Sharp uniform convergence bounds through empirical centralization

Cyrus Cousins
Department of Computer Science
Brown University
Providence, RI 02912
ccousins@cs.brown.edu

Matteo Riondato
Department of Computer Science
Amherst College
Amherst, MA 01002
mriondato@amherst.edu

Abstract

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

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 [23]. Formally, let $\mathcal{D}$ be a distribution on the domain $\mathcal{X}$ and $\boldsymbol{x} = \{x_1, \ldots, x_m\}$ be a collection of m independent samples from $\mathcal{D}$. The *Supremum Deviation (SD) of $\mathcal{F}$ on x* is the quantity

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

The sample-dependent *Empirical Rademacher Average (ERA) $\hat{\mathsf{R}}_m(\mathcal{F}, \boldsymbol{x})$ of $\mathcal{F}$ on $\boldsymbol{x}$* and its expectation, the *Rademacher Average (RA) $\mathsf{R}_m(\mathcal{F}, \mathcal{D})$ of $\mathcal{F}$* [3, 13], allow to derive upper and lower bounds to the SD (see (2)). Let $\boldsymbol{\sigma}$ be a collection of m independent Rademacher variables (i.e., uniform on $\{-1, 1\}$). These two quantities are defined as[1]

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

The RA controls the finite-sample *expected* SD as [28]

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

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

[1] The absolute value can be omitted. All we say can be adapted to this case.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

also Thm. 3). 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 [1, 21, 22, 24, 25], 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 (8)). In this work, we use various forms of *centralization* to develop such practical bounds. Define the *distributional centralization* $\mathsf{C}_\mathcal{D}(\mathcal{F})$ *w.r.t.* $\mathcal{D}$ as the family

$$\mathsf{C}_\mathcal{D}(\mathcal{F}) \doteq \{x \mapsto f(x) - \mathbb{E}_\mathcal{D}[f], f \in \mathcal{F}\} \ . \tag{3}$$

$\mathsf{C}_\mathcal{D}(\mathcal{F})$ 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 $g \in \mathsf{C}_\mathcal{D}(\mathcal{F})$. The Rademacher Average $\mathsf{R}_m(\mathsf{C}_\mathcal{D}(\mathcal{F}), \mathcal{D})$ of $\mathsf{C}_\mathcal{D}(\mathcal{F})$ *sharply* controls the finite-sample expected SD as [7, Lemma 11.4]

$$\tfrac{1}{2}\mathsf{R}_m(\mathsf{C}_\mathcal{D}(\mathcal{F}), \mathcal{D}) \le \mathbb{E}_{\boldsymbol{x}}[\mathsf{SD}(\mathcal{F}, \boldsymbol{x})] \le 2\mathsf{R}_m(\mathsf{C}_\mathcal{D}(\mathcal{F}), \mathcal{D}) \ . \tag{4}$$

Comparing (2) and (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* (see Sect. 2) 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 (6)), and the sample size m (see Thm. 4). We also show that the dependence on the wimpy variance that we obtain is optimal (Lemma 2 and Coro. 1). 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) and derive its concentration properties (Thm. 1) using tail bounds for *self-bounding functions* [4, 6]. In order to obtain fully-sample-dependent bounds, we introduce a Monte-Carlo estimation approach with novel tight deviation bounds (Thm. 5), and we also develop novel tight bounds for the empirical wimpy variance (Thm. 2), which we believe to be of independent interest. 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. Due to space restrictions, all our proofs are in the supplementary material.

2 Empirical centralization

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

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

This quantity is an empirical counterpart to the distributional centralization $\mathsf{C}_\mathcal{D}(\mathcal{F})$ of $\mathcal{F}$ (see (3)). The key quantity that we use to derive the sample-dependent probabilistic bounds to the SD (Sect. 3) is the ERA of the empirical centralization of $\mathcal{F}$, i.e., the quantity

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

This quantity is completely dependent on the realized $\boldsymbol{x}$, even more, in some sense, than a "standard" ERA (see (1)), because the considered family $\hat{\mathsf{C}}_{\boldsymbol{x}}(\mathcal{F})$ 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 $\hat{\mathsf{R}}_m(\hat{\mathsf{C}}_{\boldsymbol{x}}(\mathcal{F}), \boldsymbol{x})$ is *not* the RA of the distributional centralization of $\mathcal{F}$ (i.e., $\mathsf{R}_m(\mathsf{C}_\mathcal{D}(\mathcal{F}), \mathcal{D})$), but we now show that the bias decreases rapidly in m, i.e., $\mathsf{R}_m(\mathsf{C}_\mathcal{D}(\mathcal{F}), \mathcal{D}) \in \Theta(\mathbb{E}_{\boldsymbol{x}}[\hat{\mathsf{R}}_m(\hat{\mathsf{C}}_{\boldsymbol{x}}(\mathcal{F}), \boldsymbol{x})])$. For ease of notation, let

$$\mathsf{b}(m) \doteq \mathbb{E}_{\boldsymbol{\sigma}}\left[\left|\frac{1}{m}\sum_{i=1}^{m}\sigma_i\right|\right] \ \left(\text{which is } \Theta\left(\frac{1}{\sqrt{m}}\right)\right) \ . \tag{5}$$

Lemma 1. *Suppose $m \geq 4$. Then*

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

Concentration We now show that $\hat{\mathsf{R}}_m(\hat{\mathsf{C}}_{\boldsymbol{x}}(\mathcal{F}),\boldsymbol{x})$ is tightly concentrated around its expectation because it is a *self-bounding function* [4, 6] (see also Def. 2 in the supplementary material). We call the *widest range of* $\mathcal{F}$ the quantity

$$r \doteq \sup_{f\in\mathcal{F}}\left(\max_{x\in\mathcal{X}} f(x) - \min_{y\in\mathcal{X}} f(y)\right) \quad (\leq b-a) \ . \tag{6}$$

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

$$\mathbb{E}_{\boldsymbol{x}}[\hat{\mathsf{R}}_m(\hat{\mathsf{C}}_{\boldsymbol{x}}(\mathcal{F}),\boldsymbol{x})] \leq \hat{\mathsf{R}}_m(\hat{\mathsf{C}}_{\boldsymbol{x}}(\mathcal{F}),\boldsymbol{x}) + \frac{2r\chi\ln\frac{1}{\delta}}{3m} + \sqrt{\left(\frac{r\chi\ln\frac{1}{\delta}}{\sqrt{3}m}\right)^2 + \frac{2r\chi(\hat{\mathsf{R}}_m(\hat{\mathsf{C}}_{\boldsymbol{x}}(\mathcal{F}),\boldsymbol{x}) + r\mathsf{b}(m))\ln\frac{1}{\delta}}{m}} \ . \tag{7}$$

The ERA of $\mathcal{F}$ is a self-bounding function [5, Sect. 5.1], but proving this fact for the ERA of the empirical centralization $\hat{\mathsf{C}}_{\boldsymbol{x}}(\mathcal{F})$ of $\mathcal{F}$ is more challenging (see proof in the supplementary material), because the empirical centralization $\hat{\mathsf{C}}_{\boldsymbol{x}}(\mathcal{F})$ itself depends on the sample $\boldsymbol{x}$. This result, together with Lemma 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.

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* $\mathsf{W}^{\mathsf{r}}(\mathcal{F})$ *of* $\mathcal{F}$ and the (centralized) *wimpy variance* $\mathsf{W}(\mathcal{F})$ *of* $\mathcal{F}$ are key quantities in the study of probabilistic tail bounds to the SD [7, Ch. 11]. They are defined as

$$\mathsf{W}^{\mathsf{r}}(\mathcal{F}) \doteq \sup_{f\in\mathcal{F}} \mathbb{E}_{x\sim\mathcal{D}}\left[(f(x))^2\right], \text{ and } \mathsf{W}(\mathcal{F}) \doteq \sup_{f\in\mathcal{F}} \mathbb{E}_{x\sim\mathcal{D}}\left[(f(x)-\mathbb{E}_{\mathcal{D}}[f])^2\right] \ . \tag{8}$$

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

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

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

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

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 [18] prove that the sample variance (i.e., when $\mathcal{F}$ is a singleton) is a *weakly self-bounding function* [20]. Our result holds for general $\mathcal{F}$, and is stronger, as we show that the wimpy variance is a *(strongly) self-bounding function* [4, 6] (see also Def. 2 in the supplementary material).

