Sharper Bounds for Uniformly Stable Algorithms

Olivier Bousquet   OBOUSQUET@GOOGLE.COM
Google Research, Brain Team, Zürich

Yegor Klochkov   YK376@CAM.AC.UK
Cambridge-INET, Faculty of Economics, University of Cambridge

Nikita Zhivotovskiy   ZHIVOTOVSKIY@GOOGLE.COM
Google Research, Brain Team, Zürich

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

Deriving generalization bounds for stable algorithms is a classical question in learning theory taking its roots in the early works by Vapnik and Chervonenkis (1974) and Rogers and Wagner (1978). In a series of recent breakthrough papers by Feldman and Vondrák (2018, 2019), it was shown that the best known high probability upper bounds for uniformly stable learning algorithms due to Bousquet and Elisseef (2002) are sub-optimal in some natural regimes. To do so, they proved two generalization bounds that significantly outperform the simple generalization bound of Bousquet and Elisseef (2002). Feldman and Vondrák also asked if it is possible to provide sharper bounds and prove corresponding high probability lower bounds. This paper is devoted to these questions: firstly, inspired by the original arguments of Feldman and Vondrák (2019), we provide a short proof of the moment bound that implies the generalization bound stronger than both recent results (Feldman and Vondrák, 2018, 2019). Secondly, we prove general lower bounds, showing that our moment bound is sharp (up to a logarithmic factor) unless some additional properties of the corresponding random variables are used. Our main probabilistic result is a general concentration inequality for weakly correlated random variables, which may be of independent interest.

1. Introduction

The main motivation of our study is the analysis of *stable* learning algorithms (we recall the definition introduced in (Bousquet and Elisseef, 2002) below). We are given an i.i.d sample of points $S = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ distributed independently according to some unknown measure P on $\mathcal{X} \times \mathcal{Y}$. A learning algorithm $A : (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{Y}^{\mathcal{X}}$ maps a training sample to a function mapping the instance space $\mathcal{X}$ into the space of labels $\mathcal{Y}$. The output of the learning algorithm based on the sample S will be denoted by A_S. The quality of the function returned by the algorithm is measured using a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$. More precisely, given the random sample $S = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, the *risk* of the algorithm is defined as

$$R(A_S) = \mathbb{E}_{(X,Y) \sim P}\, \ell(A_S(X), Y)\,.$$

One of the fundamental questions in statistical learning is how to estimate the risk $R(A_S)$ of an algorithm from the sample S itself or with limited additional data. A very good estimate can be obtained if one has access to additional data from the same distribution (a test set), and in practice the so-called test error is used as such. However, the question of whether one could obtain an

accurate estimate from no or very limited additional data is essential if one wants to understand the statistical properties of the learning algorithm.

If one wishes to estimate $R(A_S)$ with no additional data, a natural quantity to consider is the so-called *empirical risk* defined as

$$R_{\text{emp}}(A_S) = \frac{1}{n} \sum_{i=1}^{n} \ell(A_S(X_i), Y_i).$$

A large body of work has been dedicated to obtaining *generalization bounds*, i.e., high probability bounds on the error of the empirical risk estimator: $R(A_S) - R_{\text{emp}}(A_S)$. A standard way to prove the generalization bounds is based on the sensitivity of the algorithm to changes in the learning sample, such as leaving one of the data points out or replacing it with a different one. To the best of our knowledge, this idea was first used by Vapnik and Chervonenkis to prove the in-expectation generalization bound for what now is known as hard-margin Support Vector Machine (Vapnik and Chervonenkis, 1974). Later works by Devroye and Wagner used a notion of stability to prove high probability generalization bounds for k-nearest neighbors (Devroye and Wagner, 1979). Bousquet and Elisseef (2002) provide an extensive analysis of different notions of stability and the corresponding (sometimes) high probability generalization bounds. In the context of the stochastic gradient method, uniform stability was shown by Hardt, Recht, and Singer (2016). Among some recent contributions on high probability upper bounds based on the notions of stability is the paper (Maurer, 2017), which studies generalization bounds for a particular case of linear regression with a strongly convex regularizer, as well as the recent works (Zhivotovskiy, 2017; Bousquet et al., 2020), which provide sharp exponential upper bounds for the SVM in the realizable case. For additional background on the topic we refer to the recent papers (Feldman and Vondrák, 2018, 2019) and references therein.

Let us define stability more precisely. For the sake of simplicity, we denote $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. The learning algorithm A is *uniformly stable* with parameter γ if given any samples

$$S = \{z_1, \ldots, z_n\} \in \mathcal{Z}^n \quad \text{and} \quad S^i = \{z_1, \ldots, z_{i-1}, z'_i, z_{i+1}, \ldots, z_n\} \in \mathcal{Z}^n,$$

for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$, we have

$$|\ell(A_S(x), y) - \ell(A_{S^i}(x), y)| \leq \gamma. \tag{1}$$

Note that this can be thought of as a deterministic version of Differential Privacy, and there is active work exploring the connections between the two notions (see e.g., (Wang, Lei, and Fienberg, 2016)).

Several works have focused on using the stability technique to derive bounds on the generalization error and we also consider this question here. This raises the question of the relevance of such an approach for studying modern machine learning models such as those used in Deep Learning. Indeed, it has been observed that very often the models can be trained to achieve zero empirical error which contradicts the possibility of having small generalization bounds. These models essentially perform *interpolation* or *memorization* of the data and the question of why this would not lead to overfitting is an active area of research (Belkin, Hsu, and Mitra, 2018). It turns out that the stability technique can also be used in the interpolation regime as was actually initially done by Devroye and Wagner (1979) when studying k-nearest neighbors (e.g., 1-nearest neighbor rules do interpolate the data, yet the stability approach can provide meaningful bound). We will return to this point in Section 5.

Let us now recall the known results about generalization bounds for uniformly stable algorithms. In order to simplify the notation in what follows, we rescale the quantity of interest by n and thus write the bounds for

$$n(R(A_S) - R_{\text{emp}}(A_S)),$$

and additionally assume that the loss function ℓ is bounded by L.

The basic and, until very recently, the best known result is the high probability upper bound in (Bousquet and Elisseef, 2002) which states that, with probability at least $1 - \delta$,

$$n(R(A_S) - R_{\text{emp}}(A_S)) \lesssim (n\sqrt{n}\gamma + L\sqrt{n})\sqrt{\log\left(\frac{1}{\delta}\right)}. \tag{2}$$

It is easy to observe that this generalization bound is tight only when $\gamma \lesssim \frac{1}{n}$, which means that only under this assumption, the generalization error $R(A_S) - R_{\text{emp}}(A_S)$ converges to zero with the optimal rate $\frac{1}{\sqrt{n}}$. However, in some applications the regime $\gamma \lesssim \frac{1}{\sqrt{n}}$ is of interest, and the bound (2) can not guarantee any convergence.

