Algorithmic Stability Unleashed: Generalization Bounds with Unbounded Losses

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

Abstract

One of the central problems of statistical learning theory is quantifying the generalization ability of learning algorithms within a probabilistic framework. Algorithmic stability is a powerful tool for deriving generalization bounds, however, it typically builds on a critical assumption that losses are bounded. In this paper, we relax this condition to unbounded loss functions with subweibull diameter. This gives new generalization bounds for algorithmic stability and also includes existing results of subgaussian and subexponential diameters as specific cases. Furthermore, we provide a refined stability analysis by developing generalization bounds which can be $\sqrt{n}$-times faster than the previous results, where n is the sample size. Our main technical contribution is general concentration inequalities for subweibull random variables, which may be of independent interest.

1. Introduction

One of the core problems in the machine learning community is to understand why the learned model of a machine learning algorithm on the training points performs well on the test points, which attracts many researchers to develop the theory of generalization bounds (Bousquet & Elisseeff, 2002; Shalev-Shwartz & Ben-David, 2014; Bartlett & Mendelson, 2002). Algorithmic stability has been a topic of growing interest in learning theory. It is a standard theoretic tool to prove the generalization bounds based on the sensitivity of the algorithm to changes in the learning sample, such as leaving one of the data points out or replacing it with a different one. This approach can be traced back to the foundational work of Vapnik & Chervonenkis (1974), where the generalization bound for the algorithm of hard-margin Support Vector Machine is analyzed. The ideas of stability were further developed by Rogers & Wagner (1978), Devroye & Wagner (1979a;b), Lugosi & Pawlak (1994) for the k-Nearest-Neighbor algorithm, k-local algorithms and potential learning rules, respectively. Interesting insights into stability have also been presented by many authors, such as Kearns & Ron (1997); Hardt et al. (2016); Gonen & Shalev-Shwartz (2017); Kuzborskij & Lampert (2018); Bassily et al. (2020); Liu et al. (2017); Maurer (2017); Foster et al. (2019); Feldman & Vondrak (2019); Bousquet et al. (2020), to mention but a few.

Stability arguments are known for typically providing in-expectation error bounds. In contrast, high probability guarantees require more effort. It merits noting that high probability bounds are necessary for inferring generalization when the algorithm is used many times, which is common in practice. Therefore, as compared to the in-expectation ones, high probability bounds are preferred in the study of the generalization performance. An extensive analysis of various notions of stability and the corresponding (sometimes) high probability generalization bounds are provided in the seminal work (Bousquet & Elisseeff, 2002). To give high probability bounds, McDiarmid's exponential inequality (McDiarmid, 1998) plays an essential role in the analysis. To satisfy the bounded difference condition in McDiarmid's inequality, a popularly used notion of stability allowing high probability upper bounds called uniform stability is introduced in (Bousquet & Elisseeff, 2002). In the context of uniform stability, a series of breakthrough papers (Feldman & Vondrak, 2018; 2019; Bousquet et al., 2020; Klochkov & Zhivotovskiy, 2021) provide sharper generalization bounds with probabilities.

However, when deriving high probability generalization bounds, the uniform stability implies the boundedness of the loss function, which might narrow the range of application of these results as the generalization analysis of unbounded losses is becoming increasingly important in many situations (Haddouche et al., 2021), such as regularized regression (Kontorovich, 2014), signal processing (Bakhshizadeh et al., 2020b), neural networks (Vladimirova 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),

[*]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 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

etc. For a relaxation, Kutin & Niyogi (2012) introduce a notion of "almost-everywhere" stability and proved valuable extensions of McDiarmid's inequality. It is shown in (Kutin & Niyogi, 2012) that the generalization error can still be bounded when the stability of the algorithm happens only on a subset of large measure. This influential result has been invoked in a number of interesting papers (El-Yaniv & Pechyony, 2006; Shalev-Shwartz et al., 2010; Hush et al., 2007; Mukherjee et al., 2002; Agarwal & Niyogi, 2009; Rubinstein & Simma, 2012; Rakhlin et al., 2005). However, as noted by Kontorovich (2014), the approach of Kutin & Niyogi (2012) entails too restrictive conditions. It is demonstrated in (Kontorovich, 2014) that the boundedness of loss functions can be dropped at the expense of a stronger notion of stability and a bounded subgaussian diameter of the underlying metric probability space. This fantastic idea is further, recently, improved to subexponential diameter by Maurer & Pontil (2021).

In this work, we move beyond the subgaussian and subexponential diameters and consider the generalization error bound under a much weaker tail assumption, so-called subweibull distribution (Kuchibhotla & Chakrabortty, 2022; Vladimirova et al., 2020). The subweibull distribution includes the subgaussian and subexponential distributions as specific cases and is inducing more and more attention in learning theory due to that it allows heavier-tailed losses than the sub-exponential and sub-Gaussian (Zhang & Wei, 2022; Bong & Kuchibhotla, 2023; Madden et al., 2020; Bakhshizadeh et al., 2020a).

In summary, our contributions are three-fold. Firstly, we provide novel concentration inequalities for general functions of independent subweibull random variables, including a moment inequality and a probabilistic inequality. The technical challenge here is that the subweibull distribution is heavy-tailed so the proof method in the related work (Kontorovich, 2014; Maurer & Pontil, 2021) does not hold in this paper. To counter this difficulty, we address it from the perspective of moment inequality. It should be noted that our concentration inequalities may be of independent interest. Secondly, we prove a high probability generalization bound for algorithmic stability with unbounded losses via our probabilistic inequality. To this end, we define the subweibull diameter of a metric probability space and prove that it can be used to relax the boundedness condition. The results here extend previous bounds in (Kontorovich, 2014; Maurer & Pontil, 2021) to the heavy-tailed subweibull diameter. Finally, we show an improved generalization bound which can be $\sqrt{n}$-times faster than the related results and our results in the previous part via our moment inequality. With an application to regularized metric regression, our generalization bound not only extends results in (Kontorovich, 2014; Maurer & Pontil, 2021) to more scenarios but also give sharper results.

The paper is organized as follows. We present our main results in Section 2. The preliminaries relevant to our discussion are stated in Section 2.1. The concentration inequalities are provided in Section 2.2. Section 2.3 is devoted to provide generalization bounds for algorithmic stability with unbounded losses. Section 2.4 aims to provide sharper generalization bounds. We give two applications of our main results in Section 2.5 and Section 2.6. All the proofs are deferred to the Appendix.

2. Main Results

In this section, we present the main results.

2.1. Preliminaries

This paper considers the metric space. A metric probability space $(\mathcal{X}, d, \mu)$ is a measurable space $\mathcal{X}$ whose Borel σ-algebra is induced by the metric d, endowed with the probability measure μ. A function $f : \mathcal{X} \to \mathbb{R}$ is L-Lipschitz if

$$|f(x) - f(x')| \leq Ld(x, x'), \quad x, x' \in \mathcal{X}.$$

Let $(\mathcal{X}_i, d_i, \mu_i)$, $i = 1, ..., n$, be a sequence of metric probability spaces. We define the product probability space

$$\mathcal{X}^n = \mathcal{X}_1 \times \mathcal{X}_2 \times ... \times \mathcal{X}_n$$

with the product measure

$$\mu^n = \mu_1 \times \mu_2 \times ... \times \mu_n$$

and ℓ_1 product metric

$$d^n(x, x') = \sum_{i=1}^{n} d_i(x_i, x_i'), \quad x, x' \in \mathcal{X}^n.$$

We write $X_i \sim \mu_i$ to mean that X_i is an $\mathcal{X}_i$-valued random variable with law μ_i, i.e., $\mathbb{P}(X_i \in A) = \mu_i(A)$ for all Borel $A \subset \mathcal{X}_i$. We use the notation $X_i^j = (X_i, ..., X_j)$ for all sequences. This notation extends naturally to sequences: $X_1^n \sim \mu^n$.