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

$$\mathsf{W}(\mathcal{F}) \leq \tfrac{m}{m-1}\widehat{\mathsf{W}}_{\boldsymbol{x}}(\mathcal{F}) + \frac{r^2\ln\frac{1}{\delta}}{m-1} + \sqrt{\left(\frac{r^2\ln\frac{1}{\delta}}{m-1}\right)^2 + \frac{2r^2\frac{m}{m-1}\widehat{\mathsf{W}}_{\boldsymbol{x}}(\mathcal{F})\ln\frac{1}{\delta}}{m-1}} \ . \tag{9}$$

Bounds to the SD Bousquet [8, Thm. 2.3 (presented here for clarity in a slightly weaker form)] uses the wimpy variance to derive concentration bounds for the SD.

Theorem 3 (8, Thm. 2.3)**.** *Let $\delta \in (0,1)$. With probability $\geq 1-\delta$ over the choice of $\boldsymbol{x}$,*

$$\mathsf{SD}(\mathcal{F},\boldsymbol{x}) \leq \mathbb{E}_{\boldsymbol{x}}\left[\mathsf{SD}(\mathcal{F},\boldsymbol{x})\right] + \frac{2r\ln\frac{1}{\delta}}{3m} + \sqrt{\frac{2\left(\mathsf{W}(\mathcal{F}) + 4r\ \mathbb{E}_{\boldsymbol{x}}\left[\mathsf{SD}(\mathcal{F},\boldsymbol{x})\right]\right)\ln\frac{1}{\delta}}{m}}\ . \tag{10}$$

By plugging the r.h.s. of the symmetrization inequalities (2) and (4) in the r.h.s. of (10), one can obtain bounds that depend on the RA of $\mathcal{F}$ or on the RA of the *distributional* centralization $\mathsf{C}_{\mathcal{D}}(\mathcal{F})$. 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 [19]'s inequality or a tail bound for self-bounding functions [6]) on the deviation of the ERA from the RA. The following result states our sample-dependent bound to the SD using the *empirical centralization* $\hat{\mathsf{C}}_{\boldsymbol{x}}(\mathcal{F})$ and tail bounds to the wimpy variance, obtained by combining Lemma 1 and Thms. 1 to 3.

Theorem 4. *Assume $m \geq 4$, and let $\eta \in (0,1)$. Take ν to be the r.h.s. of* (9) *computed with $\delta = \eta/3$, so $\Pr(\mathsf{W}(\mathcal{F}) > \nu) \leq \eta/3$, and take λ to be the r.h.s. of* (7) *computed with $\delta = \eta/3$, so $\Pr(\mathbb{E}_{\boldsymbol{x}}[\hat{\mathsf{R}}_m(\hat{\mathsf{C}}_{\boldsymbol{x}}(\mathcal{F}),\boldsymbol{x})] > \lambda) \leq \eta/3$. With probability $\geq 1-\eta$ over the choice of $\boldsymbol{x}$, it holds that*

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

The r.h.s. is

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

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

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

We shall see that the ${}^{r^2}/_{\sqrt{m}}$ and ${}^{q\sqrt{\mathsf{W}^{\mathsf{r}}(\mathcal{F})}}/_{\sqrt{m}}$ terms are incomparable, though both appear only in transient $\mathrm{O}(m^{-3/4})$ 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., [1, 21, 22, 24, 25]), 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 (10) cannot be improved: any bound to the SD of $\mathcal{F}$ must be $\Omega\sqrt{\mathsf{W}(\mathcal{F})\ln\frac{1}{\delta}/m}$, as can be shown using minimax lower bounds and median-of-means bounds [10, 15]. The question is thus whether the *complexity terms*, i.e., $\hat{\mathsf{R}}_m(\mathcal{F},\boldsymbol{x})$ and $\hat{\mathsf{R}}_m(\hat{\mathsf{C}}_{\boldsymbol{x}}(\mathcal{F}),\boldsymbol{x})$ can match this lower bound. Lemma 2 answers this question in the *negative* for $\hat{\mathsf{R}}_m(\mathcal{F})$, and in the *positive* for $\hat{\mathsf{R}}_m(\hat{\mathsf{C}}_{\boldsymbol{x}}(\mathcal{F}),\boldsymbol{x})$: the ERA of $\mathcal{F}$ is controlled (in part) by the empirical *raw* wimpy variance, whereas the ERA of $\hat{\mathsf{C}}_{\boldsymbol{x}}(\mathcal{F})$ 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 2. *For any $\boldsymbol{x} \in \mathcal{X}^m$, it holds*

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

Furthermore, it holds

$$\lim_{m\to\infty}\sqrt{m}\mathsf{R}_m(\mathcal{F},\mathcal{D}) \geq \sqrt{\frac{2}{\pi}\mathsf{W}^{\mathsf{r}}(\mathcal{F})}\ \textit{and}\ \lim_{m\to\infty}\sqrt{m}\mathsf{R}_m(\mathsf{C}_{\mathcal{D}}(\mathcal{F}),\mathcal{D}) \geq \sqrt{\frac{2}{\pi}\mathsf{W}(\mathcal{F})}\ .$$

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 $\widehat{\mathsf{W}}^{\mathrm{r}}_{\boldsymbol{x}}(\mathcal{F}) \geq c^2/4$, and from the above lemma, $\hat{\mathsf{R}}_m(\mathcal{F}, \boldsymbol{x}) \geq c/\sqrt{8m}$, but $\hat{\mathsf{R}}_m(\hat{\mathsf{C}}_{\boldsymbol{x}}(\mathcal{F}), \boldsymbol{x})$ does not suffer from this issue.

The significance of Lemma 2 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 $\hat{\mathsf{R}}_m(\hat{\mathsf{C}}_{\boldsymbol{x}}(\mathcal{F}), \boldsymbol{x})$ is an ERA, thus it can be upper-bounded using Massart's finite-class lemma [16, 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. *Assume that $\mathcal{F}$ is finite. Let $\mathcal{F}_{\pm} \doteq \mathcal{F} \cup \{-f, f \in \mathcal{F}\}$. It holds*

$$\hat{\mathsf{R}}_m(\hat{\mathsf{C}}_{\boldsymbol{x}}(\mathcal{F}), \boldsymbol{x}) \leq \sqrt{\frac{2\widehat{\mathsf{W}}_{\boldsymbol{x}}(\mathcal{F}) \ln|\hat{\mathsf{C}}_{\boldsymbol{x}}(\mathcal{F}_{\pm})|}{m}} \ . \tag{11}$$

The use of $\mathcal{F}_{\pm}$ is needed[2] to handle the absolute value in our definition of the ERA (see (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 shows that empirical centralization enables *optimal* dependence on the *centralized* wimpy variance, which *cannot be obtained without empirical centralization*, as shown in Lemma 2.

Monte-Carlo estimation The quantity $\hat{\mathsf{R}}_m(\hat{\mathsf{C}}_{\boldsymbol{x}}(\mathcal{F}), \boldsymbol{x})$ 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 Thm. 4. 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 [3].