In a series of breakthrough papers, Feldman and Vondrák (2018, 2019) managed to correctly capture this phenomenon. They first showed a generalization bound of the form

$$n(R(A_S) - R_{\text{emp}}(A_S)) \lesssim (n\sqrt{\gamma L} + L\sqrt{n})\sqrt{\log\left(\frac{1}{\delta}\right)}, \tag{3}$$

where as before, the parameter γ corresponds to the stability, and the parameter L bounds the loss function ℓ uniformly. In their second paper, Feldman and Vondrák show a stronger generalization bound

$$n(R(A_S) - R_{\text{emp}}(A_S)) \lesssim n\gamma(\log n)^2 + n\gamma \log n \log\left(\frac{1}{\delta}\right) + L\sqrt{n}\sqrt{\log\left(\frac{1}{\delta}\right)}. \tag{4}$$

Up to logarithmic factors, the bound (4) shows that with high probability in the regime $\gamma \sim \frac{1}{\sqrt{n}}$, the generalization error $R(A_S) - R_{\text{emp}}(A_S)$ converges to zero with the optimal rate $\frac{1}{\sqrt{n}}$. However, as claimed by Feldman and Vondrák, the bound (3) should not be wholly discarded since it does not contain additional logarithmic factors $\log n$ and $(\log n)^2$. More importantly, the bound (3) is *sub-gaussian*, which means that the dependence on δ comes only in the form $\sqrt{\log\left(\frac{1}{\delta}\right)}$. At the same time, the bound (4) shows both *sub-gaussian* and the *sub-exponential* regimes since it contains *two types* of terms: $\sqrt{\log\left(\frac{1}{\delta}\right)}$ and $\log\left(\frac{1}{\delta}\right)$. We will discuss the notions of *sub-gaussian* and *sub-exponential* high probability upper bounds below.

In (Feldman and Vondrák, 2019), the authors ask if their high-probability upper bounds (3) and (4) can be strengthened and if they can be matched by a high probability lower bound. In this paper, we are making some progress in answering both questions. We shortly summarize our findings:

- Our main probabilistic result is Theorem 4, presented in Section 3. As one of the immediate corollaries, it implies the risk bound of the form

$$n|R(A_S) - R_{\text{emp}}(A_S)| \lesssim n\gamma \log n \log\left(\frac{1}{\delta}\right) + L\sqrt{n}\sqrt{\log\left(\frac{1}{\delta}\right)}, \tag{5}$$

which removes the unnecessary term $n\gamma(\log n)^2$ from (4). We emphasize that our analysis is inspired by the original sample-splitting argument of Feldman and Vondrák, although our proof is significantly more straightforward. In particular, we avoid several involved technical steps, which ultimately leads us to better generalization bounds.

- Our Theorem 4 will also easily imply the sub-gaussian bound (3), which was originally shown via the techniques taking their roots from Differential Privacy. Therefore, we also make a natural bridge between the bounds of the form (3) and (4), which have different dependencies on $\log\frac{1}{\delta}$, L, and γ.

- In Section 4, we show that the bound of our Theorem 4 is tight unless some additional properties of the corresponding random variables are used. Our lower bounds are presented by some specific functions satisfying the assumptions of Theorem 4. We remark that our lower bound does not completely answer the question of the optimality of (5) for uniformly stable algorithms, as it only shows the tightness of the bound implying (5). We discuss it in more detail in Section 4.

2. Preliminaries

2.1. Notation

We provide some notation that will be used throughout the paper. The symbol $\mathbf{1}[A]$ will denote an indicator of the event A. For a pair of non-negative functions f, g the notation $f \lesssim g$ or $g \gtrsim f$ will mean that for some universal constant $c > 0$ it holds that $f \leq cg$. Similarly, we introduce $f \sim g$ to be equivalent to $g \lesssim f \lesssim g$. For $a, b \in \mathbb{R}$ we define $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$. The L_p norm of a random variable will be denoted as $\|Y\|_p = (\mathbb{E}|Y|^p)^{\frac{1}{p}}$. Let $[k]$ denote the set $\{1, \ldots, k\}$. To avoid some technical problems for $x > 0$, by $\log x$ we usually mean $\log x \vee 1$

In what follows, we work with functions of n independent variables $Z = (Z_1, \ldots, Z_n)$. For $A \subset [n]$ we will write $Z_A = (Z_j)_{j \in A}$. We also use the following notation

$$Z^i = (Z_1, \ldots, Z_{i-1}, Z_i', Z_{i+1}, \ldots, Z_n),$$

where Z_i' is an independent copy of Z_i. In addition, for $f = f(Z)$ and $A \subset [n]$ we write $\|f\|_p(Z_A) = \mathbb{E}^{1/p}[|f|^p|Z_A]$. In particular, if we have an a.s. bound $\|f\|_p(Z_A) \leq C$ for any realisation of Z_A, then by a simple integration argument we have

$$\|f\|_p = (\mathbb{E}\mathbb{E}[|f|^p|Z_A])^{1/p} \leq C, \tag{6}$$

i.e., in this sense, a conditional bound is stronger than the unconditional one. Finally, for $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ slightly abusing the notation we set $\|x\|_2 = \left(\sum_{i=1}^{d} x_i^2\right)^{1/2}$ and $\|x\|_\infty = \max_{i \in [n]} |x_i|$.

2.2. Equivalence of Tails and Moments

Probabilistic bounds in Learning Theory are often of the form

$$Y \leq a\sqrt{\log\left(\frac{1}{\delta}\right)} + b\log\left(\frac{1}{\delta}\right), \tag{7}$$

with probability at least $1 - \delta$, for any $\delta \in (0, 1)$ and some $a, b \geq 0$. Here, Y is a random variable of interest, e.g., the excess risk. The term with $\sqrt{\log\left(\frac{1}{\delta}\right)}$ is referred to as a *sub-gaussian* tail, as it matches the deviations of a Gaussian random variable. The term with $\log\left(\frac{1}{\delta}\right)$ is called a *sub-exponential* tail for a similar reason. In general, the bound above represents a *mixture of sub-gaussian and sub-exponential* tails. In particular, all the known generalization bounds (2), (3), (4) are of the form (7).

