Journal of Machine Learning Research 25 (2024) 1-33 Submitted 6/23; Revised 5/24; Published 9/24

Concentration and Moment Inequalities for General Functions of Independent Random Variables with Heavy Tails

Shaojie Li LISHAOJIE95@RUC.EDU.CN
Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China
Beijing Key Laboratory of Big Data Management and Analysis Methods, Beijing, China

Yong Liu* LIUYONGGSAI@RUC.EDU.CN
Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China
Beijing Key Laboratory of Big Data Management and Analysis Methods, Beijing, China

Editor: Benjamin Guedj

Abstract

The concentration of measure phenomenon serves an essential role in statistics and machine learning. This paper gives bounded difference-type concentration and moment inequalities for general functions of independent random variables with heavy tails. A general framework is presented, which can be used to prove inequalities for general functions once the moment inequality for sums of independent random variables is established. We illustrate the power of the framework by showing how it can be used to derive novel concentration and moment inequalities for bounded, Bernstein's moment condition, weak-exponential, and polynomial-moment random variables. Furthermore, we give potential applications of these inequalities to statistical learning theory.

Keywords: Concentration inequality, bounded difference, statistical learning theory

1. Introduction

Concentration and moment inequalities are at the heart of empirical science and form an essential toolkit in the study of natural and artificial learning systems (Boucheron et al., 2003, 2005, 2013). They have been studied for several decades and used in various areas, including convex geometry, functional analysis, statistical physics, probability theory, statistics, information theory, communications and coding theory, learning theory, and computer science (Ledoux, 2001; Raginsky and Sason, 2015, 2018).

The bounded difference inequality, also referred to as McDiarmid's inequality (McDiarmid, 1998), is one of the most popular concentration inequalities, which has been widely employed, as a powerful tool, in machine learning theory, such as algorithmic stability (Bousquet and Elisseeff, 2002; Bousquet et al., 2020) and empirical processes (Bartlett and Mendelson, 2002; Bartlett et al., 2005). Compared to the general Hoeffding- and Bernstein-type inequality (Vershynin, 2018; Wainwright, 2019), the bounded difference inequality works not only for sums but for general functions of independent random variables, which is more flexible and capable of estimating nonlinear statistics (Maurer, 2019; Maurer and Pontil, 2018, 2019).

*. Correspodning Author

Specifically, the bounded difference inequality (McDiarmid, 1998) states that

$$\mathbb{P}\left(f(X) - \mathbb{E}[f(X)] > t\right) \leq \exp\left(\frac{-2t^2}{\sum_k c_k^2}\right) \quad \forall t \geq 0,$$

where f is a real-valued function of the sequence of independent random variables $X = (X_1, ..., X_n)$, such that $|f(x) - f(x')| \leq c_k$ whenever x and x' differ only in the k-th coordinate.

Although it is pretty attractive and useful, the bounded difference inequality, however, requires the conditional ranges to be uniformly bounded, which imposes inherent limitations on their applicability to unbounded loss functions. The concentration properties of unbounded functions become important in many settings, for instance, there has been a lot of work concerned about establishing generalization bounds in unbounded settings (Cortes et al., 2021; Kontorovich, 2014; Meir and Zhang, 2003; Cortes et al., 2019; Lou et al., 2022), especially for the PAC-Bayes learning (Alquier, 2008; Haddouche and Guedj, 2023; Haddouche et al., 2021; Holland, 2019; Rivasplata et al., 2020; Casado et al., 2024; Chugg et al., 2023). To meet the growing demand, several concentration inequalities for general functions of unbounded random variables have been proposed (Kutin, 2002; Combes, 2015; Meir and Zhang, 2003; Kontorovich, 2014; Maurer and Pontil, 2021). Among these works, Kutin (2002); Kutin and Niyogi (2002) prove two extensions for strongly and weakly difference-bounded functions. Combes (2015) proposes a somewhat different extension for functions with bounded differences on a high probability set and no restriction outside this set. Although interesting, these approaches entail complex statement, and their conditions are too restrictive in practice, see a discussion in (Kontorovich, 2014). In the related work, Warnke (2016) also proposed an interesting variant of the bounded difference inequality by relaxing the worst-case changes c_k to typical changes, which has proven useful in a number of combinatorial applications (where it's often possible to distinguish between the typical- and worst-case changes), often leading to easy concentration proofs beyond the classical bounded differences inequality. However, the results of Warnke (2016) are still in the realm of bounded random variables. It may happen that the conditional ranges are infinite, but that the conditional versions (the random variables obtained by fixing all but one of the arguments of the function) have certain decay tails (Maurer and Pontil, 2021). In this context, Meir and Zhang (2003); Kontorovich (2014) give inequalities for sub-Gaussian distributions. Recently, Maurer and Pontil (2021) provide a more applicable inequality than the ones in (Meir and Zhang, 2003; Kontorovich, 2014) for the sub-Gaussian case and further study the heavier sub-exponential distributions, whose results can be seen as unbounded analogues of the bounded difference inequality under the sub-Gaussian and sub-exponential conditions.

However, both the sub-Gaussian and sub-exponential distributions are relatively light-tailed. The two distributions are characterized by their tails being upper bounded by Gaussian, respectively exponential, tails (Vladimirova et al., 2020; Wainwright, 2019). A distinctive difference between heavy-tailed distributions and sub-Gaussian and sub-exponential distributions is the moment generating function (MGF). The MGF exists in a neighborhood around zero for sub-Gaussian and sub-exponential distributions (Vershynin, 2018), while it does not exist for heavy-tailed distributions (Foss et al., 2011; Bakhshizadeh et al., 2023). Therefore, the technique to find upper bounds for the MGF, used in (Meir and Zhang, 2003; Kontorovich, 2014; Maurer and Pontil, 2021), fails for heavy-tailed distributions. However, in many applications, such as probability theory (Wong et al., 2020),

high-dimensional statistics (Kuchibhotla and Chakrabortty, 2022; Guédon et al., 2014), stochastic optimization (Gurbuzbalaban et al., 2021) and signal processing (Bakhshizadeh et al., 2020), the assumption of light-tailed sub-Gaussian and sub-exponential distributions appears to be inappropriate. Except for the concentration inequalities, recent developments in random combinatorics, statistics and empirical process theory have prompted the search to moment inequalities dealing with heavy-tailed random variables (Boucheron et al., 2013). While for bounded difference-type moment inequalities, the relevant results are very few and less than the ones of concentration inequalities. Therefore, for the sake of growing demand, we need concentration and moment inequalities for general functions of independent random variables with heavy tails.

The goal of this paper is to provide such general-purpose inequalities. Following the related work, we consider that the centered conditional versions have certain decay tails. We first provide a general framework, which can be used to establish moment inequalities for general functions of independent random variables once the moment inequality for sums of independent variables is obtained. In contrast to existing works (Maurer and Pontil, 2021; Kontorovich, 2014) that directly focus on concentration inequalities, we explore inequalities from a moment perspective. It should be noted that the moment perspective has many advantages, which enhance the flexibility of our framework and ease the derivation of both concentration and moment inequalities for heavy-tailed distributions, even those with potentially infinite variance. Then, we demonstrate the strength of this framework by applying it to bounded, Bernstein's moment condition, weak-exponential, and polynomial-moment random variables, where both the concentration and moment inequalities are presented. In the last, we examine the application of our derived concentration inequalities to some standard problems in statistical learning theory, including vector valued concentration, Rademacher complexity and generalization, and algorithmic stability and generalization. Together with the bounded difference inequality, Rademacher complexity and algorithmic stability are two fundamental tools to derive generalization bounds for various learning algorithms, but this approach typically requires the assumption of boundedness. Our inequalities significantly broaden the applicability of these results to heavy-tailed distributions.

In conclusion, this paper's contributions are threefold: presenting a general framework, deriving concentration and moment inequalities for a range of random variables, and applying these inequalities to statistical learning theory. The paper is organized as follows: Section 2 introduces the general framework. Section 3 explores the application of this framework to specific random variables, while Section 4 demonstrates the application of the derived concentration inequalities to statistical learning theory. The paper concludes with Section 5, and proofs are provided in the Appendix.

2. A General Framework

We first introduce some necessary notations, and then outline the framework.

2.1 Notations

We use uppercase letters to present random variables and vector of random variables, and use lowercase letters to present scalars and vector of scalars. Let $X = (X_1, ..., X_n)$ be a vector of independent random variables with values in a space $\mathcal{X}$, and the vector $X' = (X'_1, ..., X'_n)$

is independent and identically distributed (i.i.d.) to X. Let f be a function $f : \mathcal{X}^n \mapsto \mathbb{R}$. We will also need the following definition to characterize the fluctuations of f in the k-th variable X_k, when the other variables $(x_i : i \neq k)$ are given.

Definition 1 *If $f : \mathcal{X}^n \mapsto \mathbb{R}$, $x = (x_1, ..., x_n) \in \mathcal{X}^n$ and $X = (X_1, ..., X_n)$ is a random vector with independent components in $\mathcal{X}^n$, then the k-th centered conditional version of f is the random variable*

$$f_k(X)(x) = f(x_1, ..., x_{k-1}, X_k, x_{k+1}, ..., x_n) - \mathbb{E}\left[f\left(x_1, ..., x_{k-1}, X'_k, x_{k+1}, ..., x_n\right)\right].$$

$f_k(X)$ can be seen as a random-variable-valued-function $f_k(X) : x \in \mathcal{X}^n \to f_k(X)(x)$. Therefore,

$$f_k(X)(X) = f(X) - \mathbb{E}[f(X)|X_1, ..., X_{k-1}, X_{k+1}, ..., X_n].$$

It is clear that $f_k(X)$ does not depend on the k-th coordinate of x. For instance, consider the sum $f(x) = \sum_{i=1}^n x_i$, then $f_k(X)(x) = X_k - \mathbb{E}[X_k]$ is independent of x. The L_p norm of a real random variable Z is $\|Z\|_p = (\mathbb{E}[|Z|^p])^{\frac{1}{p}}$.

2.2 Main Results

A general framework is given, which is a moment inequality for general functions of n independent random variables. To proceed, we state two technical lemmas.

Lemma 2 *Let $h : \mathbb{R} \to \mathbb{R}$ be a convex function, $\epsilon_1, ..., \epsilon_n$ a sequence of independent Rademacher variables (i.e., with $\mathbb{P}(\epsilon_i = -1) = \mathbb{P}(\epsilon_i = 1) = 1/2$) and $a_1, ..., a_n$, $b_1, ..., b_n$ two sequences of real numbers, such that for every i $|a_i| \leq |b_i|$. Then*

$$\mathbb{E}h\left(\sum_{i=1}^n |a_i|\epsilon_i\right) \leq \mathbb{E}h\left(\sum_{i=1}^n |b_i|\epsilon_i\right).$$

Proof It is enough to prove the monotonicity of function $f(t) = \mathbb{E}h(a + |t|\epsilon_1)$, for every choice of the parameter a. By the convexity assumption we have for $|s| < |t|$

$$\frac{h(a + |t|) - h(a + |s|)}{|t| - |s|} \geq \frac{h(a - |s|) - h(a - |t|)}{|t| - |s|}.$$

Equivalently,

$$f(|s|) = \frac{1}{2}(h(a + |s|) + h(a - |s|)) \leq \frac{1}{2}(h(a + |t|) + h(a - |t|)) = f(|t|).$$

The proof is complete. ■

Lemma 3 *Let $h : \mathbb{R} \to \mathbb{R}$ be a convex function and $S = f(X_1, ..., X_{i-1}, X_i, X_{i+1}, ..., X_n)$, where $X_1, ..., X_n$ are independent random variables with values in a measurable space $\mathcal{X}$ and*

$f : \mathcal{X}^n \to \mathbb{R}$ is a measurable function. Denote as usual $S_i = f(X_1, ..., X_{i-1}, X'_i, X_{i+1}, ..., X_n)$, where $(X'_1, ..., X'_n)$ is an independent copy of $(X_1, ..., X_n)$. Assume moreover that

$$|S - S_i| \leq F_i(x_1, ..., x_{i-1}, X_i, X'_i, x_{i+1}, ..., x_n)$$

for some functions $F_i : \mathcal{X}^{n+1} \to \mathbb{R}$ that does not depend on the i-th coordinate of x, $i = 1, ..., n$, and some fixed $x \in \mathcal{X}^n$. Then,

$$\mathbb{E}h(S - \mathbb{E}S) \leq \mathbb{E}h\left(\sum_{i=1}^n \epsilon_i F_i(x_1, ..., x_{i-1}, X_i, X'_i, x_{i+1}, ..., x_n)\right),$$

where $\epsilon_1, ..., \epsilon_n$ is a sequence of independent Rademacher variables, independent of $(X_i)_{i=1}^n$ and $(X'_i)_{i=1}^n$.

Proof We will use induction with respect to n. For $n = 0$ the statement is obvious, since $\mathbb{E}h(S - \mathbb{E}S) = \mathbb{E}h\left(\sum_{i=1}^n \epsilon_i F_i(x_1, ..., x_{i-1}, X_i, X'_i, x_{i+1}, ..., x_n)\right) = h(0)$. Let us therefore assume that the Theorem is true for $n - 1$. Then

$$\begin{aligned}
\mathbb{E}h(S - \mathbb{E}S) &= \mathbb{E}h(S - \mathbb{E}_{X'_n} S_n + \mathbb{E}_{X_n} S - \mathbb{E}S) \\
&\leq \mathbb{E}h(S - S_n + \mathbb{E}_{X_n} S - \mathbb{E}S) \\
&= \mathbb{E}h(S_n - S + \mathbb{E}_{X_n} S - \mathbb{E}S) \\
&= \mathbb{E}h(\epsilon_n |S - S_n| + \mathbb{E}_{X_n} S - \mathbb{E}S) \\
&\leq \mathbb{E}h(\epsilon_n F_n(x_1, ..., x_{n-1}, X_n, X'_n) + \mathbb{E}_{X_n} S - \mathbb{E}S),
\end{aligned}$$

where the equalities follow from the symmetry, the first inequality follows from Jensen's inequality and the convexity of h, and the last inequality follows from Lemma 2. Now, denoting $Z = \mathbb{E}_{X_n} S$, $Z_i = \mathbb{E}_{X_n} S_i$, we have for $i = 1, ..., n-1$

$$|Z - Z_i| = |\mathbb{E}_{X_n} S - \mathbb{E}_{X_n} S_i| \leq \mathbb{E}_{X_n} |S - S_i| \leq F_i(x_1, ..., x_{i-1}, X_i, X'_i, x_{i+1}, ..., x_n),$$

and thus for fixed X_n, X'_n and ϵ_n, we can apply the induction assumption to the function $t \to h(\epsilon_n F_n(x_1, ..., x_{n-1}, X_n, X'_n) + t)$ instead of h and $\mathbb{E}_{X_n} S$ instead of S, to obtain

$$\mathbb{E}h(S - \mathbb{E}S) \leq \mathbb{E}h\left(\sum_{i=1}^n \epsilon_i F_i(x_1, ..., x_{i-1}, X_i, X'_i, x_{i+1}, ..., x_n)\right).$$

The proof is complete. ■

We now demonstrate the framework.

