Relative Deviation Learning Bounds and Generalization with Unbounded Loss Functions

Corinna Cortes · Spencer Greenberg · Mehryar Mohri

January, 2019

Abstract We present an extensive analysis of relative deviation bounds, including detailed proofs of two-sided inequalities and their implications. We also give detailed proofs of two-sided generalization bounds that hold in the general case of unbounded loss functions, under the assumption that a moment of the loss is bounded. We then illustrate how to apply these results in a sample application: the analysis of importance weighting.

Keywords generalization bounds · learning theory · unbounded loss functions · relative deviation bounds · importance weighting · unbounded regression · machine learning

1 Introduction

Most generalization bounds in learning theory hold only for bounded loss functions. This includes standard VC-dimension bounds [47], Rademacher complexity [30, 4, 31, 6] or local Rademacher complexity bounds [29, 5], as well as most other bounds based on other complexity terms. This assumption is typically unrelated to the statistical nature of the problem considered but it is convenient since when the loss functions are uniformly bounded, standard tools such as Hoeffding's inequality [24, 3], McDiarmid's inequality [35], or Talagrand's concentration inequality [46] apply.

There are however natural learning problems where the boundedness assumption does not hold. This includes unbounded regression tasks where the target labels are not uniformly bounded, and a variety of applications such as sample

Corinna Cortes
Google Research, 76 Ninth Avenue, New York, NY 10011
E-mail: corinna@google.com

Spencer Greenberg
E-mail: sgg247@nyu.edu

Mehryar Mohri
Courant Institute and Google Research, 251 Mercer Street, New York, NY 10012
E-mail: mohri@cims.nyu.edu

bias correction [19,25,15,45,9], domain adaptation [7,10,18,28,33,14], or the analysis of boosting [17], where the importance weighting technique is used [12]. It is therefore critical to derive learning guarantees that hold for these scenarios and the general case of unbounded loss functions.

When the class of functions is unbounded, a single function may take arbitrarily large values with arbitrarily small probabilities. This is probably the main challenge in deriving uniform convergence bounds for unbounded losses. And the methods of [32] do not seem to help in this context, for instance because the probability of the event that a function takes arbitrarily larger values is not known, and it would not be useful to rule it out with a high-probability statement.

The problem, however, can be avoided by assuming the existence of an envelope, that is a single non-negative function with a finite expectation lying above the absolute value of the loss of every function in the hypothesis set [20,40,21, 41,23]. An alternative assumption similar to Hoeffding's inequality based on the expectation of a hyperbolic function, a quantity similar to the moment-generating function, is used by [36]. However, in many problems, e.g., in the analysis of importance weighting even for common distributions, there exists no suitable envelope function [12].

Instead, the second or some other αth-moment of the loss seems to play a critical role in the analysis. Thus, we will adopt here the assumption that some αth-moment of the loss functions is bounded as in [47,48].

Relatedly, in [8] they prove a concentration inequality under a second moment assumption on the envelope function class, using the entropy method of Ledoux. But their bounds hold only for countable function classes, while our guarantees hold for arbitrary classes with finite expected shattering coefficients. They further require an assumption on the boundedness of the average of the supremum over the class of functions of the empirical error times the root of the sample size, as well as a bracketing assumption for some results.

Another approach is taken in [37], which avoids concentration based methods for the squared loss function, though the bounds produced are not straightforward to interpret. The approach taken in [34], on the other hand, is to derive general risk bounds for the empirical risk minimizer when the classification rules belong to a VC-class under margin conditions. The paper [27] gives a different perspective, giving asymptotic results for families of distributions bounded above or below by certain conditions, leading to a limiting Gumbel distribution, suggesting that under some conditions there might be alternative bounds other than sub-Gaussian ones.

In [38] a clever symmetrization approach is applied that seems to lead to self-normalized deviation bounds for VC-classes based on the mixture of the variance and empirical variance. However, the proof seems to lack some explanation, since the covering number estimate used is for VC-classes of bounded functions. Self-normalized processes have attracted attention in other domains as well, outside of the Machine Learning community, for instance in [39] and [8].

In this paper, we present in detail two-sided generalization bounds for unbounded loss functions under the assumption that some αth-moment of the loss functions, $\alpha > 1$, is bounded. The proof of these bounds makes use of relative deviation generalization bounds in binary classification, which we also prove and discuss in detail. Much of the results and material we present is not novel and the paper has therefore a survey nature. However, our presentation is motivated by the

fact that the proofs given in the past for these generalization bounds either had errors or were incomplete. We also apply these results to two sample applications of particular interest: importance weighting, and excess risk bounds in the context of the Tsybakov noise condition.

We now discuss in more detail prior results and proofs. One-side relative deviation bounds were first given by [47], later improved by a constant factor by [1]. These publications and several others have all relied on a lower bound on the probability that a binomial random variable of m trials exceeds its expected value when the bias verifies $p > \frac{1}{m}$. This also later appears in [48] and implicitly in other publications referring to the relative deviations bounds of [47]. To the best of our knowledge, no proof of this inequality was given in the past in the machine learning literature before our work [22]. One attempt was made to prove this lemma in the context of the analysis of some generalization bounds [26], but that proof was not sufficient to support the general case needed for the proof of the relative deviation bound of [47].

We present the proof of two-sided relative deviation bounds in detail using the recent results of [22]. The two-sided versions we present, as well as several consequences of these bounds, appear in [2]. However, we could not find a full proof of the two-sided bounds in any prior publication. Our presentation shows that the proof of the other side of the inequality is not symmetric and cannot be immediately obtained from that of the first side inequality. Additionally, this requires another proof related to the binomial distributions given by [22].

Relative deviation bounds are very informative guarantees in machine learning of independent interest, beyond the key role they play in the proof of unbounded loss learning bounds. They lead to sharper generalization bounds whose right-hand side is expressed as the interpolation of a $O(1/m)$ term and a $O(1/\sqrt{m})$ term that admits as a multiplier the empirical error or the generalization error. In particular, when the empirical error is zero, this leads to faster rate bounds. We present in detail the proof of this type of results as well as that of several others of interest [2]. In the form presented by [47], relative deviation bounds suffer from a discontinuity at zero (zero denominator), a problem that also affects inequalities for the other side, and which seems not to have been treated by previous work. Our proofs and results explicitly deal with this issue.

We use relative deviations bounds to give the full proofs of two-sided generalization bounds for unbounded losses with finite moments of order α, both in the case $1 < \alpha \leq 2$ and the case $\alpha > 2$. One-sided generalization bounds for unbounded loss functions were first given by [47,48] under the same assumptions and also using relative deviations. The one-sided version of our bounds for the case $1 < \alpha \leq 2$ coincides with that of [47,48] modulo a constant factor, but the proofs in both books appear to be in error.[1] The core component of our proof is based on a different technique using Hölder's inequality. We also present some more explicit bounds for the case $1 < \alpha \leq 2$ by approximating a complex term appearing in these bounds. The one-sided version of the bounds for the case $\alpha > 2$

[1] In [47, p.204–206], statement (5.37) cannot be derived from assumption (5.35) as required, and in general it does not hold: the first integral in (5.37) is restricted to a sub-domain and is thus smaller than the integral of (5.35). Furthermore, the main statement in Section (5.6.2) is not valid. In [48, p.200–202], the *Lagrange method* is invoked to show the main inequality. But with our best efforts, we could not justify some of the steps. In particular, the way function z is concluded to be equal to one over the first interval without sufficient justification.

are also due to [47,48] with similar gaps in the proofs.[2] In that case as well, we give detailed proofs using the Cauchy-Schwarz inequality in the most general case where a positive constant is used in the denominator to avoid the discontinuity at zero. These learning bounds can be used directly in the analysis of unbounded loss functions as in the case of importance weighting [12].

The remainder of this paper is organized as follows. In Section 2, we briefly introduce some definitions and notation used in the next sections. Section 3 presents in detail relative deviation bounds as well as several of their consequences. Next, in Section 4 we present generalization bounds for unbounded loss functions under the assumption that the moment of order α is bounded first in the case $1 < \alpha \leq 2$ (Section 4.1), then in the case $\alpha > 2$ (Section 4.2). Finally, in Section 5 we demonstrate two sample applications of these results.

2 Preliminaries

We consider an input space $\mathcal{X}$ and an output space $\mathcal{Y}$, which in the particular case of binary classification is $\mathcal{Y} = \{-1, +1\}$ or $\mathcal{Y} = \{0, 1\}$, or a measurable subset of $\mathbb{R}$ in regression. We denote by D a distribution over $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. For a sample S of size m drawn from D^m, we will denote by $\widehat{D}$ the corresponding empirical distribution, that is the distribution corresponding to drawing a point from S uniformly at random. Throughout this paper, H denotes a hypothesis class of functions mapping from $\mathcal{X}$ to $\mathcal{Y}$. The loss incurred by hypothesis $h \in H$ at $z \in \mathcal{Z}$ is denoted by $L(h, z)$. L is assumed to be non-negative, but not necessarily bounded. We denote by $\mathcal{L}(h)$ the expected loss or generalization error of a hypothesis $h \in H$ and by $\widehat{\mathcal{L}}_S(h)$ its empirical loss for a sample S:

$$\mathcal{L}(h) = \mathop{\mathrm{E}}_{z\sim D}[L(h, z)] \qquad \widehat{\mathcal{L}}_S(h) = \mathop{\mathrm{E}}_{z\sim \widehat{D}}[L(h, z)]. \tag{1}$$

For any $\alpha > 0$, we also use the notation $\mathcal{L}_\alpha(h) = \mathrm{E}_{z\sim D}[L^\alpha(h, z)]$ and $\widehat{\mathcal{L}}_\alpha(h) = \mathrm{E}_{z\sim \widehat{D}}[L^\alpha(h, z)]$ for the αth moments of the loss. When the loss L coincides with the standard zero-one loss used in binary classification, we equivalently use the following notation

$$R(h) = \mathop{\mathrm{E}}_{z=(x,y)\sim D}[1_{h(x)\neq y}] \qquad \widehat{R}_S(h) = \mathop{\mathrm{E}}_{z=(x,y)\sim \widehat{D}}[1_{h(x)\neq y}]. \tag{2}$$

We will sometimes use the shorthand x_1^m to denote a sample of $m > 0$ points $(x_1, \ldots, x_m) \in \mathcal{X}^m$. For any hypothesis set H of functions mapping $\mathcal{X}$ to $\mathcal{Y} = \{-1, +1\}$ or $\mathcal{Y} = \{0, 1\}$ and sample x_1^m, we denote by $\mathbb{S}_H(x_1^m)$ the number of distinct dichotomies generated by H over that sample and by $\Pi_m(H)$ the growth function:

$$\mathbb{S}_H(x_1^m) = \mathrm{Card}\left(\{(h(x_1), \ldots, h(x_m)) \colon h \in H\}\right) \tag{3}$$

$$\Pi_m(H) = \max_{x_1^m \in \mathcal{X}^m} \mathbb{S}_H(x_1^m). \tag{4}$$

[2] Several of the comments we made for the case $1 < \alpha \leq 2$ hold here as well. Some of the given steps we could not justify, even with our best efforts.

3 Relative deviation bounds

In this section we prove a series of relative deviation learning bounds which we use in the next section for deriving generalization bounds for unbounded loss functions. We will assume throughout the paper, as is common in much of learning theory, that each expression of the form $\sup_{h\in H}[...]$ is a measurable function, which is not guaranteed when H is not a countable set. This assumption holds nevertheless in most common applications of machine learning.

We start with the proof of a symmetrization lemma (Lemma 2) originally presented by [47], which is used and improved by [1]. These publications and several others have all relied on a lower bound on the probability that a binomial random variable of m trials exceeds its expected value when the bias verifies $p > \frac{1}{m}$. To our knowledge, no rigorous proof of this fact was ever provided previously in the literature in the full generality needed. The proof of this result (Lemma 1) was recently given by us [22].

Lemma 1 *Let X be a random variable distributed according to the binomial distribution $B(m,p)$ with m a positive integer (the number of trials) and $p > \frac{1}{m}$ (the probability of success of each trial). Then, the following inequality holds:*

$$\Pr\left[X \geq \mathrm{E}[X]\right] > \frac{1}{4}, \tag{5}$$

and, if instead of requiring $p > \frac{1}{m}$ we require $p < 1 - \frac{1}{m}$, then

$$\Pr\left[X \leq \mathrm{E}[X]\right] > \frac{1}{4}, \tag{6}$$

where in both cases $\mathrm{E}[X] = mp$.

The lower bound in (5) is never reached but is approached asymptotically when $m = 2$ as $p \to \frac{1}{2}$ from the right.

Our symmetrization lemma proof (Lemma 2) is more concise than that of [47]. Furthermore, our statement and proof handle the technical problem of discontinuity at zero ignored by previous authors. The denominator may in general become zero, which would lead to an undefined result. We resolve this issue by including an arbitrary positive constant τ in the denominator in most of our expressions.

For the proof of the following result, we will use the function F defined over $(0,+\infty)\times(0,+\infty)$ by $F\colon (x,y) \mapsto \frac{x-y}{\sqrt[\alpha]{\frac{1}{2}[x+y+\frac{1}{m}]}}$. By Lemma 4, $F(x,y)$ is increasing in x and decreasing in y.

