathop{\mathbb{E}}_{\omega \sim \boldsymbol{w}}\big[f_p(\alpha\mathcal{S}(\omega))\big]\right) = F\left(\mathop{\mathbb{E}}_{\omega \sim \boldsymbol{w}}\big[f_p(\alpha)\big]\right) = F\big(\mathrm{sgn}(p)\alpha^p\big) \enspace .$$

From here, we have $\alpha = F\big(\mathrm{sgn}(p)\alpha^p\big)$, thus $F^{-1}(u) = \mathrm{sgn}(p)u^p$, and consequently, $F(v) = \sqrt[p]{\mathrm{sgn}(p)v}$.

Taking $p = 0$ gets us

$$\alpha = \alpha \mathrm{M}(\mathcal{S}; \boldsymbol{w}) = F\left(\mathop{\mathbb{E}}_{\omega \sim \boldsymbol{w}}\big[\ln(\alpha\mathcal{S}(\omega))\big]\right) = F(\ln a) \enspace ,$$

from which it is clear that $F^{-1}(u) = \ln(u) \Rightarrow F(v) = \exp(v)$.

For all values of $p \in \mathbb{R}$, substituting the values of f_p and $F(\cdot)$ into item 2 yields $\mathrm{M}(\mathcal{S}; \boldsymbol{w}) = \mathrm{M}_p(\mathcal{S}; \boldsymbol{w})$ by definition.

We now show 4 and 5. These properties follow directly from 2, wherein f_p are defined, and Jensen's inequality. □

We now show corollary 3.4.2.

Corollary 3.4.2 (Statistical Estimation with Hoeffding and Bennett Bounds)**.** Suppose fair power-mean malfare $\mathrm{W}(\cdot; \cdot)$ (i.e., $p \geq 1$), *loss function* $\ell : \mathcal{X} \to [0, r]$, $\mathcal{S} \in [0, r]^g$ s.t. $\mathcal{S}_i = \mathbb{E}_{\mathcal{D}_i}[\ell]$, samples $\boldsymbol{x}_i \sim \mathcal{D}_i^m$, and take $\hat{\mathcal{S}}_i \doteq \frac{1}{m}\sum_{j=1}^m \ell(\boldsymbol{x}_{i,j})$. Then with probability at least $1 - \delta$ over choice of $\boldsymbol{x}$,

$$\left|\mathrm{M}_p(\mathcal{S}; \boldsymbol{w}) - \mathrm{M}_p(\hat{\mathcal{S}}; \boldsymbol{w})\right| \leq r\sqrt{\frac{\ln \frac{2g}{\delta}}{2m}} \enspace .$$

Alternatively, again with probability at least $1-\delta$ over choice of $\boldsymbol{x}$, we have

$$\left|\mathrm{M}_p(\mathcal{S};\boldsymbol{w}) - \mathrm{M}_p(\hat{\mathcal{S}};\boldsymbol{w})\right| \leq \frac{r\ln\frac{2g}{\delta}}{3m} + \max_{i\in 1,\dots,g}\sqrt{\frac{2\,\mathbb{V}_{\mathcal{D}_i}[\ell]\ln\frac{2g}{\delta}}{m}} \enspace .$$

Proof. This result is a corollary of lemma 3.4.1, applied to $\boldsymbol{\varepsilon}$, where we note that for $p \geq 1$, by theorem 3.3.7 item 3 (contraction) it holds that

$$\mathrm{M}_p(\hat{\mathcal{S}}+\boldsymbol{\varepsilon};\boldsymbol{w}) \leq \mathrm{M}_p(\hat{\mathcal{S}};\boldsymbol{w}) + \|\boldsymbol{\varepsilon}\|_\infty \ \& \ \mathrm{M}_p(\boldsymbol{0}\vee(\hat{\mathcal{S}}-\boldsymbol{\varepsilon});\boldsymbol{w}) \leq \mathrm{M}_p(\hat{\mathcal{S}};\boldsymbol{w}) - \|\boldsymbol{\varepsilon}\|_\infty \enspace .$$

Now, for the first bound, note that we take $\boldsymbol{\varepsilon}_i \doteq r\sqrt{\frac{\ln\frac{2g}{\delta}}{2m}}$, and by Hoeffding's inequality and the union bound, for $\Omega = \{1,\dots,n\}$, we have $\forall\omega: \mathcal{S}'(\omega) - \boldsymbol{\varepsilon}(\omega) \leq \mathcal{S}(\omega) \leq \mathcal{S}'(\omega) + \boldsymbol{\varepsilon}(\omega)$ with probability at least $1-\delta$. The result then follows via the *power-mean contraction* (theorem 3.3.7 item 3) property.

Similarly, for the second bound, note that we take $\boldsymbol{\varepsilon}_i \doteq \frac{r\ln\frac{2g}{\delta}}{3m} + \sqrt{\frac{2\,\mathbb{V}_{\mathcal{D}_i}[\ell]\ln\frac{2g}{\delta}}{m}}$, which this time follows via Bennett's inequality and the union bound. Now, we again apply lemma 3.4.1, noting that $\mathrm{M}(\boldsymbol{\varepsilon}) \leq \mathrm{M}_\infty(\boldsymbol{\varepsilon}) = \|\boldsymbol{\varepsilon}\|_\infty$ (by *power-mean monotonicity*, theorem 3.3.7 item 1), and the rest follows as in the Hoeffding case. $\square$

Appendix D

Supplementary Material for Chapter 4

D.1 Pseudocode and Ewo$k$s Algorithm Details

Algorithm 6 presents pseudocode for the Ewo$k$s. In particular, we describe the *optimization* algorithm CodecSelection(...), where the optimal $\hat{\boldsymbol{c}}_{\mathrm{W}} \in \binom{\mathcal{C}}{k}$ is selected, as well as the routines StoreMedia(...), which adds a piece of media to the database, and QueryMedia(...), which serves a user the best-available encoding for their loss function.

Algorithm 6 Pseudocode for Ewo$k$s Operations

procedure CodecSelection($\mathcal{C}, k, \mathcal{L}, \mathrm{W}, \boldsymbol{x}$) $\mapsto \hat{\boldsymbol{c}}_{\mathrm{W}}$
  Input: Codec family $\mathcal{C}$, selection size k, loss distribution $\mathcal{L}$, welfare concept W, sample $\boldsymbol{x} \in \mathcal{X}^m$
  Output: Ewo$k$s-optimal $\boldsymbol{c} \in \binom{\mathcal{C}}{k}$
  Effect Sets global codec-selection $\hat{\boldsymbol{c}}_{\mathrm{W}}$ (for use with EncodeMedia)
  for all $i \in 1, \ldots, |\mathcal{C}|$ **do** ▷ Precompute attribute values
    for all $j \in 1, \ldots, m$ **do**
      $\boldsymbol{A}_{i,j,:} \leftarrow \mathcal{C}_i(\boldsymbol{x}_j)$ ▷ $\boldsymbol{A} \in \mathbb{R}_+^{|\mathcal{C}| \times m \times a}$
    end for
  end for
  $\boldsymbol{I} \leftarrow \{i \in 1, \ldots, |\mathcal{C}| \mid \nexists i' < i \text{ s.t. } \boldsymbol{A}_{i,:,:} = \boldsymbol{A}_{i',:,:}\}$ ▷ Prune Coincident Codecs
  if Prune Monotonic **then** ▷ Prune monotonically-dominated codecs (Optional)
    $\boldsymbol{I} \leftarrow \left\{ i \in \boldsymbol{I} \,\middle|\, \nexists i' \text{ s.t. } \boldsymbol{A}_{i',:,:} \preceq \boldsymbol{A}_{i,:,:} \right\}$
  end if
  if Prune Generic **then** ▷ Prune $\mathcal{L}$-dominated codecs (Optional)
    $\boldsymbol{I} \leftarrow \left\{ i \in \boldsymbol{I} \,\middle|\, \nexists i' \neq i \text{ s.t. } \sup_{\ell \in \mathcal{L}} \max_{j \in 1, \ldots, m} \ell(\boldsymbol{A}_{i',j,:}) - \ell(\boldsymbol{A}_{i,j,:}) < 0 \right\}$
  end if
  $\boldsymbol{I}' \leftarrow \operatorname*{argmin}_{\boldsymbol{I}' \in \binom{\boldsymbol{I}}{k}} \hat{\mathrm{W}}(\{i(j) \mapsto \boldsymbol{A}_{i,j,1:a} \mid i \in \boldsymbol{I}'\}; \mathcal{L}, 1, \ldots, m)$ ▷ Combinatorial Optimization
  $\hat{\boldsymbol{c}}_{\mathrm{W}} \leftarrow \{\mathcal{C}_i \mid i \in \boldsymbol{I}'\}$
  return $\hat{\boldsymbol{c}}_{\mathrm{W}}$
end procedure
procedure EncodeMedia(x, ID) $\mapsto \emptyset$
  Input: Media x, media ID ID.
  Effect Stores each encoding $\forall c \in \hat{\boldsymbol{c}}_{\mathrm{W}} : c(x)$ of media x at index ID.
  for $c \in \boldsymbol{c}$ **do**
    $\mathrm{Store}_{\mathrm{ID},c} \leftarrow c(x)$ ▷ Store encoded media and attributes
  end for
end procedure
procedure QueryMedia(ℓ, ID) $\mapsto$ (Codec $\times$ Encoding)
  Input: Loss function ℓ, some media ID ID that has previously been passed to EncodeMedia($\cdot, \cdot$)
  Output: Optimal codec $c \in \hat{\boldsymbol{c}}_{\mathrm{W}}$ and *encoding* $c(x)$, for x the media referenced by ID
  $c \leftarrow \operatorname*{argmin}_{c' \in \boldsymbol{c}} \ell(\mathrm{Store}_{\mathrm{ID},c'})$
  return $(c, \mathrm{Store}_{\mathrm{ID},c})$ ▷ Return encoded media
end procedure

D.2 Supplementary Experiments and Methods

Here we present supplementary detail on our datasets, additional experiments in the video domain, and additional high-detail plots and extensions of experiments presented in the paper body.

D.2.1 Experimental Setup

As discussed in Section 4.3, we performed encoding experiments on audio data extracted from audio-books and music. In this section, we present additional detail on our audio and video datasets, as well as encoding parameters.

For each of our *music*, *speech*, and *video* datasets, we give the *sample count*, *sample length*, and detailed *format information* in table D.1. Each music dataset was ripped from CD, and the `au` dataset was downloaded from `librivox.org` [McGuire]. Each of the files in the original datasets are split into *sample count* contiguous nonoverlapping *sample length* segments, measured in *seconds* (s), with a possible shorter final "leftover segment" per file. In particular, we decompose the *information rate*, measured in *bit rate* bits per second (b/s), or, perhaps more conveniently, base-1000-kilobytes per second (KB/s; 1KB = 8000b), into *sample rate*, which gives the number of discrete samples per second, measured in *cycles / second* (Hertz or hz), or equivalently in *frames / second* (fps), and *sample depth*, which is the number of bits per sample. Sample depth is further decomposed by domain: audio is either *stereo*, which refers to 2-channel (left and right) sound (thus *doubling* the bitrate), or *monaural*, which contains a single channel, generally heard in both ears; similarly, *video samples* are *frames*, or *images*, each split into a resolution $x \times y$ array of *pixels* (px), each of which has 3 color values (**RGB**). Both audio waveform samples and color intensity values are represented as linearly-scaled unsigned fixed-width integers.

All audio was initially in the uncompressed `wav` format (see table D.1), and all `mp3` encoding was performed

Table D.1: Sample count, length, rate, and depth of all music, speech, and video datasets. All numbers given are for the original (uncompressed) data formats.

Modality	**Dataset**	**Count**	**Length**	**Sample Rate**	**Sample Depth**	**Information Rate**
Music	`de`	8741	2s	44100hz	16 bit Stereo (2ch)	1411200 b/s 172 KB/s
	`ex`	9443				
	`lz`	13163				
	Total	31347				
Speech	`au`	16530	10s	44100hz	16 bit Monaural (1ch)	705600 b/s 86 KB/s
Video	`video`	1033	1s	30fps	256 × 144 px 3 col (**RGB**) 8 b/col/px	26542080 b/s 3240 KB/s

with the LAME [Hegemann et al., 2017] encoder. LAME is a recursive acronym for *Lame Ain't an* `Mp3` *Encoder*, and is known for its advanced psychoacoustic modelling, flexibility, and high perceptual-quality `mp3` encodings. Details on audio encoding are given in table 4.1. The video dataset was ripped from a youtube video (compressed), but we treated it as though the original rip were the ground-truth (uncompressed) video, and measured deviation from the original as distortion, though we compute *compression ratio* as a ratio of the original compressed video size. Our experiments utilize 22 `mp4` codec variants, each a *constant rate factor* (CRF) `mp4`, encoded by FFMPEG, with *quality* parameter ranging from $\{30, \ldots, 51\}$.

D.2.2 Video Experiments

To demonstrate the flexibility of our approach, we also apply EWOkS to the selection of video codecs. Perceptual quality metrics are an area of active research, particularly in the visual domain, and in our video experiments, we characterize distortion with the *peak signal to noise ratio* (PSNR) Erfurt et al. [2019] metric. Figure D.1 visualizes the Pareto-optimal frontier for 22 `mp4` video compression codecs, which showcases similar trade-offs as those demonstrated for audio compression in figure 4.1. We use `mp4` codecs with integer-valued *Constant Rate Factor* (quality parameter) between 30 and 51 with the FFMPEG compression software. The mean performance of 22 codecs is plotted, along with Pareto optimality curves for all $k \in \{1, 2, |\mathcal{C}|\}$.

Because this experiment is a proof of concept, our dataset is limited in size and complexity. Our set of media consists of a 17:13 minutes:seconds long, 144p 30fps video *youtube video* (`https://www.youtube.com/watch?v=31OPF8ctRTM`) sliced into 1033 distinct 1-second segments. We removed the audio from the video, so the we need only compress 1-second intervals of frames of images.

We measure divergence in terms of the *peak signal-to-noise ratio* (PSNR), a common method for measuring how the *fidelity* of a compressed image (i.e., how much said image resembles its original form) [Gonzalez and Woods, 1992]. PSNR compares two images by taking the *logarithm* of the *maximum squared error* over the *mean squared error*. This scaling ensures that subtle and drastic changes to an image can be compared at the same order of magnitude with the same metric. This metric can also be applied to video compression by computing frame-by-frame differences between pixels. Because we need to restrict our attributes to the interval $[0, 1]$, and want a measure of *distortion* rather than *quality*, we take the *PSNR Divergence* to be $1 - \frac{\text{PSNR}}{\max\ \text{PSNR}}$, where $\max \text{PSNR}$ refers to the largest PSNR divergence over all our codecs for the given fixed video segment.

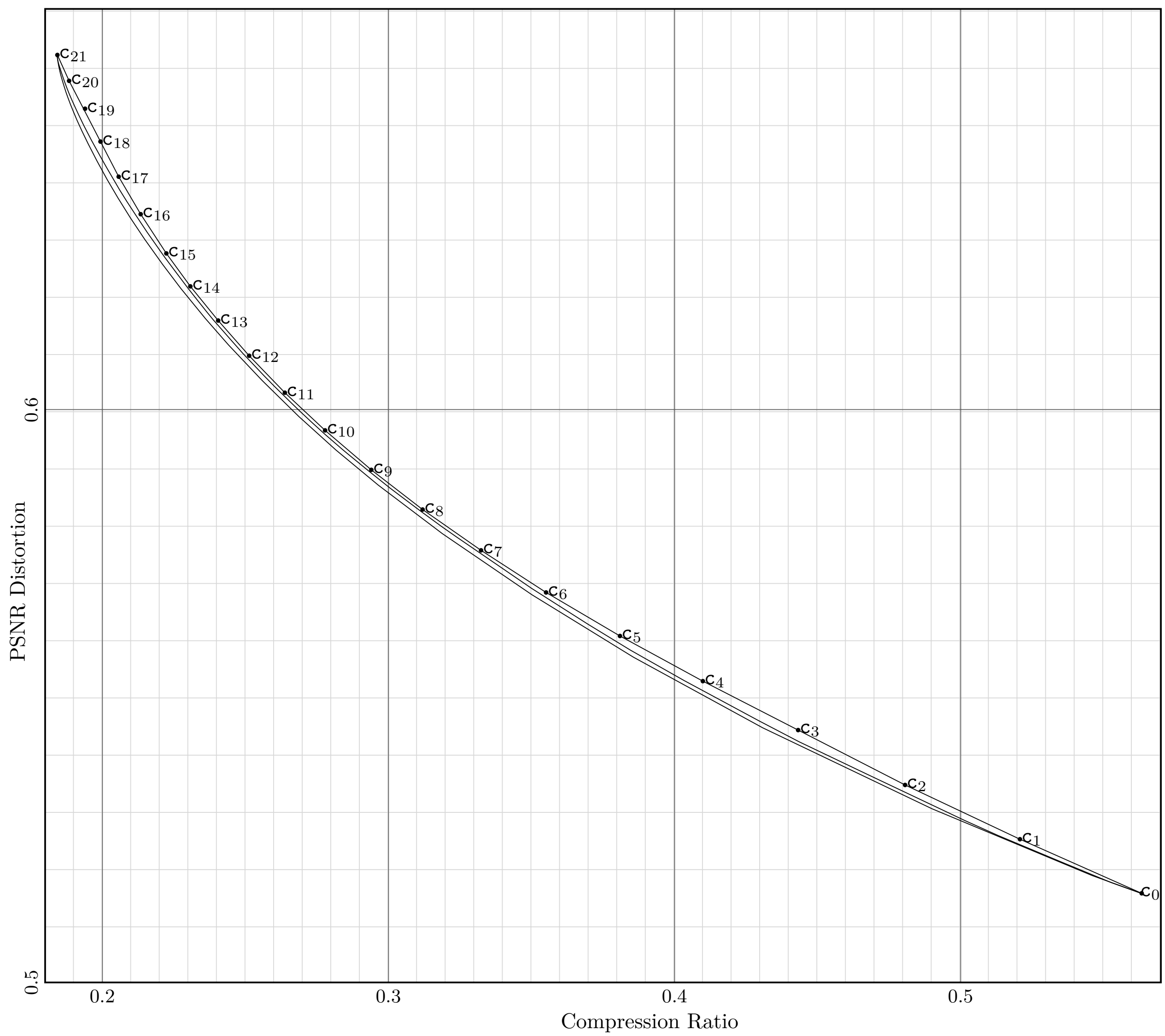


Figure D.1: Empirical Pareto-optimal frontiers for a selection of video codecs, for $k \in \{1, 2, |\mathcal{C}|\}$. Divergence measure *PSNR* (y-axis) versus *compression ratio* (x-axis) on the `video` dataset.