Theorem 4 *For all $p \geq 1$,*

$$\|f(X_1, ..., X_n) - \mathbb{E}f(X_1, ..., X_n)\|_p \leq 2 \left\| \sum_{i=1}^n \sup_{x \in \mathcal{X}^n} f_i(X)(x) \right\|_p.$$

Proof We choose $F_i(x_1, ..., x_{i-1}, X_i, X'_i, x_{i+1}, ..., x_n) = \sup_{x \in \mathcal{X}^n} |f_i(X)(x) - f_i(X')(x)|$, which equals to $\sup_{x \in \mathcal{X}^n} |f(x_1, ..., x_{i-1}, X_i, x_{i+1}, ..., x_n) - \mathbb{E}[f(x_1, ..., x_{i-1}, X'_i, x_{i+1}, ..., x_n)] -$

$f\left(x_1,...,x_{i-1},X_i',x_{i+1},...,x_n\right)+\mathbb{E}\left[f\left(x_1,...,x_{i-1},X_i,x_{i+1},...,x_n\right)\right]|$. It is clear that $|S-S_i|\leq F_i(x_1,...,x_{i-1},X_i,X_i',x_{i+1},...,x_n)$. Then, for all $p\geq 1$,

$$\begin{aligned}\|f(X_1,...,X_n)-\mathbb{E}f(X_1,...,X_n)\|_p \leq & \left\|\sum_{i=1}^n \epsilon_i F_i(x_1,...,x_{i-1},X_i,X_i',x_{i+1},...,x_n)\right\|_p \\ \leq & \left\|\sum_{i=1}^n \sup_{x\in\mathcal{X}^n}\epsilon_i|f_i(X)(x)-f_i(X')(x)|\right\|_p \\ = & \left\|\sum_{i=1}^n \sup_{x\in\mathcal{X}^n} f_i(X)(x)-f_i(X')(x)\right\|_p \\ \leq & 2\left\|\sum_{i=1}^n \sup_{x\in\mathcal{X}^n} f_i(X)(x)\right\|_p,\end{aligned}$$

where the first inequality follows from Lemma 3 with $h(t)=|t|^p$ and the last one from the triangle inequality, and the equality follows from the symmetry. The proof is complete. ■

It should be noted that the function $F_i(x_1,...,x_{i-1},X_i,X_i',x_{i+1},...,x_n)$ can be substituted with other alternative functions. For instance, in Lemma 3, if we instead assume moreover that

$$|S-S_i|\leq F_i(X_i,X_i') \tag{1}$$

for some functions $F_i:\mathcal{X}^2\to\mathbb{R}$, $i=1,...,n$, then, following the proof of Lemma 3, the conclusion in Lemma 3 is

$$\mathbb{E}[h(S-\mathbb{E}S)]\leq\mathbb{E}\left[h\left(\sum_{i=1}^n\epsilon_i F_i(X_i,X_i')\right)\right].$$

Then, the general framework in Theorem 4 changes correspondingly.

Theorem 5 *Under the condition (1), for all $p\geq 1$,*

$$\|f(X_1,...,X_n)-\mathbb{E}f(X_1,...,X_n)\|_p\leq\left\|\sum_{i=1}^n\epsilon_i F_i(X_i,X_i')\right\|_p.$$

Remark 6 Theorem 4 provides a probabilistic toolbox, which can be used to establish moment inequalities for general functions of independent variables, once the moment inequality for sums of independent variables is derived. In Theorem 4, the supremum in the sum is a major weakness of the proposed approach. Firstly, the supremum means that some sort of boundedness remains necessary. Secondly, whenever this sum of random variables satisfies a central limit theorem a corresponding variance proxy will appear in the inequalities, however, this does not imply that $f(X_1,...,X_n)$ satisfies a central limit theorem. Besides, most of the inequalities given in Section 3 based on Theorem 4 require uniform boundedness of the conditional variances due to the presence of supremum in the

sum. Fortunately, under an additional condition (1), we can bypass this issue because at this point, the general framework no longer depends on the supremum, as shown in Theorem 5. The assumption $|S - S_i| \leq F_i(X_i, X_i')$ can be seen as a Lipschitz condition when $F_i(X_i, X_i')$ is a distance function, which is a commonly used condition coupled with the bounded difference-type inequality, see Theorem 1 in (Kontorovich, 2014) and Corollary 2.21 in (Wainwright, 2019). For example, if f is L-Lipschitz with respect to the metric ρ on $\mathcal{X}^n$ defined by $\rho(x, y) = \sum_i d_i(x_i, y_i)$ for some distance functions $d_i : \mathcal{X}^2 \to \mathbb{R}$, we can substitute $F_i(X_i, X_i')$ with $Ld_i(X_i, X_i')$, $i = 1, ..., n$. Thus, in the case of Lipschitz function classes, our general framework becomes

$$\|f(X_1, ..., X_n) - \mathbb{E}f(X_1, ..., X_n)\|_p \leq L \left\| \sum_{i=1}^{n} \epsilon_i d_i(X_i, X_i') \right\|_p .$$

Remark 7 This remark discusses the technical novelty of the framework. Deriving concentration inequalities for a general function f of independent random variables is more challenging than for their sum, as it typically involves some form of decomposition of the general function. In related studies, McDiarmid (1998); Kontorovich (2014) employed the martingale approach to decompose $f - \mathbb{E}f$, whereas Maurer and Pontil (2021) used the sub-additivity of entropy to decompose $f - \mathbb{E}f$. Subsequently, McDiarmid (1998); Kontorovich (2014); Maurer and Pontil (2021) focused on establishing upper bounds for the MGF, $\mathbb{E}e^{\lambda \epsilon_i Z_i}$, of a bounded or sub-Gaussian random variable Z_i. Further, Maurer and Pontil (2021) focused on a variant of the MGF, $\mathbb{E}Z_i^2 e^{\lambda Z_i}$, of a sub-exponential variable Z_i. Through demonstrating the upper bounds of the MGF, these works derived bounded difference-type inequalities for bounded, sub-Gaussian and sub-exponential distributions. However, these distributions are light-tailed, making their MGFs bounded, while the MGF of heavy-tailed random variables (heavier than sub-exponential distributions) remains unbounded, challenging the effectiveness of standard bounding techniques. Due to the inability to directly study MGF, alternative methodologies are warranted. We need to introduce a different decomposition for the function $f - \mathbb{E}f$. To counter this difficulty, we address it through a moment inequality perspective, introducing Lemma 3. The proof of Lemma 3 is done via an induction approach, and a pivotal step is constructing the function $t \to h(\epsilon_n F_n(x_1, ..., x_{n-1}, X_n, X_n') + t)$, achieved through a conditioning strategy leveraging Jensen's inequality. With Lemma 3, we first decompose the concentration of the general function f into the sum of independent variables. Then, instead of bounding the MGF, we focus on bounding the p-th moment of the sum of variables. By comparison, our proof techniques are relatively simple and easy to follow, and the framework obtained is pretty flexible to the application of heavy-tailed distributions.

3. Applications to Concrete Random Variables

This section applies the general framework in Theorem 4 to a range of random variables, encompassing bounded, Bernstein's moment condition, weak-exponential, and polynomial-moment variables, each subsequent category exhibiting heavier tails than its predecessor. In summary, the inequalities of these variables will be obtained in two steps: (1) Given the p-th moment bound of the sum of variables, that is $\| \sum_i \sup_{x \in \mathcal{X}^n} f_i(X)(x) \|_p$, Theorem 4 yields a moment inequality $\|f - \mathbb{E}f\|_p$; (2) Given the moment inequality $\|f - \mathbb{E}f\|_p$, Markov's inequality transfers it to a tail inequality $\mathbb{P}(|f - \mathbb{E}f| \geq t)$.

3.1 Bounded Random Variables

Before presenting the results, we first introduce a Bernstein-type moment inequality for sums of bounded random variables since this forms an integral part of our proofs.

Lemma 8 (Proposition D.1 in (Kuchibhotla and Chakrabortty, 2022)) *Suppose that $X_1, X_2, ..., X_n$ are independent random variables with mean zero and uniformly bounded by b in absolute value. Then for $p \geq 1$,*

$$\left\| \sum_{i=1}^{n} X_i \right\|_p \leq \sqrt{6p} \left(\sum_{i=1}^{n} \mathbb{E}[X_i^2] \right)^{1/2} + 10pb.$$

By a combination of Theorem 4 and Lemma 8, we have the following moment inequality.

Corollary 9 *With f, $(X_i)_{i=1}^n$ and $f_i(X)(x)$ as in Theorem 4. Suppose that for all i and any $x \in \mathcal{X}^n$, $|f_i(X)(x)| \leq b$ and $\mathbb{E}[(f_i(X)(x))^2] \leq \sigma_i^2$, then for all $p \geq 1$,*

$$\|f(X_1, ..., X_n) - \mathbb{E}f(X_1, ..., X_n)\|_p \leq 2 \left(\sqrt{6p} \left(\sum_{i=1}^{n} \sigma_i^2 \right)^{1/2} + 10pb \right).$$

Remark 10 The concentration property of bounded random variables has been widely studied in the literature. Although flourishing, we notice that a sharper moment bound for general functions of bounded random variables seems to be missed. In related work, Theorem 15.4 in (Boucheron et al., 2013) provides a moment version of the bounded differences inequality. Suppose, that f satisfies the bounded differences property, namely, for any $i = 1, ..., n$ and any $x_1, ..., x_n, x_i' \in \mathcal{X}$ it holds that $|f(x_1, ..., x_n) - f(x_1, ..., x_{i-1}, x_i', x_{i+1}, ..., x_n)| \leq b$. Then, for any $p \geq 2$,

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

Consider the case $f = \sum_{i=1}^n X_i$ and $|X_i| \leq b$ a.s. and $\mathbb{E}X_i = 0$, this inequality becomes

$$\left\| \sum_{i=1}^{n} X_i \right\|_p \leq 4\sqrt{np}b,$$

which is the moment version of Hoeffding's inequality. As a comparison, our Corollary 9 provides a Bernstein's version of Theorem 15.4 of (Boucheron et al., 2013). Although the bounded random variable is not heavy-tailed, our framework gives a novel result in this case.

By Markov's inequality, we can transfer the moment inequality to a tail inequality.

Corollary 11 *With f, $(X_i)_{i=1}^n$ and $f_i(X)(x)$ as in Theorem 4. Suppose that for all i and any $x \in \mathcal{X}^n$, $|f_i(X)(x)| \leq b$ and $\mathbb{E}[(f_i(X)(x))^2] \leq \sigma_i^2$, then for all $t > 0$*

$$\mathbb{P}\left(|f(X_1, ..., X_n) - \mathbb{E}f(X_1, ..., X_n)| \geq t\right) \leq \exp\left(- \min\left\{ \frac{t^2}{96e^2 \sum_{i=1}^n \sigma_i^2}, \frac{t}{40eb} \right\} \right).$$

Remark 12 The bound is a Bernstein-type inequality, however, it merits noting that Corollary 11 exhibits certain limitations compared to Theorem 3.8 in (McDiarmid, 1998). Specifically, the constants in Corollary 11 are larger, a consequence of transitioning from moment-based to tail-based bounds. Additionally, Corollary 11 requires the uniform boundedness of all conditional variances, whereas (McDiarmid, 1998) requires only uniform boundedness of the sum of conditional variances. Nevertheless, the application of Theorem 5 can address the issue of requiring uniform boundedness of all conditional variances, referring to Remark 6.

3.2 Bernstein's moment condition

Bernstein's moment condition is satisfied by various unbounded variables, a property that lends it much broader applicability than the bounded random variable. The definition is shown below.

Definition 13 *For the centered independent random variables $X_1, ..., X_n$, we say Bernstein's moment condition with parameter b holds if*

$$\sum_{i=1}^{n} \mathbb{E}\,|X_i|^p \leq \frac{\sum_{i=1}^{n} \mathbb{E}X_i^2}{2} p! b^{p-2}, \quad \textit{for all } p = 2, 3, 4, ...$$

Lemma 14 *Suppose that $X_1, X_2, ..., X_n$ are independent random variables with mean zero and satisfy Bernstein's moment condition. Then for $p \geq 2$,*

$$\left\| \sum_{i=1}^{n} X_i \right\|_p \leq 4 \left(\sum_{i=1}^{n} \mathbb{E}[X_i^2] \right)^{1/2} \sqrt{p} + 8bp.$$

By a combination of Theorem 4 and Lemma 14, we get the following result.