An alternative way of studying tail bounds is via the moment norms. It is well-known that for a Gaussian random variable Y, there exists some $a \geq 0$ such that for any $p \geq 1$, $\|Y\|_p \leq \sqrt{p}a$, while when Y is sub-exponential, we have $\|Y\|_p \leq pb$ for some $b \geq 0$ (see e.g., Propositions 2.5.2 and 2.7.1 in (Vershynin, 2016)). In what follows, we will consider the random variables with two levels of moments, that is for some $a, b \geq 0$ that do not depend on p,

$$\|Y\|_p \leq \sqrt{p}a + pb, \qquad \forall p \geq 1.$$

In fact, the above bound and the bound (7) are equivalent up to a constant, as the following simple result suggests.

Lemma 1 (Equivalence of tails and moments) *Suppose, a random variable has a mixture of sub-gaussian and sub-exponential tails, in the sense that it satisfies for any $\delta \in (0, 1)$ with probability at least $1 - \delta$,*

$$|Y| \leq a\sqrt{\log\left(\frac{e}{\delta}\right)} + b\log\left(\frac{e}{\delta}\right),$$

for some $a, b \geq 0$. Then, for any $p \geq 1$ it holds that

$$\|Y\|_p \leq 3\sqrt{p}a + 9pb.$$

And vice versa, if $\|Y\|_p \leq \sqrt{p}a + pb$ for any $p \geq 1$ then for any $\delta \in (0, 1)$ we have, with probability at least $1 - \delta$,

$$|Y| \leq e\left(a\sqrt{\log\left(\frac{e}{\delta}\right)} + b\log\left(\frac{e}{\delta}\right)\right).$$

The proof is a simple adaptation of Theorem 2.3 from (Boucheron et al., 2013). For the sake of completeness, we present it in Appendix A. So it is quite natural to consider moment bounds and it turns out that moments are often easier to work with than deviation inequalities, for example, for lower bounds, as we will see in Section 4. We now state several well-known moment inequalities for sums and functions of independent random variables which will be our main tools. One of them is the moment version of the bounded differences inequality, which follows immediately from Theorem 15.4 in (Boucheron et al., 2013).

Lemma 2 (Bounded differences/McDiarmid's inequality) *Consider a function f of independent random variables $X_1, \ldots, X_n$ that take their values in $\mathcal{X}$. Suppose, that f satisfies the bounded differences property, namely, for any $i = 1, \ldots, n$ and any $x_1, \ldots, x_n, x_i' \in \mathcal{X}$ it holds that*

$$|f(x_1, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \leq \beta. \tag{8}$$

Then, we have for any $p \geq 2$,

$$\|f(X_1, \ldots, X_n) - \mathbb{E}f(X_1, \ldots, X_n)\|_p \leq 2\sqrt{np}\beta.$$

5

Notice that it is easy to apply the above lemma in the case that $|X_i| \leq M$ a.s. and $\mathbb{E}X_i = 0$. Since $\sum_{i=1}^{n} X_i$ satisfies the bounded differences condition with $\beta = 2M$, we have

$$\left\| \sum_{i=1}^{n} X_i \right\|_p \leq 4\sqrt{np}M. \tag{9}$$

We will refer to it as the moment version of *Hoeffding's inequality*.

Next, we use the following version of the classical Marcinkiewicz-Zygmund inequality (we also refer to Chapter 15 in (Boucheron et al., 2013) that contains similar inequalities).

Lemma 3 (Marcinkiewicz-Zygmund's inequality (Ren and Liang, 2001)) *Let $X_1, \ldots, X_n$ be independent centered random variables with a finite p-th moment for $p \geq 2$. Then,*

$$\left\| \sum_{i=1}^{n} X_i \right\|_p \leq 3\sqrt{2np} \left(\frac{1}{n} \sum_{i=1}^{n} \|X_i\|_p^p \right)^{\frac{1}{p}}.$$

3. Upper Bounds

3.1. A Moment Bound for Sums

We present here our main result which is a moment inequality for sums of functions of n independent variables.

Theorem 4 *Let $Z = (Z_1, \ldots, Z_n)$ be a vector of independent random variables each taking values in $\mathcal{Z}$, and let $g_1, \ldots, g_n$ be some functions $g_i : \mathcal{Z}^n \to \mathbb{R}$ such that the following holds for any $i \in [n]$:*

- $\left| \mathbb{E}[g_i(Z)|Z_i] \right| \leq M$ *a.s.,*

- $\mathbb{E}[g_i(Z)|Z_{[n]\setminus\{i\}}] = 0$ *a.s.,*

- g_i *has a bounded difference* (8) β *with respect to all variables except the i-th variable.*

Then, for any $p \geq 2$,

$$\left\| \sum_{i=1}^{n} g_i(Z) \right\|_p \leq 12\sqrt{2}pn\beta\lceil \log_2 n \rceil + 4M\sqrt{pn}.$$

Proof Without loss of generality, we suppose that $n = 2^k$. Otherwise, we can add extra functions equal to zero, increasing the number of terms by at most two times.

Consider a sequence of partitions $\mathcal{B}_0, \ldots, \mathcal{B}_k$ with $\mathcal{B}_0 = \{\{i\} : i \in [n]\}$, $\mathcal{B}_k = \{[n]\}$, and to get $\mathcal{B}_l$ from $\mathcal{B}_{l+1}$ we split each subset in $\mathcal{B}_{l+1}$ into two equal parts. We have

$$\mathcal{B}_0 = \{\{1\}, \ldots, \{2^k\}\}, \ \mathcal{B}_1 = \{\{1, 2\}, \{3, 4\}, \ldots, \{2^k - 1, 2^k\}\}, \ \mathcal{B}_k = \{\{1, \ldots, 2^k\}\}.$$

By construction, we have $|\mathcal{B}_l| = 2^{k-l}$ and $|B| = 2^l$ for each $B \in \mathcal{B}_l$. For each $i \in [n]$ and $l = 0, \ldots, k$, denote by $B^l(i) \in \mathcal{B}_l$ the only set from $\mathcal{B}_l$ that contains i. In particular, $B^0(i) = \{i\}$ and $B^k(i) = [n]$.