Lemma 2 *Let $1 < \alpha \leq 2$. Assume that $m\epsilon^{\frac{\alpha}{\alpha-1}} > 1$. Then, for any hypothesis set H and any $\tau > 0$, the following holds:*

$$\Pr_{S\sim D^m}\left[\sup_{h\in H}\frac{R(h)-\widehat{R}_S(h)}{\sqrt[\alpha]{R(h)+\tau}} > \epsilon\right] \leq 4\Pr_{S,S'\sim D^m}\left[\sup_{h\in H}\frac{\widehat{R}_{S'}(h)-\widehat{R}_S(h)}{\sqrt[\alpha]{\frac{1}{2}[\widehat{R}_S(h)+\widehat{R}_{S'}(h)+\frac{1}{m}]}} > \epsilon\right].$$

Proof We give a concise version of the proof given by [47]. We first show that the following implication holds for any $h \in H$:

$$\left(\frac{R(h)-\widehat{R}_S(h)}{\sqrt[\alpha]{R(h)+\tau}} > \epsilon\right) \wedge \left(\widehat{R}_{S'}(h) > R(h)\right) \Rightarrow F(\widehat{R}_{S'}(h), \widehat{R}_S(h)) > \epsilon. \tag{7}$$

The first condition can be equivalently rewritten as $\widehat{R}_S(h) < R(h) - \epsilon(R(h)+\tau)^{\frac{1}{\alpha}}$, which implies

$$\widehat{R}_S(h) < R(h) - \epsilon R(h)^{\frac{1}{\alpha}} \qquad \text{and} \qquad \epsilon^{\frac{\alpha}{\alpha-1}} < R(h), \tag{8}$$

since $\widehat{R}_S(h) \geq 0$. Assume that the antecedent of the implication (7) holds for $h \in H$. Then, in view of the monotonicity properties of function F (Lemma 4), we can write:

$$\begin{aligned}
F(\widehat{R}_{S'}(h), \widehat{R}_S(h)) &\geq F(R(h), R(h) - \epsilon R(h)^{\frac{1}{\alpha}}) \qquad (\widehat{R}_{S'}(h) > R(h) \text{ and 1st ineq. of (8)}) \\
&= \frac{R(h) - (R(h) - \epsilon R(h)^{\frac{1}{\alpha}})}{\sqrt[\alpha]{\frac{1}{2}[2R(h) - \epsilon R(h)^{\frac{1}{\alpha}} + \frac{1}{m}]}} \\
&\geq \frac{\epsilon R(h)^{\frac{1}{\alpha}}}{\sqrt[\alpha]{\frac{1}{2}[2R(h) - \epsilon^{\frac{\alpha}{\alpha-1}} + \frac{1}{m}]}} \qquad \text{(2nd ineq. of (8))} \\
&> \frac{\epsilon R(h)^{\frac{1}{\alpha}}}{\sqrt[\alpha]{\frac{1}{2}[2R(h)]}} = \epsilon, \qquad (m\epsilon^{\frac{\alpha}{\alpha-1}} > 1)
\end{aligned}$$

which proves (7). Now, by definition of the supremum, for any $\eta > 0$, there exists $h_0 \in H$ such that

$$\sup_{h \in H} \frac{R(h) - \widehat{R}_S(h)}{\sqrt[\alpha]{R(h)+\tau}} - \frac{R(h_0) - \widehat{R}_S(h_0)}{\sqrt[\alpha]{R(h_0)+\tau}} \leq \eta. \tag{9}$$

Using the definition of h_0 and implication (7), we can write

$$\begin{aligned}
&\Pr_{S,S' \sim D^m}\left[\sup_{h \in H} \frac{\widehat{R}_{S'}(h) - \widehat{R}_S(h)}{\sqrt[\alpha]{\frac{1}{2}[\widehat{R}_S(h) + \widehat{R}_{S'}(h) + \frac{1}{m}]}} > \epsilon\right] \\
&\geq \Pr_{S,S' \sim D^m}\left[\frac{\widehat{R}_{S'}(h_0) - \widehat{R}_S(h_0)}{\sqrt[\alpha]{\frac{1}{2}[\widehat{R}_S(h_0) + \widehat{R}_{S'}(h_0) + \frac{1}{m}]}} > \epsilon\right] \qquad \text{(by def. of sup)} \\
&\geq \Pr_{S,S' \sim D^m}\left[\left(\frac{R(h_0) - \widehat{R}_S(h_0)}{\sqrt[\alpha]{R(h_0)+\tau}} > \epsilon\right) \wedge \left(\widehat{R}_{S'}(h_0) > R(h_0)\right)\right] \qquad \text{(implication (7))} \\
&= \Pr_{S \sim D^m}\left[\frac{R(h_0) - \widehat{R}_S(h_0)}{\sqrt[\alpha]{R(h_0)+\tau}} > \epsilon\right] \Pr_{S' \sim D^m}\left[\widehat{R}_{S'}(h_0) > R(h_0)\right] \qquad \text{(independence)}.
\end{aligned}$$

We now show that this implies the following inequality

$$\Pr_{S,S' \sim D^m}\left[\sup_{h \in H} \frac{\widehat{R}_{S'}(h) - \widehat{R}_S(h)}{\sqrt[\alpha]{\frac{1}{2}[\widehat{R}_S(h) + \widehat{R}_{S'}(h) + \frac{1}{m}]}} > \epsilon\right] \geq \frac{1}{4} \Pr_{S \sim D^m}\left[\sup_{h \in H} \frac{R(h) - \widehat{R}_S(h)}{\sqrt[\alpha]{R(h)+\tau}} > \epsilon + \eta\right], \tag{10}$$

by distinguishing two cases. If $R(h_0) > \epsilon^{\frac{\alpha}{\alpha-1}}$, since $\epsilon^{\frac{\alpha}{\alpha-1}} > \frac{1}{m}$, by Lemma 1 the inequality $\Pr_{S' \sim D^m}\left[\widehat{R}_{S'}(h_0) > R(h_0)\right] > \frac{1}{4}$ holds, which yields immediately (10).

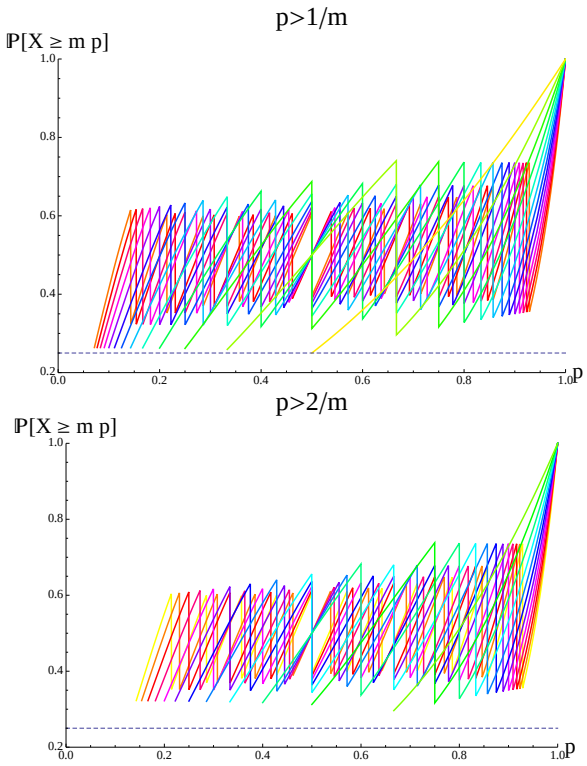

Fig. 1 These plots depict $\Pr[X \geq \mathrm{E}[X]]$, the probability that a binomially distributed random variable X exceeds its expectation, as a function of the trial success probability p. The top plot shows only regions satisfying $p > \frac{1}{m}$ whereas the bottom plot shows only regions satisfying $p > \frac{2}{m}$. Each colored line corresponds to a different number of trials, $m = 2, 3, \ldots, 14$. The dashed horizontal line at $\frac{1}{4}$ represents the value of the lower bound used in the proof of lemma 2

Otherwise we have $R(h_0) \leq \epsilon^{\frac{\alpha}{\alpha-1}}$. Then, by (8), the condition $\frac{R(h_0)-\widehat{R}_S(h_0)}{\sqrt[\alpha]{R(h_0)+\tau}} > \epsilon$ cannot hold for any sample $S \sim D^m$ which by (9) implies that the condition $\sup_{h\in H} \frac{R(h)-\widehat{R}_S(h)}{\sqrt[\alpha]{R(h)+\tau}} > \epsilon + \eta$ cannot hold for any sample $S \sim D^m$, in which case (10) trivially holds. Now, since (10) holds for all $\eta > 0$, we can take the limit $\eta \to 0$ and use the right-continuity of the cumulative distribution to obtain

$$\Pr_{S,S'\sim D^m}\left[\sup_{h\in H} \frac{\widehat{R}_{S'}(h) - \widehat{R}_S(h)}{\sqrt[\alpha]{\frac{1}{2}[\widehat{R}_S(h) + \widehat{R}_{S'}(h) + \frac{1}{m}]}} > \epsilon\right] \geq \frac{1}{4} \Pr_{S\sim D^m}\left[\sup_{h\in H} \frac{R(h) - \widehat{R}_S(h)}{\sqrt[\alpha]{R(h)+\tau}} > \epsilon\right],$$

which completes the proof of Lemma 2.

Note that the factor of 4 in the statement of lemma 2 can be modestly improved by changing the condition assumed from $\epsilon^{\frac{\alpha}{\alpha-1}} > \frac{1}{m}$ to $\epsilon^{\frac{\alpha}{\alpha-1}} > \frac{k}{m}$ for constant values of $k > 1$. This leads to a slightly better lower bound on $\Pr_{S'\sim D^m}\left[\widehat{R}_{S'}(h_0) > R(h_0)\right]$, e.g. 3.375 rather than 4 for $k = 2$, at the expense of not covering cases where the

number of samples m is less than $\frac{k}{\epsilon^{\frac{\alpha}{\alpha-1}}}$. For some values of k, e.g. $k=2$, covering these cases is not needed for the proof of our main theorem (Theorem 1) though. However, this does not seem to simplify the critical task of proving a lower bound on $\Pr_{S'\sim D^m}\left[\widehat{R}_{S'}(h_0) > R(h_0)\right]$, that is the probability that a binomial random variable $B(m,p)$ exceeds its expected value when $p > \frac{k}{m}$. One might hope that restricting the range of p in this way would help simplify the proof of a lower bound on the probability of a binomial exceeding its expected value. However, our proof in [22] suggests that this is not the case, since the regime where p is small seems to be the easiest one to analyze (Fig. 1).

The result of Lemma 2 is a one-sided inequality. The proof of a similar result (Lemma 3) with the roles of $R(h)$ and $\widehat{R}_S(h)$ interchanged makes use of (6).

The proof of the following lemma (Lemma 3) is novel.[3] While the general strategy of the proof is similar to that of Lemma 2, there are some non-trivial differences due to the requirement $p < 1 - \frac{1}{m}$ of (6). The proof is not symmetric as shown by the details given below.

Lemma 3 *Let $1 < \alpha \leq 2$. Assume that $m\epsilon^{\frac{\alpha}{\alpha-1}} > 1$. Then, for any hypothesis set H and any $\tau > 0$ the following holds:*

$$\Pr_{S\sim D^m}\left[\sup_{h\in H}\frac{\widehat{R}_S(h) - R(h)}{\sqrt[\alpha]{\widehat{R}_S(h)+\tau}} > \epsilon\right] \leq 4\Pr_{S,S'\sim D^m}\left[\sup_{h\in H}\frac{\widehat{R}_{S'}(h) - \widehat{R}_S(h)}{\sqrt[\alpha]{\frac{1}{2}[\widehat{R}_S(h)+\widehat{R}_{S'}(h)+\frac{1}{m}]}} > \epsilon\right]$$

Proof Proceeding in a way similar to the proof of Lemma 2, we first show that the following implication holds for any $h \in H$:

$$\left(\frac{\widehat{R}_S(h) - R(h)}{\sqrt[\alpha]{\widehat{R}_S(h)+\tau}} > \epsilon\right) \wedge \left(R(h) \geq \widehat{R}_{S'}(h)\right) \Rightarrow F(\widehat{R}_S(h), \widehat{R}_{S'}(h)) > \epsilon. \tag{11}$$

The first condition can be equivalently rewritten as $R(h) < \widehat{R}_S(h) - \epsilon(\widehat{R}_S(h)+\tau)^{\frac{1}{\alpha}}$, which implies

$$R(h) < \widehat{R}_S(h) - \epsilon\widehat{R}_S(h)^{\frac{1}{\alpha}} \qquad \text{and} \qquad \epsilon^{\frac{\alpha}{\alpha-1}} < \widehat{R}_S(h), \tag{12}$$

since $\widehat{R}_S(h) \geq 0$. Assume that the antecedent of the implication (11) holds for $h \in H$. Then, in view of the monotonicity properties of function F (Lemma 4), we can write:

$$\begin{aligned}
F(\widehat{R}_S(h), \widehat{R}_{S'}(h)) &\geq F(\widehat{R}_S(h), R(h)) && (R(h) \geq \widehat{R}_{S'}(h)) \\
&\geq F(\widehat{R}_S(h), \widehat{R}_S(h) - \epsilon\widehat{R}_S(h)^{\frac{1}{\alpha}}) && (\text{1st ineq. of (12)}) \\
&= \frac{\widehat{R}_S(h) - (\widehat{R}_S(h) - \epsilon\widehat{R}_S(h)^{\frac{1}{\alpha}})}{\sqrt[\alpha]{\frac{1}{2}[2\widehat{R}_S(h) - \epsilon\widehat{R}_S(h)^{\frac{1}{\alpha}} + \frac{1}{m}]}} \\
&\geq \frac{\epsilon R(h)^{\frac{1}{\alpha}}}{\sqrt[\alpha]{\frac{1}{2}[2R(h) - \epsilon^{\frac{\alpha}{\alpha-1}} + \frac{1}{m}]}} && (\text{2nd ineq. of (12)}) \\
&> \frac{\epsilon R(h)^{\frac{1}{\alpha}}}{\sqrt[\alpha]{\frac{1}{2}[2R(h)]}} = \epsilon, && (m\epsilon^{\frac{\alpha}{\alpha-1}} > 1)
\end{aligned}$$

[3] A version of this lemma is stated in [11], but no proof is given.

which proves (11). For the application of (6) to a hypothesis h, the condition $R(h) < 1 - \frac{1}{m}$ is required. Observe that this is implied by the assumptions $\widehat{R}_S(h) \geq \epsilon^{\frac{\alpha}{\alpha-1}}$ and $m\epsilon^{\frac{\alpha}{\alpha-1}} > 1$:

$$R(h) < \widehat{R}_S(h) - \epsilon \sqrt[\alpha]{\widehat{R}_S(h)} \leq 1 - \epsilon\, \epsilon^{\frac{1}{\alpha-1}} = 1 - \epsilon^{\frac{\alpha}{\alpha-1}} < 1 - \frac{1}{m}.$$

The rest of the proof proceeds nearly identically to that of Lemma 2.