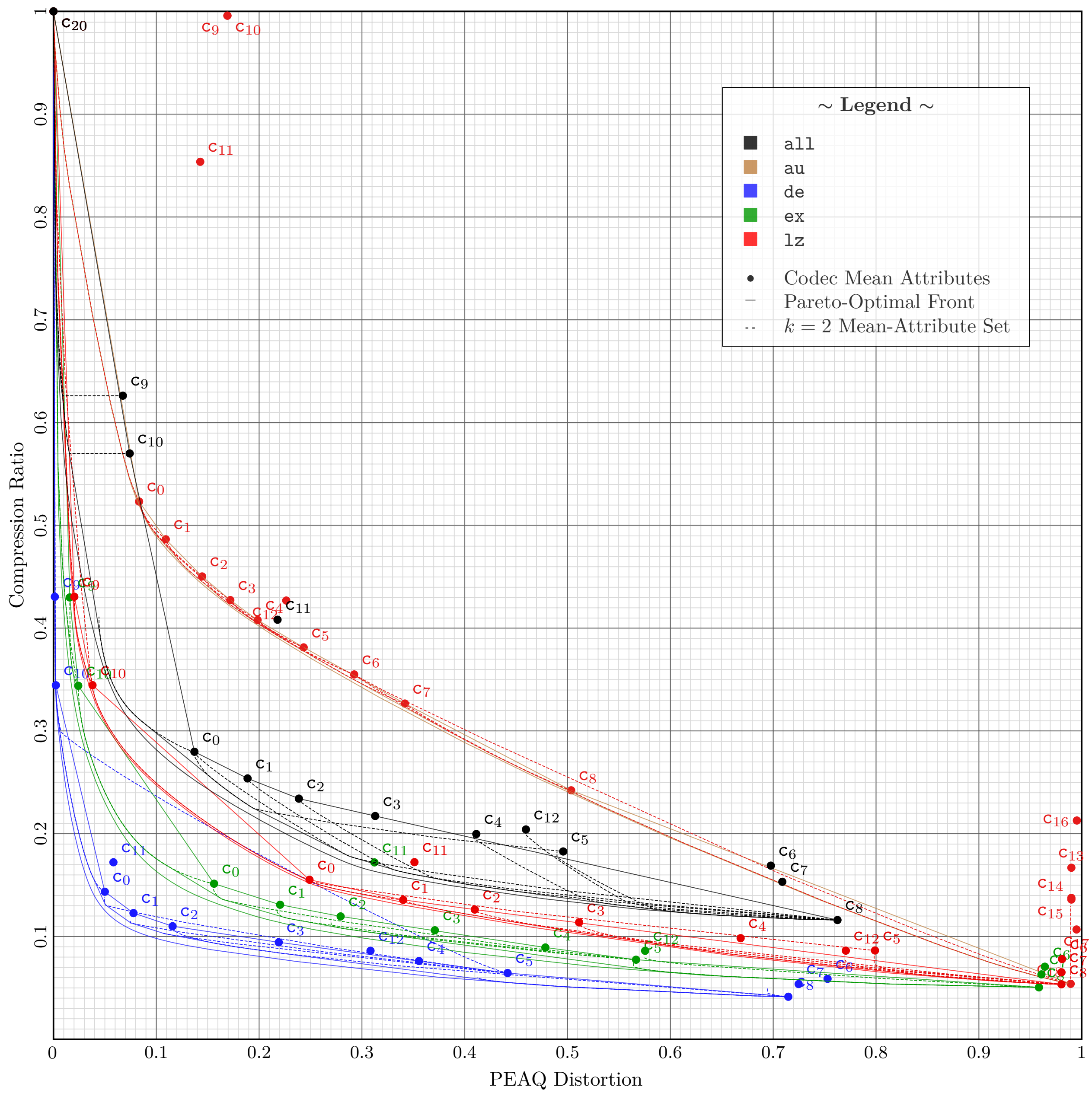


Figure D.2: Pareto optimality curves for $k \in \{1, 2, |\mathcal{C}|\}$, for all audio datasets. The figure matches figure 4.1, except here, the `music` dataset is split into its constituent datasets (`de`, `ex`, and `lz`). All codecs, Pareto-optimality curves, and Pareto-optimal pair mean-attribute-sets are color-coded by dataset (see legend in top right), and again k always increases from top-left to bottom-right.

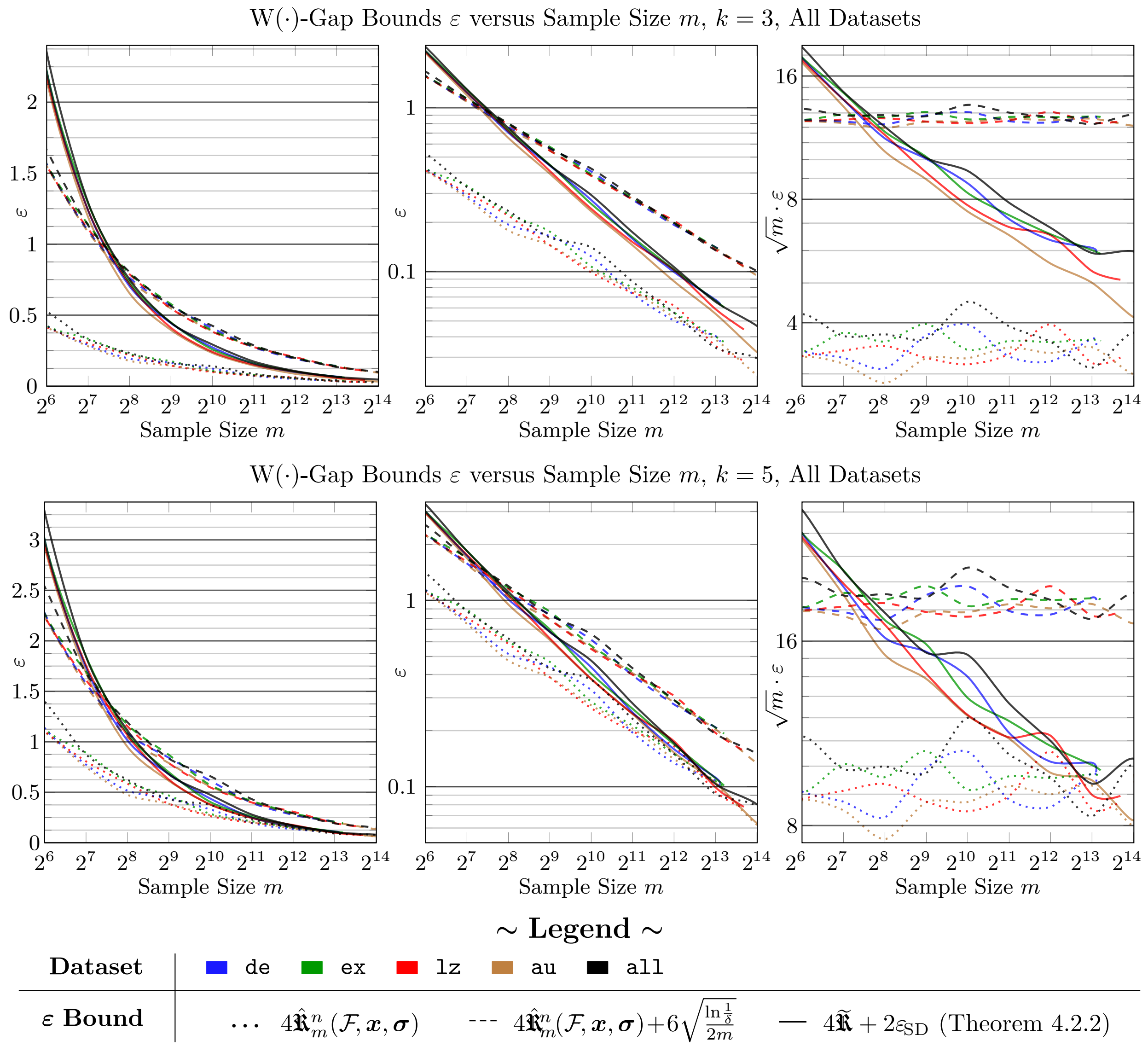


Figure D.3: Semilog, log-log, and normalized log-log plots of Ewo*k*s welfare-optimality gap bounds ε (x-axis) bounds versus sample size m (y-axis), in expectation with $4\hat{\hat{\mathfrak{R}}}^n_m(\mathcal{F}, \boldsymbol{x}, \boldsymbol{\sigma})$, and w.h.p. with McDiarmid and theorem 4.2.2. The first row presents results for $k = 3$, and the second for $k = 5$. The left, center, and right plots in each row contain the same information, and vary only in presentation. The left column contains *semilog* (log-x, natural-y axes) plots, and the center column contains *log-log* plots. The right column also contains *log-log* plots, but in this plot, all bounds are scaled by $\sqrt{m}$, to visually characterize $\boldsymbol{\Theta}\sqrt{1/m}$ rates as straight lines.

D.2.3 Supplementary Audio Experiments

Here we show additional plots and experiments on the audio datasets discussed in section 4.3.

In particular, we first complement the Pareto-optimality visualization of figure 4.1 with figure D.2, which shows the same in much greater detail (matching the inset, with mean-attribute sets of all codec pairs).

We also give in figure D.3 plots supplementing figure 4.3, now with semilog, log-log, and $\sqrt{m}$-scaled plots of the same data. We additionally present plots for $k = 3$ and $k = 5$ (only $k = 3$ is given in the paper body).

D.3 Proof Compendium

We now prove our main results stated in the paper body.

D.3.1 Proof of Theorem 4.2.2

The result is shown via a chain of union bounds on Rademacher averages and variances. We now show several lemmas in service of the main result.

Lemma D.3.1 (Empirical Wimpy Variance Expectation)**.** Suppose $\boldsymbol{x} \sim \mathcal{D}^m$, and take

$$v \doteq \sup_{f \in \mathcal{F}} \mathop{\mathbb{E}}_{x \sim \mathcal{D}} \left[\sum_{i=1}^{m} \left(f(x) - \mathop{\mathbb{E}}_{\mathcal{D}}[f] \right)^2 \right] \quad \& \quad \hat{v} \doteq \sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \left(f(\boldsymbol{x}_i) - \mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}}[f] \right)^2 \ .$$

Then $v \leq \frac{m}{m-1} \mathop{\mathbb{E}}_{\boldsymbol{x}}[\hat{v}]$.

Proof.

$$\begin{aligned}
v &= \sup_{f \in \mathcal{F}} \mathop{\mathbb{E}}_{\boldsymbol{x}} \left[\frac{1}{m} \sum_{i=1}^{m} \left(f(\boldsymbol{x}_i) - \mathop{\mathbb{E}}_{\mathcal{D}}[f] \right)^2 \right] && \textsc{Linearity} \\
&= \sup_{f \in \mathcal{F}} \mathop{\mathbb{E}}_{\boldsymbol{x}} \left[\frac{1}{m-1} \sum_{i=1}^{m} \left(f(\boldsymbol{x}_i) - \mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}}[f] \right)^2 \right] && \textsc{Bessel's Correction} \\
&\leq \mathop{\mathbb{E}}_{\boldsymbol{x}} \left[\sup_{f \in \mathcal{F}} \frac{1}{m-1} \sum_{i=1}^{m} \left(f(\boldsymbol{x}_i) - \mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}}[f] \right)^2 \right] && \textsc{Jensen's Inequality} \\
&= \tfrac{m}{m-1} \mathop{\mathbb{E}}_{\boldsymbol{x}}[\hat{v}] && \textsc{Definition of } \hat{v}
\end{aligned}$$

$\square$

Lemma D.3.2 (Empirical Wimpy Variance Concentration)**.** Suppose as in lemma D.3.1, and furthermore $\mathcal{F} \subseteq \mathcal{X} \to [0,1]$. Then

$$\mathbb{P}\left(v \geq \tfrac{m}{m-1}\hat{v} + \frac{2\ln(\frac{1}{\delta})}{m} + \sqrt{\frac{2\hat{v}\ln(\frac{1}{\delta})}{m-1}} \right) \leq \delta \enspace .$$

Proof. Suppose $\boldsymbol{X} \in \mathcal{X}^m$, let $\boldsymbol{X}_{\setminus j}$ denote the vector $\boldsymbol{X}$, missing the jth component, and take $Z(\boldsymbol{X}) \doteq \sup_{f \in \mathcal{F}} (m-1)\hat{\mathrm{V}}_{\boldsymbol{X}}[f] = \sup_{f \in \mathcal{F}} \sum_{i=1}^{m} \bigl(f(\boldsymbol{X}_i) - \hat{\mathbb{E}}_{\boldsymbol{X}}[f]\bigr)^2$, and $Z_j(\boldsymbol{X}) \doteq \sup_{f \in \mathcal{F}} (m-2)\hat{\mathrm{V}}_{\boldsymbol{X}_{\setminus j}}[f] = \sup_{f \in \mathcal{F}} \sum_{\substack{i=1 \\ i \neq j}}^{m} \bigl(f(\boldsymbol{X}_i) - \hat{\mathbb{E}}_{\boldsymbol{X}_{\setminus j}}[f]\bigr)^2$, where $\hat{\mathrm{V}}_{\boldsymbol{X}}[f]$ denotes the (unbiased) sample variance of f over $\boldsymbol{X}$.

We first show that

$$Z_j(\boldsymbol{X}) = \sup_{f \in \mathcal{F}} \left(\sum_{i=1}^{m} \bigl(f(\boldsymbol{X}_i) - \hat{\mathbb{E}}_{\boldsymbol{X}}[f]\bigr)^2 \right) - \tfrac{m-1}{m}\bigl(f(\boldsymbol{X}_j) - \hat{\mathbb{E}}_{\boldsymbol{X}_{\setminus j}}[f]\bigr)^2 \enspace ; \tag{68}$$

to see this, consider that

$$\begin{aligned}
Z_j(\boldsymbol{X}) &= \sup_{f \in \mathcal{F}} \sum_{\substack{i=1 \\ i \neq j}}^{m} \bigl(f(\boldsymbol{X}_i) - \hat{\mathbb{E}}_{\boldsymbol{X}_{\setminus j}}[f]\bigr)^2 = \sup_{f \in \mathcal{F}} \left(\sum_{i=1}^{m} \Bigl(f(\boldsymbol{X}_i) - \hat{\mathbb{E}}_{\boldsymbol{X}_{\setminus j}}[f]\Bigr)^2 \right) - \bigl(f(\boldsymbol{X}_j) - \hat{\mathbb{E}}_{\boldsymbol{X}_{\setminus j}}[f]\bigr)^2 && \textsc{Additive Identity} \\
& && \textsc{Definition of } Z_j \\
&= \sup_{f \in \mathcal{F}} \left(\sum_{i=1}^{m} \Bigl(\bigl(f(\boldsymbol{X}_i) - \hat{\mathbb{E}}_{\boldsymbol{X}}[f]\bigr) + \bigl(\hat{\mathbb{E}}_{\boldsymbol{X}}[f] - \hat{\mathbb{E}}_{\boldsymbol{X}_{\setminus j}}[f]\bigr)\Bigr)^2 \right) - \bigl(f(\boldsymbol{X}_j) - \hat{\mathbb{E}}_{\boldsymbol{X}_{\setminus j}}[f]\bigr)^2 && \textsc{Additive Identity} \\
&= \sup_{f \in \mathcal{F}} \left(\sum_{i=1}^{m} \bigl(f(\boldsymbol{X}_i) - \hat{\mathbb{E}}_{\boldsymbol{X}}[f]\bigr)^2 + 2\bigl(f(\boldsymbol{X}_i) - \hat{\mathbb{E}}_{\boldsymbol{X}}[f]\bigr)\bigl(\hat{\mathbb{E}}_{\boldsymbol{X}}[f] - \hat{\mathbb{E}}_{\boldsymbol{X}_{\setminus j}}[f]\bigr) + \bigl(\hat{\mathbb{E}}_{\boldsymbol{X}}[f] - \hat{\mathbb{E}}_{\boldsymbol{X}_{\setminus j}}[f]\bigr)^2 \right) - \bigl(f(\boldsymbol{X}_j) - \hat{\mathbb{E}}_{\boldsymbol{X}_{\setminus j}}[f]\bigr)^2 && \textsc{Unfactoring} \\
&= \sup_{f \in \mathcal{F}} \left(\sum_{i=1}^{m} \bigl(f(\boldsymbol{X}_i) - \hat{\mathbb{E}}_{\boldsymbol{X}}[f]\bigr)^2 \right) + m\bigl(\hat{\mathbb{E}}_{\boldsymbol{X}}[f] - \hat{\mathbb{E}}_{\boldsymbol{X}_{\setminus j}}[f]\bigr)^2 - \bigl(f(\boldsymbol{X}_j) - \hat{\mathbb{E}}_{\boldsymbol{X}_{\setminus j}}[f]\bigr)^2 && \sum_{i=1}^{m} \left(f(\boldsymbol{X}_i) - \hat{\mathbb{E}}_{\boldsymbol{X}}[f] \right) = 0 \\
&= \sup_{f \in \mathcal{F}} \left(\sum_{i=1}^{m} \bigl(f(\boldsymbol{X}_i) - \hat{\mathbb{E}}_{\boldsymbol{X}}[f]\bigr)^2 \right) + m\bigl(\tfrac{1}{m}f(\boldsymbol{X}_j) - \tfrac{1}{m}\hat{\mathbb{E}}_{\boldsymbol{X}_{\setminus j}}[f]\bigr)^2 - \bigl(f(\boldsymbol{X}_j) - \hat{\mathbb{E}}_{\boldsymbol{X}_{\setminus j}}[f]\bigr)^2 && \hat{\mathbb{E}}_{\boldsymbol{X}}[f] = \tfrac{1}{m}f(\boldsymbol{X}_j) + \tfrac{m-1}{m}\hat{\mathbb{E}}_{\boldsymbol{X}_{\setminus}}[f] \\
&= \sup_{f \in \mathcal{F}} \left(\sum_{i=1}^{m} \bigl(f(\boldsymbol{X}_i) - \hat{\mathbb{E}}_{\boldsymbol{X}}[f]\bigr)^2 \right) - \tfrac{m-1}{m}\bigl(f(\boldsymbol{X}_j) - \hat{\mathbb{E}}_{\boldsymbol{X}_{\setminus j}}[f]\bigr)^2 \enspace . && \textsc{Algebra}
\end{aligned}$$

The desideratum reduces to showing that $\frac{m}{m-1}Z(\cdot)$ is an $(\frac{m}{m-1}, 0)$-self-bounding function (see Boucheron et al. [2009]). *Nonnegativity* is clear from the definiton of $Z(\cdot)$. We now show *underestimation*; first consider that for any $j \in \{1, \ldots, m\}$,

$$\begin{aligned}
Z_j(\boldsymbol{X}) &= \sup_{f \in \mathcal{F}} \left(\sum_{i=1}^{m} \big(f(\boldsymbol{X}_i) - \hat{\mathop{\mathbb{E}}_{\boldsymbol{X}}}[f] \big)^2 \right) - \tfrac{m-1}{m} \big(f(\boldsymbol{X}_j) - \hat{\mathop{\mathbb{E}}_{\boldsymbol{X}_{\setminus j}}}[f] \big)^2 && \text{Eq. 68} \\
&\geq \left(\sup_{f \in \mathcal{F}} \left(\sum_{i=1}^{m} \big(f(\boldsymbol{X}_i) - \hat{\mathop{\mathbb{E}}_{\boldsymbol{X}}}[f] \big)^2 \right) \right) - \left(\sup_{f \in \mathcal{F}} \tfrac{m-1}{m} \big(f(\boldsymbol{X}_j) - \hat{\mathop{\mathbb{E}}_{\boldsymbol{X}_{\setminus j}}}[f] \big)^2 \right) && \text{Properties of Suprema} \\
&\geq Z(\boldsymbol{X}) - \tfrac{m-1}{m} \ . && \begin{array}{r}\text{Definition of } Z(\cdot) \\ \forall\, f \in \mathcal{F}\colon \ \mathrm{image}(f) \subseteq [0,1]\end{array}
\end{aligned}$$

