Distribution-dependent McDiarmid-type Inequalities for Functions of Unbounded Interaction

Shaojie Li [1 2] **Yong Liu** [1 2 *]

Abstract

The concentration of measure inequalities serves an essential role in statistics and machine learning. This paper gives unbounded analogues of the McDiarmid-type exponential inequalities for three popular classes of distributions, namely sub-Gaussian, sub-exponential and heavy-tailed distributions. The inequalities in the sub-Gaussian and sub-exponential cases are distribution-dependent compared with the recent results, and the inequalities in the heavy-tailed case are not available in the previous works. The usefulness of the inequalities is illustrated through applications to the sample mean, U-statistics and V-statistics.

1. Introduction

Concentration-of-measure inequalities are studied in order to understand the fluctuations of complicated random objects. These inequalities have made great progress over the years, playing a significant role in various fields which include functional analysis, high-dimensional geometry, high-dimensional probability and statistics, information theory, machine learning, statistical physics, stochastic analysis, and theoretical computer science (Wainwright, 2019; Ledoux, 2001; Boucheron et al., 2013; Raginsky & Sason, 2018). Specifically, concentration-of-measure inequalities provide upper bounds on the probability that a random variable deviates from its mean, median or any other typical value by a given amount.

Among these concentration inequalities, McDiarmid's inequality is a classical benchmark one, which works not only for sums but for general functions of independent random variables. It has proved to be useful in a number of applications, such as algorithmic stability (Bousquet & Elisseeff, 2002; Bousquet et al., 2020) and suprema of empirical processes (Bartlett & Mendelson, 2002; Maurer & Pontil, 2021; Maurer, 2006). McDiarmid's inequality (McDiarmid, 1998), also called bounded difference inequality, states that

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

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. Some extensions of McDiarmid's inequality to non-independent variables have also been considered (Marton, 1996; Chazottes et al., 2007; Kontorovich & Ramanan, 2008; Zhang, 2022; Paulin, 2015). Recently, a line of work by Maurer & Pontil (2019); Maurer (2019) has also introduced inequalities for general functions of independent random variables, which is Bernstein-type inequality and capable of estimating nonlinear statistics. These inequalities are pretty attractive and useful, however, unfortunately, they require the boundedness of functions, imposing inherent limitations on their applicability to unbounded loss functions.

The concentration properties of unbounded functions become important in many settings, such as signal processing (Bakhshizadeh et al., 2020b;a), neural networks (Vladimirova et al., 2020; 2019; Torralba et al., 2008), stochastic optimization (Gurbuzbalaban et al., 2021; 2022; Barsbey et al., 2021; Simsekli et al., 2019), sample bias correction (Dudík et al., 2005), domain adaptation (Cortes & Mohri, 2014; Ben-David et al., 2006; Mansour et al., 2009), boosting (Dasgupta & Long, 2003), and importance-weighting (Cortes et al., 2019; 2021). To counter this difficulty, some concentration inequalities for general functions of unbounded random variables have been proposed (Kutin, 2002; Combes, 2015; Meir & Zhang, 2003; Kontorovich, 2014; Maurer & Pontil, 2021). Among these works, Kutin (2002); Kutin & Niyogi (2002) prove two extensions of McDiarmid's inequality for strongly and weakly difference-bounded functions. However, their approach entails complex statement and proof, and their conditions are too restrictive in practice, see a discussion in (Kontorovich, 2014).

[*]Corresponding Author [1]Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China [2]Beijing Key Laboratory of Big Data Management and Analysis Methods, Beijing, China. Correspondence to: Yong Liu <liuyonggsai@ruc.edu.cn>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

Then, Meir & Zhang (2003); Kontorovich (2014) give extensions of McDiarmid's inequality for sub-Gaussian distributions. Combes (2015) proposes a somewhat different extension of McDiarmid's inequality for functions with bounded differences on a high probability set and no restriction outside this set. Very recently, Maurer & Pontil (2021), the most related work to this paper, provide a more applicable inequality than the ones in (Meir & Zhang, 2003; Kontorovich, 2014) in the sub-Gaussian case and further study the heavier sub-exponential distributions, whose results can be seen as unbounded analogues of the McDiarmid's inequality under the sub-Gaussian and sub-exponential conditions.

Although very powerful, the results of Maurer & Pontil (2021) suffer from the worst-case value of the sample space. Specifically, their variance proxy term depends on the worst-case choice of the configuration of sample space, refer to Eqs. (1) and (2); we will elaborate upon this in Section 3. In this paper, we propose distribution-dependent concentration inequalities by replacing the worst-case value (i.e. supremum) in Maurer & Pontil (2021) with an expectation. This replacement is possible if the sum of conditional variance proxies has the suitable properties of concentration around its mean, refer to Lemmas A.6 and A.7. To this end, we should control the interaction functional, which indicates that the variation of f in any given argument must not depend too much on other arguments, refer to Definition 2.3. This paper's principal contributions include providing such concentration inequalities for sub-Gaussian, sub-exponential and heavy-tailed distributions, where the McDiarmid-type exponential inequalities of heavy-tailed distributions haven't been studied before to our knowledge. Not only that, we give corresponding refined inequalities for sub-Gaussian, sub-exponential and heavy-tailed distributions by considering a weaker interaction functional, refer to Definition 5.1. Some interesting applications, the sample mean, U-statistics and V-statistics, are provided to illustrate the usefulness of the inequalities.

The paper is organized as follows. In Section 2 the preliminaries relevant to our discussion are presented. In Section 3 we provide our concentration inequalities, separated by sub-Gaussian, sub-exponential and heavy-tailed distributions. Section 4 is devoted to concrete applications (the sample mean, U-statistics and V-statistics). In Section 5 we derive refined concentration inequalities. Section 6 concludes this paper. Finally, all the proofs are given in the Appendix.

2. Preliminaries

We use uppercase letters to present random variables and vector of random variables, and 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. We consider that f is a function $f : \mathcal{X}^n \to \mathbb{R}$. In this paper, we are interested in studying the concentration of the random variable $f(X)$ with respect to (w.r.t.) its expectation, i.e.,

$$\mathbb{P}\left(f(X) - \mathbb{E}[f(X')] > t\right), \quad \forall t > 0.$$

To proceed, we need some notations 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 2.1. (Maurer, 2019) For fixed $k \in \{1, ..., n\}$ and $y, y' \in \mathcal{X}$ define the substitution operator S^k_y and the difference operator $D^k_{y,y'}$ by

$$(S^k_y f)(x_1, ..., x_n) = f(x_1, ..., x_{k-1}, y, x_{k+1}, ..., x_n)$$

and

$$D^k_{y,y'} = S^k_y - S^k_{y'}.$$

We also give a definition on the centered conditional version of f.

Definition 2.2 (Definition 1 in (Maurer & Pontil, 2021))**.** If $f : \mathcal{X}^n \to \mathbb{R}$, $x = (x_1, ..., x_n) \in \mathcal{X}^n$ 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

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

$f_k(X)$ can be seen as a random-variable-valued-function $f_k(X) : x \in \mathcal{X}^n \to f_k(X)(x)$. It is clear that S^k_y, $D^k_{y,y'}$ and $f_k(X)$ does not depend on the k-th coordinate of x. For instance, consider the summation function $f(x) = \sum_{i=1}^n x_i$, then $f_k(X)(x) = X_k - \mathbb{E}[X_k]$ is independent of x.

We also introduce some notations of the norm. If $\|\cdot\|_a$ is any given norm on random variables, then $\|f_k(X)\|_a(x) := \|f_k(X)(x)\|_a$ defines a non-negative real-valued function $\|f_k(X)\|_a$ on $\mathcal{X}^n$. Thus $\|f_k(X)\|_a(X)$ is also a random variable. If X' is i.i.d. to X then $\|f_k(X)\|_a$ is the same function as $\|f_k(X')\|_a$ and $\|f_k(X)\|_a(X')$ is i.i.d. to $\|f_k(X)\|_a(X)$. Note that

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

The L_p norms of any real random variable Z is $\|Z\|_p = (\mathbb{E}[|Z|^p])^{\frac{1}{p}}$.

Finally, we introduce the interaction functional of f, which represents the extent to which the variation of f in any given argument depends on other arguments.

Definition 2.3. The interaction functional I_a of $f : \mathcal{X}^n \to \mathbb{R}$ is defined by

$$\begin{aligned} & I_a(f) \\ =& 2\left(\sup_{x\in\mathcal{X}^n}\sum_l \sup_{z\in\mathcal{X}}\sum_{k:k\neq l}\|f_k(X)-S_z^l f_k(X)\|_a^2(x)\right)^{\frac{1}{2}}. \end{aligned}$$

From the above discussion on the summation function, one can verify that the functional $I_a(f)$ vanishes for sum. The formalism in Definition 2.3 is similar to Definition 1 in (Maurer, 2019) who defines

$$I(f) = 2\left(\sup_{x\in\mathcal{X}^n}\sum_l \sup_{z\in\mathcal{X}}\sum_{k:k\neq l}\sigma_k^2(f-S_z^l f)(x)\right)^{\frac{1}{2}},$$

where $\sigma_k^2(f) = \mathbb{E}_k[(f-\mathbb{E}_k f)^2]$ and where $\mathbb{E}_k = \mathbb{E}_{y\sim\mu}[S_y^k f]$. The two definitions are similar but different in that Maurer (2019) focuses on the conditional variance, while this paper studies the norms.

3. Main Results

This section presents our main results, organized by sub-Gaussian, sub-exponential and heavy-tailed distributions. Successively, the latter has a heavier tail than the former.

3.1. Sub-Gaussian Distributions

We define a sub-Gaussian random variable, which is a popular sub-class of unbounded random variables. This class of distributions subsumes the Gaussian random variables, as well as all the bounded ones (such as Bernoulli, uniform, and multinomial).

Definition 3.1. (Vershynin, 2018) A random variable Z is sub-Gaussian if there exists $\sigma > 0$ such that

$$\mathbb{P}(|Z|>t) \le 2\exp\left(-\frac{t^2}{2\sigma^2}\right),$$

for every $t > 0$, where the quantity σ^2 is named as the sub-Gaussian variance proxy.

We consider norms to present our concentration inequalities. For this purpose, we use the following equivalent sub-Gaussian norm (Buldygin & Kozachenko, 2000; Lei & Zhang, 2020)

$$\|Z\|_{\psi_2} = \sup_{p\ge 1}\left[\frac{\mathbb{E}Z^{2p}}{(2p-1)!!}\right]^{\frac{1}{2p}}.$$

which is equal to $\|Z\|_{\psi_2} \doteq \sup_{p\ge1}\left[\frac{2^p P!}{(2p)!}\mathbb{E}Z^{2p}\right]^{\frac{1}{2p}}$ since $(2p-1)!! = \frac{(2p)!}{2^p p!}$ for any $p \ge 1$, please refer to page 6 and Theorem 1.3 in (Buldygin & Kozachenko, 2000) for details of this sub-Gaussian norm.

We show the first concentration inequality, which assumes that $f_k(X)$ follows a sub-Gaussian distribution.

Theorem 3.2. *Let $f : \mathcal{X}^n \to \mathbb{R}$ and $X = (X_1, ..., X_n)$ be a vector of independent random variables with values in the space $\mathcal{X}$. Then for any $t > 0$, we have*

$$\begin{aligned} & \mathbb{P}\left(f(X)-\mathbb{E}[f(X')]>t\right) \\ \le & \exp\left(\frac{-t^2}{16\mathbb{E}\left[\sum_{k=1}^n\|f_k(X)\|_{\psi_2}^2(X)\right]+2\sqrt{2}I_{\psi_2}(f)t}\right). \end{aligned}$$

We give some remarks.

(1): If f is a sum of independent random variables, we recover the general Hoeffding's inequality, i.e.,

$$\mathbb{P}\left(\sum_k X_k - \mathbb{E}\Big[\sum_k X_k'\Big] > t\right) \le \exp\left(\frac{-t^2}{16\sum_k\|X_k\|_{\psi_2}^2}\right),$$

refer to Theorem 2.6.2 in (Vershynin, 2018). The constant 16 present in this bound is not optimal. Theorem 1.5 in (Buldygin & Kozachenko, 2000) gives a better bound with a constant 2, but this result is provided for the sum of independent sub-Gaussian random variables.