In the statements of all the following results, the term $\mathrm{E}_{x_1^{2m} \sim D^{2m}}[\mathbb{S}_H(x_1^{2m})]$ can be replaced by the upper bound $\Pi_{2m}(H)$ to derive simpler expressions. By Sauer's lemma [42,49], the VC-dimension d of the family H can be further used to bound these quantities since $\Pi_{2m}(H) \leq \left(\frac{2em}{d}\right)^d$ for $d \leq 2m$. The first inequality of the following theorem was originally stated and proven by [47,48], later by [1] (in the special case $\alpha = 2$) with a somewhat more favorable constant, in both cases modulo the incomplete proof of the symmetrization and the technical issue related to the denominator taking the value zero, as already pointed out. The second inequality of the theorem and its proof are novel. Our proofs benefit from the improved analysis of [1].

Theorem 1 *For any hypothesis set H of functions mapping a set $\mathcal{X}$ to $\{0,1\}$, and any fixed $1 < \alpha \leq 2$ and $\tau > 0$, the following two inequalities hold:*

$$\Pr_{S \sim D^m}\left[\sup_{h \in H} \frac{R(h) - \widehat{R}_S(h)}{\sqrt[\alpha]{R(h) + \tau}} > \epsilon\right] \leq 4\, \mathrm{E}[\mathbb{S}_H(x_1^{2m})] \exp\left(\frac{-m^{\frac{2(\alpha-1)}{\alpha}}\epsilon^2}{2^{\frac{\alpha+2}{\alpha}}}\right)$$

$$\Pr_{S \sim D^m}\left[\sup_{h \in H} \frac{\widehat{R}_S(h) - R(h)}{\sqrt[\alpha]{\widehat{R}_S(h) + \tau}} > \epsilon\right] \leq 4\, \mathrm{E}[\mathbb{S}_H(x_1^{2m})] \exp\left(\frac{-m^{\frac{2(\alpha-1)}{\alpha}}\epsilon^2}{2^{\frac{\alpha+2}{\alpha}}}\right).$$

Note that $\tau > 0$ is required to ensure that the ratio the supremum is taken over will never have its numerator and denominator be zero at the same time.

Proof We first consider the case where $m\epsilon^{\frac{\alpha}{\alpha-1}} \leq 1$, which is not covered by Lemma 2. We can then write

$$4\, \mathrm{E}[\mathbb{S}_H(x_1^{2m})] \exp\left[\frac{-m^{\frac{2(\alpha-1)}{\alpha}}\epsilon^2}{2^{\frac{\alpha+2}{\alpha}}}\right] \geq 4\, \mathrm{E}[\mathbb{S}_H(x_1^{2m})] \exp\left[\frac{-1}{2^{\frac{\alpha+2}{\alpha}}}\right] > 1,$$

for $1 < \alpha \leq 2$. Thus, the bounds of the theorem hold trivially in that case. On the other hand, when $m\epsilon^{\frac{\alpha}{\alpha-1}} \geq 1$, we can apply Lemma 2 and Lemma 3. Therefore, to prove theorem 1, it is sufficient to work with the symmetrized expression $\sup_{h \in H} \frac{\widehat{R}_{S'}(h) - \widehat{R}_S(h)}{\sqrt[\alpha]{\frac{1}{2}[\widehat{R}_S(h) + \widehat{R}_{S'}(h) + \frac{1}{m}]}}$, rather than working directly with our original expressions $\sup_{h \in H} \frac{R(h) - \widehat{R}_S(h)}{\sqrt[\alpha]{R(h)+\tau}}$ and $\sup_{h \in H} \frac{\widehat{R}(h) - R_S(h)}{\sqrt[\alpha]{\widehat{R}(h)+\tau}}$. To upper bound the probability that the symmetrized expression is larger than ϵ, we begin by introducing a vector of Rademacher random variables $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_m)$, where the σ_i are

independent, identically distributed random variables each equally likely to take the value $+1$ or -1. Using the shorthand x_1^{2m} for $(x_1, \ldots, x_{2m})$, we can then write

$$\begin{aligned}
&\Pr_{S,S' \sim D^m}\left[\sup_{h \in H} \frac{\widehat{R}_{S'}(h) - \widehat{R}_S(h)}{\sqrt[\alpha]{\frac{1}{2}[\widehat{R}_S(h) + \widehat{R}_{S'}(h) + \frac{1}{m}]}} > \epsilon\right] \\
&= \Pr_{x_1^{2m} \sim D^{2m}}\left[\sup_{h \in H} \frac{\frac{1}{m}\sum_{i=1}^m (h(x_{m+i}) - h(x_i))}{\sqrt[\alpha]{\frac{1}{2m}[\sum_{i=1}^m (h(x_{m+i}) + h(x_i)) + 1]}} > \epsilon\right] \\
&= \Pr_{x_1^{2m} \sim D^{2m}, \boldsymbol{\sigma}}\left[\sup_{h \in H} \frac{\frac{1}{m}\sum_{i=1}^m \sigma_i(h(x_{m+i}) - h(x_i))}{\sqrt[\alpha]{\frac{1}{2m}[\sum_{i=1}^m (h(x_{m+i}) + h(x_i)) + 1]}} > \epsilon\right] \\
&= \mathop{\mathrm{E}}_{x_1^{2m} \sim D^{2m}}\left[\Pr_{\boldsymbol{\sigma}}\left[\sup_{h \in H} \frac{\frac{1}{m}\sum_{i=1}^m \sigma_i(h(x_{m+i}) - h(x_i))}{\sqrt[\alpha]{\frac{1}{2m}[\sum_{i=1}^m (h(x_{m+i}) + h(x_i)) + 1]}} > \epsilon \,\middle|\, x_1^{2m}\right]\right].
\end{aligned}$$

Now, for a fixed x_1^{2m}, we have $\mathrm{E}_{\boldsymbol{\sigma}}\left[\frac{\frac{1}{m}\sum_{i=1}^m \sigma_i(h(x_{m+i}) - h(x_i))}{\sqrt[\alpha]{\frac{1}{2m}[\sum_{i=1}^m (h(x_{m+i}) + h(x_i)) + 1]}}\right] = 0$, thus, by Hoeffding's inequality, we can write

$$\begin{aligned}
&\Pr_{\boldsymbol{\sigma}}\left[\frac{\frac{1}{m}\sum_{i=1}^m \sigma_i(h(x_{m+i}) - h(x_i))}{\sqrt[\alpha]{\frac{1}{2m}[\sum_{i=1}^m (h(x_{m+i}) + h(x_i)) + 1]}} > \epsilon \,\middle|\, x_1^{2m}\right] \\
&\leq \exp\left(\frac{-[\sum_{i=1}^m (h(x_{m+i}) + h(x_i)) + 1]^{\frac{2}{\alpha}} m^{\frac{2(\alpha-1)}{\alpha}} \epsilon^2}{2^{\frac{\alpha+2}{\alpha}} \sum_{i=1}^m (h(x_{m+i}) - h(x_i))^2}\right) \\
&\leq \exp\left(\frac{-[\sum_{i=1}^m (h(x_{m+i}) + h(x_i))]^{\frac{2}{\alpha}} m^{\frac{2(\alpha-1)}{\alpha}} \epsilon^2}{2^{\frac{\alpha+2}{\alpha}} \sum_{i=1}^m (h(x_{m+i}) - h(x_i))^2}\right).
\end{aligned}$$

Since the variables $h(x_i)$, $i \in [1, 2m]$, take values in $\{0, 1\}$, we can write

$$\begin{aligned}
\sum_{i=1}^m (h(x_{m+i}) - h(x_i))^2 &= \sum_{i=1}^m h(x_{m+i}) + h(x_i) - 2h(x_{m+i})h(x_i) \\
&\leq \sum_{i=1}^m h(x_{m+i}) + h(x_i) \leq \left[\sum_{i=1}^m h(x_{m+i}) + h(x_i)\right]^{\frac{2}{\alpha}},
\end{aligned}$$

where the last inequality holds since $\alpha \leq 2$ and the sum is either zero or greater than or equal to one. In view of this identity, we can write

$$\Pr_{\boldsymbol{\sigma}}\left[\frac{\frac{1}{m}\sum_{i=1}^m \sigma_i(h(x_{m+i}) - h(x_i))}{\sqrt[\alpha]{\frac{1}{2m}[\sum_{i=1}^m (h(x_{m+i}) + h(x_i))]}} > \epsilon \,\middle|\, x_1^{2m}\right] \leq \exp\left(\frac{-m^{\frac{2(\alpha-1)}{\alpha}} \epsilon^2}{2^{\frac{\alpha+2}{\alpha}}}\right).$$

We note now that the supremum over $h \in H$ in the left-hand side expression in the statement of our theorem need not be over all hypothesis in H: without changing its value, we can replace H with a smaller hypothesis set where only one hypothesis remains for each unique binary vector $(h(x_1), h(x_2), \ldots, h(x_{2m}))$. The

number of such hypotheses is $\mathbb{S}_H(x_1^{2m})$, thus, by the union bound, the following holds:

$$\Pr_{\boldsymbol{\sigma}}\left[\sup_{h\in H}\frac{\sum_{i=1}^m \sigma_i(h(x_{m+i})-h(x_i))}{\sqrt[\alpha]{\frac{1}{2}[\sum_{i=1}^m (h(x_{m+i})+h(x_i))]}} > \epsilon \,\middle|\, x_1^{2m}\right] \leq \mathbb{S}_H(x_1^{2m})\exp\left(\frac{-m^{\frac{2(\alpha-1)}{\alpha}}\epsilon^2}{2^{\frac{\alpha+2}{\alpha}}}\right).$$

The result follows by taking expectations with respect to x_1^{2m} and applying Lemma 2 and Lemma 3 respectively.

Corollary 1 *Let $1 < \alpha \leq 2$ and let H be a hypothesis set of functions mapping $\mathcal{X}$ to $\{0,1\}$. Then, for any $\delta > 0$, each of the following two inequalities holds with probability at least $1-\delta$:*

$$R(h) - \widehat{R}_S(h) \leq 2^{\frac{\alpha+2}{2\alpha}}\sqrt[\alpha]{R(h)}\sqrt{\frac{\log \mathrm{E}[\mathbb{S}_H(x_1^{2m})] + \log\frac{4}{\delta}}{m^{\frac{2(\alpha-1)}{\alpha}}}}$$

$$\widehat{R}_S(h) - R(h) \leq 2^{\frac{\alpha+2}{2\alpha}}\sqrt[\alpha]{\widehat{R}_S(h)}\sqrt{\frac{\log \mathrm{E}[\mathbb{S}_H(x_1^{2m})] + \log\frac{4}{\delta}}{m^{\frac{2(\alpha-1)}{\alpha}}}}.$$

Proof The result follows directly from Theorem 1 by setting δ to match the upper bounds and taking the limit $\tau \to 0$.

For $\alpha = 2$, the inequalities become

$$R(h) - \widehat{R}_S(h) \leq 2\sqrt{R(h)\frac{\log \mathrm{E}[\mathbb{S}_H(x_1^{2m})] + \log\frac{4}{\delta}}{m}} \tag{13}$$

$$\widehat{R}_S(h) - R(h) \leq 2\sqrt{\widehat{R}_S(h)\frac{\log \mathrm{E}[\mathbb{S}_H(x_1^{2m})] + \log\frac{4}{\delta}}{m}}, \tag{14}$$

with the familiar dependency $O\left(\sqrt{\frac{\log(m/d)}{m/d}}\right)$. The advantage of these relative deviations is clear. For small values of $R(h)$ (or $\widehat{R}_S(h)$) these inequalities provide tighter guarantees than standard generalization bounds. Solving the corresponding second-degree inequalities in $\sqrt{R(h)}$ or $\sqrt{\widehat{R}_S(h)}$ leads to the following results.