We may thus conclude that $\frac{m}{m-1} Z_j(\boldsymbol{X}) \geq \frac{m}{m-1} Z(\boldsymbol{X}) - 1$, which satisfies 1-*underestimation*. We now show that $Z(\cdot)$, and subsequently $\frac{m}{m-1} Z(\cdot)$, obey $(\frac{m}{m-1}, 0)$-*self-boundedness*.

$$\begin{aligned}
\sum_{j=1}^{m} Z(\boldsymbol{X}) - Z_j(\boldsymbol{X}) &= mZ(\boldsymbol{X}) - \sum_{j=1}^{m} \sup_{f \in \mathcal{F}} \left(\sum_{i=1}^{m} \big(f(\boldsymbol{X}_i) - \hat{\mathop{\mathbb{E}}_{\boldsymbol{X}}}[f] \big)^2 \right) - \tfrac{m-1}{m} \big(f(\boldsymbol{X}_j) - \hat{\mathop{\mathbb{E}}_{\boldsymbol{X}_{\setminus j}}}[f] \big)^2 && \text{Eq. 68} \\
&\leq mZ(\boldsymbol{X}) - \sup_{f \in \mathcal{F}} \sum_{j=1}^{m} \left(\sum_{i=1}^{m} \big(f(\boldsymbol{X}_i) - \hat{\mathop{\mathbb{E}}_{\boldsymbol{X}}}[f] \big)^2 \right) - \tfrac{m-1}{m} \big(f(\boldsymbol{X}_j) - \hat{\mathop{\mathbb{E}}_{\boldsymbol{X}_{\setminus j}}}[f] \big)^2 && \text{Subadditivity} \\
&= mZ(\boldsymbol{X}) - \sup_{f \in \mathcal{F}} m \left(\sum_{i=1}^{m} \big(f(\boldsymbol{X}_i) - \hat{\mathop{\mathbb{E}}_{\boldsymbol{X}}}[f] \big)^2 \right) - \tfrac{m-1}{m} \sum_{j=1}^{m} \big(f(\boldsymbol{X}_j) - \hat{\mathop{\mathbb{E}}_{\boldsymbol{X}_{\setminus j}}}[f] \big)^2 && \text{Linearity} \\
&= mZ(\boldsymbol{X}) - \sup_{f \in \mathcal{F}} m \left(\sum_{i=1}^{m} \big(f(\boldsymbol{X}_i) - \hat{\mathop{\mathbb{E}}_{\boldsymbol{X}}}[f] \big)^2 \right) - \tfrac{m-1}{m} \sum_{j=1}^{m} \big(\tfrac{m}{m-1} f(\boldsymbol{X}_j) - \tfrac{m}{m-1} \hat{\mathop{\mathbb{E}}_{\boldsymbol{X}}}[f] \big)^2 && \hat{\mathop{\mathbb{E}}_{\boldsymbol{X}_{\setminus j}}}[f] = \tfrac{1}{m-1} \big(m \hat{\mathop{\mathbb{E}}_{\boldsymbol{X}}}[f] - f(\boldsymbol{X}_j) \big) \\
&= \tfrac{m(m-1)}{m-1} Z(\boldsymbol{X}) - \big(\tfrac{m(m-1)}{m-1} - \tfrac{m}{m-1} \big) Z(\boldsymbol{X}) = \tfrac{m}{m-1} Z(\boldsymbol{X}) && \text{Definition of } Z(\cdot)
\end{aligned}$$

It is now a straightforward matter to observe that $\frac{m}{m-1} Z(\cdot)$, also obeys $(\frac{m}{m-1}, 0)$-*self-boundedness*, as

$$\sum_{j=1}^{m} Z(\boldsymbol{X}) - Z_j(\boldsymbol{X}) \leq \tfrac{m}{m-1} Z(\boldsymbol{X}) \Rightarrow \tfrac{m}{m-1} Z(\boldsymbol{X}) - \tfrac{m}{m-1} Z_j(\boldsymbol{X}) \leq \tfrac{m}{m-1} \cdot \big(\tfrac{m}{m-1} Z(\boldsymbol{X}) \big) \ .$$

With all three desiderata shown, we may now conclude that $\frac{m}{m-1} Z(\cdot)$ is a 0-overestimating, 1-underestimating

$(\frac{m}{m-1}, 0)$-self-bounding function. Now, observe that

$$\begin{aligned}
\mathop{\mathbb{P}}_{\boldsymbol{x}}\Big(\mathrm{W} \geq \tfrac{m}{m-1}\widehat{\mathrm{W}} + \varepsilon\Big) &= \mathop{\mathbb{P}}_{\boldsymbol{x}}\Big(\tfrac{m}{m-1}\widehat{\mathrm{W}} \leq \mathrm{W} - \varepsilon\Big) && \textsc{Equivalent Events} \\
&\leq \mathop{\mathbb{P}}_{\boldsymbol{x}}\left(\tfrac{m}{m-1}\widehat{\mathrm{W}} \leq \tfrac{m}{m-1}\,\mathbb{E}[\widehat{\mathrm{W}}] - \varepsilon\right) && \textsc{Lemma D.3.1} \\
&= \mathop{\mathbb{P}}_{\boldsymbol{x}}\left(\tfrac{1}{m-1}Z(\boldsymbol{x}) \leq \tfrac{1}{m-1}\,\mathbb{E}[Z] - \varepsilon\right) && \widehat{\mathrm{W}} = \tfrac{1}{m}Z(\boldsymbol{x}) \\
&= \mathop{\mathbb{P}}_{\boldsymbol{x}}\left(\tfrac{1}{m}\cdot\tfrac{m}{m-1}Z(\boldsymbol{x}) \leq \tfrac{1}{m}\cdot\tfrac{m}{m-1}\,\mathbb{E}[Z] - \varepsilon\right) \ . && \textsc{Equivalent Events} \\
&\leq \exp\left(\frac{-m\varepsilon^2}{2(\frac{m}{m-1}\hat{v} + \varepsilon)}\right) = \exp\left(\frac{-m\varepsilon^2}{2(\frac{m}{m-1}\hat{v} + \varepsilon)}\right) \ . && \textsc{Sbf Inequality}
\end{aligned}$$

The step marked Sbf Inequality follows via standard self-bounding function inequalities; see Boucheron et al. [2009]. An application of the quadratic formula (positive solution) then yields the result, as

$$\varepsilon = \frac{\ln(\frac{1}{\delta})}{m} + \sqrt{\left(\frac{\ln(\frac{1}{\delta})}{m}\right)^2 + \frac{2\hat{v}\ln(\frac{1}{\delta})}{m-1}} \leq \frac{2\ln(\frac{1}{\delta})}{m} + \sqrt{\frac{2\hat{v}\ln(\frac{1}{\delta})}{m-1}} \Rightarrow \mathbb{P}\Big(v \geq \tfrac{m}{m-1}\hat{v} + \varepsilon\Big) \leq \delta \ .$$

□

With these results, we now show a variance-sensitive supremum-deviation bound in terms of the n-MCERA.

Lemma D.3.3 (Empirical-Variance Sensitive Uniform Convergence Tail Bounds)**.** Suppose as in theorem 4.2.2. It then holds with probability at least $1-\delta$ over choice of $\boldsymbol{x}, \boldsymbol{\sigma}$ that

$$\sup_{f\in\mathcal{F}}\left|\mathop{\mathbb{E}}_{\mathcal{D}}[f] - \mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}}[f]\right| \leq 2\widetilde{\mathfrak{R}} + \varepsilon_{\mathrm{SD}} \ .$$

Proof. This result follows from a chain of 4 union bounds, each of probability $\frac{\delta}{4}$, where first v is bounded in terms of $\hat{v}$ with lemma D.3.2, then $\hat{\mathfrak{R}}_m(\mathcal{F}, \boldsymbol{x})$ is bounded in terms of $\hat{\mathfrak{R}}^n_m(\mathcal{F}, \boldsymbol{x}, \boldsymbol{\sigma})$ with McDiarmid's bounded difference inequality [McDiarmid, 1989], $\mathfrak{R}_m(\mathcal{F}, \mathcal{D})$ is bounded in terms of $\hat{\mathfrak{R}}_m(\mathcal{F}, \boldsymbol{x})$ with the self-bounding function inequality [Boucheron et al., 2000, Thm. 1] (see Oneto et al. [2013] for details), and finally the supremum deviation is bounded with the symmetrization inequality equation 4 [Boucheron et al., 2013, see Lemma 11.4] and the refinement of Talagrand's inequality for empirical processes due to Bousquet [2002, Thm. 2.3]. The aforementioned results implicitly assume $r = 1$, so we first consider that case, and then note that by linear scaling, we attain the general result. In particular, while $\mathcal{F}$, the SD, and each Rademacher average and tail bound thereof scale linearly, the *variance-like* quantities ε_v, v, and $\hat{v}$ all scale *quadratically* with r^2, which has a linear effect, since these always appear inside a root in the final bound. Note also that while these bounds do place some restrictions on the failure probability δ, our 4-term union bound holds vacuously when $\delta \geq \frac{1}{4}$, and said restrictions do not apply to $\delta < \frac{1}{4}$.

All but the Monte-Carlo bounds are discussed elsewhere, so we give further detail here. Note that this result generalizes a theorem of Bartlett and Mendelson [2002, thm. 11], which holds for the special case of $n = 1$. The probability space in play is the $n \times m$ Rademacher random variables of $\boldsymbol{\sigma}$, changing any one of which produces a change of no more than $\frac{2r}{nm}$ to the n-MCERA, which yields via McDiarmid's inequality

$$\mathbb{P}_{\boldsymbol{\sigma}}\left(\hat{\mathfrak{R}}_m(\mathcal{F}, \boldsymbol{x}) \geq \hat{\mathfrak{R}}^n_m(\mathcal{F}, \boldsymbol{x}, \boldsymbol{\sigma}) + 2r\sqrt{\frac{\ln(\frac{4}{\delta})}{2nm}}\right) \leq \frac{\delta}{4} \ .$$

This is sufficient to show the stated result. A more careful analysis can improve this bound by noting that, due to the presence of the absolute value, $\hat{\mathfrak{R}}^n_m(\mathcal{F}, \boldsymbol{x}, \boldsymbol{\sigma}) = \hat{\mathfrak{R}}^n_m(\mathcal{F}, \boldsymbol{x}, -\boldsymbol{\sigma})$, and there is essentially one fewer degree of freedom as a result.

We note also that the stated bound in each of the above references is in most cases sharper than what we present; this is because we are interested in the explicit form of ε as a function of δ, and thus use the convenient sub-gamma characterizations, even when sharper sub-Poisson bounds are available. With numeric inversion, these bounds can be used with fixed δ, and are appropriate when sharper bounds are required, but they come at the cost of computational overhead, algebraic inconvenience, and lack of interpretability. Similarly, the uniform union bound of probability $\frac{\delta}{4}$ is also suboptimal, but again is algebraically convenient, and the potential improvement from nonuniform division of δ is quite small. □

With this result, we are now ready to show theorem 4.2.2.

Theorem 4.2.2 (Variance-Sensitive Bounds)**.** Suppose *distribution* $\mathcal{D}$ over $\mathcal{X}$, *training sample* $\boldsymbol{x} \sim \mathcal{D}^m$, *Monte-Carlo trial count* n, i.i.d. Rademacher matrix $\boldsymbol{\sigma} \in (\pm 1)^{n \times m}$, codec family $\mathcal{C}$, loss family $\mathcal{L} \subseteq \mathbb{R}^a_{0+} \to [-r, r]$, λ-Lipschitz *welfare* $\mathrm{W}(\cdot)$, and codec selection size k. $\forall \delta \in (0, 1)$, let $\mathcal{F} \doteq \min \circ \mathcal{L} \circ \binom{\mathcal{C}}{k}$, take $\hat{v}$ to be the *empirical wimpy variance* of $\mathcal{F}$ over $\boldsymbol{x}$ and $\hat{v}^{\mathsf{raw}}$ to be the *empirical raw wimpy variance*, and take

1. $\varepsilon_{\mathrm{ERA}} \doteq \dfrac{4r\frac{4}{\delta}}{3nm} + \sqrt{\dfrac{4\hat{v}^{\mathsf{raw}} \ln \frac{4}{\delta}}{nm}}$;
2. $\varepsilon_{\mathrm{RA}} \doteq \dfrac{2r \ln \frac{4}{\delta}}{m} + \sqrt{\dfrac{2r(\hat{\mathfrak{R}}^n_m(\mathcal{F}, \boldsymbol{x}, \boldsymbol{\sigma}) + \varepsilon_{\mathrm{ERA}}) \ln \frac{4}{\delta}}{m}}$;
3. $\tilde{\mathfrak{R}} \doteq \hat{\mathfrak{R}}^n_m(\mathcal{F}, \boldsymbol{x}, \boldsymbol{\sigma}) + \varepsilon_{\mathrm{ERA}} + \varepsilon_{\mathrm{RA}}$;
4. $\tilde{v} \doteq \frac{m}{m-1}\hat{v} + \dfrac{4r^2 \ln \frac{4}{\delta}}{m} + \sqrt{\dfrac{2r^2 \hat{v} \ln \frac{4}{\delta}}{m-1}}$; &
5. $\varepsilon_{\mathrm{SD}} \doteq \dfrac{r \ln \frac{4}{\delta}}{3m} + \sqrt{\dfrac{2(\tilde{v} + 4r\tilde{\mathfrak{R}}) \ln \frac{4}{\delta}}{m}}$.

It then holds with with pr. $\geq 1 - \delta$ over $\boldsymbol{x}, \boldsymbol{\sigma}$ that

1. $\left|\mathrm{W}(\hat{\boldsymbol{c}}_{\mathrm{W}}, \mathcal{L}, \mathcal{D}) - \hat{\mathrm{W}}(\hat{\boldsymbol{c}}_{\mathrm{W}}, \mathcal{L}, \boldsymbol{x})\right| \leq 2\lambda\widetilde{\mathfrak{R}} + \lambda\varepsilon_{\mathrm{SD}}$ &

2. $\mathrm{W}(\hat{\boldsymbol{c}}_{\mathrm{W}}, \mathcal{L}, \mathcal{D}) - \mathrm{W}(\boldsymbol{c}^*_{\mathrm{W}}, \mathcal{L}, \mathcal{D}) \leq 4\lambda\widetilde{\mathfrak{R}} + 2\lambda\varepsilon_{\mathrm{SD}}$.

Proof. This result follows via lemma D.3.3, followed by several straightforward deterministic manipulations. This result is also a probabilistic guarantee, and it suffices to show that the condition on welfare is satisfied under the assumptions and probabilistic guarantee of the aforementioned lemma. Henceforth we assume the conclusion of lemma D.3.3 holds, and thus the following hold only with probability $\geq 1 - \delta$.

By assumption, $\mathrm{W}(\boldsymbol{c}, \mathcal{L}, \boldsymbol{x})$ is λ-ℓ_∞-Lipschitz in the empirical risks (w.r.t. $\mathcal{L}$) of $\boldsymbol{c}$ on $\boldsymbol{x}$. Consequently, a uniform bound of $2\widetilde{\mathfrak{R}} + \varepsilon_{\mathrm{SD}}$ on *empirical risks* (lemma D.3.3) implies a $\lambda(2\widetilde{\mathfrak{R}} + \varepsilon_{\mathrm{SD}})$ bound on *welfare*. It therefore holds that

$$\left|\mathrm{W}(\hat{\boldsymbol{c}}_{\mathrm{W}}, \mathcal{L}, \mathcal{D}) - \hat{\mathrm{W}}(\hat{\boldsymbol{c}}_{\mathrm{W}}, \mathcal{L}, \boldsymbol{x})\right| \leq \lambda(2\widetilde{\mathfrak{R}} + \varepsilon_{\mathrm{SD}}) \ ,$$

which yields item (1). Similarly, we have

$$\left|\hat{\mathrm{W}}(\boldsymbol{c}^*_{\mathrm{W}}, \mathcal{L}, \boldsymbol{x}) - \mathrm{W}(\boldsymbol{c}^*_{\mathrm{W}}, \mathcal{L}, \mathcal{D})\right| \leq \lambda(2\widetilde{\mathfrak{R}} + \varepsilon_{\mathrm{SD}}) \ ,$$

and consequently, by the triangle inequality (noting that $\hat{\mathrm{W}}(\hat{\boldsymbol{c}}_{\mathrm{W}}, \mathcal{L}, \boldsymbol{x}) \leq \hat{\mathrm{W}}(\boldsymbol{c}^*_{\mathrm{W}}, \mathcal{L}, \boldsymbol{x})$ by definition), we have

$$\left|\mathrm{W}(\hat{\boldsymbol{c}}_{\mathrm{W}}, \mathcal{L}, \mathcal{D}) - \mathrm{W}(\boldsymbol{c}^*_{\mathrm{W}}, \mathcal{L}, \mathcal{D})\right| \leq 2\lambda(2\widetilde{\mathfrak{R}} + \varepsilon_{\mathrm{SD}}) = 4\lambda\widetilde{\mathfrak{R}} + 2\lambda\varepsilon_{\mathrm{SD}} \ ,$$

which completes the result. Note that here, the largest value the LHS can attain occurs when the empirical welfare of $\boldsymbol{c}^*_{\mathrm{W}}$ is *overestimated* by $\lambda(2\widetilde{\mathfrak{R}} + \varepsilon_{\mathrm{SD}})$, while $\hat{\boldsymbol{c}}_{\mathrm{W}}$ is *underestimated* by $\lambda(2\widetilde{\mathfrak{R}} + \varepsilon_{\mathrm{SD}})$, making $\hat{\boldsymbol{c}}_{\mathrm{W}}$ *empirically optimal* over $\boldsymbol{x}$, and no worse than $2\lambda(2\widetilde{\mathfrak{R}} + \varepsilon_{\mathrm{SD}})$-suboptimal over $\mathcal{D}$. □

D.3.2 Proof of Theorem 4.2.3

Theorem 4.2.3 combines several well-understood properties of variance with a bound on the Rademacher averages of minima of function families. We fist show lemmas regarding variances of linear families and Rademacher averages of minima. Each lemma in this subsection assumes

1. *dual norms*: p, q, s.t. $|p^{-1}| + |q^{-1}| = 1$;
2. *Codec family*: $\mathcal{C}$ such that $\sup_{c \in \mathcal{C}, x \in \mathcal{X}} \|c(x)\|_q \leq 1$; and
3. *Linear loss family* $\mathcal{L}_p = \{\, \ell(\boldsymbol{a}) \doteq \boldsymbol{w} \cdot \boldsymbol{a} \mid \boldsymbol{w} \in \mathbb{R}^a_{0+} \text{ s.t. } \|\boldsymbol{w}\|_p \leq 1 \,\}$ (as defined in eq. 34).

We first show bounds on the variances of singleton codec families.