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

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

It clearly holds $\mathbb{E}_{\boldsymbol{\sigma}}[\hat{\mathsf{R}}^n_m(\mathcal{F}, \boldsymbol{x}, \boldsymbol{\sigma})] = \hat{\mathsf{R}}_m(\mathcal{F}, \boldsymbol{x})$. Bartlett and Mendelson [3, Thm. 11] show that the MC-ERA with $n = 1$ is concentrated about the ERA as

$$\Pr_{\boldsymbol{\sigma}}\left(\left|\hat{\mathsf{R}}_m(\mathcal{F}, \boldsymbol{x}) - \hat{\mathsf{R}}^1_m(\mathcal{F}, \boldsymbol{x}, \boldsymbol{\sigma})\right| \geq \varepsilon\right) \leq 2\exp\left(\frac{-2m\varepsilon^2}{q^2}\right) \ .$$

The r.h.s. can be used in Thm. 4 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 (Thm. 5) *for general* n, where the strong dependence on q^2 of the above bound is replaced by a much weaker dependence, primarily on $\mathsf{W}(\mathcal{F})$. This change is similar to how Thm. 3 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 [26] to *upper bound* the *expectation* of suprema of empirical processes. This result is, to our knowledge, novel, and is worst-case asymptotically equivalent to the McDiarmid bounds, and improves over it when the wimpy variance is small. The bound uses the *empirical maximum* $\hat{q}_{\mathcal{F}}(\boldsymbol{x})$ *of* $\mathcal{F}$ *on* $\boldsymbol{x}$, defined as

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

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

Theorem 5. *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*

$$\hat{\mathsf{R}}_m(\mathcal{F}, \boldsymbol{x}) \leq \hat{\mathsf{R}}^n_m(\mathcal{F}, \boldsymbol{x}, \boldsymbol{\sigma}) + \frac{2\hat{q}_{\mathcal{F}}(\boldsymbol{x}) \ln \frac{1}{\delta}}{3nm} + \sqrt{\frac{4\widehat{\mathsf{W}}^{\mathsf{r}}_{\boldsymbol{x}}(\mathcal{F}) \ln \frac{1}{\delta}}{nm}} \ . \tag{12}$$

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 $\hat{q}_{\hat{\mathsf{C}}_{\boldsymbol{x}}(\mathcal{F})}(\boldsymbol{x})$, which can be a large improvement (and $\hat{q}_{\hat{\mathsf{C}}_{\boldsymbol{x}}(\mathcal{F})}(\boldsymbol{x}) < 2\hat{q}_{\mathcal{F}}(\boldsymbol{x})$ at most).

Corollary 2. *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*

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

Although $n = 1$ Monte-Carlo trials are sufficient to match the convergence rate of Thm. 3, 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 Thm. 3.

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.

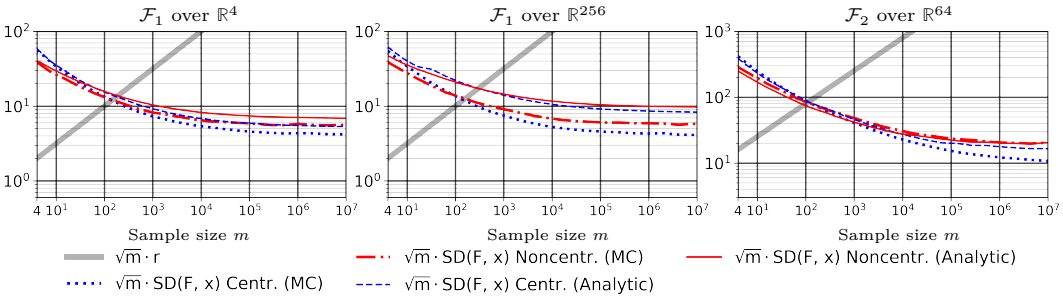

Figure 1: Comparison of SD bounds as functions of the sample size m. See the main text for an explanation of the results.

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.,

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

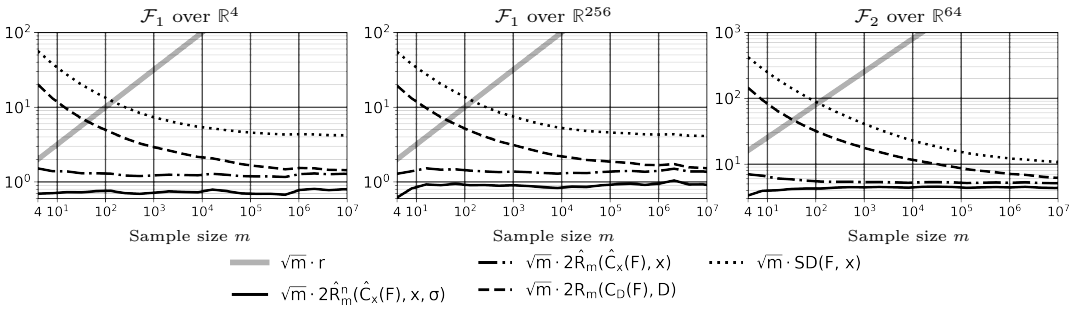

Figure 2: Upper bounds to complexity measures and SD as functions of the sample size m. See the main text for details.

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., [3] 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 [9], and to learn *equilibria* in simulation-based games [1, 2].

Analytical bounds to the ERA of $\mathcal{F}_p$ on $\boldsymbol{x}$ and Monte-Carlo estimates of it (see Def. 1) are relatively straightforward [27, Lemmas 26.10, 26.11] (see Lemma 5 in the supplementary material). The following lemma extends these results to the empirical centralization.

Lemma 3. *Let* $\bar{x} \doteq \frac{1}{m}\sum_{i=1}^{m} x_i \in \mathbb{R}^d$. *For the* ℓ_1 *norm, it holds*

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

while for the ℓ_2 *norm, it holds*

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

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 3 and Lemma 5 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) 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 $\Theta(C/\sqrt{m})$ behaviors as *straight horizontal lines*, and $o(C/\sqrt{m})$ behaviors as (transient) downward slopes. For completeness, we show plots without the scaling by $\sqrt{m}$ in the supplementary material.

Results Figure 1 compares four bounds to the SD: using the Monte-Carlo estimate $\hat{\mathsf{R}}^n_m(\hat{\mathsf{C}}_{\boldsymbol{x}}(\mathcal{F}_p), \boldsymbol{x}, \boldsymbol{\sigma})$ for the ERA $\hat{\mathsf{R}}_m(\hat{\mathsf{C}}_{\boldsymbol{x}}(\mathcal{F}_p), \boldsymbol{x})$ of the empirical centralization of $\mathcal{F}_p$ on $\boldsymbol{x}$, using the Monte-Carlo estimate $\hat{\mathsf{R}}^n_m(\mathcal{F}_p, \boldsymbol{x}, \boldsymbol{\sigma})$ for the ERA $\hat{\mathsf{R}}_m(\mathcal{F}_p, \boldsymbol{x})$ of the non-centralized $\mathcal{F}_p$, using analytical bounds to $\hat{\mathsf{R}}_m(\hat{\mathsf{C}}_{\boldsymbol{x}}(\mathcal{F}_p), \boldsymbol{x})$ from Lemma 3, and using analytical bounds to $\hat{\mathsf{R}}_m(\mathcal{F}_p, \boldsymbol{x})$ from Lemma 5 (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 (see ξ in Thm. 4) and the (fast-decaying) $\Theta(r/m^{3/4})$ term of Thm. 1. 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 $\Theta(C/\sqrt{m})$ 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 Fig. 3 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 Fig. 2, we drill down on the SD bounds using the Monte-Carlo estimate $\hat{\mathsf{R}}^n_m(\hat{\mathsf{C}}_{\boldsymbol{x}}(\mathcal{F}_p), \boldsymbol{x}, \boldsymbol{\sigma})$ for the ERA $\hat{\mathsf{R}}_m(\hat{\mathsf{C}}_{\boldsymbol{x}}(\mathcal{F}_p), \boldsymbol{x})$ 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 $\hat{\mathsf{R}}_m(\hat{\mathsf{C}}_{\boldsymbol{x}}(\mathcal{F}_p), \boldsymbol{x})$ (obtained by applying Thm. 5 to the MC-ERA), the RA $\mathsf{R}_m(\mathcal{F}_p, \mathcal{D})$ (obtained by applying Thm. 1 and Lemma 1 to the bound on the ERA),[3] and SD (obtained by applying the r.h.s. of (4) and Thm. 3 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 $\Theta(C/\sqrt{m})$, where C, which in the plots in Fig. 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 $\Theta\sqrt{\mathsf{W}(\mathcal{F})/m}$ (i.e., the MC-ERA $\to$ ERA and RA $\to$ 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 3).

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.

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 [11, 14].

[3] We do not plot a line for a bound to $\mathbb{E}_{\boldsymbol{x}}[\hat{\mathsf{R}}_m(\hat{\mathsf{C}}_{\boldsymbol{x}}(\mathcal{F}\boldsymbol{x}), \boldsymbol{x})]$ using only Thm. 1 because the fully-multiplicative correction from Lemma 1 is negligible and rapidly decaying.

