Novel Concentration of Measure Bounds with Applications to Fairness in Machine Learning

Cyrus Cousins

Ph.D. Proposal

Advisor Eli Upfal

Department of Computer Science
Brown University
Providence, Rhode Island 02912

November 2020

Abstract

I introduce novel concentration-of-measure bounds for the supremum deviation, several variance concepts, and a family of *game-theoretic welfare functions*. In particular, I introduce *empirically centralized Rademacher averages* to derive novel 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. Theoretically, the gaps between upper and lower bounds on the SD are much smaller with (empirical) centralization, with asymptotically improved dependence on various quantities of interest, and experimentally, I find that centralization improves various bounds to the SD.

I also give axiomatic justification of the *power mean* family of welfare functions, and introduce the concept of *malfare*, which measures group ill-being (instead of wellbeing), and shares the axiomatic justification of welfare. Malfare is a natural target in machine learning problems where we *minimize* (negatively connoted) loss, rather than *maximize* (positively connoted) utility. Surprisingly, welfare and malfare are not equivalent, essentially because the class of welfare (malfare) functions is not closed under *negation* or *affine transformation*. I then show statistical estimation guarantees for welfare and malfare, and develop a unified theory of fair machine learning, termed *fair PAC-learning*, in which polynomial samples are sufficient to provably ϵ-δ learn fair models. Finally, I cast a *streaming media codec selection* problem as a *fairness-sensitive* learning problem. I explore, both experimentally and theoretically, multivariate *welfare* and *Pareto* optimality concepts, and how the bias complexity tradeoff manifests in multivariate settings and with fairness issues.

Future work will investigate fair PAC-learning of generalized linear models, with experimental comparison to alternative fair learning techniques. Additionally I will apply my novel uniform convergence bounds to various sampling and data science problems, to show decreased sample complexity, thus more computationally efficient sampling algorithms.

Contents

Prelude

This thesis is organized roughly into 3 main parts. We first discuss purely statistical and learning theoretic work in chapter 1, mainly deriving new ways to bound *uniform convergence* of the *empirical mean* of each $f \in \mathcal{F}$ to its *expectation*, with brief allusions to how these results are relevant to problems in *statistical estimation*, *machine learning*, and *data science*. In my final thesis (absent from this proposal), this part will be followed by novel techniques in data science, informed by the statistical methods of chapter 1. Finally, in chapter 2, I apply the statistical methods of chapter 1 to *fairness in machine learning*. I argue that this is a particularly important setting for rigorous statistical learning guarantees, as overfitting has significant real-world implications to *fairness and justice*. I also show that, by its nature, uniform convergence techniques is particularly well-suited to answer inherently-multivariate fairness questions.

Statistical methods, concentration of measure, and uniform convergence The highly theoretical work of chapter 1 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.

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

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

[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

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

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

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

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

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

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 may nearly match variance-sensitive lower-bounds or Bennett's inequality. We defer formal definitions of Rademacher averages to chapter 1, but to briefly preview, we shall see that *centralized Rademacher averages* also play a fundamental role in lower-bounding sample complexity, which we rigorously realize through the novel *empirical centralization* strategy. We also show bounds that minimally depend on extraneous quantities, such as ranges and raw-variances, while depending on centralized Rademacher averages and (centralized) variances (as asymptotically necessary). Furthermore, in practice it is often unreasonable to assume variances and other properties beyond ranges are known *a priori*, so we additionally show that variances, Rademacher averages, and all relevant statistics may be bounded w.h.p. from a finite sample.

Characterizing fair machine learning, with statistical guarantees In chapter 2, I define a novel concept of *fairness* based on the idea of the *social planner's problem* in welfare theory. While a *welfare* function $\mathrm{W}(\dots)$ summarizes *overall wellbeing* (e.g., income, happiness, or life expectancy) across a population, I introduce the parallel notion of *malfare* $\reflectbox{W}(\dots)$, which summarizes *overall illbeing* (e.g., suffering, harm, or disease); both described in definition 2.1.1. Naturally, we seek to *minimize malfare* in learning problems (often posed as *mechanism design* problems) where one would generally *maximize welfare*. We justify this fairness concept via the same axiomatization as welfare theory. Although the decision to minimize harm seems like a macabre twist on maximizing utility, and they share an axiomatic justification, surprisingly, the two concepts are in general not isomorphic.

In this framework, lemma 2.3.1 shows highly generic conditions under which malfare and welfare concepts can be estimated from a sample, i.e., I address the *evaluation problem*. I then generalize the notion of *PAC-learnability* to *fair PAC-learnability* (definition 2.3.4) to optimize these objectives, rather than the empirical risk of a single distribution, and show sufficient conditions under which a model is fair-PAC-learnable. Finally, I also show data-dependent generalization bounds using Rademacher averages that are on-par with our state-of-the-art statistical learning methodology of chapter 1.

In chapter 3, I perform an extended study in *codec selection*, i.e., selecting a set of *media encodings* (trading off between fidelity and bandwidth) to satisfy a diverse population of users. This is an *accessibility issue*, as it concerns optimizing a system to make it usable and useful to groups with varying desiderata. In particular, I define and analyze the *Empirical Welfare-Optimal k codec Selection* (EWO*k*S) algorithm, showing rigorous data-dependent generalization guarantees, ensuring fairness not only on the training data, but also on the *underlying data distribution*. I show how to optimize for and consider the multivariate nature of various fairness issues (e.g., welfare optimization and Pareto optimality).

I note that while fair-learning via *empirical welfare maximization* (EWM) and *empirical malfare minimization* (EMM) are analogous to learning via *empirical risk minimization* (ERM) [Vapnik, 1992], there are some key differences, and neither EWM nor EMM reduces to ERM. In general, empirical welfare and malfare are *biased estimates* of true welfare and malfare (respectively), whereas empirical risk is an *unbiased estimate* of true risk, making even *evaluating a single model* more challenging. Furthermore, while the ERM function is an M-estimator, the EWM and EMM functions are not (due to nonlinearity in malfare or welfare functions), which complicates understanding *generalization error*. Despite these differences, I show high-probability bounds on welfare and malfare, and generalization guarantees using well-known *concentration inequalities* and *Rademacher averages*, thus the additional statistical challenge is not insurmountable. Similarly, while I show conditions under which fair-PAC-learning is computationally tractable, many of which resemble standard conditions for traditional PAC-learning, it remains an open question which (if any) classes of model are PAC-learnable but not fair-PAC learnable.

Chapter 1

Concentration of Measure and Uniform Convergence Bounds

1.1 Introduction

The *supremum deviation* of the empirical means of functions in a family $\mathcal{F} \subseteq \mathcal{X} \to [a,b] \subset \mathbb{R}$ 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 $\boldsymbol{x} = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_m\}$ be a collection of m independent samples from $\mathcal{D}$. The *Supremum Deviation (SD) of* $\mathcal{F}$ *on* $\boldsymbol{x}$ is the quantity

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

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 equation 1.1.2). Let $\boldsymbol{\sigma}$ be a collection of m independent Rademacher variables (i.e., uniform on $\{-1, 1\}$). These two quantities are defined as

$$\hat{\mathfrak{R}}_m(\mathcal{F}, \boldsymbol{x}) \doteq \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{m} \sum_{i=1}^{m} \boldsymbol{\sigma}_i f(\boldsymbol{x}_i) \right| \right], \text{ and } \mathfrak{R}_m(\mathcal{F}, \mathcal{D}) \doteq \mathop{\mathbb{E}}_{\boldsymbol{x}} \big[\hat{\mathfrak{R}}_m(\mathcal{F}, \boldsymbol{x}) \big] \ . \tag{1.1.1}$$

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

$$\tfrac{1}{2} \mathfrak{R}_m(\mathcal{F}, \mathcal{D}) - \tfrac{1}{\sqrt{m}} \sup_{f \in \mathcal{F}} \|f\|_\infty \leq \mathop{\mathbb{E}}_{\boldsymbol{x}} [\mathsf{SD}(\mathcal{F}, \boldsymbol{x})] \leq 2 \mathfrak{R}_m(\mathcal{F}, \mathcal{D}) \ . \tag{1.1.2}$$

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 equation 1.3.1). 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

$$\mathrm{C}_{\mathcal{D}}(\mathcal{F}) \doteq \{x \mapsto f(x) - \mathop{\mathbb{E}}_{\mathcal{D}}[f], f \in \mathcal{F}\} \ . \tag{1.1.3}$$

$\mathrm{C}_{\mathcal{D}}(\mathcal{F})$ 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]

$$\tfrac{1}{2} \mathfrak{R}_m(\mathrm{C}_{\mathcal{D}}(\mathcal{F}), \mathcal{D}) \leq \mathop{\mathbb{E}}_{\boldsymbol{x}} [\mathsf{SD}(\mathcal{F}, \boldsymbol{x})] \leq 2 \mathfrak{R}_m(\mathrm{C}_{\mathcal{D}}(\mathcal{F}), \mathcal{D}) \ . \tag{1.1.4}$$

Comparing equation 1.1.2 and equation 1.1.4, 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 equation 1.2.2), 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

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

This quantity is an empirical counterpart to the distributional centralization $\mathrm{C}_{\mathcal{D}}(\mathcal{F})$ of $\mathcal{F}$ (see equation 1.1.3). 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

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

This quantity is completely dependent on the realized $\boldsymbol{x}$, even more, in some sense, than a "standard" ERA (see equation 1.1.1), 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

$$\mathrm{b}(m) \doteq \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\left| \frac{1}{m} \sum_{i=1}^{m} \boldsymbol{\sigma}_i \right| \right] \ \left(\text{which is } \boldsymbol{\Theta}\left(\frac{1}{\sqrt{m}} \right) \right) \ . \tag{1.2.1}$$

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

$$r \doteq \sup_{f\in\mathcal{F}} \big(\max_{x\in\mathcal{X}} f(x) - \min_{y\in\mathcal{X}} f(y)\big) \quad (\leq b-a) \ . \tag{1.2.2}$$

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*

$$\mathop{\mathbb{E}}_{\boldsymbol{x}}[\hat{\mathfrak{R}}_m(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}), \boldsymbol{x})] \leq \hat{\mathfrak{R}}_m(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}), \boldsymbol{x}) + \frac{2r\chi \ln\frac{1}{\delta}}{3m} + \sqrt{\left(\frac{r\chi\ln\frac{1}{\delta}}{\sqrt{3}m}\right)^2 + \frac{2r\chi(\hat{\mathfrak{R}}_m(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}), \boldsymbol{x}) + r\mathrm{b}(m))\ln\frac{1}{\delta}}{m}} \ . \tag{1.2.3}$$

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

$$\mathrm{W}^{\mathrm{r}}(\mathcal{F}) \doteq \sup_{f\in\mathcal{F}} \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\big(f(x)\big)^2\right], \text{ and } \mathrm{W}(\mathcal{F}) \doteq \sup_{f\in\mathcal{F}} \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\left(f(x) - \mathop{\mathbb{E}}_{\mathcal{D}}[f]\right)^2\right] \ . \tag{1.3.1}$$

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 $\boldsymbol{x}$,*

$$\mathrm{W}(\mathcal{F}) \leq \tfrac{m}{m-1}\widehat{\mathrm{W}}_{\boldsymbol{x}}(\mathcal{F}) + \frac{r^2 \ln \frac{1}{\delta}}{m-1} + \sqrt{\left(\frac{r^2 \ln \frac{1}{\delta}}{m-1}\right)^2 + \frac{2r^2 \frac{m}{m-1}\widehat{\mathrm{W}}_{\boldsymbol{x}}(\mathcal{F}) \ln \frac{1}{\delta}}{m-1}} \ . \tag{1.3.2}$$

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 $\boldsymbol{x}$,*

$$\mathsf{SD}(\mathcal{F},\boldsymbol{x}) \leq \mathop{\mathbb{E}}_{\boldsymbol{x}}\left[\mathsf{SD}(\mathcal{F},\boldsymbol{x})\right] + \frac{2r \ln \frac{1}{\delta}}{3m} + \sqrt{\frac{2\left(\mathrm{W}(\mathcal{F}) + 4r\ \mathbb{E}_{\boldsymbol{x}}\left[\mathsf{SD}(\mathcal{F},\boldsymbol{x})\right]\right)\ln \frac{1}{\delta}}{m}} \ . \tag{1.3.3}$$

By plugging the r.h.s. of the symmetrization inequalities equation 1.1.2 and equation 1.1.4 in the r.h.s. of equation 1.3.3, 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 equation* 1.3.2 *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 equation* 1.2.3 *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*

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

The r.h.s. is

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

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

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

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 equation 1.3.3 cannot be improved: any bound to the SD of $\mathcal{F}$ must be $\boldsymbol{\Omega}\sqrt{\mathrm{W}(\mathcal{F})\ln \frac{1}{\delta}/m}$, 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 $\hat{\mathfrak{R}}_m(\mathcal{F})$, 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*

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

Furthermore, it holds

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

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*

$$\hat{\mathfrak{R}}_m(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}), \boldsymbol{x}) \leq \sqrt{\frac{2\widehat{\mathrm{W}}_{\boldsymbol{x}}(\mathcal{F}) \ln|\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}_\pm)|}{m}} \ . \tag{1.3.4}$$

The use of $\mathcal{F}_\pm$ is needed to handle the absolute value in our definition of the ERA (see equation 1.1.1). 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

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

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. 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

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

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*

$$\hat{\mathfrak{R}}_{m}(\mathcal{F},\boldsymbol{x}) \leq \hat{\mathfrak{R}}^{n}_{m}(\mathcal{F},\boldsymbol{x},\boldsymbol{\sigma}) + \frac{2\hat{q}_{\mathcal{F}}(\boldsymbol{x})\ln\frac{1}{\delta}}{3nm} + \sqrt{\frac{4\widehat{\mathrm{W}}^{\mathrm{r}}_{\boldsymbol{x}}(\mathcal{F})\ln\frac{1}{\delta}}{nm}} \ . \tag{1.3.5}$$

Empirical centralization obtains a dependence on the empirical wimpy variance of $\mathcal{F}$, rather than on the raw (i.e., non-centralized) one. 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*

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

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, i.e., 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.