Corollary 15 *With f, $(X_i)_{i=1}^n$ and $f_i(X)(x)$ as in Theorem 4. Suppose that for all i and any $x \in \mathcal{X}^n$, the $f_i(X)(x)$ satisfy Bernstein's moment condition and $\mathbb{E}[(f_i(X)(x))^2] \leq \sigma_i^2$, then for all $p \geq 2$,*

$$\|f(X_1, ..., X_n) - \mathbb{E}f(X_1, ..., X_n)\|_p \leq 2 \left(4\sqrt{p} \left(\sum_{i=1}^{n} \sigma_i^2 \right)^{1/2} + 8pb \right).$$

Corollary 16 *With f, $(X_i)_{i=1}^n$ and $f_i(X)(x)$ as in Theorem 4. Suppose that for all i and any $x \in \mathcal{X}^n$, the $f_i(X)(x)$ satisfy Bernstein's moment condition and $\mathbb{E}[(f_i(X)(x))^2] \leq \sigma_i^2$, then for all $t > 0$,*

$$\mathbb{P}(|f(X_1, ..., X_n) - \mathbb{E}f(X_1, ..., X_n)| \geq t) \leq \exp\left(- \min\left\{ \frac{t^2}{256e^2 \sum_{i=1}^{n} \sigma_i^2}, \frac{t}{32eb} \right\} \right).$$

Remark 17 The bound is a Bernstein-type inequality, exhibiting a mixture of two tails, a sub-Gaussian tail governed by the variance-proxy $\sum_{i=1}^{n} \sigma_i^2$ for small deviations, and a sub-exponential tail governed by the scale-proxy b for large deviations. For the bounded difference-type inequality of Bernstein's moment variables, we have not found related results in the literature.

3.3 Weak-Exponential Random Variables

We introduce the definition of such a class of random variables.

Definition 18 *For $\alpha > 0$, define the function $\psi_\alpha : \mathbb{R}_+ \to \mathbb{R}_+$ with the formula $\psi_\alpha(x) = \exp(x^\alpha) - 1$. For a random variable X, define also the Orlicz norm*

$$\|X\|_{\psi_\alpha} = \inf\{\lambda > 0 : \mathbb{E}\psi_\alpha(|X|/\lambda) \leq 1\}.$$

Then, by Chebyshev's inequality, we say the random variable X is weak-exponential if for $t \geq 0$,

$$\mathbb{P}(|X| \geq t) \leq 2\exp\left(-\left(\frac{t}{\|X\|_{\psi_\alpha}}\right)^\alpha\right).$$

Remark 19 For $\alpha < 1$ the above definition does not give a norm but only a quasi-norm. It can be fixed by changing the function ψ_α near zero, to make it convex (which would give an equivalent norm) (Adamczak, 2007). It is however widely accepted in literature to use the word norm also for the quasi-norm given by our definition. In Definition 18, this class of variables is also referred to as sub-Weibull variables (Vladimirova et al., 2020; Kuchibhotla and Chakrabortty, 2022), which is a popular sub-class of unbounded random variables. It is parameterized by a positive tail index α, and a higher tail parameter α indicates a lighter tail. The weak-exponential distributions are reduced to sub-Gaussian distributions for $\alpha = 2$, to sub-exponential distributions for $\alpha = 1$, and to bounded variables for $\alpha = \infty$. The sub-Gaussian distributions subsume the Gaussian random variables, as well as all the bounded ones (such as Bernoulli, uniform, and multinomial). Sub-exponential random variables have heavier tails than sub-Gaussian variables and include the exponential, chi-squared, and Poisson distributions (Vershynin, 2018). Therefore, the weak-exponential distributions fall under a broad class of unbounded and heavy-tailed distributions, see (Vladimirova et al., 2020; Kuchibhotla and Chakrabortty, 2022).

Before showing the main results, we need to provide moment inequalities for the sum of weak-exponential random variables from (Kuchibhotla and Chakrabortty, 2022). There will be a transition for the moment bounds at $\alpha = 1$ due to the fact that weak-exponential random variables are log-convex for $\alpha \leq 1$ and log-concave for $\alpha \geq 1$.

Lemma 20 (Weak-exponential Variables with $0 < \alpha \leq 1$) *Suppose $X_1, X_2, ..., X_n$ are mean zero and independent weak-exponential random variables such that $0 < \alpha \leq 1$. Then for $p \geq 1$,*

$$\left\|\sum_{i=1}^n X_i\right\|_p \leq \sqrt{6p}\left(\sum_{i=1}^n \mathbb{E}[X_i^2]\right)^{1/2} + C_\alpha K_\alpha \max_{1\leq i\leq n} \|X_i\|_{\psi_\alpha}(\log(n+1))^{1/\alpha} p^{1/\alpha},$$

where C_α and K_α are constants depending only on α.

Lemma 21 (Weak-exponential Variables with $\alpha \geq 1$) *Suppose $X_1, X_2, ..., X_n$ are mean zero and independent weak-exponential random variables such that $\alpha \geq 1$. Then for $p \geq 1$,*

$$\left\|\sum_{i=1}^n X_i\right\|_p \leq \sqrt{6p}\left(\sum_{i=1}^n \mathbb{E}[X_i^2]\right)^{1/2} + C_\alpha \max_{1\leq i\leq n} \|X_i\|_{\psi_\alpha}(\log(n+1))^{1/\alpha} p,$$

where C_α is a constant depending only on α.

Remark 22 Note that for the weak-exponential distribution, when $0 < \alpha < 1$, such a class of distribution is heavy-tailed. In this case, its MGFs do not exist, and the technique to find upper bounds for the MGF fails.

By a combination of Theorem 4 and Lemma 20 and Lemma 21, we get the following result.

Corollary 23 *With f, $(X_i)_{i=1}^n$ and $f_i(X)(x)$ as in Theorem 4. Suppose that for all i and any $x \in \mathcal{X}^n$, $\|f_i(X)(x)\|_{\psi_\alpha} \leq b$ and $\mathbb{E}[(f_i(X)(x))^2] \leq \sigma_i^2$, then for all $p \geq 1$, if $0 < \alpha \leq 1$*

$$\|f(X_1, ..., X_n) - \mathbb{E}f(X_1, ..., X_n)\|_p \leq 2\left(\sqrt{6p}\left(\sum_{i=1}^n \sigma_i^2\right)^{1/2} + C_\alpha K_\alpha b(\log(n+1))^{1/\alpha} p^{1/\alpha}\right),$$

if $\alpha \geq 1$

$$\|f(X_1, ..., X_n) - \mathbb{E}f(X_1, ..., X_n)\|_p \leq 2\left(\sqrt{6p}\left(\sum_{i=1}^n \sigma_i^2\right)^{1/2} + C_\alpha b(\log(n+1))^{1/\alpha} p\right).$$

Corollary 24 *With f, $(X_i)_{i=1}^n$ and $f_i(X)(x)$ as in Theorem 4. Suppose that for all i and any $x \in \mathcal{X}^n$, $\|f_i(X)(x)\|_{\psi_\alpha} \leq b$ and $\mathbb{E}[(f_i(X)(x))^2] \leq \sigma_i^2$, then for all $t > 0$, if $0 < \alpha \leq 1$*

$$\mathbb{P}\left(|f(X) - \mathbb{E}f(X')| \geq t\right) \leq \exp\left(-\min\left\{\frac{t^2}{96e^2\sum_{i=1}^n \sigma_i^2}, \frac{t^\alpha}{(4eC_\alpha K_\alpha)^\alpha \log(n+1) b^\alpha}\right\}\right),$$

if $\alpha \geq 1$

$$\mathbb{P}\left(|f(X) - \mathbb{E}f(X')| \geq t\right) \leq \exp\left(-\min\left\{\frac{t^2}{96e^2\sum_{i=1}^n \sigma_i^2}, \frac{t}{4eC_\alpha(\log(n+1))^{1/\alpha} b}\right\}\right).$$

Remark 25 When $0 < \alpha \leq 1$, the bound exhibits a mixture of two tails. One is the sub-Gaussian tail $\exp\left(-\frac{t^2}{\sum_{i=1}^n \sigma_i^2}\right)$ for small deviations, which is induced from the central limit theorem, and the other is the weak-exponential tail $\exp\left(-\frac{t^\alpha}{b^\alpha}\right)$ for large deviations, which is expected from the weak-exponential distributions. When $\alpha \geq 1$, the bound also has two tails, a sub-Gaussian tail $\exp\left(-\frac{t^2}{\sum_{i=1}^n \sigma_i^2}\right)$ governed by the variance-proxy $\sum_{i=1}^n \sigma_i^2$ for small deviations, and a sub-exponential tail $\exp\left(-\frac{t}{b}\right)$ governed by the scale-proxy b for large deviations. For the bounded difference-type inequality of weak-exponential random variables, we have not found results comparable in the literature. The concentration inequality for the weak-exponential random variables is mainly devoted to the sum of independent variables (Kuchibhotla and Chakrabortty, 2022; Zhang and Wei, 2022; Bong and Kuchibhotla, 2023). If f is a sum of independent random variables, we recover the inequalities in the related works (Kuchibhotla and Chakrabortty, 2022; Zhang and Wei, 2022; Bong and Kuchibhotla, 2023) up to constants, i.e., if for all i, $\|X_i - \mathbb{E}[X_i]\|_{\psi_\alpha} \leq b$ and $\mathbb{E}[(X_i - \mathbb{E}[X_i])^2] \leq \sigma_i^2$, then for all $t > 0$, if $0 < \alpha \leq 1$

$$\mathbb{P}\left(\left|\sum_{i=1}^n X_i - \mathbb{E}\sum_{i=1}^n X_i\right| \geq t\right) \leq \exp\left(-\min\left\{\frac{t^2}{96e^2\sum_{i=1}^n \sigma_i^2}, \frac{t^\alpha}{(4eC_\alpha K_\alpha)^\alpha \log(n+1) b^\alpha}\right\}\right), \tag{2}$$

if $\alpha \geq 1$

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} X_i - \mathbb{E}\sum_{i=1}^{n} X_i\right| \geq t\right) \leq \exp\left(-\min\left\{\frac{t^2}{96e^2\sum_{i=1}^{n}\sigma_i^2}, \frac{t}{4eC_\alpha(\log(n+1))^{1/\alpha}b}\right\}\right). \quad (3)$$

Remark 26 We compare Corollary 24 with the work (Maurer and Pontil, 2021). Theorem 4 of (Maurer and Pontil, 2021) shows that for all $t > 0$,

$$\begin{aligned}&\mathbb{P}\left(f(X) - \mathbb{E}f(X') \geq t\right)\\ &\leq \exp\left(\frac{-t^2}{4e^2\sup_{x\in\mathcal{X}^n}\sum_{i=1}^{n}\|f_i(X)(x)\|_{\psi_1}^2 + 2e\max_i\sup_{x\in\mathcal{X}^n}\|f_i(X)(x)\|_{\psi_1}t}\right),\end{aligned}$$

where $\|\cdot\|_{\psi_1}$ is the sub-exponential norm defined by $\|Z\|_{\psi_1} = \sup_{p\geq 1}\frac{\|Z\|_p}{p}$ for a real random variable Z. In the case of $\alpha = 1$, our definition of the sub-exponential norm is equivalent to this one, differing from each other by at most an absolute constant factor, referring to Proposition 2.7.1 in (Vershynin, 2018). In the above inequality, both the variance-proxy and the scale proxy depend on the sub-exponential norm. However, a well known two-tailed bound for sums of bounded variables, Bernstein's inequality (Vershynin, 2018), has the variance proxy depending on $\|\cdot\|_2$ and the scale-proxy on $\|\cdot\|_\infty$. When $\|\cdot\|_2 \ll \|\cdot\|_\infty$, this leads to tighter bounds, whenever the inequality is operating in the sub-Gaussian regime, which often happens for large sample-sizes. In this spirit, Theorem 5 in (Maurer and Pontil, 2021) further shows that let $p, q \in (1, \infty)$ satisfy $p^{-1} + q^{-1} = 1$, then for all $t > 0$,

$$\begin{aligned}&\mathbb{P}\left(f(X) - \mathbb{E}f(X') \geq t\right)\\ &\leq \exp\left(\frac{-t^2}{2\sup_{x\in\mathcal{X}^n}\sum_{i=1}^{n}\|f_i(X)(x)\|_{2p}^2 + 2eq\max_i\sup_{x\in\mathcal{X}^n}\|f_i(X)(x)\|_{\psi_1}t}\right).\end{aligned}$$

Although this inequality gives substantial improvements over their Theorem 4, we cannot let $p \to 1$ to recover the behavior of Bernstein's inequality in the sub-Gaussian regime, because this would drive the scale-proxy to infinity. As a comparison, in the sub-Gaussian regime of our Corollary 24, the variance proxy $\sum_{i=1}^{n}\sigma_i^2$ successfully depends on $\|\cdot\|_2$. However, in the weak-exponential regime, our inequality entails a logarithmic term $\log(n+1)$, which is a disadvantage compared to (Maurer and Pontil, 2021). Moreover, the approach of (Maurer and Pontil, 2021) places the supremum over x outside the sum in the variance proxy, which is beneficial compared to Corollary 24. Nevertheless, we note that applying Theorem 5 addresses the issue of having the supremum over x inside the variance proxy, thereby eliminating the supremum. In such cases, our inequality recovers the behavior of Bernstein's inequality in the sub-Gaussian regime.

3.4 Polynomial-Moment Random Variables

The final class of distributions that we consider are those that satisfy a polynomial-moment bound, as specified by the following condition: there exists some $p \geq 2$ such that

$$\mathbb{E}[|X|^p] \leq b < \infty.$$

Distributions satisfying the above condition fall under a broad class of heavy-tailed distributions, including those with infinite variance (Nair, 2012; Nagaev, 1979).

Next, we introduce a moment inequality for the sum of variables satisfying the polynomial-moment condition, which involves the variance, known as Rosenthal-type inequality.

Lemma 27 (Theorem 15.11 in (Boucheron et al., 2005)) *Suppose that $X_1, X_2, ..., X_n$ are independent centered random variables. Then for any integer $p \geq 2$*

$$\left\| \sum_{i=1}^{n} X_i \right\|_p \leq \sqrt{2\kappa(2+\theta)p} \left(\sum_{i=1}^{n} \mathbb{E}|X_i|^2 \right)^{\frac{1}{2}} + p\kappa\sqrt{1+\frac{1}{\theta}} \left(\sum_{i=1}^{n} \mathbb{E}|X_i|^p \right)^{\frac{1}{p}},$$

where $\theta \in (0,1)$ and $\kappa = \frac{\sqrt{e}}{2(\sqrt{e}-1)} < 1.271$.