Lemma D.3.4 (Variance of Linear Families)**.**

$$\sup_{c \in \mathcal{C}} \sup_{\ell \in \mathcal{L}} \hat{\mathbb{V}}_{\boldsymbol{x}}[\ell \circ c] = \sup_{c \in \mathcal{C}} \left\| \hat{\mathbb{C}}_{\boldsymbol{x}}[c] \right\|_{q \to p}$$

Proof. First note that this result essentially holds componentwise over each $c \in \mathcal{C}$. For a single c, in the special case of $p = q = 2$, it is the standard Rayleigh decomposition argument used in *singular value decomposition* (SVD) and *principal component analysis* (PCA), and for arbitrary $p \neq q \neq 2$ dual norms, the spectral norm is generalized to an operator norm. We introduce the unit-sphere notation

$$\mathcal{S}_p^{a-1} \doteq \{ \boldsymbol{w} \in \mathbb{R}^a \mid \|\boldsymbol{w}\|_p = 1 \} \ ,$$

and now show the result proper:

$$\begin{aligned}
\sup_{c \in \mathcal{C}} \sup_{\ell \in \mathcal{L}} \hat{\mathbb{V}}_{\boldsymbol{x}}[\ell \circ c] &= \sup_{c \in \mathcal{C}} \sup_{\boldsymbol{w} \in \mathcal{S}_p^{a-1}} \hat{\mathbb{V}}_{\boldsymbol{x}}[\boldsymbol{w} \cdot c(\cdot)] && \text{Definition of } \mathcal{L} \\
&= \sup_{c \in \mathcal{C}} \sup_{\boldsymbol{w} \in \mathcal{S}_p^{a-1}} \boldsymbol{w} \, \hat{\mathbb{C}}_{\boldsymbol{x}}[c] \boldsymbol{w}^\top && \text{Variance Properties} \\
&= \sup_{c \in \mathcal{C}} \left\| \hat{\mathbb{C}}_{\boldsymbol{x}}[c] \right\|_{p \to q} \ . && \begin{array}{c} \left\| \hat{\mathbb{C}}_{\boldsymbol{x}}[c] \boldsymbol{w}^\top \right\|_q \leq \|\boldsymbol{w}\|_p \left\| \hat{\mathbb{C}}_{\boldsymbol{x}}[c] \right\|_{p \to q} \\ \text{Hölder's Inequality} \end{array}
\end{aligned}$$

Note that Hölder's inequality, which generalizes the Cauchy-Schwarz inequality to $\boldsymbol{a} \cdot \boldsymbol{b} \leq \|\boldsymbol{a}\|_p \|\boldsymbol{b}\|_q$, holds with equality for appropriate choice of $\boldsymbol{w}$, which the supremum selects, yielding (strict) equality. □

We now show compositional bounds on the Rademacher averages of codec selection.

Lemma D.3.5 (Bounding Rademacher Averages of Minima)**.** Suppose $\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_k \subseteq \mathcal{X} \to \mathbb{R}$, and let

$$\mathcal{F}_{\min} \doteq \{ \min_{i \in 1, \ldots, k} \boldsymbol{f}_i(x) \mid \boldsymbol{f} \in \mathcal{F}_1 \times \cdots \times \mathcal{F}_k \} \ .$$

Then $\hat{\mathfrak{R}}_m(\mathcal{F}_{\min}, \boldsymbol{x}) \leq \displaystyle\sum_{i=1}^{m} \hat{\mathfrak{R}}_m(\mathcal{F}_i, \boldsymbol{x})$.

Proof. We first show the result for $k = 2$.

$$
\begin{aligned}
\hat{\mathfrak{R}}_m(\mathcal{F}_{\min}, \boldsymbol{x}) &= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{f_1\in\mathcal{F}_1, f_2\in\mathcal{F}_2}\left|\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{\sigma}_i \min\bigl(f_1(\boldsymbol{x}_i), f_2(\boldsymbol{x}_i)\bigr)\right|\right] && \textsc{Definition} \\
&= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{f_1\in\mathcal{F}_1, f_2\in\mathcal{F}_2}\left|\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{\sigma}_i\left(\tfrac{f_1+f_2-|f_1-f_2|}{2}\right)(\boldsymbol{x}_i)\right|\right] && \textsc{Minimum Identity} \\
&\leq \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\substack{f_1\in\mathcal{F}_1\\ f_2\in\mathcal{F}_2}}\left|\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{\sigma}_i\left(\tfrac{f_1+f_2}{2}\right)(\boldsymbol{x}_i)\right|\right] + \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\substack{f_1\in\mathcal{F}_1\\ f_2\in\mathcal{F}_2}}\left|\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{\sigma}_i\left(\tfrac{|f_1-f_2|}{2}\right)(\boldsymbol{x}_i)\right|\right] && \textsc{Triangle Inequality} \\
&= \hat{\mathfrak{R}}_m(\tfrac{f_1+f_2}{2}, \boldsymbol{x}) + \hat{\mathfrak{R}}_m(\tfrac{|\mathcal{F}_1\ominus\mathcal{F}_2|}{2}, \boldsymbol{x}) && \textsc{Definition} \\
&= \hat{\mathfrak{R}}_m(\tfrac{f_1+f_2}{2}, \boldsymbol{x}) + \hat{\mathfrak{R}}_m(|\cdot|\circ\tfrac{\mathcal{F}_1\ominus\mathcal{F}_2}{2}, \boldsymbol{x}) && \textsc{Algebra} \\
&\leq \hat{\mathfrak{R}}_m(\tfrac{f_1+f_2}{2}, \boldsymbol{x}) + \hat{\mathfrak{R}}_m(\tfrac{\mathcal{F}_1\ominus\mathcal{F}_2}{2}, \boldsymbol{x}) && \textsc{Talagrand's Contraction Inequality} \\
&\leq \hat{\mathfrak{R}}_m(\mathcal{F}_1, \boldsymbol{x}) + \hat{\mathfrak{R}}_m(\mathcal{F}_2, \boldsymbol{x}) && \textsc{Direct Summation}
\end{aligned}
$$

The result now follows by induction. Note that the contraction inequality avoids the usual 2-factor gap, as the $|\cdot|$ function is *odd*. □

With these ingredients, we now restate and show theorem 4.2.3.

Theorem 4.2.3 (Bounds for Linear Loss Families)**.** Suppose p, q, s.t. $|p^{-1}| + |q^{-1}| = 1$, *codec family* $\mathcal{C}$, $\mathcal{L}_p$ as in (34), and assume $\sup_{c\in\mathcal{C}, x\in\mathcal{X}} \|c(x)\|_q \leq s$. For all $k \in 1, \ldots |\mathcal{C}|$, take $\mathcal{F}_k \doteq \min\circ\mathcal{L}_p\circ\binom{\mathcal{C}}{k}$, and $\hat{v}_k$ to be the *empirical wimpy variance* of $\mathcal{F}_k$ on *sample* $\boldsymbol{x}$. Taking $\hat{\mathbb{C}}_{\boldsymbol{x}}[c] \in \mathbb{R}^{a\times a}$ to denote the *empirical covariance matrix* of the attributes of codec c over $\boldsymbol{x}$, and $\|\cdot\|_{p\to q}$ to be the ℓ_p-ℓ_q *operator norm*, we have

$$
\hat{v}_1 = \sup_{c\in\mathcal{C}}\sup_{\ell\in\mathcal{L}}\hat{\mathop{\mathbb{V}}_{\boldsymbol{x}}}[\ell\circ c] \leq \sup_{c\in\mathcal{C}}\left\|\hat{\mathop{\mathbb{C}}_{\boldsymbol{x}}}[c]\right\|_{q\to p} \ , \ \& \quad \hat{\mathfrak{R}}^n_m(\mathcal{F}_1, \boldsymbol{x}, \boldsymbol{\sigma}) \leq \frac{1}{n}\sum_{j=1}^{n}\sup_{c\in\mathcal{C}}\left\|\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{\sigma}_{j,i}c(\boldsymbol{x}_i)\right\|_q \ .
$$

Furthermore, for any $a, b \in \mathbb{N}$ s.t. $k = a + b$, we have

$$
\hat{v}_k \leq \hat{v}_a + \hat{v}_b \leq k\hat{v}_1 \ , \ \& \quad \hat{\mathfrak{R}}_m(\mathcal{F}_k, \boldsymbol{x}) \leq \hat{\mathfrak{R}}_m(\mathcal{F}_a, \boldsymbol{x}) + \hat{\mathfrak{R}}_m(\mathcal{F}_b, \boldsymbol{x}) \leq k\hat{\mathfrak{R}}_m(\mathcal{F}_1, \boldsymbol{x}) \ .
$$

Proof. We first consider the bounds on $\hat{v}_1$ and $\hat{\mathfrak{R}}^1_m(\mathcal{F}_1, \boldsymbol{x}, \boldsymbol{\sigma})$. Both are inequalities, as they are shown by defining a linear loss family without nonnegativity constraints $\mathcal{L}_{p\pm}$, and bounding the corresponding quantities for $\mathcal{L}_{p\pm}$. In particular, take

$$\mathcal{L}_{p\pm} \doteq \{\, \ell(\boldsymbol{a}) \doteq \boldsymbol{w} \cdot \boldsymbol{a} \mid \boldsymbol{w} \in \mathbb{R}^a \text{ s.t. } \|\boldsymbol{w}\|_p \leq 1 \,\} \enspace ,$$

which differs from $\mathcal{L}_p$ in that $\boldsymbol{w} \in \mathbb{R}^a_{0+}$ is relaxed to $\boldsymbol{w} \in \mathbb{R}^a$, thus $\mathcal{L}_p \subseteq \mathcal{L}_{p\pm}$. Now take $\mathcal{F}_{1\pm} \doteq \mathcal{L}_p \circ \mathcal{C}$, and note that $\mathcal{F}_1 \subseteq \mathcal{F}_{1\pm}$. It thus holds that the empirical wimpy variance and 1-ERA of $\mathcal{F}_{1\pm}$ are at least as great as the corresponding quantities for $\mathcal{F}_1$.

Consequently, from the definition of operator norm, we have

$$\hat{v}_1 = \sup_{c \in \mathcal{C}} \sup_{\ell \in \mathcal{L}_p} \hat{\mathbb{V}}_{\boldsymbol{x}}[\ell \circ c] \leq \sup_{c \in \mathcal{C}} \sup_{\ell \in \mathcal{L}_{p\pm}} \hat{\mathbb{V}}_{\boldsymbol{x}}[\ell \circ c] = \sup_{c \in \mathcal{C}} \left\| \hat{\mathbb{C}}_{\boldsymbol{x}}[c] \right\|_{q \to p} \enspace .$$

Similarly, via dual-norm properties, we have

$$\hat{\mathfrak{R}}^1_m(\mathcal{F}_1, \boldsymbol{x}, \boldsymbol{\sigma}) \leq \hat{\mathfrak{R}}^1_m(\mathcal{F}_{1\pm}, \boldsymbol{x}, \boldsymbol{\sigma}) = \sup_{c \in \mathcal{C}} \left\| \frac{1}{m} \sum_{i=1}^{m} \boldsymbol{\sigma}_i c(\boldsymbol{x}_i) \right\|_q \enspace .$$

Now, note that the property

$$\hat{v}_{a+b} \leq \hat{v}_a + \hat{v}_b \leq (a+b)\hat{v}_1$$

then follows by subadditivity of the variance of minima, and finally

$$\hat{\mathfrak{R}}_m(\mathcal{F}_{a+b}, \boldsymbol{x}) \leq \hat{\mathfrak{R}}_m(\mathcal{F}_a, \boldsymbol{x}) + \hat{\mathfrak{R}}_m(\mathcal{F}_b, \boldsymbol{x}) \leq (a+b)\hat{\mathfrak{R}}_m(\mathcal{F}_1, \boldsymbol{x})$$

follows from lemma D.3.5. □

Corollary D.3.6 (Bounds for Linear Loss Families)**.** Suppose as in theorem 4.2.3, and also $p = q = 2$, and all attribute values are contained by the unit sphere. Then

$$\hat{v}_k \leq k \sup_{c \in \mathcal{C}} \left\| \hat{\mathbb{C}}_{\boldsymbol{x}}[c] \right\|_{2 \to 2} \enspace ,$$

where $\|\cdot\|_{2\to 2}$ is the *spectral norm* (largest eigenvalue). Additionally, with pr. $\geq 1 - \delta$ over $\boldsymbol{x}, \boldsymbol{\sigma}$, we have

$$\begin{aligned}
\hat{\mathfrak{R}}_m(\mathcal{F}_k, \boldsymbol{x}) &\leq k \left(\hat{\mathfrak{R}}^n_m(\mathcal{F}_1, \boldsymbol{x}, \boldsymbol{\sigma}) + \sqrt{\frac{\ln \frac{1}{\delta}}{2nm}} \right) \\
&\leq k \left(\frac{1}{n} \sum_{i=1}^{n} \sup_{c \in \mathcal{C}} \left\| \frac{1}{m} \sum_{j=1}^{m} \boldsymbol{\sigma}_{i,j} c(\boldsymbol{x}_j) \right\|_2 + \sqrt{\frac{\ln \frac{1}{\delta}}{2nm}} \right) \enspace .
\end{aligned}$$

Consequently, theorem 4.2.2 holds with

$$\begin{aligned}\varepsilon_{\mathrm{RA}} &\leq \frac{\ln\frac{4}{\delta}}{3m} + \sqrt{\frac{2k(\hat{\mathfrak{R}}^n_m(\mathcal{F}_1, \boldsymbol{x}, \boldsymbol{\sigma}) + \varepsilon_{\mathrm{ERA}})\ln\frac{4}{\delta}}{m}};\\ \tilde{\mathfrak{R}} &\leq k\big(\hat{\mathfrak{R}}^n_m(\mathcal{F}_1, \boldsymbol{x}, \boldsymbol{\sigma}) + \varepsilon_{\mathrm{ERA}}\big) + \varepsilon_{\mathrm{RA}};\ \&\\ \tilde{v} &\leq \tfrac{m}{m-1}k\hat{v}_1 + \frac{2\ln\frac{4}{\delta}}{m} + \sqrt{\frac{2k\hat{v}_1\ln\frac{4}{\delta}}{m-1}}\ .\end{aligned}$$

Proof. This result follows directly from theorem 4.2.3, noting that $\|\cdot\|_{2\to 2}$ is the *spectral norm*, and applying the McDiarmid argument of lemma D.3.3 to (probabilistically) upper-bound $\hat{\mathfrak{R}}_m(\mathcal{F}_1, \boldsymbol{x})$ as $\hat{\mathfrak{R}}^n_m(\mathcal{F}_1, \boldsymbol{x}, \boldsymbol{\sigma}) + \varepsilon_{\mathrm{ERA}}$, and then noting that $\hat{\mathfrak{R}}_m(\mathcal{F}_k, \boldsymbol{x}) \leq k\hat{\mathfrak{R}}_m(\mathcal{F}_1, \boldsymbol{x})$. □

D.3.3 Justifications of Equations

We now justify or further elucidate several numbered equations found in the paper body.

Asymptotic Uniform Convergence Bounds We first consider equations 33 and 35. These follow from theorem 4.2.2 and corollary D.3.6, respectively. Equation 33 assumes $n \in \mathbf{\Omega}(\frac{r^2}{v})$, which is required to ensure that the contributions of the term $\varepsilon_{\mathrm{ERA}}$ are dominated by $\varepsilon_{\mathrm{SD}}$, as

$$\varepsilon_{\mathrm{ERA}} = \sqrt{\frac{r^2}{nm}} \in \mathbf{O}\sqrt{\frac{v}{m}}\ .$$

Note also that the ERA itself dominates the $\varepsilon_{\mathrm{ERA}}$ terms, as by the Khintchine inequality [see Haagerup, 1982], we have

$$\hat{\mathfrak{R}}_m(\mathcal{F}, \boldsymbol{x}) \geq \sqrt{\frac{\hat{v}}{m}} \underset{m\to\infty}{\leadsto} \sqrt{\frac{v}{m}}\ ,$$

where the $\hat{v} \leadsto v$ is asymptotic in m for fixed r. Essentially the same arguments hold for equation 35, now assuming $n \in \mathbf{\Omega}(\frac{1}{v_1})$, as we assume (corollary D.3.6) $r^2 = r = 1$, and the analysis now operates on $\mathcal{F}_1$, thus v_1 is the variance-concept of interest.

Pareto Optimality Equations We now consider equations 37 and 38. Note that we define the Pareto-optimal fronts as the LCH$(\cdot)$ *of the Pareto-optimal codecs / codec sets*, but the expressions given merely take the LCH$(\cdot)$ of *all codecs / codec sets*. This is because any Pareto-suboptimal codecs or codec sets included in the LCH$(\cdot)$ by definition have no impact on the LCH$(\cdot)$, thus the two formulations are identical.

Appendix E

Supplementary Material for Chapter 5

E.1 Efficient FPAC-Learning

We now show theorem 5.3.1.

Theorem 5.3.1 (Efficient FPAC Learning via Convex Optimization)**.** Suppose each hypothesis space $\mathcal{H}_d \in \mathcal{H}$ is indexed by $\Theta_d \subseteq \mathbb{R}^{\mathrm{Poly}(d)}$, i.e., $\mathcal{H}_d = \{h(\cdot;\theta) \mid \theta \in \Theta_d\}$, s.t. (Euclidean) $\mathrm{Diam}(\Theta_d) \in \mathrm{Poly}(d)$, and $\forall x \in \mathcal{X}, \theta \in \Theta_d$, $h(x;\theta)$ can be evaluated in $\mathrm{Poly}(d)$ time, and $\tilde{\theta} \in \mathbb{R}^{\mathrm{Poly}(d)}$ can be Euclidean-projected onto Θ_d in $\mathrm{Poly}(d)$ time. Suppose also ℓ such that $\forall x \in \mathcal{X}, y \in \mathcal{Y}: \ \theta \mapsto \ell(y, h(x;\theta))$ is a *convex function*, and suppose Lipschitz constants $\lambda_\ell, \lambda_{\mathcal{H}} \in \mathrm{Poly}(d)$ and some norm $\|\cdot\|_{\mathcal{Y}}$ over $\mathcal{Y}$ s.t. ℓ is λ_ℓ-$\|\cdot\|_{\mathcal{Y}}$-$|\cdot|$-Lipschitz in $\hat{y}$, i.e.,

$$\forall y, \hat{y}, \hat{y}' \in \mathcal{Y}: \ \left|\ell(y,\hat{y}) - \ell(y,\hat{y}')\right| \leq \lambda_\ell \left\|\hat{y} - \hat{y}'\right\|_{\mathcal{Y}} \ ,$$

and also that each $\mathcal{H}_d$ is $\lambda_{\mathcal{H}}$-$\|\cdot\|_2$-$\|\cdot\|_{\mathcal{Y}}$-Lipschitz in θ, i.e.,

$$\forall x \in \mathcal{X}, \theta, \theta' \in \Theta_d: \ \left\|h(x;\theta) - h(x;\theta')\right\|_{\mathcal{Y}} \leq \lambda_{\mathcal{H}} \left\|\theta - \theta'\right\|_2 \ .$$

Finally, assume $\ell \circ \mathcal{H}_d$ exhibits ε-δ *uniform convergence* with sample complexity $\mathrm{m}_{\mathrm{UC}}(\varepsilon,\delta,d) \in \mathrm{Poly}(\frac{1}{\varepsilon}, \frac{1}{\delta}, d)$.