We now define the subweibull random variable. A real-valued random variable X is said to be subweibull if it has a bounded ψ_α-norm. The ψ_α-norm of X for any $\alpha > 0$ is defined as

$$\|X\|_{\psi_\alpha} = \inf \left\{ \sigma \in (0, \infty) : \mathbb{E} \exp \left(\left(\frac{|X|}{\sigma} \right)^\alpha \right) \leq 2 \right\}.$$

As shown in (Kuchibhotla & Chakrabortty, 2022; Vladimirova et al., 2020), the subweibull random variable is characterized by the right tail of the Weibull distribution and generalizes subgaussian and subexponential distributions. Particularly, when $\alpha = 1$ or 2, subweibull random

variables reduce to subexponential or subgaussian random variables, respectively. It is obvious that the smaller α is, the heavier tail the random variable has. Further, we define the subweibull diameter $\Delta_\alpha(\mathcal{X}_i)$ of the metric probability space $(\mathcal{X}_i, d_i, \mu_i)$ as

$$\Delta_\alpha(\mathcal{X}_i) = \|d_i(X_i, X_i')\|_{\psi_\alpha},$$

where $X_i, X_i' \sim \mu_i$ are independent.

In this paper, the standard order of magnitude notation such as $O(\cdot)$ and $\Omega(\cdot)$ will be used. For a pair of non-negative functions f, g the notation $f \lesssim g$ will mean that for some universal constant $c > 0$ it holds that $f \leq cg$. The Γ function is defined by the integral formula $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$, $x > 0$. The L_p norm of a real-valued random variable Z is denoted as $\|Z\|_p = (\mathbb{E}|Z|^p)^{\frac{1}{p}}$.

2.2. Concentration Inequalities

This section presents our main concentration inequalities, which will be used as an fundamental tool in the following sections. Our first result is a moment inequality, which is essential for Section 2.4.

Theorem 2.1. *Let $X_1, ..., X_n$ 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 $g = f(X_1, ..., X_{i-1}, X_i, X_{i+1}, ..., X_n)$ and $g_i = f(X_1, ..., X_{i-1}, X_i', X_{i+1}, ..., X_n)$, where $(X_1', ..., X_n')$ is an independent copy of $(X_1, ..., X_n)$. Assume moreover that*

$$|g - g_i| \leq H_i(X_i, X_i')$$

for some functions $H_i : \mathcal{X}^2 \to \mathbb{R}$, $i = 1, ..., n$. Suppose that $\|H_i(X_i, X_i')\|_{\psi_\alpha} < \infty$ for all $i = 1, ..., n$. For any $p \geq 2$,

1.) if $0 < \alpha \leq 1$, let $c_\alpha = 2\sqrt{2}((\log 2)^{1/\alpha} + e^3\Gamma^{1/2}(\frac{2}{\alpha} + 1) + e^3 3^{\frac{2-\alpha}{3\alpha}} \sup_{p\geq 2} p^{\frac{-1}{\alpha}}\Gamma^{1/p}(\frac{p}{\alpha} + 1))$, we have

$$\begin{aligned}&\|f(X_1, ..., X_n) - \mathbb{E}f(X_1, ..., X_n)\|_p \\ \leq& c_\alpha \left(\sqrt{p} \left(\sum_{i=1}^n \|H_i(X_i, X_i')\|_{\psi_\alpha}^2 \right)^{\frac{1}{2}} \right. \\ &\left. + p^{1/\alpha} \max_{1\leq i\leq n} \|H_i(X_i, X_i')\|_{\psi_\alpha} \right);\end{aligned}$$

2.) if $\alpha > 1$, let $1/\alpha^* + 1/\alpha = 1$ and $c_\alpha' = 8e + 2(\log 2)^{1/\alpha}$, and let $(\|H(X, X')\|_{\psi_\alpha}) = (\|H_1(X_1, X_1')\|_{\psi_\alpha}, ..., \|H_n(X_n, X_n')\|_{\psi_\alpha}) \in \mathbb{R}^n$, we have

$$\begin{aligned}&\|f(X_1, ..., X_n) - \mathbb{E}f(X_1, ..., X_n)\|_p \\ \leq& c_\alpha' \left(\sqrt{p} \left(\sum_{i=1}^n \|H_i(X_i, X_i')\|_{\psi_\alpha}^2 \right)^{\frac{1}{2}} \right. \\ &\left. + p^{1/\alpha} \|(\|H(X, X')\|_{\psi_\alpha})\|_{\alpha^*} \right).\end{aligned}$$

Our second result is a probabilistic inequality, which is essential for Section 2.3.

Theorem 2.2. *Under the settings of Theorem 2.1, for any $0 < \delta < 1/e^2$, with probability at least $1 - \delta$*

1.) if $0 < \alpha \leq 1$, we have

$$\begin{aligned}&|f(X_1, ..., X_n) - \mathbb{E}f(X_1, ..., X_n)| \\ \leq& c_\alpha \left(\sqrt{\log(\frac{1}{\delta})} \left(\sum_{i=1}^n \|H_i(X_i, X_i')\|_{\psi_\alpha}^2 \right)^{\frac{1}{2}} \right. \\ &\left. + \log^{1/\alpha}(\frac{1}{\delta}) \max_{1\leq i\leq n} \|H_i(X_i, X_i')\|_{\psi_\alpha} \right);\end{aligned}$$

2.) if $\alpha > 1$, we have

$$\begin{aligned}&|f(X_1, ..., X_n) - \mathbb{E}f(X_1, ..., X_n)| \\ \leq& c_\alpha' \left(\sqrt{\log(\frac{1}{\delta})} \left(\sum_{i=1}^n \|H_i(X_i, X_i')\|_{\psi_\alpha}^2 \right)^{\frac{1}{2}} \right. \\ &\left. + \log^{1/\alpha}(\frac{1}{\delta}) \|(\|H(X, X')\|_{\psi_\alpha})\|_{\alpha^*} \right).\end{aligned}$$

Remark 2.3. H_i is an arbitrary function satisfying the subweibull condition $\|H_i(X_i, X_i')\|_{\psi_\alpha} \leq \infty$. By considering H_i as a metric function, in the context of the subweibull diameter, Theorem 2.2 becomes (1) if $0 < \alpha \leq 1$, we have

$$\begin{aligned}&|f(X_1, ..., X_n) - \mathbb{E}f(X_1, ..., X_n)| \\ \leq& c_\alpha \Big(\sqrt{\log(\frac{1}{\delta})} \Big(\sum_{i=1}^n \Delta_\alpha^2(\mathcal{X}_i) \Big)^{\frac{1}{2}} + \log^{\frac{1}{\alpha}}(\frac{1}{\delta}) \max_{1\leq i\leq n} \Delta_\alpha(\mathcal{X}_i) \Big);\end{aligned}$$

(2) if $\alpha > 1$, let $(\Delta_\alpha(\mathcal{X})) = (\Delta_\alpha(\mathcal{X}_1), ..., \Delta_\alpha(\mathcal{X}_n))$,

$$\begin{aligned}&|f(X_1, ..., X_n) - \mathbb{E}f(X_1, ..., X_n)| \\ \leq& c_\alpha' \Big(\sqrt{\log(\frac{1}{\delta})} \Big(\sum_{i=1}^n \Delta_\alpha^2(\mathcal{X}_i) \Big)^{\frac{1}{2}} + \log^{\frac{1}{\alpha}}(\frac{1}{\delta}) \|(\Delta_\alpha(\mathcal{X}))\|_{\alpha^*} \Big).\end{aligned}$$

We now provide some discussions on the optimality of our bound. In Theorem 2.2, our inequality shows a mixture of subgaussian $\sqrt{\log(\frac{1}{\delta})}$ and subweibull $\log^{1/\alpha}(\frac{1}{\delta})$ tails. The subgaussian tail is of course expected from the central limit theorem, and the subweibull tail captures the right

decaying rate of the subweibull random variable. Therefore, our inequality successfully captures the right subgaussian tail for small deviations and the right subweibull tail for large deviations, which also implies that the convergence rate will be faster for small deviations and will be slower for large deviations.