By a combination of Theorem 4 and Lemma 27, we get the following result.

Corollary 28 *With f, $(X_i)_{i=1}^n$ and $f_i(X)(x)$ as in Theorem 4. Suppose that for all i and any $x \in \mathcal{X}^n$, $\mathbb{E}[|f_i(X)(x)|^p] \leq b_i$ for some $p \geq 2$ and $\mathbb{E}[(f_i(X)(x))^2] \leq \sigma_i^2$, then*

$$\|f(X_1, ..., X_n) - \mathbb{E}f(X_1, ..., X_n)\|_p \leq 2\sqrt{2\kappa(2+\theta)p} \left(\sum_{i=1}^{n} \sigma^2 \right)^{\frac{1}{2}} + 2p\kappa\sqrt{1+\frac{1}{\theta}} \left(\sum_{i=1}^{n} b_i \right)^{\frac{1}{p}}.$$

Corollary 29 *With f, $(X_i)_{i=1}^n$ and $f_i(X)(x)$ as in Theorem 4. Suppose that for all i and any $x \in \mathcal{X}^n$, $\mathbb{E}[|f_i(X)(x)|^p] \leq b_i$ for some $p \geq 2$ and $\mathbb{E}[(f_i(X)(x))^2] \leq \sigma_i^2$, then for all $t > 0$,*

$$\mathbb{P}\left(|f(X) - \mathbb{E}f(X')| \geq t\right) \leq \exp\left(-\frac{t^2}{16e^2(2\kappa(2+\theta))\sum_{i=1}^{n}\sigma_i^2} \right) + \frac{(4p\kappa\sqrt{1+\frac{1}{\theta}})^p \sum_{i=1}^{n} b_i}{t^p}.$$

Remark 30 The bound exhibits a mixture of two tails, a sub-Gaussian tail $\exp\left(-\frac{t^2}{\sum_{i=1}^n \sigma_i^2}\right)$ for small deviations, which is induced from the central limit theorem, and a power-type tail $\frac{\sum_{i=1}^n b_i}{t^p}$ for large deviations, which is expected from the polynomial-moment condition. For the bounded difference-type inequality of polynomial-moment random variables, we have not found results comparable to Corollary 29 in the literature.

Remark 31 The classical Nagaev's inequality states that if for all i, $\mathbb{E}[|X_i - \mathbb{E}[X_i]|^p] \leq b_i$ for some for some $p > 2$ and $\mathbb{E}[(X_i - \mathbb{E}[X_i])^2] \leq \sigma_i^2$, then for all $t > 0$,

$$\mathbb{P}\left(\left| \sum_{i=1}^{n} X_i - \mathbb{E}\sum_{i=1}^{n} X_i \right| \geq t \right) \leq 2\exp\left(-\frac{2e^{-p}(p+2)^{-2}t^2}{\sum_{i=1}^{n}\sigma_i^2} \right) + \frac{(1+\frac{2}{p})^p \sum_{i=1}^{n} b_i}{t^p},$$

referring to Corollary 1.8 in (Nagaev, 1979). Nagaev's inequality has raised interesting applications in machine learning, such as stochastic optimization (Lou et al., 2022). As a comparison, if f is a sum of independent random variables, we recover Nagaev's inequality up to constants.

Then, we consider a different moment inequality for the polynomial-moment condition, which does not require the bounded variance, providing concentration and moment inequalities for distributions with infinite variance.

Lemma 32 (Marcinkiewicz-Zygmund's inequality (Ren and Liang, 2001)) *Suppose $X_1, X_2, ..., X_n$ are independent centered random variables with a finite p-th moment for $p \geq 2$. Then*

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

By a combination of Theorem 4 and Lemma 32, we get the following result.

Corollary 33 *With f, $(X_i)_{i=1}^n$ and $f_i(X)(x)$ as in Theorem 4. Suppose that for all i and any $x \in \mathcal{X}^n$, $\mathbb{E}[|f_i(X)(x)|^p] \leq b$ for some $p \geq 2$, then*

$$\|f(X_1, ..., X_n) - \mathbb{E}f(X_1, ..., X_n)\|_p \leq 6\sqrt{2np}b^{1/p}.$$

Corollary 34 *With f, $(X_i)_{i=1}^n$ and $f_i(X)(x)$ as in Theorem 4. Suppose that for all i and any $x \in \mathcal{X}^n$, $\mathbb{E}[|f_i(X)(x)|^p] \leq b$ for some $p \geq 2$, then for all $t > 0$,*

$$\mathbb{P}\left(|f(X_1, ..., X_n) - \mathbb{E}f(X_1, ..., X_n)| \geq t\right) \leq \frac{(6\sqrt{2np}b^{1/p})^p}{t^p}.$$

4. Applications to Statistical Learning Theory

In this section, we explore the application of these inequalities to statistical learning theory, showcasing their use in vector-valued concentration and different approaches to prove generalization bounds. We primarily focus on weak-exponential random variables for conciseness. The inequalities of other types random variables in this paper can also be applied to these applications by the reader following the same pattern. The results in this section provide a substantial extension of the existing ones, the classical bounded results and the sub-Gaussian/sub-exponential results in (Maurer and Pontil, 2021; Kontorovich, 2014), to heavy-tailed distributions. Since the improvement is clear, we will not list the relevant results in (Maurer and Pontil, 2021; Kontorovich, 2014) for comparison.

4.1 Vector Valued Concentration

We study the concentration of vectors in a normed space $(\mathcal{X}, \|\cdot\|)$ and demonstrate results of the weak-exponential variables.

Theorem 35 *Suppose the X_i are independent random variables with values in a normed space $(\mathcal{X}, \|\cdot\|)$ such that $\|\|X_i\|\|_{\psi_\alpha} \leq \infty$.*
(i) Then for all $t > 0$, if $0 < \alpha \leq 1$

$$\begin{aligned}
&\mathbb{P}\left(\left\| \left\| \sum_{i=1}^{n} X_i \right\| - \mathbb{E} \left\| \sum_{i=1}^{n} X_i \right\| \right\| \geq t \right) \\
&\leq \exp\left(- \min\left\{ \frac{t^2}{96e^2 4 \sum_{i=1}^{n} \|\|X_i\|\|_2^2}, \frac{t^\alpha}{(4eC_\alpha K_\alpha)^\alpha \log(n+1)(2\max_k \|\|X_k\|\|_{\psi_\alpha})^\alpha} \right\} \right),
\end{aligned}$$

if $\alpha \geq 1$

$$\mathbb{P}\left(\left|\left\|\sum_{i=1}^{n} X_i\right\| - \mathbb{E}\left\|\sum_{i=1}^{n} X_i\right\|\right| \geq t\right)$$
$$\leq \exp\left(-\min\left\{\frac{t^2}{96e^2 4\sum_{i=1}^{n} \|\|X_i\|\|_2^2}, \frac{t}{4eC_\alpha(\log(n+1))^{1/\alpha} 2\max_k \|\|X_k\|\|_{\psi_\alpha}}\right\}\right).$$

(ii) If $\mathcal{X}$ *is a Hilbert space, the* X_i *are i.i.d., then for all* $t > 0$*, if* $0 < \alpha \leq 1$

$$\mathbb{P}\left(\left\|\sum_{i=1}^{n}(X_i - \mathbb{E}X_1')\right\| \geq t + \sqrt{n}\|\|X_1\|\|_2\right)$$
$$\leq \exp\left(-\min\left\{\frac{t^2}{96e^2 4n\|\|X_1\|\|_2^2}, \frac{t^\alpha}{(4eC_\alpha K_\alpha)^\alpha \log(n+1)(2\|\|X_1\|\|_{\psi_\alpha})^\alpha}\right\}\right),$$

if $\alpha \geq 1$

$$\mathbb{P}\left(\left\|\sum_{i=1}^{n}(X_i - \mathbb{E}X_1')\right\| \geq t + \sqrt{n}\|\|X_1\|\|_2\right)$$
$$\leq \exp\left(-\min\left\{\frac{t^2}{96e^2 4n\|\|X_1\|\|_2^2}, \frac{t}{4eC_\alpha(\log(n+1))^{1/\alpha} 2\|\|X_1\|\|_{\psi_\alpha}}\right\}\right).$$

Remark 36 The vector valued inequality has wide applications in learning theory, such as the generalization error analysis in Hilbert Space, see (Smale and Zhou, 2007).

As an illustration, we apply the derived vector valued concentration inequality above to study the principal subspace selection, often called PCA (principal component analysis), with weak-exponential data. In PCA we seek a projection onto a d-dimensional subspace that most faithfully represents the data. Let H be a Hilbert-space, X_i i.i.d. with values in H and $\mathcal{P}_d$ the set of d-dimensional orthogonal projection operators in H. For $x \in H$ and $P \in \mathcal{P}_d$ the reconstruction error is $\ell(P, x) := \|Px - x\|_H^2$. We provide a bound, uniformly for projections in $\mathcal{P}_d$, on the estimation error between the expected and the empirical reconstruction error.

Corollary 37 *With* $X = (X_1, ..., X_n)$ *i.i.d., then for all* $t > 0$*, if* $0 < \alpha \leq 1$

$$\mathbb{P}\left(\sup_{P\in\mathcal{P}_d} \frac{1}{n}\sum_i \mathbb{E}[\ell(P, X_1)] - \ell(P, X_i) \geq t + \frac{\sqrt{d}}{\sqrt{n}}\|\|X_1\|^2\|_2\right)$$
$$\leq 2\exp\left(-\min\left\{\frac{nt^2}{96e^2 4d\|\|X_1\|^2\|_2^2}, \frac{(nt)^\alpha}{(4eC_\alpha K_\alpha)^\alpha \log(n+1)(2\sqrt{d}\|\|X_1\|^2\|_{\psi_\alpha})^\alpha}\right\}\right),$$

if $\alpha \geq 1$

$$\mathbb{P}\left(\sup_{P\in\mathcal{P}_d} \frac{1}{n}\sum_i \mathbb{E}[\ell(P, X_1)] - \ell(P, X_i) \geq t + \frac{\sqrt{d}}{\sqrt{n}}\|\|X_1\|^2\|_2\right)$$
$$\leq 2\exp\left(-\min\left\{\frac{nt^2}{96e^2 4d\|\|X_1\|^2\|_2^2}, \frac{nt}{4eC_\alpha(\log(n+1))^{1/\alpha} 2\sqrt{d}\|\|X_1\|^2\|_{\psi_\alpha}}\right\}\right).$$

4.2 Rademacher Complexity and Generalization

Rademacher complexity is a popular notion of complexity that is distribution dependent. Suppose that $\mathcal{G}$ is a class of function $g : \mathcal{X} \to \mathbb{R}$. The Rademacher complexity of $\mathcal{G}$ is defined as

$$\mathcal{R}(\mathcal{G}) = \mathbb{E}\left[\frac{1}{n}\mathbb{E}\left[\sup_{g\in\mathcal{G}}\sum_i \epsilon_i g(X_i)|X\right]\right],$$

where $\epsilon_1, ..., \epsilon_n$ are independent random variables uniformly chosen from $\{-1, 1\}$. Together with the symmetrization argument, it leads to an expected bound, uniformly for functions in $\mathcal{G}$

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_i g(X_i) - \mathbb{E}[g(X_i')]\right] \le 2\mathcal{R}(\mathcal{G}).$$

The classical method in statistical learning uses the bounded difference inequality to show that $\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_i g(X_i) - \mathbb{E}[g(X_i')]$ is sharply concentrated about its mean $\mathbb{E}[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_i g(X_i) - \mathbb{E}[g(X_i')]]$, resulting in the following generalization bound: if $g : \mathcal{X} \to [0, 1]$, for any $t \ge 0$

$$\mathbb{P}\left(\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_i g(X_i) - \mathbb{E}[g(X_i')] > 2\mathcal{R}(\mathcal{G}) + t\right) \le \exp\left(-2nt^2\right).$$

Although this approach is very fundamental, it relies on the $g(X_i)$ being bounded, a condition dictated by the bounded difference inequality. However, providing bounds on the Rademacher complexity does not necessarily require the boundedness, and Lipschitz properties are often more prevalent. We now demonstrate that the boundedness can be relaxed by heavy-tailed distributions for uniformly Lipschitz function classes and present results of the weak-exponential variables.

Theorem 38 *Let $X = (X_1, ..., X_n)$ be a vector of i.i.d. weak-exponential random variables with values in a Banach space $(\mathcal{X}, \|\cdot\|)$ and let $\mathcal{G}$ be a class of function $g : \mathcal{X} \to \mathbb{R}$ such that $g(x) - g(y) \le L\|x - y\|$ for all $g \in \mathcal{G}$ and that $x, y \in \mathcal{X}$. Then, for all $t > 0$, if $0 < \alpha \le 1$*

$$\begin{aligned}&\mathbb{P}\left(\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_i g(X_i) - \mathbb{E}[g(X_i')] \ge t + 2\mathcal{R}(\mathcal{G})\right)\\ \le &\exp\left(-\min\left\{\frac{t^2}{96e^2\frac{16L^2}{n}\|\|X_1\|\|_2^2}, \frac{t^\alpha}{(4eC_\alpha K_\alpha)^\alpha \log(n+1)(\frac{4L}{n}\|\|X_1\|\|_{\psi_\alpha})^\alpha}\right\}\right),\end{aligned}$$

if $\alpha \ge 1$

$$\begin{aligned}&\mathbb{P}\left(\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_i g(X_i) - \mathbb{E}[g(X_i')] \ge t + 2\mathcal{R}(\mathcal{G})\right)\\ \le &\exp\left(-\min\left\{\frac{t^2}{96e^2\frac{16L^2}{n}\|\|X_1\|\|_2^2}, \frac{t}{4eC_\alpha(\log(n+1))^{1/\alpha}\frac{4L}{n}\|\|X_1\|\|_{\psi_\alpha}}\right\}\right).\end{aligned}$$

Remark 39 In the application of Theorem 38 to give generalization bounds for a learning problem, the key step is to prove the upper bound for the Rademacher complexity $\mathcal{R}(\mathcal{G})$. Typically, $\mathcal{R}(\mathcal{G})$ is of the order $O(1/\sqrt{n})$, see Corollary 40.