Statement of broader impact

The goal of our work is to make it possible to get the best possible bounds to the SD as possible from the available data. By reducing the amount of data needed to achieve a certain bound to the SD, we essentially increase the value of each data point, or on the flip side, make each "unit of bound" cheaper. We make essentially no assumption on the family of functions we consider, thus our results are very broadly applicable. As concepts and results from uniform convergence are being used in fields very different than learning (e.g., graph analysis, statistical hypothesis testing, and more, see the Introduction for some references), we believe that enabling machine learning practitioners to *better understand their models*, and scientists in other fields to *make better use of their data* is a positive effort. Certainly, we cannot predict possible misuse of our results, either voluntary or involuntary, in the same way that theoretical results of general applicability are often misused or misapplied (as an example, consider secure cryptographic ciphers that are implemented in the wrong way or used in a cryptographic system is not end-to-end secure).

Acknowledgments and Disclosure of Funding

This material is based upon work supported by the National Science Foundation under grants 2006765 and RI-1813444, as well as DARPA/AFRL grant FA8750.

References

[1] E. Areyan Viqueira, A. Greenwald, C. Cousins, and E. 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.

[2] E. Areyan Viqueira, C. Cousins, and A. 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.

[3] P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

[4] S. Boucheron, G. Lugosi, and P. Massart. A sharp concentration inequality with applications. *Random Structures & Algorithms*, 16(3):277–292, 2000.

[5] S. Boucheron, G. Lugosi, and P. Massart. Concentration inequalities using the entropy method. *The Annals of Probability*, 31(3):1583–1614, 2003.

[6] S. Boucheron, G. Lugosi, and P. Massart. On concentration of self-bounding functions. *Electronic Journal of Probability*, 14:1884–1899, 2009.

[7] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[8] O. Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematique*, 334(6):495–500, 2002.

[9] J. Bradley and C. Guestrin. Sample complexity of composite likelihood. In *Artificial Intelligence and Statistics*, pages 136–160, 2012.

[10] L. Devroye, M. Lerasle, G. Lugosi, and R. I. Oliveira. Sub-Gaussian mean estimators. *The Annals of Statistics*, 44(6):2695–2725, 2016.

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

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

[13] V. Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902–1914, 2001.

[14] V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006.

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

[16] P. Massart. Some applications of concentration inequalities to statistics. In *Annales-Faculte des Sciences Toulouse Mathematiques*, volume 9, pages 245–303. Université Paul Sabatier, 2000.

[17] A. Maurer. Concentration inequalities for functions of independent variables. *Random Structures & Algorithms*, 29(2):121–138, 2006.

[18] A. Maurer and M. Pontil. Empirical Bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

[19] C. McDiarmid. On the method of bounded differences. *Surveys in combinatorics*, 141 (1):148–188, 1989.

[20] C. McDiarmid and B. Reed. Concentration for self-bounding functions and an inequality of Talagrand. *Random Structures & Algorithms*, 29(4):549–557, 2006.

[21] L. Pellegrina, M. Riondato, and F. 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.

[22] L. Pellegrina, C. Cousins, F. Vandin, and M. 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.

[23] D. Pollard. *Convergence of Stochastic Processes*. Springer, New York, 1984.

[24] M. Riondato and E. 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.

[25] M. Riondato and E. 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.

[26] P.-M. 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.

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

[28] A. Van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes*. Springer, New York, 1996.

Supplementary Material
for
Sharp uniform convergence bounds
through empirical centralization

A Proofs

Lemma 1. *Suppose $m \geq 4$. Then*

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

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

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

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

$$\mathsf{R}_m(\mathsf{C}_{\mathcal{D}}(\mathcal{F}), \mathcal{D}) \leq \mathbb{E}_{\boldsymbol{\sigma}, \boldsymbol{x}}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{m} \sum_{i=1}^{m} \sigma_i (f(x_i) - \hat{\mathbb{E}}_{\boldsymbol{x}}[f]) \right| \right] + \mathbb{E}_{\boldsymbol{\sigma}, \boldsymbol{x}}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{m} \sum_{i=1}^{m} \sigma_i (\hat{\mathbb{E}}_{\boldsymbol{x}}[f] - \mathbb{E}_{\mathcal{D}}[f]) \right| \right] \ .$$

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

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

Each function in $\mathcal{K}_{\boldsymbol{x}}$ is *constant*. Thus, we can write

$$\mathsf{R}_m(\mathsf{C}_{\mathcal{D}}(\mathcal{F}), \mathcal{D}) \leq \mathbb{E}_{\boldsymbol{x}}\Big[\hat{\mathsf{R}}_m(\hat{\mathsf{C}}_{\boldsymbol{x}}(\mathcal{F}), \boldsymbol{x})\Big] + \mathbb{E}_{\boldsymbol{x}}\Big[\hat{\mathsf{R}}_m(\mathcal{K}_{\boldsymbol{x}}, \boldsymbol{x})\Big] \ . \tag{13}$$

Using (5) and the linearity of expectation we have that, for each $\boldsymbol{x} \in \mathcal{X}^m$, it holds

$$\hat{\mathsf{R}}_m(\mathcal{K}_{\boldsymbol{x}}, \boldsymbol{x}) = \sup_{f \in \mathcal{F}} |\hat{\mathbb{E}}_{\boldsymbol{x}}[f] - \mathbb{E}_{\mathcal{D}}[f]| \mathsf{b}(m) = \mathsf{SD}(\mathcal{F}, \boldsymbol{x}) \mathsf{b}(m) = \mathsf{SD}(\mathsf{C}_{\mathcal{D}}(\mathcal{F}), \boldsymbol{x}) \mathsf{b}(m), \tag{14}$$

where in the last step we use the fact that the SD is invariant to shifting of functions. Continuing from (13) and using (14) and the rightmost inequality of (4), we obtain

$$\mathsf{R}_m(\mathsf{C}_{\mathcal{D}}(\mathcal{F}), \mathcal{D}) \leq \mathbb{E}_{\boldsymbol{x}}\Big[\hat{\mathsf{R}}_m(\hat{\mathsf{C}}_{\boldsymbol{x}}(\mathcal{F}), \boldsymbol{x})\Big] + 2\mathsf{R}_m(\mathsf{C}_{\mathcal{D}}(\mathcal{F}), \mathcal{D}) \mathsf{b}(m) \ .$$

The hypothesis $m \geq 4$ implies $1 - 2\mathsf{b}(m) > 0$ (see (5)), so we can rewrite the above as

$$\mathsf{R}_m(\mathsf{C}_{\mathcal{D}}(\mathcal{F}), \mathcal{D}) \leq \frac{1}{1 - 2\mathsf{b}(m)} \mathbb{E}_{\boldsymbol{x}}\Big[\hat{\mathsf{R}}_m(\hat{\mathsf{C}}_{\boldsymbol{x}}(\mathcal{F}), \boldsymbol{x})\Big],$$

which completes the proof of the upper bound.

We next show the lower bound. Starting from the definition of $\hat{\mathsf{R}}_m(\hat{\mathsf{C}}_{\boldsymbol{x}}(\mathcal{F}), \boldsymbol{x})$ and subtracting and adding $\mathbb{E}_{\mathcal{D}}[f]$, it holds

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

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

$$\begin{aligned} \mathbb{E}_{\boldsymbol{x}}[\hat{\mathsf{R}}_m(\hat{\mathsf{C}}_{\boldsymbol{x}}(\mathcal{F}), \boldsymbol{x})] \leq & \mathbb{E}_{\boldsymbol{\sigma}, \boldsymbol{x}}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{m} \sum_{i=1}^{m} \sigma_i (f(x_i) - \mathbb{E}_{\mathcal{D}}[f]) \right| \right] \\ & + \mathbb{E}_{\boldsymbol{\sigma}, \boldsymbol{x}}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{m} \sum_{i=1}^{m} \sigma_i (\mathbb{E}_{\mathcal{D}}[f] - \hat{\mathbb{E}}_{\boldsymbol{x}}[f]) \right| \right] \ . \end{aligned} \tag{15}$$

The first term on the r.h.s. is the RA of the *distributional* centralization of $\mathcal{F}$, i.e., it is $\mathsf{R}_m(\mathsf{C}_{\mathcal{D}}(\mathcal{F}), \mathcal{D})$. The second term is the expectation w.r.t. $\boldsymbol{x}$ of the ERA on $\boldsymbol{x}$ of the family

