

Supplementary Material  
for  
Sharp uniform convergence bounds  
through empirical centralization

## A Proofs

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

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

*Proof.* We first show the rightmost inequality. Starting from the definition of the RA of the distributional centralization, and then subtracting and adding  $\hat{\mathbb{E}}_{\mathbf{x}}[f]$ , it holds

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

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

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

Both terms on the r.h.s. can be seen as expectations w.r.t.  $\mathbf{x}$  of the ERAs on  $\mathbf{x}$  of two sample-dependent families: the empirical centralization of  $\mathcal{F}$ , and the family

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

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

$$\mathbf{R}_m(\mathbf{C}_{\mathcal{D}}(\mathcal{F}), \mathcal{D}) \leq \mathbb{E}_{\mathbf{x}} \left[ \hat{\mathbf{R}}_m(\hat{\mathbf{C}}_{\mathbf{x}}(\mathcal{F}), \mathbf{x}) \right] + \mathbb{E}_{\mathbf{x}} \left[ \hat{\mathbf{R}}_m(\mathcal{K}_{\mathbf{x}}, \mathbf{x}) \right] . \quad (13)$$

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

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

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

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

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

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

which completes the proof of the upper bound.

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

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

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

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

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

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

Each function in  $\mathcal{Z}_{\mathbf{x}}$  is *constant*. Proceeding in exactly the same way as we did for the family  $\mathcal{K}_{\mathbf{x}}$  in the proof of the upper bound, we can write

$$\hat{R}_m(\mathcal{Z}_{\mathbf{x}}, \mathbf{x}) = \text{SD}(\mathcal{C}_{\mathcal{D}}(\mathcal{F}), \mathbf{x})\mathbf{b}(m) . \quad (16)$$

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

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

and our proof is complete.  $\square$

**Definition 2.** A function  $Z \in \mathcal{X}^m \rightarrow \mathbb{R}$  is  $(\alpha, \beta)$ -self-bounding with scale  $\gamma$ , for some  $\alpha > 0$ ,  $\beta \geq 0$ ,  $\gamma \geq 0$  if for each  $j = 1, \dots, m$ , there exists a function  $Z_j \in \mathcal{X}^m \rightarrow \mathbb{R}$  such that, for any  $\mathbf{x} \in \mathcal{X}^m$  it holds that

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

Additionally, the functions  $Z_j$ ,  $j = 1, \dots, m$ , must be such that, for any  $\mathbf{x} \in \mathcal{X}^m$ , it holds

$$\sum_{j=1}^m \left( Z(\mathbf{x}) - Z_j(\mathbf{x}) \right) \leq \alpha Z(\mathbf{x}) + \beta .$$

**Theorem 6.** Let  $Z$  be a function from  $\mathcal{X}^m$  to  $\mathbb{R}$  that is  $(\alpha, \beta)$ -self-bounding with scale  $\gamma$ , for  $\alpha \geq 1/3$ . Let  $\delta \in (0, 1)$  and let  $\mathbf{x}$  be a collection of  $m$  i.i.d. samples from  $\mathcal{X}$ . With probability at least  $1 - \delta$  over the choice of  $\mathbf{x}$ , it holds

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

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

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

*Proof.* In both cases, we will assume WLOG  $\gamma = 1$ . The results then hold by linearity, noting that if  $Z(\cdot)$  is  $\alpha$ - $\beta$  self-bounding, with scale  $\gamma$ , then  $\frac{1}{\gamma}Z(\cdot)$  is  $\alpha$ - $\beta/\gamma$  self-bounding, with scale 1; the general case thus follows by dividing out  $\gamma$ , obtaining a bound, and then multiplying through by  $\gamma$ .

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]

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

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

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

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

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

Finally, in the general case, with  $\gamma$ -scaling, we have

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

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*  $1-\beta$  self-bounding function inequality [4, Thm. 1].

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

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

which by the quadratic formula, yields

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

The general result then follows via  $\gamma$ -scaling. □

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

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

*Proof.* This proof proceeds by showing that  $\hat{\mathbf{R}}_m(\hat{\mathbf{C}}_{\mathbf{x}}(\mathcal{F}), \mathbf{x})$  is a  $(1, r\mathbf{b}(m))$ -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  $\mathbf{x} \in \mathcal{X}^m$ , let

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

and let  $\mathbf{x}_{\setminus j}$  (resp.  $\sigma_{\setminus j}$ ) denote the  $m - 1$ -dimensional vector of all but the  $j$ -th element of  $\mathbf{x}$  (resp.  $\sigma$ ). Define

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

We define these functions for convenience of notation. They will be handy when we later introduce the functions  $\mathbf{Z}$  and  $\mathbf{Z}_j$ ,  $j = 1, \dots, m$  that we want to show to be self-bounding.