As an illustration, we apply the derived inequalities on Rademacher complexity above to provide generalization bounds for linear regression in the context of potentially unbounded data. Let $\mathcal{X} = (H, \mathbb{R})$, where H is a Hilbert-space with inner product $\langle \cdot, \cdot \rangle$ and norm $\| \cdot \|_H$, and let X_1 and Z_1 be each weak-exponential random variables in H and $\mathbb{R}$ respectively. The pair (X_1, Z_1) represents the joint occurrence of input-vectors X_1 and real outputs Z_1. Within $\mathcal{X}$ we consider the class $\mathcal{G}$ of functions $\mathcal{G} = \{(x, z) \to g(x, z) = \ell(\langle w, x \rangle - z) : \|w\|_H \leq L\}$, where ℓ is a 1-Lipschitz loss function, such as the absolute error or the Huber loss. The generalization bound is shown below.

Corollary 40 *Let $\mathcal{X}$ and $\mathcal{G}$ be as above and $(X, Z) = ((X_1, Z_1), ..., (X_n, Z_n))$ be an i.i.d. sample of random variables in $\mathcal{X}$. Then, for all $t > 0$, if $0 < \alpha \leq 1$*

$$\mathbb{P}\left(\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_i g(X_i) - \mathbb{E}[g(X_i')] \geq t + \frac{4}{\sqrt{n}}\delta_{(X_1,Z_1)}\right)$$
$$\leq \exp\left(-\min\left\{\frac{t^2}{96e^2\frac{16}{n}(\delta_{(X_1,Z_1)})^2}, \frac{t^\alpha}{(4eC_\alpha K_\alpha \frac{4}{n})^\alpha \log(n+1)(L|||X_1|||_{\psi_\alpha} + |||Z_1|||_{\psi_\alpha})^\alpha}\right\}\right),$$

if $\alpha \geq 1$

$$\mathbb{P}\left(\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_i g(X_i) - \mathbb{E}[g(X_i')] \geq t + \frac{4}{\sqrt{n}}\delta_{(X_1,Z_1)}\right)$$
$$\leq \exp\left(-\min\left\{\frac{t^2}{96e^2\frac{16}{n}(\delta_{(X_1,Z_1)})^2}, \frac{t}{4eC_\alpha(\log(n+1))^{1/\alpha}\frac{4}{n}(L|||X_1|||_{\psi_\alpha} + |||Z_1|||_{\psi_\alpha})}\right\}\right),$$

where $\delta_{(X_1,Z_1)} = L|||X_1|||_2 + |||Z_1|||_2$.

4.3 Algorithmic Stability and Generalization

Algorithmic stability is receiving increasing attention in the generalization analysis of machine learning algorithms. The classical statistical learning method uses the bounded difference inequality to show that $f(X) - \mathbb{E}[f(X')]$ is sharply concentrated about its mean. Together with certain measures of algorithmic stability, this method gives stability bounds for the mean, which in turn lead to stability-based generalization bounds (Bousquet and Elisseeff, 2002). However, it requires boundedness. We now review two pivotal works that broaden the scope of classical stability theory to accommodate unbounded scenarios. If $(\mathcal{X}, d, \mu)$ is a metric probability space and $X, X' \sim \mu$ are i.i.d. random variables with values in $\mathcal{X}$. Kontorovich (2014) studies the sub-Gaussian tail of $d(X, X')$. Maurer and Pontil (2021) extend the method of (Kontorovich, 2014) from sub-Gaussian to sub-exponential distributions. They work with sub-Gaussian and sub-exponential norms defined respectively as $\|d(X, X')\|_{\psi_2}$ and $\|d(X, X')\|_{\psi_1}$ for independent $X', X \sim \mu$. Our results can extend the method of (Maurer and Pontil, 2021; Kontorovich, 2014) to heavy-tailed distributions.

Theorem 41 *For $1 \leq i \leq n$, let X_i be independent weak-exponential random variables distributed as μ_i in $\mathcal{X}$, $X = (X_1, ..., X_n)$, X' i.i.d. to X, and let $f : \mathcal{X}^n \to \mathbb{R}$ have Lipschitz constant L with respect to the metric ρ on $\mathcal{X}^n$ defined by $\rho(x, y) = \sum_i d(x_i, y_i)$. Then for all $t > 0$, if $0 < \alpha \leq 1$*

$$\mathbb{P}\left(|f(X) - \mathbb{E}f(X')| \geq t\right)$$
$$\leq \exp\left(-\min\left\{\frac{t^2}{96e^2\sum_{i=1}^n L^2\|d(X_i, X_i')\|_2^2}, \frac{t^\alpha}{(4eC_\alpha K_\alpha)^\alpha \log(n+1)L^\alpha \max_k \|d(X_k, X_k')\|_{\psi_\alpha}^\alpha}\right\}\right),$$

if $\alpha \geq 1$

$$\mathbb{P}\left(|f(X) - \mathbb{E}f(X')| \geq t\right)$$
$$\leq \exp\left(-\min\left\{\frac{t^2}{96e^2\sum_{i=1}^n L^2\|d(X_i, X_i')\|_2^2}, \frac{t}{4eC_\alpha L \max_k \|d(X_k, X_k')\|_{\psi_\alpha}(\log(n+1))^{1/\alpha}}\right\}\right).$$

As an illustration, we can apply the derived inequalities on algorithmic stability above to establish generalization bounds using the notion of total Lipschitz stability (Maurer and Pontil, 2021; Kontorovich, 2014) for weak-exponential distributions. Readers may refer to the proof steps in Section 5 of (Kontorovich, 2014). To keep the paper within a reasonable length, we omit the results of this part.

Also note that in Section 4, the norm $\|\cdot\|_2$ can be replaced with $\|\cdot\|_{\psi_\alpha}$ for weak-exponential random variables, which can be obtained from the following steps

$$\mathbb{E}[Y^2] = \int_0^\infty \mathbb{P}(|Y|^2 > t)dt = \int_0^\infty \mathbb{P}\left(|Y| > t^{1/2}\right)dt \leq \int_0^\infty 2\exp\left(-(t^{1/2}/\|Y\|_{\psi_\alpha})^\alpha\right)dt$$
$$= \int_0^\infty \frac{4}{\alpha}e^{-u}u^{\frac{2}{\alpha}-1}\|Y\|_{\psi_\alpha}^2 dt = \frac{4}{\alpha}\|Y\|_{\psi_\alpha}^2\Gamma\left(\frac{2}{\alpha}\right) = 2\|Y\|_{\psi_\alpha}^2\Gamma\left(\frac{2}{\alpha}+1\right),$$

where Y is a weak-exponential random variable, and where the function Γ is defined by the integral formula $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$, $x > 0$.

5. Conclusion

This paper presented bounded difference-type concentration and moment inequalities for general functions of heavy-tailed independent variables. We provided a probabilistic toolbox that is general and flexible to derive bounded difference-type concentration and moment inequalities for heavy-tailed distributions. We illustrated this framework to bounded, Bernstein's moment condition, weak-exponential, and polynomial-moment variables. We then illustrated these inequalities with applications to some standard problems in learning theory. We hope that future work will reveal more interesting applications of these inequalities.

Acknowledgments

We thank the editor and anonymous reviewers for their valuable suggestions. This work is supported by the Beijing Natural Science Foundation (No.4222029); the National Natural Science Foundation of China (NO.62076234, 62476277); the National Key Research and Development Project (No.2022YFB2703102).

Appendix A. Proofs of Section 3

A.1 Proof of Corollary 11

Proof For any $t > 0$, by Markov's inequality

$$\begin{aligned}&\mathbb{P}\left(|f(X_1,...,X_n) - \mathbb{E}f(X_1,...,X_n)| \geq t\right)\\ \leq &\frac{\|f(X_1,...,X_n) - \mathbb{E}f(X_1,...,X_n)\|_p^p}{t^p} \leq \frac{2^p\left(\sqrt{6p}\left(\sum_{i=1}^n \sigma_i^2\right)^{1/2} + 10pb\right)^p}{t^p}.\end{aligned}$$

Setting t such that

$$\exp(-p) = 2^p\left(\sqrt{6p}\left(\sum_{i=1}^n \sigma_i^2\right)^{1/2} + 10pb\right)^p / t^p.$$

By the inequality $a+b \leq 2\max\{a,b\}$ for $a,b>0$, it is clear that if $\sqrt{6p}\left(\sum_{i=1}^n \sigma_i^2\right)^{1/2} \geq 10pb$, we have

$$\mathbb{P}\left(|f(X_1,...,X_n) - \mathbb{E}f(X_1,...,X_n)| \geq 4e\sqrt{6p}\left(\sum_{i=1}^n \sigma_i^2\right)^{1/2}\right) \leq \exp(-p).$$

Put $t = 4e\sqrt{6p}\left(\sum_{i=1}^n \sigma_i^2\right)^{1/2}$, we have

$$\mathbb{P}\left(|f(X_1,...,X_n) - \mathbb{E}f(X_1,...,X_n)| \geq t\right) \leq \exp\left(-\frac{t^2}{96e^2\sum_{i=1}^n \sigma_i^2}\right).$$

While if $\sqrt{6p}\left(\sum_{i=1}^n \sigma_i^2\right)^{1/2} \leq 10pb$, we have

$$\mathbb{P}\left(|f(X_1,...,X_n) - \mathbb{E}f(X_1,...,X_n)| \geq 40ebp\right) \leq \exp(-p).$$

Put $t = 40ebp$, we have

$$\mathbb{P}\left(|f(X_1,...,X_n) - \mathbb{E}f(X_1,...,X_n)| \geq t\right) \leq \exp\left(-\frac{t}{40eb}\right).$$

Combining the two cases, the proof is complete. ■

A.2 Proof of Lemma 14

Proof From Theorem 2.10 of (Boucheron et al., 2013), for all $t \geq 0$ there holds

$$\mathbb{P}\left(\left|\sum_{i=1}^n X_i\right| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2(\sum_{i=1}^n \mathbb{E}[X_i^2] + bt)}\right).$$

Set $t_0 = \sum_{i=1}^n \mathbb{E}[X_i^2]/b$. Now observe that for $p \geq 2$,

$$\begin{aligned}
&\mathbb{E}\left[\left|\sum_{i=1}^n X_i\right|^p\right] = \int_0^\infty pt^{p-1}\mathbb{P}\left(\left|\sum_{i=1}^n X_i\right| \geq t\right) dt \\
\leq& 2\int_0^\infty pt^{p-1}\exp\left(-\frac{t^2}{2(\sum_{i=1}^n \mathbb{E}[X_i^2] + bt)}\right) dt \\
=& 2\int_0^{t_0} pt^{p-1}\exp\left(-\frac{t^2}{4\sum_{i=1}^n \mathbb{E}[X_i^2]}\right) dt + 2\int_{t_0}^\infty pt^{p-1}\exp\left(-\frac{t^2}{4bt}\right) dt \\
\leq& 2\int_0^\infty pt^{p-1}\exp\left(-\frac{t^2}{4\sum_{i=1}^n \mathbb{E}[X_i^2]}\right) dt + 2\int_0^\infty pt^{p-1}\exp\left(-\frac{t}{4b}\right) dt \\
:=& A + B.
\end{aligned}$$

Considering the first term A, by a change of variable, we have

$$\begin{aligned}
A =& 2\int_0^\infty pt^{p-1}\exp\left(-\frac{t^2}{4\sum_{i=1}^n \mathbb{E}[X_i^2]}\right) dt \\
=& 2\left(\sqrt{4\sum_{i=1}^n \mathbb{E}[X_i^2]}\right)^p \int_0^\infty pz^{p-1}\exp\left(-z^2\right) dz \\
=& \left(\sqrt{4\sum_{i=1}^n \mathbb{E}[X_i^2]}\right)^p \Gamma\left(\frac{p}{2}+1\right) \\
\leq& \left(\sqrt{4\sum_{i=1}^n \mathbb{E}[X_i^2]}\right)^p \sqrt{2\pi\frac{p}{2}}\left(\frac{p}{2}\right)^{\frac{p}{2}}\exp^{-\frac{p}{2}+\frac{1}{12\frac{p}{2}}},
\end{aligned}$$

where the last inequality follows from the Stirling formula: $n! = \sqrt{2\pi n}n^n e^{-n+\theta_n}$, $|\theta_n| \leq \frac{1}{12n}$, $n > 1$. Simplifying the above bound for $p \geq 2$, we get

$$\begin{aligned}
A^{1/p} \leq& \sqrt{4\sum_{i=1}^n \mathbb{E}[X_i^2]}(\pi p)^{\frac{1}{2p}}\sqrt{\frac{p}{2}}\exp^{-\frac{1}{2}+\frac{1}{6p^2}} \\
\leq& \sqrt{4\sum_{i=1}^n \mathbb{E}[X_i^2]}(\pi)^{\frac{1}{4}}(p)^{\frac{1}{2p}}\sqrt{\frac{p}{2}}\exp^{-\frac{1}{2}+\frac{1}{24}} \leq 4\sqrt{\sum_{i=1}^n \mathbb{E}[X_i^2]}\sqrt{p},
\end{aligned}$$

where the last inequality uses the fact that $p^{1/p} \leq e^{1/e}$. To bound B, note that by change of variable

$$\begin{aligned}
B =& 2\int_0^\infty pt^{p-1}\exp\left(-\frac{t}{4b}\right) dt \\
=& 2(4b)^p\int_0^\infty pz^{p-1}\exp\left(-z\right) dz \\
=& 2(4b)^p\Gamma(p+1) \\
\leq& 2(4b)^p\sqrt{2\pi p}p^p\exp^{-p+\frac{1}{12p}}.
\end{aligned}$$