It then holds that, for arbitrary initial guess $\theta_0 \in \Theta_d$, for any group distributions $\mathcal{D}_{1:g}$, group weights $\boldsymbol{w}$, and fair malfare function $\text{M}(\cdot;\cdot)$, the algorithm (see algorithm 2)

$$\mathcal{A}(\mathcal{D}_{1:g}, \boldsymbol{w}, \text{M}(\cdot;\cdot), \varepsilon, \delta, d) \doteq \mathcal{A}_{\mathrm{PSG}}\big(\ell, \mathcal{H}_d, \theta_0, \mathrm{m}_{\mathrm{UC}}(\cdot,\cdot,d), \mathcal{D}_{1:g}, \boldsymbol{w}, \text{M}(\cdot;\cdot), \varepsilon, \delta\big)$$

fair-PAC-learns $(\mathcal{H}, \ell)$ with sample complexity $\mathrm{m}(\varepsilon,\delta,d,g) = g \cdot \mathrm{m}_{\mathrm{UC}}(\frac{\varepsilon}{3}, \frac{\delta}{g}, d)$, and (training) time-complexity

$\in \mathrm{Poly}(\frac{1}{\varepsilon}, \frac{1}{\delta}, d, g)$, thus $(\mathcal{H}, \ell) \in \mathrm{FPAC}^{\mathrm{Agn}}_{\mathrm{Poly}}$.

Proof. We now show that this subgradient-method construction of $\mathcal{A}$ requires $\mathrm{Poly}(\frac{1}{\varepsilon}, \frac{1}{\delta}, d, g)$ time to identify an ε-δ-$\mathrm{M}_p(\cdot;\cdot)$-optimal $\tilde{\theta} \in \Theta_d$, and thus fair-PAC-learns $(\mathcal{H}, \ell)$. This essentially boils down to showing that (1) the empirical malfare objective is *convex* and *Lipschitz continuous*, and (2) that algorithm 2 runs sufficiently many subgradient-update steps, with appropriate step size, on a sufficiently large training set, to yield the appropriate guarantees, and that each step of the subgradient method, of which there are polynomially many, itself requires polynomial time.

First, note that by theorem 3.3.8 items 3 and 5, we may assume that $\mathrm{M}(\cdot;\cdot)$ can be expressed as a p-power mean with $p \geq 1$; thus henceforth we refer to it as $\mathrm{M}_p(\cdot;\cdot)$. Now, recall that the empirical malfare objective (given $\theta \in \Theta_d$ and training sets $\boldsymbol{z}_{1:g}$) is defined as

$$\mathrm{M}_p\big(i \mapsto \hat{\mathrm{R}}(h(\cdot;\theta); \ell, \boldsymbol{z}_i); \boldsymbol{w}\big) \enspace .$$

We first show that empirical malfare is convex in Θ_d. By assumption and positive linear closure, $\hat{\mathrm{R}}(h(\cdot;\theta'); \ell, \boldsymbol{z}_i)$ is convex in $\theta \in \Theta_d$. The objective of interest is the composition of $\mathrm{M}_p(\cdot;\boldsymbol{w})$ with this quantity evaluated on each of g training sets. By theorem 3.3.7 item 4, $\mathrm{M}_p(\cdot;\boldsymbol{w})$ is convex $\forall p \in [1,\infty]$ in $\mathbb{R}^g_{0+}$, and by the monotonicity axiom, it is monotonically increasing. Composition of a monotonically increasing convex function on $\mathbb{R}^g_{0+}$ with convex functions on Θ_d yields a convex function, thus we conclude the empirical malfare objective is convex in Θ_d.

We now show that empirical malfare is Lipschitz-continuous. Now, note that for any $p \geq 1$, $\boldsymbol{w}$,

$$\forall \mathcal{S}, \mathcal{S}' : \big|\mathrm{M}_p(\mathcal{S};\boldsymbol{w}) - \mathrm{M}_p(\mathcal{S}';\boldsymbol{w})\big| \leq 1 \big\|\mathcal{S} - \mathcal{S}'\big\|_\infty \enspace ,$$

i.e., $\mathrm{M}_p(\cdot;\boldsymbol{w})$ is 1-$\|\cdot\|_\infty$-$|\cdot|$-Lipschitz in *empirical risks* (see theorem 3.3.7 item 3), and thus by Lipschitz composition, we have Lipschitz property

$$\forall \theta, \theta' \in \Theta_d : \Big|\mathrm{M}_p\big(i \mapsto \hat{\mathrm{R}}(h(\cdot;\theta); \ell, \boldsymbol{z}_i); \boldsymbol{w}\big) - \mathrm{M}_p\big(i \mapsto \hat{\mathrm{R}}(h(\cdot;\theta'); \ell, \boldsymbol{z}_i); \boldsymbol{w}\big)\Big| \leq \lambda_\ell \lambda_{\mathcal{H}} \big\|\theta - \theta'\big\|_2 \enspace .$$

We now show that algorithm 2 FPAC-learns $(\mathcal{H}, \ell)$. As above, take $m \doteq \mathrm{m}_{\mathrm{UC}}(\frac{\varepsilon}{3}, \frac{\delta}{g}, d)$. Our algorithm shall operate on a training sample $\boldsymbol{z}_{1:g,1:m} \sim \mathcal{D}_1^m \times \cdots \times \mathcal{D}_g^m$.

First note that evaluating a subgradient (via forward finite-difference estimation or automated subdifferentiation) requires $(\dim(\Theta_d) + 1)m$ evaluations of $h(\cdot;\cdot)$, which by assumption is possible in $\mathrm{Poly}(d, m) =$

$\mathrm{Poly}(\frac{1}{\varepsilon}, \frac{1}{\delta}, d, g)$ time.

The subgradient method produces $\tilde{\theta}$ approximating the empirically-optimal $\hat{\theta}$ such that [see Shor, 2012]

$$f(\tilde{\theta}) \leq f(\hat{\theta}) + \frac{\left\|\theta_0 - \hat{\theta}\right\|_2^2 + \Lambda^2\alpha^2 n}{2\alpha n} \leq \frac{\mathrm{Diam}^2(\Theta_d) + \Lambda^2\alpha^2 n}{2\alpha n} \enspace ,$$

for Λ-$\|\cdot\|_2$-$|\cdot|$-Lipschitz objective f, thus taking $\alpha \doteq \frac{\mathrm{Diam}(\Theta_d)}{\Lambda\sqrt{n}}$ yields

$$f(\tilde{\theta}) - f(\hat{\theta}) \leq \frac{\mathrm{Diam}(\Theta_d)\Lambda}{\sqrt{n}} \enspace .$$

As shown above, $\Lambda = \lambda_\ell \lambda_{\mathcal{H}}$, thus we may guarantee *optimization error*

$$\varepsilon_{\mathrm{opt}} \doteq f(\hat{\theta}) - f(\theta^*) \leq \frac{\varepsilon}{3}$$

if we take iteration count

$$n \geq \frac{9\,\mathrm{Diam}^2(\Theta_d)\lambda_\ell^2\lambda_{\mathcal{H}}^2}{\varepsilon^2} = \left(\frac{3\,\mathrm{Diam}(\Theta_d)\lambda_\ell\lambda_{\mathcal{H}}}{\varepsilon}\right)^2 \in \mathrm{Poly}(\tfrac{1}{\varepsilon}, d) \enspace .$$

As each iteration requires $m \cdot \mathrm{Poly}(d) \subseteq \mathrm{Poly}(\frac{1}{\varepsilon}, \frac{1}{\delta}, d, g)$ time, the subgradient method identifies an $\frac{\varepsilon}{3}$-empirical-malfare-optimal $\tilde{\theta} \in \Theta_d$ in $\mathrm{Poly}(\frac{1}{\varepsilon}, \frac{1}{\delta}, d, g)$ time.

As m was selected to ensure $\frac{\varepsilon}{3}$-$\frac{\delta}{g}$ uniform convergence, we thus have that by uniform convergence, and union bound (over g groups), with probability at least $1 - \delta$ over choice of $\boldsymbol{z}_{1:g}$, we have

$$\forall i \in \{1, \dots, g\}, \theta \in \Theta_d : \left| \mathrm{M}_p(i \mapsto \hat{\mathrm{R}}(h(\cdot;\theta); \ell, \boldsymbol{z}_i); \boldsymbol{w}) - \mathrm{M}_p(i \mapsto \mathrm{R}(h(\cdot;\theta); \ell, \mathcal{D}_i); \boldsymbol{w}) \right| \leq \frac{\varepsilon}{3} \enspace .$$

Combining *estimation* and *optimization* errors, we get that with probability at least $1 - \delta$, the approximate-EMM-optimal $h(\cdot;\tilde{\theta})$ obeys

$$\begin{aligned}
\mathrm{M}_p(i \mapsto \mathrm{R}(h(\cdot;\tilde{\theta}); \ell, \mathcal{D}_i); \boldsymbol{w}) &\leq \mathrm{M}_p(i \mapsto \hat{\mathrm{R}}(h(\cdot;\tilde{\theta}); \ell, \boldsymbol{z}_i); \boldsymbol{w}) + \tfrac{\varepsilon}{3} \\
&\leq \mathrm{M}_p(i \mapsto \hat{\mathrm{R}}(h(\cdot;\hat{\theta}); \ell, \boldsymbol{z}_i); \boldsymbol{w}) + \tfrac{2\varepsilon}{3} \\
&\leq \mathrm{M}_p(i \mapsto \hat{\mathrm{R}}(h(\cdot;\theta^*); \ell, \boldsymbol{z}_i); \boldsymbol{w}) + \tfrac{2\varepsilon}{3} \\
&\leq \mathrm{M}_p(i \mapsto \mathrm{R}(h(\cdot;\theta^*); \ell, \mathcal{D}_i); \boldsymbol{w}) + \varepsilon \enspace .
\end{aligned}$$

We may thus conclude that $\mathcal{A}$ fair-PAC learns $\mathcal{H}$ with sample complexity $gm = g \cdot \mathrm{m}_{\mathrm{UC}}(\frac{\varepsilon}{3}, \frac{\delta}{g}, d)$. Furthermore, as the entire operation requires polynomial time, we have $(\mathcal{H}, \ell) \in \mathrm{PAC}^{\mathrm{Agn}}_{\mathrm{Poly}}$. $\square$

We now work towards proof of theorem 5.3.2. We begin with a technical lemma deriving relevant properties of the cover employed in the main result.

Lemma E.1.1 (Group Cover Properties)**.** Suppose loss function ℓ of bounded codomain (i.e., $\|\ell\|_\infty$ is bounded), hypothesis class $\mathcal{H} \subseteq \mathcal{X} \to \mathcal{Y}$, and per-group samples $\boldsymbol{z}_{1:g,1:m} \in (\mathcal{X} \times \mathcal{Y})^{g \times m}$, letting $\bigcirc_{i=1}^{g} \boldsymbol{z}_i$ denote their *concatenation*. Now define

$$\hat{\mathcal{C}}_{\cup(1:g)} \doteq \bigcup_{i=1}^{g} \hat{\mathcal{C}}\left(\ell \circ \mathcal{H}, \boldsymbol{z}_i, \gamma\right) \quad \& \quad \hat{\mathcal{C}}_{\circ(1:g)} \doteq \hat{\mathcal{C}}\left(\ell \circ \mathcal{H}, \bigcirc_{i=1}^{g} \boldsymbol{z}_i, \tfrac{\gamma}{\sqrt{g}}\right) \ .$$

Then, letting $\hat{\mathcal{C}}$ refer generically to either $\hat{\mathcal{C}}_{\cup(1:g)}$ or $\hat{\mathcal{C}}_{\circ(1:g)}$, the following hold.

1. If $\hat{\mathcal{C}}$ is of *minimal cardinality*, then

$$\left|\hat{\mathcal{C}}_{\cup(1:g)}\right| \leq g \mathcal{N}(\ell \circ \mathcal{H}, m, \gamma) \quad \& \quad \left|\hat{\mathcal{C}}_{\circ(1:g)}\right| \leq \mathcal{N}(\ell \circ \mathcal{H}, gm, \tfrac{\gamma}{\sqrt{g}}) \ .$$

2. $\displaystyle \sup_{\mathcal{D} \text{ over } \mathcal{X} \times \mathcal{Y}} \mathfrak{R}_m(\ell \circ \mathcal{H}, \mathcal{D}) \leq \inf_{\gamma \geq 0} \gamma + \|\ell\|_\infty \sqrt{\frac{\ln \mathcal{N}(\ell \circ \mathcal{H}, \gamma)}{2m}}.$

3. Suppose $\ln \mathcal{N}(\ell \circ \mathcal{H}, \gamma) \in \mathrm{Poly}(\frac{1}{\gamma})$. Then the uniform-convergence sample-complexity of $\ell \circ \mathcal{H}$ over g groups obeys

$$\begin{aligned} \mathrm{m}_{\mathrm{UC}}(\ell \circ \mathcal{H}, \varepsilon, \delta, g) &\doteq \operatorname{argmin}\left\{ m \,\middle|\, \sup_{\mathcal{D}_{1:g} \text{ over } (\mathcal{X} \times \mathcal{Y})^g} \mathbb{P}\left(\max_{i \in 1,\dots,g} \sup_{h \in \mathcal{H}} \left| \mathop{\mathbb{E}}_{\mathcal{D}_i}[\ell \circ h] - \mathop{\hat{\mathbb{E}}}_{\boldsymbol{z}_i \sim \mathcal{D}_i^m}[\ell \circ h] \right| > \varepsilon \right) \leq \delta \right\} \\ &\leq \frac{8\|\ell\|_\infty^2 \ln\left(\sqrt[4]{\frac{2g}{\delta}}\, \mathcal{N}(\ell \circ \mathcal{H}, \frac{\varepsilon}{4})\right)}{\varepsilon^2} \\ &\in \mathbf{O}\left(\frac{\ln \frac{g \mathcal{N}(\ell \circ \mathcal{H}, \varepsilon)}{\delta}}{\varepsilon^2} \right) \subset \mathrm{Poly}\left(\frac{1}{\varepsilon}, \exp \frac{1}{\delta}, \exp g\right) \ . \end{aligned}$$

4. For the sample $\boldsymbol{z}_i$ associated with each group $i \in 1, \dots, g$, $\hat{\mathcal{C}}$ is a γ-uniform-approximation of *empirical risk* $\hat{\mathrm{R}}(h; \ell, \boldsymbol{z}_i)$, and a γ-ℓ_2 cover of the *loss family* $\ell \circ \mathcal{H}$, as

$$\max_{i \in 1,\dots,g} \min_{h_\gamma \in \hat{\mathcal{C}}} \left| \hat{\mathrm{R}}(h; \ell, \boldsymbol{z}_i) - \hat{\mathrm{R}}(h_\gamma; \ell, \boldsymbol{z}_i) \right| \leq \max_{i \in 1,\dots,g} \min_{h_\gamma \in \hat{\mathcal{C}}} \sqrt{\frac{1}{m} \sum_{j=1}^{m} \big((\ell \circ h)(\boldsymbol{z}_{i,j}) - (\ell \circ h_\gamma)(\boldsymbol{z}_{i,j}) \big)^2} \leq \gamma \ .$$

5. $\hat{\mathcal{C}}_{\circ(1:g)}$, but not necessarily $\hat{\mathcal{C}}_{\cup(1:g)}$, *simultaneously* (across all groups) γ-uniformly-approximates *empirical risk*, and is a γ-ℓ_2 cover of the *loss family* $\ell \circ \mathcal{H}$, as

$$\min_{h_\gamma \in \hat{\mathcal{C}}_{\circ(1:g)}} \max_{i \in 1,\dots,g} \left| \hat{\mathrm{R}}(h; \ell, \boldsymbol{z}_i) - \hat{\mathrm{R}}(h_\gamma; \ell, \boldsymbol{z}_i) \right| \leq \min_{h_\gamma \in \hat{\mathcal{C}}_{\circ(1:g)}} \max_{i \in 1,\dots,g} \sqrt{\frac{1}{m} \sum_{j=1}^{m} \big((\ell \circ h)(\boldsymbol{z}_{i,j}) - (\ell \circ h_\gamma)(\boldsymbol{z}_{i,j}) \big)^2} \leq \gamma \ .$$

Proof. We first show items 1 to 3, followed by a key intermediary relating risk values and ℓ_2 distances, and close by showing items 4 and 5.

We begin with item 1. Both bounds follow directly from the definition of uniform covering numbers.

We now show item 2. This result follows via a standard sequence of operations over the Rademacher average. In particular, observe

$$\begin{aligned}
\sup_{\mathcal{D} \text{ over } \mathcal{X} \times \mathcal{Y}} \mathfrak{R}_m(\ell \circ \mathcal{H}, \mathcal{D}) &= \sup_{\mathcal{D} \text{ over } \mathcal{X} \times \mathcal{Y}} \mathop{\mathbb{E}}_{\boldsymbol{z} \sim \mathcal{D}^m} \left[\hat{\mathfrak{R}}_m(\ell \circ \mathcal{H}, \boldsymbol{z}) \right] && \textsc{Definition of } \mathfrak{R} \\
&\leq \sup_{\mathcal{D} \text{ over } \mathcal{X} \times \mathcal{Y}} \mathop{\mathbb{E}}_{\boldsymbol{z} \sim \mathcal{D}^m} \left[\inf_{\gamma \geq 0} \gamma + \hat{\mathfrak{R}}_m(\mathcal{C}^*(\ell \circ \mathcal{H}, \boldsymbol{z}, \gamma), \boldsymbol{z}) \right] && \textsc{Discretization} \\
&\leq \inf_{\gamma \geq 0} \gamma + \sup_{\mathcal{D} \text{ over } \mathcal{X} \times \mathcal{Y}} \mathop{\mathbb{E}}_{\boldsymbol{z} \sim \mathcal{D}^m} \left[\|\ell\|_\infty \sqrt{\frac{\ln \left| \mathcal{C}^*(\ell \circ \mathcal{H}, \boldsymbol{z}, \gamma) \right|}{2m}} \right] && \textsc{Massart's Inequality} \\
&\leq \inf_{\gamma \geq 0} \gamma + \|\ell\|_\infty \sqrt{\frac{\ln \mathcal{N}(\ell \circ \mathcal{H}, \gamma)}{2m}} \ , && \textsc{Definition of } \mathcal{N}
\end{aligned}$$

where the Massart's Inequality step follows via *Massart's finite class inequality* [Massart, 2000, lemma 1], and the Discretization step via *Dudley's discretization argument*.