Corollary 2 *Let H be a hypothesis set of functions mapping $\mathcal{X}$ to $\{0,1\}$. Then, for any $\delta > 0$, each of the following two inequalities holds with probability at least $1-\delta$:*

$$R(h) \leq \widehat{R}_S(h) + 2\sqrt{\widehat{R}_S(h)\frac{\log \mathrm{E}[\mathbb{S}_H(x_1^{2m})] + \log\frac{4}{\delta}}{m}} + 4\frac{\log \mathrm{E}[\mathbb{S}_H(x_1^{2m})] + \log\frac{4}{\delta}}{m}$$

$$\widehat{R}_S(h) \leq R(h) + 2\sqrt{R(h)\frac{\log \mathrm{E}[\mathbb{S}_H(x_1^{2m})] + \log\frac{4}{\delta}}{m}} + 4\frac{\log \mathrm{E}[\mathbb{S}_H(x_1^{2m})] + \log\frac{4}{\delta}}{m}.$$

Proof The second-degree inequality corresponding to (13) can be written as

$$\sqrt{R(h)}^2 - 2\sqrt{R(h)}u - \widehat{R}_S(h) \leq 0,$$

with $u = \sqrt{\frac{\log \mathrm{E}[\mathbb{S}_H(x_1^{2m})] + \log \frac{4}{\delta}}{m}}$, and implies $\sqrt{R(h)} \leq u + \sqrt{u^2 + \widehat{R}_S(h)}$. Squaring both sides gives:

$$\begin{aligned} R(h) \leq \left[u + \sqrt{u^2 + \widehat{R}_S(h)}\right]^2 &= u^2 + 2u\sqrt{u^2 + \widehat{R}_S(h)} + u^2 + \widehat{R}_S(h) \\ &\leq u^2 + 2u\left(\sqrt{u^2} + \sqrt{\widehat{R}_S(h)}\right) + u^2 + \widehat{R}_S(h) \\ &= 4u^2 + 2u\sqrt{\widehat{R}_S(h)} + \widehat{R}_S(h). \end{aligned}$$

The second inequality can be proven in the same way from (14).

The learning bounds of the corollary make clear the presence of two terms: a term in $O(1/m)$ and a term in $O(1/\sqrt{m})$ which admits as a factor $\widehat{R}_S(h)$ or $R(h)$ and which for small values of these terms can be more favorable than standard bounds. Theorem 1 can also be used to prove the following relative deviation bounds.

The following theorem and its proof assuming the result of Theorem 1 were given by [2].

Theorem 2 *For all $0 < \epsilon < 1$, $\nu > 0$, the following inequalities hold:*

$$\begin{aligned} \Pr_{S\sim D^m}\left[\sup_{h\in H} \frac{R(h) - \widehat{R}_S(h)}{R(h) + \widehat{R}_S(h) + \nu} > \epsilon\right] &\leq 4\,\mathrm{E}[\mathbb{S}_H(x_1^{2m})] \exp\left(\frac{-m\nu\epsilon^2}{2(1-\epsilon^2)}\right) \\ \Pr_{S\sim D^m}\left[\sup_{h\in H} \frac{\widehat{R}_S(h) - R(h)}{R(h) + \widehat{R}_S(h) + \nu} > \epsilon\right] &\leq 4\,\mathrm{E}[\mathbb{S}_H(x_1^{2m})] \exp\left(\frac{-m\nu\epsilon^2}{2(1-\epsilon^2)}\right). \end{aligned}$$

Proof We prove the first statement, the proof of the second statement is identical modulo the permutation of the roles of $R(h)$ and $\widehat{R}_S(h)$. To do so, it suffices to determine $\epsilon' > 0$ such that

$$\Pr_{S\sim D^m}\left[\sup_{h\in H} \frac{R(h) - \widehat{R}_S(h)}{R(h) + \widehat{R}_S(h) + \nu} > \epsilon\right] \leq \Pr_{S\sim D^m}\left[\sup_{h\in H} \frac{R(h) - \widehat{R}_S(h)}{\sqrt[\alpha]{R(h) + \tau}} > \epsilon'\right],$$

since we can then apply theorem 1 with $\alpha = 2$ to bound the right-hand side and take the limit as $\tau \to 0$ to eliminate the τ-dependence. To find such a choice of ϵ', we begin by observing that for any $h \in H$,

$$\frac{R(h) - \widehat{R}_S(h)}{R(h) + \widehat{R}_S(h) + \nu} \leq \epsilon \Leftrightarrow R(h) \leq \frac{1+\epsilon}{1-\epsilon}\widehat{R}_S(h) + \frac{\epsilon}{1-\epsilon}\nu. \tag{15}$$

Assume now that $\frac{R(h) - \widehat{R}_S(h)}{\sqrt{R(h)+\tau}} \leq \epsilon'$ for some $\epsilon' > 0$, which is equivalent to $R(h) \leq \widehat{R}_S(h) + \epsilon'\sqrt{R(h)+\tau}$. We will prove that this implies (15).

To start, we see that for all $\theta > 0$

$$\epsilon'\sqrt{R(h)+\tau} = \sqrt{\frac{\epsilon'^2}{\theta}\theta(R(h)+\tau)} \leq \frac{1}{2\theta}\epsilon'^2 + \frac{\theta}{2}(R(h)+\tau)$$

by the inequality of arithmetic and geometric means. Hence:

$$R(h) \leq \widehat{R}_S(h) + \frac{1}{2\theta}\epsilon'^2 + \frac{\theta}{2}(R(h)+\tau)$$

and therefore, by rearranging to isolate $R(h)$ on one side,

$$R(h) \leq \frac{1}{1-\theta/2}\widehat{R}_S(h) + \frac{1}{1-\theta/2}\left(\frac{1}{2\theta}\epsilon'^2 + \frac{\theta}{2}\tau\right).$$

We now only need to choose ϵ', τ and θ such that

$$\frac{1}{1-\theta/2} \leq \frac{1+\epsilon}{1-\epsilon}$$

$$\frac{1}{1-\theta/2}\left(\frac{1}{2\theta}\epsilon'^2 + \frac{\theta}{2}\tau\right) \leq \frac{\epsilon}{1-\epsilon}\nu.$$

It is sufficient to choose

$$\theta = \frac{4\epsilon}{1+\epsilon}$$

$$\epsilon'^2 \leq \frac{8\epsilon^2}{(1+\epsilon)^2}(\nu - 2\tau)$$

which establishes that

$$R(h) \leq \frac{1+\epsilon}{1-\epsilon}\widehat{R}_S(h) + \frac{\epsilon}{1-\epsilon}\nu.$$

With these choices, the following inequality holds for all $h \in H$:

$$\frac{R(h) - \widehat{R}_S(h)}{\sqrt{R(h)+\tau}} \leq \epsilon' \Rightarrow \frac{R(h) - \widehat{R}_S(h)}{R(h) + \widehat{R}_S(h) + \nu} \leq \epsilon,$$

concluding the proof.

The following corollary was given by [2].

Corollary 3 *For all $\epsilon > 0$, $v > 0$, the following inequality holds:*

$$\Pr_{S\sim D^m}\left[\sup_{h\in H} R(h) - (1+v)\widehat{R}_S(h) > \epsilon\right] \leq 4\,\mathrm{E}[\mathbb{S}_H(x_1^{2m})]\exp\left(\frac{-mv\epsilon}{4(1+v)}\right).$$

Proof Observe that

$$\begin{aligned}\frac{R(h) - \widehat{R}_S(h)}{R(h) + \widehat{R}_S(h) + \nu} > \epsilon &\Leftrightarrow R(h) - \widehat{R}_S(h) > (R(h) + \widehat{R}_S(h) + \nu)\epsilon \\ &\Leftrightarrow R(h) > \frac{1+\epsilon}{1-\epsilon}\widehat{R}_S(h) + \frac{\epsilon\nu}{1-\epsilon}.\end{aligned}$$

To derive the statement of the corollary from that of Theorem 2, we identify $\frac{1+\epsilon}{1-\epsilon}$ with $1+v$, which gives $\epsilon(2+v) = v$, that is we choose $\epsilon = \frac{v}{2+v}$, and similarly identify $\frac{\epsilon\nu}{1-\epsilon}$ with ϵ', that is $\epsilon' = \frac{\frac{v}{2+v}}{\frac{2}{2+v}}\nu = \frac{v}{2}\nu$, thus we choose $\nu = \frac{2}{v}\epsilon'$. With these choices of ϵ' and ν, the coefficient in the exponential appearing in the bounds of Theorem 2 can be rewritten as follows: $\frac{v\epsilon^2}{2(1-\epsilon^2)} = \frac{2\epsilon'}{2v}\frac{\frac{v^2}{(2+v)^2}}{\frac{4v+4}{(2+v)^2}} = \frac{\epsilon'}{v}\frac{v^2}{4(v+1)} = \frac{\epsilon' v}{4(v+1)}$, which concludes the proof.

The result of Corollary 3 is remarkable since it shows that a fast convergence rate of $O(1/m)$ can be achieved provided that we settle for a slightly larger value than the empirical error, one differing by a fixed factor $(1 + v)$. The following is an immediate corollary when $\widehat{R}_S(h) = 0$, where we take $v \to \infty$.

Corollary 4 *For all $\epsilon > 0$, the following inequality holds:*

$$\Pr_{S\sim D^m}\left[\exists h \in H\colon R(h) > \epsilon \wedge \widehat{R}_S(h) = 0\right] \leq 4\,\mathrm{E}[\mathbb{S}_H(x_1^{2m})]\exp\left(\frac{-m\epsilon}{4}\right).$$

This is the familiar fast rate convergence result for separable cases.

4 Generalization bounds for unbounded losses

In this section we will make use of the relative deviation bounds given in the previous section to prove generalization bounds for unbounded loss functions under the assumption that the moment of order α of the loss is bounded. We will start with the case $1 < \alpha \leq 2$ and then move on to considering the case when $\alpha > 2$. As already indicated earlier, the one-sided version of the results presented in this section were given by [47] with slightly different constants, but the proofs do not seem to be correct or complete. The second statements in all these results (other side of the inequality) are new. Our proofs for both sets of results are new.

4.1 Bounded moment with $1 < \alpha \leq 2$

Our first theorem reduces the problem of deriving a relative deviation bound for an unbounded loss function with $\mathcal{L}_\alpha(h) = \mathrm{E}_{z\sim D}[L(h,z)^\alpha] < +\infty$ for all $h \in H$, to that of relative deviation bound for binary classification. To simplify the presentation of the results, in what follows we will use the shorthand $\Pr[L(h,z) > t]$ instead of $\Pr_{z\sim D}[L(h,z) > t]$, and similarly $\widehat{\Pr}[L(h,z) > t]$ instead of $\Pr_{z\sim \widehat{D}}[L(h,z) > t]$.

Theorem 3 *Let $1 < \alpha \leq 2$, $0 < \epsilon \leq 1$, and $0 < \tau^{\frac{\alpha-1}{\alpha}} < \epsilon^{\frac{\alpha}{\alpha-1}}$. For any loss function L (not necessarily bounded) and hypothesis set H such that $\mathcal{L}_\alpha(h) < +\infty$ for all $h \in H$, the following two inequalities hold:*

$$\Pr\left[\sup_{h\in H}\frac{\mathcal{L}(h) - \widehat{\mathcal{L}}_S(h)}{\sqrt[\alpha]{\mathcal{L}_\alpha(h) + \tau}} > \Gamma(\alpha,\epsilon)\,\epsilon\right] \leq \Pr\left[\sup_{h\in H, t\in\mathbb{R}}\frac{\Pr[L(h,z) > t] - \widehat{\Pr}[L(h,z) > t]}{\sqrt[\alpha]{\Pr[L(h,z) > t] + \tau}} > \epsilon\right]$$

$$\Pr\left[\sup_{h\in H}\frac{\mathcal{L}(h) - \widehat{\mathcal{L}}_S(h)}{\sqrt[\alpha]{\mathcal{L}_\alpha(h) + \tau}} > \Gamma(\alpha,\epsilon)\,\epsilon\right] \leq \Pr\left[\sup_{h\in H, t\in\mathbb{R}}\frac{\widehat{\Pr}[L(h,z) > t] - \Pr[L(h,z) > t]}{\sqrt[\alpha]{\widehat{\Pr}[L(h,z) > t] + \tau}} > \epsilon\right],$$

with $\Gamma(\alpha,\epsilon) = \frac{\alpha-1}{\alpha}(1+\tau)^{\frac{1}{\alpha}} + \frac{1}{\alpha}\left(\frac{\alpha}{\alpha-1}\right)^{\alpha-1}(1 + \left(\frac{\alpha-1}{\alpha}\right)^\alpha \tau^{\frac{1}{\alpha}})^{\frac{1}{\alpha}}\left[1 + \frac{\log(1/\epsilon)}{\left(\frac{\alpha}{\alpha-1}\right)^{\alpha-1}}\right]^{\frac{\alpha-1}{\alpha}}$
.

The proof of Theorem 3 is given in the Appendix.

The next corollary follows immediately by upper bounding the right-hand side of the learning bounds of theorem 3 using theorem 1. It provides learning bounds for unbounded loss functions in terms of the growth functions in the case $1 < \alpha \leq 2$.