Remark 2.4. Let us see how Theorem 2.1 compares to previous results on some examples. Theorem 1 in (Kontorovich, 2014) states that if f is 1-Lipschitz function, then for any $t > 0$

$$\begin{aligned}&\mathbb{P}(|f(X_1, ..., X_n) - \mathbb{E}f(X_1, ..., X_n)| > t)\\ \leq& 2\exp\left(-\frac{t^2}{2\sum_{i=1}^n \Delta_2^2(\mathcal{X}_i)}\right).\end{aligned}$$

Theorem 11 in (Maurer & Pontil, 2021) shows that if f is 1-Lipschitz function, a one-sided inequality holds for any $t > 0$

$$\begin{aligned}&\mathbb{P}(f(X_1, ..., X_n) - \mathbb{E}f(X_1, ..., X_n) > t)\\ \leq& \exp\left(-\frac{t^2}{4e\sum_{i=1}^n \Delta_1^2(\mathcal{X}_i) + 2e\max_{1\leq i\leq n}\Delta_1(\mathcal{X}_i)t}\right).\end{aligned}$$

By comparison, it is clear that when $\alpha = 2$ or $\alpha = 1$, our inequalities, respectively, reduce to the ones in (Kontorovich, 2014; Maurer & Pontil, 2021), respectively, up to constants.

Remark 2.5. In this remark, we parse α on the concentration inequality. For the subweibull diameter $\Delta_\alpha(\mathcal{X}_i)$, according to its definition, we know that the smaller α is, the heavier tail the random variable has and the bigger this subweibull diameter is. For the subweibull tail $\log^{1/\alpha}(\frac{1}{\delta})$, it is clear that a smaller α leads to a bigger term. As for the constant c_α, according to the property of Gamma function, $\Gamma(\frac{2}{\alpha}+1)$ becomes bigger as α becomes smaller. For $\sup_{p\geq 2} p^{\frac{-1}{\alpha}}\Gamma^{1/p}(\frac{p}{\alpha}+1)$, a fine-grained analysis by Stirling formula gives us a concise form that does not depend on p, and we can find that the smaller α is, the bigger this term is. Specifically, by the Stirling formula

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

we get the following result

$$\begin{aligned}&\sup_{p\geq 2} p^{-\frac{1}{\alpha}}\Gamma^{1/p}(\frac{p}{\alpha}+1)\\ \leq& \sup_{p\geq 2} p^{-\frac{1}{\alpha}}\left(\sqrt{2\pi p/\alpha}(\frac{p}{e\alpha})^{\frac{p}{\alpha}}e^{\frac{\alpha}{12p}}\right)^{1/p}\\ =& \sup_{p\geq 2} p^{-\frac{1}{\alpha}}\left(\sqrt{\frac{2\pi}{\alpha}}\right)^{1/p} p^{1/2p}(\frac{p}{\alpha})^{1/\alpha}e^{\alpha/12p^2-1/\alpha}\\ \leq& \sup_{p\geq 2}\left(\sqrt{\frac{2\pi}{\alpha}}\right)^{1/p} e^{1/2e}\frac{1}{(e\alpha)^{1/\alpha}}e^{\alpha/12p^2}\\ \leq& \left(\sqrt{\frac{2\pi}{\alpha}}\right)^{1/2} e^{1/2e}\frac{1}{(e\alpha)^{1/\alpha}}e^{\alpha/48},\end{aligned}$$

where the first inequality uses the Stirling formula and the second inequality uses the fact that $p^{1/p} \leq e^{1/e}$. These results are consistent with a plain intuition: a heavier-tailed distribution, i.e., smaller α, will result in a worse upper bound.

Remark 2.6. The assumption $|g - g_i| \leq H_i(X_i, X_i')$ is mild. It is a Lipschitz-type condition if we consider that $H_i(X_i, X_i')$ is a metric function. We will give an application of this condition to ℓ_1-regularized loss functions in Section 2.5. Here, we show that this condition also holds for ℓ_2-regularized loss functions. Let's consider that the loss function is $f(x, y) = (h(x) - y)^2 + \|w\|_2^2$, the function $h : \mathcal{X} \to [0, 1]$ is a Lipschitz function with Lipschitz constant L, i.e., $h(x) - h(y) \leq Ld(x, y)$, and $(\mathcal{Z}, d_2)$ is the metric space where $\mathcal{Z} = \mathcal{X} \times [0, 1]$ and $d_2((x, y), (x', y')) = (d(x, x')^2 + |y - y'|^2)^{1/2}$. Let us take $\psi[0, 1]^2 \to \mathbb{R}$ to be $\psi(h, y) = (h - y)^2$, which has the property $\max_{(h,y)\in[0,1]^2} \|\nabla\psi(h, y)\|_2 = 2^{3/2}$. It follows that

$$\begin{aligned}&|f(x, y) - f(x', y')|\\ =& |(h(x) - y)^2 - (h(x') - y')^2|\\ \leq& 2^{3/2}((h(x) - h(x'))^2 + (y - y')^2)^{1/2}\\ \leq& 2^{3/2}(L^2 d(x, x')^2 + (y - y')^2)^{1/2}\\ \leq& 2^{3/2}\max\{1, L\}d_2((x, y), (x', y')).\end{aligned}$$

Matching f to g and $2^{3/2}\max\{1, L\}d_2((x, y), (x', y'))$ to $H_i(X_i, X_i')$), we can conclude that the condition $|g - g_i| \leq H_i(X_i, X_i')$ holds for ℓ_2-regularized loss functions.

2.3. Algorithmic Stability with Unbounded Losses

In this section, the metric probability space $(\mathcal{Z}_i, d_i, \mu_i)$ will have the structure $\mathcal{Z}_i = \mathcal{X}_i \times \mathcal{Y}_i$ where $\mathcal{X}_i$ and $\mathcal{Y}_i$ are the instance and label space of the i-th example, respectively. Under the i.i.d assumption, the $(\mathcal{Z}_i, d_i, \mu_i)$ are identical for all $i \in \mathbb{N}$, and so we will henceforth drop the subscript i from these.

Assume we are given a training dataset $S = Z_1^n \sim \mu^n$, a learning algorithm $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{Y}^{\mathcal{X}}$ maps S to a function mapping the instance space $\mathcal{X}$ into the label space $\mathcal{Y}$. The output of the learning algorithm based on the sample S will be denoted by $\mathcal{A}_S$. Following the previous literature, we assume that $\mathcal{A}$ is symmetric, which means that $\mathcal{A}$ is invariant under permutations of S. The quality of the function returned by the algorithm is measured using a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$.