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

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

$$\hat{\mathsf{R}}_m(\mathcal{Z}_{\boldsymbol{x}}, \boldsymbol{x}) = \mathsf{SD}(\mathsf{C}_{\mathcal{D}}(\mathcal{F}), \boldsymbol{x})\mathsf{b}(m) \ . \tag{16}$$

Continuing from (15) and using (16) and the rightmost inequality of (4), we obtain

$$\mathbb{E}_{\boldsymbol{x}}[\hat{\mathsf{R}}_m(\hat{\mathsf{C}}_{\boldsymbol{x}}(\mathcal{F}), \boldsymbol{x})] \leq \mathsf{R}_m(\mathsf{C}_{\mathcal{D}}(\mathcal{F}), \mathcal{D}) + 2\mathsf{R}_m(\mathsf{C}_{\mathcal{D}}(\mathcal{F}), \mathcal{D})\mathsf{b}(m) \leq (1 + 2\mathsf{b}(m))\mathsf{R}_m(\mathsf{C}_{\mathcal{D}}(\mathcal{F}), \mathcal{D}),$$

and our proof is complete. □

Definition 2. A function $\mathsf{Z} \in \mathcal{X}^m \to \mathbb{R}$ 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 $\mathsf{Z}_j \in \mathcal{X}^m \to \mathbb{R}$ such that, for any $\boldsymbol{x} \in \mathcal{X}^m$ it holds that

1. $\mathsf{Z}_j(\boldsymbol{x})$ does not depend on the j-th component x_j of $\boldsymbol{x}$; and
2. it holds $\mathsf{Z}_j(\boldsymbol{x}) \leq \mathsf{Z}(\boldsymbol{x}) \leq \mathsf{Z}_j(\boldsymbol{x}) + \gamma$;

Additionally, the functions Z_j, $j = 1, \ldots, m$, must be such that, for any $\boldsymbol{x} \in \mathcal{X}^m$, it holds $\sum_{j=1}^{m} \left(\mathsf{Z}(\boldsymbol{x}) - \mathsf{Z}_j(\boldsymbol{x}) \right) \leq \alpha \mathsf{Z}(\boldsymbol{x}) + \beta$.

Theorem 6. *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*

$$\mathbb{E}_{\boldsymbol{x}}\left[\mathsf{Z}(\boldsymbol{x})\right] \leq \mathsf{Z}(\boldsymbol{x}) + \alpha\gamma \ln \frac{1}{\delta} + \sqrt{\left(\alpha\gamma \ln \frac{1}{\delta}\right)^2 + 2\gamma(\alpha \mathsf{Z}(\boldsymbol{x}) + \beta) \ln \frac{1}{\delta}} \ . \tag{17}$$

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

$$\mathbb{E}_{\boldsymbol{x}}\left[\mathsf{Z}(\boldsymbol{x})\right] \leq \mathsf{Z}(\boldsymbol{x}) + \frac{2}{3}\gamma \ln \frac{1}{\delta} + \sqrt{\left(\frac{1}{\sqrt{3}}\gamma \ln \frac{1}{\delta}\right)^2 + 2\gamma(\mathsf{Z}(\boldsymbol{x}) + \beta) \ln \frac{1}{\delta}} \ . \tag{18}$$

Proof. In both cases, we will assume WLOG $\gamma = 1$. The results then hold by linearity, noting that if $\mathsf{Z}(\cdot)$ is α-β self-bounding, with scale γ, then $\frac{1}{\gamma}\mathsf{Z}(\cdot)$ 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 eq. (17). Assume scale $\gamma = 1$. It is known that for $\gamma = 1$, we have for all $\alpha \geq \frac{1}{3}$, as described in [6, Thm. 1], which improves the earlier bounds of [17]

$$\Pr\left(\mathsf{Z}(\boldsymbol{x}) \leq \mathbb{E}_{\boldsymbol{x}}[\mathsf{Z}(\boldsymbol{x})] - \varepsilon\right) \leq \exp\left(\frac{-\varepsilon^2}{2(\alpha \mathbb{E}_{\boldsymbol{x}}[\mathsf{Z}(\boldsymbol{x})] + \beta)}\right) \ . \tag{19}$$

Now, taking δ equal to the RHS of (19), and solving for ε, this implies that with probability at least $1 - \delta$, we have

$$\mathsf{Z}(\boldsymbol{x}) + \frac{\beta}{\alpha} \geq \mathbb{E}_{\boldsymbol{x}}\left[\mathsf{Z}(\boldsymbol{x})\right] + \frac{\beta}{\alpha} - \sqrt{2(\alpha \mathbb{E}_{\boldsymbol{x}}\left[\mathsf{Z}(\boldsymbol{x})\right] + \beta) \ln \frac{1}{\delta}} \ .$$

Note that this is a quadratic inequality in $\sqrt{\mathbb{E}_{\boldsymbol{x}}\left[\mathsf{Z}(\boldsymbol{x})\right] + \frac{\beta}{\alpha}}$, solving for which (via the quadratic formula) yields nondegenerate solution

$$\mathbb{E}_{\boldsymbol{x}}\left[\mathsf{Z}(\boldsymbol{x})\right] \leq \mathsf{Z}(\boldsymbol{x}) + \alpha \ln \frac{1}{\delta} + \sqrt{\left(\alpha \ln \frac{1}{\delta}\right)^2 + 2\alpha(\mathbb{E}_{\boldsymbol{x}}\left[\mathsf{Z}(\boldsymbol{x})\right] + \beta) \ln \frac{1}{\delta}} \ .$$

Finally, in the general case, with γ-scaling, we have

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

We now show eq. (18) (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. [7, Ch. 2.1], section 2.1) of the stronger *sub-Poisson* 1-β self-bounding function inequality [4, Thm. 1].

In particular, here we have that with probability at least $1-\delta$,

$$\mathsf{Z}(\boldsymbol{x}) \geq \mathbb{E}_{\boldsymbol{x}}\left[\mathsf{Z}(\boldsymbol{x})\right] + \frac{1}{3}\ln\frac{1}{\delta} - \sqrt{2(\mathbb{E}_{\boldsymbol{x}}\left[\mathsf{Z}(\boldsymbol{x})\right] + \beta)\ln\frac{1}{\delta}} \ ,$$

which by the quadratic formula, yields

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

The general result then follows via γ-scaling.

□

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

$$\mathbb{E}_{\boldsymbol{x}}[\hat{\mathsf{R}}_m(\hat{\mathsf{C}}_{\boldsymbol{x}}(\mathcal{F}),\boldsymbol{x})] \leq \hat{\mathsf{R}}_m(\hat{\mathsf{C}}_{\boldsymbol{x}}(\mathcal{F}),\boldsymbol{x}) + \frac{2r\chi\ln\frac{1}{\delta}}{3m} + \sqrt{\left(\frac{r\chi\ln\frac{1}{\delta}}{\sqrt{3}m}\right)^2 + \frac{2r\chi(\hat{\mathsf{R}}_m(\hat{\mathsf{C}}_{\boldsymbol{x}}(\mathcal{F}),\boldsymbol{x}) + r\mathsf{b}(m))\ln\frac{1}{\delta}}{m}} \ . \quad (7)$$

Proof. This proof proceeds by showing that $\hat{\mathsf{R}}_m(\hat{\mathsf{C}}_{\boldsymbol{x}}(\mathcal{F}),\boldsymbol{x})$ is a $(1, r\mathsf{b}(m))$-self-bounding function with scale $r\chi/m$, then applying (18) from Thm. 6. 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

$$\mathsf{Y}(\boldsymbol{x}) \doteq \hat{\mathsf{R}}_m(\hat{\mathsf{C}}_{\boldsymbol{x}}(\mathcal{F}),\boldsymbol{x}),$$

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

$$\mathsf{Y}_j(\boldsymbol{x}) \doteq \frac{m-1}{m}\hat{\mathsf{R}}_{m-1}\left(\hat{\mathsf{C}}_{\boldsymbol{x}_{\setminus j}}(\mathcal{F}),\boldsymbol{x}_{\setminus j}\right) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\left|\sum_{i=1,i\neq j}^{m}\sigma_i\left(f(x_i) - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f]\right)\right|\right] \ .$$