Corollary 5 *Let $\epsilon < 1$, $1 < \alpha \leq 2$, and $0 < \tau^{\frac{\alpha-1}{\alpha}} < \epsilon^{\frac{\alpha}{\alpha-1}}$. For any loss function L (not necessarily bounded) and hypothesis set H such that $\mathcal{L}_\alpha(h) < +\infty$ for all $h \in H$, the following inequalities hold:*

$$\Pr\left[\sup_{h\in H} \frac{\mathcal{L}(h) - \widehat{\mathcal{L}}(h)}{\sqrt[\alpha]{\mathcal{L}_\alpha(h) + \tau}} > \Gamma(\alpha,\epsilon)\epsilon\right] \leq 4\ \mathrm{E}[\mathbb{S}_Q(z_1^{2m})] \exp\left(\frac{-m^{\frac{2(\alpha-1)}{\alpha}}\epsilon^2}{2^{\frac{\alpha+2}{\alpha}}}\right)$$

$$\Pr\left[\sup_{h\in H} \frac{\widehat{\mathcal{L}}(h) - \mathcal{L}(h)}{\sqrt[\alpha]{\widehat{\mathcal{L}}_\alpha(h) + \tau}} > \Gamma(\alpha,\epsilon)\epsilon\right] \leq 4\ \mathrm{E}[\mathbb{S}_Q(z_1^{2m})] \exp\left(\frac{-m^{\frac{2(\alpha-1)}{\alpha}}\epsilon^2}{2^{\frac{\alpha+2}{\alpha}}}\right),$$

where Q is the set of functions $Q = \{z \mapsto 1_{L(h,z)>t} \mid h \in H, t \in \mathbb{R}\}$, and $\Gamma(\alpha,\epsilon) = \frac{\alpha-1}{\alpha}(1+\tau)^{\frac{1}{\alpha}} + \frac{1}{\alpha}\left(\frac{\alpha}{\alpha-1}\right)^{\alpha-1}(1 + \left(\frac{\alpha-1}{\alpha}\right)^{\alpha}\tau^{\frac{1}{\alpha}})^{\frac{1}{\alpha}}\left[1 + \frac{\log(1/\epsilon)}{\left(\frac{\alpha}{\alpha-1}\right)^{\alpha-1}}\right]^{\frac{\alpha-1}{\alpha}}$.

For comparison with other results in situations where the 2nd moment of the loss is finite, the following corollary gives the explicit $\alpha = 2$ case of the above corollary. Note that in the $\alpha = 2$ case the expression $\sqrt[\alpha]{\mathcal{L}_\alpha(h) + \tau}$ relates directly to the standard deviation, and also coincides with the corresponding expression used by Vapnik.

Corollary 6 *Let $\epsilon < 1$ and $0 < \tau < \epsilon^4$. For any loss function L (not necessarily bounded) and hypothesis set H such that $\mathcal{L}_2(h) < +\infty$ for all $h \in H$, the following inequalities hold:*

$$\Pr\left[\sup_{h\in H} \frac{\mathcal{L}(h) - \widehat{\mathcal{L}}(h)}{\sqrt{\mathcal{L}_2(h) + \tau}} > \Gamma(2,\epsilon)\epsilon\right] \leq 4\ \mathrm{E}[\mathbb{S}_Q(z_1^{2m})] \exp\left(\frac{-m\epsilon^2}{4}\right)$$

$$\Pr\left[\sup_{h\in H} \frac{\widehat{\mathcal{L}}(h) - \mathcal{L}(h)}{\sqrt{\widehat{\mathcal{L}}_2(h) + \tau}} > \Gamma(2,\epsilon)\epsilon\right] \leq 4\ \mathrm{E}[\mathbb{S}_Q(z_1^{2m})] \exp\left(\frac{-m\epsilon^2}{4}\right),$$

with $\Gamma(2,\epsilon) = \left(\frac{\sqrt{1+\tau}}{2} + \sqrt{1 + \frac{1}{4}\sqrt{\tau}}\sqrt{1 + \frac{1}{2}\log\frac{1}{\epsilon}}\right)$ and Q the set of functions $Q = \{z \mapsto 1_{L(h,z)>t} \mid h \in H, t \in \mathbb{R}\}$.

Corollary 7 *Let L be a loss function (not necessarily bounded) and H a hypothesis set such that $\mathcal{L}_2(h) < +\infty$ for all $h \in H$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, each of the following inequalities holds for all $h \in H$:*

$$\mathcal{L}(h) \leq \widehat{\mathcal{L}}_S(h) + 2\sqrt{\mathcal{L}_2(h)}\sqrt{\frac{2\log \mathrm{E}[\mathbb{S}_Q(z_1^{2m})] + \log\frac{1}{\delta}}{m}}\,\Gamma_0\left(2, 2\sqrt{\frac{2\log \mathrm{E}[\mathbb{S}_Q(z_1^{2m})] + \log\frac{1}{\delta}}{m}}\right)$$

$$\widehat{\mathcal{L}}_S(h) \leq \mathcal{L}(h) + 2\sqrt{\widehat{\mathcal{L}}_2(h)}\sqrt{\frac{2\log \mathrm{E}[\mathbb{S}_Q(z_1^{2m})] + \log\frac{1}{\delta}}{m}}\,\Gamma_0\left(2, 2\sqrt{\frac{2\log \mathrm{E}[\mathbb{S}_Q(z_1^{2m})] + \log\frac{1}{\delta}}{m}}\right),$$

where Q is the set of functions $Q = \{z \mapsto 1_{L(h,z)>t} \mid h \in H, t \in \mathbb{R}\}$ and $\Gamma_0(2,\epsilon) = \frac{1}{2} + \sqrt{1 + \frac{1}{2}\log\frac{1}{\epsilon}}$.

Proof For any $\epsilon > 0$, let $f(\epsilon) = \Gamma_0(2,\epsilon)\epsilon$ (which is invertible and approximately linear for ϵ slightly larger than 0). Then, by Corollary 6,

$$\Pr\left[\sup_{h\in H} \frac{\mathcal{L}(h) - \widehat{\mathcal{L}}(h)}{\sqrt{\mathcal{L}_2(h) + \tau}} > \epsilon\right] \leq 4\ \mathrm{E}[\mathbb{S}_Q(z_1^{2m})] \exp\left(\frac{-m[f^{-1}(\epsilon)]^2}{4}\right).$$

Setting the right-hand side to ϵ and using inversion yields immediately the first inequality. The second inequality is proven in the same way.

Observe that, modulo the factors in Γ_0, the bounds of the corollary admit the standard $(1/\sqrt{m})$ dependency and that the factors in Γ_0 are only logarithmic in m.

4.2 Bounded moment with $\alpha > 2$

This section gives two-sided generalization bounds for unbounded losses with finite moments of order α, with $\alpha > 2$. As for the case $1 < \alpha < 2$, the one-sided version of our bounds coincides with that of [47,48] modulo a constant factor, but, here again, the proofs given in both books seem to be incorrect.

Proposition 1 *Let $\alpha > 2$. For any loss function L (not necessarily bounded) and hypothesis set H such that $0 < \mathcal{L}_\alpha(h) < +\infty$ for all $h \in H$, the following two inequalities hold:*

$$\int_0^{+\infty} \sqrt{\Pr[L(h,z) > t]}dt \leq \Psi(\alpha) \sqrt[\alpha]{\mathcal{L}_\alpha(h)} \quad \textit{and} \quad \int_0^{+\infty} \sqrt{\widehat{\Pr}[L(h,z) > t]}dt \leq \Psi(\alpha) \sqrt[\alpha]{\widehat{\mathcal{L}}_\alpha(h)},$$

where $\Psi(\alpha) = \left(\frac{1}{2}\right)^{\frac{2}{\alpha}} \left(\frac{\alpha}{\alpha-2}\right)^{\frac{\alpha-1}{\alpha}}$.

Note that the difference between the two inequalities above is that the right inequality involves the empirical distribution (i.e. drawing a data point uniformly at random from our fixed data set), whereas the left inequality involves the true, underlying distribution from which the data points were drawn. The proof of Proposition 1 is given in the Appendix.

Theorem 4 *Let $\alpha > 2$, $0 < \epsilon \leq 1$, and $0 < \tau \leq \epsilon^2$. Then, for any loss function L (not necessarily bounded) and hypothesis set H such that $\mathcal{L}_\alpha(h) < +\infty$ and $\widehat{\mathcal{L}}_\alpha(h) < +\infty$ for all $h \in H$, the following two inequalities hold:*

$$\Pr\left[\sup_{h\in H} \frac{\mathcal{L}(h) - \widehat{\mathcal{L}}(h)}{\sqrt[\alpha]{\mathcal{L}_\alpha(h) + \tau}} > \Lambda(\alpha)\epsilon\right] \leq \Pr\left[\sup_{h\in H, t\in\mathbb{R}} \frac{\Pr[L(h,z) > t] - \widehat{\Pr}[L(h,z) > t]}{\sqrt{\Pr[L(h,z) > t] + \tau}} > \epsilon\right]$$

$$\Pr\left[\sup_{h\in H} \frac{\widehat{\mathcal{L}}(h) - \mathcal{L}(h)}{\sqrt[\alpha]{\widehat{\mathcal{L}}_\alpha(h) + \tau}} > \Lambda(\alpha)\epsilon\right] \leq \Pr\left[\sup_{h\in H, t\in\mathbb{R}} \frac{\widehat{\Pr}[L(h,z) > t] - \Pr[L(h,z) > t]}{\sqrt{\widehat{\Pr}[L(h,z) > t] + \tau}} > \epsilon\right],$$

where $\Lambda(\alpha) = \left(\frac{1}{2}\right)^{\frac{2}{\alpha}} \left(\frac{\alpha}{\alpha-2}\right)^{\frac{\alpha-1}{\alpha}} + \frac{\alpha}{\alpha-1}\tau^{\frac{\alpha-2}{2\alpha}}$.

Proof We prove the first statement since the second one can be proven in a very similar way. Assume that $\sup_{h,t} \frac{\Pr[L(h,z)>t]-\widehat{\Pr}[L(h,z)>t]}{\sqrt{\Pr[L(h,z)>t]+\tau}} \leq \epsilon$. Fix $h \in H$, let $J = \int_0^{+\infty} \sqrt{\Pr\left[L(h,z) > t\right]}\, dt$ and $\nu = \mathcal{L}_\alpha(h)$. By Markov's inequality, for any $t > 0$, $\Pr[L(h,z) > t] = \Pr[L^\alpha(h,z) > t^\alpha] \leq \frac{\mathcal{L}_\alpha(h)}{t^\alpha} = \frac{\nu}{t^\alpha}$. Using this inequality, for any $t_0 > 0$, we can write

$$\begin{aligned}
\mathcal{L}(h) - \widehat{\mathcal{L}}(h) &= \int_0^{+\infty} (\Pr[L(h,z) > t] - \widehat{\Pr}[L(h,z) > t])\, dt \\
&= \int_0^{t_0} (\Pr[L(h,z) > t] - \widehat{\Pr}[L(h,z) > t])\, dt \\
&\quad + \int_{t_0}^{+\infty} (\Pr[L(h,z) > t] - \widehat{\Pr}[L(h,z) > t])\, dt \\
&\leq \epsilon \int_0^{t_0} \sqrt{\Pr[L(h,z) > t] + \tau}\, dt + \int_{t_0}^{+\infty} \Pr[L(h,z) > t]\, dt \\
&\leq \epsilon \int_0^{t_0} (\sqrt{\Pr[L(h,z) > t]} + \sqrt{\tau})\, dt + \int_{t_0}^{+\infty} \frac{\nu}{t^\alpha}\, dt \\
&\leq \epsilon J + \epsilon\sqrt{\tau} t_0 + \frac{\nu}{(\alpha-1)t_0^{\alpha-1}}.
\end{aligned}$$

Choosing t_0 to minimize the right-hand side yields $t_0 = \left(\frac{\nu}{\epsilon\sqrt{\tau}}\right)^{\frac{1}{\alpha}}$ and gives

$$\mathcal{L}(h) - \widehat{\mathcal{L}}(h) \leq \epsilon J + \frac{\alpha}{\alpha-1} \nu^{\frac{1}{\alpha}} (\epsilon\sqrt{\tau})^{\frac{\alpha-1}{\alpha}}.$$

Since $\tau \leq \epsilon^2$, $(\epsilon\sqrt{\tau})^{\frac{\alpha-1}{\alpha}} = [\epsilon \tau^{\frac{1}{2(\alpha-1)}} \tau^{\frac{\alpha-2}{2(\alpha-1)}}]^{\frac{\alpha-1}{\alpha}} \leq [\epsilon \epsilon^{\frac{1}{(\alpha-1)}} \tau^{\frac{\alpha-2}{2(\alpha-1)}}]^{\frac{\alpha-1}{\alpha}} = \epsilon \tau^{\frac{\alpha-2}{2\alpha}}$. Thus, by Proposition 1, the following holds:

$$\frac{\mathcal{L}(h) - \widehat{\mathcal{L}}(h)}{\sqrt[\alpha]{\mathcal{L}_\alpha(h) + \tau}} \leq \epsilon \Psi(\alpha) \frac{\nu^{\frac{1}{\alpha}}}{(\nu+\tau)^{\frac{1}{\alpha}}} + \frac{\alpha}{\alpha-1} \epsilon \tau^{\frac{\alpha-2}{2\alpha}} \frac{\nu^{\frac{1}{\alpha}}}{(\nu+\tau)^{\frac{1}{\alpha}}} \leq \epsilon \Psi(\alpha) + \frac{\alpha}{\alpha-1} \epsilon \tau^{\frac{\alpha-2}{2\alpha}},$$

which concludes the proof.

Combining Theorem 4 with Theorem 1 leads immediately to the following two results.