The empirical risk $R_n(\mathcal{A}, S)$ is typically defined as

$$R_n(\mathcal{A}, S) = \frac{1}{n}\sum_{i=1}^n \ell(\mathcal{A}_S, Z_i)$$

and the population risk $R(\mathcal{A}, S)$ as

$$R(\mathcal{A}, S) = \mathbb{E}_{Z\sim\mu}[\ell(\mathcal{A}_S, Z)].$$

A large body of work has been dedicated to obtaining high probability generalization bounds, i.e., giving bounds on the error $R(\mathcal{A}, S) - R_n(\mathcal{A}, S)$ with probabilities.

The widely used notion of stability allowing high probability upper bounds is called uniform stability. We mention a variant of uniform stability provided in (Rakhlin et al., 2005), which is slightly more general than the original notion in (Bousquet & Elisseeff, 2002).

Definition 2.7. (Rakhlin et al., 2005) The algorithm $\mathcal{A}$ is said to be γ-uniform stable if for any $\bar{z} \in \mathcal{Z}$, the function $f : \mathcal{Z}^n \to \mathbb{R}$ given by $f(z) = \ell(\mathcal{A}_z, \bar{z})$ is γ-Lipschitz with respect to the Hamming metric on $\mathcal{Z}^n$:

$$\forall z, z' \in \mathcal{Z}^n, \forall \bar{z} \in \mathcal{Z} : |f(z) - f(z')| \leq \gamma \sum_{i=1}^{n} \mathbb{I}_{\{z_i \neq z'_i\}}.$$

Most previous work based on the uniform stability required the loss to be bounded by some constant $M < \infty$. We make no such restriction in this paper. To relax the boundedness condition, we use a different notion of stability proposed in (Kontorovich, 2014).

Definition 2.8. (Kontorovich, 2014) The algorithm $\mathcal{A}$ is said to be γ-totally Lipschitz stable if the function $f : \mathcal{Z}^{n+1} \to \mathbb{R}$ given by $f(z_1^{n+1}) = \ell(\mathcal{A}_{z_1^n}, z_{n+1})$ is γ-Lipschitz with respect to the ℓ_1 product metric on $\mathcal{Z}^{n+1}$:

$$\forall z, z' \in \mathcal{Z}^{n+1} : |f(z) - f(z')| \leq \gamma \sum_{i=1}^{n+1} d(z_i, z'_i).$$

Based on this stability, we first give an in-expectation generalization bound for stable algorithms.

Lemma 2.9. *Suppose $\mathcal{A}$ is a symmetric, γ-totally Lipschitz stable learning algorithm over the metric probability space $(\mathcal{Z}, d, \mu)$ with $\Delta_\alpha(\mathcal{Z}) < \infty$. Then*

$$\mathbb{E}[R(\mathcal{A}, S) - R_n(\mathcal{A}, S)] \leq c(\alpha)\gamma\Delta_\alpha(\mathcal{Z}),$$

where $c(\alpha) = (\log 2)^{1/\alpha}$ if $\alpha > 1$ and $c(\alpha) = 2\Gamma(\frac{1}{\alpha} + 1)$ if $0 < \alpha \leq 1$.

Remark 2.10. In this proof, the heavy tailedness of subweibull distributions hinders standard proof techniques, such as Jensen's inequality.

The next lemma discusses the Lipschitz continuity.

Lemma 2.11 (Lemma 2 in (Kontorovich, 2014))**.** *Suppose $\mathcal{A}$ is a symmetric, γ-totally Lipschitz stable learning algorithm and define the function $f : \mathcal{Z}^n \to \mathbb{R}$ by $f(z) = R(\mathcal{A}, z) - R_n(\mathcal{A}, z)$. Then f is 3γ-Lipschitz.*

Now, combining Lemma 2.11 with the probabilistic inequality in Theorem 2.2 and further together with the in-expectation generalization bound in Lemma 2.9 yields the following high probability generalization bound.

Theorem 2.12. *Suppose $\mathcal{A}$ is a symmetric, γ-totally Lipschitz stable learning algorithm over the metric probability space $(\mathcal{Z}, d, \mu)$ with $\Delta_\alpha(\mathcal{Z}) < \infty$. Then, for training samples $S \sim \mu^n$ and any $0 < \delta < 1/e^2$, with probability at least $1 - \delta$*

1.) if $0 < \alpha \leq 1$, let $c_\alpha = 2\sqrt{2}((\log 2)^{1/\alpha} + e^3\Gamma^{1/2}(\frac{2}{\alpha} + 1) + e^3 3^{\frac{2-\alpha}{3\alpha}} \sup_{p\geq 2} p^{\frac{-1}{\alpha}}\Gamma^{1/p}(\frac{p}{\alpha} + 1))$, we have

$$R(\mathcal{A}, S) - R_n(\mathcal{A}, S) \leq c(\alpha)\gamma\Delta_\alpha(\mathcal{Z}) + 3\gamma c_\alpha \left(\sqrt{n\log(\frac{1}{\delta})}\Delta_\alpha(\mathcal{Z}) + \log^{\frac{1}{\alpha}}(\frac{1}{\delta})\Delta_\alpha(\mathcal{Z})\right);$$

2.) if $\alpha > 1$, let $1/\alpha^* + 1/\alpha = 1$ and $c'_\alpha = 8e + 2(\log 2)^{1/\alpha}$, we have

$$R(\mathcal{A}, S) - R_n(\mathcal{A}, S) \leq c(\alpha)\gamma\Delta_\alpha(\mathcal{Z}) + 3\gamma c'_\alpha \left(\sqrt{n\log(\frac{1}{\delta})}\Delta_\alpha(\mathcal{Z}) + \log^{\frac{1}{\alpha}}(\frac{1}{\delta})n^{\frac{1}{\alpha^*}}\Delta_\alpha(\mathcal{Z})\right),$$

where $c(\alpha) = (\log 2)^{1/\alpha}$ if $\alpha > 1$ and $c(\alpha) = 2\Gamma(\frac{1}{\alpha} + 1)$ if $0 < \alpha \leq 1$.

Remark 2.13. The relationship between α and the generalization bound follows the analysis in Remark 2.5. Next, we compare Theorem 2.12 with relevant results. In the related work, the basic and the best known result is the high probability upper bound in (Bousquet & Elisseeff, 2002) which states that, with probability at least $1 - \delta$,

$$R(\mathcal{A}, S) - R_n(\mathcal{A}, S) \lesssim \left(\gamma\sqrt{n} + \frac{M}{\sqrt{n}}\right)\sqrt{\log(\frac{1}{\delta})}, \quad (1)$$

where γ denotes the uniform stability and M is the upper bound of the loss ℓ. Kontorovich (2014) extends this bound to the unbounded loss with subgaussian diameter, and their Theorem 2 states that, with probability at least $1 - \delta$,

$$R(\mathcal{A}, S) - R_n(\mathcal{A}, S) \lesssim \gamma^2\Delta_2^2(\mathcal{Z}) + \gamma\Delta_2(\mathcal{Z})\sqrt{n\log(\frac{1}{\delta})},$$

where, in this case, γ denotes the totally Lipschitz stability and $\Delta_2(\mathcal{Z})$ is the subgaussian diameter. If we instead consider the subexponential distribution, the generalization bound in (Maurer & Pontil, 2021) is with probability at least $1 - \delta$,

$$R(\mathcal{A}, S) - R_n(\mathcal{A}, S) \lesssim \gamma\Delta_1(\mathcal{Z}) + \gamma\Delta_1(\mathcal{Z})\sqrt{n\log(\frac{1}{\delta})} + \gamma\Delta_1(\mathcal{Z})\log(\frac{1}{\delta}).$$