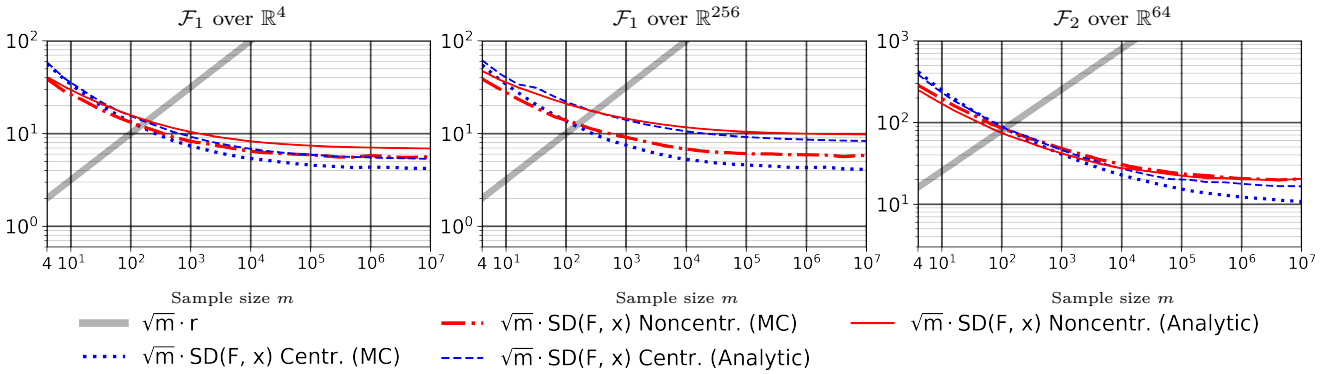

Figure 1.1: Comparison of SD bounds as functions of the sample size m. See the main text for an explanation of the results.

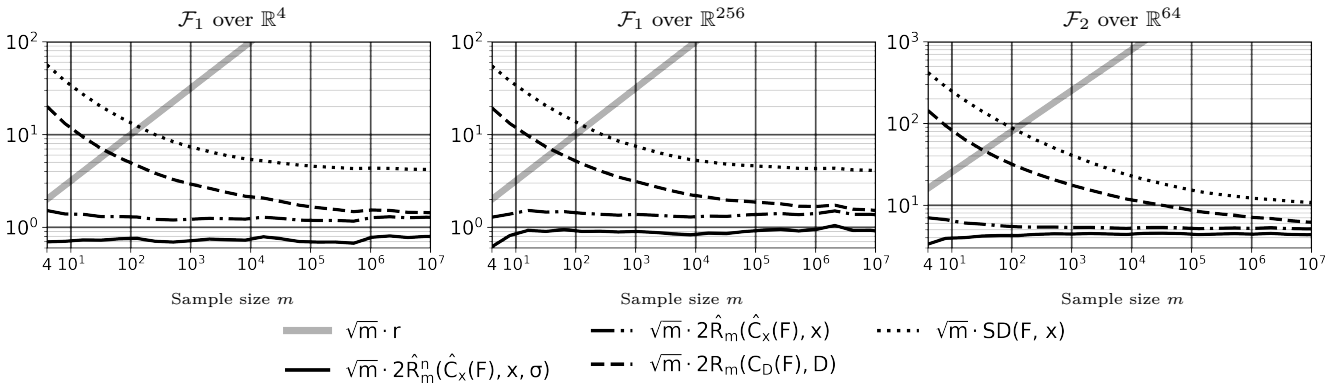

Figure 1.2: Upper bounds to complexity measures and SD as functions of the sample size m. See the main text for details.

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*

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

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 (e.g., q, $\hat{q}$, 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 equation 1.1.4 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 $\boldsymbol{\Theta}(C/\sqrt{m})$, 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 $\boldsymbol{\Theta}\sqrt{\mathrm{W}(\mathcal{F})/m}$ (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

Fairness with Population Means: Malfare and Welfare

2.1 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 2.1.1 (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 (i.e., 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 (i.e., loss or risk), we call it a* malfare function*, written* $\mathrm{W}(\mathcal{S}; \boldsymbol{w})$*.*

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 $\mathrm{M}(\mathcal{S}; \boldsymbol{w})$ is an M (*inverted* W), to emphasize its inverted nature.

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

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

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 are needed to represent cases where populations differ in size, and ensure that the welfare or malfare of *weight-preserving decompositions* of groups into subgroups with equal risk or utility remains constant.

Example 2.1.2 (Utilitarian Welfare)**.** *Suppose a 3-group population* $\Omega \doteq \{\omega_1, \omega_2, \omega_3\}$ *consisting of three groups, encompassing* 30%*,* 30%*, and* 40% *of the total population, respectively. We now take* weights measure $\boldsymbol{w}$ *s.t.* $\boldsymbol{w}(\omega_i) = (0.3, 0.3, 0.4)_i$*, representing the relative sizes of each group in the population.*

Now suppose individuals reside in some space $\mathcal{X}$*, where* distributions $\mathcal{D}_{1:3}$ *over domain* $\mathcal{X}$ *describe the 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 e.g., w.r.t. some particular* situation*,* allocation*, or* classifier*. We now take* sentiment function *to be the* mean utility *(per-group), i.e.,*

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

We then define the utilitarian welfare *as*

$$\mathrm{W}_1(\mathcal{S};\boldsymbol{w}) \doteq \sum_{i=1}^{3} \boldsymbol{w}(\omega_i)\mathcal{S}(\omega_i) = \mathop{\mathbb{E}}_{\omega\sim\boldsymbol{w}}[\mathcal{S}(\omega)] = \mathop{\mathbb{E}}_{\boldsymbol{w}}[\mathcal{S}] \ .$$

Of course, in statistical, sampling, and machine learning contexts, $\mathcal{D}_{1:n}$ and $\boldsymbol{w}$ may be unknown, so we now discuss an *empirical analogue*. Section 2.3 is then devoted to showing how and when empirical population means well-approximate their true counterparts.

Example 2.1.3 (Empirical Utilitarian Welfare)**.** *Now suppose $\mathcal{D}_{1:n}$ are unknown, but instead, we are given a* sample *$\boldsymbol{x}_{1:n,1:m} \in \mathcal{X}^{n\times m}$, where $\boldsymbol{x}_{i,1:m} \sim \mathcal{D}_{1:m}$. We define an* empirical analogue *of the utilitarian welfare as in example 2.1.2, instead taking*

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

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.*

2.1.1 Axioms of Cardinal Welfare (or Malfare)

Definition 2.1.4 (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 $\forall\mathcal{S},\mathcal{S}' \in \Omega\to\mathbb{R}_{0+}$, $\forall$ permutations π over Ω, and probability measure $\boldsymbol{w}$ over Ω.*

1. *Monotonicity: $\forall\boldsymbol{\varepsilon}:\Omega\to\mathbb{R}_{0+} : \mathrm{M}(\mathcal{S};\boldsymbol{w}) \le \mathrm{M}(\mathcal{S}+\boldsymbol{\varepsilon};\boldsymbol{w})$.*
2. *Symmetry: $\mathrm{M}(\mathcal{S};\boldsymbol{w}) = \mathrm{M}(\pi(\mathcal{S});\pi(\boldsymbol{w}))$.*
3. *Continuity: $\forall\mathcal{S} : \{\mathcal{S}' \mid \mathrm{M}(\mathcal{S}';\boldsymbol{w}) \le \mathrm{M}(\mathcal{S};\boldsymbol{w})\}$ is a closed set.*
4. *Independence of unconcerned agents: Suppose $\alpha,\beta\in\mathbb{R}_{0+}$, and take subpopulation $\Omega'\subseteq\Omega$. Then*

$$\mathrm{M}\left(\begin{cases} p\in\Omega' & \alpha \\ p\notin\Omega' & \mathcal{S}(p)\end{cases};\boldsymbol{w}\right) \le \mathrm{M}\left(\begin{cases} p\in\Omega' & \alpha \\ p\notin\Omega' & \mathcal{S}'(p)\end{cases};\boldsymbol{w}\right) \implies \mathrm{M}\left(\begin{cases} p\in\Omega' & \beta \\ p\notin\Omega' & \mathcal{S}(p)\end{cases};\boldsymbol{w}\right) \le \mathrm{M}\left(\begin{cases} p\in\Omega' & \beta \\ p\notin\Omega' & \mathcal{S}'(p)\end{cases};\boldsymbol{w}\right) \ .$$

5. *Independence of common scale: Suppose scalar $\alpha\ge 0$. Then $\mathrm{M}(\mathcal{S},\boldsymbol{w}) \le \mathrm{M}(\mathcal{S}',\boldsymbol{w}) \implies \mathrm{M}(\alpha\mathcal{S},\boldsymbol{w}) \le \mathrm{M}(\alpha\mathcal{S}',\boldsymbol{w})$*
6. *Multiplicative Linearity: Suppose $\alpha\ge 0$. Then $\mathrm{M}(\alpha\mathcal{S};\boldsymbol{w}) = \alpha\mathrm{M}(\mathcal{S};\boldsymbol{w})$.*
7. Unit Scale*: $\mathrm{M}(\mathbf{1};\boldsymbol{w}) = \mathrm{M}(1,\dots,1;\boldsymbol{w}) = 1$.*
8. *Pigou-Dalton transfer principle: Suppose $\mathcal{S},\mathcal{S}'$ s.t. $\mu = \mathbb{E}_{\boldsymbol{w}}[\mathcal{S}] = \mathbb{E}_{\boldsymbol{w}}[\mathcal{S}']$, 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 the hypothesis of 8, but instead require the conclusion $\mathrm{M}(\mathcal{S}';\boldsymbol{w}) \le \mathrm{M}(\mathcal{S};\boldsymbol{w})$.*

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 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*.

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

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

While seemingly quite important to early welfare theorists (i.e., 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. 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 2.1.3, and directly compared with the additively-separable form in section 2.1.4.

2.1.2 The Power Mean

We now define the *p-power mean* $\mathrm{M}_p(\ldots)$, 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 2.1.4, and as we shall see in theorem 2.1.7, are a particularly important class of population means.

Definition 2.1.6 (Power-Mean Welfare and Malfare)**.** *Suppose* $p \in \mathbb{R} \cup \pm\infty$*. We first define the*

unweighted power mean *of sentiment vector* $\mathcal{S} \in \mathbb{R}_{0+}^n$ *as*

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

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*

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

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* $\mathrm{M}_f(\mathcal{S}; \boldsymbol{w}) \doteq f^{-1}(\mathbb{E}_{\omega \sim \boldsymbol{w}}[\ln \mathcal{S}(\omega)])$*.*

Theorem 2.1.7 (Properties of the Power-Mean)**.** *Suppose* $\mathcal{S}, \boldsymbol{\varepsilon}$ *are loss values* $\Omega \to \mathbb{R}_{0+}$*, and* $\boldsymbol{w}$ *is a probability measures over some space* 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:* $\forall p \geq 1 : \left|\mathrm{M}_p(\mathcal{S}; \boldsymbol{w}) - \mathrm{M}_p(\mathcal{S}'; \boldsymbol{w})\right| \leq \mathrm{M}_p\left(\left|\mathcal{S} - \mathcal{S}'\right|; \boldsymbol{w}\right) \leq \left\|\mathcal{S} - \mathcal{S}'\right\|_\infty$*.*
4. *Curvature:* $\mathrm{M}_p(\mathcal{S})$ *is convex for* $p \in [1, \infty]$ *and concave for* $p \in [-\infty, 1]$*.*

Proof. We omit proof of item 1, as this is a standard property of power-means (generally termed the *power mean inequality* [Bullen, 2013, Ch. 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

$$\begin{aligned} \mathrm{M}_p(\mathcal{S}; \boldsymbol{w}) &= \mathrm{M}_p(\mathcal{S}' + \boldsymbol{\varepsilon}; \boldsymbol{w}) && \text{Definition of } \boldsymbol{\varepsilon} \\ &\leq \mathrm{M}_p(\mathcal{S}' + \boldsymbol{\varepsilon}_+; \boldsymbol{w}) && \text{Monotonicity} \\ &\leq \mathrm{M}_p(\mathcal{S}'; \boldsymbol{w}) + \mathrm{M}_p(\boldsymbol{\varepsilon}_+; \boldsymbol{w}) && \text{item 2} \\ &\leq \mathrm{M}_p(\mathcal{S}'; \boldsymbol{w}) + \mathrm{M}_p\left(\left|\mathcal{S} - \mathcal{S}'\right|; \boldsymbol{w}\right) \ . && \text{Monotonicity} \end{aligned}$$

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

$$\mathrm{M}_1(\mathcal{S}; \boldsymbol{w}) = \mathop{\mathbb{E}}_{\omega \sim \boldsymbol{w}}[\mathcal{S}(\omega)] = \underbrace{\mathop{\mathbb{E}}_{\omega \sim \boldsymbol{w}}[\sqrt[p]{\mathcal{S}^p(\omega)}] \leq \sqrt[p]{\mathop{\mathbb{E}}_{\omega \sim \boldsymbol{w}}[\mathcal{S}^p(\omega)]}}_{\text{Definition of Convexity}} = \mathrm{M}_p(\mathcal{S}; \boldsymbol{w}) \ ,$$

i.e., convexity, and similarly, for $p \leq 1$, $p \neq 0$, we have by convexity of $\sqrt[p]{\cdot}$, we have

$$\mathrm{M}_1(\mathcal{S}; \boldsymbol{w}) = \mathop{\mathbb{E}}_{\omega \sim \boldsymbol{w}}[\mathcal{S}(\omega)] = \underbrace{\mathop{\mathbb{E}}_{\omega \sim \boldsymbol{w}}[\sqrt[p]{\mathcal{S}^p(\omega)}] \geq \sqrt[p]{\mathop{\mathbb{E}}_{\omega \sim \boldsymbol{w}}[\mathcal{S}^p(\omega)]}}_{\text{Definition of Concavity}} = \mathrm{M}_p(\mathcal{S}; \boldsymbol{w}) \ .$$

Similar reasoning, now by convexity of $\ln(\cdot)$, shows the case of $p = 0$, which completes the proof. □

2.1.3 Properties of Welfare and Malfare Functions

We now show that definition 2.1.4 are sufficient to characterize many properties of Welfare and Malfare.