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 $\mathsf{Y}_j(\boldsymbol{x}) \leq \mathsf{Y}(\boldsymbol{x}) + r/m\mathsf{b}(m)$. Starting from the definition of $\mathsf{Y}_j(\boldsymbol{x})$ and adding and subtracting $(f(x_j) - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f])/2m$ to the argument of the supremum, it holds

$$\mathsf{Y}_j(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{\sigma}_{\setminus j}}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\left|\left(\sum_{\substack{i=1\\i\neq j}}^{m}\sigma_i\left(f(x_i) - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f]\right)\right) + \frac{1}{2}(f(x_j) - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f]) - \frac{1}{2}(f(x_j) - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f])\right|\right] \ .$$

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

$$\mathsf{Y}_j(\boldsymbol{x}) \leq \mathbb{E}_{\boldsymbol{\sigma}_{\setminus j}}\left[\begin{array}{ll} & \frac{1}{2}\left(\sup_{f\in\mathcal{F}}\frac{1}{m}\left|\sum_{\substack{i=1\\i\neq j}}^{m}\sigma_i\Big(f(x_i) - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f]\Big) + \Big(f(x_j) - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f]\Big)\right|\right) \\ + & \frac{1}{2}\left(\sup_{f\in\mathcal{F}}\frac{1}{m}\left|\sum_{\substack{i=1\\i\neq j}}^{m}\sigma_i\Big(f(x_i) - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f]\Big) - \Big(f(x_j) - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f]\Big)\right|\right) \end{array}\right] \ .$$

The two-term sum forming the argument of the outermost expectation is the expectation *w.r.t. only* σ_j (i.e., *conditioned* on $\boldsymbol{\sigma}_{\setminus j}$) of the quantity

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

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

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

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

$$\mathsf{Y}_j(\boldsymbol{x}) \leq \underbrace{\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\left|\sum_{i=1}^{m} \sigma_i \left(f(x_i) - \hat{\mathbb{E}}_{\boldsymbol{x}}[f]\right)\right|\right]}_{=\mathsf{Y}(\boldsymbol{x})} + \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\left|\sum_{i=1}^{m} \sigma_i \left(\hat{\mathbb{E}}_{\boldsymbol{x}}[f] - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f]\right)\right|\right] \ . \tag{20}$$

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

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

Each function in $\mathcal{W}_{\boldsymbol{x}}$ is *constant*. Using (5) and the linearity of expectation, like we did in the proof of Lemma 1 for the family $\mathcal{K}_{\boldsymbol{x}}$ (see (14)), it holds

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

Thus, continuing from (20) by incorporating the above fact, it holds

$$\mathsf{Y}_j(\boldsymbol{x}) \leq \mathsf{Y}(\boldsymbol{x}) + \frac{r}{m}\mathsf{b}(m) \ . \tag{21}$$

We now show that $\mathsf{Y}_j(\boldsymbol{x}) \geq \mathsf{Y}(\boldsymbol{x}) - (1+\mathsf{b}(m))r/m$. Starting from the definition of Y_j and adding and removing

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

to the argument of the supremum, it holds

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

Then, from the triangle inequality and the fact that

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

we obtain

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

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

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

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 (5), it follows that

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

Combining the above and (21), we obtain

$$\mathsf{Y}(\boldsymbol{x}) - (1+\mathsf{b}(m))\frac{r}{m} \leq \mathsf{Y}_j(\boldsymbol{x}) \leq \mathsf{Y}(\boldsymbol{x}) + \frac{r}{m}\mathsf{b}(m) \ . \tag{22}$$

We now show that

$$\sum_{j=1}^{m} \big(\mathsf{Y}(\boldsymbol{x}) - \mathsf{Y}_j(\boldsymbol{x})\big) \leq \mathsf{Y}(\boldsymbol{x}) \ . \tag{23}$$

Starting from the definition of the Y_j functions, and using the linearity of expectation and the subadditivity of the supremum

$$\begin{aligned}\sum_{j=1}^{m} \mathsf{Y}_j(\boldsymbol{x}) &= \sum_{j=1}^{m} \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\left|\sum_{i=1,i\neq j}^{m} \sigma_i\left(f(x_i) - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f]\right)\right|\right] \\ &\geq \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\left|\sum_{j=1}^{m}\sum_{i=1,i\neq j}^{m} \sigma_i\left(f(x_i) - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f]\right) \ .\right|\right] \ .\end{aligned}$$

We rearrange the terms in the double sums, and use the linearity of expectation to obtain

$$\begin{aligned}\sum_{j=1}^{m} \mathsf{Y}_j(\boldsymbol{x}) &\geq \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\left|(m-1)\sum_{i=1}^{m} \sigma_i\left(f(x_i) - \hat{\mathbb{E}}_{\boldsymbol{x}}[f]\right)\right|\right] \\ &\geq (m-1)\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\left|\sum_{i=1}^{m} \sigma_i\left(f(x_i) - \hat{\mathbb{E}}_{\boldsymbol{x}}[f]\right)\right|\right] ,\end{aligned}$$

which completes our proof of (23), as the last expectation is $\mathsf{Y}(\boldsymbol{x})$.

Define now the functions

$$\mathsf{Z}(\boldsymbol{x}) \doteq \mathsf{Y}(\boldsymbol{x}) \text{ and } \mathsf{Z}_j(\boldsymbol{x}) \doteq \mathsf{Y}_j(\boldsymbol{x}) - \frac{r}{m}\mathsf{b}(m) \text{ for each } j = 1, \dots, m \ .$$

The value of $\mathsf{Z}_j(\boldsymbol{x})$ clearly does not dependent on the j-th component of $\boldsymbol{x}$. Also, from (22) it follows that

$$\mathsf{Z}_j(\boldsymbol{x}) \leq \mathsf{Z}(\boldsymbol{x}) \leq \mathsf{Z}_j(\boldsymbol{x}) + (1+2\mathsf{b}(m))\frac{r}{m} \text{ for each } j = 1, \dots, m \ .$$

A consequence of (23) is finally that

$$\sum_{j=1}^{m} \big(\mathsf{Z}(\boldsymbol{x}) - \mathsf{Z}_j(\boldsymbol{x})\big) \leq \mathsf{Z}(\boldsymbol{x}) + r\mathsf{b}(m) \ .$$

Thus Z, i.e., $\hat{\mathsf{R}}_m(\hat{\mathsf{C}}_{\boldsymbol{x}}(\mathcal{F}), \boldsymbol{x})$, is a $(1, r\mathsf{b}(m))$-self-bounding function with scale $(1+2\mathsf{b}(m))r/m$. An application of (18) from Thm. 6 completes the proof. □

Before proving Thm. 2, we need the following lemma.

Lemma 4. *It holds*

$$\mathsf{W}(\mathcal{F}) \leq \frac{m}{m-1}\mathbb{E}_{\boldsymbol{x}}[\widehat{\mathsf{W}}_{\boldsymbol{x}}(\mathcal{F})] \ .$$

Proof. Using Bessel's correction, we can rewrite the definition of wimpy variance to use the empirical expectation as

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

An application of Jensen's inequality gives

$$\mathsf{W}(\mathcal{F}) \leq \mathbb{E}_{\boldsymbol{x}}\Bigg[\underbrace{\sup_{f\in\mathcal{F}} \frac{1}{m-1}\sum_{i=1}^{m}\big(f(x_i) - \hat{\mathbb{E}}_{\boldsymbol{x}}[f]\big)^2}_{=\frac{m}{m-1}\widehat{\mathsf{W}}_{\boldsymbol{x}}(\mathcal{F})}\Bigg] \ .$$

□

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

$$\mathsf{W}(\mathcal{F}) \leq \tfrac{m}{m-1}\widehat{\mathsf{W}}_{\boldsymbol{x}}(\mathcal{F}) + \frac{r^2 \ln \frac{1}{\delta}}{m-1} + \sqrt{\left(\frac{r^2 \ln \frac{1}{\delta}}{m-1}\right)^2 + \frac{2r^2 \frac{m}{m-1}\widehat{\mathsf{W}}_{\boldsymbol{x}}(\mathcal{F}) \ln \frac{1}{\delta}}{m-1}} \ . \tag{9}$$

Proof. This proof proceeds by showing that $\widehat{\mathsf{W}}_{\boldsymbol{x}}(\mathcal{F})$ is a $(m/m-1, 0)$-self-bounding with scale r^2/m, then applying Lemma 4, and finally (17) from Thm. 6.