(2): Theorem 3 in Maurer & Pontil (2021) shows

$$\begin{aligned} & \mathbb{P}\left(f(X)-\mathbb{E}[f(X')]>t\right) \\ \le & \exp\left(\frac{-t^2}{32e\|\sum_{k=1}^n\|f_k(X)\|_{\psi_2'}^2\|_\infty}\right), \end{aligned} \quad (1)$$

where the sub-Gaussian norm in (Maurer & Pontil, 2021) is defined as $\|Z\|_{\psi_2'} \doteq \sup_{p\ge1}\frac{\|Z\|_p}{\sqrt{p}}$, and where $\|f_k(X)\|_{\psi_2'}^2\|_\infty$ is the essential supremum of $\|f_k(X)\|_{\psi_2'}^2(X)$. Clearly, $\mathbb{E}[\sum_{k=1}^n\|f_k(X)\|_{\psi_2}^2(X)] \le \|\sum_{k=1}^n\|f_k(X)\|_{\psi_2'}^2\|_\infty$. Thus, the variance proxy term in (1) can never be small than what we get in Theorem 3.2. Indeed, the supremum can be very large in some extreme cases, and we successfully replace it with an expectation, giving a distribution-dependent inequality.

(3): The bound has two tails. The sub-Gaussian tail is of course expected from the sub-Gaussian distribution. The sub-exponential tail is produced by the concentration property of weakly self-bounded functions, refer to (9) in Lemma A.6, where a^2 corresponds to the interaction functional $I_{\psi_2}(f)$ in Theorem 3.2. In the applications to U-statistics and V-statistics respectively, we show that the interaction functional $I_a(f)$ is only of order m^2/n and $m(m-1)/n$, which will approach to 0 as $n \to \infty$, where m is the degree of the two statistics.

3.2. Sub-Exponential Distributions

The second class of random variables that we treat in this paper are sub-exponential. Although the class of sub-Gaussian distributions is natural and quite large, it leaves out some important distributions whose tails are heavier than Gaussian. For instance, all sub-Gaussian random variables are sub-exponential. Products and squares of sub-Gaussian variables are sub-exponential, no more sub-Gaussian. Apart from that, sub-exponential distributions include the exponential, chi-squared and Poisson distributions.

Definition 3.3. (Vershynin, 2018) A random variable Z is sub-exponential if there exists $c > 0$ such that

$$\mathbb{P}(|Z| > t) \leq 2\exp(-ct),$$

for every $t > 0$.

Here, we use the following equivalent sub-exponential norm (Buldygin & Kozachenko, 2000; Yang et al., 2022)

$$\|Z\|_{\psi_1} = \sup_{p\geq 1}\left[\frac{\mathbb{E}|Z|^p}{p!}\right]^{\frac{1}{p}},$$

please refer to page 23 and Remark 3.2 in (Buldygin & Kozachenko, 2000) for details of this sub-exponential norm.

We now show the concentration inequality assuming that $f_k(X)$ follows a sub-exponential distribution.

Theorem 3.4. *Let $f : \mathcal{X}^n \to \mathbb{R}$ and $X = (X_1, ..., X_n)$ be a vector of independent random variables with values in the space $\mathcal{X}$. Then for any $t > 0$, we have*

$$\begin{aligned}&\mathbb{P}\left(f(X) - \mathbb{E}[f(X')] > t\right)\\ \leq&\exp\left(\frac{-t^2}{4\mathbb{E}\left[\sum_{k=1}^n \|f_k(X)\|^2_{\psi_1}(X)\right] + 2\left(M + \frac{I_{\psi_1}(f)}{\sqrt{2}}\right)t}\right),\end{aligned}$$

where $M = \max_k \|\|\|f_k(X)\|_{\psi_1}\|_\infty$ and $\|\|\|f_k(X)\|_{\psi_1}\|_\infty$ is the essential supremum of $\|f_k(X)\|_{\psi_1}(X)$.

We give some remarks.

(1): If f is a sum, we recover Bernstein's inequality for sub-exponential random variables, refer to Theorem 2.8.1 in (Vershynin, 2018).

(2): Theorem 4 in Maurer & Pontil (2021) shows

$$\begin{aligned}&\mathbb{P}\left(f(X) - \mathbb{E}[f(X')] > t\right)\\ \leq&\exp\left(\frac{-t^2}{4e^2\|\sum_{k=1}^n \|f_k(X)\|^2_{\psi_1'}\|_\infty + 2eMt}\right), \qquad (2)\end{aligned}$$

where $M = \max_k \|\|\|f_k(X)\|_{\psi_1'}\|_\infty$ and $\|\|\|f_k(X)\|_{\psi_1'}\|_\infty$ is the essential supremum of $\|f_k(X)\|_{\psi_1'}(X)$, and where the sub-exponential norm in (Maurer & Pontil, 2021) is defined as $\|Z\|_{\psi_1'} \doteq \sup_{p\geq 1}\frac{\|Z\|_p}{p}$. As a comparison, we improve the supremum variance proxy $\|\sum_{k=1}^n \|f_k(X)\|^2_{\psi_1'}\|_\infty$ in (2) to an expectation $\mathbb{E}[\sum_{k=1}^n \|f_k(X)\|^2_{\psi_1}(X)]$, giving a distribution-dependent McDiarmid-type inequality.

(3): The bound exhibits a mixture of two tails, a sub-Gaussian tail governed by the variance-proxy $\mathbb{E}[\sum_{k=1}^n \|f_k(X)\|^2_{\psi_1}(X)]$ for small deviations, and a sub-exponential tail governed by the scale-proxy $\max_k \|\|\|f_k(X)\|_{\psi_1}\|_\infty + I_{\psi_1}(f)$ for large deviations.

(4): We now discuss the improvement of the constants in our bound from two perspectives, one is giving proofs and the other is providing some examples.

(i) We first consider the sub-Gaussian case. In Theorem 1.3 of (Buldygin & Kozachenko, 2000), it states that: suppose that Z is a zero-mean random variable, in order for Z to be sub-Gaussian, it is necessary and sufficient that $\|Z\|_{\psi_2} < \infty$ or $\|Z\|'_{\psi_2} < \infty$. And in this case, the following inequality hold

$$\|Z\|_{\psi_2} \leq e^{9/16}\|Z\|_{\psi_2'}. \qquad (3)$$

Substituting the above inequality (3) into our Theorem 3.2, we get

$$\begin{aligned}&\mathbb{P}\left(f(X) - \mathbb{E}[f(X')] > t\right) \leq\\ &\exp\left(\frac{-t^2}{16e^{9/8}\mathbb{E}\left[\sum_{k=1}^n \|f_k(X)\|^2_{\psi_2'}(X)\right] + 2\sqrt{2}I_{\psi_2}(f)t}\right). \qquad (4)\end{aligned}$$

Now the definition of the sub-Gaussian norm in (4) is equal to the one in (Maurer & Pontil, 2021). Comparing (4) with (1), since $16e^{9/8} < 32e$, our bound gives an improvement in the constant.

We then consider the sub-exponential case. In Theorem 3.2 and Remark 3.2 of (Buldygin & Kozachenko, 2000), it states that: suppose that Z is a zero-mean random variable, in order for Z to be sub-exponential, it is necessary and sufficient that $\|Z\|_{\psi_1} < \infty$. And in this case, a similar inequality to (3) is

$$\|Z\|_{\psi_1} \leq \frac{e}{\sqrt{2\pi}}\|Z\|_{\psi_1'}. \qquad (5)$$

We now give the proof. By the Stirling formula

$$n! = \sqrt{2\pi n}n^n e^{-n+\theta_n}, \quad |\theta_n| < \frac{1}{12n}, n > 1,$$

we obtain the inequality

$$\begin{aligned}&\left[\frac{1}{p!}\right]^{\frac{1}{p}} = \left[\frac{e^{p-\theta_p}}{\sqrt{2\pi p}p^p}\right]^{\frac{1}{p}}\\ \leq&\frac{e}{(2\pi p)^{1/2p}p} = \frac{1}{\pi^{1/2p}}\frac{1}{(2p)^{1/2p}}\frac{e}{p} \leq \frac{e}{\sqrt{2\pi}p}\end{aligned}$$

for all positive integers p, giving $\|Z\|_{\psi_1} \leq \frac{e}{\sqrt{2\pi}} \|Z\|_{\psi_1'}$. Substituting the above inequality (5) into our Theorem 3.4, we get

$$\mathbb{P}\left(f(X) - \mathbb{E}[f(X')] > t\right) \leq$$
$$\exp\left(\frac{-t^2}{\frac{4e^2}{2\pi}\mathbb{E}\left[\sum_{k=1}^n \|f_k(X)\|_{\psi_1'}^2(X)\right] + 2\left(M + \frac{I_{\psi_1}(f)}{\sqrt{2}}\right)t}\right), \tag{6}$$

where $M = \frac{e}{\sqrt{2\pi}} \max_k \|\|\|f_k(X)\|_{\psi_1'}\|_\infty$. Now the definition of the sub-exponential norm in (6) is equal to the one in (Maurer & Pontil, 2021). Comparing (6) with (2), since $\frac{4e^2}{2\pi} < 4e^2$ and $2\frac{e}{\sqrt{2\pi}} \leq 2e$, our bound gives an improvement in the constants. These proofs confirm our improvements in the constants compared to (Maurer & Pontil, 2021).

(ii) We now list some examples to support the improvement in the constants. For example, the constant in Theorem 3 in (Maurer & Pontil, 2021) for Gaussian distribution (i.e. r.v. $X \sim N(\mu, \sigma^2)$ with mean μ and variance σ^2) is

$$\begin{aligned} &32en\left(\sup_{p\geq 1} \sigma\sqrt{2}\frac{[\Gamma(\frac{(1+p)}{2})]^{1/p}}{\sqrt{p}\sqrt{\pi}^{1/p}}\right)^2 \\ =&64en\sigma^2 \sup_{p\geq 1} \frac{[\Gamma(\frac{(1+p)}{2})]^{2/p}}{p\pi^{1/p}} = \frac{64en\sigma^2}{\pi}. \end{aligned}$$

Note that the moments of the normal distribution $X \sim N(\mu, \sigma^2)$ are $\mathbb{E}[|X-\mu|^p] = \sigma^p 2^{p/2}\frac{\Gamma(\frac{(1+p)}{2})}{\sqrt{\pi}}$ and $\mathbb{E}[|X-\mu|^{2p}] = \sigma^{2p}\frac{(2p)!}{2^p p!}$. As a comparison, in Gaussian distribution, our Theorem 3.2 improves the constants as

$$16n\left(\sup_{p\geq 1}\left[\frac{2^p p!}{(2p)!}\sigma^{2p}\frac{(2p)!}{2^p p!}\right]^{1/(2p)}\right)^2 = 16n\sigma^2.$$

The constant in Theorem 4 in (Maurer & Pontil, 2021) for Laplace distribution (i.e. r.v. $X \sim Laplace(\mu, \lambda)$ with mean μ and $|X-\mu| \sim exponential(\lambda^{-1})$ are

$$\begin{aligned} &4e^2 n\left(\sup_{p\geq 1}\frac{\lambda(p!)^{1/p}}{p}\right)^2 + 2e\left(\sup_{p\geq 1}\frac{\lambda(p!)^{1/p}}{p}\right) \\ =&4e^2 n\lambda^2 \sup_{p\geq 1}\frac{(p!)^{2/p}}{p^2} + 2e\lambda \sup_{p\geq 1}\frac{(p!)^{1/p}}{p} \\ =&4e^2 n\lambda^2 + 2e\lambda. \end{aligned}$$

Note that the moments of the exponential distribution $|X-\mu| \sim exponential(\lambda^{-1})$ is $\mathbb{E}[|X-\mu|^p] = p!\lambda^p$. As a comparison, in Laplace distribution, our Theorem 3.4 improves the constants as

$$4n\left(\sup_{p\geq 1}\frac{\lambda(p!)^{1/p}}{(p!)^{1/p}}\right)^2 + 2\left(\sup_{p\geq 1}\frac{\lambda(p!)^{1/p}}{(p!)^{1/p}}\right) = 4n\lambda^2 + 2\lambda.$$

These examples confirm our improvements in the constants compared to (Maurer & Pontil, 2021).