Theorem 2.1.8 (Population Mean Properties (Weighted))**.** *Suppose population-mean function* $\mathrm{M}(\mathcal{S}; \boldsymbol{w})$*. If* $\mathrm{M}(\cdot;\cdot)$ *satisfies (subsets of) the population-mean axioms (see definition 2.1.4), we have that* $\mathrm{M}(\cdot;\cdot)$ *exhibits the following properties. For each, assume arbitrary sentiment-value function* $\mathcal{S} : \Omega \to \mathbb{R}_{0+}$*, weights measure* $\boldsymbol{w}$*.*

1. Identity*: Axioms 6 & 7 imply* $\mathrm{M}(\omega \mapsto \alpha, \boldsymbol{w}) = \alpha$*.*
2. *Axioms 1-5 imply* $\exists p \in \mathbb{R}$*, monotonically increasing* $F : \mathbb{R} \to \mathbb{R}_{0+}$ *s.t.*

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

3. *Axioms 1-7 imply* $\exists p \in \mathbb{R}$ *s.t.* $\mathrm{M}(\mathcal{S}; \boldsymbol{w}) = \mathrm{M}_p(\mathcal{S}; \boldsymbol{w})$ *(i.e.,* $\mathrm{M}(\cdot;\cdot)$ *is a* weighted power-mean *with finite p).*
4. *Axioms 1-5 and 8 imply* $p \in [1, \infty)$*.*
5. *Axioms 1-5 and 9 imply* $p \in (-\infty, 1]$*.*

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 6 and item 7, for all $p \neq 0$, we have:

$$\alpha = \alpha\mathrm{M}(\mathcal{S}; \boldsymbol{w}) = \mathrm{M}(\alpha\mathcal{S}; \boldsymbol{w}) = F(\mathop{\mathbb{E}}_{\omega \sim \boldsymbol{w}}[f_p(\alpha\mathcal{S}(\omega))]) = F(\mathop{\mathbb{E}}_{\omega \sim \boldsymbol{w}}[f_p(\alpha)]) = \begin{cases} p > 0 & F(\mathop{\mathbb{E}}_{\omega \sim \boldsymbol{w}}[(\alpha\mathcal{S}(\omega))^p]) = F(\alpha^p) \\ p < 0 & F(\mathop{\mathbb{E}}_{\omega \sim \boldsymbol{w}}[-(\alpha\mathcal{S}(\omega))^p]) = F(-\alpha^p) \end{cases} .$$

From here, it is clear that for $p > 0$, $\alpha = F(\alpha^p)$, thus $F^{-1}(u) = u^p$ and consequently $F(v) = \sqrt[p]{v}$, and similarly, for $p < 0$, $\alpha = F(-\alpha^p)$, thus $F^{-1}(u) = -u^p$, and consequently $F(v) = \sqrt[p]{-v}$.

Taking $p = 0$ gets us

$$\alpha = \alpha\mathrm{M}(\mathcal{S}; \boldsymbol{w}) = F(\mathop{\mathbb{E}}_{\omega \sim \boldsymbol{w}}[\ln(\alpha\mathcal{S}(\omega))]) = F(\ln a) \ ,$$

from which it is clear that $F^{-1}(u) = \ln(u) \implies F(v) = \exp(v)$.

In all cases of p, substituting the values of f_p and $F(\cdot)$ into 2 yields $\mathrm{M}(\mathcal{S}; \boldsymbol{w}) = \mathrm{M}_p(\mathcal{S}; \boldsymbol{w})$.

We now show 4 and 5. These properties follow directly from 2, wherein f_p are defined, and Jensen's inequality.

□

Taken together, the items of theorem 2.1.8 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* (i.e., nats or binits 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 Malfare function (even without multiplicative linearity, axioms 1-5 imply a population mean functions is still a *monotonic transformation* of a power-mean). Furthermore, the entirely milquetoast *unit scale* axiom 7 implies that sentiment values and population means have the same *scale* (making comparisons like "the loss of group i is above / 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 suffering from worse-off groups to better-off groups.

2.1.4 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

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

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 2.1.7, both of which 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 *mean concept*. This yields desirable statistical estimation and learnability properties (see section 2.3), 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 dimension (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*.

2.1.5 Relating Power Means and Inequality Indices

We now discuss and define *relative inequality indices*, which have been employed in the literature to construct *welfare functions* of the form

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

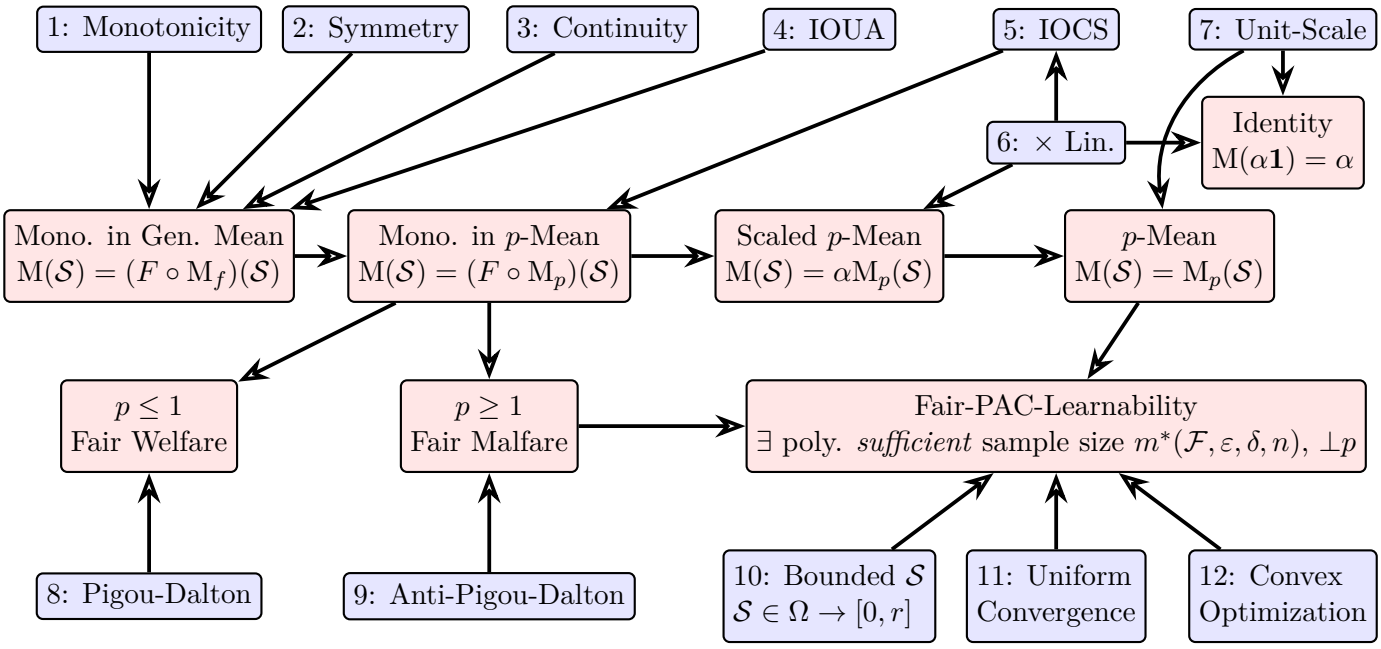

Figure 2.1: Relationships between population-mean axioms and properties. Assumptions and axioms shown in pastel blue, and properties shown in pastel red. Equivalent properties hold for weighted population mean functions.

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 $[0, 1]$.

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

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

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

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

Proof. This is a direct consequence of definition 2.1.9, noting $p = 1 - \varepsilon \Leftrightarrow \varepsilon = 1 - p$. □

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 2.1.8), 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.

2.2 Welfare and Fairness in Machine Learning

Most prior-work in fair machine learning arguably falls into a single category: adding *fairness constraints* to existing *risk minimization* objectives. Constraints like *equalized odds*, *equality of opportunity*, *equality of outcome*, and many others, are imposed on various measures of *utility* $\mathrm{U}(\dots)$ between fairness-sensitive groups. In particular, given *labeled samples* $\boldsymbol{z}_{1:n}$ for groups $1, \dots, n$ we take *empirical risk*

$$\hat{\mathrm{R}}(h; \boldsymbol{z}, \ell) \doteq \hat{\mathbb{E}}_{(x,y) \in \boldsymbol{z}_1 \cup \dots \cup \boldsymbol{z}_n} \left[\ell(y, h(x)) \right] \ ,$$

where sample $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{x}) \in (\mathcal{X} \times \mathcal{Y})^m$. Now assume some *utility function*, usually posed as $\mathrm{U}(\hat{y}, y)$, representing the *value* an individual places on receiving predicted label $\hat{y}$, given true label y, and impose *constraints* of the form

$$C\left(\hat{\mathbb{E}}_{(x,y) \in \boldsymbol{z}_1} [\mathrm{U}(h(x), y)], \dots, \hat{\mathbb{E}}_{(x,y) \in \boldsymbol{z}_n} [\mathrm{U}(h(x), y))] \right) \leq \tau \ ,$$

i.e., some function of per-group empirical utilities is constrained.

I argue that under this setup, choice of $\mathrm{U}(\dots)$ is often rather ad-hoc, and, interpreted literally, leads to absurdity; for instance, if a classifier is to decide whether x committed a crime, and thus is to be jailed, does the fact of whether x is guilty have any impact on the utility they derive from punishment? Does the answer change if we cast the potentially rehabilitating effects of imprisonment inside, and ask if instead x is to be executed? The question is then naturally, whose utility do we seek to optimize? [Kasy and Abebe, 2020] raise objections to this method of constrained optimization, asking whether it is the mechanism designer, rather than the groups under consideration, who benefit from such approaches.

Furthermore, except in trivial circumstances, many such constraints are *mutually unsatisfiable* [Kleinberg et al., 2017]. [Hu and Chen, 2020] also criticize the approach, showing that in certain cases, increasing the value of a constraint can actually lead to *worse outcomes* for disadvantaged groups. One approach, taken by [Zafar et al., 2017], is to *relax* such constraints, to instead require only that each group is *at least as well off* as if such constraints had been imposed. This notion of envy-free Pareto-like dominance seemingly lies between constraint-based formulations and direct welfare-maximization.

Previous authors have also considered welfare *as a constraint* in fair machine learning methods, although the mechanism by which welfare enters the equation differs substantially from our approach. In particular, without a notion of *malfare*, there is no clear way to convert *loss*, as measured by various machine learning methods, to *utility*, as required by *welfare concepts*.

Now, assuming utility $\mathrm{U}(\dots)$ and welfare function $\mathrm{W}(\cdot; \boldsymbol{w})$, take *empirical welfare*

$$\hat{\mathrm{W}}(h; \boldsymbol{z}_{1:n}, \mathrm{U}) \doteq \mathrm{W}\left(i \mapsto \hat{\mathbb{E}}_{(x,y) \in \boldsymbol{z}_i} [\mathrm{U}(y, h(x))]; \boldsymbol{w} \right) \ .$$

We may then use $\hat{\mathrm{W}}(h; \boldsymbol{z}_{1:n}, \mathrm{U}) \leq \tau$ as a *constraint* in the standard constraint-based formulation.

This approach is taken, e.g., by [Speicher et al., 2018, Heidari et al., 2018] propose a notion, wherein we tradeoff between *classifier accuracy* and *welfare* (fairness) via a *constrained optimization problem*, where *empirical risk* (across all groups) is minimized, *subject to a constraint* on *empirical welfare*; i.e., $\hat{\mathrm{W}}(h; \boldsymbol{z}_{1:n}, \mathrm{U}) \geq \tau$. They assume axioms 1-4 and 8, and consider the *additively separable* family of welfare concepts $\{\mathrm{W}_p^p(\mathcal{S}) \mid p \in [0, 1]\}$. However, despite this strong theoretical grounding in welfare-economics,

and their consideration of efficient *computational routines* for this constrained minimization, they don't consider the *statistical aspects* of this problem, leaving open the door to *overfitting*. This is particularly dangerous in this context, as not only may we overfit to *loss* (i.e., true risk is higher than empirical), but also to *fairness* (i.e., true welfare is smaller than empirical), giving a false sense of security as to the fairness of an algorithm or model.

We propose instead to consider the *empirical risk* of a group to be their *sentiment value* (i.e., $\mathcal{S}(i) \mapsto \hat{\mathrm{R}}(h; \boldsymbol{z}, \ell)$), and then *directly* minimize some (empirical) malfare concept $\reflectbox{M}(\dots)$. We term this objective *empirical malfare minimization* (EMM), in line with the *empirical risk minimization* (ERM) standard in statistical learning theory. There is no tradeoff between the inequality concept W and the inequality constraint τ, as we have reduced to a single interpretable parameter p.

In generic classification tasks, where a group's satisfaction is a function of the performance of the classifier, this is a very natural way of measuring their satisfaction. For instance, image recognition systems may perform poorly on certain groups, which is reflected in higher risk values on said groups, and EMM seeks to learn a model that performs well across all groups. This is a serious issue with many commercial facial recognition tools [Cook et al., 2019], and furthermore, as image recognition sees increasingly use in law-enforcement, if accuracy of image recognition systems varies across groups, then false arrests can become more likely in particular groups, which again is captured by high risk for these groups [Berk et al., 2018]. Note that the malfare-minimization setup is sufficiently general that one may define arbitrary loss functions, i.e., applying reweighting, confusion matrices, or other arbitrary measures of displeasure, to capture aspects of a scenario that don't depend simply on *model accuracy*; thus despite our criticism of ad-hoc utility function construction, our setup retains the *generality* to perform equivalent *ad-hoc* loss function construction.

Perhaps the most similar to our work is a method of [Hu and Chen, 2020], wherein they *directly maximize* empirical utility over linear (halfspace) classifiers; however again an appropriate utility function must be selected. 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.

2.3 Statistical Estimation and Learning Theory

We first illustrate the ease of which p-power means can be estimated, in contrast to the standard additive welfare formulations.

Lemma 2.3.1 (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 $\forall\omega : \mathcal{S}'(\omega) - \boldsymbol{\varepsilon}(\omega) \le \mathcal{S}(\omega) \le \mathcal{S}'(\omega) + \boldsymbol{\varepsilon}(\omega)$, then with said probability, we have*

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

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 and Bennett bounds to show concentration and derive an explicit form for $\boldsymbol{\varepsilon}$.

Corollary 2.3.2 (Statistical Estimation with Hoeffding and Bennett Bounds)**.** *Suppose fair power-mean Malfare* $\Mal(\dots)$ *(i.e.,* $p \geq 1$*),* loss function $\ell : \mathcal{X} \to [0, r]$*,* $\mathcal{S} \in [0, r]^n$ *s.t.* $\mathcal{S}_i = \mathbb{E}_{\mathcal{D}_i}[\ell]$*, samples* $\boldsymbol{x}_i \sim \mathcal{D}_i^m$ *and* $\hat{\mathcal{S}}_i = \frac{1}{m}\sum_{j=1}^m \ell(\boldsymbol{x}_{i,j})$*. Then with probability at least* $1-\delta$ *over choice of* $\boldsymbol{x}$*,*

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

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

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

Proof. This result is a corollary of lemma 2.3.1, applied to $\boldsymbol{\varepsilon}$, where we note that for $p \geq 1$, by theorem 2.1.7 item 3 (contraction) it holds that

$$\mathrm{M}_p(\hat{\mathcal{S}} + \boldsymbol{\varepsilon}; \boldsymbol{w}) \leq \mathrm{M}_p(\hat{\mathcal{S}}; \boldsymbol{w}) + \|\boldsymbol{\varepsilon}\|_\infty \ \& \ \mathrm{M}_p(\mathbf{0} \vee (\hat{\mathcal{S}} - \boldsymbol{\varepsilon}); \boldsymbol{w}) \leq \mathrm{M}_p(\hat{\mathcal{S}}; \boldsymbol{w}) - \|\boldsymbol{\varepsilon}\|_\infty \ .$$

Now, for the first bound, note that we take $\boldsymbol{\varepsilon}_i \doteq r\sqrt{\frac{\ln \frac{2n}{\delta}}{2m}}$, 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 2.1.7 item 3) property.