Corollary 8 *Let $\alpha > 2$, $0 < \epsilon \leq 1$, and $0 < \tau \leq \epsilon^2$. Then, for any loss function L (not necessarily bounded) and hypothesis set H such that $\mathcal{L}_\alpha(h) < +\infty$ and $\widehat{\mathcal{L}}_\alpha(h) < +\infty$ for all $h \in H$, the following two inequalities hold:*

$$\Pr\left[\sup_{h\in H} \frac{\mathcal{L}(h) - \widehat{\mathcal{L}}(h)}{\sqrt[\alpha]{\mathcal{L}_\alpha(h) + \tau}} > \Lambda(\alpha)\epsilon\right] \leq 4\, \mathrm{E}[\mathbb{S}_Q(z_1^{2m})] \exp\left(\frac{-m\epsilon^2}{4}\right)$$

$$\Pr\left[\sup_{h\in H} \frac{\widehat{\mathcal{L}}(h) - \mathcal{L}(h)}{\sqrt[\alpha]{\widehat{\mathcal{L}}_\alpha(h) + \tau}} > \Lambda(\alpha)\epsilon\right] \leq 4\, \mathrm{E}[\mathbb{S}_Q(z_1^{2m})] \exp\left(\frac{-m\epsilon^2}{4}\right),$$

where $\Lambda(\alpha) = \left(\frac{1}{2}\right)^{\frac{2}{\alpha}} \left(\frac{\alpha}{\alpha-2}\right)^{\frac{\alpha-1}{\alpha}} + \frac{\alpha}{\alpha-1} \tau^{\frac{\alpha-2}{2\alpha}}$ and where Q is the set of functions $Q = \{z \mapsto 1_{L(h,z)>t} \mid h \in H, t \in \mathbb{R}\}$.

In the following result, $\mathrm{Pdim}(G)$ denotes the pseudo-dimension of a family of real-valued functions G [40,41,47], which coincides with the VC-dimension of the corresponding thresholded functions:

$$\mathrm{Pdim}(G) = \mathrm{VCdim}\left(\left\{(x,t) \mapsto 1_{(g(x)-t)>0} \colon g \in G\right\}\right) . \tag{16}$$

Corollary 9 *Let $\alpha > 2$, $0 < \epsilon \leq 1$. Let L be a loss function (not necessarily bounded) and H a hypothesis set such that $\mathcal{L}_\alpha(h) < +\infty$ for all $h \in H$, and $d = \mathrm{Pdim}(\{z \mapsto L(h,z) \mid h \in H\}) < +\infty$. Then, for any $\delta > 0$, with probability at least $1-\delta$, each of the following inequalities holds for all $h \in H$:*

$$\mathcal{L}(h) \leq \widehat{\mathcal{L}}(h) + 2\Lambda(\alpha) \sqrt[\alpha]{\mathcal{L}_\alpha(h)} \sqrt{\frac{d \log \frac{2em}{d} + \log \frac{4}{\delta}}{m}}$$

$$\widehat{\mathcal{L}}(h) \leq \mathcal{L}(h) + 2\Lambda(\alpha) \sqrt[\alpha]{\widehat{\mathcal{L}}_\alpha(h)} \sqrt{\frac{d \log \frac{2em}{d} + \log \frac{4}{\delta}}{m}}$$

where $\Lambda(\alpha) = \left(\frac{1}{2}\right)^{\frac{2}{\alpha}} \left(\frac{\alpha}{\alpha-2}\right)^{\frac{\alpha-1}{\alpha}}$.

5 Applications

We now apply the above results to a couple sample applications of particular interest.

5.1 Importance weighting

As an example application of the results above, consider the case of deriving learning bounds for importance weighting. The importance weighting technique involves adjusting the cost of each error made on training data points so as to correct for differences between the distributions from which training points and testing points are drawn. It has applications in areas such as sample bias correction and adaptation [43,16].

Let $Q(z)$ be the probability distribution from which training samples are drawn and $P(z)$ be the probability distribution from which testing samples are drawn. Then, by weighting the error made on each point, z, by $w(z) = P(z)/Q(z)$ in our empirical loss function, we achieve an unbiased estimator of the generalization error on the distribution $P(z)$, even though the training data is drawn according to $Q(z)$. Hence, we replace our loss function $L(h,z)$ with the weighted loss function $w(z)L(h,z)$.

Producing deviation bounds in the case of importance weighting is complicated by the fact that the weights $w(z)$ may be unbounded (if the distribution $Q(z)$ approaches 0 faster than $P(z)$ in some regions). Assume that our loss function (prior to weighting) is restricted to the range $[0,1]$, but our entire loss $w(z)L(h,z)$ is still potentially unbounded due to the unbounded weights. Then, assuming $d_\alpha(P||Q)$ is bounded, we note that:

$$\mathcal{L}_\alpha(h) = \mathop{\mathrm{E}}_{z\sim D}[w(z)^\alpha L^\alpha(h,z)] \leq \mathop{\mathrm{E}}_{z\sim Q}[(P(z)/Q(z))^\alpha] = d_\alpha(P||Q)^{\alpha-1}$$

where $d_\alpha(P||Q)$ is the exponential base 2 of the Rényi divergence.

Corollary 9 then immediately tells us that when $\mathcal{L}_\alpha(h) < +\infty$, we have for $\alpha > 2$, $0 < \epsilon \leq 1$, and for all $h \in H$ and any $\delta > 0$, with probability at least $1 - \delta$ the following inequality holds for all $h \in H$:

$$\mathcal{L}(h) \leq \widehat{\mathcal{L}}(h) + 2\Lambda(\alpha)\, d_\alpha(P||Q)^{\frac{\alpha-1}{\alpha}} \sqrt{\frac{d \log \frac{2em}{d} + \log \frac{4}{\delta}}{m}}.$$

Note that when α is slightly bigger than 2 this result is extremely similar to Theorem 3 of [13], but with an improved rate of convergence with respect to the sample size m.

What's more, Corollary 7 tells us that for the $\alpha = 2$ case, with probability at least $1 - \delta$ the following inequality holds for all $h \in H$:

$$\mathcal{L}(h) \leq \widehat{\mathcal{L}}_S(h) + 2\sqrt{d_2(P||Q)} \sqrt{\frac{2 \log \mathrm{E}[\mathbb{S}_Q(z_1^{2m})] + \log \frac{1}{\delta}}{m}} \Gamma_0 \left(2, 2\sqrt{\frac{2 \log \mathrm{E}[\mathbb{S}_Q(z_1^{2m})] + \log \frac{1}{\delta}}{m}} \right)$$

where Q is the set of functions $Q = \{z \mapsto 1_{w(z)L(h,z)>t} \mid h \in H, t \in \mathbb{R}\}$ and $\Gamma_0(2, \epsilon) = \frac{1}{2} + \sqrt{1 + \frac{1}{2} \log \frac{1}{\epsilon}}$. These bounds are again better than those in [13].

Note further that in this $\alpha = 2$ case, we have:

$$\mathcal{L}_2(h) = \mathop{\mathrm{E}}_{z \sim D}[w(z)^2 L^2(h, z)] \leq \mathop{\mathrm{E}}_{z \sim D}[w(z)^2] = \mathop{\mathrm{E}}_{z \sim Q}[(P(z)/Q(z))^2] = d_2(P||Q)$$

so $d_2(P||Q)$ simply coincides with the second moment of w.

5.2 The Tsybakov noise condition and excess risk bounds

The Tsybakov noise condition [34,44,50] provides a means to study excess risk bounds. Corollary 1, if it is extended to cover the case of hypotheses with absolute values in $\{0, 1\}$, can be used along with the Tsybakov noise condition to achieve fast rates for excess risk in a classification setting, by comparing the Bayes classifier $h^*(x)$ (which by definition minimizes the classification error) to the empirical risk minimizer $\widehat{h}(x)$ (i.e. the hypothesis with the lowest error on the training data).

We proceed by defining

$$R(h) = \mathrm{E}[1_{h(x) \neq y}]$$

$$\widehat{R}_S(h) = \mathrm{E}[1_{h(x) \neq y}].$$

and define our loss function as

$$L(h(x), y) = 1_{h(x) \neq y} - 1_{h^*(x) \neq y}$$

which is 0 when $h(x)$ and $h^*(x)$ agree, 1 when $h(x)$ makes a classification mistake that $h^*(x)$ doesn't, and -1 when $h^*(x)$ makes a classification mistake that $h(x)$ doesn't. Now we apply an extended form of Corollary 1 to $L(h(x), y)$ for any fixed

α satisfying $1 < \alpha \leq 2$. Then, for any $\delta > 0$, the following inequality holds with probability at least $1 - \delta$:

$$R(h) - R(h^*) \leq \widehat{R}_S(h) - \widehat{R}_S(h^*) + 2^{\frac{\alpha+2}{2\alpha}} \sqrt[\alpha]{\mathrm{E}[|L(h(x), y)|]} \sqrt{\frac{\log \mathrm{E}[\mathbb{S}_H(x_1^{2m})] + \log \frac{4}{\delta}}{m^{\frac{2(\alpha-1)}{\alpha}}}}.$$

Now, since the above holds for any h in our hypothesis set H, we can choose h to be $\widehat{h}$, the empirical risk minimizer. Furthermore, if we assume that $h^* \in H$ (a strong assumption, but standard in this context), we have:

$$\widehat{R}_S(\widehat{h}) - \widehat{R}_S(h^*) \leq 0$$

since by definition $\widehat{h}$ achieves the minimum classification error on S of any $h \in H$. Therefore:

$$R(\widehat{h}) - R(h^*) \leq 2^{\frac{\alpha+2}{2\alpha}} \sqrt[\alpha]{\mathrm{E}[|L(h(x), y)|]} \sqrt{\frac{\log \mathrm{E}[\mathbb{S}_H(x_1^{2m})] + \log \frac{4}{\delta}}{m^{\frac{2(\alpha-1)}{\alpha}}}}$$

$$= 2^{\frac{\alpha+2}{2\alpha}} \sqrt{\mathrm{E}[|L(h(x), y)|]^{\frac{2}{\alpha}} \frac{\log \mathrm{E}[\mathbb{S}_H(x_1^{2m})] + \log \frac{4}{\delta}}{m^{\frac{2(\alpha-1)}{\alpha}}}}.$$

By applying the Tsybakov noise assumption we have:

$$\mathrm{E}[|L(h(x), y)|] = \mathrm{E}[|1_{h(x)\neq y} - 1_{h^*(x)\neq y}|] = \mathrm{E}[(1_{h(x)\neq y} - 1_{h^*(x)\neq y})^2]$$

$$\leq \beta_0 (R(\widehat{h}) - R(h^*))^{\gamma}$$

for some $\gamma \in [0, 1]$ and some $\beta_0 > 0$. Therefore:

$$R(\widehat{h}) - R(h^*) \leq \sqrt{\beta (R(\widehat{h}) - R(h^*))^{\frac{2\gamma}{\alpha}} \frac{\log \mathrm{E}[\mathbb{S}_H(x_1^{2m})] + \log \frac{4}{\delta}}{m^{\frac{2(\alpha-1)}{\alpha}}}}.$$

for some $\beta > 0$ (dependent on α) and

$$R(\widehat{h}) - R(h^*) \leq \left(\beta \frac{\log \mathrm{E}[\mathbb{S}_H(x_1^{2m})] + \log \frac{4}{\delta}}{m^{\frac{2(\alpha-1)}{\alpha}}} \right)^{\frac{1}{2-2\frac{\gamma}{\alpha}}}.$$

Choosing $\alpha = 2$ we have

$$R(\widehat{h}) - R(h^*) \leq \left(\beta \frac{\log \mathrm{E}[\mathbb{S}_H(x_1^{2m})] + \log \frac{4}{\delta}}{m} \right)^{\frac{1}{2-\gamma}},$$

which proves a fast rate of convergence in $m^{\frac{-1}{2-\gamma}}$.

6 Conclusion

We presented a series of results for relative deviation bounds used to prove generalization bounds for unbounded loss functions. These learning bounds can be used in a variety of applications to deal with the more general unbounded case. The relative deviation bounds are of independent interest and can be further used for a sharper analysis of guarantees in binary classification and other tasks.

Acknowledgments

We thank the reviewers for several careful and very useful comments. This work was partly funded by NSF CCF-1535987 and NSF IIS-1618662.