We now show item 3. By the *symmetrization inequality*, and a 2-tailed application of McDiarmid's bounded difference inequality [McDiarmid, 1989], where changing any $\boldsymbol{z}_{i,j}$ has bounded difference $\frac{\|\ell\|_\infty}{m}$, we have that

$$\forall i : \ \mathbb{P} \left(\sup_{h \in \mathcal{H}} \left| \mathop{\mathbb{E}}_{\mathcal{D}_i} [\ell \circ h] - \mathop{\hat{\mathbb{E}}}_{\boldsymbol{z}_i \sim \mathcal{D}_i^m} [\ell \circ h] \right| > 2\mathfrak{R}_m(\ell \circ \mathcal{H}, \mathcal{D}_i) + \|\ell\|_\infty \sqrt{\frac{\ln \frac{2}{\delta}}{2m}} \right) \leq \delta$$

thus by union bound over g groups, we have

$$\mathbb{P} \left(\max_{i \in 1,\dots,g} \sup_{h \in \mathcal{H}} \left| \mathop{\mathbb{E}}_{\mathcal{D}_i} [\ell \circ h] - \mathop{\hat{\mathbb{E}}}_{\boldsymbol{z}_i \sim \mathcal{D}_i^m} [\ell \circ h] \right| > \sup_{\mathcal{D} \text{ over } \mathcal{X} \times \mathcal{Y}} 2\mathfrak{R}_m(\ell \circ \mathcal{H}, \mathcal{D}) + \|\ell\|_\infty \sqrt{\frac{\ln \frac{2g}{\delta}}{2m}} \right) \leq \delta \ .$$

Now, let *estimation error* bound $\varepsilon_{\mathrm{est}} \doteq \sup_{\mathcal{D} \text{ over } \mathcal{X}\times\mathcal{Y}} 2\mathfrak{R}_m(\ell \circ \mathcal{H}, \mathcal{D}) + \|\ell\|_\infty \sqrt{\frac{\ln \frac{2g}{\delta}}{2m}}$, and observe that via item 2,

$$\varepsilon_{\mathrm{est}} \leq \inf_{\gamma \geq 0} 2\gamma + 2\|\ell\|_\infty \sqrt{\frac{\ln \mathcal{N}(\ell \circ \mathcal{H}, \gamma)}{2m}} + \|\ell\|_\infty \sqrt{\frac{\ln \frac{2g}{\delta}}{2m}} \enspace .$$

From here, we solve for an upper-bound on sample-size m to get

$$\begin{aligned}
\mathrm{m}_{\mathrm{UC}}(\ell \circ \mathcal{H}, \varepsilon, \delta, g) \leq \inf_{\gamma \geq 0} \frac{4\|\ell\|_\infty^2 \ln \mathcal{N}(\ell \circ \mathcal{H}, \gamma) + \|\ell\|_\infty^2 \ln \frac{2g}{\delta}}{2(\varepsilon - 2\gamma)^2} &= \frac{2\|\ell\|_\infty^2 \ln\left(\sqrt[4]{\frac{2g}{\delta}}\, \mathcal{N}(\ell \circ \mathcal{H}, \gamma)\right)}{(\varepsilon - 2\gamma)^2} \\
\leq \frac{8\|\ell\|_\infty^2 \ln\left(\sqrt[4]{\frac{2g}{\delta}}\, \mathcal{N}(\ell \circ \mathcal{H}, \frac{\varepsilon}{4})\right)}{\varepsilon^2} & \qquad \textsc{Set } \gamma = \frac{\varepsilon}{4} \\
\in \mathbf{O}\left(\frac{\ln \frac{g\mathcal{N}(\ell\circ\mathcal{H},\varepsilon)}{\delta}}{\varepsilon^2}\right) \subset \mathrm{Poly}\left(\frac{1}{\varepsilon}, \exp\frac{1}{\delta}, \exp g\right) \enspace . & \qquad \mathcal{N}(\ell \circ \mathcal{H}, \varepsilon) \in \mathrm{Poly}\,\frac{1}{\varepsilon}
\end{aligned}$$

We now show an intermediary which immediately implies the left inequalities of both items 4 and 5. In particular, we may relate these empirical risk gaps to (size-normalized) ℓ_2 distance, as (for each i) we have $\forall h \in \mathcal{H}, h_\gamma \in \hat{\mathcal{C}}$ that

$$\left|\hat{\mathrm{R}}(h; \ell, \boldsymbol{z}_i) - \hat{\mathrm{R}}(h_\gamma; \ell, \boldsymbol{z}_i)\right| \leq \frac{1}{m}\sum_{j=1}^m \left|(\ell \circ h)(\boldsymbol{z}_{i,j}) - (\ell \circ h_\gamma)(\boldsymbol{z}_{i,j})\right| \leq \sqrt{\frac{1}{m}\sum_{j=1}^m \big((\ell \circ h)(\boldsymbol{z}_{i,j}) - (\ell \circ h_\gamma)(\boldsymbol{z}_{i,j})\big)^2} \enspace .$$

Here the last inequality holds since we divide m *inside the* $\sqrt{\cdot}$. The opposite inequality holds for standard ℓ_1 and Euclidean distance, where m is not divided, essentially because the ℓ_1 and ℓ_2 distances differ by up to a factor $\sqrt{m}$, but this form may be familiar as the relationship between the *mean* and *root mean square* errors. The unconvinced reader may note that this size-normalized ℓ_2 distance is in fact the (unweighted) $p = 2$ power-mean, and thus this step follows via theorem 3.3.7 item 1.

We now show the right inequality of item 4. Note that for the case of $\hat{\mathcal{C}}_{\cup(1:g)}$, the result is almost tautological, as it holds per group by the union-based construction of $\hat{\mathcal{C}}_{\cup(1:g)}$. The case of $\hat{\mathcal{C}}_{\circ(1:g)}$ is more subtle, but we defer its proof to the final item, as it then follows as an immediate consequence of the *max-min inequality*, i.e., $\forall h \in \mathcal{H}$,

$$\max_{i \in 1,\dots,g} \min_{h_\gamma \in \hat{\mathcal{C}}_{\circ(1:g)}} \sqrt{\frac{1}{m}\sum_{j=1}^m \big((\ell \circ h)(\boldsymbol{z}_{i,j}) - (\ell \circ h_\gamma)(\boldsymbol{z}_{i,j})\big)^2} \leq \min_{h_\gamma \in \hat{\mathcal{C}}_{\circ(1:g)}} \max_{i \in 1,\dots,g} \sqrt{\frac{1}{m}\sum_{j=1}^m \big((\ell \circ h)(\boldsymbol{z}_{i,j}) - (\ell \circ h_\gamma)(\boldsymbol{z}_{i,j})\big)^2} \enspace .$$

We now show item 5. Suppose (by way of contradiction) that there exists some $h \in \mathcal{H}$ such that

$$\min_{h_\gamma \in \hat{\mathcal{C}}_{\circ(1:g)}} \max_{i \in 1,\dots,g} \sqrt{\frac{1}{m}\sum_{j=1}^{m}\bigl((\ell \circ h)(\boldsymbol{z}_{i,j}) - (\ell \circ h_\gamma)(\boldsymbol{z}_{i,j})\bigr)^2} > \gamma \enspace .$$

One then need only consider the summands associated with a maximal i to observe that this implies

$$\min_{h_\gamma \in \hat{\mathcal{C}}_{\circ(1:g)}} \sqrt{\frac{1}{mg}\sum_{i=1}^{g}\sum_{j=1}^{m}\bigl((\ell \circ h)(\boldsymbol{z}_{i,j}) - (\ell \circ h_\gamma)(\boldsymbol{z}_{i,j})\bigr)^2} > \frac{\gamma}{\sqrt{g}} \enspace ,$$

thus $\hat{\mathcal{C}}_{\circ(1:g)}$ is not a $\frac{\gamma}{\sqrt{g}}$-ℓ_2 cover of $\bigcirc_{i=1}^{g}\boldsymbol{z}_i$, which contradicts its very definition. We thus conclude

$$\forall h \in \mathcal{H} : \min_{h_\gamma \in \hat{\mathcal{C}}_{\circ(1:g)}} \max_{i \in 1,\dots,g} \sqrt{\frac{1}{m}\sum_{j=1}^{m}\bigl((\ell \circ h)(\boldsymbol{z}_{i,j}) - (\ell \circ h_\gamma)(\boldsymbol{z}_{i,j})\bigr)^2} \leq \gamma \enspace .$$

□

With lemma E.1.1 in hand, we are now ready to show theorem 5.3.2.

Theorem 5.3.2 (Efficient FPAC-Learning by Covering)**.** Suppose loss function ℓ of bounded codomain (i.e., $\|\ell\|_\infty$ is bounded), and hypothesis class sequence $\mathcal{H}$, s.t. $\forall m, d \in \mathbb{N}$, $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^m$, there exist

1. a γ-ℓ_2 cover $\mathcal{C}^*(\ell \circ \mathcal{H}_d, \boldsymbol{z}, \gamma)$, where $\left|\mathcal{C}^*(\ell \circ \mathcal{H}_d, \boldsymbol{z}, \gamma)\right| \leq \mathcal{N}(\ell \circ \mathcal{H}_d, \gamma) \in \mathrm{Poly}(\frac{1}{\gamma}, d)$; and
2. an algorithm to enumerate a γ-ℓ_2 cover $\hat{\mathcal{C}}(\ell \circ \mathcal{H}_d, \boldsymbol{z}, \gamma)$ of size $\mathrm{Poly}\,\mathcal{N}(\ell \circ \mathcal{H}_d, m, \gamma)$ in $\mathrm{Poly}(m, \frac{1}{\gamma}, d)$ time.

Then $(\ell, \mathcal{H}) \in \mathrm{FPAC}_{\mathrm{Poly}}$, where (1) is sufficient to show that $(\ell, \mathcal{H})$ is fair-PAC learnable with polynomial *sample complexity*, and (2) is required only to show polynomial *training time complexity*.

Proof. We now constructively show the existence of a fair-PAC-learner $\mathcal{A}$ for $(\ell, \mathcal{H})$ over domain $\mathcal{X}$ and codomain $\mathcal{Y}$. As in theorem 5.3.1, we first note that by theorem 3.3.8 items 3 and 5, under the conditions of FPAC learning, this reduces to showing that we can learn any malfare concept $\mathrm{M}_p(\cdot;\cdot)$ that is a p-power mean with $p \geq 1$.

We first assume a training sample $\boldsymbol{z}_{1:g,1:m} \sim \mathcal{D}_1^m \times \cdots \times \mathcal{D}_g^m$, i.e., a collection of m draws from each of the g groups. In particular, we shall select m to guarantee that the *estimation error* for the malfare does not exceed $\frac{\varepsilon}{3}$ with probability at least $1 - \delta$, i.e., we require that with said probability,

$$\varepsilon_{\mathrm{est}} \doteq \left|\mathrm{M}_p(i \mapsto \hat{\mathrm{R}}(h; \ell, \boldsymbol{z}_i); \boldsymbol{w}) - \mathrm{M}_p(i \mapsto \mathrm{R}(h; \ell, \mathcal{D}_i); \boldsymbol{w})\right| \leq \frac{\varepsilon}{3} \enspace .$$

Now, note that by theorem 3.3.7 item 3 (contraction), we have

$$\left|\mathrm{M}_p(i \mapsto \hat{\mathrm{R}}(h;\ell,\boldsymbol{z}_i);\boldsymbol{w}) - \mathrm{M}_p(i \mapsto \mathrm{R}(h;\ell,\mathcal{D}_i);\boldsymbol{w})\right| \le \max_{i\in 1,\dots,g} \sup_{h\in\mathcal{H}_d} \left|\mathrm{R}(h;\ell,\mathcal{D}_i);\boldsymbol{w}) - \hat{\mathrm{R}}(h;\ell,\boldsymbol{z}_i);\boldsymbol{w})\right| \ ,$$

and by lemma E.1.1 item 3, a sample of size

$$m = \left\lceil \frac{81\|\ell\|_\infty^2 \ln\left(\sqrt[4]{\frac{2g}{\delta}}\mathcal{N}(\ell\circ\mathcal{H},\frac{\varepsilon}{12})\right)}{\varepsilon^2} \right\rceil \in \mathbf{O}\left(\frac{\ln\frac{g\mathcal{N}(\ell\circ\mathcal{H},\varepsilon)}{\delta}}{\varepsilon^2}\right) \subset \mathrm{Poly}\left(\frac{1}{\varepsilon}, \exp\frac{1}{\delta}, \exp g\right)$$

suffices to ensure that

$$\mathbb{P}\left(\max_{i\in 1,\dots,g} \sup_{h\in\mathcal{H}_d} \left|\mathrm{R}(h;\ell,\mathcal{D}_i);\boldsymbol{w}) - \hat{\mathrm{R}}(h;\ell,\boldsymbol{z}_i);\boldsymbol{w})\right| > \frac{\varepsilon}{3}\right) \le \delta \ ,$$

thus guaranteeing the stated estimation error bound.

With our sample size and *estimation error* guarantee, we now define the *learning algorithm* and bound its *optimization error*. Take *cover precision* $\gamma \doteq \frac{\varepsilon}{3\sqrt{g}}$. By assumption, for each $d \in \mathbb{N}$, we may enumerate a γ-cover $\hat{\mathcal{C}}(\ell\circ\mathcal{H}_d, \bigcirc_{i=1}^{g}\boldsymbol{z}_i, \gamma)$, where $\bigcirc_{i=1}^{g}\boldsymbol{z}_i$ denotes the *concatenation* of each $\boldsymbol{z}_i$, in $\mathrm{Poly}(gm, \frac{1}{\gamma}, d) = \mathrm{Poly}(\frac{1}{\varepsilon}, \frac{1}{\delta}, d, g)$ time. For the remainder of this proof, we refer to this cover as $\hat{\mathcal{C}}$.

Now, we take the learning algorithm to be *empirical malfare minimization* over $\hat{\mathcal{C}}$. Let

$$\hat{h} \doteq \operatorname*{argmin}_{h_\gamma\in\hat{\mathcal{C}}} \mathrm{M}_p(i \mapsto \hat{\mathrm{R}}(h_\gamma;\ell,\boldsymbol{z}_i);\boldsymbol{w}) \quad \& \quad \tilde{h} \doteq \operatorname*{argmin}_{h\in\mathcal{H}_d} \mathrm{M}_p(i \mapsto \hat{\mathrm{R}}(h;\ell,\boldsymbol{z}_i);\boldsymbol{w}) \ ,$$

where ties may be broken arbitrarily. Note that via standard covering properties, that the *optimization error* is bounded as

$$\begin{aligned}
\varepsilon_{\mathrm{opt}} &\doteq \mathrm{M}_p\big(i \mapsto \hat{\mathrm{R}}(\hat{h};\ell,\boldsymbol{z}_i);\boldsymbol{w}\big) - \mathrm{M}_p\big(i \mapsto \hat{\mathrm{R}}(\tilde{h};\ell,\boldsymbol{z}_i);\boldsymbol{w}\big) && \textsc{Definition} \\
&= \sup_{h\in\mathcal{H}_d} \min_{h_\gamma\in\hat{\mathcal{C}}} \mathrm{M}_p\big(i \mapsto \hat{\mathrm{R}}(h_\gamma;\ell,\boldsymbol{z}_i);\boldsymbol{w}\big) - \mathrm{M}_p\big(i \mapsto \hat{\mathrm{R}}(h;\ell,\boldsymbol{z}_i);\boldsymbol{w}\big) && \begin{array}{r}\textsc{Properties of Suprema} \\ \textsc{Definition of } \hat{h}, \tilde{h}\end{array} \\
&\le \sup_{h\in\mathcal{H}_d} \min_{h_\gamma\in\hat{\mathcal{C}}} \max_{i\in\{1,\dots,g\}} \left|\hat{\mathrm{R}}(h_\gamma;\ell,\boldsymbol{z}_i) - \hat{\mathrm{R}}(h;\ell,\boldsymbol{z}_i)\right| && \textsc{Item 3 (Contraction)} \\
&\le \sqrt{g}\gamma = \frac{\varepsilon}{3} \ . && \textsc{Lemma E.1.1 Item 5}
\end{aligned}$$

This controls for *optimization error* between the true and approximate EMM solutions $\tilde{h}$ and $\hat{h}$.

We now combine the optimization and estimation error inequalities, letting

$$h^* \doteq \operatorname*{argmin}_{h \in \mathcal{H}_d} \mathrm{M}_p(i \mapsto \mathrm{R}(h; \ell, \mathcal{D}_i); \boldsymbol{w}) \enspace ,$$

denote the true *malfare optimal* solution, over *distributions* rather than *samples*, breaking ties arbitrarily. We then derive

$$\begin{aligned}
&\mathrm{M}_p(i \mapsto \mathrm{R}(h^*; \ell, \mathcal{D}_i); \boldsymbol{w}) - \mathrm{M}_p(i \mapsto \mathrm{R}(\hat{h}; \ell, \mathcal{D}_i); \boldsymbol{w}) & \\
&\quad = \Big(\mathrm{M}_p(i \mapsto \mathrm{R}(h^*; \ell, \mathcal{D}_i); \boldsymbol{w}) - \mathrm{M}_p(i \mapsto \hat{\mathrm{R}}(h^*; \ell, \boldsymbol{z}_i); \boldsymbol{w}) \Big) & \leq \varepsilon_{\mathrm{est}} \\
&\quad\quad + \Big(\mathrm{M}_p(i \mapsto \hat{\mathrm{R}}(h^*; \ell, \boldsymbol{z}_i); \boldsymbol{w}) - \mathrm{M}_p(i \mapsto \hat{\mathrm{R}}(\tilde{h}; \ell, \boldsymbol{z}_i); \boldsymbol{w}) \Big) & \leq 0 \\
&\quad\quad + \Big(\mathrm{M}_p(i \mapsto \hat{\mathrm{R}}(\tilde{h}; \ell, \boldsymbol{z}_i); \boldsymbol{w}) - \mathrm{M}_p(i \mapsto \hat{\mathrm{R}}(\hat{h}; \ell, \boldsymbol{z}_i); \boldsymbol{w}) \Big) & \leq \varepsilon_{\mathrm{opt}} \\
&\quad\quad + \Big(\mathrm{M}_p(i \mapsto \hat{\mathrm{R}}(\hat{h}; \ell, \boldsymbol{z}_i); \boldsymbol{w}) - \mathrm{M}_p(i \mapsto \mathrm{R}(\hat{h}; \ell, \mathcal{D}_i); \boldsymbol{w}) \Big) & \leq \varepsilon_{\mathrm{est}} \\
&\quad \leq \varepsilon_{\mathrm{opt}} + 2\varepsilon_{\mathrm{est}} = \frac{\varepsilon}{3} + \frac{2\varepsilon}{3} = \varepsilon \enspace . & \textsc{See Above}
\end{aligned}$$

We thus conclude that this algorithm produces an ε-$\mathrm{M}_p(\cdot;\cdot)$ optimal solution with probability at least $1 - \delta$, and furthermore both the sample complexity and time complexity of this algorithm are $\mathrm{Poly}(\frac{1}{\varepsilon}, \frac{1}{\delta}, g, d)$. Hence, as we have constructed a polynomial-time fair-PAC learner for $(\mathcal{H}, \ell)$, we may conclude $(\mathcal{H}, \ell) \in \mathrm{FPAC}_{\mathrm{Poly}}$.

$\square$

Appendix F

Supplementary Material for Chapter 7

F.1 A Compendium of Missing Proofs

Definition F.1.1. Consider $\boldsymbol{X}_{1:T} : X_1, X_2, \ldots X_T$. We define C_i to be the autocovariace of two steps of $\mathcal{M}$ (at stationarity) being i apart. i.e., $C_i \doteq \mathrm{Cov}(X_1, X_{1+i})$.