Similarly, for the second bound, note that we take $\boldsymbol{\varepsilon}_i \doteq \frac{r \ln \frac{2n}{\delta}}{3m} + \sqrt{\frac{2\,\mathbb{V}_{\mathcal{D}_i}[\ell] \ln \frac{2n}{\delta}}{m}}$, which this time follows via Bennett's inequality and the union bound. Now, we again apply lemma 2.3.1, noting that $\mathrm{M}(\boldsymbol{\varepsilon}) \leq \mathrm{M}_\infty(\boldsymbol{\varepsilon}) = \|\boldsymbol{\varepsilon}\|_\infty$ (by *power-mean monotonicity*, item 1 item 1), and the rest follows as in the Hoeffding case. □

As corollary 2.3.2 follows directly from lemma 2.3.1, with Hoeffding and Bennett inequalities applied to derive $\boldsymbol{\varepsilon}$ bounds, similar results are immediately possible with arbitrary concentration inequalities. In particular, we can show *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 2.3.2 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, i.e., with Rademacher averages and appropriate concentration-of-measure bounds.

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).

2.3.1 Characterizing Fair Learnability

We now characterize a notion of *fair-learnability* in machine learning, reminiscent of standard *probably approximately correct (PAC) learning* guarantees of Valiant [1984]. In particular, we generalize to our definition to *arbitrary supervised learning problems* (i.e., beyond binary classification with 0-1 loss), and

to *agnostic learning* problems, and settings with non-zero Bayes risk (i.e., so-called *noisy* problems, where x does not uniquely determine y). Following the original conception of PAC-learnability, we also include *computational considerations* in our definition; i.e., it is predicated on the existence of an efficient training algorithm. Some authors, e.g., Shalev-Shwartz and Ben-David [2014, Definition 3.1] and Mitzenmacher and Upfal [2017a, Definition 14.9], consider *computational complexity* a separate issue, and instead define PAC-learnability as requiring polynomial *sample complexity*, though doing so leads to a breakdown in the learnability hierarchy (i.e., finite VC dimension, PAC-learnability, and agnostic PAC-learnability become equivalent, see Shalev-Shwartz and Ben-David [2014, thm. 6.7]). Our definition requires a computationally efficient training procedure, but this requirement may be removed by assuming a training oracle. Following our definition, we show sufficient conditions under which an algorithm is fair-PAC-learnable.

For context, we first present a generalized notion of PAC-learnability, which we then generalize to fair-PAC-learnability. In the realizable case, we assume $\exists h^* \in \mathcal{H}$ s.t. $\mathrm{R}_{\mathcal{D}}(h^*; \ell) = 0$, which restricts the space of possible probability distributions. For reasons that shall later be made clear, we consider only the *agnostic learning case*, where this assumption is dropped.

Definition 2.3.3 (Agnostic PAC-Learnability)**.** *Suppose a* hypothesis class $\mathcal{H} \subseteq \mathcal{X} \to \mathcal{Y}$*, 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* A*, such that* $\forall$*:*

1. *instance distributions* $\mathcal{D}$ *over* $\mathcal{X} \times \mathcal{Y}$*;*
2. *additive approximation error* $\varepsilon > 0$*;*
3. *failure probability* $\delta \in (0, 1)$*;*

it holds that A *can identify a hypothesis* $\hat{h} \in \mathcal{H}$ *such that*

1. A *has* $\mathrm{Poly}(1/\varepsilon, 1/\delta)$ *expected sample complexity;*
2. *with probability at least* $1 - \delta$ *(over randomness of* A*),* $\hat{h}$ *obeys*

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_i}\big[\ell(y, \hat{h}(x))\big] \leq \operatorname*{argmin}_{h\in\mathcal{H}} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_i}\big[\ell(y, h(x))\big] + \varepsilon \ .$$

Note that this substantially differs from the original presentation of PAC-learning [Valiant, 1984], which handles only *binary classification* with 0-1 loss, and we consider PAC-learnability *w.r.t.* a particular (arbitrary) loss function. Note that PAC-learnability is a property of both $\mathcal{H}$ and ℓ, and we can not hope for uniform guarantees over *arbitrary loss functions*.

We now generalize PAC-learnability to fair-PAC-learnability.

Definition 2.3.4 (Fair Agnostic PAC-Learnability)**.** *Suppose a* hypothesis class $\mathcal{H} \subseteq \mathcal{X} \to \mathcal{Y}$*, 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 n *and* loss function ℓ *if* $\exists$ *a (randomized) algorithm* A*, such that* $\forall$*:*

1. *instance distributions* $\mathcal{D}_{1:n}$ *over* $(\mathcal{X} \times \mathcal{Y})^n$*;*
2. *group weights measure* $\boldsymbol{w}$ *over* $\{1, \ldots, n\}$*;*
3. *malfare concept* $\Mal(\ldots)$ *satisfying axioms 1-7 and 9;*
4. *additive approximation error* $\varepsilon > 0$*;*
5. *failure probability* $\delta \in (0, 1)$*;*

it holds that A *can identify a hypothesis* $\hat{h} \in \mathcal{H}$ *such that*

1. A *has* $\mathrm{Poly}(1/\varepsilon, 1/\delta)$ *expected sample complexity.*

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

$$\mathbb{M}\left(i \mapsto \left(\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_i}\left[\ell(y,\hat{h}(x))\right]\right); \boldsymbol{w}\right) \leq \operatorname*{argmin}_{h\in\mathcal{H}} \mathbb{M}\left(i \mapsto \left(\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_i}\left[\ell(y,h(x))\right]\right); \boldsymbol{w}\right) + \varepsilon \ .$$

Note that this definition is extremely specific, and perhaps our axiomatization and all previous relevant results should be viewed as leading up to it. We first discuss a few highly precise definitional choices, and explain why they are equivalent to classes that perhaps seems broader or more specific at first glance. We then discuss the *motivations* for particular assumptions, and how they relate to our previous analysis of the cardinal welfare axioms and their statistical estimation properties. First note that assuming an instance distribution $\mathcal{D}_{1:n}$ over $(\mathcal{X}\times\mathcal{Y})^n$ is important to remain general over sampling strategies. In many cases, each marginal distribution $\mathcal{D}_i$ is available separately, and we sample from them *independently*, however in some learning problems, it may make more sense to sample in a dependent manner. As the definition is not sensitive to *polynomial* complexity changes, being unable to directly sample the marginal distributions is inconsequential.

We now discuss the deeper motivation and consequences of definition 2.3.4 Fair-PAC-learnability must hold *uniformly* over the choice of *instance distributions* $\mathcal{D}_{1:n}$; i.e., the concept is inherently *distribution-free*, in the sense that we require a fair algorithm to be able to (agnostically) learn fairly under any circumstances. To this end we also require uniformity over *fairness concepts* (item 2; choice of $\boldsymbol{w}$, and item 2; choice of $\mathbb{M}(\cdot;\boldsymbol{w})$); consequently, (upper-bounds to) time, space, entropy, and sample complexity must hold *for all* $p\geq 1$, i.e., they *may not* depend on p. This is perhaps our strongest use of the *unit scale* and *multiplicative linearity* axioms, without which welfare and malfare functions may be arbitrarily scaled and monotonically transformed. From our statistical bounds (lemma 2.3.1), we see that learning becomes statistically more challenging for as p increases, due to dependence on p-norms and monotonicity in p of $\mathbb{M}_p(\cdot;\boldsymbol{w})$, but statistical learning theory is well-equipped to handle even the limiting case of $p=\infty$. We soon reify this definition by showing highly generic conditions under which classifiers guarantee fair PAC-learnability.

Fair PAC-learning and computational efficiency Note that we specifically assume *fixed* n, and in the case of $n=1$, the definition coincides with that of (agnostic) PAC-learnability. Furthermore, if we wish to consider *computational efficiency*, we may strengthen condition 1 to "A has $\mathrm{Poly}(1/\varepsilon, 1/\delta)$ expected *time* complexity." We term this stronger requirement *efficient fair-PAC-learnability*, following the convention of [Shalev-Shwartz and Ben-David, 2014, Mitzenmacher and Upfal, 2017a]. This convention is convenient, as we shall see that additional conditions are required to show efficiency, so we may distinguish between *sample-efficient* and *computationally efficient* fair learning algorithms. An even stronger requirement would be to require A to work uniformly over choice of n, and require "A has $\mathrm{Poly}(1/\varepsilon, 1/\delta, n)$ expected *time* complexity."

A note on realizability The natural generalization of realizable PAC-learning to fairness is to assume $\exists h^*\in\mathcal{H}$ s.t. $\forall i\in\{1,\ldots,n\}: \mathrm{R}_{\mathcal{D}_i}(h^*;\ell)=0$, in which case no real tradeoffs exist, as all groups may be perfectly satisfied simultaneously. This h^* is thus optimal for any malfare concepts satisfying axioms 1-7 and 9, and may be identified via (standard) PAC-learning over the mixture distribution $\mathcal{D}_{1:n}$, thus realizable *fair* PAC-learning reduces to realizable PAC-learning. The concept is also philosophically uninteresting, as *scarcity begets conflict*, and only then must fairness-sensitive tradeoffs occur. We thus consider only *agnostic learning*, wherein no such mutual satisfiability assumption is made.

Sufficient conditions for fair PAC-learnability As noted by [Blumer et al., 1989], PAC-learnability and finite VC-dimension [Vapnik and Chervonenkis, 1968] are essentially equivalent (subject to basic regularity conditions). We now extend this analogy to fair-PAC-learning.

Theorem 2.3.5 (Fair PAC-Learning with Vapnik-Chervonenkis Classes and Covering Numbers)**.** *Suppose $\mathcal{H}$ has finite VC-dimension d, and the* projection *of $\mathcal{H}$ onto $\boldsymbol{x} \in \mathcal{X}^m$ can be enumerated in* $\text{Poly}(m)$ *time (wherein d is a constant). Then $\mathcal{H}$ is fair-PAC-learnable.*

Furthermore, suppose generic $\mathcal{H} \subseteq \mathcal{X} \to \mathcal{Y}$, $\ell \in \mathcal{Y} \times \mathcal{Y} \to [0, 1]$, and assume that

1. *$\mathcal{Y}$ is a set of diameter $D_{\mathcal{Y}}$, and ℓ is λ_ℓ-Lipschitz continuous;*
2. *the ℓ_2 covering number $\mathcal{N}(\mathcal{H}, \boldsymbol{x}, \gamma) \in \text{Poly}(m, 1/\gamma)$;*
3. *a* γ-precision cover *$\mathcal{C}(\mathcal{H}, \boldsymbol{x}, \gamma)$ of size $\mathbf{O}(\mathcal{N}(\mathcal{H}, \boldsymbol{x}, \gamma))$ may be enumerated in* $\text{Poly}(m, 1/\gamma)$ *time.*

Then $\mathcal{H}$ is efficiently fair-PAC-learnable. Items (1) and (2) are sufficient to show that $\mathcal{H}$ is fair-PAC learnable; (3) is required only to show efficiency*.*

Proof Sketch. First, note that the Vapnik-Chervonenkis conditions on $\mathcal{H}$ immediately imply that, given samples $\boldsymbol{x}_{1:n}$, each of size m, each covering number $\mathcal{N}(\mathcal{H}, \boldsymbol{x}_i, \gamma) \in \text{Poly}(m, \frac{1}{\gamma}) \leq \sum_{i=1}^{d} \binom{m}{i} \leq m^d$, by the Sauer-Shelah lemma, where the VC-dimension d is *constant* (w.r.t. fixed $\mathcal{H}$), and also, by assumption, this cover may be enumerated in $\text{Poly}(m) = \text{Poly}(m, 1/\gamma)$ time. The first case thus reduces to a special case of the second.

I now sketch a proof of the general case. We first take *cover precision* $\gamma \doteq \frac{\varepsilon}{3D_{\mathcal{Y}}\lambda_\ell}$. We then show that, taking $\hat{h}$ to be the EMM solution w.r.t. *a cover* of this precision, and $\tilde{h}$ to be the exact EMM solution w.r.t. *all of* $\mathcal{H}$, it holds that the empirical malfare of $\hat{h}$ does not exceed that of $\tilde{h}$ by more than $\frac{\varepsilon}{3}$. We then apply standard covering-number bounds on Rademacher averages to show that a (polynomial) sufficient m exists to guarantee $\tilde{h}$ generalizes with $\frac{2\varepsilon}{3}$-δ malfare error guarantees. In particular, by the Dudley discretization method, a $\frac{\varepsilon}{3}$ *discretization error* term appears, and we set m such that a $\frac{\varepsilon}{3}$ covering number term appears, for a total of $\frac{2\varepsilon}{3}$. Combining these bounds on the *optimization* and *approximation* errors yields the desired additive ϵ-δ guarantees.

Note that (3) implies that the cover may be efficiently enumerated, thus guaranteeing polynomial time complexity. □

This immediately implies $\mathcal{H}$ that are *finite*, of bounded *pseudodimension*, or bounded *fat-shattering dimension* are fair-PAC-learnable. For instance, this gives us *classifiers* such as *all possible languages* of *Boolean formulae* over d variables, or *halfspaces* (i.e., linear hard classifiers $\mathcal{H} \doteq \{\vec{x} \mapsto \text{sgn}(\vec{x} \cdot \vec{w}) \mid \vec{w} \in \mathbb{R}^d\}$), as well as *generalized linear models*, 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 fair-PAC learnable, but it says nothing about their performance as $d \to \infty$.

Note also that theorem 2.3.5 leverages *covering arguments* in both their *statistical* and *computational* capacity. 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, optimization methods (e.g., gradient descent, 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.

Despite these limitations, this theorem characterizes a large class in which fair learning is tractable (both statistically and computationally). In particular, it contains VC, bounded pseudodimension, and bounded fat-shattering dimension classes in which ε-ERM is efficient.

We now show alternative meta-conditions, which leverage convex-optimization tools to derive highly-general EMM algorithms and show fair-PAC learnability. This is the most practical of our fair-PAC-learnability results, as the algorithm employed by the constructive proof is quite efficient, and we

recommend its use in practice. However, it is also the most restrictive, as we must assume a lot about the structure of $\mathcal{H}$ to ensure efficient convex optimization.