Similarly, simplifying the above bound for $p \geq 2$, we get

$$B^{1/p} \leq 4b(2)^{1/p}(2\pi p)^{\frac{1}{p}} p \exp^{-1+\frac{1}{12p^2}} \leq 2b(4\pi)^{\frac{1}{2}}(p)^{\frac{1}{p}} p \exp^{-1+\frac{1}{48}} \leq 8bp.$$

Therefore, for $p \geq 2$,

$$\left(\mathbb{E}\left[\left|\sum_{i=1}^{n} X_i\right|^p\right]\right)^{1/p} \leq 4\sqrt{\sum_{i=1}^{n} \mathbb{E}[X_i^2]}\sqrt{p} + 8bp.$$

The proof is complete. ■

A.3 Proof of Corollary 16

Proof For any $t > 0$, by Markov's inequality

$$\begin{aligned}
&\mathbb{P}\left(|f(X_1, ..., X_n) - \mathbb{E}f(X_1, ..., X_n)| \geq t\right) \\
\leq &\frac{\|f(X_1, ..., X_n) - \mathbb{E}f(X_1, ..., X_n)\|_p^p}{t^p} \leq \frac{2^p\left(4\sqrt{p}\left(\sum_{i=1}^n \sigma_i^2\right)^{1/2} + 8pb\right)^p}{t^p}.
\end{aligned}$$

Setting t such that

$$\exp(-p) = 2^p\left(4\sqrt{p}\left(\sum_{i=1}^{n}\sigma_i^2\right)^{1/2} + 8pb\right)^p / t^p.$$

By the inequality $a + b \leq 2\max\{a, b\}$ for $a, b > 0$, it is clear that if $4\sqrt{p}\left(\sum_{i=1}^n \sigma_i^2\right)^{1/2} \geq 8pb$, we have

$$\mathbb{P}\left(|f(X_1, ..., X_n) - \mathbb{E}f(X_1, ..., X_n)| \geq 4e4\sqrt{p}\left(\sum_{i=1}^{n}\sigma_i^2\right)^{1/2}\right) \leq \exp(-p).$$

Put $t = 4e4\sqrt{p}\left(\sum_{i=1}^n \sigma_i^2\right)^{1/2}$, we have

$$\mathbb{P}\left(|f(X_1, ..., X_n) - \mathbb{E}f(X_1, ..., X_n)| \geq t\right) \leq \exp\left(-\frac{t^2}{256e^2\sum_{i=1}^n \sigma_i^2}\right).$$

While if $4\sqrt{p}\left(\sum_{i=1}^n \sigma_i^2\right)^{1/2} \leq 8pb$, we have

$$\mathbb{P}\left(|f(X_1, ..., X_n) - \mathbb{E}f(X_1, ..., X_n)| \geq 32ebp\right) \leq \exp(-p).$$

Put $t = 32ebp$, we have

$$\mathbb{P}\left(|f(X_1, ..., X_n) - \mathbb{E}f(X_1, ..., X_n)| \geq t\right) \leq \exp\left(-\frac{t}{32eb}\right).$$

Combining the two cases, the proof is complete. ■

A.4 Proof of Lemma 20

Proof The proof method is a combination of truncation and Hoffmann-Jorgensen's inequality, referring to (Kuchibhotla and Chakrabortty, 2022). Define

$$Z = \max_{1\le i\le n} |X_i|, \rho = 8\mathbb{E}[Z], X_{i,1} = X_i\mathbb{I}\{|X_i| \le \rho\} - \mathbb{E}[X_i\mathbb{I}\{|X_i| \le \rho\}], \text{ and } X_{i,2} = X - X_{i,1}.$$

It is clear that $X_i = X_{i,1} + X_{x,2}$ and $|X_{i,1}| \le 2\rho$ for $1 \le i \le n$. Also by by triangle inequality, for $p \ge 1$,

$$\left\|\sum_{i=1}^n X_i\right\|_p \le \left\|\sum_{i=1}^n X_{i,1}\right\|_p + \left\|\sum_{i=1}^n X_{i,2}\right\|_p.$$

Now for $1 \le i \le n$,

$$\mathbb{E}[X_{i,1}^2] = Var(X_{i,1}) = Var(X_i\mathbb{I}\{|X_i| \le \rho\}) \le \mathbb{E}[X_i^2].$$

Thus, applying Bernstein's inequality of Lemma 8, for $p \ge 1$,

$$\left\|\sum_{i=1}^n X_{i,1}\right\|_p \le \sqrt{6p}\left(\sum_{i=1}^n \mathbb{E}[X_i^2]\right)^{1/2} + 20p\rho.$$

By Hoffmann-Jorgensen's inequality, Proposition 6.8 of (Ledoux and Talagrand, 1991), and by the choice of ρ,

$$\left\|\sum_{i=1}^n X_{i,2}\right\|_1 \le 2\left\|\sum_{i=1}^n |X_i|\mathbb{I}\{|X_i| \ge \rho\}\right\|_1 \le 16\|Z\|_1,$$

since

$$\mathbb{P}\left(\max_{1\le k\le n}\sum_{i=1}^k |X_i|\mathbb{I}\{|X_i| \ge \rho\} > 0\right) \le \mathbb{P}\left(Z \ge \rho\right) \le \frac{1}{8}.$$

Therefore, by Theorem 6.21 of (Ledoux and Talagrand, 1991),

$$\left\|\sum_{i=1}^n X_{i,2}\right\|_{\psi_\alpha} \le 17K_\alpha\|Z\|_{\psi_\alpha},$$

where the constant K_α is given in Theorem 6.21 of (Ledoux and Talagrand, 1991). Hence, for $p \ge 1$,

$$\left\|\sum_{i=1}^n X_{i,2}\right\|_p \le C_\alpha K_\alpha \max_{1\le i\le n}\|X_i\|_{\psi_\alpha}(\log(n+1))^{1/\alpha}p^{1/\alpha},$$

for some constant $C_\alpha > 0$ depending on α. Therefore, for $p \ge 1$,

$$\left\|\sum_{i=1}^n X_i\right\|_p \le \sqrt{6p}\left(\sum_{i=1}^n \mathbb{E}[X_i^2]\right)^{1/2} + C_\alpha K_\alpha \max_{1\le i\le n}\|X_i\|_{\psi_\alpha}(\log(n+1))^{1/\alpha}p^{1/\alpha},$$

for some constant $C_\alpha > 0$ (which may be different from the previous term). The proof is complete. ■

A.5 Proof of Lemma 21

Proof The proof follows the same technique as that of Lemma 20, which is also a combination of truncation and Hoffmann-Jorgensen's inequality, referring to (Kuchibhotla and Chakrabortty, 2022). Define

$$Z = \max_{1\le i\le n} |X_i|, \rho = 8\mathbb{E}[Z], X_{i,1} = X_i\mathbb{I}\{|X_i| \le \rho\} - \mathbb{E}[X_i\mathbb{I}\{|X_i| \le \rho\}], \text{ and } X_{i,2} = X - X_{i,1}.$$

Also by triangle inequality, for $p \ge 1$,

$$\left\|\sum_{i=1}^n X_i\right\|_p \le \left\|\sum_{i=1}^n X_{i,1}\right\|_p + \left\|\sum_{i=1}^n X_{i,2}\right\|_p.$$

Following the same argument as in the proof of Lemma 20, for $p \ge 1$,

$$\left\|\sum_{i=1}^n X_{i,1}\right\|_p \le \sqrt{6p}\left(\sum_{i=1}^n \mathbb{E}[X_i^2]\right)^{1/2} + 20p\rho. \tag{4}$$

By Hoffmann-Jorgensen's inequality, Proposition 6.8 of (Ledoux and Talagrand, 1991), and by the choice of ρ,

$$\left\|\sum_{i=1}^n X_{i,2}\right\|_1 \le 2\left\|\sum_{i=1}^n |X_i|\mathbb{I}\{|X_i| \ge \rho\}\right\|_1 \le 16\|Z\|_1,$$

since

$$\mathbb{P}\left(\max_{1\le k\le n}\sum_{i=1}^k |X_i|\mathbb{I}\{|X_i| \ge \rho\} > 0\right) \le \mathbb{P}\left(Z \ge \rho\right) \le \frac{1}{8}.$$

Therefore, by Theorem 6.21 of (Ledoux and Talagrand, 1991), with $\alpha = 1$,

$$\left\|\sum_{i=1}^n X_{i,2}\right\|_{\psi_1} \le K_1[16\|Z\|_{\psi_1} + \|Z\|_{\psi_1}] \le 17K_1\|Z\|_{\psi_1}.$$

By Problem 5 of Chapter 2.2 of (Van Der Vaart and Wellner, 1996), for $\alpha \ge 1$,

$$\|Z\|_{\psi_1} \le \|Z\|_{\psi_\alpha}(\log 2)^{1/\alpha - 1}$$

and so,

$$\left\|\sum_{i=1}^n X_{i,2}\right\|_{\psi_1} \le 17K_1\|Z\|_{\psi_\alpha}(\log 2)^{1/\alpha-1} \le C_\alpha(\log(n+1))^{1/\alpha}\max_{1\le i\le n}\|X_i\|_{\psi_\alpha}, \tag{5}$$

for some constant $C_\alpha > 0$ depending only on α. Therefore, combining inequalities (4) and (5) with $\rho \le 8C_\alpha(\log(n+1))^{1/\alpha}$ for $p \ge 1$

$$\left\|\sum_{i=1}^n X_i\right\|_p \le \sqrt{6p}\left(\sum_{i=1}^n \mathbb{E}[X_i^2]\right)^{1/2} + C_\alpha p(\log(n+1))^{1/\alpha}\max_{1\le i\le n}\|X_i\|_{\psi_\alpha},$$

for some constant $C_\alpha > 0$ depending only on α (which may be different from the previous term). The proof is complete. ■

A.6 Proof of Corollary 24

Proof We first consider the case $0 < \alpha \leq 1$. For any $t > 0$, by Markov's inequality

$$\begin{aligned}&\mathbb{P}\left(|f(X_1,...,X_n) - \mathbb{E}f(X_1,...,X_n)| \geq t\right) \leq \frac{\|f(X_1,...,X_n) - \mathbb{E}f(X_1,...,X_n)\|_p^p}{t^p} \\ &\leq \frac{2^p\left(\sqrt{6p}\left(\sum_{i=1}^n \sigma_i^2\right)^{1/2} + C_\alpha K_\alpha b(\log(n+1))^{1/\alpha}p^{1/\alpha}\right)^p}{t^p}.\end{aligned}$$

Setting t such that

$$\exp(-p) = 2^p\left(\sqrt{6p}\left(\sum_{i=1}^n \sigma_i^2\right)^{1/2} + C_\alpha K_\alpha b(\log(n+1))^{1/\alpha}p^{1/\alpha}\right)^p/t^p.$$

By the inequality $a + b \leq 2\max\{a, b\}$ for $a, b > 0$, it is clear that if $\sqrt{6p}\left(\sum_{i=1}^n \sigma_i^2\right)^{1/2} \geq C_\alpha K_\alpha b(\log(n+1))^{1/\alpha}p^{1/\alpha}$, we have

$$\mathbb{P}\left(|f(X_1,...,X_n) - \mathbb{E}f(X_1,...,X_n)| \geq 4e\sqrt{6p}\left(\sum_{i=1}^n \sigma_i^2\right)^{1/2}\right) \leq \exp(-p).$$

Put $4e\sqrt{6p}\left(\sum_{i=1}^n \sigma_i^2\right)^{1/2} = t$, we have

$$\mathbb{P}\left(|f(X_1,...,X_n) - \mathbb{E}f(X_1,...,X_n)| \geq t\right) \leq \exp\left(-\frac{t^2}{96e^2\sum_{i=1}^n \sigma_i^2}\right).$$

While if $\sqrt{6p}\left(\sum_{i=1}^n \sigma_i^2\right)^{1/2} \leq C_\alpha K_\alpha b(\log(n+1))^{1/\alpha}p^{1/\alpha}$

$$\mathbb{P}\left(|f(X_1,...,X_n) - \mathbb{E}f(X_1,...,X_n)| \geq 4eC_\alpha K_\alpha b(\log(n+1))^{1/\alpha}p^{1/\alpha}\right) \leq \exp(-p).$$

Put $4eC_\alpha K_\alpha b(\log(n+1))^{1/\alpha}p^{1/\alpha} = t$, we have

$$\mathbb{P}\left(|f(X_1,...,X_n) - \mathbb{E}f(X_1,...,X_n)| \geq t\right) \leq \exp\left(-\frac{t^\alpha}{(4eC_\alpha K_\alpha)^\alpha \log(n+1)b^\alpha}\right).$$

Combining the two cases, the proof of the case $0 < \alpha \leq 1$ is complete.