3.3. Heavy-tailed Distributions

The previous two sections give concentration inequalities of magnitude $\mathbb{P}\left(f(X) - \mathbb{E}[f(X')] > t\right) = O\left(\exp(-ct^\alpha)\right)$, $t \to \infty$, where $\alpha \in \{1, 2\}$ and $c > 0$ is a constant. It is natural to ask under what condition we have the following exponential decay rate

$$\mathbb{P}\left(f(X) - \mathbb{E}[f(X')] > t\right) = O\left(\exp(-ct^\alpha)\right) \quad t \to \infty, \tag{7}$$

where $\alpha \in (0, 1)$ is given and $c > 0$ is a constant. We should mention that distributions satisfying (7) fall under the broad class of heavy-tailed distributions (see (Vladimirova et al., 2020; Kuchibhotla & Chakrabortty, 2022)), such as Weibull distributions. The standard technique, i.e. finding upper bounds for the moment generating function, fails for the heavy-tailed distributions whose moment generating functions do not exist. In this paper, we give a sufficient condition in order that (7) holds. Specifically, for any real random variable Z, we give the following moment condition

$$\mathbb{E}\left[\exp\left((Z^+)^{\frac{2\alpha}{1-\alpha}}\right)\right] \leq c, \tag{8}$$

where $Z^+ = \max\{Z, 0\}$. Clearly, $\mathbb{E}[\exp((Z^+)^{\frac{2\alpha}{1-\alpha}})] \leq \mathbb{E}[\exp(|Z|^{\frac{2\alpha}{1-\alpha}})]$. The result in this section also holds for this condition $\mathbb{E}[\exp(|Z|^{\frac{2\alpha}{1-\alpha}})] \leq c$.

We now show the concentration inequality assuming that $f_k(X)$ satisfies the moment condition (8).

Theorem 3.5. *Let $f : \mathcal{X}^n \to \mathbb{R}$ and $X = (X_1, ..., X_n)$ be a vector of independent random variables with values in the space $\mathcal{X}$. Then for any $t > 0$, we have*

$$\mathbb{P}\left(f(X) - \mathbb{E}[f(X')] > t\right) \leq nc\exp(-t^\alpha) +$$
$$\exp\left(\frac{-t^2}{2\mathbb{E}\left[\sum_{k=1}^n \|f_k(X)\|_2^2(X)\right] + 2(2t^{1-\alpha}/3 + I_2(f)/2)t}\right).$$

We give some remarks.

(1): If f is a sum of independent random variables, we get

$$\begin{aligned} &\mathbb{P}\left(\sum_k X_k - \mathbb{E}\Big[\sum_k X_k'\Big] > t\right) \leq nc\exp(-t^\alpha) \\ &+ \exp\left(\frac{-t^2}{2\sum_{k=1}^n \|X_k\|_2^2 + 4t^{2-\alpha}/3}\right), \end{aligned}$$

which exhibits a mixture of two tails. One is the sub-Gaussian tail for small deviations, which is of course expected from the central limit theorem, and the other has the tail of magnitude $O(\exp(-ct^\alpha))$ for large deviations.

(2): For concentration inequalities of general functions, we haven't seen related results of heavy-tailed distributions.

(3): Theorem 3.5 gives a distribution-dependent McDiarmid-type inequality and exhibits a tail of magnitude $O(\exp(-ct^{\alpha}))$ for large deviations.

(4): Similar motivation and techniques to this paper have been used in (Maurer, 2019) exploring Bernstein-type inequalities with bounded functions. We first consider the sub-Gaussian case to compare the technical difference. According to the entropy method, the proof techniques can be divided into four steps. Central to the entropy method is the entropy and the sub-additivity of entropy. The first step is to give a bound of the entropy $S(\gamma f(X))$ in terms of the sum of the conditional entropies

$$S(\gamma f(X)) \leq 4\gamma^2 \mathbb{E}_{\gamma f(X)} \left[\sum_{k=1}^{n} \|f_k(X)\|_{\psi_2}^2 (X) \right].$$

In bounding the entropy, the techniques of unbounded random variables are different from the bounded one, refer to Lemma 2 in (Maurer, 2019) for the bounded case and our Lemma A.8 for the sub-Gaussian case. After the entropy bound, the second step is to use a decoupling technique to decouple the expectation functional $\mathbb{E}_{\gamma f(X)} \left[\sum_{k=1}^{n} \|f_k(X)\|_{\psi_2}^2 (X) \right]$. The aim is to give a bound on the entropy $S(\gamma f(X))$ in terms of the $\ln \mathbb{E} \left[e^{\theta \sum_{k=1}^{n} \|f_k(X)\|_{\psi_2}^2 (X)} \right]$. In this proof, the techniques of unbounded random variables are also different from the bounded one. The third step is to prove the weakly self-boundedness of $\sum_{k=1}^{n} \|f_k(X)\|_a^2 (X)$, where $a = \{\psi_2, \psi_1, 2\}$. Since we deal with norms, the proof here is different from (Maurer, 2019) that deal with real-valued functions, refer to the proof of Lemma A.7 and the proof of Proposition 2 in (Maurer, 2019). After the three steps, we obtain the entropy bound. The forth step is use standard log-Laplace transform argument (Theorem 7 in (Maurer, 2012)) to give the concentration inequality. The above analysis of sub-Gaussian variables hold for the sub-exponential case, while for the heavy-tailed variables, an clipping argument on random variables is additionally introduced.

4. Applications

This section applies our concentration inequalities to the sample mean, U-statistics and V-statistics.

4.1. Sample Mean

The sample mean is $f(x) = \frac{1}{n} \sum_{i=1}^{n} x_i$. In this case, $f_k(X)(x) = \frac{1}{n}(X_k - \mathbb{E}[X_k])$ is independent of x, and thus $I_a(f) = 0$. The sample mean is the most natural choice of a mean estimator and, particularly, is closely related to empirical risk in learning theory (Mohri et al., 2018). We now give its results, which would be useful in the generalization error analysis in learning theory.

Theorem 4.1. *Let $f(x)$ be as defined above and $X = (X_1, ..., X_n)$ be a vector of independent random variables with values in the space $\mathcal{X}$. Then for any $t > 0$, we have*

(1) for the sub-Gaussian case

$$\begin{aligned} &\mathbb{P}\left(f(X) - \mathbb{E}[f(X')] > t\right) \\ &\leq \exp\left(\frac{-t^2}{\frac{64}{n^2} \sum_{k=1}^{n} \|X_k\|_{\psi_2}^2} \right). \end{aligned}$$

(2) for the sub-exponential case

$$\begin{aligned} &\mathbb{P}\left(f(X) - \mathbb{E}[f(X')] > t\right) \\ &\leq \exp\left(\frac{-t^2}{\frac{16}{n^2} \sum_{k=1}^{n} \|X_k\|_{\psi_1}^2 + \frac{4}{n} \max_k \|X_k\|_{\psi_1} t} \right). \end{aligned}$$

(3) for the heavy-tailed case

$$\begin{aligned} &\mathbb{P}\left(f(X) - \mathbb{E}[f(X')] > t\right) \\ &\leq nc \exp(-t^{\alpha}) + \exp\left(\frac{-t^2}{\frac{8}{n^2} \sum_{k=1}^{n} \|X_k\|_2^2 + 4t^{2-\alpha}/3} \right). \end{aligned}$$

4.2. U-statistics

Let f be a measurable, symmetric kernel $f : \mathcal{X}^m \to \mathbb{R}$ with $1 < m < n$. Then, we define U-statistics as

$$U(x) = \binom{n}{m}^{-1} \sum_{1 \leq j_1 < \cdots < j_m \leq n} f(x_{j_1}, \cdots, x_{j_m}).$$

We introduce some notations. If B is a set and $m \in \mathbb{N}$, then let S_B^m denote the set of all those subsets of B which have cardinality m. Also, if $S \subseteq \{1, ..., n\}$ and $x \in \mathcal{X}^n$, we use x_S to denote the vector $(x_{j_1}, ..., x_{j_{|S|}}) \in \mathcal{X}^{|S|}$, where $\{j_1, ..., j_{|S|}\} = S$ and the j_k are increasingly ordered. For $y, z \in \mathcal{X}$ we use (y, x_S) and (y, z, x_S) to denote the vectors $(y, x_{j_1}, ..., x_{j_{|S|}}) \in \mathcal{X}^{|S|+1}$ and $(y, z, x_{j_1} ..., x_{j_{|S|}}) \in \mathcal{X}^{|S|+2}$ respectively. With these notations, we have

$$U(x) = \binom{n}{m}^{-1} \sum_{S \in \mathcal{S}_{\{1,\ldots,n\}}^m} f(x_S).$$

Let $f_k(X)(x_S) := f(x_S, X_k) - \mathbb{E}[f(x_S, X_k)]$, where $S \in \mathcal{S}_{\{1,\ldots,n\} \setminus k}^{m-1}$. We assume $\max_S \max_k \|f_k(X)\|_a(x_S) \leq b$, which is weaker than the boundedness of f, and $\mathbb{E}\|f_k(X)\|_a^2(X_S) \leq \sigma^2$ for any $k = \{1, ..., n\}$. Note that the two assumptions are intended to present the following results in a concise form. The results allow for weaker and fine-grained assumptions.

Theorem 4.2. *Let $U(x)$ be as defined above and $X = (X_1, ..., X_n)$ be a vector of independent random variables with values in the space $\mathcal{X}$. Then for any $t > 0$, we have*

(1) for the sub-Gaussian case

$$\mathbb{P}\left(U(X) - \mathbb{E}[U(X')] > t\right) \le \exp\left(\frac{-t^2}{\frac{16m^2\sigma^2}{n} + \frac{8\sqrt{2}m^2 b}{n}t}\right).$$

(2) for the sub-exponential case

$$\mathbb{P}\left(U(X) - \mathbb{E}[U(X')] > t\right) \le \exp\left(\frac{-t^2}{\frac{4m^2}{n}\sigma^2 + 2(\frac{m}{n}b + \frac{4m^2}{n\sqrt{2}}b)t}\right).$$

(3) for the heavy-tailed case

$$\mathbb{P}\left(U(X) - \mathbb{E}[U(X')] > t\right) \le nc\exp(-t^\alpha) + \exp\left(\frac{-t^2}{\frac{2m^2}{n}\sigma^2 + 2(2t^{1-\alpha}/3 + \frac{2m^2}{n}b)t}\right).$$

U-statistics are closely relevant to ranking (clémençon et al., 2008) and pairwise learning (U-statistic of order 2), where the latter instantiates many well-known machine learning tasks, for instance, similarity and metric learning (Cao et al., 2016; Maurer et al., 2021), AUC maximization (Cortes & Mohri, 2003; Ying et al., 2016), bipartite ranking (clémençon et al., 2005), gradient learning (Mukherjee & Wu, 2006), minimum error entropy principle (Hu et al., 2013), multiple kernel learning (Kumar et al., 2012), and preference learning (Fürnkranz & Hüllermeier, 2010).

4.3. V-statistics

V-statistics is introduced in (Mises, 1947). Let $f : \mathcal{X}^m \to \mathbb{R}$ with $1 < m < n$. The V-statistic is defined by

$$V(x) = \frac{1}{n^m} \sum_{\{j_1,\dots,j_m\}\in\{1,\dots,n\}^m} f(x_{j_1}, \cdots, x_{j_m}).$$

$V(x)$ receives contributions from multi-indices with multiple occurrences of individual indices, different from $U(x)$ that avoids multi-indices with multiple occurrences of indices. We then introduce some notations. If multi-index $S = \{j_1, ..., j_{|S|}\} \in \{1, ..., n\}^{|S|}$ and $x \in \mathcal{X}^n$, we use x_S to denote the vector $(x_{j_1}, ..., x_{j_{|S|}}) \in \mathcal{X}^{|S|}$. For $y, z \in \mathcal{X}$ we also use (y, x_S) and (y, z, x_S) to denote respectively, the vectors $(y, x_{j_1}, ..., x_{j_{|S|}}) \in \mathcal{X}^{|S|+1}$ and $(y, z, x_{j_1} ..., x_{j_{|S|}}) \in \mathcal{X}^{|S|+2}$. With these notations, we have

$$V(x) = \binom{n}{m}^{-1} \sum_{S\in\{1,\dots,n\}^m} f(x_S).$$

Let $f_k(X)(x_S) := f(x_S, X_k) - \mathbb{E}[f(x_S, X_k)]$, where $S \in \{1, ..., n\}^{m-1}$. We assume $\max_S \max_k \|f_k(X)\|_a(x_S) \le b$, which is weaker than the boundedness of f, and $\mathbb{E}\|f_k(X)\|_a^2(X_S) = \sigma^2$ for any $k = \{1, ..., n\}$. We give results of the V-statistics, which also allows for weaker and fine-grained assumptions.