Theorem 2.3.6 (Fair PAC-Learning with Convex Optimization)**.** *Suppose $\mathcal{H}$ is agnostic-PAC-learnable, and $\exists$ algorithm A to agnostic-PAC-learn $\mathcal{H}$ w.r.t. some loss function ℓ, where A simply applies a generic* optimization routine *O that requires only that the objective is* bounded*,* smooth*, and* convex*. Then $\mathcal{H}$ is fair-PAC learnable w.r.t. ℓ.*

Proof Sketch. We show this result constructively. By assumption, we know there exist A, O, such that A PAC-learns $\mathcal{H}$ by drawing a polynomially large sample $(\boldsymbol{x}, \boldsymbol{y})$, and applying O to optimize empirical risk. In our construction, we instead optimize empirical malfare, and the properties of A ensure efficient approximate optimality. □

To close this section, we now show a convenient, generic, and modular constructive characterization that yields fair-PAC-learnability for an extremely broad class of learning problem. The constructive training algorithm is also *efficient* for fixed group-count n, though its runtime is exponential in n. This construction makes no structural assumptions about $\mathcal{H}$, and assumes only that loss values of ERM solutions are *stable* under small perturbations to training data (i.e., no structural assumptions, e.g., convexity on $\mathcal{H}$, nor assumptions on the structure of the PAC-learning algorithm, are required). Then if $\mathcal{H}$ is PAC-learnable w.r.t. ℓ, it is also fair-PAC-learnable.

Theorem 2.3.7 (Fair PAC-Learning with Grid Reweighting)**.** *Suppose loss function ℓ and hypothesis class $\mathcal{H}$, such that for any $\boldsymbol{x} \in \mathcal{X}^m$, $\boldsymbol{x}' = \boldsymbol{x}_1, \ldots, \boldsymbol{x}_m, x'$, taking ERM $\hat{h}, \hat{h}'$ over $\boldsymbol{x}, \boldsymbol{x}'$, respectively, it holds that, for some* stability constant *B we have*

$$\hat{\mathrm{R}}(\hat{h}; \boldsymbol{x}) - \hat{\mathrm{R}}(\hat{h}'; \boldsymbol{x}) \leq \tfrac{B}{m} \ .$$

Then if $\mathcal{H}$ is PAC-learnable w.r.t. ℓ, then $\mathcal{H}$ fair PAC-learnable for bounded *n w.r.t. ℓ. Furthermore, if $\mathcal{H}$ is efficiently-PAC-learnable w.r.t. ℓ, then $\mathcal{H}$ fair PAC-learnable* for bounded *n w.r.t. ℓ.*

Proof Sketch. The strategy here is to create a *grid-cover* of *reweightings* of the n groups. We construct a cover of reweightings of the n groups, and show the $\mathcal{S}$-weighted malfare is Lipschitz continuous in reweighting $\mathcal{S}'$. We then PAC-learn each reweighting, and take the empirically-malfare-minimal solution.

Note that the Lipschitz or stability condition $\hat{\mathrm{R}}(\hat{h}; \boldsymbol{x}) - \hat{\mathrm{R}}(\hat{h}'; \boldsymbol{x}) \leq \frac{B}{m}$ is required to bound the size of the grid-cover; without this assumption, we can't guarantee that a finite-grid is sufficient. □

Nonlearnability and a warning in closing We caution against blindly applying such methods in other settings without learnability guarantees, such as learning with *local optimization* or *proxy losses*, without carefully considering the fairness implications. For instance it may be the case that training an SVM with *hinge loss* as a proxy for *ramp loss* or *0-1 loss* results in a few high-hinge-loss outliers for group A, thus *unfairly* incentivizing better treatment of group A, at the cost of performance for group B. Similarly, optimizing a neural network may converge to a *locally optimal* solution that favors some *easily mutually satisfied* subset of groups, and thus may terminate before identifying a solution that works for all groups. In this sense, without learnability guarantees, we are unable to ensure fairness, making similar claims of fairness in these settings naively optimistic at best, and shallow platitudes at worst.

Chapter 3

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

3.1 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, 2017b], 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.

3.2 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* $\mathbb{R}_{0+}^a$. 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 $\mathcal{L} \subseteq \mathbb{R}_{0+}^a \to \mathbb{R}_{0+}$ 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*

$$\mathrm{R}(\boldsymbol{c}; \ell, \mathcal{D}) \doteq \mathop{\mathbb{E}}_{x \sim \mathcal{D}} \left[\min_{c \in \boldsymbol{c}} (\ell \circ c)(x) \right] . \tag{3.2.1}$$

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

3.2.1 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

$$\boldsymbol{c}^*_\ell \doteq \operatorname*{argmin}_{\boldsymbol{c} \in \binom{\mathcal{C}}{k}} \mathrm{R}(\boldsymbol{c}; \ell, \mathcal{D}) \ . \tag{3.2.2}$$

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 over $\mathcal{L}$. Welfare concepts capture various aspects of multivariate 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 2.1.6. 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

$$\boldsymbol{c}^*_{\mathrm{M}} \doteq \operatorname*{argmin}_{\boldsymbol{c} \in \binom{\mathcal{C}}{k}} \mathrm{M}(\boldsymbol{c}; \mathcal{L}, \mathcal{D}) \ . \tag{3.2.3}$$

3.2.2 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*

$$\hat{\mathrm{R}}(\boldsymbol{c}; \ell, \boldsymbol{x}) \doteq \hat{\mathbb{E}}_{x \in \boldsymbol{x}} \left[\min_{c \in \boldsymbol{c}} (\ell \circ c)(x) \right] \ , \tag{3.2.4}$$

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

$$\hat{\boldsymbol{c}}_\ell \doteq \operatorname*{argmin}_{\boldsymbol{c} \in \binom{\mathcal{C}}{k}} \hat{\mathrm{R}}(\boldsymbol{c}; \ell, \boldsymbol{x}) \ , \tag{3.2.5}$$

corresponding to eqs. 3.2.1 and 3.2.2, 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 3.2.3, as

$$\hat{\boldsymbol{c}}_{\mathrm{M}} \doteq \operatorname*{argmin}_{\boldsymbol{c}\in\binom{\mathcal{C}}{k}} \hat{\mathrm{M}}(\boldsymbol{c};\mathcal{L},\boldsymbol{x}) \quad . \tag{3.2.6}$$

The *Empirical Welfare-Optimal k codec Selection* (EWOkS) 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 EWOkS 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 3.2.3 we show tail bounds on the welfare-estimation error, thus controlling for overfitting and providing methods to guarantee that EWOkS solutions are statistically significant.

Note that k is a fixed hyperparameter of the EWOkS 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].

3.2.3 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

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

$\mathcal{F}$ 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 EWOkS solutions.

Definition 3.2.1 (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

$$\mathop{\mathbb{E}}_{\boldsymbol{x}}\Big[\underbrace{\sup_{f\in\mathcal{F}} \left| \mathop{\mathbb{E}}_{\mathcal{D}}[f] - \hat{\mathop{\mathbb{E}}}_{\boldsymbol{x}}[f] \right|}_{\text{Supremum Deviation}} \Big] \leq 2\mathfrak{R}_m(\mathcal{F},\mathcal{D}) \ , \tag{3.2.7}$$

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

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

$$4\hat{\mathfrak{R}}^n_m(\mathcal{F},\boldsymbol{x},\boldsymbol{\sigma}) + \boldsymbol{\Theta}\left(\frac{r\ln\frac{1}{\delta}}{m} + \sqrt{\frac{(v + r\mathfrak{R}_m(\mathcal{F},\mathcal{D}))\ln\frac{1}{\delta}}{m}}\right) . \tag{3.2.8}$$

With worst-case variance $v = r^2/4$, we recover the slow-decaying $\boldsymbol{\Theta}\sqrt{r^2\ln\frac{1}{\delta}/m}$ McDiarmid [1989] 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 3.3.3).

3.2.4 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

$$\mathcal{L}_p \doteq \{\, \ell(\boldsymbol{a}) \doteq \boldsymbol{w} \cdot \boldsymbol{a} \mid \boldsymbol{w} \in \mathbb{R}^a_{0+} \text{ s.t. } \|\boldsymbol{w}\|_p \leq 1 \,\} \tag{3.2.9}$$

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 3.2.3 (Bounds for Linear Loss Families)**.** *Suppose p, q, s.t. $|p^{-1}| + |q^{-1}| = 1$, codec family $\mathcal{C}$, $\mathcal{L}_p$ as in equation* 3.2.9*, 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*

$$\hat{v}_1 = \sup_{c\in\mathcal{C}} \sup_{\ell\in\mathcal{L}} \hat{\mathbb{V}}_{\boldsymbol{x}}[\ell \circ c] \leq \sup_{c\in\mathcal{C}} \big\|\hat{\mathbb{C}}_{\boldsymbol{x}}[c]\big\|_{q\to p} \;,\; \& \quad \hat{\mathfrak{R}}^n_m(\mathcal{F}_1, \boldsymbol{x}, \boldsymbol{\sigma}) \leq \tfrac{1}{n}\sum_{j=1}^{n} \sup_{c\in\mathcal{C}} \left\| \tfrac{1}{m}\sum_{i=1}^{m} \boldsymbol{\sigma}_{j,i} c(\boldsymbol{x}_i) \right\|_q \;.$$

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

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

Here we see that linearity of $\mathcal{L}_p$ is convenient, though it *does not* in general convert the EWOKS 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 3.2.3 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 3.2.3 with a simple corollary for the ℓ_2 linear loss family.

Corollary 3.2.4 (Bounds for Linear Loss Families)**.** *Suppose as in theorem 3.2.3, and also $p = q = 2$, and all attribute values are contained by the unit sphere. Then*

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

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

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

Consequently, theorem 3.2.2 holds with

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

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

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 3.2.4 & theorem 3.2.2 imply that for 1-Lipschitz welfare $\mathrm{M}(\cdot)$, $r \in \boldsymbol{\Theta}(1)$, with pr. $\geq 1-\delta$, we have $\mathrm{M}(\hat{\boldsymbol{c}}_{\mathrm{M}}, \mathcal{L}, \mathcal{D}) - \mathrm{M}(\boldsymbol{c}^*_{\mathrm{M}}, \mathcal{L}, \mathcal{D}) \leq$

$$4k\hat{\mathfrak{R}}^{n}_{m}(\mathcal{F}_1, \boldsymbol{x}, \boldsymbol{\sigma}) + \boldsymbol{\Theta}\left(\frac{\ln\frac{1}{\delta}}{m} + \sqrt{\frac{k(v_1 + \mathfrak{R}_m(\mathcal{F}_1, \mathcal{D}))\ln\frac{1}{\delta}}{m}}\right). \tag{3.2.10}$$

3.3 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 3.1 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 $[0, 1]$.

3.3.1 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 3.2.9) $\ell \in \mathcal{L}_2$ for which the risk (equation 3.2.1) of $\boldsymbol{c}$ is minimal among $\binom{\mathcal{C}}{k}$, i.e.,

$$\mathrm{Pareto}\big(\tbinom{\mathcal{C}}{k}, \mathcal{D}\big) \doteq \bigcup_{\ell \in \mathcal{L}_2} \operatorname*{argmin}_{\boldsymbol{c} \in \binom{\mathcal{C}}{k}} \mathrm{R}(\boldsymbol{c}; \ell, \mathcal{D}) \ . \tag{3.3.1}$$

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

$$\mathrm{PFrnt}(\mathcal{C}, \mathcal{D}) \doteq \mathrm{LCH}\Big\{ \mathop{\mathbb{E}}_{x \sim \mathcal{D}}[c(x)] \ \Big| \ c \in \mathcal{C} \Big\} \ . \tag{3.3.2}$$

This notion is easily visualized (see figure 3.1), 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

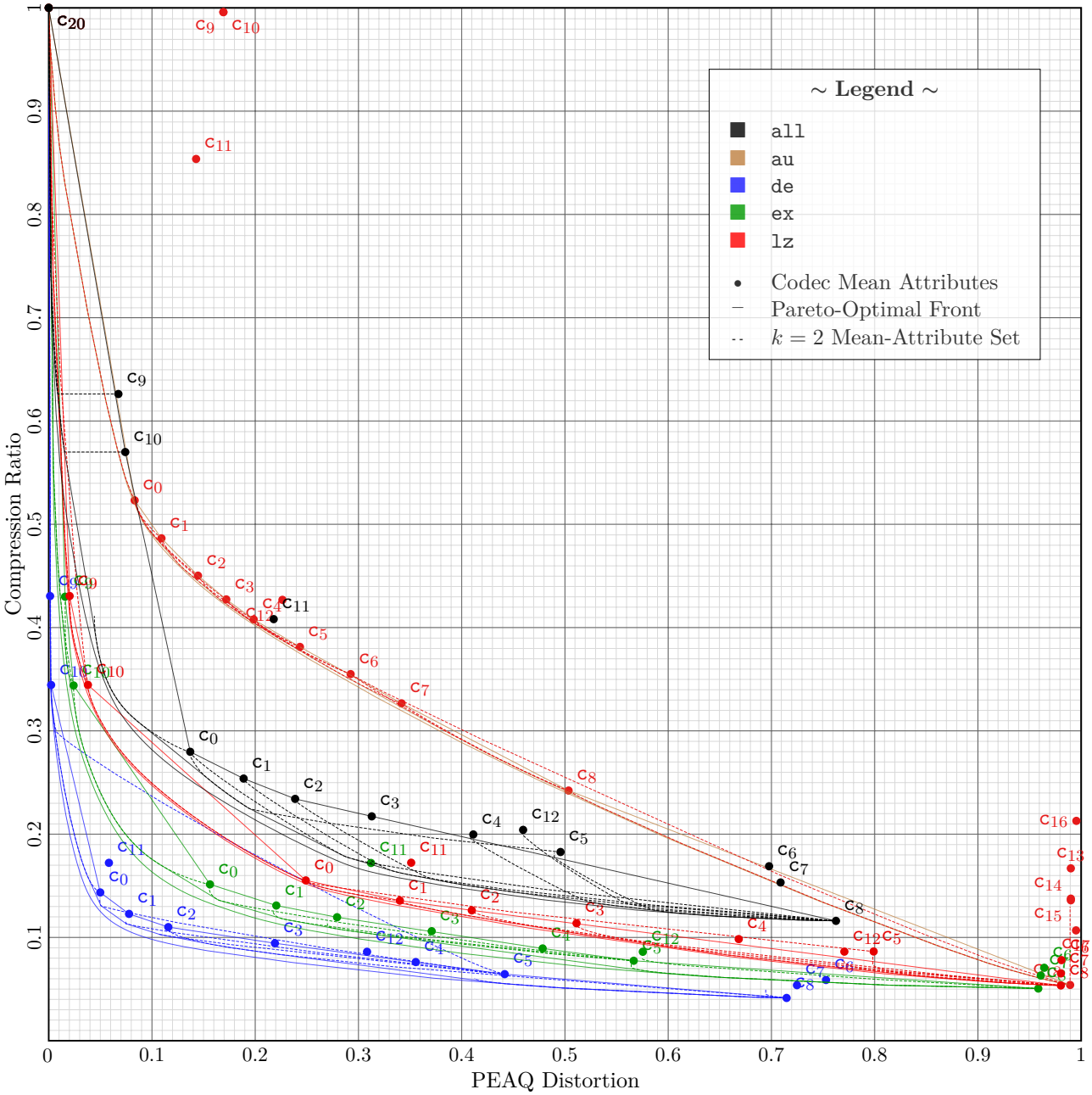

Figure 3.1: Pareto-optimal frontiers for all datasets. Mean distortion (x axis) and compression ratio (y axis) shown for each codec •, with Pareto-optimal frontiers for codec families of sizes $k \in \{1, 2, |\mathcal{C}|\}$, color-coded by dataset. Curves for each k are unambiguous, since k *increases* from top-right to bottom-left. Mean-attribute sets for Pareto-optimal $k = 2$ pairs are shown in the inset as dotted curves.