As shown in the above three bounds and related results on algorithmic stability, the stability γ is required at the least of the order $1/\sqrt{n}$ for nontrivial convergence decay. By comparison to the relevant bounds in (Bousquet & Elisseeff, 2002; Kontorovich, 2014; Maurer & Pontil, 2021), our generalization bound in Theorem 2.12 give results for unbounded loss functions with subweibull diameter: (1) if $0 < \alpha \leq 1$,

$$R(\mathcal{A}, S) - R_n(\mathcal{A}, S) \lesssim \gamma\Delta_\alpha(\mathcal{Z}) + \gamma\Delta_\alpha(\mathcal{Z})\left(\sqrt{n\log(\frac{1}{\delta})} + \log^{\frac{1}{\alpha}}(\frac{1}{\delta})\right); \quad (2)$$

(2) if $\alpha > 1$, let $1/\alpha^* + 1/\alpha = 1$

$$R(\mathcal{A}, S) - R_n(\mathcal{A}, S) \lesssim \gamma\Delta_\alpha(\mathcal{Z}) + \gamma\Delta_\alpha(\mathcal{Z})\left(\sqrt{n\log(\frac{1}{\delta})} + \log^{\frac{1}{\alpha}}(\frac{1}{\delta})n^{\frac{1}{\alpha^*}}\right), \quad (3)$$

which includes the results of Kontorovich (2014); Maurer & Pontil (2021) as specific cases and substantially extends the existing results to a large broad class of unbounded losses.

Remark 2.14. We can derive numerical evaluations of $\Delta_\alpha(\mathcal{Z})$. Recall that the subweibull diameter $\Delta_\alpha(\mathcal{Z})$ of the metric probability space $(\mathcal{Z}, d, \mu)$ is defined as

$$\Delta_\alpha(\mathcal{Z}) = \|d(Z, Z')\|_{\psi_\alpha},$$

where $Z, Z' \sim \mu$ are independent. If the distribution μ is known through sampling, we can derive the upper bound of $\Delta_\alpha(\mathcal{Z})$. We take the following setting as an example to show the proof roadmap: $\mathcal{Z} = \mathbb{R}$, metric $d(Z, Z') = |Z - Z'|$ and μ is the standard Gaussian probability measure $d\mu = (2\pi)^{-\frac{1}{2}}e^{-\frac{x^2}{2}}dx$. In this case, we have $\Delta_\alpha(\mathcal{Z}) \leq 2\|Z\|_{\psi_\alpha}$. Since

$$\|Z\|_{\psi_\alpha} = \inf\left\{\sigma \in (0, \infty) : \mathbb{E}\exp\left(\left(\frac{|Z|}{\sigma}\right)^\alpha\right) \leq 2\right\},$$

when we know the distribution of Z, the upper bound of $\|Z\|_{\psi_\alpha}$ can be derived exactly. For instance, if $\alpha = 2$ and $Z \sim N(0, 1)$, we get $\|Z\|_p = \sqrt{2}(\frac{\Gamma((1+p)/2)}{\Gamma(1/2)})^{1/p}$ for each $p \geq 1$, which implies that $\|Z\|_p \leq 3\sqrt{p}$ since $\Gamma(x) \leq 3x^x$ for all $x \geq 1/2$. Further, we get $\mathbb{E}[Z^{2p}] \leq (18p)^p$. Since Stirling's approximation yields $p! \geq (p/e)^p$, and recalling the Taylor series expansion of the exponential function, we have

$$\mathbb{E}\exp(\lambda^2 Z^2) = \mathbb{E}\left[1 + \sum_{p=1}^{\infty}\frac{(\lambda^2 Z^2)^p}{p!}\right]$$
$$= 1 + \sum_{p=1}^{\infty}\frac{\lambda^{2p}\mathbb{E}[Z^{2p}]}{p!} \leq 1 + \sum_{p=1}^{\infty}\frac{(18\lambda^2 p)^p}{(p/e)^p}$$
$$= \sum_{p=0}^{\infty}(18e\lambda^2)^p = \frac{1}{1 - 18e\lambda^2},$$

which provided that when $18e\lambda^2 < 1$, the geometric series above converges. To bound this quantity further, we can use the numeric inequality $1/(1 - x) \leq e^{2x}$ which is valid for $x \in [0, 1/2]$. It follows that

$$\mathbb{E}\exp(\lambda^2 Z^2) \leq \exp(36e\lambda^2),$$

for all λ satisfying $|\lambda| \leq \frac{1}{6\sqrt{e}}$. Setting $\lambda^2 = \frac{1}{64e}$ gives $\mathbb{E}\exp(Z^2/64e) \leq 2$. According to the above definition of $\|Z\|_{\psi_\alpha}$, we have $\|Z\|_{\psi_\alpha} \leq 8\sqrt{e}$, which means that $\Delta_\alpha(\mathcal{Z}) \leq 16\sqrt{e}$ in this case.

2.4. Sharper Bounds for Algorithmic Stability with Unbounded Losses

Recently, via a novel sample-splitting argument, Bousquet et al. (2020) provided a general moment inequality for weakly correlated random variables, refer to their Theorem 4. This moment inequality is equipped to establish the following moment bound of γ-uniformly stable algorithms for all $p \geq 2$:

$$\|R(\mathcal{A}, S) - R_n(\mathcal{A}, S)\|_p \lesssim p\gamma\log n + M\sqrt{\frac{p}{n}},$$

where M is the upper bound of the loss ℓ. According to the equivalence between tails and moments, this generalization bound implies that for any $\delta \in (0, 1)$, the following deviation bound holds with probability at least $1 - \delta$ over the draw of S

$$|R(\mathcal{A}, S) - R_n(\mathcal{A}, S)| \lesssim \gamma\log n\log(\frac{1}{\delta}) + M\sqrt{\frac{\log(\frac{1}{\delta})}{n}}. \quad (4)$$

This probabilistic result substantially improves the classical result (1) of Bousquet & Elisseeff (2002) by enhancing $\sqrt{n\log(\frac{1}{\delta})}$ to $\log n\log(\frac{1}{\delta})$. Up to logarithmic factors on sample size and tail bounds, the rate in (4) is nearly optimal in the sense of a lower bound by Bousquet et al. (2020). While powerful, this bound requires the loss functions to be bounded, which may largely narrow its applications.

In this section, we provide sharper generalization bounds than the results of Section 2.3, which is also an unbounded analogue of Bousquet et al. (2020). The following lemma establishes a moment inequality for a summation of weakly-dependent and unbounded random variables.