Theorem 4.3. *Let $V(x)$ be as defined above and $X = (X_1, ..., X_n)$ be a vector of independent random variables with values in the space $\mathcal{X}$. Then for any $t > 0$, we have*

(1) for the sub-Gaussian case

$$\mathbb{P}\left(V(X) - \mathbb{E}[V(X')] > t\right) \le \exp\left(\frac{-t^2}{\frac{16m^2\sigma^2}{n} + \frac{8\sqrt{2}m(m-1)b}{n}t}\right).$$

(2) for the sub-exponential case

$$\mathbb{P}\left(V(X) - \mathbb{E}[V(X')] > t\right) \le \exp\left(\frac{-t^2}{\frac{4m^2}{n}\sigma^2 + 2(\frac{m}{n}b + \frac{4m(m-1)}{n\sqrt{2}}b)t}\right).$$

(3) for the heavy-tailed case

$$\mathbb{P}\left(V(X) - \mathbb{E}[V(X')] > t\right) \le nc\exp(-t^\alpha) + \exp\left(\frac{-t^2}{\frac{2m^2}{n}\sigma^2 + 2(2t^{1-\alpha}/3 + \frac{2m(m-1)}{n}b)t}\right).$$

V-statistics reveal some interesting machine learning applications, such as auto-encoding variational Bayes (Lopez et al., 2018), off-policy evaluation in reinforcement learning (Feng et al., 2020), and kernel learning (Shen et al., 2020).

Next, we take the sub-exponential case as an example to discuss how these bounds can be compared to the inequalities of (Maurer & Pontil, 2021).

(1) We first consider the application to the sample mean. In this case, $f_k(X)(x) = \frac{1}{n}(X_k - \mathbb{E}[X_k])$ is independent of x. Following the proof in Appendix B.1, we can deduce $\|\|f_k(X)\|_{\psi_1'}^2\|_\infty \le \frac{4}{n^2}\|X_k\|_{\psi_1'}^2$ and $\max_k \|\|f_k(X)\|_{\psi_1'}\|_\infty \le \frac{2}{n}\max_k \|X_k\|_{\psi_1'}$. Plugging these bounds into Theorem 4 of (Maurer & Pontil, 2021) gives

$$\mathbb{P}\left(f(X) - \mathbb{E}[f(X')] > t\right) \le \exp\left(\frac{-t^2}{\frac{16e^2}{n^2}\sum_{k=1}^n \|X_k\|_{\psi_1'}^2 + \frac{4e}{n}\max_k \|X_k\|_{\psi_1'} t}\right).$$

By comparison, our bound in Theorem 4.1 has an improvement in the constants.

(2) We then consider the application to U-statistics. Since the variance proxy term in (Maurer & Pontil, 2021) depends on the worst-case choice of the configuration of

sample space, the upper bound of their variance proxy term would be very large. Following the proof in Appendix B.2, we can deduce $\max_k \||\|U_k(X)\|_{\psi_1'}\|_\infty \leq \frac{m}{n}b$ and $\|\sum_{k=1}^n \|U_k(X)\|^2_{\psi_1'}\|_\infty \leq \frac{m^2}{n}b^2$, where b is from the assumption $\max_S \max_k \|f_k(X)\|_{\psi_1'}(x_S) \leq b$. Plugging these bounds into Theorem 4 of (Maurer & Pontil, 2021) gives

$$\mathbb{P}\left(U(X) - \mathbb{E}[U(X')] > t\right) \leq \left(\frac{-t^2}{\frac{4e^2m^2}{n}b^2 + 2e\frac{m}{n}bt}\right).$$

In the regime of sub-Gaussian tail, our variance proxy term $\frac{4m^2}{n}\sigma^2$ in Theorem 4.2 is sharper than the result of (Maurer & Pontil, 2021) since b^2 depends on the worst case and may be very larger than σ^2. In the regime of sub-exponential tail, our scale-proxy term $2(\frac{m}{n}b + \frac{4m^2}{n\sqrt{2}}b)$ would have similar performance to the bound $2e\frac{m}{n}b$ of (Maurer & Pontil, 2021), especially for the widely studied pairwise learning in machine learning (i.e., $m = 2$).

(3) We further consider the application to V-statistics. Following the proof in Appendix B.3, we can deduce that $\max_k \||\|V_k(X)\|_{\psi_1'}\|_\infty \leq \frac{m}{n}b$ and $\|\sum_{k=1}^n \|V_k(X)\|^2_{\psi_1'}\|_\infty \leq \frac{m^2}{n}b^2$, where b is from the assumption $\max_S \max_k \|f_k(X)\|_{\psi_1'}(x_S) \leq b$. Plugging these bounds into Theorem 4 of (Maurer & Pontil, 2021) gives

$$\mathbb{P}\left(V(X) - \mathbb{E}[V(X')] > t\right) \leq \left(\frac{-t^2}{\frac{4e^2m^2}{n}b^2 + 2e\frac{m}{n}bt}\right).$$

Since b^2 depends on the worst-case choice of the configuration of sample space and may be very larger than σ^2, in the regime of sub-Gaussian tail, our variance proxy term $\frac{4m^2}{n}\sigma^2$ is sharper than the result of (Maurer & Pontil, 2021). In the regime of sub-exponential tail, our scale-proxy term $2(\frac{m}{n}b + \frac{4m(m-1)}{n\sqrt{2}}b)$ would also have similar performance to the bound $2e\frac{m}{n}b$ of (Maurer & Pontil, 2021), especially for the widely studied pairwise learning (i.e., $m = 2$).

5. Refined Results

This section gives refined concentration inequalities by considering a weaker interaction functional, inspired from (Maurer, 2017). Let μ be a probability measure defined on the sample space $\mathcal{X}$. The definition of this interaction functional is introduced below.

Definition 5.1. The interaction functional I'_a of $f : \mathcal{X}^n \to \mathbb{R}$ is defined by

$$I'_a(f) = 2\left(\sup_{x\in\mathcal{X}^n}\sum_l \mathbb{E}_{Z\sim\mu}\left[\sum_{k:k\neq l}\|f_k(X) - S^l_Z f_k(X)\|^2_a(x)\mathbb{I}_{A_l}(Z)\right]\right)^{\frac{1}{2}},$$

where $\mathbb{E}_{Z\sim\mu}$ denotes the expectation w.r.t. the random variable Z drawn from μ, $\mathbb{I}_{A_l}(Z)$ is the indicator function, and A_l is the subset of $\mathcal{X}$ defined by

$$A_l = \Big\{Z \in \mathcal{X} : S^l_Z\sum_{k=1}^n \|f_k(X)\|^2_a(X) \leq \sum_{k=1}^n \|f_k(X)\|^2_a(X)\Big\}.$$

Clearly, we have $I'_a(f) \leq I_a(f)$ for any f. By this weaker interaction functional, we get improved concentration inequalities with $I_a(f)$ replaced by $I'_a(f)$.

Theorem 5.2. *Let $f : \mathcal{X}^n \to \mathbb{R}$ and $X = (X_1, ..., X_n)$ be a vector of independent random variables with values in the space $\mathcal{X}$. Then for any $t > 0$, we have*

(1) for the sub-Gaussian case

$$\mathbb{P}\left(f(X) - \mathbb{E}[f(X')] > t\right) \leq \exp\left(\frac{-t^2}{16\mathbb{E}\left[\sum_{k=1}^n \|f_k(X)\|^2_{\psi_2}(X)\right] + 2\sqrt{2}I'_{\psi_2}(f)t}\right).$$

(2) for the sub-exponential case

$$\mathbb{P}\left(f(X) - \mathbb{E}[f(X')] > t\right) \leq \exp\left(\frac{-t^2}{4\mathbb{E}\left[\sum_{k=1}^n \|f_k(X)\|^2_{\psi_1}(X)\right] + 2\left(M + \frac{I'_{\psi_1}(f)}{\sqrt{2}}\right)t}\right).$$

(3) for the heavy-tailed case

$$\mathbb{P}\left(f(X) - \mathbb{E}[f(X')] > t\right) \leq nc\exp(-t^\alpha) + \exp\left(\frac{-t^2}{2\mathbb{E}\left[\sum_{k=1}^n \|f_k(X)\|^2_2(X)\right] + 2(2t^{1-\alpha}/3 + I'_2(f)/2)t}\right).$$

6. Conclusion

This paper gave distribution-dependent McDiarmid-type concentration inequalities for sub-Gaussian, sub-exponential and heavy-tailed distributions. The results sharpen existing inequalities by replacing the supremum with an expectation. This paper then illustrated these inequalities with applications to the sample mean, U-statistics and V-statistics.

Acknowledgements

We thank the anonymous reviewers for their valuable and constructive suggestions and comments. This work is supported by the Beijing Natural Science Foundation (No.4222029); the National Natural Science Foundation of China (NO.62076234); the National Key Research and Development Project (No.2022YFB2703102); the "Intelligent Social Governance Interdisciplinary Platform, Major Innovation & Planning Interdisciplinary Platform for the "Double-First Class" Initiative, Renmin University of China"; the Beijing Outstanding Young Scientist Program (NO.BJJWZYJH012019100020098); the Public Computing Cloud, Renmin University of China; the Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China (NO.2021030199); the Huawei-Renmin University joint program on Information Retrieval; the Unicom Innovation Ecological Cooperation Plan; and the CCF-Huawei Populus Grove Fund.

References

Bakhshizadeh, M., Maleki, A., and de la Pena, V. H. Sharp concentration results for heavy-tailed distributions. *arXiv preprint arXiv:2003.13819*, 2020a.

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

Barsbey, M., Sefidgaran, M., Erdogdu, M. A., Richard, G., and Simsekli, U. Heavy tails in sgd and compressibility of overparametrized neural networks. *Advances in Neural Information Processing Systems*, 34:29364–29378, 2021.

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

Ben-David, S., Blitzer, J., Crammer, K., and Pereira, F. Analysis of representations for domain adaptation. In *Advances in neural information processing systems*, 2006.

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

Bousquet, O. and Elisseeff, A. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

Bousquet, O., Klochkov, Y., and Zhivotovskiy, N. Sharper bounds for uniformly stable algorithms. In *Conference on Learning Theory*, pp. 610–626, 2020.

Buldygin, V. V. and Kozachenko, I. V. *Metric characterization of random variables and random processes*, volume 188. American Mathematical Soc., 2000.

Cao, Q., Guo, Z.-C., and Ying, Y. Generalization bounds for metric and similarity learning. *Machine Learning*, 102 (1):115–132, 2016.

Chazottes, J.-R., Collet, P., Külske, C., and Redig, F. Concentration inequalities for random fields via coupling. *Probability Theory and Related Fields*, 137(1):201–225, 2007.

clémençon, S., Lugosi, G., and Vayatis, N. Ranking and scoring using empirical risk minimization. In *Conference on Learning Theory*, pp. 1–15, 2005.

clémençon, S., Lugosi, G., and Vayatis, N. Ranking and empirical minimization of u-statistics. *Annals of Statistics*, 36(2):844–874, 2008.

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

Cortes, C. and Mohri, M. Auc optimization vs. error rate minimization. In *Advances in Neural Information Processing Systems*, pp. 313–320, 2003.

Cortes, C. and Mohri, M. Domain adaptation and sample bias correction theory and algorithm for regression. *Theoretical Computer Science*, 519:103–126, 2014.

Cortes, C., Greenberg, S., and Mohri, M. Relative deviation learning bounds and generalization with unbounded loss functions. *Annals of Mathematics and Artificial Intelligence*, 85(1):45–70, 2019.

Cortes, C., Mohri, M., and Suresh, A. T. Relative deviation margin bounds. In *International Conference on Machine Learning*, pp. 2122–2131. PMLR, 2021.

Dasgupta, S. and Long, P. M. Boosting with diverse base classifiers. In *Conference on Computational Learning Theory*, pp. 273, 2003.

Dudík, M., Phillips, S., and Schapire, R. E. Correcting sample selection bias in maximum entropy density estimation. In *Advances in neural information processing systems*, 2005.