Table 3.2: 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**

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

$$\bigcup_{\boldsymbol{c} \in \binom{\mathcal{C}}{k}} \left\{ \mathop{\mathbb{E}}_{x \sim \mathcal{D}} \left[\mathop{\mathrm{argmin}}_{c(x) \in \boldsymbol{c}(x)} \ell(c(x)) \right] \,\middle|\, \ell \in \mathcal{L}_2 \right\} \quad . \tag{3.3.3}$$

Observe that equation 3.3.3 generalizes equation 3.3.2, 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 3.2.1, 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 3.2.3 to bound the gap between empirical and true Pareto-optimaltity concepts.

In figure 3.1, 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.

3.3.2 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 be required to well-optimize many loss functions *simultaneously*; consequently, we expect various *welfare concepts* to continue improving beyond $k = 2$. In table 3.2 & figure 3.2, 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 3.2.9), 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 3.2.9) $\ell_{1/2}(\boldsymbol{a}) \doteq (1/2, 1/2) \cdot \boldsymbol{a}$, *utilitarian welfare* $\mathrm{M}_1(\cdot)$ on the loss distribution over $\mathcal{L}_1$ with density

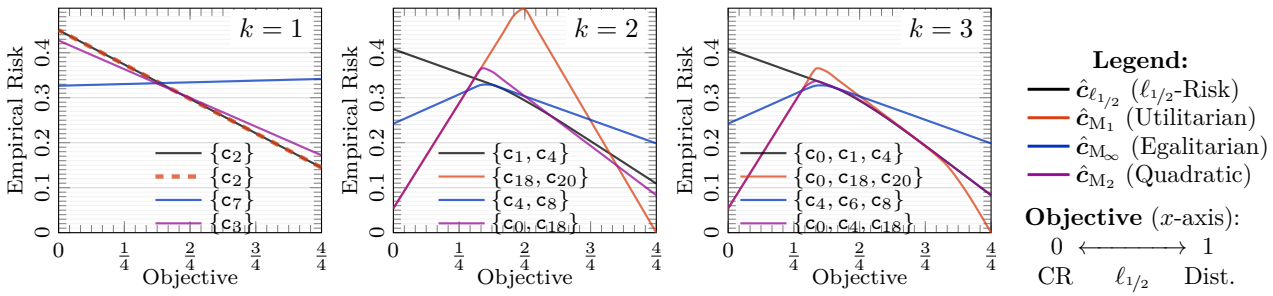

Figure 3.2: Empirical risk (y-axis) plotted against linear objective family $\mathcal{L}_1$ (x-axis), with *distortion weight* $\boldsymbol{w}_1 = x$ and *compression ratio weight* $\boldsymbol{w}_2 = 1 - x$, on au, for $k \in \{1, 2, 3\}$. Empirically-optimal codec sets $\hat{\boldsymbol{c}}_o \in \binom{\mathcal{C}}{k}$ for each objective o are also given.

$\propto |\boldsymbol{w}_1 - 1/2|^2$ (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 3.2a shows the raw empirical objective values of each optimality concept, for $k \in \{1, 2, 3, |\mathcal{C}|\}$, and table 3.2b 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 3.2 plots the *empirical risk* (see equation 3.2.4) 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.

From the *objective values* matrix in table 3.2, 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 3.2a). 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 3.2b) 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 vs 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.

3.3.3 Uniform Convergence Bounds

We now show that our uniform convergence bounds yield valuable conclusions as to the approximate optimality of EWOkS solutions, and outperform McDiarmid bounds on real data. In figure 3.3, we plot EWOkS 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 equation 3.2.9), $n = 100$ 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 3.2.2) are significantly sharper than the McDiarmid bounds, yielding valuable conclusions after only $m \approx 1000$ samples. The *mixed convergence rates* (see theorem 3.2.2) 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

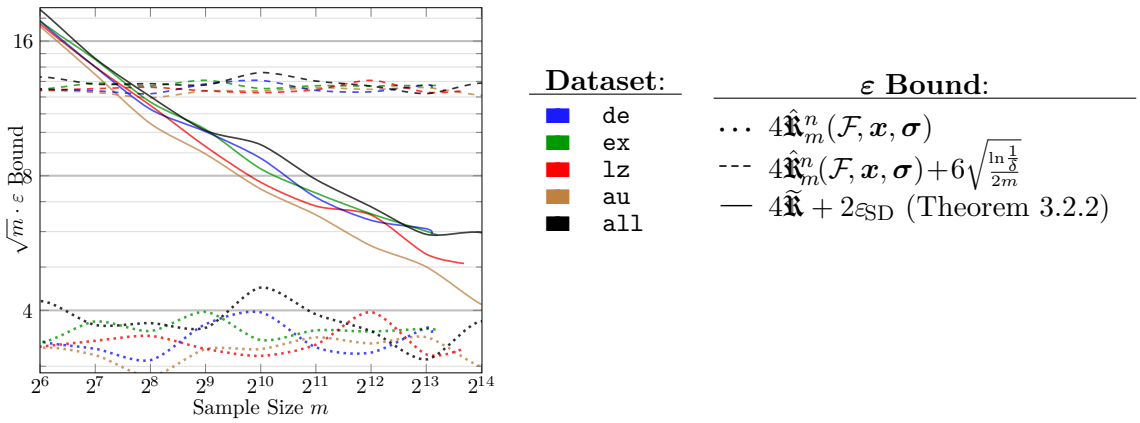

Figure 3.3: Log-log plots of EWOkS $\sqrt{m}$-scaled M$(\cdot)$-optimality gap bound $\sqrt{m}\varepsilon$ (x-axis), vs. sample size m (y-axis), in expectation with $4\hat{\mathfrak{R}}^n_m(\mathcal{F}, \boldsymbol{x}, \boldsymbol{\sigma})$, and w.h.p. with McDiarmid and theorem 3.2.2.

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 $\widetilde{\mathfrak{R}}$).

3.4 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* (EWOkS) 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 3.3.1), particularly for heterogenous (mixed speech / music) data, it greatly improves *welfare-objectives* over *loss families* (section 3.3.2). 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 EWOkS (section 3.2.3), 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 3.3.3) 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 4

Conclusion

Chapter 1 introduces new statistical techniques for variance-insensitive uniform convergence bounds. The centralization strategy achieves asymptotically-optimal 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.

Following the pure statistical methods of chapter 1, in chapter 2, I define *malfare* parallel to *welfare* (definition 2.1.1), and show that subject to several intuitive and basic axioms (section 2.1.1), all welfare and malfare are *power means* (definition 2.1.6, see theorem 2.1.8). I then argue that fair ML should seek to *minimize malfare*, and characterize learnability as such in the fair PAC-learning framework (definition 2.3.4). I then combine the uniform convergence guarantees of chapter 1 with the fairness setting of chapter 2 in chapter 3 the *fair codec selection problem*. In the full thesis, I will expand upon the fair-PAC-learnability setting, and better characterize necessary and sufficient conditions for various flavors of learnability. I will additionally describe practical efficient optimization and data-dependent sample-complexity bounds for *fair generalized linear models*, as well as extending the methodology to other fairness-sensitive areas, such as *mechanism design* and *fair machine learning*.

The full thesis will contain many additional applications of concentration of measure and uniform convergence guarantees in data science, as well as additional fairness applications in generalized linear models and reinforcement learning. In particular, I shall present results on *mean estimation* in *block databases* with arbitrary dependence structure, and for *Monte Carlo Markov Chain* processes; in both case using optimal empirical variance-sensitive bounds to reduce dependence on *a priori* knowledge, and achieve asymptotically optimal (Gaussian-type) rates, with finite-sample guarantees. I will also present results on improved sample-complexity guarantees for estimating *frequent itemsets* [Pellegrina et al., 2020] and *betweenness centrality*.

I have additionally worked in several areas that don't fit cleanly into this thesis, *anomaly detection* [Cousins et al., 2017], *automated machine learning* [Binnig et al., 2015, 2018], *empirical game theory* and *mechanism design* [Viqueira and Cousins, Viqueira et al., 2020]. Of special note are two somewhat more theoretical papers in pure machine learning. First [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; the analysis uses the *half-binomial expectation*, which also plays a key role in the analysis of centralized Rademacher averages, particularly in lemma 1.2.1. Second [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.

Bibliography

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.

Anthony B Atkinson et al. On the measurement of inequality. *Journal of Economic Theory*, 2(3):244–263, 1970.

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.

George Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57(297):33–45, 1962.

Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, page 0049124118782533, 2018.

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.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the Vapnik-Chervonekis dimension. *Journal of the Association for Computing Machinery*, 36(4):929–965, 1989.

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.

Peter S Bullen. *Handbook of means and their inequalities*, volume 560. Springer Science & Business Media, 2013.

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 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 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.

Hugh Dalton. The measurement of the inequality of incomes. *The Economic Journal*, 30(119):348–361, 1920.

Gerard Debreu. Topological methods in cardinal utility theory. Technical report, Cowles Foundation for Research in Economics, Yale University, 1959.

Luc Devroye, Matthieu Lerasle, Gábor Lugosi, and Roberto I. Oliveira. Sub-Gaussian mean estimators. *The Annals of Statistics*, 44(6):2695–2725, 2016.

Robert Gibbons. *Game theory for applied economists*. Princeton University Press, 1992.

Evarist Giné and Vladimir Koltchinskii. Concentration inequalities and asymptotic results for ratio type empirical processes. *The Annals of Probability*, 34(3):1143–1216, 2006.

William M Gorman. The structure of utility functions. *The Review of Economic Studies*, 35(4):367–390, 1968.

Uffe Haagerup. The best constants in the Khintchine inequality. *Studia Mathematica*, 70(3):231–283, 1982.

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.

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.

Maximilian Kasy and Rediet Abebe. Fairness, equality, and power in algorithmic decision making. Technical report, Working paper, 2020.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent Trade-Offs in the Fair Determination of Risk Scores. In Christos H. Papadimitriou, editor, *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, volume 67 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 43:1–43:23, Dagstuhl, Germany, 2017. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. ISBN 978-3-95977-029-3. doi: 10.4230/LIPIcs.ITCS.2017.43. URL `http://drops.dagstuhl.de/opus/volltexte/2017/8156`.

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.

Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019.

Pascal Massart. Some applications of concentration inequalities to statistics. In *Annales-Faculte des Sciences Toulouse Mathematiques*, volume 9, pages 245–303. Université Paul Sabatier, 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.

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.

Michael Mitzenmacher and Eli Upfal. *Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis*. Cambridge university press, second edition, 2017a.

Michael Mitzenmacher and Eli Upfal. *Probability and Computing: Randomization and Probabilistic Techniques in Algorithms and Data Analysis*. Cambridge University Press, second edition edition, 2017b.

Hervé Moulin. *Fair division and collective welfare*. MIT Press, 2004.

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*. ACM, 2020.

Arthur Cecil Pigou. *Wealth and welfare*. Macmillan and Company, limited, 1912.

David Pollard. *Convergence of Stochastic Processes*. Springer, New York, 1984.

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.

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.

Amartya Sen. On weights and measures: informational constraints in social welfare analysis. *Econometrica: Journal of the Econometric Society*, pages 1539–1572, 1977.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

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.

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.

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.

Vladimir Vapnik. Principles of risk minimization for learning theory. In *Advances in neural information processing systems*, pages 831–838, 1992.

Vladimir Naumovich Vapnik and Aleksei Yakovlevich Chervonenkis. The uniform convergence of frequencies of the appearance of events to their probabilities. In *Doklady Akademii Nauk*, volume 181, pages 781–783. Russian Academy of Sciences, 1968.

Enrique Areyan Viqueira and Cyrus Cousins. Learning simulation-based games from data. In *Proceeding AAMAS'19 Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*.

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.

Muhammad Bilal Zafar, Isabel Valera, Manuel Rodriguez, Krishna Gummadi, and Adrian Weller. From parity to preference-based notions of fairness in classification. In *Advances in Neural Information Processing Systems*, pages 229–239, 2017.

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

$$\mathfrak{R}_m(\mathrm{C}_{\mathcal{D}}(\mathcal{F}), \mathcal{D}) \leq \mathop{\mathbb{E}}_{\boldsymbol{x}}\Big[\hat{\mathfrak{R}}_m(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}), \boldsymbol{x})\Big] + \mathop{\mathbb{E}}_{\boldsymbol{x}}\Big[\hat{\mathfrak{R}}_m(\mathcal{K}_{\boldsymbol{x}}, \boldsymbol{x})\Big] \ . \tag{A.1.1}$$

Using equation 1.2.1 and the linearity of expectation we have that, for each $\boldsymbol{x} \in \mathcal{X}^m$, it holds

$$\hat{\mathfrak{R}}_m(\mathcal{K}_{\boldsymbol{x}}, \boldsymbol{x}) = \sup_{f \in \mathcal{F}} |\hat{\mathop{\mathbb{E}}_{\boldsymbol{x}}}[f] - \mathop{\mathbb{E}}_{\mathcal{D}}[f]| \mathrm{b}(m) = \mathsf{SD}(\mathcal{F}, \boldsymbol{x}) \mathrm{b}(m) = \mathsf{SD}(\mathrm{C}_{\mathcal{D}}(\mathcal{F}), \boldsymbol{x}) \mathrm{b}(m), \tag{A.1.2}$$

where in the last step we use the fact that the SD is invariant to shifting of functions. Continuing from equation A.1.1 and using equation A.1.2 and the rightmost inequality of equation 1.1.4, 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 equation 1.2.1), 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

$$\mathop{\mathbb{E}}_{\boldsymbol{x}}\big[\hat{\mathfrak{R}}_m(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}), \boldsymbol{x})\big] = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}, \boldsymbol{x}}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{\sigma}_i\Big(\big(f(\boldsymbol{x}_i) - \mathop{\mathbb{E}}_{\mathcal{D}}[f]\big) + \big(\mathop{\mathbb{E}}_{\mathcal{D}}[f] - \mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}}[f]\big)\Big)\right|\right] \ .$$