References

1. M. Anthony and J. Shawe-Taylor. A result of Vapnik with applications. *Discrete Applied Mathematics*, 47:207 – 217, 1993.
2. Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.
3. Kazuoki Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal*, 19(3):357–367, 1967.
4. Peter L. Bartlett, Stéphane Boucheron, and Gábor Lugosi. Model selection and error estimation. *Machine Learning*, 48:85–113, September 2002.
5. Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Localized Rademacher complexities. In *COLT*, volume 2375, pages 79–97. Springer-Verlag, 2002.
6. Peter L. Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3, 2002.
7. S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira. Analysis of representations for domain adaptation. *NIPS*, 2007.
8. Bernard Bercu, Elisabeth Gassiat, Emmanuel Rio, et al. Concentration inequalities, large and moderate deviations for self-normalized empirical processes. *The Annals of Probability*, 30(4):1576–1604, 2002.
9. Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning for differing training and test distributions. In *ICML*, pages 81–88, 2007.
10. J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Wortman. Learning bounds for domain adaptation. *NIPS 2007*, 2008.
11. Stphane Boucheron, Olivier Bousquet, and Gbor Lugosi. Theory of classification: a survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323375, 2005.
12. Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. In *Advances in Neural Information Processing Systems (NIPS 2010)*, Vancouver, Canada, 2010. MIT Press.
13. Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. In *Advances in neural information processing systems*, pages 442–450, 2010.
14. Corinna Cortes and Mehryar Mohri. Domain adaptation and sample bias correction theory and algorithm for regression. *Theoretical Computer Science*, 9474, 2013.
15. Corinna Cortes, Mehryar Mohri, Michael Riley, and Afshin Rostamizadeh. Sample selection bias correction theory. In *ALT*, 2008.
16. Corinna Cortes, Mehryar Mohri, Michael Riley, and Afshin Rostamizadeh. Sample selection bias correction theory. In *International Conference on Algorithmic Learning Theory*, pages 38–53. Springer, 2008.
17. Sanjoy Dasgupta and Philip M. Long. Boosting with diverse base classifiers. In *COLT*, 2003.
18. Hal Daumé III and Daniel Marcu. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, 26:101–126, 2006.
19. Miroslav Dudík, Robert E. Schapire, and Steven J. Phillips. Correcting sample selection bias in maximum entropy density estimation. In *NIPS*, 2006.
20. R. M. Dudley. A course on empirical processes. *Lecture Notes in Mathematics*, 1097:2 – 142, 1984.
21. R. M. Dudley. Universal Donsker classes and metric entropy. *Annals of Probability*, 14(4):1306 – 1326, 1987.
22. S. Greenberg and M. Mohri. Tight lower bound on the probability of a binomial exceeding its expectation. *Technical Report 2013-957, Courant Institute, New York, New York*, 2013.
23. David Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Inf. Comput.*, 100(1):78–150, 1992.
24. Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

25. Jiayuan Huang, Alexander J. Smola, Arthur Gretton, Karsten M. Borgwardt, and Bernhard Schölkopf. Correcting sample selection bias by unlabeled data. In *NIPS*, volume 19, pages 601–608, 2006.
26. Savina Andonova Jaeger. Generalization bounds and complexities based on sparsity and clustering for convex combinations of functions from random classes. *Journal of Machine Learning Research*, 6:307–340, 2005.
27. Dankwart Jaeschke. The asymptotic distribution of the supremum of the standardized empirical distribution function on subintervals. *The Annals of Statistics*, pages 108–115, 1979.
28. Jing Jiang and ChengXiang Zhai. Instance Weighting for Domain Adaptation in NLP. In *ACL*, 2007.
29. V. Koltchinskii. Local rademacher complexities and oracle inequalities in risk minimization. *Annals of Statistics*, 34(6), 2006.
30. Vladimir Koltchinskii and Dmitry Panchenko. Rademacher processes and bounding the risk of function learning. In *High Dimensional Probability II*, pages 443–459. Birkhäuser, 2000.
31. Vladmir Koltchinskii and Dmitry Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, 30, 2002.
32. Samuel Kutin and Partha Niyogi. Almost-everywhere algorithmic stability and generalization error. *arXiv preprint arXiv:1301.0579*, 2012.
33. Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *COLT*, 2009.
34. Pascal Massart, Élodie Nédélec, et al. Risk bounds for statistical learning. *The Annals of Statistics*, 34(5):2326–2366, 2006.
35. Colin McDiarmid. On the method of bounded differences. *Surveys in Combinatorics*, 141(1):148–188, 1989.
36. Ron Meir and Tong Zhang. Generalization Error Bounds for Bayesian Mixture Algorithms. *Journal of Machine Learning Research*, 4:839–860, 2003.
37. Shahar Mendelson. Learning without concentration. In *Conference on Learning Theory*, pages 25–39, 2014.
38. Dmitry Panchenko. Symmetrization approach to concentration inequalities for empirical processes. *Annals of Probability*, pages 2068–2081, 2003.
39. Victor H Peña, Tze Leung Lai, and Qi-Man Shao. *Self-Normalized Processes*. Springer, 2008.
40. David Pollard. *Convergence of Stochastic Processess*. Springer, New York, 1984.
41. David Pollard. Asymptotics via empirical processes. *Statistical Science*, 4(4):341 – 366, 1989.
42. Norbert Sauer. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13(1):145–147, 1972.
43. Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.
44. Ingo Steinwart, Clint Scovel, et al. Fast rates for support vector machines using gaussian kernels. *The Annals of Statistics*, 35(2):575–607, 2007.
45. M. Sugiyama, S. Nakajima, H. Kashima, P. von Bünau, and M. Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *NIPS*, 2008.
46. Michael Talagrand. Sharper bounds for gaussian and empirical processes. *Annals of Probability*, 22(1):28–76, 1994.
47. Vladimir N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, 1998.
48. Vladimir N. Vapnik. *Estimation of Dependences Based on Empirical Data, second edition*. Springer, Berlin, 2006.
49. Vladimir N. Vapnik and Alexey Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, 16:264, 1971.
50. Yining Wang and Aarti Singh. Noise-adaptive margin-based active learning and lower bounds under tsybakov noise condition. In *AAAI*, pages 2180–2186, 2016.

Lemmas in Support of Section 3

Lemma 4 *Let $1 < \alpha \leq 2$ and for any $\eta > 0$, let $f\colon (0,+\infty) \times (0,+\infty) \to \mathbb{R}$ be the function defined by $f\colon (x,y) \mapsto \frac{x-y}{\sqrt[\alpha]{x+y+\eta}}$. Then, f is a strictly increasing function of x and a strictly decreasing function of y.*

Proof f is differentiable over its domain of definition and for all $(x,y) \in (0,+\infty) \times (0,+\infty)$,

$$\frac{\partial f}{\partial x}(x,y) = \frac{(x+y+\eta)^{\frac{1}{\alpha}} - \frac{x-y}{\alpha}(x+y+\eta)^{\frac{1}{\alpha}-1}}{(x+y+\eta)^{\frac{2}{\alpha}}} = \frac{\frac{\alpha-1}{\alpha}x + \frac{\alpha+1}{\alpha}y + \eta}{(x+y+\eta)^{1+\frac{1}{\alpha}}} > 0$$
$$\frac{\partial f}{\partial y}(x,y) = \frac{-(x+y+\eta)^{\frac{1}{\alpha}} - \frac{x-y}{\alpha}(x+y+\eta)^{\frac{1}{\alpha}-1}}{(x+y+\eta)^{\frac{2}{\alpha}}} = -\frac{\frac{\alpha+1}{\alpha}x + \frac{\alpha-1}{\alpha}y + \eta}{(x+y+\eta)^{1+\frac{1}{\alpha}}} < 0.$$

Proofs in Support of Section 4

Proof of Theorem 3

Proof We prove the first statement. The second statement can be shown in a very similar way. Fix $1 < \alpha \leq 2$ and $\epsilon > 0$ and $\mathcal{S}$ assume that for any $h \in H$ and $t \geq 0$, the following holds:

$$\frac{\Pr[L(h,z) > t] - \widehat{\Pr}[L(h,z) > t]}{\sqrt[\alpha]{\Pr[L(h,z) > t] + \tau}} \leq \epsilon. \tag{17}$$

We show that this implies that for any $h \in H$, $\frac{\mathcal{L}(h) - \widehat{\mathcal{L}}_S(h)}{\sqrt[\alpha]{\mathcal{L}_\alpha(h) + \tau}} \leq \Gamma(\alpha,\epsilon)\epsilon$. By the properties of the Lebesgue integral, we can write

$$\mathcal{L}(h) = \mathrm{E}_{z\sim D}[L(h,z)] = \int_0^{+\infty} \Pr[L(h,z) > t]\, dt$$
$$\widehat{\mathcal{L}}(h) = \mathrm{E}_{z\sim \widehat{D}}[L(h,z)] = \int_0^{+\infty} \widehat{\Pr}[L(h,z) > t]\, dt,$$

and, similarly,

$$\mathcal{L}_\alpha(h) = \mathcal{L}_\alpha(h) = \int_0^{+\infty} \Pr[L^\alpha(h,z) > t]\, dt = \int_0^{+\infty} \alpha t^{\alpha-1} \Pr[L(h,z) > t]\, dt.$$

In what follows, we use the notation $I_\alpha = \mathcal{L}_\alpha(h) + \tau$. Let $t_0 = s I_\alpha^{\frac{1}{\alpha}}$ and $t_1 = t_0 \left[\frac{1}{\epsilon}\right]^{\frac{1}{\alpha-1}}$ for $s > 0$. To bound $\mathcal{L}(h) - \widehat{\mathcal{L}}(h)$, we simply bound $\Pr[L(h,z) > t] - \widehat{\Pr}[L(h,z) > t]$ by $\Pr[L(h,z) > t]$ for large values of t, that is $t > t_1$, and use inequality (17) for smaller values of t:

$$\begin{aligned}\mathcal{L}(h) - \widehat{\mathcal{L}}(h) &= \int_0^{+\infty} \Pr[L(h,z) > t] - \widehat{\Pr}[L(h,z) > t]\, dt \\ &\leq \int_0^{t_1} \epsilon \sqrt[\alpha]{\Pr[L(h,z) > t] + \tau}\, dt + \int_{t_1}^{+\infty} \Pr[L(h,z) > t]\, dt.\end{aligned}$$

For relatively small values of t, $\Pr[L(h,z) > t]$ is close to one. Thus, we can write

$$\begin{aligned}\mathcal{L}(h) - \widehat{\mathcal{L}}(h) &\leq \int_0^{t_0} \epsilon \sqrt[\alpha]{1+\tau}\, dt + \int_{t_0}^{t_1} \epsilon \sqrt[\alpha]{\Pr[L(h,z) > t] + \tau}\, dt + \int_{t_1}^{+\infty} \Pr[L(h,z) > t] dt \\ &= \int_0^{+\infty} f(t) g(t)\, dt,\end{aligned}$$

with

$$f(t) = \begin{cases} \gamma_1 I_\alpha^{\frac{\alpha-1}{\alpha^2}} \epsilon \sqrt[\alpha]{1+\tau} & \text{if } 0 \le t \le t_0 \\ \gamma_2 \left[\alpha t^{\alpha-1}(\Pr[L(h,z) > t] + \tau)\right]^{\frac{1}{\alpha}} \epsilon & \text{if } t_0 < t \le t_1 \\ \gamma_2 \left[\alpha t^{\alpha-1} \Pr[L(h,z) > t]\right]^{\frac{1}{\alpha}} \epsilon & \text{if } t_1 < t. \end{cases} \quad g(t) = \begin{cases} \frac{1}{\gamma_1 I_\alpha^{\frac{\alpha-1}{\alpha^2}}} & \text{if } 0 \le t \le t_0 \\ \frac{1}{\gamma_2 (\alpha t^{\alpha-1})^{\frac{1}{\alpha}}} & \text{if } t_0 < t \le t_1 \\ \frac{\Pr[L(h,z)>t]^{\frac{\alpha-1}{\alpha}}}{\gamma_2 (\alpha t^{\alpha-1})^{\frac{1}{\alpha}}} \frac{1}{\epsilon} & \text{if } t_1 < t, \end{cases}$$

where γ_1, γ_2 are positive parameters that we shall select later. Now, since $\alpha > 1$, by Hölder's inequality,

$$\mathcal{L}(h) - \widehat{\mathcal{L}}(h) \le \left[\int_0^{+\infty} f(t)^\alpha \, dt\right]^{\frac{1}{\alpha}} \left[\int_0^{+\infty} g(t)^{\frac{\alpha}{\alpha-1}} \, dt\right]^{\frac{\alpha-1}{\alpha}}.$$

The first integral on the right-hand side can be bounded as follows:

$$\begin{aligned} \int_0^{+\infty} f(t)^\alpha \, dt &= \int_0^{t_0} (1+\tau)(\gamma_1 I_\alpha^{\frac{\alpha-1}{\alpha^2}} \epsilon)^\alpha \, dt + \gamma_2^\alpha \epsilon^\alpha \tau \int_{t_0}^{t_1} \alpha t^{\alpha-1} dt + \gamma_2^\alpha \int_{t_0}^{+\infty} \alpha t^{\alpha-1} \Pr[L(h,z) > t] \epsilon^\alpha \, dt \\ &\le (1+\tau)\gamma_1^\alpha I_\alpha^{\frac{\alpha-1}{\alpha}} t_0 \epsilon^\alpha + \gamma_2^\alpha \epsilon^\alpha \tau (t_1^\alpha - t_0^\alpha) + \gamma_2^\alpha \epsilon^\alpha I_\alpha \\ &\le (\gamma_1^\alpha (1+\tau) s + \gamma_2^\alpha (1 + s^\alpha (1/\epsilon)^{\frac{\alpha}{\alpha-1}} \tau)) \epsilon^\alpha I_\alpha \\ &\le (\gamma_1^\alpha (1+\tau) s + \gamma_2^\alpha (1 + s^\alpha \tau^{\frac{1}{\alpha}})) \epsilon^\alpha I_\alpha. \end{aligned}$$