Lemma 2.15. *Let $Z = \{Z_1, ..., Z_n\}$ be a set of independent random variables each taking values in $\mathcal{Z}$. Define $Z\backslash\{Z_i\}$ be set $\{Z_1, ..., Z_{i-1}, Z_{i+1}, ..., Z_n\}$ and $Z^i = \{Z_1, ..., Z_{i-1}, Z_i', Z_{i+1}, ..., Z_n\}$, where $(Z_1', ..., Z_n')$ is an independent copy of $(Z_1, ..., Z_n)$. Let $g_1, ..., g_n$ be some functions $g_i : \mathcal{Z}^n \to \mathbb{R}$ such that the following inequalities hold for any $i \in \{1, ..., n\}$,*

- $\left\|\mathbb{E}_{Z\setminus\{Z_i\}}[g_i(Z)]\right\|_{\psi_\alpha} \leq M$ *almost surely (a.s.),*
- $\mathbb{E}_{Z_i}[g_i(Z)] = 0$ *a.s.,*
- *for any* $j \in [n]$ *with* $j \neq i$

$$\left|g_i(Z) - g_i(Z^j)\right| \leq H_j(Z_j, Z_j'),$$
$$\text{and} \quad \|H_j(Z_j, Z_j')\|_{\psi_\alpha} \leq \beta.$$

Then, for any $p \geq 2$

1.) if $0 < \alpha \leq 1$, *let* $c_\alpha = 2\sqrt{2}((\log 2)^{1/\alpha} + e^3\Gamma^{1/2}(\frac{2}{\alpha} + 1) + e^3 3^{\frac{2-\alpha}{3\alpha}} \sup_{p\geq 2} p^{\frac{-1}{\alpha}}\Gamma^{1/p}(\frac{p}{\alpha}+1))$, *we have*

$$\Big\|\sum_{i=1}^n g_i(Z)\Big\|_p \leq c_\alpha\left(\sqrt{pn}M + p^{1/\alpha}M\right) + 3\sqrt{2}c_\alpha\left(2np\beta\lceil\log_2 n\rceil + 2np^{\frac{1}{\alpha}+\frac{1}{2}}\beta\right);$$

2.) if $\alpha > 1$, *let* $1/\alpha^* + 1/\alpha = 1$ *and* $c_\alpha' = 8e + 2(\log 2)^{1/\alpha}$, *we have*

$$\Big\|\sum_{i=1}^n g_i(Z)\Big\|_p \leq c_\alpha'\left(\sqrt{pn}M + p^{1/\alpha}n^{\frac{1}{\alpha^*}}M\right) + 3\sqrt{2}c_\alpha'\left(2np\beta\lceil\log_2 n\rceil + 3(2n)^{(\frac{1}{\alpha^*}+\frac{1}{2})}p^{\frac{1}{\alpha}+\frac{1}{2}}\beta\right).$$

Remark 2.16. This lemma extends the moment inequality of Theorem 4 of Bousquet et al. (2020) from the bounded random variable to the subweibull variable. Lemma 2.15 is proved by our proposed moment inequality in Theorem 2.1. We emphasize here that the moment inequality we proposed for unbounded random variables is crucial for providing sharper bounds. In contrast, the probabilistic inequalities proved in (Kontorovich, 2014; Maurer & Pontil, 2021) fail to give Lemma 2.15 as they are not moment inequality. Of course, the technique of Kontorovich (2014); Maurer & Pontil (2021) fails to give a sharper generalization bound.

As an important consequence of this lemma, we can give the following sharper generalization bounds for algorithmic stability with unbounded losses.

Theorem 2.17. *Suppose* $\mathcal{A}$ *is a symmetric,* γ*-totally Lipschitz stable learning algorithm over the metric probability space* $(\mathcal{Z}, d, \mu)$ *with* $\Delta_\alpha(\mathcal{Z}) < \infty$. *Then, for training samples* $S \sim \mu^n$ *and any* $0 < \delta < 1/e^2$, *with probability at least* $1-\delta$

1.) if $0 < \alpha \leq 1$, *let* $c_\alpha = 2\sqrt{2}((\log 2)^{1/\alpha} + e^3\Gamma^{1/2}(\frac{2}{\alpha} + 1) + e^3 3^{\frac{2-\alpha}{3\alpha}} \sup_{p\geq 2} p^{\frac{-1}{\alpha}}\Gamma^{1/p}(\frac{p}{\alpha}+1))$, *we have*

$$\begin{aligned}
&|R(\mathcal{A},S) - R_n(\mathcal{A},S)| \\
\leq& 6\sqrt{2}c_\alpha\gamma\Delta_\alpha(\mathcal{Z})\left(2\lceil\log_2 n\rceil\log(\frac{1}{\delta}) + 2\log^{\frac{1}{\alpha}+\frac{1}{2}}(\frac{1}{\delta})\right) \\
&+ \frac{2c_\alpha}{n}\gamma\Delta_\alpha(\mathcal{Z})\left(\sqrt{n}\log^{\frac{1}{2}}(\frac{1}{\delta}) + \log^{1/\alpha}(\frac{1}{\delta})\right) \\
&+ 2\gamma\Delta_\alpha(\mathcal{Z})(2\Gamma(\log(\frac{1}{\delta})\frac{1}{\alpha}+1))^{\frac{1}{\log(\frac{1}{\delta})}};
\end{aligned}$$

2.) if $\alpha > 1$, *let* $1/\alpha^* + 1/\alpha = 1$ *and* $c_\alpha' = 8e + 2(\log 2)^{1/\alpha}$, *we have*

$$\begin{aligned}
&|R(\mathcal{A},S) - R_n(\mathcal{A},S)| \leq 6\sqrt{2}c_\alpha'\gamma\Delta_\alpha(\mathcal{Z})\times \\
&\left(2\lceil\log_2 n\rceil\log(\frac{1}{\delta}) + 3(2)^{(\frac{1}{\alpha^*}+\frac{1}{2})}\frac{1}{n^{\frac{1}{2}-\frac{1}{\alpha^*}}}\log^{\frac{1}{\alpha}+\frac{1}{2}}(\frac{1}{\delta})\right) \\
&+ \frac{2c_\alpha'}{n}\gamma\Delta_\alpha(\mathcal{Z})\left(\sqrt{n}\log^{\frac{1}{2}}(\frac{1}{\delta}) + \log^{1/\alpha}(\frac{1}{\delta})n^{\frac{1}{\alpha^*}}\right) \\
&+ 2\gamma\Delta_\alpha(\mathcal{Z})(2\Gamma(\log(\frac{1}{\delta})\frac{1}{\alpha}+1))^{\frac{1}{\log(\frac{1}{\delta})}}.
\end{aligned}$$

Remark 2.18. Theorem 2.17 implies the following inequalities (1) if $0 < \alpha \leq 1$,

$$\begin{aligned}
&|R(\mathcal{A},S) - R_n(\mathcal{A},S)| \\
\lesssim& \gamma\Delta_\alpha(\mathcal{Z})\Big(\log n\log(\frac{1}{\delta}) + \log^{\frac{1}{\alpha}+\frac{1}{2}}(\frac{1}{\delta})\Big) \\
&+ \frac{\gamma\Delta_\alpha(\mathcal{Z})}{n}\Big(\sqrt{n}\log^{\frac{1}{2}}(\frac{1}{\delta}) + \log^{\frac{1}{\alpha}}(\frac{1}{\delta})\Big);
\end{aligned}$$

(2) if $\alpha > 1$, let $1/\alpha^* + 1/\alpha = 1$

$$\begin{aligned}
&|R(\mathcal{A},S) - R_n(\mathcal{A},S)| \\
\lesssim& \gamma\Delta_\alpha(\mathcal{Z})\Big(\log n\log(\frac{1}{\delta}) + \frac{1}{n^{\frac{1}{2}-\frac{1}{\alpha^*}}}\log^{\frac{1}{\alpha}+\frac{1}{2}}(\frac{1}{\delta})\Big) \\
&+ \frac{\gamma\Delta_\alpha(\mathcal{Z})}{n}\Big(\sqrt{n}\log^{\frac{1}{2}}(\frac{1}{\delta}) + \log^{1/\alpha}(\frac{1}{\delta})n^{\frac{1}{\alpha^*}}\Big).
\end{aligned}$$

To compare with the generalization bound in Eqs. (2) and (3), our bound in Theorem 2.17 substantially improves the dominated term from $\gamma\Delta_\alpha(\mathcal{Z})\sqrt{n}$ to $\gamma\Delta_\alpha(\mathcal{Z})$, which successfully gives sharper generalization bounds. By comparison to the relevant bounds in (Kontorovich, 2014; Maurer & Pontil, 2021), our bound in Theorem 2.17 also improve their results in the case $\alpha = 2$ and $\alpha = 1$, respectively, which can be $\sqrt{n}$-times faster than their results.