We then consider the case $\alpha \geq 1$. Following a similar pattern, if $\sqrt{6p}\left(\sum_{i=1}^n \sigma_i^2\right)^{1/2} \geq C_\alpha pb(\log(n+1))^{1/\alpha}$, we have

$$\mathbb{P}\left(|f(X_1,...,X_n) - \mathbb{E}f(X_1,...,X_n)| \geq 4e\sqrt{6p}\left(\sum_{i=1}^n \sigma_i^2\right)^{1/2}\right) \leq \exp(-p).$$

Put $4e\sqrt{6p}\left(\sum_{i=1}^n \sigma_i^2\right)^{1/2} = t$, we have

$$\mathbb{P}\left(|f(X_1,...,X_n) - \mathbb{E}f(X_1,...,X_n)| \geq t\right) \leq \exp\left(-\frac{t^2}{96e^2\sum_{i=1}^n \sigma_i^2}\right).$$

While if $\sqrt{6p}\left(\sum_{i=1}^n \sigma_i^2\right)^{1/2} \leq C_\alpha pb(\log(n+1))^{1/\alpha}$

$$\mathbb{P}\left(|f(X_1,...,X_n) - \mathbb{E}f(X_1,...,X_n)| \geq 4eC_\alpha pb(\log(n+1))^{1/\alpha}\right) \leq \exp(-p).$$

Put $4eC_\alpha pb(\log(n+1))^{1/\alpha} = t$, we have

$$\mathbb{P}\left(|f(X_1,...,X_n) - \mathbb{E}f(X_1,...,X_n)| \geq t\right) \leq \exp\left(-\frac{t}{4eC_\alpha b(\log(n+1))^{1/\alpha}}\right).$$

Combining the two cases, the proof of the case $\alpha \geq 1$ is complete. ■

A.7 Proof of Corollary 29

Proof For any $t > 0$, by Markov's inequality

$$\mathbb{P}\left(|f(X_1,...,X_n) - \mathbb{E}f(X_1,...,X_n)| \geq t\right) \leq \frac{\|f(X_1,...,X_n) - \mathbb{E}f(X_1,...,X_n)\|_p^p}{t^p}$$
$$\leq \frac{\left(2\sqrt{2\kappa(2+\theta)p}\left(\sum_{i=1}^n \sigma^2\right)^{\frac{1}{2}} + 2p\kappa\sqrt{1+\frac{1}{\theta}}\left(\sum_{i=1}^n b_i\right)^{\frac{1}{p}}\right)^p}{t^p}.$$

If $2p\kappa\sqrt{1+\frac{1}{\theta}}\left(\sum_{i=1}^n b_i\right)^{\frac{1}{p}} \geq 2\sqrt{2\kappa(2+\theta)p}\left(\sum_{i=1}^n \sigma^2\right)^{\frac{1}{2}}$, we have

$$\mathbb{P}\left(|f(X_1,...,X_n) - \mathbb{E}f(X_1,...,X_n)| \geq t\right) \leq \frac{(4p\kappa\sqrt{1+\frac{1}{\theta}})^p \sum_{i=1}^n b_i}{t^p}.$$

While if $2p\kappa\sqrt{1+\frac{1}{\theta}}\left(\sum_{i=1}^n b_i\right)^{\frac{1}{p}} \leq 2\sqrt{2\kappa(2+\theta)p}\left(\sum_{i=1}^n \sigma^2\right)^{\frac{1}{2}}$, we have

$$\mathbb{P}\left(|f(X_1,...,X_n) - \mathbb{E}f(X_1,...,X_n)| \geq t\right) \leq \frac{(4\sqrt{2\kappa(2+\theta)p})^p \left(\sum_{i=1}^n \sigma^2\right)^{\frac{p}{2}}}{t^p}.$$

By setting t such that

$$\exp(-p) = (4\sqrt{2\kappa(2+\theta)p})^p \left(\sum_{i=1}^n \sigma^2\right)^{\frac{p}{2}} / t^p,$$

we have

$$\mathbb{P}\left(|f(X_1,...,X_n) - \mathbb{E}f(X_1,...,X_n)| \geq e4\sqrt{2\kappa(2+\theta)p}\left(\sum_{i=1}^n \sigma^2\right)^{\frac{1}{2}}\right) \leq \exp(-p).$$

Put $e4\sqrt{2\kappa(2+\theta)p}\left(\sum_{i=1}^n \sigma^2\right)^{\frac{1}{2}} = t$, we have

$$\mathbb{P}\left(|f(X_1,...,X_n) - \mathbb{E}f(X_1,...,X_n)| \geq t\right) \leq \exp\left(-\frac{t^2}{16e^2(2\kappa(2+\theta))\sum_{i=1}^n \sigma_i^2}\right).$$

Combining the two cases, the proof is complete. ■

A.8 Proof of Corollary 34

Proof For any $t > 0$, by Markov's inequality

$$\begin{aligned}&\mathbb{P}\left(|f(X_1, ..., X_n) - \mathbb{E}f(X_1, ..., X_n)| \geq t\right)\\ \leq &\frac{\|f(X_1, ..., X_n) - \mathbb{E}f(X_1, ..., X_n)\|_p^p}{t^p} \leq \frac{(6\sqrt{2np}b^{1/p})^p}{t^p}.\end{aligned}$$

The proof is complete. ■

Appendix B. Proofs of Section 4

To proceed, we introduce a lemma.

Lemma 42 (Lemma 6 in Maurer and Pontil (2021)) *Let X, X' be i.i.d. with values in $\mathcal{X}$, $\phi : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is measurable. Then*

$$\left\|\mathbb{E}[\phi(X, X')|X]\right\|_p \leq \|\phi(X, X')\|_p.$$

B.1 Proof of Theorem 35

Proof (i) We look at the function $f(x) = \|\sum_{i=1}^n x_i\|$. Then

$$|f_k(X)(x)| = \left|\left\|\sum_{i\neq k} x_i + X_k\right\| - \mathbb{E}\left[\left\|\sum_{i\neq k} x_i + X_k'\right\|\right]\right| \leq \mathbb{E}\left[\left\|X_k - X_k'\right\| |X_k\right].$$

Observe that the upper bound on $f_k(X)(x)$ is independent of x. By Lemma 42, we get $\|\mathbb{E}[\|X_k - X_k'\||X]\|_p \leq 2\|\|X_k\|\|_p$ and thus $\|f_k(X)(x)\|_p \leq 2\|\|X_k\|\|_p$. By Theorem 2.1 in (Vladimirova et al., 2020), we also have $\|f_k(X)(x)\|_{\psi_\alpha} \leq 2\|\|X_k\|\|_{\psi_\alpha}$.

Plugging these bounds into Corollary 24, if $0 < \alpha \leq 1$

$$\begin{aligned}&\mathbb{P}\left(\left|\left\|\sum_{i=1}^n X_i\right\| - \mathbb{E}\left\|\sum_{i=1}^n X_i\right\|\right| \geq t\right)\\ \leq &\exp\left(-\min\left\{\frac{t^2}{96e^2 4\sum_{i=1}^n \|\|X_i\|\|_2^2}, \frac{t^\alpha}{(4eC_\alpha K_\alpha)^\alpha \log(n+1)(2\max_k \|\|X_k\|\|_{\psi_\alpha})^\alpha}\right\}\right),\end{aligned}$$

if $\alpha \geq 1$

$$\begin{aligned}&\mathbb{P}\left(\left|\left\|\sum_{i=1}^n X_i\right\| - \mathbb{E}\left\|\sum_{i=1}^n X_i\right\|\right| \geq t\right)\\ \leq &\exp\left(-\min\left\{\frac{t^2}{96e^2 4\sum_{i=1}^n \|\|X_i\|\|_2^2}, \frac{t}{4eC_\alpha(\log(n+1))^{1/\alpha} 2\max_k \|\|X_k\|\|_{\psi_\alpha}}\right\}\right).\end{aligned}$$

(ii) We look at the function $f(x) = \|\sum_{i=1}^n (x_i - \mathbb{E}X_1')\|$. By the i.i.d. property of the X_i and Jensen's inequality, we have

$$\mathbb{E}\left\|\sum_{i=1}^n (X_i - \mathbb{E}X_i')\right\| \leq \left(n\mathbb{E}\left[\left\|X_1 - \mathbb{E}X_1'\right\|^2\right]\right)^{1/2} \leq \sqrt{n}\|\|X_1\|\|_2.$$

Then we have

$$|f_k(X)(x)| = \left| \left\| \sum_{i\neq k} x_i + X_k - n\mathbb{E}X_1' \right\| - \mathbb{E}\left[\left\| \sum_{i\neq k} x_i + X_k' - n\mathbb{E}X_1' \right\| \right] \right| \leq \mathbb{E}\left[\left\| X_k - X_k' \right\| | X_k \right].$$

Similarly, by Lemma 42, we have $\|f_k(X)(x)\|_p \leq 2\|\|X_k\|\|_p$ and $\|f_k(X)(x)\|_{\psi_\alpha} \leq 2\|\|X_k\|\|_{\psi_\alpha}$.

Plugging these bounds into Corollary 24, if $0 < \alpha \leq 1$

$$\begin{aligned}
&\mathbb{P}\left(\left\| \sum_{i=1}^n (X_i - \mathbb{E}X_1') \right\| > t + \sqrt{n}\|\|X_1\|\|_2 \right) \leq \mathbb{P}\left(\left\| \sum_{i=1}^n (X_i - \mathbb{E}X_1') \right\| > t + \mathbb{E}\left\| \sum_{i=1}^n (X_i - \mathbb{E}X_i') \right\| \right) \\
&\leq \exp\left(-\min\left\{ \frac{t^2}{96e^2 4\sum_{i=1}^n \|\|X_i\|\|_2^2}, \frac{t^\alpha}{(4eC_\alpha K_\alpha)^\alpha \log(n+1)(2\max_k \|\|X_k\|\|_{\psi_\alpha})^\alpha} \right\} \right) \\
&= \exp\left(-\min\left\{ \frac{t^2}{96e^2 4n\|\|X_1\|\|_2^2}, \frac{t^\alpha}{(4eC_\alpha K_\alpha)^\alpha \log(n+1)(2\|\|X_1\|\|_{\psi_\alpha})^\alpha} \right\} \right),
\end{aligned}$$

if $\alpha \geq 1$

$$\begin{aligned}
&\mathbb{P}\left(\left\| \sum_{i=1}^n (X_i - \mathbb{E}X_1') \right\| > t + \sqrt{n}\|\|X_1\|\|_2 \right) \leq \mathbb{P}\left(\left\| \sum_{i=1}^n (X_i - \mathbb{E}X_1') \right\| > t + \mathbb{E}\left\| \sum_{i=1}^n (X_i - \mathbb{E}X_i') \right\| \right) \\
&\leq \exp\left(-\min\left\{ \frac{t^2}{96e^2 4\sum_{i=1}^n \|\|X_i\|\|_2^2}, \frac{t}{4eC_\alpha (\log(n+1))^{1/\alpha} 2\max_k \|\|X_k\|\|_{\psi_\alpha}} \right\} \right) \\
&= \exp\left(-\min\left\{ \frac{t^2}{96e^2 4n\|\|X_1\|\|_2^2}, \frac{t}{4eC_\alpha (\log(n+1))^{1/\alpha} 2\|\|X_1\|\|_{\psi_\alpha}} \right\} \right).
\end{aligned}$$

The proof is complete. ■

B.2 Proof of Corollary 37

Proof If H is a Hilbert space, then the Hilbert space of Hilbert-Schmidt operators $HS(H)$ is the set of bounded operators T on H satisfying $\|T\|_{HS} = \sqrt{\sum_{i,j} \langle Te_i, e_j\rangle_H^2} < \infty$ with inner product $\langle T, S\rangle_{HS} = \sum_{i,j} \langle Te_i, e_j\rangle_H \langle Se_i, e_j\rangle_H$, where (e_i) is an orthonormal basis. For $x \in H$ the operator $Q_x \in HS(H)$ is defined by $Q_x y = \langle y, x\rangle x$, and it can be shown that $\|Q_x\|_{HS} = \|x\|_H^2$.

Working within the space of Hilbert-Schmidt operators, $HS(H)$, allows us to express $\ell(P, x) = \|Q_x\|_{HS} - \langle P, Q_x\rangle_{HS}$. Then

$$\begin{aligned}
&\sup_{P\in\mathcal{P}_d} \frac{1}{n}\sum_i \mathbb{E}[\ell(P, X_1)] - \ell(P, X_i) \\
&= \sup_{P\in\mathcal{P}_d} \left\langle P, \frac{1}{n}\sum_i (Q_{X_i} - \mathbb{E}[Q_{X_i}]) \right\rangle_{HS} + \left(\mathbb{E}\|Q_{X_i}\|_{HS} - \frac{1}{n}\sum_i \|Q_{X_i}\|_{HS} \right).
\end{aligned}$$

For the first term, since for $P \in \mathcal{P}_d$ we have $\|P\|_{HS} = \sqrt{d}$, we can use Cauchy-Schwarz and use Theorem 35.(ii) for the random variable $\frac{\sqrt{d}}{n} Q_{X_i}$ to give the following bound: if $0 < \alpha \leq 1$

$$\begin{aligned}
&\mathbb{P}\left(\sqrt{d}\left\|\frac{1}{n}\sum_i (Q_{X_i} - \mathbb{E}[Q_{X_i}])\right\|_{HS} > t + \frac{\sqrt{d}}{\sqrt{n}}\|\|Q_{X_1}\|_{HS}\|_2\right) \\
\leq &\exp\left(-\min\left\{\frac{nt^2}{96e^2 4d\|\|Q_{X_1}\|_{HS}\|_2^2}, \frac{t^\alpha}{(4eC_\alpha K_\alpha)^\alpha \log(n+1)(2\sqrt{d}\|\|Q_{X_1}\|_{HS}\|_{\psi_\alpha}/n)^\alpha}\right\}\right),
\end{aligned}$$

if $\alpha \geq 1$

$$\begin{aligned}
&\mathbb{P}\left(\sqrt{d}\left\|\frac{1}{n}\sum_i (Q_{X_i} - \mathbb{E}[Q_{X_i}])\right\|_{HS} > t + \frac{\sqrt{d}}{\sqrt{n}}\|\|Q_{X_1}\|_{HS}\|_2\right) \\
\leq &\exp\left(-\min\left\{\frac{nt^2}{96e^2 4d\|\|Q_{X_1}\|_{HS}\|_2^2}, \frac{nt}{4eC_\alpha(\log(n+1))^{1/\alpha} 2\sqrt{d}\|\|Q_{X_1}\|_{HS}\|_{\psi_\alpha}}\right\}\right).
\end{aligned}$$

Using (2) and (3), the second term can be bounded by applying the same result to the random vectors $\|Q_{X_i}\|_{HS}$. Note that $\|\|Q_{X_1}\|_{HS}\|_{\psi_\alpha} = \|\|X_1\|^2\|_{\psi_\alpha}$. Finally, the result follows from combining both bounds in a union bound. ■