Feng, Y., Ren, T., Tang, Z., and Liu, Q. Accountable off-policy evaluation with kernel bellman statistics. In *International Conference on Machine Learning*, pp. 3102–3111. PMLR, 2020.

Fürnkranz, J. and Hüllermeier, E. Preference learning and ranking by pairwise comparison. *Preference Learning*, pp. 65–82, 2010.

Gurbuzbalaban, M., Simsekli, U., and Zhu, L. The heavy-tail phenomenon in sgd. In *International Conference on Machine Learning*, pp. 3964–3975. PMLR, 2021.

Gurbuzbalaban, M., Hu, Y., Simsekli, U., Yuan, K., and Zhu, L. Heavy-tail phenomenon in decentralized sgd. *arXiv preprint arXiv:2205.06689*, 2022.

Hu, T., Fan, J., Wu, Q., and Zhou, D.-X. Learning theory approach to minimum error entropy criterion. *Journal of Machine Learning Research*, 14(1):377–397, 2013.

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

Kontorovich, L. A. and Ramanan, K. Concentration inequalities for dependent random variables via the martingale method. *The Annals of Probability*, 36(6):2126–2158, 2008.

Kuchibhotla, A. K. and Chakrabortty, A. 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.

Kumar, A., Niculescu-mizil, A., Kavukcuoglu, K., and Daume, H. A binary classification framework for two-stage multiple kernel learning. In *International Conference on Machine Learning*, pp. 1331–1338, 2012.

Kutin, S. 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.

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

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

Lei, X. and Zhang, H. Non-asymptotic optimal prediction error for rkhs-based partially functional linear models. *arXiv preprint arXiv:2009.04729*, 2020.

Lopez, R., Regier, J., Jordan, M. I., and Yosef, N. Information constraints on auto-encoding variational bayes. *Advances in neural information processing systems*, 31, 2018.

Mansour, Y., Mohri, M., and Rostamizadeh, A. Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430*, 2009.

Marton, K. Bounding $\bar{d}$-distance by informational divergence: a method to prove measure concentration. *The Annals of Probability*, 24(2):857–866, 1996.

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

Maurer, A. Thermodynamics and concentration. *Bernoulli*, 18(2):434–454, 2012.

Maurer, A. A note on exponential efron-stein inequalities. 2017. URL http://www.andreas-maurer.eu/expoefronstein.pdf.

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

Maurer, A. and Pontil, M. Uniform concentration and symmetrization for weak interactions. In *Conference on Learning Theory*, pp. 2372–2387, 2019.

Maurer, A. and Pontil, M. Concentration inequalities under sub-gaussian and sub-exponential conditions. In *Advances in Neural Information Processing Systems*, 2021.

Maurer, A., Parletta, D. A., Paudice, A., and Pontil, M. Robust unsupervised learning via l-statistic minimization. In *International Conference on Machine Learning*, pp. 7524–7533, 2021.

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

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

Mises, R. v. On the asymptotic distribution of differentiable statistical functions. *The annals of mathematical statistics*, 18(3):309–348, 1947.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of machine learning*. MIT press, 2018.

Mukherjee, S. and Wu, Q. Estimation of gradients and coordinate covariation in classification. *Journal of Machine Learning Research*, 7(88):2481–2514, 2006.

Paulin, D. Concentration inequalities for markov chains by marton couplings and spectral methods. *Electron. J. Probab*, 20(79):1–32, 2015.

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

Shen, Y., Han, F., and Witten, D. Exponential inequalities for dependent v-statistics via random fourier features. *Electronic Journal of Probability*, 25:1–18, 2020.

Simsekli, U., Sagun, L., and Gurbuzbalaban, M. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning*, pp. 5827–5837. PMLR, 2019.

Torralba, A., Fergus, R., and Freeman, W. T. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 30(11):1958–1970, 2008.

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

Vladimirova, M., Verbeek, J., Mesejo, P., and Arbel, J. Understanding priors in bayesian neural networks at the unit level. In *International Conference on Machine Learning*, pp. 6458–6467. PMLR, 2019.

Vladimirova, M., Girard, S., Nguyen, H., and Arbel, J. Sub-weibull distributions: Generalizing sub-gaussian and sub-exponential properties to heavier tailed distributions. *Stat*, 9(1):e318, 2020.

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

Yang, X., Liu, X., and Wei, H. Concentration inequalities of mle and robust mle. *arXiv preprint arXiv:2210.09398*, 2022.

Ying, Y., Wen, L., and Lyu, S. Stochastic online auc maximization. In *International Conference on Neural Information Processing Systems*, pp. 451–459, 2016.

Zhang, R.-R. When janson meets mcdiarmid: Bounded difference inequalities under graph-dependence. *Statistics & Probability Letters*, 181:109272, 2022.

A. Proof of Main Results

The proof of our main results uses the entropy method. We introduce some necessary tools useful to our proof, which are collected from (Maurer, 2012; Maurer & Pontil, 2021).

The entropy $S(Z)$ of a real valued random variable Z is defined as

$$S(Z) = \mathbb{E}_Z[Z] - \ln \mathbb{E}[e^Z],$$

where the expectation functional $\mathbb{E}_Z$ is defined as $\mathbb{E}_Z[Y] = \mathbb{E}[Ye^Z]/\mathbb{E}[e^Z]$. Note that the meaning of this notation $\mathbb{E}_Z$ is different from the notation $\mathbb{E}_{Z\sim\mu}$ in Definition 5.1. Besides, we have the following fluctuation representation of the entropy.

Lemma A.1 (Theorem 3 in (Maurer, 2012))**.** *For $\gamma > 0$, we have*

$$S(\gamma Z) = \int_0^\gamma \left(\int_t^\gamma \mathbb{E}_{sZ}[(Z - \mathbb{E}_{sZ}[Z])^2] ds \right) dt.$$

We present an important Lemma that shows how bounds on the entropy can lead to concentration results.

Lemma A.2 (Theorem 1 in (Maurer, 2012))**.** *For any $f : \mathcal{X}^n \to \mathbb{R}$ and $\beta > 0$, we have*

$$\ln \mathbb{E}\left[e^{\beta(f(X)-\mathbb{E}[f(X')])}\right] = \beta \int_0^\beta \frac{S(\gamma f(X))}{\gamma^2} d\gamma,$$

and, for any $t \geq 0$,

$$\mathbb{P}\left(f(X) - \mathbb{E}[f(X')] > t\right) \leq \exp\left(\beta \int_0^\beta \frac{S(\gamma f(X)) d\gamma}{\gamma^2} - \beta t\right).$$

The following is an important lemma which shows the sub-additivity of entropy and states that the total entropy is no greater than the thermal average of the sum of the conditional entropies. Before presenting it, we introduce the conditional entropy. If $f : \mathcal{X}^n \to \mathbb{R}$, X and f_k are as in Section 2 then the conditional entropy is the function $S_{f,k} : \mathcal{X}^n \to \mathbb{R}$ defined by $S_{f,k}(x) = S(f_k(X)(x))$ for $x \in \mathcal{X}^n$.

Lemma A.3 (Theorem 6 in (Maurer, 2012))**.** *The sub-additivity of entropy is*

$$S(f(X)) \leq \mathbb{E}_{f(X)}\left[\sum_{k=1}^n S_{f,k}(X)\right].$$

In our proofs, we also use the following decoupling technique.

Lemma A.4 (Lemma 5 in (Maurer, 2019))**.** *We have for any function $g : \mathcal{X}^n \to \mathbb{R}$ that*

$$\mathbb{E}_{\gamma f}[g] \leq S(\gamma f) + \ln \mathbb{E}[e^g].$$

The following Lemma considers the weakly self-boundedness of f. We first give a definition, and then show the Lemma.

Definition A.5. Define an operator D of $f : \mathcal{X}^n \to \mathbb{R}$ as

$$Df = \sum_k (f - \inf_{y \in \mathcal{X}} S_y^k f)^2.$$

Lemma A.6 (Theorem 19 in (Maurer, 2012))**.** *Suppose that*

$$Df \leq a^2 f.$$

Then for $\beta \in (0, 2/a^2)$

$$\ln \mathbb{E}[e^{\beta f}] \leq \frac{\beta \mathbb{E} f}{1 - a^2\beta/2},$$

and for any $t > 0$,

$$\mathbb{P}\left(f - \mathbb{E}[f] > t\right) \leq \exp\left(\frac{-t^2}{2a^2\mathbb{E}[f] + a^2 t}\right). \tag{9}$$

We then show the weakly self-boundedness of $\sum_{k=1}^n \|f_k(X)\|_a^2(X)$, where $a = \{\psi_2, \psi_1, 2\}$. For brevity, we denote the sum of conditional variance proxies $\sum_{k=1}^n \|f_k(X)\|_a^2(X)$ as $\sum(f)^2$, which together with Lemma A.6 shows that the sum of conditional variance proxies has the suitable properties of concentration around its mean.

Lemma A.7. *We have $D(\sum(f)^2) \leq I_a^2(f) \sum(f)^2$ for any $f : \mathcal{X}^n \to \mathbb{R}$, where $a = \{\psi_2, \psi_1, 2\}$.*

Proof. For $l \in \{1, ..., n\}$ let $z_l \in \mathcal{X}$ be a minimizer in z of $S_z^l \sum(f)^2$, so that

$$\inf_{z \in \mathcal{X}} S_z^l \sum(f)^2 = S_{z_l}^l \sum(f)^2 = \sum S_{z_l}^l \|f_k(X)\|_a^2(X) = \sum_{k:k\neq l} S_z^l \|f_k(X)\|_a^2(X) + \|f_l(X)\|_a^2(X),$$

where the last step uses the fact that $S_{z_l}^l \|f_l(X)\|_a^2(X) = \|f_l(X)\|_a^2(X)$, because for $l \in \{1, ..., n\}$, the substitution operator S_z^l is homomorphism (linear and multiplicative) w.r.t. f and the identity w.r.t. the l-th coordinate.

Thus we get

$$\begin{aligned}
&D(\textstyle\sum(f)^2)\\
=&\sum_l \Big(\textstyle\sum(f)^2 - S_{z_l}^l \sum(f)^2\Big)^2\\
=&\sum_l \Big(\sum_{k:k\neq l} (\|f_k(X)\|_a^2(X) - S_{z_l}^l \|f_k(X)\|_a^2(X))^2\Big)^2\\
=&\sum_l \Big(\sum_{k:k\neq l} (\|f_k(X)\|_a^2(X) - \|S_{z_l}^l f_k(X)\|_a^2(X))^2\Big)^2\\
=&\sum_l \Big(\sum_{k:k\neq l} (\|f_k(X)\|_a(X) - \|S_{z_l}^l f_k(X)\|_a(X)) \times (\|f_k(X)\|_a(X) + \|S_{z_l}^l f_k(X)\|_a(X))\Big)^2\\
\leq&\sum_l \sum_{k:k\neq l} \Big(\|f_k(X)\|_a(X) - \|S_{z_l}^l f_k(X)\|_a(X)\Big)^2 \times \sum_{k:k\neq l} \Big(\|f_k(X)\|_a(X) + \|S_{z_l}^l f_k(X)\|_a(X)\Big)^2,
\end{aligned}$$

where the third step uses the fact that $S_{z_l}^l \|f_k(X)\|_a^2(X) = \|S_{z_l}^l f_k(X)\|_a^2(X)$ and the last inequality uses the Cauchy-Schwarz inequality. Then, we can get

$$\begin{aligned}
&\sum_{k:k\neq l} \Big(\|f_k(X)\|_a(X) + \|S_{z_l}^l f_k(X)\|_a(X)\Big)^2 \leq 2 \sum_{k:k\neq l} \|f_k(X)\|_a^2(X) + \|S_{z_l}^l f_k(X)\|_a^2(X)\\
\leq& 2(\textstyle\sum(f)^2 + S_{z_l}^l \sum(f)^2) \leq 4 \sum(f)^2,
\end{aligned}$$

where the first inequality uses $(a + b)^2 \leq 2(a^2 + b^2)$.