Remark 2.19. Based on the technique of Bousquet et al. (2020); Klochkov & Zhivotovskiy (2021), many recent works develop sharper generalization bounds for alternative settings, see (Lei & Ying, 2020; Lei et al., 2021; Yuan & Li, 2023), to mention but a few. Among them, Yuan & Li (2023) provide in-expectation generalization bounds in the sense of Bousquet et al. (2020); Klochkov & Zhivotovskiy (2021) for ℓ_q stability. Their results also allow sharper high

probability generalization bounds for unbounded losses up to subexponential distributions (i.e., $\alpha = 1$), see the analysis below their Theorem 2. As a comparison, our generalization bounds allow heavy-tailed distributions, i.e., $0 < \alpha < 1$.

2.5. Application to Regularized Nearest-neighbor Regression

We first give some necessary notations of the regularized regression. We assume the label space $\mathcal{Y}$ to be all of $\mathbb{R}$. A simple no-free-lunch argument shows that it is impossible to learn functions with arbitrary oscillation, and hence Lipschitzness is a natural and commonly used regularization constraint (Shalev-Shwartz & Ben-David, 2014; Tsybakov, 2003; Wasserman, 2006). We will denote by $\mathcal{F}_\lambda$ the collection of all λ-Lipschitz functions $f : \mathcal{X} \to \mathbb{R}$. The learning algorithm $\mathcal{A}$ maps the sample $S = Z_{i=1}^n$, with $Z_i = (X_i, Y_i) \in \mathcal{X} \times \mathbb{R}$, to the function $\hat{f} \in \mathcal{F}_\lambda$ by minimizing the empirical risk

$$\hat{f} = \arg\min_{f \in \mathcal{F}_\lambda} \frac{1}{n} \sum_{i=1}^n |f(X_i) - Y_i|$$

over all $f \in \mathcal{F}_\lambda$, where we have chosen the absolute loss $\ell(y, y') = |y - y'|$. In the general metric space, Gottlieb et al. (2017) proposed an efficient algorithm for regression via Lipschitz extension, a method that can be traced back to the seminal work (von Luxburg & Bousquet, 2004), which is algorithmically realized by 1-nearest neighbors. This approach very facilitates generalization analysis. For any metric space $(\mathcal{X}, d)$, we associate it to a metric space $(\mathcal{Z}, \bar{d})$, where $\mathcal{Z} = \mathcal{X} \times \mathbb{R}$ and $\bar{d}((x, y), (x', y')) = d(x, x') + |y - y'|$, and we suppose that $(\mathcal{Z}, \bar{d})$ is endowed with a measure μ such that $\Delta_\alpha(\mathcal{Z}) = \Delta_\alpha(\mathcal{Z}, \bar{d}, \mu) < \infty$.

We follow the analysis of Kontorovich (2014) on the stability of 1-NN regression regularized by Lipschitz continuity λ. If none of the $n + 1$ points (n sample and 1 test) is too isolated from the rest, Kontorovich (2014) shows that the regression algorithm is $\gamma = O(\lambda/n)$-totally Lipschitz stable. In the case of subgaussian distribution, with probability $1 - n\exp(-\Omega(n))$, each of the $n+1$ points is within distance $O(\Delta_2(\mathcal{Z}))$ of another point. Hence, Kontorovich (2014) states that, with probability at least $1 - n\exp(-\Omega(n)) - \delta$

$$\begin{aligned} & R(\mathcal{A}, S) - R_n(\mathcal{A}, S) \\ \lesssim & \left(\frac{\lambda}{n}\Delta_2(\mathcal{Z})\right)^2 + \frac{\lambda}{\sqrt{n}}\Delta_2(\mathcal{Z})\sqrt{\log(\frac{1}{\delta})}. \end{aligned}$$

While in the case of subweibull distribution, according to Theorem 2.1 in (Vladimirova et al., 2020), with probability $1 - n\exp(-\Omega(n^\alpha))$, each of the $n + 1$ points is within distance $O(\Delta_\alpha(\mathcal{Z}))$ of another point. Thus, by Theorem 2.17, our bound is, with probability at least $1 - n\exp(-\Omega(n^\alpha)) - \delta$, (1) if $0 < \alpha \leq 1$,

$$\begin{aligned} & |R(\mathcal{A}, S) - R_n(\mathcal{A}, S)| \\ \lesssim & \frac{\lambda}{n}\Delta_\alpha(\mathcal{Z})\left(\log n \log(\frac{1}{\delta}) + \log^{\frac{1}{\alpha}+\frac{1}{2}}(\frac{1}{\delta})\right) \\ & + \frac{\lambda\Delta_\alpha(\mathcal{Z})}{n^2}\left(\sqrt{n}\log^{\frac{1}{2}}(\frac{1}{\delta}) + \log^{\frac{1}{\alpha}}(\frac{1}{\delta})\right); \end{aligned}$$

(2) if $\alpha > 1$, let $1/\alpha^* + 1/\alpha = 1$

$$\begin{aligned} & |R(\mathcal{A}, S) - R_n(\mathcal{A}, S)| \\ \lesssim & \frac{\lambda}{n}\Delta_\alpha(\mathcal{Z})\left(\log n \log(\frac{1}{\delta}) + \frac{1}{n^{\frac{1}{2}-\frac{1}{\alpha^*}}}\log^{\frac{1}{\alpha}+\frac{1}{2}}(\frac{1}{\delta})\right) \\ & + \frac{\lambda\Delta_\alpha(\mathcal{Z})}{n^2}\left(\sqrt{n}\log^{\frac{1}{2}}(\frac{1}{\delta}) + \log^{1/\alpha}(\frac{1}{\delta})n^{\frac{1}{\alpha^*}}\right). \end{aligned}$$

As a comparison, our results allow a substantial extension of existing generalization bounds to heavy-tailed distributions and improve the bound from the order $O(1/\sqrt{n})$ in (Kontorovich, 2014) to a sharper order $O(1/n)$. Our bound reveals that the generalization bound of regularized nearest-neighbor regression enjoys a faster rate.

2.6. Application to Rademacher Complexity

In this section, we show the strength of our probabilistic inequality, which, other than algorithmic stability, is also capable of giving generalization bounds for Rademacher complexity with unbounded losses. Rademacher complexity is also a popular tool of learning theory to reason about the generalization.

Suppose that $\mathcal{F}$ is a class of function $f : \mathcal{X} \to \mathbb{R}$, the Rademacher complexity of $\mathcal{F}$ is defined as $\mathcal{R}(\mathcal{F}) = \mathbb{E}\left[\frac{1}{n}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^n \epsilon_i f(X_i)|X\right]\right]$, where $\epsilon_1, ..., \epsilon_n$ are independent Rademacher variables. Using McDiarmid's exponential inequality, Theorem 8 in (Bartlett & Mendelson, 2002) establishes a high probability generalization error bound for Rademacher complexity, however, this approach requires the function f to be bounded, induced by McDiarmid's exponential inequality. In this section, we show that the boundedness can be relaxed by unbounded sub-weibull distributions for uniformly Lipschitz function classes. Indeed, Theorem 2.20 also holds for the metric space. We study the Banach space as it simplifies the proof.