For each $i \in [n]$ and every $l = 0, \ldots, k$ consider the random variables

$$g_i^l = g_i^l(Z_i, Z_{[n] \setminus B^l(i)}) = \mathbb{E}[g_i | Z_i, Z_{[n] \setminus B^l(i)}],$$

i.e. conditioned on Z_i and all the variables that are not in the same set as Z_i in the partition $\mathcal{B}_l$. In particular, $g_i^0 = g_i$ and $g_i^k = \mathbb{E}[g_i | Z_i]$. We can write a telescopic sum for each $i \in [n]$,

$$g_i - \mathbb{E}[g_i | Z_i] = \sum_{l=0}^{k-1} g_i^l - g_i^{l+1},$$

and the total sum of interest satisfies by the triangle inequality

$$\left\| \sum_{i=1}^n g_i \right\|_p \leq \left\| \sum_{i=1}^n \mathbb{E}[g_i | Z_i] \right\|_p + \sum_{l=0}^{k-1} \left\| \sum_{i=1}^n g_i^l - g_i^{l+1} \right\|_p. \tag{10}$$

Since $|\mathbb{E}[g_i | Z_i]| \leq M$ and $\mathbb{E}(\mathbb{E}[g_i | Z_i]) = 0$, by applying (9) we have

$$\left\| \sum_{i=1}^n \mathbb{E}[g_i | Z_i] \right\|_p \leq 4\sqrt{pn}M. \tag{11}$$

The only non-trivial part is the second term of the r.h.s. of (10). Observe that

$$g_i^{l+1}(Z_i, Z_{[n] \setminus B^{l+1}(i)}) = \mathbb{E}[g_i^l(Z_i, Z_{[n] \setminus B^l(i)}) | Z_i, Z_{[n] \setminus B^{l+1}(i)}],$$

that is, the expectation is taken w.r.t. the variables $Z_j, j \in B^{l+1}(i) \setminus B^l(i)$. It is also not hard to see that the function g_i^l preserves the bounded differences property, just like the the function g_i. Therefore, if we apply Lemma 2 conditioned on $Z_i, Z_{[n] \setminus B^{l+1}(i)}$, we obtain a uniform bound

$$\|g_i^l - g_i^{l+1}\|_p(Z_i, Z_{[n] \setminus B^{l+1}(i)}) \leq 2\sqrt{p2^l}\beta, \qquad \forall p \geq 2,$$

as there are 2^l indices in $B^{l+1}(i) \setminus B^l(i)$. We have as well $\|g_i^l - g_i^{l+1}\|_p \leq 2\sqrt{p2^l}\beta$ by (6).

Let us take a look at the sum $\sum_{i \in B^l} g_i^l - g_i^{l+1}$ for $B^l \in \mathcal{B}_l$. Since $g_i^l - g_i^{l+1}$ for $i \in B^l$ depends only on $Z_i, Z_{[n] \setminus B^l}$, the terms are independent and zero mean conditioned on $Z_{[n] \setminus B^l}$. Applying Lemma 3, we have for any $p \geq 2$,

$$\left\| \sum_{i \in B^l} g_i^l - g_i^{l+1} \right\|_p^p (Z_{[n] \setminus B^l}) \leq (3\sqrt{2p2^l})^p \frac{1}{2^l} \sum_{i \in B^l} \|g_i^l - g_i^{l+1}\|_p^p(Z_{[n] \setminus B^l}).$$

Integrating with respect to $(Z_{[n] \setminus B^l})$ and using $\|g_i^l - g_i^{l+1}\|_p \leq 2\sqrt{p2^l}\beta$, we have

$$\left\| \sum_{i \in B^l} g_i^l - g_i^{l+1} \right\|_p \leq 3\sqrt{2p2^l} \times 2\sqrt{p2^l}\beta = 6\sqrt{2}p2^l\beta.$$

It is left to use the triangle inequality over all sets $B^l \in \mathcal{B}_l$. We have

$$\left\| \sum_{i \in [n]} g_i^l - g_i^{l+1} \right\|_p \leq \sum_{B^l \in \mathcal{B}_l} \left\| \sum_{i \in B^l} g_i^l - g_i^{l+1} \right\|_p \leq 2^{k-l} \times 6\sqrt{2}p2^l\beta = 6\sqrt{2}p2^k\beta.$$

Recall, that $2^k < 2n$ due to the possible extension of the sample. Then

$$\sum_{l=0}^{k-1} \left\| \sum_{i=1}^{n} g_i^l - g_i^{l+1} \right\|_p \leq 12\sqrt{2}pn\beta\lceil\log_2 n\rceil .$$

Plugging the above bound together with (11) into (10), we get the required bound. ∎

Remark 5 *We remark that a version of Theorem 4 holds if $\mathbb{E}[g_i|Z_i]$ is sub-Gaussian for $i \in [n]$, but for the sake of presentation we focus on this simplified form.*

Remark 6 *The strategy of the proof of Theorem 4 is inspired by the original approach of Feldman and Vondrák. Their* clamping *can be related to the analysis of the terms $g_i^l - g_i^{l+1}$. It is important to notice that the truncation part of their analysis creates some technical difficulties since it introduces some bias and changes the stability parameter. In particular, the truncation brings an unnecessary logarithmic factor. We entirely avoid these steps by a simple application of the Marcinkiewicz-Zygmund inequality. The analog of the* dataset reduction *step of Feldman and Vondrák is our nested partition scheme. However, the recursive structure of their approach is replaced by an application of telescopic sums, whereas the union bound, which also brings a logarithmic factor, is replaced by the triangle inequality for L_p norms. Apart from a much shorter proof, our analysis leads to a better result: we eliminate the unnecessary $n\beta(\log n)^2$ term.*

3.2. Consequences for Uniformly Stable Algorithms

In order to make use of Theorem 4 to obtain generalization bounds, we will consider the following functions:

$$g_i = g_i(Z_1, \ldots, Z_n) = \mathbb{E}_{(X_i', Y_i')} \left(\mathbb{E}_{(X,Y)} \ell(A_{S^i}(X), Y) - \ell(A_{S^i}(X_i), Y_i) \right) , \qquad (12)$$

where we recall that

$$S^i = \{(X_1, Y_1), \ldots, (X_{i-1}, Y_{i-1}), (X_i', Y_i'), (X_{i+1}, Y_{i+1}), \ldots, (X_n, Y_n)\},$$

and $Z_i' = (X_i', Y_i')$ is an independent copy of (X_i, Y_i). The key observation is captured by the following simple lemma.

Lemma 7 *Under the uniform stability condition with parameter γ (1) and uniform boundedness of the loss function $\ell(\cdot, \cdot) \leq L$ we have for g_i defined by (12), that*

$$\left| |n(R(A_S) - R_{emp}(A_S))| - \left| \sum_{i=1}^{n} g_i \right| \right| \leq 2\gamma n.$$

Moreover, we have a. s. for $i \in [n]$, $|g_i| \leq L$ and $\mathbb{E}[\mathbb{E}[g_i|Z_{[n]\setminus i}]] = 0$.