Now we obtain that

$$\begin{aligned}
&D(\textstyle\sum(f)^2)\\
\leq& 4\sum_l \sum_{k:k\neq l} \Big(\|f_k(X)\|_a(X) - \|S_{z_l}^l f_k(X)\|_a(X)\Big)^2 \textstyle\sum(f)^2\\
\leq& 4\sum_l \sum_{k:k\neq l} \|f_k(X) - S_{z_l}^l f_k(X)\|_a^2(X) \textstyle\sum(f)^2\\
\leq& 4\sum_l \sup_{z\in\mathcal{X}} \sum_{k:k\neq l} \|f_k(X) - S_z^l f_k(X)\|_a^2(X) \textstyle\sum(f)^2\\
\leq& 4 \sup_{x\in\mathcal{X}^n} \sum_l \sup_{z\in\mathcal{X}} \sum_{k:k\neq l} \|f_k(X) - S_z^l f_k(X)\|_a^2(x) \textstyle\sum(f)^2\\
\leq& I_a^2(f) \textstyle\sum(f)^2,
\end{aligned}$$

where the second inequality uses the norm's triangle inequality. The proof is complete. □

A.1. Proof of Theorem 3.2

We first give a bound on the entropy of a sub-Gaussian random variable.

Lemma A.8. *For any centered random variable Z, if Z is sub-Gaussian and β is real, then we have $S(\beta Z) \leq 4\beta^2 \|Z\|_{\psi_2}^2$.*

Proof. We have

$$S(Z) = \mathbb{E}_Z[Z] - \ln \mathbb{E}[e^Z] = \mathbb{E}_Z\left[\ln \frac{e^Z}{\mathbb{E}e^Z}\right] \leq \ln \mathbb{E}_Z\left[\frac{e^Z}{\mathbb{E}e^Z}\right] = \ln \mathbb{E}[e^{2Z}] - 2\ln \mathbb{E}[e^Z] \leq \ln \mathbb{E}[e^{2Z}],$$

where the first and second inequalities use the Jensen's inequality. We next focus on the bound of $\mathbb{E}[e^{\beta Z}]$ for all $\beta \in \mathbb{R}$. By Cauchy's inequality and arithmetic-geometric mean inequality, we have

$$\mathbb{E}|\beta Z|^{2m+1} \leq (\mathbb{E}|\beta Z|^{2m}\mathbb{E}|\beta Z|^{2m+2})^{\frac{1}{2}} \leq \frac{1}{2}(\beta^{2m}\mathbb{E}Z^{2m} + \beta^{2m+2}\mathbb{E}Z^{2m+2}),$$

which gives

$$\mathbb{E}[e^{\beta Z}] = 1 + \sum_{m=2}^{\infty} \frac{\beta^m \mathbb{E}Z^m}{m!} \leq 1 + \left(\frac{1}{2} + \frac{1}{2\times 3!}\right)\beta^2\mathbb{E}Z^2 + \sum_{m=2}^{\infty}\left(\frac{1}{(2m)!} + \frac{1}{2}\left[\frac{1}{(2m-1)!} + \frac{1}{(2m+1)!}\right]\right)\beta^{2m}\mathbb{E}Z^{2m}$$
$$\leq \sum_{m=0}^{\infty} 2^m \frac{\beta^{2m}\mathbb{E}Z^{2m}}{(2m)!} \leq \exp(\beta^2\|Z\|_{\psi_2}^2).$$

Thus, we get

$$S(\beta Z) \leq 4\beta^2\|Z\|_{\psi_2}^2.$$

The proof is complete. □

We begin to prove Theorem 3.2.

Proof. By Lemma A.8, we have

$$S_{\gamma f,k}(X) = S(\gamma f_k(X)(X)) \leq 4\gamma^2\|f_k(X)\|_{\psi_2}^2(X).$$

Together with Lemma A.3, we get

$$S(\gamma f(X)) \leq \mathbb{E}_{\gamma f(X)}\left[\sum_{k=1}^n S_{\gamma f,k}(X)\right] \leq 4\gamma^2\mathbb{E}_{\gamma f(X)}\left[\sum_{k=1}^n \|f_k(X)\|_{\psi_2}^2(X)\right].$$

Again, denote $\sum_{k=1}^n \|f_k(X)\|_{\psi_2}^2(X)$ as $\sum(f)^2$. Let $0 < \gamma \leq \beta < \frac{1}{\sqrt{2}a}$ and take $\theta := \frac{1}{a}2\sqrt{2}\gamma$, we can get

$$\theta < \frac{2}{a^2} \quad \text{and} \quad \theta > 4\gamma^2.$$

Now, using Lemma A.4, we have

$$\theta S(\gamma f(X)) \leq 4\gamma^2\mathbb{E}_{\gamma f(X)}\left[\theta \sum (f)^2\right] \leq 4\gamma^2\left(S(\gamma f(X)) + \ln \mathbb{E}\left[e^{\theta\sum(f)^2}\right]\right) \leq 4\gamma^2 S(\gamma f(X)) + \frac{4\gamma^2\theta\mathbb{E}\sum(f)^2}{1-a^2\theta/2},$$

where the last inequality follows from Lemma A.6.

Thus, we have

$$S(\gamma f(X)) \leq \frac{1}{\theta - 4\gamma^2}\frac{4\gamma^2\theta\mathbb{E}\sum(f)^2}{1-a^2\theta/2}) = \frac{4\gamma^2\mathbb{E}\sum(f)^2}{(1-a\sqrt{2}\gamma)^2},$$

which implies that

$$\int_0^\beta \frac{S(\gamma f(X))d\gamma}{\gamma^2} \leq \int_0^\beta \frac{4\mathbb{E}\sum(f)^2}{(1-a\sqrt{2}\gamma)^2}d\gamma = \frac{\beta 4\mathbb{E}\sum(f)^2}{1-a\sqrt{2}\beta}.$$

To proceed, we need an optimization Lemma.

Lemma A.9 (Lemma 14 in (Maurer & Pontil, 2021))**.** *Let C and b denote two positive real numbers, $t > 0$. Then*

$$\inf_{\beta \in [0,1/b)} \left(-\beta t + \frac{C\beta^2}{1 - b\beta} \right) \leq \frac{-t^2}{2(2C + bt)}.$$

Now, by Lemma A.2, we get

$$\begin{aligned}
&\mathbb{P}\left(f(X) - \mathbb{E}[f(X')] > t\right) \leq \inf_{\beta > 0} \exp\left(\beta \int_0^\beta \frac{S(\gamma f(X)) d\gamma}{\gamma^2} - \beta t \right) \\
\leq& \inf_{\beta > 0} \exp\left(\frac{4\beta^2 \mathbb{E} \sum (f)^2}{1 - a\sqrt{2}\beta} - \beta t \right) \leq \exp\left(\frac{-t^2}{2(8\mathbb{E}[\sum (f)^2] + a\sqrt{2}t)} \right),
\end{aligned}$$

where the last inequality uses Lemma A.9. According to Lemma A.7, by substituting $I_{\psi_2}(f)$ for a, we finally get

$$\begin{aligned}
&\mathbb{P}\left(f(X) - \mathbb{E}[f(X')] > t\right) \leq \exp\left(\frac{-t^2}{2(8\mathbb{E}[\sum (f)^2] + I_{\psi_2}(f)\sqrt{2}t)} \right) \\
=& \exp\left(\frac{-t^2}{2(8\mathbb{E}[\sum_{k=1}^n \|f_k(X)\|_{\psi_2}^2(X)] + I_{\psi_2}(f)\sqrt{2}t)} \right).
\end{aligned}$$

The proof is complete. □

A.2. Proof of Theorem 3.4

We first give a bound on the entropy of a sub-exponential random variable.

Lemma A.10. *For any centered random variable Z, if Z is sub-exponential and $\|Z\|_{\psi_1} \leq 1$, then we have $S(Z) \leq \frac{\|Z\|_{\psi_1}^2}{(1 - \|Z\|_{\psi_1})^2}$.*

Proof. According to Lemma A.1, we have

$$S(\gamma Z) = \int_0^\gamma \left(\int_t^\gamma \mathbb{E}_{sZ}[(Z - \mathbb{E}_{sZ}[Z])^2] ds \right) dt.$$

Next, we show

$$\begin{aligned}
&\mathbb{E}_{sZ}[(Z - \mathbb{E}_{sZ}[Z])^2] \leq \mathbb{E}_{sZ} Z^2 = \frac{\mathbb{E}[Z^2 e^{sZ}]}{\mathbb{E}[e^{sZ}]} \leq \mathbb{E}[Z^2 e^{sZ}] \\
=& \mathbb{E}\left[\sum_{m=0}^\infty \frac{s^m Z^{m+2}}{m!} \right] \leq \sum_{m=0}^\infty \frac{s^m (m+2)! \|Z\|_{\psi_1}^{m+2}}{m!} = \sum_{m=0}^\infty s^m (m+2)(m+1) \|Z\|_{\psi_1}^{m+2},
\end{aligned}$$

where the first inequality follows from the variational property of variance, the second from Jensen's inequality and $\mathbb{E}Z = 0$, and the third from the definition of the sub-exponential norm. Together with this inequality gives

$$\begin{aligned}
S(Z) =& \int_0^1 \left(\int_t^1 \sum_{m=0}^\infty s^m (m+2)(m+1) \|Z\|_{\psi_1}^{m+2} ds \right) dt \\
\leq& \sum_{m=0}^\infty (m+1) \|Z\|_{\psi_1}^{m+2} = \|Z\|_{\psi_1}^2 \sum_{m=0}^\infty (m+1) \|Z\|_{\psi_1}^m = \frac{\|Z\|_{\psi_1}^2}{(1 - \|Z\|_{\psi_1})^2}.
\end{aligned}$$

The proof is complete. □

We begin to prove Theorem 3.4.

Proof. By Lemma A.10, we first show that

$$S_{\gamma f,k}(X) = S(\gamma f_k(X)(X)) \leq \frac{\gamma^2 \|f_k(X)\|_{\psi_1}^2(X)}{(1-\gamma\|f_k(X)\|_{\psi_1}(X))^2} \leq \frac{\gamma^2 \|f_k(X)\|_{\psi_1}^2(X)}{(1-\gamma M)^2}.$$

By Lemma A.3, we get

$$S(\gamma f(X)) \leq \mathbb{E}_{\gamma f(X)}\left[\sum_{k=1}^{n} S_{\gamma f,k}(X)\right] \leq \frac{\gamma^2 \mathbb{E}_{\gamma f(X)}\left[\sum_{k=1}^{n} \|f_k(X)\|_{\psi_1}^2(X)\right]}{(1-\gamma M)^2}.$$

Again, denote $\sum_{k=1}^{n}\|f_k(X)\|_{\psi_1}^2(X)$ as $\sum(f)^2$. Let $0<\gamma\leq\beta<\frac{1}{M+a/\sqrt{2}}$ and take $\theta := \frac{\sqrt{2}\gamma}{a(1-\gamma M)}$, we can get

$$\theta < \frac{2}{a^2} \quad \text{and} \quad \theta > \frac{\gamma^2}{(1-\gamma M)^2}.$$

Now, using Lemma A.4, we have

$$\begin{aligned}\theta S(\gamma f(X)) &\leq \frac{\gamma^2 \mathbb{E}_{\gamma f(X)}[\theta\sum(f)^2]}{(1-\gamma M)^2}\\ &\leq \frac{\gamma^2}{(1-\gamma M)^2}\left(S(\gamma f(X)) + \ln \mathbb{E}[e^{\theta\sum(f)^2}]\right) \leq \frac{\gamma^2}{(1-\gamma M)^2}\left(S(\gamma f(X)) + \frac{\theta\mathbb{E}\sum(f)^2}{1-a^2\theta/2}\right),\end{aligned}$$

where the last inequality follows from Lemma A.6.