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 Thm. 1. Let $\hat{\mathsf{V}}_{\boldsymbol{x}}[f]$ denote the (unbiased) sample variance of f over $\boldsymbol{x}$, i.e.,

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

Define

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

and

$$\mathsf{Z}_j(\boldsymbol{x}) \doteq \sup_{f\in\mathcal{F}} \frac{1}{m-1}\sum_{i=1, i\neq j}^{m}\left(f(x_i) - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f]\right)^2 \ . \tag{24}$$

We first show that

$$\mathsf{Z}_j(\boldsymbol{x}) = \sup_{f\in\mathcal{F}}\left[\hat{\mathsf{V}}_{\boldsymbol{x}}[f] - \frac{1}{m}\left(f(x_j) - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f]\right)^2\right], \tag{25}$$

as this form comes in handy many times. Starting from the definition of Z_j in (24), we add and subtract $\frac{1}{m-1}\left(f(x_j) - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f]\right)^2$ to the argument of the supremum, and then add and subtract $\hat{\mathbb{E}}_{\boldsymbol{x}}[f]$ to the argument of the sum, to obtain:

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

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 $\sum_{i=1}^{m}(f(x_i) - \hat{\mathbb{E}}_{\boldsymbol{x}}[f]) = 0$, and another has argument $\left(\hat{\mathbb{E}}_{\boldsymbol{x}}[f] - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f]\right)^2$ independent from i, we obtain

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

It holds $\hat{\mathbb{E}}_{\boldsymbol{x}}[f] = \frac{1}{m}f(x_j) + \frac{m-1}{m}\hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f]$, so we have

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

The identity in (25) 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 Def. 2). By definition of Z_j in (24), the value of $\mathsf{Z}_j(\boldsymbol{x})$ does not depend on the j-th component of $\boldsymbol{x}$, as required by the first point in Def. 2.

We now show that, for any $j = 1, \ldots, m$, it holds,

$$\mathsf{Z}_j(\boldsymbol{x}) \leq \mathsf{Z}(\boldsymbol{x}) \leq \mathsf{Z}_j(\boldsymbol{x}) + \frac{r^2}{m} \text{ for any } \boldsymbol{x} \in \mathcal{X}^m, \tag{26}$$

as required by the second point in Def. 2. The leftmost inequality follows from the definitions of Z and Z_j. To show the rightmost inequality, we start from (25), and use the subadditivity of the supremum to obtain

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

The rightmost supremum is always smaller than r^2/m because $|f(x_j) - \hat{\mathbb{E}}_{\boldsymbol{x}_{\setminus j}}[f]| \leq r$, thus we have obtained the rightmost inequality in (26).

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

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

as in the last requirement of Def. 2. Starting again from (25) and using the subadditivity of the supremum, it holds

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

By simple algebra we then get

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

From here, we use the fact that

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

to get

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

Now by simplifying some terms on the r.h.s., we obtain

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

Collecting terms and using the original definition of Z results in

$$\sum_{j=1}^{m} \mathsf{Z}_j(\boldsymbol{x}) \geq \left(m - \frac{m}{m-1} \right) \mathsf{Z}(\boldsymbol{x}) \ .$$

Thus,

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

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 4 is

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

From here, we use the definition

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

and apply (17) from Thm. 6 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 2. *For any $\boldsymbol{x} \in \mathcal{X}^m$, it holds*

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

Furthermore, it holds

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

Proof. From the subadditivity of the supremum, it holds that

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

An application of Khintchine's inequality [12] gives

$$\hat{\mathsf{R}}_m(\mathcal{F},\boldsymbol{x}) \geq \sup_{f\in\mathcal{F}} \frac{1}{\sqrt{2}}\sqrt{\frac{\|f(\boldsymbol{x})\|_2^2}{m^2}},$$

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

$$\widehat{\mathsf{W}}^{\mathsf{r}}_{\boldsymbol{x}}(\mathcal{F}) = \frac{\|f(\boldsymbol{x})\|_2^2}{m} \ .$$

The rightmost inequality is then a corollary, using the identity $\widehat{\mathsf{W}}^{\mathsf{r}}_{\boldsymbol{x}}(\hat{\mathsf{C}}_{\boldsymbol{x}}(\mathcal{F})) = \widehat{\mathsf{W}}_{\boldsymbol{x}}(\mathcal{F})$.

The asymptotic lower bounds follow by replacing the Khintchine's inequality step with an application of the central limit theorem. □

Before proving Thm. 5 we need to introduce an important technical result. For any $u \in \mathbb{R}$, let $\mathsf{h}(u) \doteq (1+u)\ln(1+u) - u$, and let $(u)_+ \doteq \max(0,u)$.

Theorem 7 (Samson's bound, [7, Thm. 12.11])**.** *Let $\mathcal{Q}_1,\dots,\mathcal{Q}_m$ 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 $i \in \{1,\dots,m\}$, it holds $\mathbb{E}_{\mathcal{Q}_i}[g] = 0$. Now, for any $\boldsymbol{y} \in \mathcal{Y}^m$, let*

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

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[4]

$$\Pr_{\boldsymbol{y}}\left(\mathsf{Z}(\boldsymbol{y}) \leq \mathbb{E}_{\mathcal{Q}_{1:m}}[\mathsf{Z}] - \varepsilon\right) \leq \exp\left(-\frac{S^2}{4}\mathsf{h}\left(\frac{2\varepsilon}{S^2}\right)\right) \ . \tag{27}$$

Theorem 5. *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*

$$\hat{\mathsf{R}}_m(\mathcal{F},\boldsymbol{x}) \leq \hat{\mathsf{R}}^n_m(\mathcal{F},\boldsymbol{x},\boldsymbol{\sigma}) + \frac{2\hat{q}_{\mathcal{F}}(\boldsymbol{x})\ln\frac{1}{\delta}}{3nm} + \sqrt{\frac{4\widehat{\mathsf{W}}^{\mathsf{r}}_{\boldsymbol{x}}(\mathcal{F})\ln\frac{1}{\delta}}{nm}} \ . \tag{12}$$

[4] To be precise, this is an immediate consequence of the statement of [7, Thm. 2.11], through an application of the Chernoff method to the moment generating function given therein.

Proof. Without loss of generality, we assume that $\hat{q}_{\mathcal{F}}(\boldsymbol{x}) = 1$. The general case then follows via scaling.

Let

$$\mathsf{Z}(\boldsymbol{\sigma}) \doteq nm\hat{\mathsf{R}}_m^n(\mathcal{F}, \boldsymbol{x}, \boldsymbol{\sigma}) = \sum_{j=1}^{n} \sup_{f\in\mathcal{F}} \left| \sum_{i=1}^{m} \sigma_{j,i} f(x_i) \right| \ .$$

It holds $\mathbb{E}_{\boldsymbol{\sigma}}[\mathsf{Z}] = nm\hat{\mathsf{R}}_m(\mathcal{F}, \boldsymbol{x})$.

We first show that we can apply Samson's bound (Thm. 7) to Z, i.e., to the scaled MC-ERA. Consider the function family $\mathcal{F}_{\pm}$ introduced in Coro. 1, 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

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

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

$$\mathsf{Z}(\boldsymbol{\sigma}) = \sup_{\boldsymbol{f}\in(\mathcal{F}_{\pm})^n} \sum_{j=1}^{n}\sum_{i=1}^{m} \sigma_{j,i} f_j(x_i) = \sup_{g\in\mathcal{G}} \sum_{(j,i)\in\{1,\ldots n\}\times\{1,\ldots,m\}} g(\sigma_{j,i}) \ . \tag{28}$$

Thus Z has the form required by Thm. 7.

Let $\boldsymbol{\sigma}'$ denote a second $n \times m$ i.i.d. Rademacher matrix (like $\boldsymbol{\sigma}$), and define

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

It holds

$$S^2 \leq 2nm\widehat{\mathsf{W}}^{\mathsf{r}}_{\boldsymbol{x}}(\mathcal{F}) \ . \tag{29}$$

For each $g \in \mathcal{G}$, $g(\sigma_{j,i})$ and $g(\sigma_{j',i'})$ are *independent*, though not necessarily *identically distributed*, for $(j,i) \neq (j',i')$, due to the dependence of $g(\sigma_{j,i})$ on indices (j,i). It also holds, for each $g \in \mathcal{G}$, and indices (j,i), that $\mathbb{E}_{\sigma_{i,j}}[g(\sigma_{i,j})] = 0$, simply due to multiplication by symmetric (Rademacher) r.v.'s.