B.3 Proof of Theorem 38

Proof The vector space

$$\mathcal{B} = \left\{p : \mathcal{G} \to \mathbb{R} : \sup_{g\in\mathcal{G}} |p(g)| < \infty\right\}$$

becomes a normed space with norm $\|p\|_{\mathcal{B}} = \sup_{g\in\mathcal{G}} |p(g)|$. For each X_i define $\bar{X}_i \in \mathcal{B}$ by $\bar{X}_i(g) = \frac{1}{n}\left(g(X_i) - \mathbb{E}[g(X_i')]\right)$. Then the $\bar{X}_i$ are zero mean random variable in $\mathcal{B}$ and

$$\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_i g(X_i) - \mathbb{E}[g(X_i')] = \left\|\sum_i \bar{X}_i\right\|_{\mathcal{B}}.$$

With Lemma 42 and the i.i.d. assumption, we have

$$\begin{aligned}
\|\|\bar{X}_i\|_{\mathcal{B}}\|_p &= \frac{1}{n}\left\|\sup_g(\mathbb{E}[g(X_i) - g(X_i')]|X)\right\|_p \\
&\leq \frac{L}{n}\left\|\mathbb{E}[\|X_i - X_i'\|]|X\right\|_p \leq \frac{2L}{n}\|\|X_i\|\|_p = \frac{2L}{n}\|\|X_1\|\|_p,
\end{aligned}$$

where the first inequality uses the Lipschitz condition. By Theorem 2.1 in (Vladimirova et al., 2020), we also have $\|\|\bar{X}_i\|_{\mathcal{B}}\|_{\psi_\alpha} \leq \frac{2L}{n}\|\|X_1\|\|_{\psi_\alpha}$. Thus, from Theorem 35.(ii), we have

if $0 < \alpha \leq 1$

$$\mathbb{P}\left(\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_i g(X_i) - \mathbb{E}[g(X_i')] - \mathbb{E}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_i g(X_i) - \mathbb{E}[g(X_i')]\right] > t\right)$$
$$\leq \exp\left(-\min\left\{\frac{t^2}{96e^2\frac{16L^2}{n}|||X_1|||_2^2}, \frac{t^\alpha}{(4eC_\alpha K_\alpha)^\alpha \log(n+1)(\frac{4L}{n}|||X_1|||_{\psi_\alpha})^\alpha}\right\}\right),$$

if $\alpha \geq 1$

$$\mathbb{P}\left(\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_i g(X_i) - \mathbb{E}[g(X_i')] - \mathbb{E}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_i g(X_i) - \mathbb{E}[g(X_i')]\right] > t\right)$$
$$\leq \exp\left(-\min\left\{\frac{t^2}{96e^2\frac{16L^2}{n}|||X_1|||_2^2}, \frac{t}{4eC_\alpha(\log(n+1))^{1/\alpha}\frac{4L}{n}|||X_1|||_{\psi_\alpha}}\right\}\right).$$

The proof is complete. ■

B.4 Proof of Corollary 40

Proof In the setting we considered, $\mathcal{X}$ becomes a Banach space with the norm $\|(x,z)\| = L\|x\|_H + |z|$. It is clear that $|||(X_1,Z_1)|||_{\psi_\alpha} \leq L|||X_1|||_{\psi_\alpha} + |||Z_1|||_{\psi_\alpha}$. Then for $g \in \mathcal{G}$

$$\begin{aligned} g(x,z) - g(x',z') =& \ell(\langle w, x\rangle - z) - \ell(\langle w, x'\rangle - z') \\ \leq& L\|x - x'\|_H + |z - z'| \leq \|(x,z) - (x',z')\|, \end{aligned}$$

so $\mathcal{G}$ is uniformly Lipschitz with constant 1. Also for an i.i.d. sample $(X,Z) \in \mathcal{X}^n$, by employing the Lipschitz property of ℓ, combining the triangle inequality and Jensen's inequality, it becomes straightforward to observe that

$$\begin{aligned} \mathcal{R}(\mathcal{G}) &\leq \frac{2}{n}\mathbb{E}\left(L\sqrt{\sum_i \|X_i\|_H^2} + \sqrt{\sum_i |Z_i|^2}\right) \\ &\leq \frac{2}{n}\left(L\sqrt{\mathbb{E}\sum_i \|X_i\|_H^2} + \sqrt{\mathbb{E}\sum_i |Z_i|^2}\right) \\ &\leq \frac{2}{\sqrt{n}}\left(L|||X_1|||_2 + |||Z_1|||_2\right). \end{aligned}$$

Substituting these results into Theorem 38 gives the final inequalities. ■

B.5 Proof of Theorem 41

Proof We derive

$$\begin{aligned}\|f_k(X)(x)\|_p =& \|f(x_1,...,X_k,x_{k+1},...,x_n) - \mathbb{E}[f(x_1,...,X_k,x_{k+1},...,x_n)]\|_p \\ =& \left\|\mathbb{E}[f(x_1,...,X_k,x_{k+1},...,x_n) - f(x_1,...,X_k',x_{k+1},...,x_n)|X_k]\right\|_p \\ \leq& L\|\mathbb{E}[d(X_k,X_k')|X_k]\|_p \\ \leq& L\|d(X_k,X_k')\|_p,\end{aligned}$$

where the first inequality uses the Lipschitz condition and the last inequality uses Lemma 42. By Theorem 2.1 in (Vladimirova et al., 2020), we also have $\|f_k(X)(x)\|_{\psi_\alpha} \leq L\|d(X_k,X_k')\|_{\psi_\alpha}$.

Plugging these bounds into Corollary 24, we obtain if $0 < \alpha \leq 1$

$$\begin{aligned}&\mathbb{P}\left(|f(X) - \mathbb{E}f(X')| \geq t\right) \\ \leq& \exp\left(-\min\left\{\frac{t^2}{96e^2\sum_{i=1}^n L^2\|d(X_i,X_i')\|_2^2}, \frac{t^\alpha}{(4eC_\alpha K_\alpha)^\alpha \log(n+1)L^\alpha \max_k \|d(X_k,X_k')\|_{\psi_\alpha}^\alpha}\right\}\right),\end{aligned}$$

if $\alpha \geq 1$

$$\begin{aligned}&\mathbb{P}\left(|f(X) - \mathbb{E}f(X')| \geq t\right) \\ \leq& \exp\left(-\min\left\{\frac{t^2}{96e^2\sum_{i=1}^n L^2\|d(X_i,X_i')\|_2^2}, \frac{t}{4eC_\alpha L \max_k \|d(X_k,X_k')\|_{\psi_\alpha}(\log(n+1))^{1/\alpha}}\right\}\right).\end{aligned}$$

The proof is complete. ■

References

Radoslaw Adamczak. A tail inequality for suprema of unbounded empirical processes with applications to markov chains. *Electronic Journal of Probability*, 13:1000–1034, 2007.

Pierre Alquier. Pac-bayesian bounds for randomized empirical risk minimizers. *Mathematical Methods of Statistics*, 17:279–304, 2008.

Milad Bakhshizadeh, Arian Maleki, and Shirin Jalali. Using black-box compression algorithms for phase retrieval. *IEEE Transactions on Information Theory*, 66(12):7978–8001, 2020.

Milad Bakhshizadeh, Arian Maleki, and Victor H De La Pena. Sharp concentration results for heavy-tailed distributions. *Information and Inference: A Journal of the IMA*, 12(3): 1655–1685, 2023.

Peter Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *Annals of Statistics*, 33(4):1497–1537, 2005.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Heejong Bong and Arun Kumar Kuchibhotla. Tight concentration inequality for sub-weibull random variables with generalized bernstien orlicz norm. *arXiv preprint arXiv:2302.03850*, 2023.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. Concentration inequalities using the entropy method. *The Annals of Probability*, 31(3):1583–1614, 2003.

Stéphane Boucheron, Olivier Bousquet, Gábor Lugosi, and Pascal Massart. Moment inequalities for functions of independent random variables. *Annals of Probability*, 33: 514–560, 2005.

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

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

Olivier Bousquet, Yegor Klochkov, and Nikita Zhivotovskiy. Sharper bounds for uniformly stable algorithms. In *Conference on Learning Theory*, pages 610–626. PMLR, 2020.

Ioar Casado, Luis A Ortega, Andrés R Masegosa, and Aritz Pérez. Pac-bayes-chernoff bounds for unbounded losses. *arXiv preprint arXiv:2401.01148*, 2024.

Ben Chugg, Hongjian Wang, and Aaditya Ramdas. A unified recipe for deriving (time-uniform) pac-bayes bounds. *Journal of Machine Learning Research*, 24(372):1–61, 2023.

Richard Combes. An extension of mcdiarmid's inequality. *arXiv preprint arXiv:1511.05240*, 2015.

Corinna Cortes, Spencer Greenberg, and Mehryar Mohri. Relative deviation learning bounds and generalization with unbounded loss functions. *Annals of Mathematics and Artificial Intelligence*, 85:45–70, 2019.

Corinna Cortes, Mehryar Mohri, and Ananda Theertha Suresh. Relative deviation margin bounds. In *International Conference on Machine Learning*, pages 2122–2131, 2021.

Sergey Foss, Dmitry Korshunov, Stan Zachary, et al. *An introduction to heavy-tailed and subexponential distributions*, volume 6. Springer, 2011.

Olivier Guédon, Alexander E Litvak, Alain Pajor, and Nicole Tomczak-Jaegermann. Restricted isometry property for random matrices with heavy-tailed columns. *Comptes Rendus Mathematique*, 352(5):431–434, 2014.

Mert Gurbuzbalaban, Umut Simsekli, and Lingjiong Zhu. The heavy-tail phenomenon in sgd. In *International Conference on Machine Learning*, pages 3964–3975, 2021.

Maxime Haddouche and Benjamin Guedj. Pac-bayes generalisation bounds for heavy-tailed losses through supermartingales. *Transactions on Machine Learning Research*, 2023.

Maxime Haddouche, Benjamin Guedj, Omar Rivasplata, and John Shawe-Taylor. Pac-bayes unleashed: Generalisation bounds with unbounded losses. *Entropy*, 23(10):1330, 2021.

Matthew Holland. Pac-bayes under potentially heavy tails. In *Advances in Neural Information Processing Systems*, 2019.

Aryeh Kontorovich. Concentration in unbounded metric spaces and algorithmic stability. In *International Conference on Machine Learning*, pages 28–36, 2014.

Arun Kumar Kuchibhotla and Abhishek Chakrabortty. Moving beyond sub-gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression. *Information and Inference: A Journal of the IMA*, 11(4):1389–1456, 2022.

Samuel Kutin. Extensions to mcdiarmid's inequality when differences are bounded with high probability. *Dept. Comput. Sci., Univ. Chicago, Chicago, IL, USA, Tech. Rep. TR-2002-04*, 2002.

Samuel Kutin and Partha Niyogi. Almost-everywhere algorithmic stability and generalization error. *arXiv preprint arXiv:1301.0579*, 2002.

Michel Ledoux. *The concentration of measure phenomenon*. Number 89. American Mathematical Soc., 2001.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*, volume 23. Springer Science & Business Media, 1991.

Zhipeng Lou, Wanrong Zhu, and Wei Biao Wu. Beyond sub-gaussian noises: Sharp concentration analysis for stochastic gradient descent. *Journal of Machine Learning Research*, 23:1–22, 2022.

Andreas Maurer. A bernstein-type inequality for functions of bounded interaction. *Bernoulli*, 25(2):1451–1471, 2019.

Andreas Maurer and Massimiliano Pontil. Empirical bounds for functions with weak interactions. In *Conference On Learning Theory*, pages 987–1010, 2018.

Andreas Maurer and Massimiliano Pontil. Uniform concentration and symmetrization for weak interactions. In *Conference on Learning Theory*, pages 2372–2387, 2019.

Andreas Maurer and Massimiliano Pontil. Concentration inequalities under sub-gaussian and sub-exponential conditions. *Advances in Neural Information Processing Systems*, 34: 7588–7597, 2021.

Colin McDiarmid. Concentration. In *Probabilistic methods for algorithmic discrete mathematics*, pages 195–248. Springer, 1998.

Ron Meir and Tong Zhang. Generalization error bounds for bayesian mixture algorithms. *Journal of Machine Learning Research*, 4(Oct):839–860, 2003.

Sergey V Nagaev. Large deviations of sums of independent random variables. *The Annals of Probability*, pages 745–789, 1979.

Jayakrishnan U Nair. *Scheduling for heavy-tailed and light-tailed workloads in queueing systems*. California Institute of Technology, 2012.

Maxim Raginsky and Igal Sason. Concentration of measure inequalities and their communication and information-theoretic applications. *arXiv preprint arXiv:1510.02947*, 2015.

Maxim Raginsky and Igal Sason. *Concentration of Measure Inequalities in Information Theory, Communications, and Coding*. 2018.

Yao-Feng Ren and Han-Ying Liang. On the best constant in marcinkiewicz–zygmund inequality. *Statistics & probability letters*, 53(3):227–233, 2001.

Omar Rivasplata, Ilja Kuzborskij, Csaba Szepesvári, and John Shawe-Taylor. Pac-bayes analysis beyond the usual bounds. In *Advances in Neural Information Processing Systems*, pages 16833–16845, 2020.

Steve Smale and Ding-Xuan Zhou. Learning theory estimates via integral operators and their approximations. *Constructive approximation*, 26(2):153–172, 2007.

Aad W Van Der Vaart and Jon A Wellner. *Weak convergence and empirical processes*. Springer, 1996.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Mariia Vladimirova, Stéphane Girard, Hien Nguyen, and Julyan Arbel. Sub-weibull distributions: Generalizing sub-gaussian and sub-exponential properties to heavier tailed distributions. *Stat*, 9(1):e318, 2020.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Lutz Warnke. On the method of typical bounded differences. *Combinatorics, Probability and Computing*, 25(2):269–299, 2016.

Kam Chung Wong, Zifan Li, and Ambuj Tewari. Lasso guarantees for β-mixing heavy-tailed time series. *The Annals of Statistics*, 48(2):1124–1142, 2020.

Huiming Zhang and Haoyu Wei. Sharper sub-weibull concentrations. *Mathematics*, 10(13): 2252, 2022.