We now get

$$\left(\theta - \frac{\gamma^2}{(1-\gamma M)^2}\right) S(\gamma f(X)) \leq \frac{\gamma^2}{(1-\gamma M)^2}\frac{\theta\mathbb{E}\sum(f)^2}{1-a^2\theta/2},$$

which implies that

$$S(\gamma f(X)) \leq \frac{a\gamma}{\sqrt{2}(1-M\gamma-a\gamma/\sqrt{2})}\frac{\theta\mathbb{E}\sum(f)^2}{1-a^2\theta/2} = \frac{\gamma^2\mathbb{E}\sum(f)^2}{(1-M\gamma-a\gamma/\sqrt{2})^2}.$$

Thus we get

$$\int_0^\beta \frac{S(\gamma f(X))d\gamma}{\gamma^2} \leq \int_0^\beta \frac{\mathbb{E}\sum(f)^2}{(1-M\gamma-a\gamma/\sqrt{2})^2}d\gamma = \frac{\beta\mathbb{E}\sum(f)^2}{1-M\beta-a\beta/\sqrt{2}}.$$

Further, by Lemma A.2, we obtain

$$\begin{aligned}\mathbb{P}\left(f(X)-\mathbb{E}[f(X')]>t\right) &\leq \inf_{\beta>0}\exp\left(\beta\int_0^\beta \frac{S(\gamma f(X))d\gamma}{\gamma^2} - \beta t\right)\\ &\leq \inf_{\beta>0}\exp\left(\frac{\beta^2\mathbb{E}\sum(f)^2}{1-M\beta-a\beta/\sqrt{2}} - \beta t\right) \leq \exp\left(\frac{-t^2}{2(2\mathbb{E}\sum(f)^2+(M+a/\sqrt{2})t)}\right),\end{aligned}$$

where the last inequality uses Lemma A.9. According to Lemma A.7, by substituting $I_{\psi_1}(f)$ for a, we finally get

$$\begin{aligned}\mathbb{P}\left(f(X)-\mathbb{E}[f(X')]>t\right) &\leq \exp\left(\frac{-t^2}{2(2\mathbb{E}\sum(f)^2+(M+I_{\psi_1}(f)/\sqrt{2})t)}\right)\\ &\leq \exp\left(\frac{-t^2}{2(2\mathbb{E}[\sum_{k=1}^{n}\|f_k(X)\|_{\psi_1}^2(X)]+(M+I_{\psi_1}(f)/\sqrt{2})t)}\right).\end{aligned}$$

The proof is complete. □

A.3. Proof of Theorem 3.5

We first give a bound on the entropy of a bounded random variable.

Lemma A.11. *For any centered random variable Z, if $Z \leq b$ where $b > 0$, then we have $S(\gamma Z) \leq \Big(\frac{1}{b}e^{b\gamma}\gamma - \frac{1}{b^2}e^{b\gamma} + \frac{1}{b^2}\Big)\|Z\|_2^2$.*

Proof. By Lemma A.1, we have

$$S(\gamma Z) = \int_0^\gamma \left(\int_t^\gamma \mathbb{E}_{sZ}[(Z - \mathbb{E}_{sZ}[Z])^2]ds\right)dt.$$

Next,

$$\mathbb{E}_{sZ}[(Z - \mathbb{E}_{sZ}[Z])^2] \leq \mathbb{E}[Z^2 e^{sZ}] \leq e^{bs}\mathbb{E}[Z^2] = e^{bs}\|Z\|_2^2.$$

Together with this inequality gives

$$S(\gamma Z) = \int_0^\gamma \left(\int_t^\gamma e^{bs}\mathbb{E}[Z^2]ds\right)dt = \Big(\frac{1}{b}e^{b\gamma}\gamma - \frac{1}{b^2}e^{b\gamma} + \frac{1}{b^2}\Big)\|Z\|_2^2.$$

The proof is complete. □

We begin to prove Theorem 3.5.

Proof. Given $u > 0$, set

$$Z' := Z\mathbb{I}_{Z \leq u} \quad \text{and} \quad Z'' := Z\mathbb{I}_{Z > u}.$$

Then, we define three events $E_1 : f(X) - \mathbb{E}[f(X')] > t$, $E_2 : \exists k, f_k(X)(X) > u$ and E_3 as the mutually exclusive event of E_1. We decude

$$\mathbb{P}\left(E_1\right) \leq \mathbb{P}(E_1 E_3) + \mathbb{P}(E_2).$$

We first focus on $\mathbb{P}(E_1E_3)$. The proof follows the proof of Theorem 1 in (Maurer, 2019). Without losing generality, we consider the case $u = 1$. Then, by Lemma A.11,

$$S_{\gamma f,k}(X) = S(\gamma f_k(X)(X)) \leq (e^\gamma\gamma - e^\gamma + 1)\|f_k(X)\|_2^2(X).$$

Using Lemma A.3, we get

$$S(\gamma f(X)) \leq \mathbb{E}_{\gamma f(X)}\left[\sum_{k=1}^n S_{\gamma f,k}(X)\right] \leq (e^\gamma\gamma - e^\gamma + 1)\mathbb{E}_{\gamma f(X)}\left[\sum_{k=1}^n \|f_k(X)\|_2^2(X)\right].$$

Again, denote $\sum_{k=1}^n \|f_k(X)\|_2^2(X)$ as $\sum(f)^2$. Let $0 < \gamma \leq \beta < \frac{1}{1/3 + a/2}$ and take $\theta := \frac{\sqrt{2(e^\gamma\gamma - e^\gamma + 1)}}{a}$, according to Lemma 6 in (Maurer, 2019), we have

$$\theta < \frac{2}{a^2} \quad \text{and} \quad \theta > (e^\gamma\gamma - e^\gamma + 1).$$

Now, using Lemma A.4, we have

$$\begin{aligned}\theta S(\gamma f(X)) \leq (e^\gamma\gamma - e^\gamma + 1)\mathbb{E}_{\gamma f(X)}[\theta \textstyle\sum(f)^2] \leq & (e^\gamma\gamma - e^\gamma + 1)(S(\gamma f(X)) + \ln\mathbb{E}[e^{\theta\sum(f)^2}]) \\ \leq & (e^\gamma\gamma - e^\gamma + 1)\left(S(\gamma f(X)) + \frac{\theta\mathbb{E}\sum(f)^2}{1 - a^2\theta/2}\right),\end{aligned}$$

where the last inequality follows from Lemma A.6.

We now get

$$(\theta - (e^\gamma \gamma - e^\gamma + 1))S(\gamma f(X)) \leq (e^\gamma \gamma - e^\gamma + 1)\frac{\theta \mathbb{E}\sum(f)^2}{1 - a^2\theta/2},$$

which implies that

$$S(\gamma f(X)) \leq (e^\gamma \gamma - e^\gamma + 1)\frac{\theta \mathbb{E}\sum(f)^2}{1 - a^2\theta/2} = \frac{(e^\gamma \gamma - e^\gamma + 1)\mathbb{E}\sum(f)^2}{(1 - a\sqrt{(e^\gamma \gamma - e^\gamma + 1)/2})^2}.$$

Thus we get

$$\int_0^\beta \frac{S(\gamma f(X))d\gamma}{\gamma^2} \leq \int_0^\beta \frac{(e^\gamma \gamma - e^\gamma + 1)\mathbb{E}\sum(f)^2 d\gamma}{\gamma^2(1 - a\sqrt{(e^\gamma \gamma - e^\gamma + 1)/2})^2} \leq \int_0^\beta \frac{\mathbb{E}\sum(f)^2 d\gamma}{2(1 - (1/3 + a/2)\gamma)^2} = \frac{\mathbb{E}\sum(f)^2}{2}\frac{\beta}{1 - (1/3 + a/2)\beta},$$

where the second inequality uses Lemma 6 in (Maurer, 2019). Further, by Lemma A.2, we obtain the following inequality for the case $u = 1$

$$\begin{aligned}\mathbb{P}(E_1 E_3) \leq \inf_{\beta > 0} \exp\left(\beta \int_0^\beta \frac{S(\gamma f(X))d\gamma}{\gamma^2} - \beta t\right) &\leq \inf_{\beta > 0} \exp\left(\frac{\mathbb{E}\sum(f)^2}{2}\frac{\beta^2}{1 - (1/3 + a/2)\beta} - \beta t\right) \\ &\leq \exp\left(\frac{-t^2}{2(\mathbb{E}\sum(f)^2 + (1/3 + a/2)t)}\right),\end{aligned}$$

where the last inequality uses Lemma A.9. According to Lemma A.7, by substituting $I_2(f)$ for a, we finally get

$$\mathbb{P}(E_1 E_3) \leq \exp\left(\frac{-t^2}{2(\mathbb{E}\sum(f)^2 + (1/3 + I_2(f)/2)t)}\right) \leq \exp\left(\frac{-t^2}{2(\mathbb{E}[\sum_{k=1}^n \|f_k(X)\|_2^2(X)] + (1/3 + I_2(f)/2)t)}\right).$$

By rescaling, we get the following inequality for the general case of u

$$\mathbb{P}\left(E_1 E_3\right) \leq \exp\left(\frac{-t^2}{2(\mathbb{E}[\sum_{k=1}^n \|f_k(X)\|_2^2(X)] + (2u/3 + I_2(f)/2)t)}\right).$$

We then focus on $\mathbb{P}(E_2)$. Using the exponential Markov's inequality and the condition $\mathbb{E}\exp((Z^+)^{\frac{\alpha}{1-\alpha}}) \leq c$ for $\alpha \in (0, 1)$ and some constant $c \in (0, +\infty)$, we get

$$\mathbb{P}(E_2) \leq \sum_{k=1}^n \mathbb{P}(f_k(X)(X) > u) \leq \sum_{k=1}^n \mathbb{E}\exp\left(((f_k(X)(X))^+)^{\frac{\alpha}{1-\alpha}} - u^{\frac{\alpha}{1-\alpha}}\right) \leq nc\exp\left(-u^{\frac{\alpha}{1-\alpha}}\right).$$

Combining the inequalities of $\mathbb{P}(E_1 E_3)$ and $\mathbb{P}(E_2)$, we obtain

$$\mathbb{P}\left(f(X) - \mathbb{E}[f(X')] > t\right) \leq \exp\left(\frac{-t^2}{2(\mathbb{E}[\sum_{k=1}^n \|f_k(X)\|_2^2(X)] + (2u/3 + I_2(f)/2)t)}\right) + nc\exp(-u^{\frac{\alpha}{1-\alpha}}).$$

Taking $u = t^{1-\alpha}$, we get

$$\mathbb{P}\left(f(X) - \mathbb{E}[f(X')] > t\right) \leq \exp\left(\frac{-t^2}{2(\mathbb{E}[\sum_{k=1}^n \|f_k(X)\|_2^2(X)] + (2t^{1-\alpha}/3 + I_2(f)/2)t)}\right) + nc\exp(-t^\alpha).$$

The proof is complete. □

B. Proof of Applications

B.1. Sample Mean

Proof. For any given k, we have

$$\|f_k(X)\|_a^2(X) = \left\|\frac{1}{n}(X_k - \mathbb{E}[X_k])\right\|_a^2 = \frac{1}{n^2}\|(X_k - \mathbb{E}[X_k])\|_a^2 \leq \frac{1}{n^2}(2\|X_k\|_a)^2 \leq \frac{4}{n^2}\|X_k\|_a^2,$$

where the first inequality uses Lemma 6 (ii) in (Maurer & Pontil, 2021). And with a similar proof, we have

$$M = \max_k \|\|\|f_k(X)\|_{\psi_1}\|_\infty \leq \frac{2}{n}\max_k \|X_k\|_{\psi_1}.$$

Plugging these inequalities into the concentration inequalities gives the results of Theorem 4.1. □

B.2. U-Statistics

We have the following lemma.

Lemma B.1. *(i)* $I_a(U) \leq \frac{4m^2}{n}b$*, (ii)* $\max_k \|\|\|U_k(X)\|_{\psi_1}\|_\infty \leq \frac{m}{n}b$*, and (iii)* $\mathbb{E}[\sum_{k=1}^n \|U_k(X)\|_a^2(X)] \leq \frac{m^2}{n}\sigma^2$.

Proof. (i) For any given k, we have a decomposition

$$U(x) = \binom{n}{m}^{-1}\sum_{S\in\mathcal{S}^{m-1}_{\{1,\dots,n\}\backslash k}} f(x_S, x_k) + \binom{n}{m}^{-1}\sum_{S\in\mathcal{S}^{m}_{\{1,\dots,n\}}, k\notin S} f(x_S).$$

Thus, we get

$$U_k(X)(x) = \binom{n}{m}^{-1}\sum_{S\in\mathcal{S}^{m-1}_{\{1,\dots,n\}\backslash k}} f(x_S, X_k) - \mathbb{E}[f(x_S, X_k)]. \tag{10}$$

We deduce that

$$\begin{aligned}\|U_k(X)\|_a(x) = &\binom{n}{m}^{-1}\Big\|\sum_{S\in\mathcal{S}^{m-1}_{\{1,\dots,n\}\backslash k}} f(x_S, X_k) - \mathbb{E}[f(x_S, X_k)]\Big\|_a \\ \leq &\binom{n}{m}^{-1}\sum_{S\in\mathcal{S}^{m-1}_{\{1,\dots,n\}\backslash k}} \Big\|f(x_S, X_k) - \mathbb{E}[f(x_S, X_k)]\Big\|_a \leq \binom{n}{m}^{-1}\sum_{S\in\mathcal{S}^{m-1}_{\{1,\dots,n\}\backslash k}} b = \frac{m}{n}b,\end{aligned} \tag{11}$$

where the second inequality uses the assumption $\max_S \max_k \|f_k(X)\|_a(x_S) \leq b$.