Since $t_1/t_0 = (1/\epsilon)^{\frac{1}{\alpha-1}}$, the second one can be computed and bounded following

$$\begin{aligned} \int_0^{+\infty} g(t)^{\frac{\alpha}{\alpha-1}} \, dt &= \int_0^{t_0} \frac{dt}{\gamma_1^{\frac{\alpha}{\alpha-1}} I_\alpha^{\frac{1}{\alpha}}} + \int_{t_0}^{t_1} \frac{1}{\gamma_2^{\frac{\alpha}{\alpha-1}} \alpha^{\frac{1}{\alpha-1}}} \frac{dt}{t} + \int_{t_1}^{+\infty} \frac{\Pr[L(h,z) > t]}{\gamma_2^{\frac{\alpha}{\alpha-1}} \alpha^{\frac{1}{\alpha-1}} \epsilon^{\frac{\alpha}{\alpha-1}} t} dt \\ &= \frac{s}{\gamma_1^{\frac{\alpha}{\alpha-1}}} + \frac{1}{\gamma_2^{\frac{\alpha}{\alpha-1}} (\alpha-1) \alpha^{\frac{1}{\alpha-1}}} \log \frac{1}{\epsilon} + \int_{t_1}^{+\infty} \frac{\alpha t^{\alpha-1} \Pr[L(h,z) > t]}{\gamma_2^{\frac{\alpha}{\alpha-1}} (\alpha\epsilon)^{\frac{\alpha}{\alpha-1}} t^\alpha} dt \\ &\le \frac{s}{\gamma_1^{\frac{\alpha}{\alpha-1}}} + \frac{1}{\gamma_2^{\frac{\alpha}{\alpha-1}} (\alpha-1) \alpha^{\frac{1}{\alpha-1}}} \log \frac{1}{\epsilon} + \int_{t_1}^{+\infty} \frac{\alpha t^{\alpha-1} \Pr[L(h,z) > t]}{\gamma_2^{\frac{\alpha}{\alpha-1}} (\alpha\epsilon)^{\frac{\alpha}{\alpha-1}} t_1^\alpha} dt \\ &\le \frac{s}{\gamma_1^{\frac{\alpha}{\alpha-1}}} + \frac{1}{\gamma_2^{\frac{\alpha}{\alpha-1}} (\alpha-1) \alpha^{\frac{1}{\alpha-1}}} \log \frac{1}{\epsilon} + \frac{I_\alpha}{\gamma_2^{\frac{\alpha}{\alpha-1}} (\alpha\epsilon)^{\frac{\alpha}{\alpha-1}} s^\alpha I_\alpha (\frac{1}{\epsilon})^{\frac{\alpha}{\alpha-1}}} \\ &= \frac{s}{\gamma_1^{\frac{\alpha}{\alpha-1}}} + \frac{1}{\gamma_2^{\frac{\alpha}{\alpha-1}}} \left(\frac{1}{(\alpha-1)\alpha^{\frac{1}{\alpha-1}}} \log \frac{1}{\epsilon} + \frac{1}{\alpha^{\frac{\alpha}{\alpha-1}} s^\alpha} \right). \end{aligned}$$

Combining the bounds obtained for these integrals yields directly

$$\begin{aligned} &\mathcal{L}(h) - \widehat{\mathcal{L}}(h) \\ &\le \left[(\gamma_1^\alpha (1+\tau) s + \gamma_2^\alpha (1 + s^\alpha \tau^{\frac{1}{\alpha}})) \epsilon^\alpha I_\alpha\right]^{\frac{1}{\alpha}} \left[\frac{s}{\gamma_1^{\frac{\alpha}{\alpha-1}}} + \frac{1}{\gamma_2^{\frac{\alpha}{\alpha-1}}} \left(\frac{1}{(\alpha-1)\alpha^{\frac{1}{\alpha-1}}} \log \frac{1}{\epsilon} + \frac{1}{\alpha^{\frac{\alpha}{\alpha-1}} s^\alpha} \right)\right]^{\frac{\alpha-1}{\alpha}} \\ &= (\gamma_1^\alpha (1+\tau) s + \gamma_2^\alpha (1 + s^\alpha \tau^{\frac{1}{\alpha}}))^{\frac{1}{\alpha}} \left[\frac{s}{\gamma_1^{\frac{\alpha}{\alpha-1}}} + \frac{1}{\gamma_2^{\frac{\alpha}{\alpha-1}}} \left(\frac{1}{(\alpha-1)\alpha^{\frac{1}{\alpha-1}}} \log \frac{1}{\epsilon} + \frac{1}{\alpha^{\frac{\alpha}{\alpha-1}} s^\alpha} \right)\right]^{\frac{\alpha-1}{\alpha}} \epsilon I_\alpha^{\frac{1}{\alpha}}. \end{aligned}$$

Observe that the expression on the right-hand side can be rewritten as $\|\mathbf{u}\|_\alpha \|\mathbf{v}\|_{\frac{\alpha}{\alpha-1}} \, \epsilon I_\alpha^{\frac{1}{\alpha}}$ where the vectors $\mathbf{u}$ and $\mathbf{v}$ are defined by $\mathbf{u} = (\gamma_1(1+\tau)^{\frac{1}{\alpha}} s^{\frac{1}{\alpha}}, \gamma_2 (1 + s^\alpha \tau^{\frac{1}{\alpha}})^{\frac{1}{\alpha}})$ and $\mathbf{v} =$

$(v_1, v_2) = \left(\frac{s^{\frac{\alpha-1}{\alpha}}}{\gamma_1}, \frac{1}{\gamma_2} \left[\frac{1}{(\alpha-1)\alpha^{\frac{1}{\alpha-1}}} \log \frac{1}{\epsilon} + \frac{1}{\alpha^{\frac{\alpha}{\alpha-1}} s^\alpha} \right]^{\frac{\alpha-1}{\alpha}} \right)$. The inner product $\mathbf{u} \cdot \mathbf{v}$ does not depend on γ_1 and γ_2, and equality holds if and only if the vectors $\mathbf{u}$ and $\mathbf{v}' = (v_1^{\frac{1}{\alpha-1}}, v_2^{\frac{1}{\alpha-1}})$ are collinear (as we can see by applying Hölder's inequality).

γ_1 and γ_2 can be chosen so that $\det(\mathbf{u}, \mathbf{v}') = 0$, since this condition can be rewritten as

$$s^{\frac{1}{\alpha}} (1+\tau)^{\frac{1}{\alpha}} \frac{\gamma_1}{\gamma_2^{\frac{1}{\alpha-1}}} \left[\frac{1}{(\alpha-1)\alpha^{\frac{1}{\alpha-1}}} \log \frac{1}{\epsilon} + \frac{1}{\alpha^{\frac{\alpha}{\alpha-1}} s^\alpha} \right]^{\frac{1}{\alpha}} - s^{\frac{1}{\alpha}} (1 + s^\alpha \tau^{\frac{1}{\alpha}})^{\frac{1}{\alpha}} \frac{\gamma_2}{\gamma_1^{\frac{1}{\alpha-1}}} = 0, \quad (18)$$

or equivalently,

$$\left(\frac{\gamma_1}{\gamma_2} \right)^{\frac{\alpha}{\alpha-1}} \left[\frac{1}{(\alpha-1)\alpha^{\frac{1}{\alpha-1}}} \log \frac{1}{\epsilon} + \frac{1}{\alpha^{\frac{\alpha}{\alpha-1}} s^\alpha} \right]^{\frac{1}{\alpha}} - (1 + s^\alpha \tau^{\frac{1}{\alpha}})^{\frac{1}{\alpha}} = 0. \quad (19)$$

Thus, for such values of γ_1 and γ_2, the following inequality holds:

$$\mathcal{L}(h) - \widehat{\mathcal{L}}(h) \leq (\mathbf{u} \cdot \mathbf{v})\, \epsilon I_\alpha^{\frac{1}{\alpha}} = f(s)\, \epsilon I_\alpha^{\frac{1}{\alpha}},$$

with

$$\begin{aligned} f(s) &= (1+\tau)^{\frac{1}{\alpha}} s + (1 + s^\alpha \tau^{\frac{1}{\alpha}})^{\frac{1}{\alpha}} \left[\frac{1}{(\alpha-1)\alpha^{\frac{1}{\alpha-1}}} \log \frac{1}{\epsilon} + \frac{1}{\alpha^{\frac{\alpha}{\alpha-1}} s^\alpha} \right]^{\frac{\alpha-1}{\alpha}} \\ &= (1+\tau)^{\frac{1}{\alpha}} s + \frac{(1 + s^\alpha \tau^{\frac{1}{\alpha}})^{\frac{1}{\alpha}}}{\alpha} \left[\frac{\alpha}{(\alpha-1)} \log \frac{1}{\epsilon} + \frac{1}{s^\alpha} \right]^{\frac{\alpha-1}{\alpha}}. \end{aligned}$$

Setting $s = \frac{\alpha-1}{\alpha}$ yields the statement of the theorem.

Proof of Proposition 1

Proof We prove the first inequality. The second can be proven in a very similar way. Fix $\alpha > 2$ and $h \in H$. As in the proof of Theorem 3, we bound $\Pr[L(h,z) > t]$ by 1 for t close to 0, say $t \leq t_0$ for some $t_0 > 0$ that we shall later determine. We can write

$$\int_0^{+\infty} \sqrt{\Pr[L(h,z) > t]} dt \leq \int_0^{t_0} 1 dt + \int_{t_0}^{+\infty} \sqrt{\Pr[L(h,z) > t]} dt = \int_0^{+\infty} f(t) g(t) dt,$$

with functions f and g defined as follows:

$$f(t) = \begin{cases} \gamma I_\alpha^{\frac{\alpha-1}{2\alpha}} & \text{if } 0 \leq t \leq t_0 \\ \alpha^{\frac{1}{2}} t^{\frac{\alpha-1}{2}} \Pr[L(h,z) > t]^{\frac{1}{2}} & \text{if } t_0 < t. \end{cases} \qquad g(t) = \begin{cases} \frac{1}{\gamma I_\alpha^{\frac{\alpha-1}{2\alpha}}} & \text{if } 0 \leq t \leq t_0 \\ \frac{1}{\alpha^{\frac{1}{2}} t^{\frac{\alpha-1}{2}}} & \text{if } t_0 < t, \end{cases}$$

where $I_\alpha = \mathcal{L}_\alpha(h)$ and where γ is a positive parameter that we shall select later. By the Cauchy-Schwarz inequality,

$$\int_0^{+\infty} \sqrt{\Pr[L(h,z) > t]} dt \leq \left(\int_0^{+\infty} f(t)^2 dt \right)^{\frac{1}{2}} \left(\int_0^{+\infty} g(t)^2 dt \right)^{\frac{1}{2}}.$$

Thus, we can write

$$\begin{aligned} &\int_0^{+\infty} \sqrt{\Pr[L(h,z) > t]} dt \\ &\leq \left(\gamma^2 I_\alpha^{\frac{\alpha-1}{\alpha}} t_0 + \int_{t_0}^{+\infty} \alpha t^{\alpha-1} \Pr[L(h,z) > t] dt \right)^{\frac{1}{2}} \left(\frac{t_0}{\gamma^2 I_\alpha^{\frac{\alpha-1}{\alpha}}} + \int_{t_0}^{+\infty} \frac{1}{\alpha t^{\alpha-1}} dt \right)^{\frac{1}{2}} \\ &\leq \left(\gamma^2 I_\alpha^{\frac{\alpha-1}{\alpha}} t_0 + I_\alpha \right)^{\frac{1}{2}} \left(\frac{t_0}{\gamma^2 I_\alpha^{\frac{\alpha-1}{\alpha}}} + \frac{1}{\alpha(\alpha-2) t_0^{\alpha-2}} \right)^{\frac{1}{2}}. \end{aligned}$$

Introducing t_1 with $t_0 = I_\alpha^{1/\alpha} t_1$ leads to

$$\int_0^{+\infty} \sqrt{\Pr[L(h,z) > t]} dt \leq \left(\gamma^2 I_\alpha t_1 + I_\alpha\right)^{\frac{1}{2}} \left(\frac{t_1}{\gamma^2 {I_\alpha}^{\frac{\alpha-2}{\alpha}}} + \frac{1}{\alpha(\alpha-2) t_1^{\alpha-2} {I_\alpha}^{\frac{\alpha-2}{\alpha}}} \right)^{\frac{1}{2}}$$
$$\leq \left(\gamma^2 t_1 + 1\right)^{\frac{1}{2}} \left(\frac{t_1}{\gamma^2} + \frac{1}{\alpha(\alpha-2) t_1^{\alpha-2}} \right)^{\frac{1}{2}} I_\alpha^{\frac{1}{\alpha}}.$$

We now seek to minimize the expression $\left(\gamma^2 t_1 + 1\right)^{\frac{1}{2}} \left(\frac{t_1}{\gamma^2} + \frac{1}{\alpha(\alpha-2) t_1^{\alpha-2}} \right)^{\frac{1}{2}}$, first as a function of γ. This expression can be viewed as the product of the norms of the vectors $\mathbf{u} = (\gamma t_1^{\frac{1}{2}}, 1)$ and $\mathbf{v} = (\frac{t_1^{\frac{1}{2}}}{\gamma}, \frac{1}{\sqrt{\alpha(\alpha-2)} {t_1}^{\frac{\alpha-2}{2}}})$, with a constant inner product (not depending on γ). Thus, by the properties of the Cauchy-Schwarz inequality, it is minimized for collinear vectors and in that case equals their inner product:

$$\mathbf{u} \cdot \mathbf{v} = t_1 + \frac{1}{\sqrt{\alpha(\alpha-2)} {t_1}^{\frac{\alpha-2}{2}}}.$$

Differentiating this last expression with respect to t_1 and setting the result to zero gives the minimizing value of t_1: $(\frac{2}{\alpha-2}\sqrt{\alpha(\alpha-2)})^{-\frac{2}{\alpha}} = \left(\frac{1}{2}\sqrt{\frac{\alpha-2}{\alpha}}\right)^{\frac{2}{\alpha}}$. For that value of t_1,

$$\mathbf{u} \cdot \mathbf{v} = \left(1 + \frac{2}{\alpha-2}\right) t_1 = \frac{\alpha}{\alpha-2} \left(\frac{1}{2}\sqrt{\frac{\alpha-2}{\alpha}}\right)^{\frac{2}{\alpha}} = \left(\frac{1}{2}\right)^{\frac{2}{\alpha}} \left(\frac{\alpha-2}{\alpha}\right)^{\frac{1-\alpha}{\alpha}},$$

which concludes the proof.