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

$$\begin{aligned}\mathop{\mathbb{E}}_{\boldsymbol{x}}\big[\hat{\mathfrak{R}}_m(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}), \boldsymbol{x})\big] \leq \mathop{\mathbb{E}}_{\boldsymbol{\sigma}, \boldsymbol{x}}&\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{\sigma}_i\big(f(\boldsymbol{x}_i) - \mathop{\mathbb{E}}_{\mathcal{D}}[f]\big)\right|\right] \\ &+ \mathop{\mathbb{E}}_{\boldsymbol{\sigma}, \boldsymbol{x}}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{\sigma}_i\big(\mathop{\mathbb{E}}_{\mathcal{D}}[f] - \mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}}[f]\big)\right|\right] \ . \end{aligned} \tag{A.1.3}$$

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

$$\mathcal{Z}_{\boldsymbol{x}} \doteq \{x \mapsto \mathop{\mathbb{E}}_{\mathcal{D}}[f] - \mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}}[f], f \in \mathcal{F}\} \ .$$

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

$$\hat{\mathfrak{R}}_m(\mathcal{Z}_{\boldsymbol{x}}, \boldsymbol{x}) = \mathsf{SD}(\mathrm{C}_{\mathcal{D}}(\mathcal{F}), \boldsymbol{x})\mathrm{b}(m) \ . \tag{A.1.4}$$

Continuing from equation A.1.3 and using equation A.1.4 and the rightmost inequality of equation 1.1.4, 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*

$$\mathop{\mathbb{E}}_{\boldsymbol{x}}\big[\mathrm{Z}(\boldsymbol{x})\big] \leq \mathrm{Z}(\boldsymbol{x}) + \alpha\gamma\ln\frac{1}{\delta} + \sqrt{\left(\alpha\gamma\ln\frac{1}{\delta}\right)^2 + 2\gamma(\alpha\mathrm{Z}(\boldsymbol{x}) + \beta)\ln\frac{1}{\delta}} \ . \tag{A.1.5}$$

Additionally, when $\alpha = 1$, *we may improve the constants to*

$$\mathop{\mathbb{E}}_{\boldsymbol{x}}\big[\mathrm{Z}(\boldsymbol{x})\big] \leq \mathrm{Z}(\boldsymbol{x}) + \frac{2}{3}\gamma\ln\frac{1}{\delta} + \sqrt{\left(\frac{1}{\sqrt{3}}\gamma\ln\frac{1}{\delta}\right)^2 + 2\gamma(\mathrm{Z}(\boldsymbol{x}) + \beta)\ln\frac{1}{\delta}} \ . \tag{A.1.6}$$

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 A.1.5. 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]

$$\mathbb{P}\left(\mathrm{Z}(\boldsymbol{x}) \leq \mathop{\mathbb{E}}_{\boldsymbol{x}}[\mathrm{Z}(\boldsymbol{x})] - \varepsilon\right) \leq \exp\left(\frac{-\varepsilon^2}{2(\alpha\,\mathbb{E}_{\boldsymbol{x}}[\mathrm{Z}(\boldsymbol{x})] + \beta)}\right) \ . \tag{A.1.7}$$

Now, taking δ equal to the RHS of equation A.1.7, 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 A.1.6 (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.

□

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*

$$\mathop{\mathbb{E}}_{\boldsymbol{x}}[\hat{\mathfrak{R}}_m(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}), \boldsymbol{x})] \leq \hat{\mathfrak{R}}_m(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}), \boldsymbol{x}) + \frac{2r\chi\ln\frac{1}{\delta}}{3m} + \sqrt{\left(\frac{r\chi\ln\frac{1}{\delta}}{\sqrt{3}m}\right)^2 + \frac{2r\chi(\hat{\mathfrak{R}}_m(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}), \boldsymbol{x}) + r\mathrm{b}(m))\ln\frac{1}{\delta}}{m}} \ . \tag{1.2.3}$$

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 equation A.1.6 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

$$\sup_{f\in\mathcal{F}}\frac{1}{m}\left|\sum_{i=1}^{m}\boldsymbol{\sigma}_i\left(f(\boldsymbol{x}_i) - \mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}_{\setminus j}}[f]\right)\right| .$$

Thus, using the law of total expectation, we can write

$$\mathrm{Y}_j(\boldsymbol{x}) \leq \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\left|\sum_{i=1}^{m}\boldsymbol{\sigma}_i\left(f(\boldsymbol{x}_i) - \mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}_{\setminus j}}[f]\right)\right|\right] .$$

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

$$\mathrm{Y}_j(\boldsymbol{x}) \leq \underbrace{\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\left|\sum_{i=1}^{m}\boldsymbol{\sigma}_i\left(f(\boldsymbol{x}_i) - \mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}}[f]\right)\right|\right]}_{=\mathrm{Y}(\boldsymbol{x})} + \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\left|\sum_{i=1}^{m}\boldsymbol{\sigma}_i\left(\mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}}[f] - \mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}_{\setminus j}}[f]\right)\right|\right] . \tag{A.1.8}$$

The first term on the r.h.s. is $\mathrm{Y}(\boldsymbol{x})$. The second term is the ERA of the sample-dependent family

$$\mathcal{W}_{\boldsymbol{x}} \doteq \left\{ y \mapsto \frac{1}{m}(f(\boldsymbol{x}_j) - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f]), f \in \mathcal{F} \right\} \ .$$

Each function in $\mathcal{W}_{\boldsymbol{x}}$ is *constant*. Using equation 1.2.1 and the linearity of expectation, like we did in the proof of lemma 1.2.1 for the family $\mathcal{K}_{\boldsymbol{x}}$ (see equation A.1.2), it holds

$$\hat{\mathfrak{R}}_m(\mathcal{W}_{\boldsymbol{x}}, \boldsymbol{x}) = \frac{1}{m} \sup_{f \in \mathcal{F}} |f(\boldsymbol{x}_j) - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f]| \mathrm{b}(m) \leq \frac{r}{m} \mathrm{b}(m) \ .$$

Thus, continuing from equation A.1.8 by incorporating the above fact, it holds

$$\mathrm{Y}_j(\boldsymbol{x}) \leq \mathrm{Y}(\boldsymbol{x}) + \frac{r}{m} \mathrm{b}(m) \ . \tag{A.1.9}$$

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

$$\mathrm{Y}_j(\boldsymbol{x}) = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \left| \left(\sum_{\substack{i=1 \\ i \neq j}}^{m} \boldsymbol{\sigma}_i \left(f(\boldsymbol{x}_i) - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f] \right) \right) + \boldsymbol{\sigma}_j \big(f(\boldsymbol{x}_j) - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f] \big) - \boldsymbol{\sigma}_j \big(f(\boldsymbol{x}_j) - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f] \big) \right| \right] \ .$$

Then, from the triangle inequality and the fact that

$$\sup_{f \in \mathcal{F}} \big| \boldsymbol{\sigma}_j \big(f(\boldsymbol{x}_j) - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f] \big) \big| \leq r,$$

we obtain

$$\mathrm{Y}_j(\boldsymbol{x}) \geq \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \left| \sum_{i=1}^{m} \boldsymbol{\sigma}_i \left(f(\boldsymbol{x}_i) - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f] \right) \right| \right] - \frac{r}{m} \ .$$

From here, we add and subtract $\sigma_i \hat{\mathbb{E}}_{\boldsymbol{x}}[f]$ to each term of the sum, and then use the triangle inequality, the subadditivity of the supremum, and the linearity of expectation, to obtain

$$\mathrm{Y}_j(\boldsymbol{x}) \geq \underbrace{\mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \left| \sum_{i=1}^{m} \boldsymbol{\sigma}_i \left(f(\boldsymbol{x}_i) - \hat{\mathbb{E}}_{\boldsymbol{x}}[f] \right) \right| \right]}_{= \mathrm{Y}(\boldsymbol{x})} - \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \left| \sum_{i=1}^{m} \boldsymbol{\sigma}_i \left(\hat{\mathbb{E}}_{\boldsymbol{x}}[f] - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f] \right) \right| \right] - \frac{r}{m} \ .$$

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 equation 1.2.1, it follows that

$$\mathrm{Y}_j(\boldsymbol{x}) \geq \mathrm{Y}(\boldsymbol{x}) - (1 + \mathrm{b}(m)) \frac{r}{m} \ .$$

Combining the above and equation A.1.9, we obtain

$$\mathrm{Y}(\boldsymbol{x}) - (1 + \mathrm{b}(m)) \frac{r}{m} \leq \mathrm{Y}_j(\boldsymbol{x}) \leq \mathrm{Y}(\boldsymbol{x}) + \frac{r}{m} \mathrm{b}(m) \ . \tag{A.1.10}$$

We now show that

$$\sum_{j=1}^{m} \big(\mathrm{Y}(\boldsymbol{x}) - \mathrm{Y}_j(\boldsymbol{x})\big) \leq \mathrm{Y}(\boldsymbol{x}) \ . \tag{A.1.11}$$

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 equation A.1.11, 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 equation A.1.10 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 equation A.1.11 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 equation A.1.6 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

$$\mathrm{W}(\mathcal{F}) = \sup_{f \in \mathcal{F}} \mathop{\mathbb{E}}_{\boldsymbol{x}} \left[\frac{1}{m} \sum_{i=1}^{m} \big(f(\boldsymbol{x}_i) - \mathop{\mathbb{E}}_{\mathcal{D}}[f]\big)^2 \right] = \sup_{f \in \mathcal{F}} \mathop{\mathbb{E}}_{\boldsymbol{x}} \left[\frac{1}{m-1} \sum_{i=1}^{m} \big(f(\boldsymbol{x}_i) - \mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}}[f]\big)^2 \right] \ .$$

An application of Jensen's inequality gives

$$\mathrm{W}(\mathcal{F}) \leq \mathop{\mathbb{E}}_{\boldsymbol{x}}\Bigg[\underbrace{\sup_{f\in\mathcal{F}} \frac{1}{m-1}\sum_{i=1}^{m}\left(f(\boldsymbol{x}_i) - \hat{\mathbb{E}}_{\boldsymbol{x}}[f]\right)^2}_{=\frac{m}{m-1}\widehat{\mathrm{W}}_{\boldsymbol{x}}(\mathcal{F})} \Bigg] \ . \qquad \square$$

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

$$\mathrm{W}(\mathcal{F}) \leq \tfrac{m}{m-1}\widehat{\mathrm{W}}_{\boldsymbol{x}}(\mathcal{F}) + \frac{r^2 \ln\frac{1}{\delta}}{m-1} + \sqrt{\left(\frac{r^2\ln\frac{1}{\delta}}{m-1}\right)^2 + \frac{2r^2\frac{m}{m-1}\widehat{\mathrm{W}}_{\boldsymbol{x}}(\mathcal{F})\ln\frac{1}{\delta}}{m-1}} \ . \tag{1.3.2}$$

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 equation A.1.5 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.,

$$\hat{\mathrm{V}}_{\boldsymbol{x}}[f] \doteq \frac{1}{m-1}\sum_{i=1}^{m}\left(f(\boldsymbol{x}_i) - \hat{\mathbb{E}}_{\boldsymbol{x}}[f]\right)^2 \ .$$

Define

$$\mathrm{Z}(\boldsymbol{x}) \doteq \frac{m}{m-1}\widehat{\mathrm{W}}_{\boldsymbol{x}}(\mathcal{F}) = \sup_{f\in\mathcal{F}}\hat{\mathrm{V}}_{\boldsymbol{x}}[f] = \sup_{f\in\mathcal{F}}\frac{1}{m-1}\sum_{i=1}^{m}\left(f(\boldsymbol{x}_i) - \hat{\mathbb{E}}_{\boldsymbol{x}}[f]\right)^2$$

and

$$\mathrm{Z}_j(\boldsymbol{x}) \doteq \sup_{f\in\mathcal{F}}\frac{1}{m-1}\sum_{i=1,i\neq j}^{m}\left(f(\boldsymbol{x}_i) - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f]\right)^2 \ . \tag{A.1.12}$$

We first show that

$$\mathrm{Z}_j(\boldsymbol{x}) = \sup_{f\in\mathcal{F}}\left[\hat{\mathrm{V}}_{\boldsymbol{x}}[f] - \frac{1}{m}\left(f(x_j) - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f]\right)^2\right], \tag{A.1.13}$$

as this form comes in handy many times. Starting from the definition of Z_j in equation A.1.12, we add and subtract $\frac{1}{m-1}(f(x_j) - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f])^2$ 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:

$$\begin{aligned}
\mathrm{Z}_j(\boldsymbol{x}) &= \sup_{f\in\mathcal{F}}\frac{1}{m-1}\left[\left(\sum_{i=1}^{m}\left(f(\boldsymbol{x}_i) - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f]\right)^2\right) - \left(f(\boldsymbol{x}_j) - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f]\right)^2\right] \\
&= \sup_{f\in\mathcal{F}}\frac{1}{m-1}\left[\left(\sum_{i=1}^{m}\left(\left(f(\boldsymbol{x}_i) - \hat{\mathbb{E}}_{\boldsymbol{x}}[f]\right) + \left(\hat{\mathbb{E}}_{\boldsymbol{x}}[f] - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f]\right)\right)^2\right) - \left(f(\boldsymbol{x}_j) - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f]\right)^2\right]
\end{aligned}$$

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

$$\mathrm{Z}_j(\boldsymbol{x}) = \sup_{f\in\mathcal{F}}\frac{1}{m-1}\Bigg[\underbrace{\left(\sum_{i=1}^{m}\left(f(\boldsymbol{x}_i) - \hat{\mathbb{E}}_{\boldsymbol{x}}[f]\right)^2\right)}_{=(m-1)\hat{\mathrm{V}}_{\boldsymbol{x}}[f]} + m\left(\hat{\mathbb{E}}_{\boldsymbol{x}}[f] - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f]\right)^2 - \left(f(\boldsymbol{x}_j) - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f]\right)^2\Bigg] \ .$$

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

$$\mathrm{Z}_j(\boldsymbol{x}) = \sup_{f \in \mathcal{F}} \frac{1}{m-1} \left[(m-1)\hat{\mathrm{V}}_{\boldsymbol{x}}[f] + m\left(\frac{1}{m} f(\boldsymbol{x}_j) - \frac{1}{m} \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f] \right)^2 - \left(f(\boldsymbol{x}_j) - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f] \right)^2 \right] .$$

The identity in equation A.1.13 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 equation A.1.12, 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 $j = 1, \ldots, m$, it holds,

$$\mathrm{Z}_j(\boldsymbol{x}) \leq \mathrm{Z}(\boldsymbol{x}) \leq \mathrm{Z}_j(\boldsymbol{x}) + \frac{r^2}{m} \text{ for any } \boldsymbol{x} \in \mathcal{X}^m, \tag{A.1.14}$$

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 equation A.1.13, and use the subadditivity of the supremum to obtain

$$\mathrm{Z}_j(\boldsymbol{x}) \geq \left[\underbrace{\left(\sup_{f \in \mathcal{F}} \hat{\mathrm{V}}_{\boldsymbol{x}}[f] \right)}_{=\mathrm{Z}(\boldsymbol{x})} - \left(\sup_{f \in \mathcal{F}} \frac{1}{m} \left(f(\boldsymbol{x}_j) - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f] \right)^2 \right) \right] .$$

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 equation A.1.14.