Finally, as a deterministic function $g_i(z_1, \ldots, z_n)$ satisfies the bounded difference condition (8) with $\beta = 2\gamma$ for all except the i-th variable.

The proof is straightforward and we postpone it to Appendix B. Using this result, we can now obtain our main generalization bound.

Corollary 8 *Under the uniform stability condition* (1) *with parameter γ and the uniform bound-edness of the loss function $\ell(\cdot, \cdot) \leq L$, we have that for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$|n(R(A_S) - R_{emp}(A_S))| \lesssim n\gamma \log n \log\left(\frac{1}{\delta}\right) + L\sqrt{n \log\left(\frac{1}{\delta}\right)}.$$

The last bound is an improvement of the recent upper bound for uniformly stable algorithms by Feldman and Vondrák (4). To be precise, we removed the unnecessary $n\gamma(\log n)^2$ term.

Proof Combining Lemma 7 and Theorem 4 with g_i defined in (12), $M = L$, and $\beta = 2\gamma$, we have for any $p \geq 2$,

$$\|n(R - R_{emp})\|_p \lesssim pn\gamma \log n + L\sqrt{pn}.$$

The deviation bound now follows immediately from Lemma 1. ∎

4. Lower Bounds

Since the bound of Theorem 4 implies the best known risk bound, it is natural to ask if it can be improved in general. By Lemma 7, we know that the analysis of the generalization bounds is closely related to the analysis of the functions satisfying the assumptions of Theorem 4. Therefore, it is interesting to know how sharp the general bound (17) is. Recall that

$$\left\|\sum_{i=1}^{n} g_i(Z)\right\|_p \lesssim (pn\beta \log n + M\sqrt{pn}) \wedge nL,$$

where, as before, L is a uniform bound on $|g_i|$. In this section, we prove that one can not improve the bound of Theorem 4, apart from the $\log n$-factor, and the bound is tight with respect to the parameters $M, \beta, n, \log\frac{1}{\delta}$ in some regimes. We notice, however, that this does not completely answer the question of the optimality of the risk bound of Corollary 8 for uniformly stable algorithms, but shows that this is the best we can hope for as long as our upper bound is based only on the parameters $M, \beta, n, \log\frac{1}{\delta}$. In particular, Theorem 4 disregards the condition $|g_i| \leq L$. We discuss this in more detail in what follows.

Proposition 9 (The lower bound, $p \leq n$) *Let $Z_1, \ldots, Z_n$ be i.i.d. Rademacher signs. There is an absolute constant $\kappa > 0$ and functions $g_i : \{-1, 1\}^n \to \mathbb{R}$ that satisfy the conditions of Theorem 4 with the parameters β, M, such that we have for any $\kappa \leq p \leq n$,*

$$\left\|\sum_{i=1}^{n} g_i(Z_1, \ldots, Z_n)\right\|_p \gtrsim pn\beta + M\sqrt{pn}. \tag{13}$$

The proof is based on a moment version of the Montgomery-Smith bound (originally from (Montgomery-Smith, 1990)) which is due to Hitczenko (1993). We postpone the proof to Appendix C. We note additionally that the case $M = 0$ follows immediately from Corollary 1, Example 2 by Latała (1999).

The lower bound of Proposition 9 matches the result of Theorem 4 up to the logarithmic factor in the regime $p \leq n$. In particular, it means that in this regime, the bound has to be sub-exponential

unless we use some properties of the functions g_i, other than mentioned in Theorem 4. We additionally note that our moment lower bounds imply the deviation lower bounds. We can show that there are absolute constants $c_1, c_2 > 0$ such that the functions defined in (19) satisfy for any $\delta \in (e^{-c_1 n}, 1)$,

$$\mathbb{P}\left(\left|\sum_{i=1}^n g_i\right| \lesssim n\beta \log\left(\frac{1}{\delta}\right) + M\sqrt{n \log\left(\frac{1}{\delta}\right)}\right) \leq 1 - \min(c_2, \delta). \tag{14}$$

This bound can be derived through the Paley-Zygmund inequality, e.g., using the standard arguments as in (Gluskin and Kwapień, 1995). For the sake of completeness, we derive this inequality in Appendix D.

Besides, when $p > n$, a trivial bound $\left\|\sum_{i=1}^n g_i\right\|_p \lesssim nL$ is the best one can have. Like in (20), consider the functions $g_i = LZ_i$, where Z_i are i.i.d. Rademacher signs (it corresponds to a learning algorithm that always outputs the same classifier). By Lemma 10 we have for $p > n$,

$$\left\|\sum_{i=1}^n g_i\right\|_p = L\left\|\sum_{i=1}^n Z_i\right\|_p \gtrsim Ln.$$

5. Discussions

On the Sub-Optimality of (2)   At first, we prove an exact moment analog of (2) for $\left|\sum_{i=1}^n g_i\right|$, for g_i defined by (12). We have in mind the illustrative regime of $L = 1$ and $\gamma = \frac{1}{\sqrt{n}}$, this is exactly when the bound (4) balances the two terms. By the triangle inequality we have

$$\left\|\sum_{i=1}^n g_i\right\|_p \leq \sum_{i=1}^n \|g_i - \mathbb{E}[g_i|Z_i]\|_p + \left\|\sum_{i=1}^n \mathbb{E}[g_i|Z_i]\right\|_p$$
$$\leq 4n\sqrt{np}\gamma + 4\sqrt{pn}L.$$

where we used that conditioned on Z_i the random variable $g_i - \mathbb{E}[g_i|Z_i]$ is centered and combined this fact with Lemma 2. Since $\sum_{i=1}^n \mathbb{E}[g_i|Z_i]$ is a sum of independent centered bounded random variables, Hoeffding's inequality (9) is applicable to $\left\|\sum_{i=1}^n \mathbb{E}[g_i|Z_i]\right\|_p$.

Observe that we lose a lot by replacing $\left\|\sum_{i=1}^n (g_i - \mathbb{E}[g_i|Z_i])\right\|_p$ with $\sum_{i=1}^n \|g_i - \mathbb{E}[g_i|Z_i]\|_p$. Indeed, it is easy to see that the random variables g_i, as well as $g_i - \mathbb{E}[g_i|Z_i]$, are weakly correlated. For $i \neq j$, using $\mathbb{E}[g_i|Z_{[n]\setminus\{i\}}] = 0$ and $\mathbb{E}[g_j|Z_{[n]\setminus\{j\}}] = 0$ together with the bounded difference property, we have

$$|\mathbb{E}g_i g_j| = |\mathbb{E}(g_i(Z) - g_i(Z^j))(g_j(Z) - g_j(Z^i))| \leq 4\gamma^2. \tag{15}$$

This suggests that for $\gamma = \frac{1}{\sqrt{n}}$, the random variables g_i and g_j have small correlation. However, the original argument in (Bousquet and Elisseef, 2002) does not take this into account and would give the same bound even if all $g_i - \mathbb{E}[g_i|Z_i]$ were replaced by the same random variable $g_1 - \mathbb{E}[g_1|Z_1]$.