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

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

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

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

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

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

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

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

By subtracting and adding  $\hat{\mathbb{E}}_{\mathbf{x}}[f]$  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

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

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

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

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

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

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

$$Y_j(\mathbf{x}) \leq Y(\mathbf{x}) + \frac{r}{m} \mathbf{b}(m) . \quad (21)$$

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

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

to the argument of the supremum, it holds

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

Then, from the triangle inequality and the fact that

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

we obtain

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

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

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

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

$$Y_j(\mathbf{x}) \geq Y(\mathbf{x}) - (1 + \mathbf{b}(m)) \frac{r}{m} .$$

Combining the above and (21), we obtain

$$\mathbf{Y}(\mathbf{x}) - (1 + \mathbf{b}(m))\frac{r}{m} \leq \mathbf{Y}_j(\mathbf{x}) \leq \mathbf{Y}(\mathbf{x}) + \frac{r}{m}\mathbf{b}(m) . \quad (22)$$

We now show that

$$\sum_{j=1}^m (\mathbf{Y}(\mathbf{x}) - \mathbf{Y}_j(\mathbf{x})) \leq \mathbf{Y}(\mathbf{x}) . \quad (23)$$

Starting from the definition of the  $\mathbf{Y}_j$  functions, and using the linearity of expectation and the subadditivity of the supremum

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

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

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

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

Define now the functions

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

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

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

A consequence of (23) is finally that

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

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

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

**Lemma 4.** *It holds*

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

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

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

An application of Jensen's inequality gives

$$\mathbf{W}(\mathcal{F}) \leq \mathbb{E}_{\mathbf{x}} \left[ \underbrace{\sup_{f \in \mathcal{F}} \frac{1}{m-1} \sum_{i=1}^m (f(x_i) - \hat{\mathbb{E}}_{\mathbf{x}}[f])^2}_{= \frac{m}{m-1} \widehat{\mathbf{W}}_{\mathbf{x}}(\mathcal{F})} \right] . \quad \square$$

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

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

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

Let  $\mathbf{x}_{\setminus j}$  denote the vector  $\mathbf{x}$  with the  $j$ -th component removed, as we defined it also in the proof for Thm. 1. Let  $\widehat{\mathbb{V}}_{\mathbf{x}}[f]$  denote the (unbiased) sample variance of  $f$  over  $\mathbf{x}$ , i.e.,

$$\widehat{\mathbb{V}}_{\mathbf{x}}[f] \doteq \frac{1}{m-1} \sum_{i=1}^m (f(x_i) - \widehat{\mathbb{E}}_{\mathbf{x}}[f])^2.$$

Define

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

and

$$Z_j(\mathbf{x}) \doteq \sup_{f \in \mathcal{F}} \frac{1}{m-1} \sum_{i=1, i \neq j}^m (f(x_i) - \widehat{\mathbb{E}}_{\mathbf{x}_{\setminus j}}[f])^2. \quad (24)$$

We first show that

$$Z_j(\mathbf{x}) = \sup_{f \in \mathcal{F}} \left[ \widehat{\mathbb{V}}_{\mathbf{x}}[f] - \frac{1}{m} (f(x_j) - \widehat{\mathbb{E}}_{\mathbf{x}_{\setminus j}}[f])^2 \right], \quad (25)$$

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

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

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

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

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

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

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  $Z_j(\mathbf{x})$  does not depend on the  $j$ -th component of  $\mathbf{x}$ , as required by the first point in Def. 2.

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

$$Z_j(\mathbf{x}) \leq Z(\mathbf{x}) \leq Z_j(\mathbf{x}) + \frac{r^2}{m} \text{ for any } \mathbf{x} \in \mathcal{X}^m, \quad (26)$$

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

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

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

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

$$\sum_{i=1}^m (Z(\mathbf{x}) - Z_j(\mathbf{x})) \leq \frac{m}{m-1} Z(\mathbf{x}),$$

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

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

By simple algebra we then get

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

From here, we use the fact that

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

to get

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

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

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

Collecting terms and using the original definition of  $Z$  results in

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

Thus,

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

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

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

From here, we use the definition

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

and apply (17) from Thm. 6 to obtain the thesis.  $\square$

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  $\alpha$  exceeds 1.