Thus, we can use Samson's bound (Thm. 7) 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

$$\Pr_{\boldsymbol{\sigma}} \left(\hat{\mathsf{R}}_m(\mathcal{F}, \boldsymbol{x}) \geq \hat{\mathsf{R}}_m^n(\mathcal{F}, \boldsymbol{x}, \boldsymbol{\sigma}) + \varepsilon \right) = \Pr_{\boldsymbol{\sigma}} \left(\mathbb{E}[\mathsf{Z}] \geq \mathsf{Z}(\boldsymbol{\sigma}) + nm\varepsilon \right) \leq \exp\left(-\frac{S^2}{4} \mathsf{h}\left(\frac{2nm\varepsilon}{S^2} \right) \right) \ .$$

The function

$$g(x) \doteq x\mathsf{h}\left(\frac{2nm\varepsilon}{x} \right)$$

is monotonically decreasing in its argument. Thus, using (29) gives

$$\Pr_{\boldsymbol{\sigma}} \left(\hat{\mathsf{R}}_m(\mathcal{F}, \boldsymbol{x}) \geq \hat{\mathsf{R}}_m^n(\mathcal{F}, \boldsymbol{x}, \boldsymbol{\sigma}) + \varepsilon \right) \leq \exp\left(\frac{-nm\widehat{\mathsf{W}}^{\mathsf{r}}_{\boldsymbol{x}}(\mathcal{F})}{2} \mathsf{h}\left(\frac{\varepsilon}{\widehat{\mathsf{W}}^{\mathsf{r}}_{\boldsymbol{x}}(\mathcal{F})} \right) \right) \ .$$

Now, for $u > -1/2$, define the function

$$\mathsf{h}_1(u) \doteq 1 + u - \sqrt{1 + 2u} \ .$$

Using the fact (see Boucheron et al. [7, Ch. 2.4]) that

$$\mathsf{h}(u) \geq 9\mathsf{h}_1\left(\frac{u}{3} \right) \text{ for every } u \in (-1, +\infty),$$

we obtain

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

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 Thm. 3). The variance factors $\widehat{\mathsf{W}}^{\mathsf{r}}_{\boldsymbol{x}}(\mathcal{F})$ and $\widehat{\mathsf{W}}_{\boldsymbol{x}}(\mathcal{F})$ 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. Thm. 3, both are divided by MC-trial count n, so for $n \geq 2$ trials, the Monte-Carlo error terms become negligible.

B Details on the Experimental Evaluation

As mentioned in the main text, Lemma 3 is a consequence of [27, Lemmas 26.11, 26.10], reported here for completeness.[5]

Lemma 5 (27, Lemmas 26.11, 26.10)**.** *It holds*

$$\hat{\mathsf{R}}_m(\mathcal{F}_1,\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\left\|\frac{1}{m}\sum_{i=1}^m \sigma_i x_i\right\|_\infty\right] \leq \max_i \|x_i\|_\infty \sqrt{\frac{2\ln(2d)}{m}},$$

and

$$\hat{\mathsf{R}}_m(\mathcal{F}_2,\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\left\|\frac{1}{m}\sum_{i=1}^m \sigma_i x_i\right\|_2\right] \leq \max_i \|x_i\|_2 \frac{1}{\sqrt{m}} \ .$$

We now show the centralized variants.

Lemma 3. *Let $\bar{x} \doteq \frac{1}{m}\sum_{i=1}^m x_i \in \mathbb{R}^d$. For the ℓ_1 norm, it holds*

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

while for the ℓ_2 norm, it holds

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

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 $\hat{\mathsf{R}}_m(\hat{\mathsf{C}}_{\boldsymbol{x}}(\mathcal{F}_2),\boldsymbol{x})$ is

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

where

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

Using linearity we then get

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

[5] The identities in the lemma are not reported in the original, but can be easily obtained through a slightly more refined proof than the one presented in the original. See the proof of Lemma 3 for intuition.

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

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

We thus obtain

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

From here, we can proceed as in the second part of the proof of [27, 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 [27, Lemma 26.11]), we get that

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

□

B.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 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 $\sqrt{\chi_j^2}$, where χ_k^2 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.,

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

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, $r = q = \sqrt{d}$.

B.2 Supplementary Plots

Figure 3 shows the same results as Fig. 1 (in the main text), but without the scaling of the quantities by $\sqrt{m}$. Similarly, Fig. 4 shows the same results as Fig. 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 $\hat{\mathsf{R}}^1_m(\mathcal{F}, \boldsymbol{x}, \boldsymbol{\sigma})$. 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.

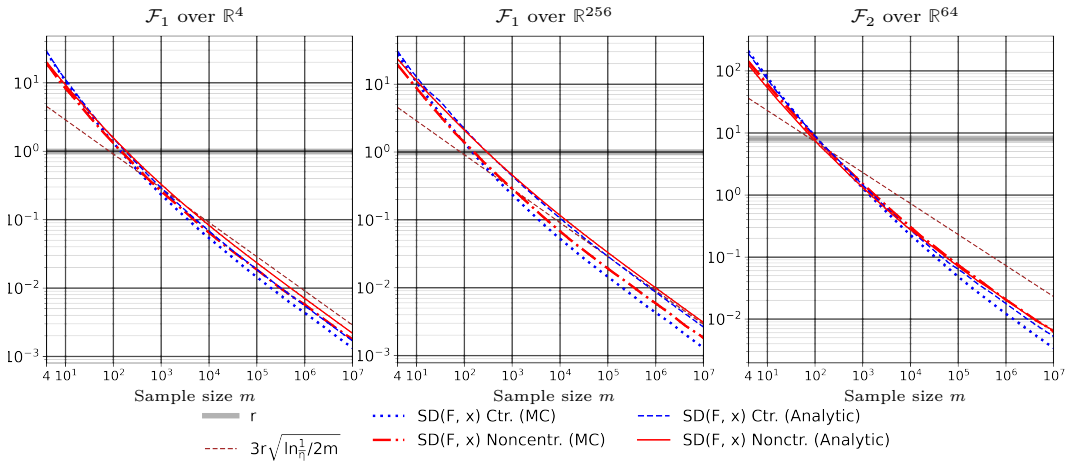

Figure 3: Comparison of SD bounds as functions of the sample size m. See the main text for an explanation of the results.

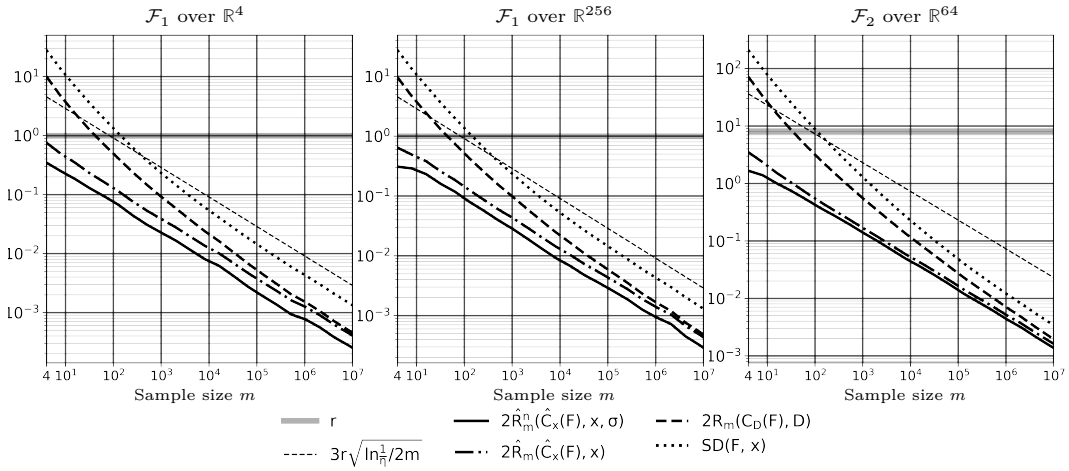

Figure 4: Comparison of SD bounds as functions of the sample size m. See the main text for an explanation of the results.