Note that for i.i.d samples the autocovariance of any pair of samples X_i and X_j is zero. We do not have independence here nevertheless for reversible Makrov chains (See for example Equation 12.9 from Levin and Peres [2017]) we have $C_i \leq \lambda^i \sqrt{\mathbb{V}_\pi[f]\,\mathbb{V}_\pi[f]} = \lambda^i v$. The following lemmas will be used throughout:

Lemma F.1.2. Suppose $\boldsymbol{X}_{1:T} : X_1, X_2, \ldots X_T$ is a trace of length T of $\mathcal{M}$. using the above definition for C_i, the trace variance are related as

$$v_T = \frac{1}{T} v_\pi + \frac{2}{T^2} \sum_{i=1}^{T-1} (T-i) C_i \ . \tag{69}$$

Proof. The trace variance is

$$v_T = \mathbb{E}\left[\left(\frac{1}{T}\sum_{i=1}^{T}(f(X_i)-\mu)\right)^2\right] = \mathbb{E}\left[\frac{1}{T^2}\sum_{i=1}^{T}\sum_{j=1}^{T}(f(X_i)-\mu)(f(X_j)-\mu)\right] = \frac{1}{T}v_\pi + \frac{2}{T^2}\sum_{i=1}^{T-1}(T-i)C_i.$$

□

Lemma F.1.3. For any lazy reversible chain $\mathcal{M}$, we have $v_\pi \leq 2v_2 \leq 3v_3 \leq \ldots$, i.e, for any T and T' satisfying $T < T'$ we have, $Tv_T \leq T'v_{T'}$.

Lemma F.1.4. For $1 \leq i \leq \frac{n}{2}$, let f_i be defined as in lemma 7.2.3, we have $v_T(f_i) \leq \Theta({}^{i^2}\!/\!{}_2)$.

Proof. We first prove that $v_T(f_i) \leq \Theta(i^2)/T$. Note that the function $x \mapsto x \bmod 2i$ partitions the set $[n]$ to $2i$ partitions, and the relaxation time of the projection chain $\mathcal{M}_{f_i}$ is $\Theta(i^2)$.

We first assume WLOG that i divides 3: this is so we may analyze an convenient integer-length trace, but the remaining cases hold with constant-factor differences. Note that $\tau_{\mathrm{rel}} \in \boldsymbol{\Theta}(n^2)$, and our proof operates by analyzing I-traces, for $I \doteq {}^{i^2}\!/\!{}_9$, and then applying trace-variance inequalities to draw the desired conclusions.

Assume $X_1, X_2, \ldots, X_{i^2}$ is a trace of the unbiased walk on the cycle, and define $Y_k = X_{k+1} - X_k$. Thus, $X_k = \sum_{j=1}^{k-1} Y_j + s_0$ where s_0 is the starting point. Note that Y_j is a symmetric random variable (thus has equal mean and median), which we shall use to apply Lévy's inequality. The proof strategy here is to lower-bound $v_{{i^2}/9}$ by showing that a constant fraction of stationary traces in $\mathcal{M}^{({i^2}/9)}$ see only one color. We call such traces *homogeneous*, and note that for such traces $\boldsymbol{X}$, we have $(f_{\mathrm{avg}}(\boldsymbol{X}) - \mu)^2 = \frac{1}{4}$ (as $f_{\mathrm{avg}}(\boldsymbol{X}) \in \{0, 1\}$, and $\mu = \frac{1}{2}$), and from there we bound trace variances as appropriate.

We begin with a key step in deriving a lower-bound on the proportion of such homogeneous traces. In particular, take $S_0 \doteq \{{}^{i}\!/\!{}_3 + 1, {}^{i}\!/\!{}_3 + 2, \ldots, {}^{2i}\!/\!{}_3\} \subseteq [n]$, and similarly take $S_k \doteq \{ki + {}^{i}\!/\!{}_3 + 1, ki + {}^{i}\!/\!{}_3 + 2, \ldots, ki + {}^{2i}\!/\!{}_3\} \subseteq [n]$, in other words each S_k is the *middle third* of the kth contiguous color region. Finally, take $S \doteq \bigcup_{k=1}^{n/i} S_k$. These regions are depicted graphically in figure 7.2.

Let $\mathrm{SD}(\boldsymbol{Y}) \doteq \max\limits_{k \in 1, \ldots, {i^2}/9} \left| \sum\limits_{j=1}^{I} Y_k \right|$, i.e., $\mathrm{SD}(\boldsymbol{Y}) = \max\limits_{k \in 1, \ldots, {i^2}/9} \Delta_\circ(X_1, X_k)$, for $\Delta_\circ$ the shortest-path distance on the cycle, and observe

$$\begin{aligned}
\mathbb{P}\left(\mathrm{SD}(\boldsymbol{Y}) \geq \frac{i}{3}\right) &\leq 2\,\mathbb{P}\left(\left|\sum_{j=1}^{{i^2}/9} Y_j\right| \geq \frac{i}{3}\right) && \text{Lévy's inequality} \\
&\leq 4\exp\left(\frac{-2\left(\frac{i}{3}\right)^2}{i^2/9}\right) && \text{Hoeffding's inequality} \\
&= 4\exp(-2)\ .
\end{aligned}$$

From here, we decompose the trace variance and bound it as

$$
\begin{aligned}
v_I &= \mathop{\mathbb{E}}_{\boldsymbol{Y}} \left(\frac{1}{i^2/9} \sum_{j=1}^{i^2/9} Y_j - \mu \right)^2 && \textsc{By Definition} \\
&\geq \mathbb{P}(X_1 \in S)\, \mathbb{P}\Big(\mathrm{SD}(\boldsymbol{Y}) \leq \frac{i}{3} \Big| X_1 \in S \Big) \mathop{\mathbb{E}}_{\boldsymbol{Y} | \mathrm{SD}(\boldsymbol{Y}) \leq \frac{i}{3}} \left(\frac{1}{i^2/9} \sum_{j=1}^{i^2/9} Y_j - \mu \right)^2 && \textsc{Law of Total Expectation} \\
&\geq \frac{1}{3}\big(1 - 4\exp(-2)\big)\frac{1}{4} > 0.03822 \ . && \mathbb{P}(X_1 \in S) = \tfrac{1}{3} \quad \textsc{See Above} \quad (0-\tfrac{1}{2})^2 = (1-\tfrac{1}{2})^2 = \tfrac{1}{4}
\end{aligned}
$$

We thus conclude $v_I \in \Theta(1)$, therefore via lemma F.1.3, $v_T \in \boldsymbol{\Theta}(i^2/T)$. □

F.1.1 Correctness and Efficiency of DynaMITE

A Note on Nonstationarity

Note that theorems 7.2.5 and 7.2.6 assume *stationarity* i.e., they consider traces $\boldsymbol{X}_{1:m} : X_1, X_2, \ldots X_m$ assuming $X_1 \sim \pi$. This assumption is often prohibitive, as drawing even a single such sample can be computationally infeasible. We overcome this problem using the following equation for $\boldsymbol{X}_{1:m} : X_1, X_2, \ldots X_m$, $X_1 \sim \nu$.

$$
\mathop{\mathbb{P}}_{\substack{\boldsymbol{X} \\ X_1 \sim \nu}} \big(|\hat{\mu} - \mu| \geq \varepsilon \big) \leq \left\| \frac{\partial \nu}{\partial \pi} \right\|_{\pi,\infty} \mathop{\mathbb{P}}_{\substack{\boldsymbol{X} \\ X_1 \sim \pi}} \big(|\hat{\mu} - \mu| \geq \varepsilon \big) \quad ; \tag{70}
$$

see, e.g., [Fan et al., 2018], proof of theorem 2.3. Note that by eq. 70, it is sufficient to have $\left\| \frac{\partial \nu}{\partial \pi} \right\|_{\pi,\infty} = \operatorname{ess\,sup}_{\omega \in \Omega} |\frac{\nu(x)}{\pi(x)}| \in \mathcal{O}(1)$. This can generally be accomplished straightforwardly with a *warm-start* by selecting an arbitrary fixed $\omega \in \Omega$, taking ν to be the distribution reached after running $\mathcal{M}$ for $\tau_{\mathrm{rel}}(\mathcal{M}) \ln(1/\pi_{\min})$ steps (see, e.g., Levin and Peres [2017]). With this in mind, for simplicity we assume stationarity, knowing that our proofs and algorithm generalize with trivial modifications.

Stationarity and mixing time of trace chain $\mathcal{M}^{(T)}$

Lemma F.1.5 (Trace Chain Mixing)**.** Having a Markov chain $\mathcal{M}$ with stationary distribution π and mixing time τ_{mix}, the trace chain $\mathcal{M}^{(T)}$'s stationary distribution is $\pi^{(T)}$ where $\pi^{(T)}(\boldsymbol{a}) \doteq \pi(a_1) \prod_{i=1}^{k-1} \mathcal{M}(a_i, a_{i+1})$ and its mixing time is bounded as $\tau_{\mathrm{mix}}(\mathcal{M}^{(T)}) \leq 1 + \frac{1}{T}\tau_{\mathrm{mix}}(\mathcal{M})$. In particular for $T \geq \tau_{\mathrm{mix}}(\mathcal{M})$, we have that the second largest eigenvalue of $\mathcal{M}^{(T)}$ is at most $4/5$.

Proof of Lemma F.1.5. We first remark that it is not difficult to prove by induction that

$$\left(\mathcal{M}^{(T)}\right)^j(\boldsymbol{a},\boldsymbol{b}) = \mathcal{M}^{(j-1)T+1}(a_T, b_1)\prod_{i=1}^{T-1}\mathcal{M}(b_i, b_{i+1}).$$

It is not difficult to see that with the above definitions, for any $\boldsymbol{a}$ a trace of length T we have: $\pi^{(T)}(\boldsymbol{a}) = \sum_{\boldsymbol{b}\in\Omega^T}\pi^{(T)}(\boldsymbol{a},\boldsymbol{b})\mathcal{M}^{(T)}(\boldsymbol{a},\boldsymbol{b})$.

We now show that $\mathcal{M}^{(T)}$ is close to $\pi^{(T)}$ after $(\tau(\varepsilon)/T)+1$ steps . Let $X_0, X_1, X_2 \ldots,$ be the trace of the original random walk, thus each X_i is a random variable having distribution $X_i \sim \mathcal{M}^i\mu$. We partition this trace into blocks of length T as follows: $B^j = (X_j, X_{j+1}, \ldots X_{j+T-1})$. Thus the trace of $\mathcal{M}^{(T)}$ will be $B^0, B^1, B^2, \ldots.$

Since the mixing time of $\mathcal{M}$ is τ_{mix}, we have: for any starting distribution ν over Ω, and $\tau \geq \tau_{\text{mix}}\log(\varepsilon)$, $\text{TVD}(\mathcal{M}^\tau(\nu), \pi) \leq \varepsilon$. Assume $\mathcal{M}^{(T)}$ has started at initial distribution ν', looking at the distribution of $B_{\frac{\tau(\varepsilon)}{T}+1}$ we will have:

$$\begin{aligned}
\text{TVD}(B_{\frac{\tau}{T}+1}, \pi^{(T)}) &= \frac{1}{2}\sum_{a\in\Omega^T}\left|(\mathcal{M}^{(T)})^{\frac{\tau}{T}+1}\nu'(a) - \pi^{(T)}(a)\right| \\
&= \frac{1}{2}\sum_{a\in\Omega^T}\left|(\mathcal{M}^\tau)(\nu'_T, a_1)\prod_{i=1}^{T-1}\mathcal{M}(a_i, a_{i+1}) - \pi^{(T)}(a)\right| \\
&= \frac{1}{2}\sum_{a\in\Omega^T}\left|(\mathcal{M}^\tau)(\nu'_T, a_1)\prod_{i=1}^{T-1}\mathcal{M}(a_i, a_{i+1}) - \pi(a_1)\prod_{i=1}^{T-1}\mathcal{M}(a_i, a_{i+1})\right| \\
&\leq \varepsilon\prod_{i=1}^{T-1}\mathcal{M}(a_i, a_{i+1}) \leq \varepsilon.
\end{aligned}$$

The bound on λ then follows since $\left(\tau_{\text{rel}}(\mathcal{M}) - 1\right)\ln(2) \leq \tau_{\text{mix}}(\mathcal{M}) \leq \tau_{\text{rel}}(\mathcal{M})\ \ln\!\left(\frac{2}{\sqrt{\pi_{\min}}}\right)$.

$\square$

Trace Variance Estimation

Lemma 7.3.5 (Trace Variance Estimation)**.** Let $\hat{v}$ be defined as in definition 7.3.4, it then holds that, for any $\delta \in (0,1)$, we have

$$\mathbb{P}\left(v_\pi \geq \hat{v} + \frac{(11+\sqrt{21})(1+\frac{\lambda}{\sqrt{21}})r^2 \ln\frac{1}{\delta}}{(1-\lambda)m} + \sqrt{\frac{(1+\lambda)r^2\hat{v}\ln\frac{1}{\delta}}{(1-\lambda)m}} \right) \leq \delta \enspace .$$

Proof. The strategy here is to apply a Bernstein-type inequality to the *empirical variance estimate* $\hat{v}$. As Bernstein's inequality bounds an *expectation*, we construct a *tensor product chain* $\mathcal{M} \otimes \mathcal{M}$, and define a function $g : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$, such that $\mathbb{E}_{\mathcal{M}\times\mathcal{M}}[g] = \mathbb{V}_{\mathcal{M}} f$.

In particular, we first note that by the tensor product chain rule (see Ex. 12.6 of [Levin and Peres, 2017]), the *spectral gap* $(1-\lambda)$ of $\mathcal{M} \otimes \mathcal{M}$ equals that of $\mathcal{M}$. We now define the function $g : \mathcal{X} \times \mathcal{X} \mapsto [0, \frac{1}{2}r^2]$ as $g(x_1, x_2) \doteq \frac{1}{2}(x_1 - x_2)^2$.

We first show that g is an *unbiased estimator* of the variance of f, i.e., $\mathbb{E}_{\mathcal{M}\times\mathcal{M}}[g] = \mathbb{V}_{\mathcal{M}}[f]$:

$$\begin{aligned}
\mathbb{E}[g] &= \mathbb{E}\left[\frac{1}{2}(X_1 - X_2)^2\right] && \textsc{Definition of } g \\
&= \frac{1}{2}\,\mathbb{E}\left[(X_1 - X_2)^2\right] && \textsc{Stationarity} \\
&= \frac{1}{2}\left(\mathbb{E}[X_1^2 + X_2^2] - 2\,\mathbb{E}[X_1 X_2]\right) && \textsc{Linearity} \\
&= \mathbb{E}[X_1^2] - (\mathbb{E}[X_1])^2 && \textsc{Independence} \\
&= \mathbb{V}[X_1] = v_\pi && \textsc{Variance Properties}
\end{aligned}$$

We now seek to apply the Bernstein bound to g on $\mathcal{M} \otimes \mathcal{M}$. Note that the *range* of g is $[0, \frac{1}{2}r^2]$. We require also a bound on the *variance of the variance* $\mathbb{V}[g]$. We break the infinite regress here by noting that

$$\begin{aligned}
\mathbb{V}[g] &= \mathbb{E}[g^2] - (\mathbb{E}[g])^2 \\
&\leq \mathbb{E}[g \cdot g] \\
&\leq \mathbb{E}[g\tfrac{1}{2}r^2] \\
&= \tfrac{1}{2}r^2\,\mathbb{E}[g] \enspace . \\
&= \tfrac{1}{2}r^2 v_\pi \enspace .
\end{aligned}$$

Note that this is effectively the argument of Bernstein's inequality that removes higher moments (beyond the second) from the exponential-sum decomposition in the Chernoff-MGF bound.

Applying the Bernstein inequality (theorem 7.2.6) then yields

$$\mathbb{P}\left(v_\pi \geq \hat{v} + \varepsilon\right) \leq \delta \ ,$$

for

$$\begin{aligned} \varepsilon &\leq \frac{10r^2 \ln \frac{1}{\delta}}{2(1-\lambda)m} + \sqrt{\frac{2(1+\lambda)\,\mathbb{V}[g] \ln \frac{1}{\delta}}{(1-\lambda)m}} && \text{Range}(g) = \tfrac{1}{2}r^2 \\ &\leq \frac{5r^2 \ln \frac{1}{\delta}}{(1-\lambda)m} + \sqrt{\frac{(1+\lambda)r^2 v_\pi \ln \frac{1}{\delta}}{(1-\lambda)m}} && \mathbb{V}[g] \leq \tfrac{1}{2}r^2 v_\pi \end{aligned}$$

We now seek a form that depends only on the *empirical variance*, which we derive via the quadratic formula.

$$\begin{aligned} \varepsilon &\leq \frac{(11+\lambda)r^2 \ln \frac{1}{\delta}}{2(1-\lambda)m} + \sqrt{\left(\frac{(1+\lambda)r^2 \ln \frac{1}{\delta}}{2(1-\lambda)m}\right)^2 + \frac{5r^2 \ln \frac{1}{\delta}}{(1-\lambda)m} \cdot \frac{(1+\lambda)r^2 \ln \frac{1}{\delta}}{(1-\lambda)m} + \frac{(1+\lambda)r^2 \hat{v} \ln \frac{1}{\delta}}{(1-\lambda)m}} && \textsc{Quadratic Formula} \\ &= \frac{(11+\lambda)r^2 \ln \frac{1}{\delta}}{2(1-\lambda)m} + \sqrt{\left(\frac{(1+\lambda)r^2 \ln \frac{1}{\delta}}{2(1-\lambda)m}\right)^2 + \left(\frac{\sqrt{5(1+\lambda)}r^2 \ln \frac{1}{\delta}}{(1-\lambda)m}\right)^2 + \frac{(1+\lambda)r^2 \hat{v} \ln \frac{1}{\delta}}{(1-\lambda)m}} && \textsc{Algebra} \\ &= \frac{(11+\lambda)r^2 \ln \frac{1}{\delta}}{2(1-\lambda)m} + \sqrt{((2\sqrt{5})^2 + (1+\lambda))\left(\frac{\sqrt{1+\lambda}r^2 \ln \frac{1}{\delta}}{2(1-\lambda)m}\right)^2 + \frac{(1+\lambda)r^2 \hat{v} \ln \frac{1}{\delta}}{(1-\lambda)m}} && ((2\sqrt{5})^2 + (1+\lambda)) = 21 + \lambda \\ &\leq \frac{(11+\lambda)r^2 \ln \frac{1}{\delta}}{2(1-\lambda)m} + \sqrt{\left(\frac{(\sqrt{21} + \frac{11}{\sqrt{21}}\lambda)r^2 \ln \frac{1}{\delta}}{2(1-\lambda)m}\right)^2 + \frac{(1+\lambda)r^2 \hat{v} \ln \frac{1}{\delta}}{(1-\lambda)m}} && \sqrt{21+\lambda}\sqrt{1+\lambda} \leq \sqrt{21} + \frac{11}{\sqrt{21}}\lambda \\ &\leq \frac{(11 + \sqrt{21} + (1 + \frac{11}{\sqrt{21}})\lambda)r^2 \ln \frac{1}{\delta}}{2(1-\lambda)m} + \sqrt{\frac{(1+\lambda)r^2 \hat{v} \ln \frac{1}{\delta}}{(1-\lambda)m}} && \sqrt{a^2+b} \leq a + \sqrt{b} \\ &= \frac{((11+\sqrt{21})(1 + \frac{\lambda}{\sqrt{21}})r^2 \ln \frac{1}{\delta}}{2(1-\lambda)m} + \sqrt{\frac{(1+\lambda)r^2 \hat{v} \ln \frac{1}{\delta}}{(1-\lambda)m}} \ . && \textsc{Algebra} \end{aligned}$$

This concludes the result. □

As an intermediary, we obtain the following, which shall prove useful in deriving efficiency guarantees.