A Better Second Moment Bound Actually it turns out that using the property of weak correlation between the g_i functions one can obtain a tight upper bound on the second moment. Indeed, (15) with $\gamma = \beta/2$ together with the bounded difference property for g_i readily gives

$$\left\| \sum_{i=1}^{n} g_i \right\|_2 = \sqrt{\sum_{i \neq j} \mathbb{E} g_i g_j + \sum_{i=1}^{n} \mathbb{E} g_i^2} \leq n\beta + \sqrt{n} \max_{i \in [n]} \|g_i\|_2$$

$$\leq n\beta + \sqrt{n} \max_{i \in [n]} \left(\|g_i - \mathbb{E}[g_i | Z_i]\|_2 + \|\mathbb{E}[g_i | Z_i]\|_2 \right)$$

$$\leq (1 + 2\sqrt{2})n\beta + \sqrt{n}M.$$

We also note that an argument analogous to (15) was first used in (Feldman and Vondrák, 2018) to prove the following variance bound

$$\text{Var}(n(R(A_S) - R_{\text{emp}}(A_S))) \lesssim n^2 \gamma^2 + nL^2. \tag{16}$$

Note that our second moment bound is a little better as it uses M instead of L.

Recovering (3) Another interesting direction is the analysis of the first bound of Feldman and Vondrák (3), which was originally proved by the techniques taking their roots from Differential Privacy (see the discussions on various ways to prove this bound in (Feldman and Vondrák, 2019)). As already noticed in (Feldman and Vondrák, 2019), the bound (3) should not be discarded due to the fact that it does not contain additional $\log n$-factors and, more importantly, it has the *sub-gaussian* form, since it depends only on $\sqrt{\log \frac{1}{\delta}}$. Recall that the second bound (4) is a mixture of *sub-gaussian* $\sqrt{\log \frac{1}{\delta}}$ and *sub-exponential* $\log \frac{1}{\delta}$ tails. Although we can adapt the moment technique to prove (3), we instead come to the following more general observation:

The bound of Theorem 4 is strong enough to almost recover the sub-gaussian bound (3).

In order to show this, we have by Theorem 4, provided that $|g_i| \leq L$ almost surely

$$\left\| \sum_{i=1}^{n} g_i(Z) \right\|_p \leq \left(12\sqrt{2}pn\beta \lceil \log_2 n \rceil + 4M\sqrt{pn} \right) \wedge nL. \tag{17}$$

Since $M \leq L$ and for $a, b \geq 0$, $a \wedge b \leq \sqrt{ab}$ (which is rather crude) we have for $p \leq n$,

$$\left\| \sum_{i=1}^{n} g_i(Z) \right\|_p \leq (12\sqrt{2}pn\beta \lceil \log_2 n \rceil \wedge Ln) + 4L\sqrt{pn}$$

$$\lesssim n\sqrt{p\beta L \log n} + L\sqrt{pn}.$$

Similarly to the proof of Corollary 8, it immediately implies

$$n|R(A_S) - R_{\text{emp}}(A_S)| \lesssim (n\sqrt{\gamma L \log n} + L\sqrt{n})\sqrt{\log \frac{1}{\delta}}, \tag{18}$$

which is (3) up to an unnecessary $\sqrt{\log n}$ factor. The latter is clearly an artifact of the proof in our case.

Interpolation Algorithms   We observe here that our results can potentially be applied to interpolation algorithms. As mentioned earlier, in this case, since $R_{emp}(A_S) = 0$ one cannot hope to get any non-trivial bound on the generalization error (since there are distributions for which $\mathbb{E}R(A_S)$ cannot go to zero). The idea is then to use another estimator of the risk, namely the leave-one-out estimator, which can be defined as

$$R_{loo}(A_S) = \frac{1}{n} \sum_{i=1}^{n} \ell\left(A_{S^{\backslash i}}(X_i), Y_i\right) ,$$

where $S^{\backslash i}$ is $S \setminus \{(X_i, Y_i)\}$. Then an analogue of Lemma 7 can be obtained with R_{loo} instead of R_{emp}. Note that one also needs to modify the notion of uniform stability to exclude the case where the test point is in the training set. However, proving non-trivial stability results for realistic interpolation algorithms or extending our results to use something like hypothesis stability (which is an in-expectation stability instead of the worst-case one) would still require more work.

6. Open Questions

We have presented some progress towards narrowing down the exact behaviour of the difference between true and empirical risk for uniformly stable algorithms with almost matching upper and lower bounds. But there are a few remaining open questions regarding the tightness of our results. For the upper bound, since it is possible to get rid of the factor $\log n$ in the second moment, as demonstrated by (16), there is hope that this factor can be completely removed for $p > 2$ in Theorem 4 as well as in (5).

For the lower bound, it would be interesting to find a learning algorithm with uniformly bounded functions $|g_i| \leq L$ and a generalization bound matching the high probability lower bound (14). Indeed, Proposition 9 shows that the high probability upper bound of Theorem 4 can not be improved in general (apart from the logarithmic factor). At the same time, our bound does not take into account that L can actually be of the same order as M, which happens in the context of uniformly stable learning algorithms. In particular, the example of Proposition (14) has $L = \beta(n-1)$, which can be much larger than M in the regime of $\beta = \frac{1}{\sqrt{n}}$. Therefore, in this context it will be interesting to find (if it is possible at all) a lower bound of the form

$$\left\| \sum_{i=1}^{n} g_i(Z_1, \ldots, Z_n) \right\|_p \gtrsim pn\beta + L\sqrt{pn},$$

for the functions g_i corresponding to a uniformly stable algorithm.

Finally, some extensions of Lemma 2 are known in the unbounded case (see e.g., (Kontorovich, 2014)). This may lead to the generalization bounds for the unbounded loss functions using our Theorem 4. Following Feldman and Vondrák (2019), there is a renewed interest in the use of stability bounds for studying learning algorithms, for example (Foster, Greenberg, Kale, Luo, Mohri, and Sridharan, 2019), whose results can possibly be further improved using the techniques we presented.