**Lemma 2.** *For any  $\mathbf{x} \in \mathcal{X}^m$ , it holds*

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

Furthermore, it holds

$$\lim_{m \rightarrow \infty} \sqrt{m} R_m(\mathcal{F}, \mathcal{D}) \geq \sqrt{\frac{2}{\pi} W^r(\mathcal{F})} \text{ and } \lim_{m \rightarrow \infty} \sqrt{m} R_m(C_{\mathcal{D}}(\mathcal{F}), \mathcal{D}) \geq \sqrt{\frac{2}{\pi} W(\mathcal{F})} .$$

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

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

An application of Khintchine's inequality [12] gives

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

where  $f(\mathbf{x})$  denotes the  $m$ -dimensional vector of values of  $f$  on  $\mathbf{x}$ . The proof of the leftmost inequality in the thesis ends by noting that

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

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

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

Before proving Thm. 5 we need to introduce an important technical result. For any  $u \in \mathbb{R}$ , let  $h(u) \doteq (1 + u) \ln(1 + u) - u$ , 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  $\mathcal{G} \subseteq \mathcal{X} \rightarrow [-1, 1]$ . 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  $\mathbf{y} \in \mathcal{Y}^m$ , let*

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

Let  $\mathbf{y} \in \mathcal{Y}^m$ , with each  $y_i \sim \mathcal{Q}_i$ , independently (but not necessarily identically, since the distributions may be different). It holds<sup>4</sup>

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

**Theorem 5.** *Let  $\sigma \in (\pm 1)^{n \times m}$  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  $\sigma$ , it holds*

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

<sup>4</sup>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  $\hat{q}_{\mathcal{F}}(\mathbf{x}) = 1$ . The general case then follows via scaling.

Let

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

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

We first show that we can apply Samson's bound (Thm. 7) to  $\mathbf{Z}$ , 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

$$(\mathcal{F}_{\pm})^n = \underbrace{\mathcal{F}_{\pm} \times \dots \times \mathcal{F}_{\pm}}_{n \text{ times}} .$$

We use  $\mathbf{f} = (f_1, \dots, f_n)$  to denote an element of  $(\mathcal{F}_{\pm})^n$ . Now, define the family

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

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

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

Thus  $\mathbf{Z}$  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

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

It holds

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

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  $\mathcal{G}$ ,  $\mathbf{Z}$ , 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

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

The function

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

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

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

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

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

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

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

we obtain

$$\Pr_{\boldsymbol{\sigma}} \left( \widehat{\mathbf{R}}_m(\mathcal{F}, \mathbf{x}) \geq \widehat{\mathbf{R}}_m^n(\mathcal{F}, \mathbf{x}, \boldsymbol{\sigma}) + \varepsilon \right) \leq \exp \left( -\frac{9}{2} nm \widehat{\mathbf{W}}_x^r(\mathcal{F}) h_1 \left( \frac{\varepsilon}{\widehat{\mathbf{W}}_x^r(\mathcal{F})} \right) \right) .$$

The result for  $\widehat{q}_{\mathcal{F}}(\mathbf{x}) = 1$  is obtained by imposing that the r.h.s. be at most  $\delta$  and solving for  $\varepsilon$  using standard sub-gamma inequalities. The general case then follows via linear scaling.  $\square$

This bound is quite comparable to Bousquet's bound on the SD (see Thm. 3). The variance factors  $\widehat{\mathbf{W}}_x^r(\mathcal{F})$  and  $\widehat{\mathbf{W}}_x(\mathcal{F})$  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.<sup>5</sup>

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

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

and

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

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  $\ell_1$  norm, it holds*

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

while for the  $\ell_2$  norm, it holds

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

*Proof.* We show the  $\ell_2$  case in detail; the reasoning for the  $\ell_1$  case is essentially the same (see details at the end of the proof). The definition of  $\widehat{\mathbf{R}}_m(\widehat{\mathbf{C}}_{\mathbf{x}}(\mathcal{F}_2), \mathbf{x})$  is

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

where

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

Using linearity, we then get

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

<sup>5</sup>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  $u \doteq \frac{1}{m} \sum_{i=1}^m \sigma_i(x_i - \bar{x})$ . The supremum is realized when

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

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

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

We thus obtain

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

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

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

## B.1 Data Generation

Our data distributions for both the  $\ell_1$  and  $\ell_2$  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  $d$ .

**$\ell_1$  Datasets** In our  $\ell_1$  experiments, each  $x_j$  is independently Beta-distributed, thus  $\mathbf{x} \sim \mathbf{B}(\boldsymbol{\alpha}_1, \boldsymbol{\beta}_1) \times \dots \times \mathbf{B}(\boldsymbol{\alpha}_d, \boldsymbol{\beta}_d)$ . 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  $\chi_k^2$  is the  $\chi^2$  distribution with  $k$  degrees of freedom. In these datasets,  $r = q = 1$ .