Further, for $k \neq l$ we have

$$\begin{aligned}&S^l_{z'}U_k(X)(x) - S^l_z U_k(X)(x) \\ =&\binom{n}{m}^{-1}\sum_{S\in\mathcal{S}^{m-2}_{\{1,\dots,n\}\backslash k,l}} f(z', x_S, X_k) - \mathbb{E}[f(z', x_S, X_k)] - [f(z, x_S, X_k) - \mathbb{E}[f(z, x_S, X_k)]].\end{aligned}$$

This gives

$$\begin{aligned}&\|U_k(X) - S^l_z U_k(X)\|_a(x) \leq \sup_{z'\in\mathcal{X}} \|S^l_{z'}U_k(X) - S^l_z U_k(X)\|_a(x) \\ \leq& \sup_{z\in\mathcal{X}}\binom{n}{m}^{-1}\sum_{S\in\mathcal{S}^{m-2}_{\{1,\dots,n\}\backslash k,l}} 2\|f(z, x_S, X_k) - \mathbb{E}[f(z, x_S, X_k)]\|_a \leq \binom{n}{m}^{-1}\sum_{S\in\mathcal{S}^{m-2}_{\{1,\dots,n\}\backslash k,l}} 2b \leq \frac{m(m-1)}{n(n-1)}2b,\end{aligned}$$

where the second inequality uses the norm's triangle inequality and the third inequality uses the assumption $\max_S \max_k \|f_k(X)\|_a(x_S) \le b$. Thus,

$$I_a(U) = 2\Big(\sup_{x\in\mathcal{X}^n}\sum_l \sup_{z\in\mathcal{X}}\sum_{k:k\neq l}\|U_k(X) - S_z^l U_k(X)\|_a^2(x)\Big)^{\frac{1}{2}} \le \frac{4m(m-1)}{\sqrt{n(n-1)}}b \le \frac{4m^2}{n}b.$$

(ii) With a proof similar to (11), $\max_k \|\|U_k(X)\|_{\psi_1}\|_\infty \le \frac{m}{n}b$.

(iii) According to (10), we deduce

$$U_k(X)(X) = \binom{n}{m}^{-1}\sum_{S\in\mathcal{S}^{m-1}_{\{1,\dots,n\}\setminus k}} f(X_S, X_k) - \mathbb{E}[f(X_S, X_k)].$$

This gives

$$\|U_k(X)\|_a(X) \le \binom{n}{m}^{-1}\sum_{S\in\mathcal{S}^{m-1}_{\{1,\dots,n\}\setminus k}}\|f_k(X)\|_a(X_S).$$

Thus, using the assumption $\mathbb{E}\|f_k(X)\|_a^2(X_S) \le \sigma^2$ for any $k = \{1, ..., n\}$,

$$\mathbb{E}\left[\sum_{k=1}^n \|U_k(X)\|_a^2(X)\right] \le \frac{m^2}{n}\sigma^2.$$

The proof is complete. □

Plugging Lemma B.1 into the concentration inequalities gives the results of Theorem 4.2.

B.3. V-Statistics

We have the following lemma.

Lemma B.2. *(i)* $I_a(V) \le \frac{4m(m-1)}{n}b$*, (ii)* $\max_k \|\|V_k(X)\|_{\psi_1}\|_\infty \le \frac{m}{n}b$*, and (iii)* $\mathbb{E}[\sum_{k=1}^n \|V_k(X)\|_a^2(X)] \le \frac{m^2}{n}\sigma^2$.

Proof. The proof follows the proof of Lemma B.1.

(i) The index k can appear with m possibilities and the remaining indices can assume all values in $\{1, ..., n\}$. Thus, we get

$$V_k(X)(x) = n^{-m}\sum_{k=1}^m \sum_{S\in\{1,\dots,n\}^{m-1}} f(x_S, X_k) - \mathbb{E}[f(x_S, X_k)],$$

which implies that

$$\|V_k(X)\|_a(x) \le n^{-m}\sum_{k=1}^m \sum_{S\in\{1,\dots,n\}^{m-1}}\|f_k(X)\|_a(x_S) \le n^{-m}\sum_{k=1}^m \sum_{S\in\{1,\dots,n\}^{m-1}} b = \frac{m}{n}b, \tag{12}$$

where the second inequality uses the assumption $\max_S \max_k \|f_k(X)\|_a(x_S) \le b$.

Further, for $k \neq l$ we have a decomposition

$$\|V_k(X) - S_z^l V_k(X)\|_a(x) \le \sup_{z'}\|S_{z'}^l V_k(X) - S_z^l V_k(X)\|_a(x) \le n^{-m}\sum_l \sum_{k,k\neq l}\sum_{S\in\{1,\dots,n\}^{m-2}} 2b = \frac{m(m-1)}{n^2}2b.$$

Thus,

$$I_a(V) = 2\Big(\sup_{x\in\mathcal{X}^n}\sum_l \sup_{z\in\mathcal{X}}\sum_{k:k\neq l}\|V_k(X) - S_z^l V_k(X)\|_a^2(x)\Big)^{\frac{1}{2}} \le 4\frac{m(m-1)}{n}b.$$

(ii) With a proof similar to (12), $\max_k \|\|V_k(X)\|_{\psi_1}\|_\infty \le \frac{m}{n}b$.

(iii) Similar to the proof of Lemma B.1, we deduce

$$\mathbb{E}\left[\sum_{k=1}^{n} \|V_k(X)\|_a^2(X)\right] \le \frac{m^2}{n}\sigma^2.$$

The proof is complete. □

Plugging Lemma B.2 into the concentration inequalities gives the results of Theorem 4.3.

C. Proof of Refined Results

C.1. Proof of Theorem 5.2

To proceed, we first introduce an operator.

Definition C.1. Define an operator V^+ of $f : \mathcal{X}^n \to \mathbb{R}$ as

$$V^+ f = \sum_k \mathbb{E}_{Z\sim\mu}(f - S_Z^k f)_+^2,$$

where $Z_+ = \max\{Z, 0\}$.

The following Lemma also considers the weakly self-boundedness of f in terms of V^+.

Lemma C.2 (Corollary 4 in (Maurer, 2017))**.** *Suppose that*

$$V^+ f \le a^2 f.$$

Then for $\beta \in (0, 2/a^2)$

$$\ln \mathbb{E}[e^{\beta f}] \le \frac{\beta \mathbb{E} f}{1 - a^2\beta/2}.$$

For brevity, we denote $\sum_{k=1}^{n} \|f_k(X)\|_a^2(X)$ as $\sum(f)^2$.

Lemma C.3. *We have $V^+(\sum(f)^2) \le (I_a'(f))^2 \sum(f)^2$ for any $f : \mathcal{X}^n \to \mathbb{R}$.*

Proof. This proof follows the proof of Lemma A.7 and Proposition 6 in (Maurer, 2017). For $l \in \{1, ..., n\}$ and any $z \in \mathcal{X}$

$$S_z^l \sum(f)^2 = \sum_{k:k\neq l} S_z^l \|f_k(X)\|_a^2(X) + \|f_l(X)\|_a^2(X),$$

where we use the fact that $S_z^l \|f_l(X)\|_a^2(X) = \|f_l(X)\|_a^2(X)$, because for $l \in \{1, ..., n\}$, the substitution operator S_z^l is homomorphism (linear and multiplicative) w.r.t. f and the identity w.r.t. the l-th coordinate.

Thus we get

$$\begin{aligned}
V^+(\textstyle\sum(f)^2) &= \sum_l \mathbb{E}_{Z\sim\mu}\Big[\Big(\textstyle\sum(f)^2 - S_Z^l \sum(f)^2\Big)^2 \mathbb{I}_{A_l}(Z)\Big] \\
&= \sum_l \mathbb{E}_{Z\sim\mu}\Big[\Big(\sum_{k:k\neq l}(\|f_k(X)\|_a^2(X) - S_Z^l\|f_k(X)\|_a^2(X))^2\Big)^2 \mathbb{I}_{A_l}(Z)\Big] \\
&= \sum_l \mathbb{E}_{Z\sim\mu}\Big[\Big(\sum_{k:k\neq l}(\|f_k(X)\|_a^2(X) - \|S_Z^l f_k(X)\|_a^2(X))^2\Big)^2 \mathbb{I}_{A_l}(Z)\Big] \\
&= \sum_l \mathbb{E}_{Z\sim\mu}\Big[\Big(\sum_{k:k\neq l}(\|f_k(X)\|_a(X) - \|S_Z^l f_k(X)\|_a(X)) \times (\|f_k(X)\|_a(X) + \|S_Z^l f_k(X)\|_a(X))\Big)^2 \mathbb{I}_{A_l}(Z)\Big] \\
&\le \sum_l \mathbb{E}_{Z\sim\mu}\Big[\sum_{k:k\neq l}\Big(\|f_k(X)\|_a(X) - \|S_Z^l f_k(X)\|_a(X)\Big)^2 \times \sum_{k:k\neq l}\Big(\|f_k(X)\|_a(X) + \|S_Z^l f_k(X)\|_a(X)\Big)^2 \mathbb{I}_{A_l}(Z)\Big],
\end{aligned}$$

where the third step uses the fact that $S_Z^l\|f_k(X)\|_a^2(X) = \|S_Z^l f_k(X)\|_a^2(X)$ and the last inequality uses the Cauchy-Schwarz inequality. Using Hölder inequality, we have

$$\begin{aligned}
&V^+(\sum(f)^2)\\
\leq&\sum_l \mathbb{E}_{Z\sim\mu}\Big[\sum_{k:k\neq l}\Big(\|f_k(X)\|_a(X) - \|S_Z^l f_k(X)\|_a(X)\Big)^2 \mathbb{I}_{A_l}(Z)\Big] \times \sup_{Z\in A_l}\sum_{k:k\neq l}\Big(\|f_k(X)\|_a(X) + \|S_Z^l f_k(X)\|_a(X)\Big)^2.
\end{aligned}$$

Then, we can get

$$\begin{aligned}
&\sup_{Z\in A_l}\sum_{k:k\neq l}\Big(\|f_k(X)\|_a(X) + \|S_Z^l f_k(X)\|_a(X)\Big)^2\\
\leq&2\sum_{k:k\neq l}\|f_k(X)\|_a^2(X) + \sup_{Z\in A_l}\|S_Z^l f_k(X)\|_a^2(X)\\
\leq&2\Big(\sum(f)^2 + \sup_{Z\in A_l} S_Z^l\sum(f)^2\Big)\\
\leq&4\sum(f)^2,
\end{aligned}$$

where the first inequality uses $(a+b)^2 \leq (a^2+b^2)$, and the last uses the definition of A_l.

Now we obtain that

$$\begin{aligned}
&V^+(\sum(f)^2)\\
\leq&4\sum_l \mathbb{E}_{Z\sim\mu}\Big[\sum_{k:k\neq l}\Big(\|f_k(X)\|_a(X) - \|S_Z^l f_k(X)\|_a(X)\Big)^2 \mathbb{I}_{A_l}(Z)\Big]\sum(f)^2\\
\leq&4\sum_l \mathbb{E}_{Z\sim\mu}\Big[\sum_{k:k\neq l}\|f_k(X) - S_Z^l f_k(X)\|_a^2(X)\mathbb{I}_{A_l}(Z)\Big]\sum(f)^2\\
\leq&4\sup_{x\in\mathcal{X}^n}\sum_l \mathbb{E}_{Z\sim\mu}\Big[\sum_{k:k\neq l}\|f_k(X) - S_Z^l f_k(X)\|_a^2(x)\mathbb{I}_{A_l}(Z)\Big]\sum(f)^2\\
\leq&(I_a')^2(f)\sum(f)^2,
\end{aligned}$$

where the second inequality uses the norm's triangle inequality. The proof is complete. □

Following the proof in Section 3 and then replacing I_a by I_a', we get the results in Theorem 5.2.