Corollary F.1.6 (Trace Variance Estimation)**.**

$$\mathbb{P}\left(v_\pi \geq \hat{v} + \frac{5r^2 \ln \frac{1}{\delta}}{(1-\lambda)m} + \sqrt{\frac{(1+\lambda)r^2 v_\pi \ln \frac{1}{\delta}}{(1-\lambda)m}}\right) \leq \delta \ .$$

DynaMITE's correctness and efficiency

Theorem 7.3.1 (Correctness of DYNAMITE)**.** Consider a Markov chain $\mathcal{M}$ over Ω, and assume its second absolute eigenvalue is less than Λ. Assume we have function $f : \Omega \to [a, b]$ with $r \doteq b - a$, and we want to estimate *true mean*, $\mu \doteq \mathbb{E}_\pi[f]$ with precision $\varepsilon > 0$, and confidence $(1 - \delta) \in (0, 1)$.

If we start at stationarity, i.e., $(x_0, x_1) \sim \pi(\mathcal{M} \otimes \mathcal{M})$, we take *mean estimate* $\hat{\mu}$ as either

1. $\hat{\mu} \leftarrow$ MCMCPRO$((x_0, x_1), \mathcal{M}, \Lambda, f, \varepsilon, \delta)$; or
2. $\hat{\mu} \leftarrow$ DYNAMITE$((x_0, x_1), \mathcal{M}, \Lambda, f, \varepsilon, \delta)$ for lazy $\mathcal{M}$.

More generally, for a nonstationary start from any $\omega \in \mathrm{Support}(\pi(\mathcal{M}))$, given minimum supported stationary probability at least $\pi_{\min}$, we may take

3. $\hat{\mu} \leftarrow$ WARMSTARTDYNAMITE$(\omega, \mathcal{M}, \Lambda, \pi_{\min}, f, \varepsilon, \delta)$ for lazy, reversible $\mathcal{M}$.

In all cases, with probability at least $1 - \delta$, we have $|\hat{\mu} - \mu| \leq \varepsilon$.

Proof. We first show (1), i.e., correctness of MCMCPRO, and then show that correctness of WARMSTARTDYNAMITE easily follows from that of DYNAMITE, which follows from that of MCMCPRO.

We now show claim (1). First note that in initialization (independent of any sampling), MCMCPROcomputes *iteration count* N and *initial sample size* α (line 4), which determine the *schedule* of *sample sizes* and *probabilistic bounds*. Over the course of the algorithm, at each of N timesteps, a 1-tail (upper) bound on *variance* $\boldsymbol{u}_i$ is computed (line 12), and a 2-tail bound on *mean* $\boldsymbol{\mu}_i$ is computed (line 13), for a total of $3N$ tail bounds. Each tail is bounded with probability $1 - \frac{\delta}{3N}$, thus by union bound, all hold simultaneously with probability $1 - \delta$. We assume henceforth that app tail bounds hold, thus all conclusions are thus qualified as holding *with probability* $\geq 1 - \delta$. Now, note that termination (line 14) occurs for one of two reasons: either $i = N$, or $\hat{\boldsymbol{\varepsilon}}_i \leq \varepsilon$. We analyze these cases separately.

In case 1, we have termination at $i < N$, i.e., $\hat{\boldsymbol{\varepsilon}}_i \leq \varepsilon$. As assumed above, all tail bounds at each iteration hold by union bound. Thus for the sample drawn at iteration i, in particular the *variance bounds* (line 12) hold, as $\hat{\boldsymbol{v}}_i$ is an *unbiased estimate* of v_T, and by lemma 7.3.5, w.h.p., $v_T \leq \boldsymbol{u}_i$. Similarly, the *mean bounds* (13) hold via theorem 7.2.6 (noting that, by *averaging over* a pair of *independent chains*, the variance proxy of interest is $\frac{1}{2}v_T$). Note that while both tail bounds are taken over the tensor-product chain $\mathcal{M} \otimes \mathcal{M}$, it holds that $\lambda(\mathcal{M} \otimes \mathcal{M}) = \lambda(\mathcal{M})$, so the bound remains valid.

Consequently, when it holds that $\hat{\boldsymbol{\varepsilon}}_i \leq \varepsilon$, the algorithm returns the estimate $\hat{\boldsymbol{\mu}}_i$ (line 15), which by the above is sufficiently accurate to satisfy the stated guarantees.

We now consider case 2, wherein we have termination at step $i = N$. When this occurs, it holds that

$$m_i = m_N = \left\lceil \alpha 2^N \right\rceil \geq m_H(\lambda, r, \varepsilon, \tfrac{\delta}{3I}) \ ,$$

and thus by Hoeffding's inequality for mixing processes (theorem 7.2.5), we have $|\mu - \hat{\boldsymbol{\mu}}_N| \leq \varepsilon$, (i.e., the schedule was selected exactly to ensure m_N samples would be sufficient, regardless of early termination and variance.

We now proceed to show claim (2). To see this result, note that

$$\lambda(\mathcal{M}^{(T)}) \leq \lambda^T(\mathcal{M}) \ ,$$

and thus claim (2) follows directly from claim (1).

Finally, note that claim (3) follows from claim (2), paired with Eq. 70. In particular, note that the uniform mixing time is selected such that we have

$$\forall \omega : \left\| \frac{\partial \mathcal{M}^{\tau_{\mathrm{unif}}}(\omega)}{\partial \pi} \right\|_{\pi,\infty} \leq 2 \ ,$$

i.e., $\mathcal{M}$ is uniformly mixed. This further implies that the trace chain is uniformly mixed, as both $\pi_{\min}$ and Λ are exponential in the trace-length T, which cancels out in the uniform-mixing bound (see Eq. 12.13 of [Levin and Peres, 2017]). However, we still apply nonstationarity correction $\frac{\delta}{4}$ instead of $\frac{\delta}{2}$, to account for the fact that the *tensor product chain* ($\mathcal{M} \otimes \mathcal{M}$) *uniformly mixes* slightly slower than $\mathcal{M}$, even though it *relaxes* and *mixes* as quickly as $\mathcal{M}$. ☐

Before showing efficiency, we state a lemma that allows us to relate Tv_T with $\tau_{\mathrm{rel}} v_{\tau\mathrm{rel}}$.

Lemma F.1.7 (Trace Variance of Large Traces)**.** Suppose a lazy reversible Markov chain $\mathcal{M}$. The product of trace-size and trace-variance, i.e., Tv_T, is *asymptotically equivalent* for $T \geq \tau_{\mathrm{rel}}$. In other words, we have

$$\Theta(\tau_{\mathrm{rel}} v_{\tau\mathrm{rel}}) = \Theta(Tv_T) = \Theta\left(\lim_{t\to\infty} tv_t\right) \ .$$

Proof. We show this result by deriving a bidirectional inequality chain. We begin with the upper-bounds. By monotonicity, we have

$$\tau_{\mathrm{rel}} v_{\tau\mathrm{rel}} \leq Tv_T \leq \lim_{t\to\infty} tv_t \ .$$

We now show the lower-bounds. This direction is a bit subtler. Assume for now that $T = n\tau_{\mathrm{rel}}$ for some integer n. We then have

$$
\begin{aligned}
Tv_T(\mathcal{M}) &= n\tau_{\mathrm{rel}} v_n(\mathcal{M}^{(\tau_{\mathrm{rel}})}) && \textsc{Identities} \\
&\leq n\tau_{\mathrm{rel}} \frac{2\tau_{\mathrm{rel}}(\mathcal{M}^{(\tau_{\mathrm{rel}})})v_{\tau\mathrm{rel}}}{n} && \textsc{Eq. 3.14 of [Paulin, 2015]} \\
&= n\tau_{\mathrm{rel}} \frac{2\tau_{\mathrm{rel}}(\mathcal{M}^{(\tau_{\mathrm{rel}})})v_{\tau\mathrm{rel}}(\mathcal{M})}{n} && v_\pi(\mathcal{M}^{(\tau_{\mathrm{rel}})}) = v_{\tau\mathrm{rel}}(\mathcal{M}) \\
&= 2\tau_{\mathrm{rel}}\tau_{\mathrm{rel}}(\mathcal{M}^{(\tau_{\mathrm{rel}})})v_{\tau\mathrm{rel}}(\mathcal{M}) && \textsc{Algebra} \\
&\in \Theta(\tau_{\mathrm{rel}} v_{\tau\mathrm{rel}}(\mathcal{M})) \ . && \tau_{\mathrm{rel}}(\mathcal{M}^{(\tau_{\mathrm{rel}})}) \in \Theta(1)
\end{aligned}
$$

Correcting for non-integer n is simple, as we may apply the result to the nearest integer n, and the ratio $\frac{v_T}{v_{T'}}$ is bounded when $\frac{T}{T'}$ is near 1.

Finally, note that

$$\lim_{t \to \infty} tv_t \in \Theta(\tau_{\mathrm{rel}} v_{\tau\mathrm{rel}})$$

follows similarly via Eq. 15 of [Paulin, 2015]. □

We now present and prove an extended statement of theorem 7.3.2, which provides finite-sample and asymptotic sample complexity bounds to McmcPro, DynaMITE, and WarmStartDynaMITE.

Theorem F.1.8 (Efficiency of DynaMITE)**.** Suppose as in theorem 7.3.1, and take $N \doteq \log_2\left(\frac{r}{2\varepsilon}\right)$ and $T \doteq \lceil \frac{1+\Lambda}{1-\Lambda} \ln \sqrt{2} \rceil$. Then with probability at least $1 - \frac{\delta}{3N}$, each mean-estimation algorithm runs for no more than $\hat{m}$ steps (individually), where $\hat{m}$ is

1. for McmcPro:

$$
\begin{aligned}
\hat{m} &\leq 4T \ln \frac{3N}{\delta} \left(\frac{5(6+\Lambda)r}{2(1-\Lambda)\varepsilon} + \frac{(1+\Lambda)v_T}{(1-\Lambda)\varepsilon^2} \right) \\
&\in \mathcal{O}\left(\frac{1}{1-\Lambda} \log\left(\frac{\log(r/\varepsilon)}{\delta} \right) \left(\frac{r}{\varepsilon} + \frac{v_\pi}{\varepsilon^2} \right) \right) \ ;
\end{aligned}
$$

2. for DynaMITE:

$$
\begin{aligned}
\hat{m} &\leq 2T \ln \frac{3N}{\delta}\left(\frac{65r}{\varepsilon} + \frac{12v_T}{\varepsilon^2}\right) \\
&\in \mathcal{O}\left(T \log\left(\frac{\log(r/\varepsilon)}{\delta}\right)\left(\frac{r}{\varepsilon} + \frac{v_T}{\varepsilon^2}\right)\right) \\
&= \mathcal{O}\left(\log\left(\frac{\log(r/\varepsilon)}{\delta}\right)\left(\frac{r}{(1-\Lambda)\varepsilon} + \frac{\tau_{\mathrm{rel}} v_{\tau\mathrm{rel}}}{\varepsilon^2}\right)\right) \quad ; \ \&
\end{aligned}
$$

3. for WarmStartDynaMITE:

$$
\begin{aligned}
\hat{m} &\leq 2\lceil \frac{\ln \frac{1}{\pi_{\min}}}{\ln \frac{1}{\Lambda}} \rceil + 2T \ln \frac{12N}{\delta}\left(\frac{65r}{\varepsilon} + \frac{12v_T}{\varepsilon^2}\right) \\
&\in \mathcal{O}\left(\frac{\ln \frac{1}{\pi_{\min}}}{\ln \frac{1}{\Lambda}} + T \log\left(\frac{\log(r/\varepsilon)}{\delta}\right)\left(\frac{r}{\varepsilon} + \frac{v_T}{\varepsilon^2}\right)\right) \\
&= \mathcal{O}\left(\frac{\ln \frac{1}{\pi_{\min}}}{\ln \frac{1}{\Lambda}} + \log\left(\frac{\log(r/\varepsilon)}{\delta}\right)\left(\frac{r}{(1-\Lambda)\varepsilon} + \frac{\tau_{\mathrm{rel}} v_{\tau\mathrm{rel}}}{\varepsilon^2}\right)\right) \quad .
\end{aligned}
$$

Proof. The strategy here is to derive a sample size m', dependent on $r, v_T, \varepsilon, \delta$, s.t. w.h.p., each algorithm will terminate after drawing a sample of at least m' traces. We then bound the total number of samples drawn over the course of this process, and make some substitutions to derive the result. For brevity, throughout this result we take $\eta \doteq \ln\frac{3N}{\delta}$.

We show the result for McmcPro, using second absolute eigenvalue bound λ, as it immediately implies the corresponding results for DynaMITE and WarmStartDynaMITE.

We first show that, with high probability, the *empirical variance* is not much larger than the true variance, and thus with high probability, the variance-bounds used by McmcPro are not loose. Let

$$
\varepsilon_{v,1} \doteq \frac{5r^2\eta}{(1-\Lambda)m}, \quad \varepsilon_{v,2} \doteq \sqrt{\frac{(1+\Lambda)r^2 v_\pi \eta}{(1-\Lambda)m}} \quad .
$$

Now, note that by corollary F.1.6,[1] we have for any sample size m that

$$
\mathbb{P}\left(\hat{\boldsymbol{v}}_i \geq v_\pi + \varepsilon_{v,1} + \varepsilon_{v,2}\right) \leq \frac{\delta}{3N} \quad .
$$

[1]Note that here we take a lower-tail bound, rather than an upper-tail bound; the constants are identical and the result similarly follows from the Bernstein inequality.

We now consider the first iteration i such that $m_i \geq m'$, letting $m \doteq m_i$ and $\hat{v} \doteq \hat{\boldsymbol{v}}$. On line 12 of McmcPro, we have (w.h.p.)

$$
\begin{aligned}
\boldsymbol{u}_i &\leq \hat{v} + \frac{(11 + \sqrt{21} + (1 + \frac{11}{\sqrt{21}})\Lambda)r^2\eta}{(1-\Lambda)m} + \sqrt{\frac{(1+\Lambda)r^2\hat{v}\eta}{(1-\Lambda)m}} \\
&\leq v_\pi + \varepsilon_{v,1} + \varepsilon_{v,2} + \frac{(11+\sqrt{21}+(1+\frac{11}{\sqrt{21}})\Lambda)r^2\eta}{(1-\Lambda)m} + \sqrt{\frac{(1+\Lambda)r^2(v_\pi+\varepsilon_{v,1}+\varepsilon_{v,2})\eta}{(1-\Lambda)m}} && \hat{v} \leq v_\pi + \varepsilon_{v,1} + \varepsilon_{v,2} \text{ (w.h.p.)} \\
&= v_\pi + \varepsilon_{v,2} + \frac{(16 + \sqrt{21} + (1 + \frac{11}{\sqrt{21}})\Lambda)r^2\eta}{(1-\Lambda)m} + \sqrt{\frac{(1+\Lambda)r^2(v_\pi + \varepsilon_{v,1} + \varepsilon_{v,2})\eta}{(1-\Lambda)m}} \\
&= v_\pi + \varepsilon_{v,2} + \frac{(16 + \sqrt{21} + (1 + \frac{11}{\sqrt{21}})\Lambda)r^2\eta}{(1-\Lambda)m} + \sqrt{\frac{(1+\Lambda)r^2\eta}{(1-\Lambda)m}}\sqrt{v_\pi + \varepsilon_{v,1} + \varepsilon_{v,2}} && \textsc{Algebra} \\
&< v_\pi + \varepsilon_{v,2} + \frac{(16 + \sqrt{21} + (1 + \frac{11}{\sqrt{21}})\Lambda)r^2\eta}{(1-\Lambda)m} + \sqrt{\frac{(1+\Lambda)r^2\eta}{(1-\Lambda)m}}\left(\sqrt{v_\pi} + \sqrt{\varepsilon_{v,1}}\right) && \sqrt{a+b+2\sqrt{ab}} = \sqrt{a} + \sqrt{b} \\
&\leq v_\pi + \varepsilon_{v,2} + \frac{(16 + \sqrt{5} + \sqrt{21} + (1 + \frac{\sqrt{5}}{2} + \frac{11}{\sqrt{21}})\Lambda)r^2\eta}{(1-\Lambda)m} + \sqrt{\frac{(1+\Lambda)r^2 v_\pi\eta}{(1-\Lambda)m}} \\
&= v_\pi + \underbrace{\frac{(16 + \sqrt{5} + \sqrt{21} + (1 + \frac{\sqrt{5}}{2} + \frac{11}{\sqrt{21}})\Lambda)r^2\eta}{(1-\Lambda)m}}_{\doteq \varepsilon_{v,3}} + \underbrace{2\sqrt{\frac{(1+\Lambda)r^2 v_\pi\eta}{(1-\Lambda)m}}}_{\doteq \varepsilon_{v,4}}
\end{aligned}
$$

Substitution into the Bernstein bound (line 13), and similar algebra, gives us

$$
\begin{aligned}
\hat{\boldsymbol{\varepsilon}}_i &= \frac{10r\eta}{(1-\Lambda)m_i} + \sqrt{\frac{(1+\Lambda)\boldsymbol{u}_i\eta}{(1-\Lambda)m_i}} \\
&\leq \frac{10r\eta}{(1-\Lambda)m_i} + \sqrt{\frac{(1+\Lambda)(v_\pi + \varepsilon_{v,3} + \varepsilon_{v,4})\eta}{(1-\Lambda)m_i}} && \textsc{See Above} \\
&< \frac{5(6+\Lambda)r\eta}{2(1-\Lambda)m_i} + \sqrt{\frac{(1+\Lambda)v_\pi\eta}{(1-\Lambda)m_i}} && \begin{array}{r}\sqrt{a+b+2\sqrt{ab}} = \sqrt{a} + \sqrt{b} \\ \textsc{Algebraic Bounds}\end{array}
\end{aligned}
$$

We terminate when $\hat{\boldsymbol{\varepsilon}}_i \leq \varepsilon$, thus this implies sufficient sample size

$$
m' \leq +\eta\left(\frac{5(6+\Lambda)r}{2(1-\Lambda)\varepsilon} + \frac{(1+\Lambda)v_\pi}{(1-\Lambda)\varepsilon^2}\right) \quad .
$$

Now, due to the doubling geometric grid, we must have $m_i \in [m', 2m']$, thus we have

$$m_i \leq 2\eta \left(\frac{5(6+\Lambda)r}{2(1-\Lambda)\varepsilon} + \frac{(1+\Lambda)v_\pi}{(1-\Lambda)\varepsilon^2} \right) \enspace .$$

Now, each step of the tensor-product chain $(\mathcal{M} \otimes \mathcal{M})$ requires two steps of $\mathcal{M}$, so we conclude (1) by noting that $4m'$ samples suffice.

Now, to get (2), note that in DynaMITE, we take $T \doteq \lceil \frac{1+\Lambda}{1-\Lambda} \ln \sqrt{2} \rceil$, and thus $\lambda(\mathcal{M}^{(T)}) \leq \lambda^T \leq \frac{1}{2}$. The finite-sample bound then follows from (1) as DynaMITE simply calls McmcPro (see line 24, multiplying total sample complexity by T, as each step in $(\mathcal{M}^{(T)} \otimes \mathcal{M}^{(T)})$ (i.e., the tensor-product trace-chain) takes T steps in $\mathcal{M}$ for every step in the $\mathcal{M} \otimes \mathcal{M}$ chain of McmcPro. Finally, applying lemma F.1.7 yields the result.

We now show (3). In WarmStartDynaMITE, the nonstationarity correction appears as a factor-4 increase on $\frac{1}{\delta}$ (see line 31), and an additive sampling cost for the warm start (see line 29). □