**$\ell_2$  Datasets** In our  $\ell_2$  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  $\mathbf{x}' \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ , and finally obtain sample  $\mathbf{x}$  by projecting  $\mathbf{x}'$  to the nonnegative hyperquadrant of the radius  $\sqrt{d}$   $\ell_2$  sphere; i.e.,

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

Taking  $\mathbf{I}_d$  to be the identity matrix, we sample  $\boldsymbol{\mu} \sim \mathcal{N}(\mathbf{1}, \mathbf{I}_d)$ , and taking  $\mathbf{a} \sim \mathcal{U}(0, 1)^{d \times d}$ , we let  $\boldsymbol{\Sigma} \doteq \frac{\mathbf{a} \mathbf{a}^\top}{d} + \mathbf{I}_d$ . 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  $\hat{\mathbf{R}}_m^1(\mathcal{F}, \mathbf{x}, \boldsymbol{\sigma})$ . We stress that this term *does not* include the MC-ERA itself, 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 improvement of *variance-sensitive* bounds over the (range-only) McDiarmid bounds.

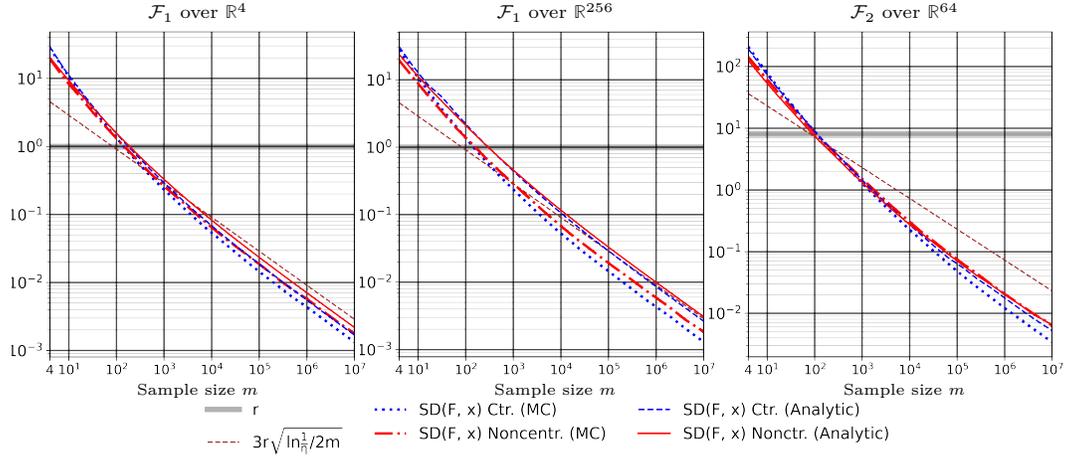


Figure 3: Comparison of SD bounds as functions of the sample size  $m$ . See the main text for an explanation of the results.

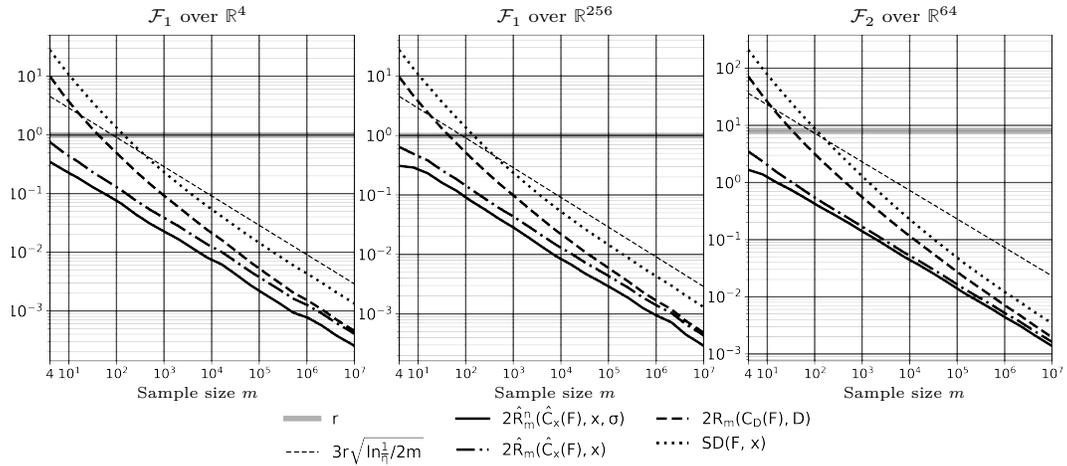


Figure 4: Comparison of SD bounds as functions of the sample size  $m$ . See the main text for an explanation of the results.