Acknowledgments

References

Mikhail Belkin, Daniel J Hsu, and Partha Mitra. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. In *Advances in Neural Information Processing Systems*, pages 2300–2311, 2018.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Olivier Bousquet and Andre Elisseef. Stability and generalization. *Journal of Machine Learning Research*, 2(Mar):499–526, 2002.

Olivier Bousquet, Steve Hanneke, Shay Moran, and Nikita Zhivotovskiy. Proper learning, Helly number, and an optimal SVM bound. *arXiv:2005.11818. To appear in proceedings of COLT*, 2020.

L. Devroye and T. J. Wagner. Distribution-free performance bounds with the resubstitution error estimate. *IEEE Trans. Information Theory*, 25(2):601–604, 1979.

Vitaly Feldman and Jan Vondrák. Generalization bounds for uniformly stable algorithms. *In proceedings of NeurIPS*, 2018.

Vitaly Feldman and Jan Vondrák. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. *arxiv:1902.10710. Extended abstract in proceedings of COLT*, 2019.

Dylan J Foster, Spencer Greenberg, Satyen Kale, Haipeng Luo, Mehryar Mohri, and Karthik Sridharan. Hypothesis set stability and generalization. *arXiv:1904.04755*, 2019.

Efim D Gluskin and Stanisław Kwapień. Tail and moment estimates for sums of independent random variables with logarithmically concave tails. *Studia Mathematica*, 3(114):303–309, 1995.

Moritz Hardt, Benjamin Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. *In proceedings of ICML*, pages 1225–1234, 2016.

Paweł Hitczenko. Domination inequality for martingale transforms of a Rademacher sequence. *Israel Journal of Mathematics*, 84(1-2):161–178, 1993.

Aryeh Kontorovich. Concentration in unbounded metric spaces and algorithmic stability. In *proceedings of ICML*, pages 28–36, 2014.

Rafał Latała. Tail and moment estimates for some types of chaos. *Studia Mathematica*, 135(1): 39–53, 1999.

Andreas Maurer. A second-order look at stability and generalization. *In proceedings of COLT*, 2017.

Stephen J Montgomery-Smith. The distribution of Rademacher sums. *In proceedings of the American Mathematical Society*, 109(2):517–522, 1990.

Yao-Feng Ren and Han-Ying Liang. On the best constant in Marcinkiewicz–Zygmund inequality. *Statistics & Probability Letters*, 53(3):227–233, 2001.

W. H. Rogers and T. J. Wagner. A finite sample distribution-free performance bound for local discrimination rules. *The Annals of Statistics*, 6(3):506–514, 1978.

Vladimir Vapnik and Alexey Chervonenkis. *Theory of Pattern Recognition*. Nauka. Moscow., 1974.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, 2016.

Yu-Xiang Wang, Jing Lei, and Stephen E. Fienberg. Learning with Differential Privacy: Stability, learnability and the sufficiency and necessity of ERM principle. *Journal of Machine Learning Research*, 17(183):1–40, 2016.

Nikita Zhivotovskiy. Optimal learning via local entropies and sample compression. *In proceedings of COLT*, 2017.

Appendix A. Proof of Lemma 1

It is easy to see that the random variable $(|Y| - a - b)_+$ satisfies the requirements of Theorem 2.3 in (Boucheron et al., 2013). Therefore, we have for each integer $p \geq 1$,

$$\|(|Y| - a - b)_+\|_{2p} \leq \sqrt[2p]{p!(2a)^{2p} + (2p)!(4b)^{2p}}$$
$$\leq \sqrt{2p}a + 4pb,$$

and the first bound follows from the trivial $\|Y\|_p \leq \|(|Y| - a - b)_+\|_{2p} + a + b$.

On the other hand, if $\|Y\|_p \leq \sqrt{p}a + pb$ for any $p \geq 1$, by Markov's inequality for any $\delta \in (0, 1)$,

$$\mathbb{P}\left(|Y| > \|Y\|_p e^{\frac{\log\left(\frac{1}{\delta}\right)}{p}}\right) \leq \left(\frac{\|Y\|_p}{\|Y\|_p e^{\frac{\log\left(\frac{1}{\delta}\right)}{p}}}\right)^p = \delta.$$

It remains to pick $p = \log\left(\frac{e}{\delta}\right) \geq 1$, so that the exponent disappears and we get the bound. ∎

Appendix B. Proof of Lemma 7

By uniform stability, we can write the following decomposition

$$|n(R(A_S) - R_{\text{emp}}(A_S))| = \left|\sum_{i=1}^{n} \left(\mathbb{E}_{(X,Y)}\ell(A_S(X), Y) - \ell(A_S(X_i), Y_i)\right)\right|$$
$$\leq 2\gamma n + \left|\sum_{i=1}^{n} \mathbb{E}_{(X_i', Y_i')}\left(\mathbb{E}_{(X,Y)}\ell(A_{S^i}(X), Y) - \ell(A_{S^i}(X_i), Y_i)\right)\right| = 2\gamma n + \left|\sum_{i=1}^{n} g_i\right|.$$

Similarly to the computations above we have

$$|n(R(A_S) - R_{\text{emp}}(A_S))| \geq \left| \sum_{i=1}^{n} g_i \right| - 2\gamma n.$$

The remaining properties can be immediately verified. ∎

Appendix C. Proof of Proposition 9

As before, we need two well-known facts from probability theory. The first lemma is a moment version of the Montgomery-Smith bound (originally from (Montgomery-Smith, 1990)) which is due to Hitczenko (1993). It characterizes the moments of Rademacher sums up to a multiplicative constant factor.

Lemma 10 (**Moments of weighted Rademacher sums, (Hitczenko, 1993)**) *Let $a_1 \geq \ldots \geq a_n$ be a non-increasing sequence of non-negative numbers and let $\varepsilon_1, \ldots, \varepsilon_n$ denote i.i.d. Rademacher signs. Then, it holds that*

$$\left\| \sum_{i=1}^{n} a_i \varepsilon_i \right\|_p \sim \sum_{i=1}^{p} a_i + \sqrt{p} \left(\sum_{i=p+1}^{n} a_i^2 \right)^{1/2}.$$

The next lemma is Chebyshev's association inequality, see e.g., Theorem 2.14 in (Boucheron et al., 2013).