We now show that, for any $\boldsymbol{x} \in \mathcal{X}^m$, it holds

$$\sum_{i=1}^{m} \left(\mathrm{Z}(\boldsymbol{x}) - \mathrm{Z}_j(\boldsymbol{x}) \right) \leq \frac{m}{m-1} \mathrm{Z}(\boldsymbol{x}),$$

as in the last requirement of definition A.1.1. Starting again from equation A.1.13 and using the subadditivity of the supremum, it holds

$$\sum_{j=1}^{m} \mathrm{Z}_j(\boldsymbol{x}) = \sum_{j=1}^{m} \sup_{f \in \mathcal{F}} \left[\hat{\mathrm{V}}_{\boldsymbol{x}}[f] - \frac{1}{m} \left(f(\boldsymbol{x}_j) - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f] \right)^2 \right] \geq \sup_{f \in \mathcal{F}} \sum_{j=1}^{m} \left[\hat{\mathrm{V}}_{\boldsymbol{x}}[f] - \frac{1}{m} \left(f(\boldsymbol{x}_j) - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f] \right)^2 \right] .$$

By simple algebra we then get

$$\sum_{j=1}^{m} \mathrm{Z}_j(\boldsymbol{x}) \geq \sup_{f \in \mathcal{F}} \left[m\hat{\mathrm{V}}_{\boldsymbol{x}}[f] - \frac{1}{m} \sum_{j=1}^{m} \left(f(\boldsymbol{x}_j) - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f] \right)^2 \right] .$$

From here, we use the fact that

$$\hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f] = \frac{1}{m-1} \left(m \hat{\mathbb{E}}_{\boldsymbol{x}}[f] - f(\boldsymbol{x}_j) \right),$$

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

$$\mathbb{P}_{\boldsymbol{x}} \left(\widehat{\mathrm{W}}_{\boldsymbol{x}}(\mathcal{F}) \leq \mathrm{W}(\mathcal{F}) - \varepsilon \right) \leq \mathbb{P}_{\boldsymbol{x}} \left(\widehat{\mathrm{W}}_{\boldsymbol{x}}(\mathcal{F}) \leq \frac{m}{m-1} \mathbb{E}_{\boldsymbol{x}}[\widehat{\mathrm{W}}_{\boldsymbol{x}}(\mathcal{F})] - \varepsilon \right) .$$

From here, we use the definition

$$\mathrm{Z}(\boldsymbol{x}) = \frac{m}{m-1} \widehat{\mathrm{W}}_{\boldsymbol{x}}(\mathcal{F})$$

and apply equation A.1.5 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

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

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. □

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*

$$\mathrm{Z}(\boldsymbol{y}) \doteq \sup_{g \in \mathcal{G}} \sum_{i=1}^{m} g(y_i) \ \textit{and}\ S^2 \doteq \mathop{\mathbb{E}}_{\boldsymbol{y}} \left[\sup_{g \in \mathcal{F}} \sum_{i=1}^{m} \mathop{\mathbb{E}}_{y_i' \sim \mathcal{Q}_i} \left[\left(\left(g(y_i) - g(y_i') \right)_+ \right)^2 \right] \right] \ .$$

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

$$\mathop{\mathbb{P}}_{\boldsymbol{y}} \left(\mathrm{Z}(\boldsymbol{y}) \leq \mathop{\mathbb{E}}_{\mathcal{Q}_{1:m}} [\mathrm{Z}] - \varepsilon \right) \leq \exp\left(-\frac{S^2}{4} \mathrm{h}\left(\frac{2\varepsilon}{S^2} \right) \right) \ . \tag{A.1.15}$$

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*

$$\hat{\mathfrak{R}}_m(\mathcal{F}, \boldsymbol{x}) \leq \hat{\mathfrak{R}}^n_m(\mathcal{F}, \boldsymbol{x}, \boldsymbol{\sigma}) + \frac{2\hat{q}_{\mathcal{F}}(\boldsymbol{x}) \ln \frac{1}{\delta}}{3nm} + \sqrt{\frac{4\widehat{\mathrm{W}}^{\mathrm{r}}_{\boldsymbol{x}}(\mathcal{F}) \ln \frac{1}{\delta}}{nm}} \ . \tag{1.3.5}$$

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

$$\mathcal{G} \doteq \{ g(\boldsymbol{\sigma}_{j,i}) \doteq \boldsymbol{\sigma}_{j,i} f_j(\boldsymbol{x}_i), \boldsymbol{f} \in (\mathcal{F}_\pm)^n \} \ .$$

The functions in $\mathcal{G}$ have domain $\mathcal{Y} = \{-1, 1\}$ and values in $[-1, 1]$. It holds

$$\mathrm{Z}(\boldsymbol{\sigma}) = \sup_{\boldsymbol{f} \in (\mathcal{F}_\pm)^n} \sum_{j=1}^{n} \sum_{i=1}^{m} \boldsymbol{\sigma}_{j,i} f_j(\boldsymbol{x}_i) = \sup_{g \in \mathcal{G}} \sum_{(j,i) \in \{1, \ldots n\} \times \{1, \ldots, m\}} g(\boldsymbol{\sigma}_{j,i}) \ . \tag{A.1.16}$$

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

$$S^2 \doteq \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{\boldsymbol{f} \in (\mathcal{F}_\pm)^n} \sum_{j=1}^{n} \sum_{i=1}^{m} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}'_{j,i}} \left[\left(\left(\boldsymbol{\sigma}_{j,i} f_j(\boldsymbol{x}_i) - \boldsymbol{\sigma}'_{j,i} f_j(\boldsymbol{x}_i) \right)_+ \right)^2 \right] \right]$$
$$= n \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{f \in \mathcal{F}_\pm} \sum_{i=1}^{m} 2 \left(\left(\boldsymbol{\sigma}_{1,i} f(\boldsymbol{x}_i) \right)_+ \right)^2 \right] \ .$$

It holds

$$S^2 \leq 2nm \widehat{\mathrm{W}}^{\mathrm{r}}_{\boldsymbol{x}}(\mathcal{F}) \ . \tag{A.1.17}$$

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 equation A.1.17 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

$$\mathop{\mathbb{P}}_{\boldsymbol{\sigma}} \left(\hat{\mathfrak{R}}_m(\mathcal{F}, \boldsymbol{x}) \geq \hat{\mathfrak{R}}^n_m(\mathcal{F}, \boldsymbol{x}, \boldsymbol{\sigma}) + \varepsilon \right) \leq \exp\left(-\frac{9}{2} nm\widehat{\mathrm{W}}^{\mathrm{r}}_{\boldsymbol{x}}(\mathcal{F}) \mathrm{h}_1\left(\frac{\varepsilon}{\widehat{\mathrm{W}}^{\mathrm{r}}_{\boldsymbol{x}}(\mathcal{F})} \right) \right) \ .$$

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.

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.

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

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

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

$$\hat{\mathfrak{R}}_m(\hat{\mathrm{C}}_{\boldsymbol{x}}(\mathcal{F}_2), \boldsymbol{x}) = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{w : \|w\|_2 \leq 1} \left| \frac{1}{m} \sum_{i=1}^{m} \boldsymbol{\sigma}_i \big(w \cdot \boldsymbol{x}_i - \hat{\mathop{\mathbb{E}}}_{\boldsymbol{x}}[w] \big) \right| \right],$$

where

$$\hat{\mathop{\mathbb{E}}}_{\boldsymbol{x}}[w] = \frac{1}{m} \sum_{i=1}^{m} (w \cdot \boldsymbol{x}_i) = w \cdot \bar{\boldsymbol{x}} \ .$$

Using linearity we then get

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

Now, for ease of notation let $v = \frac{1}{m} \sum_{i=1}^{m} \boldsymbol{\sigma}_i (\boldsymbol{x}_i - \bar{\boldsymbol{x}})$. The supremum is realized when

$$w = \frac{v}{\|v\|_2},$$

because in this case the vector w has the same direction as w and the largest possible norm $\|w\|_2 = 1$. Since the two vectors w and v are proportional to each other, Cauchy-Schwarz inequality holds with equality and we have

$$w \cdot v = \|w\|_2 \|v\|_2 = \|v\|_2 = \left\| \frac{1}{m} \sum_{i=1}^{m} \boldsymbol{\sigma}_i (\boldsymbol{x}_i - \bar{\boldsymbol{x}}) \right\|_2 .$$

We thus obtain

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

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

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

□

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 $\boldsymbol{x}' \sim \mathcal{N}(\mu, \Sigma)$, and finally obtain sample $\boldsymbol{x}$ by projecting $\boldsymbol{x}'$ to the nonnegative hyperquadrant of the radius $\sqrt{d}$ ℓ_2 sphere; i.e.,

$$\boldsymbol{x} = \mathop{\mathrm{argmin}}_{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_2 \leq \sqrt{d} \wedge \mathbf{0} \preceq \boldsymbol{x}} \|\boldsymbol{x} - \boldsymbol{x}'\|_2 .$$

Taking I_d to be the identity matrix, we sample $\boldsymbol{\mu} \sim \mathcal{N}(\mathbf{1}, I_d)$, and taking $\boldsymbol{a} \sim \mathcal{U}(0,1)^{d \times d}$, we let $\boldsymbol{\Sigma} = \frac{\boldsymbol{a}\boldsymbol{a}^\top}{d} + I_d$. 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 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 further reinforces the importance of *variance-sensitive* bounds over the (range-only) McDiarmid bounds.

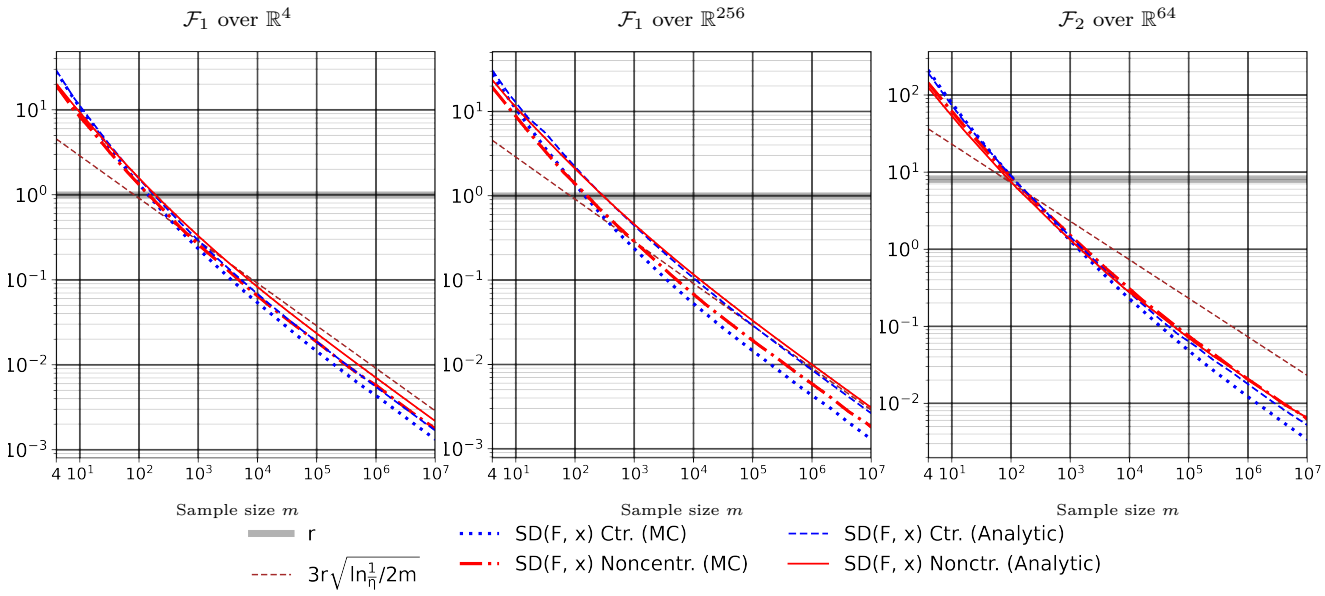

Figure A.1: Comparison of SD bounds as functions of the sample size m. See the main text for an explanation of the results.

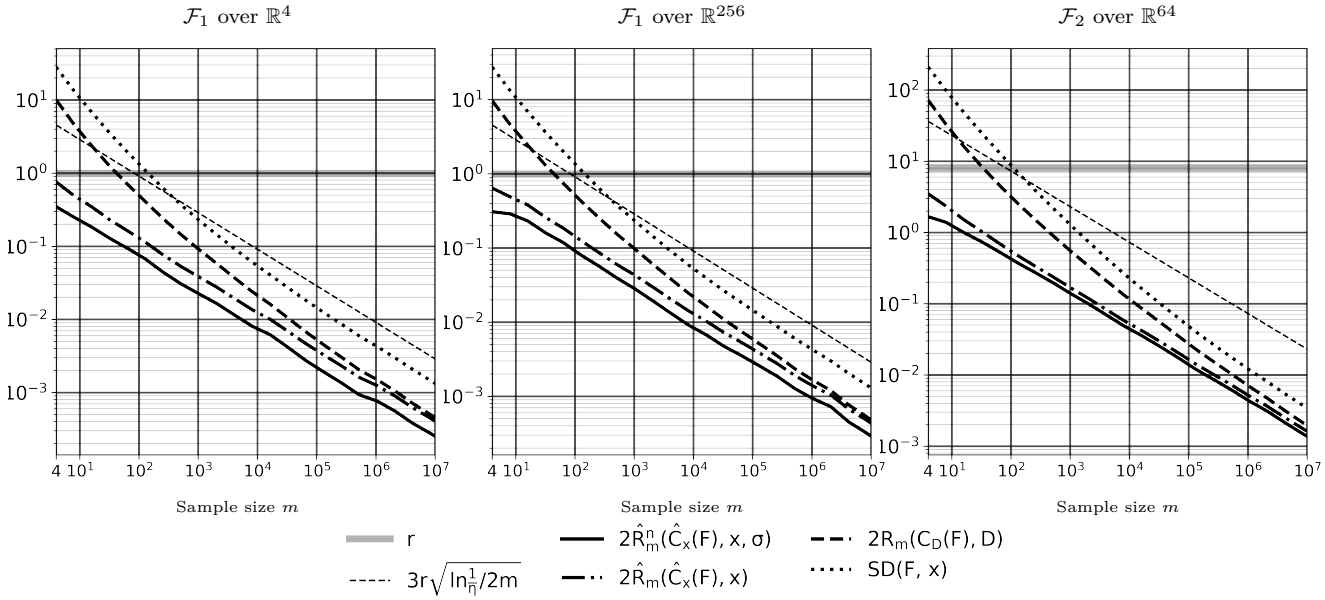

Figure A.2: Comparison of SD bounds as functions of the sample size m. See the main text for an explanation of the results.