Lemma 11 *Let f and h be non-decreasing real-valued functions defined on the real line. If X is a real-valued random variable, then*

$$\mathbb{E}g(X)h(X) \geq \mathbb{E}g(X)\mathbb{E}h(X).$$

Consider the functions

$$g_i = g_i(Z_1, \ldots, Z_n) = MZ_i + \frac{\beta}{2} Z_i \left(\sum_{j \neq i} Z_j \right). \tag{19}$$

It is easy to check that $\mathbb{E}[g_i | Z_{[n] \setminus \{i\}}] = 0$ and $\mathbb{E}[g_i | Z_i] = M$ a.s. Moreover, each g_i satisfies the bounded difference property with parameter β w.r.t. all except the i-th variable. Denoting $S = \sum_{i=1}^{n} Z_i$ we have

$$\sum_{i=1}^{n} g_i = MS + \frac{\beta}{2} S^2 - \frac{\beta}{2} n. \tag{20}$$

By the triangle inequality we have

$$\left\| \sum_{i \leq n} g_i \right\|_p \geq \left\| MS + \frac{\beta}{2} S^2 \right\|_p - \frac{\beta}{2} n.$$

For $S_+ = S\mathbf{1}[S \geq 0]$ we obviously have $\|MS + \beta S^2/2\|_p \geq \|MS_+ + \beta S_+^2/2\|_p$. By Lemma 11 and since both functions x^p and $(M + \beta x/2)^p$ are non-decreasing for non-negative x, we have

$$\|MS_+ + \beta S_+^2/2\|_p \geq \|S_+\|_p\|M + \beta S_+/2\|_p \geq \|S_+\|_p(M \vee \beta\|S_+\|_p/2).$$

Finally, due to the symmetry of S and Lemma 10, we have for $p \leq n$,

$$\|S_+\|_p \geq \frac{1}{2}\|S\|_p \gtrsim \sqrt{pn}.$$

Therefore, for some $c > 0$, our construction implies the following lower bound

$$\left\|\sum_{i=1}^n g_i\right\|_p \geq c\sqrt{pn}(M + \beta\sqrt{pn}/2) - n\beta/2 = (cp - 1)\beta n/2 + \sqrt{pn}M.$$

∎

Appendix D. Proof of the Lower Tail (14)

By the Paley-Zygmund inequality (see e.g., (Gluskin and Kwapień, 1995)) we have for $f = \sum_{i=1}^n g_i$, where g_i are defined by (19),

$$\mathbb{P}\left(|f| \geq \frac{1}{2}\|f\|_p\right) = \mathbb{P}\left(|f|^p \geq \frac{1}{2^p}\mathbb{E}|f|^p\right) \geq \left(1 - \frac{1}{2^p}\right)\frac{(\mathbb{E}|f|^p)^2}{\mathbb{E}|f|^{2p}}$$

$$\geq \left(1 - \frac{1}{2^p}\right)\left(\frac{\|f\|_p}{\|f\|_{2p}}\right)^{2p} \geq \left(\frac{\|f\|_p^2}{2\|f\|_{2p}^2}\right)^p.$$

In Proposition 9, we derived a lower bound $\|f\|_p \gtrsim pn\beta + M\sqrt{pn}$ for $\kappa \leq p \leq n$. In our case, it is also not hard to get a matching upper bound without the logarithm via Latała's bound for Rademacher chaos. We note that if we are only interested in the upper bound, we can alternatively use a moment version of the Hanson-Wright inequality (see e.g., (Vershynin, 2016)).

Lemma 12 (Corollary 1 and Example 2 in (Latała, 1999)) *Let $(a_{i,j})$ be an $n \times n$ real symmetric matrix with $a_{i,i} = 0$ for all $i \in [n]$ and let $\varepsilon_1, \ldots, \varepsilon_n$ denote i.i.d. Rademacher signs. We have that for any $p \geq 1$,*

$$\left\|\sum_{i,j} a_{i,j}\varepsilon_i\varepsilon_j\right\|_p \sim \sup\left\{\sum_{i,j} a_{i,j}x_iy_j : \|x\|_2, \|y\|_2 \leq \sqrt{p}, \|x\|_\infty, \|y\|_\infty \leq 1\right\}$$

$$+ \sum_{i=1}^p A_i^* + \sqrt{p}\left(\sum_{i>p}^n (A_i^*)^2\right)^{1/2},$$

where A_i^* denotes the non-decreasing rearrangement of the sequence $A_i = \left(\sum_j^n a_{i,j}^2\right)^{1/2}$.

Using Lemma [12](#) we can easily control $\left\|\sum_{i \neq j} Z_i Z_j\right\|_p$ for $p \leq n$. Observe that in our case $a_{i,j} = 1$ for $i \neq j$ and $A_i = \sqrt{n-1}$ for $i \in [n]$. It implies that for any $p \geq 1$,

$$\left\|\sum_{i \neq j} Z_i Z_j\right\|_p \lesssim pn + p\sqrt{n} + \sqrt{p}n, \tag{21}$$

where for the first term in (21) we used $\sup\left\{\sum_{i \neq j} x_i y_j : \|x\|_2, \|y\|_2 \leq \sqrt{p}\right\} \leq pn$. Finally, we have for any $p \geq 1$,

$$\|f\|_p \leq M \left\|\sum_{i=1}^n Z_i\right\|_p + \frac{\beta}{2}\left\|\sum_{i \neq j} Z_i Z_j\right\|_p \lesssim M\sqrt{pn} + pn\beta.$$

Since $\frac{a\sqrt{p}+bp}{a\sqrt{2p}+2bp} \geq \frac{1}{2}$, we have that for any p such that $\kappa \leq p \leq n/2$, it holds that $\|f\|_p \geq c\|f\|_{2p}$, where $c \leq 1$ is an absolute constant. Therefore,

$$\mathbb{P}\left(|f| \geq \frac{1}{2}\|f\|_p\right) \geq \left(\frac{c^2}{2}\right)^p.$$

Let us pick $p = \kappa \vee p_\delta$, where $p_\delta = (\log 2c^{-2})^{-1} \log\left(\frac{1}{\delta}\right)$, so that $(c^2/2)^{p_\delta} = \delta$. Assuming that $\delta \geq e^{-(\log 2c^{-2})^{-1}n}$, we have $p_\delta \leq n$. Since $\kappa \leq p \leq n$ and $c^2/2 < 1$, we have

$$\mathbb{P}\left(|f| \geq \|f\|_{p_\delta}\right) \geq \mathbb{P}\left(|f| \geq \|f\|_p\right) \geq \left(\frac{c^2}{2}\right)^\kappa \wedge \delta.$$

It is left to notice that our lower bound implies

$$\|f\|_{p_\delta} \gtrsim n\beta \log\left(\frac{1}{\delta}\right) + M\sqrt{n\log\left(\frac{1}{\delta}\right)}.$$

Therefore, (14) holds for $c_1 = (\log 2c^{-2})^{-1}$ and $c_2 = (c^2/2)^\kappa$.   ∎