Theorem 2.20. *Let $X = (X_1, ..., X_n)$ be a vector of independent subweibull random variables with values in a Banach space $(\mathcal{X}, \|\cdot\|)$ and let $\mathcal{F}$ be a class of function $f : \mathcal{X} \to \mathbb{R}$ such that $f(x) - f(y) \leq L\|x - y\|$ for all $f \in \mathcal{F}$ and $x, y \in \mathcal{X}$. Then, for any $0 < \delta < 1/e^2$, with probability at least $1 - \delta$*

1.) if $0 < \alpha \leq 1$, let $c_\alpha = 2\sqrt{2}((\log 2)^{1/\alpha} + e^3\Gamma^{1/2}(\frac{2}{\alpha} +$

$1) + e^3 3^{\frac{2-\alpha}{3\alpha}} \sup_{p\geq 2} p^{\frac{-1}{\alpha}} \Gamma^{1/p}(\frac{p}{\alpha}+1))$, *we have*

$$\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X_i')] \leq 2\mathcal{R}(\mathcal{F}) + \frac{4Lc_\alpha}{n}\Big(\sqrt{\log(\frac{1}{\delta})}\Big(\sum_{i=1}^{n} |||X_i|||_{\psi_\alpha}^2\Big)^{\frac{1}{2}} + \log^{1/\alpha}(\frac{1}{\delta}) \max_{1\leq i\leq n} |||X_i|||_{\psi_\alpha}\Big);$$

2.) if $\alpha \geq 1$, *let* $1/\alpha^* + 1/\alpha = 1$, $c_\alpha' = 8e + 2(\log 2)^{1/\alpha}$ *and* $(|||X|||_{\psi_\alpha}) = (|||X_1|||_{\psi_\alpha}, ..., |||X_n|||_{\psi_\alpha})$, *we have*

$$\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X_i')] \leq 2\mathcal{R}(\mathcal{F}) + \frac{4Lc_\alpha'}{n}\Big(\sqrt{\log(\frac{1}{\delta})}\Big(\sum_{i=1}^{n} |||X_i|||_{\psi_\alpha}^2\Big)^{\frac{1}{2}} + \log^{1/\alpha}(\frac{1}{\delta})\|(|||X|||_{\psi_\alpha})\|_{\alpha^*}\Big).$$

3. Conclusions

In this paper, we provided generalization bounds for algorithmic stability with unbounded losses. The technical contribution is general concentration inequalities for subweibull random variables. In future work, it would be important to show that some other common learning algorithms, such as stochastic gradient descent, are also stable in the notion of totally Lipschitz stability.

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; and the Unicom Innovation Ecological Cooperation Plan.

Impact Statement

This work studies the theory of algorithmic stability, which does not present any foreseeable societal consequence.

References

Agarwal, S. and Niyogi, P. Generalization bounds for ranking algorithms via algorithmic stability. *Journal of Machine Learning Research*, 10(2), 2009.

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.

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.

Bassily, R., Feldman, V., Guzmán, C., and Talwar, K. Stability of stochastic gradient descent on nonsmooth convex losses. *Advances in Neural Information Processing Systems*, 33:4381–4391, 2020.

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

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

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. PMLR, 2020.

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.

De la Pena, V. and Giné, E. *Decoupling: from dependence to independence*. Springer Science & Business Media, 2012.

Devroye, L. and Wagner, T. Distribution-free inequalities for the deleted and holdout error estimates. *IEEE Transactions on Information Theory*, 25(2):202–207, 1979a.

Devroye, L. and Wagner, T. Distribution-free performance bounds for potential function rules. *IEEE Transactions on Information Theory*, 25(5):601–604, 1979b.

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.

El-Yaniv, R. and Pechyony, D. Stable transductive learning. In *Learning Theory: 19th Annual Conference on Learning Theory, COLT 2006, Pittsburgh, PA, USA, June 22-25, 2006. Proceedings 19*, pp. 35–49. Springer, 2006.

Feldman, V. and Vondrak, J. Generalization bounds for uniformly stable algorithms. *Advances in Neural Information Processing Systems*, 31, 2018.

Feldman, V. and Vondrak, J. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In *Conference on Learning Theory*, pp. 1270–1279. PMLR, 2019.

Foster, D. J., Greenberg, S., Kale, S., Luo, H., Mohri, M., and Sridharan, K. Hypothesis set stability and generalization. *Advances in Neural Information Processing Systems*, 32, 2019.

Gonen, A. and Shalev-Shwartz, S. Average stability is invariant to data preconditioning: Implications to exp-concave empirical risk minimization. *The Journal of Machine Learning Research*, 18(1):8245–8257, 2017.

Gottlieb, L.-A., Kontorovich, A., and Krauthgamer, R. Efficient regression in metric spaces via approximate lipschitz extension. *IEEE Transactions on Information Theory*, 63(8):4838–4849, 2017.

Haddouche, M., Guedj, B., Rivasplata, O., and Shawe-Taylor, J. Pac-bayes unleashed: Generalisation bounds with unbounded losses. *Entropy*, 23(10):1330, 2021.

Hardt, M., Recht, B., and Singer, Y. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pp. 1225–1234, 2016.

Hush, D., Scovel, C., and Steinwart, I. Stability of unstable learning algorithms. *Machine learning*, 67:197–206, 2007.

Kearns, M. and Ron, D. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. In *Proceedings of the tenth annual conference on Computational learning theory*, pp. 152–162, 1997.

Klochkov, Y. and Zhivotovskiy, N. Stability and deviation optimal risk bounds with convergence rate $o(1/n)$. *Advances in Neural Information Processing Systems*, 34: 5065–5076, 2021.

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

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.

Kutin, S. and Niyogi, P. Almost-everywhere algorithmic stability and generalization error. *arXiv preprint arXiv:1301.0579*, 2012.

Kuzborskij, I. and Lampert, C. Data-dependent stability of stochastic gradient descent. In *International Conference on Machine Learning*, pp. 2815–2824, 2018.

Latała, R. Estimation of moments of sums of independent real random variables. *The Annals of Probability*, 25(3): 1502–1513, 1997.

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

Lei, Y. and Ying, Y. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *International Conference on Machine Learning*, pp. 5809–5819, 2020.

Lei, Y., Liu, M., and Ying, Y. Generalization guarantee of sgd for pairwise learning. In *Advances in Neural Information Processing Systems*, 2021.

Liu, T., Lugosi, G., Neu, G., and Tao, D. Algorithmic stability and hypothesis complexity. In *International Conference on Machine Learning*, pp. 2159–2167, 2017.

Lugosi, G. and Pawlak, M. On the posterior-probability estimate of the error rate of nonparametric classification rules. *IEEE Transactions on Information Theory*, 40(2): 475–481, 1994.

Madden, L., Dall'Anese, E., and Becker, S. High-probability convergence bounds for non-convex stochastic gradient descent. *arXiv e-prints*, pp. arXiv–2006, 2020.

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

Maurer, A. A second-order look at stability and generalization. In *Conference on learning theory*, pp. 1461–1475, 2017.

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

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

Mukherjee, S., Niyogi, P., Poggio, T., and Rifkin, R. Statistical learning: Stability is necessary and sufficient for consistency of empirical risk minimization. *CBCL Paper*, 23, 2002.

Rakhlin, A., Mukherjee, S., and Poggio, T. Stability results in learning theory. *Analysis and Applications*, 3(04):397–417, 2005.

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

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

Rubinstein, B. I. and Simma, A. On the stability of empirical risk minimization in the presence of multiple risk minimizers. *IEEE transactions on information theory*, 58 (7):4160–4163, 2012.

Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Shalev-Shwartz, S., Shamir, O., Srebro, N., and Sridharan, K. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 11:2635–2670, 2010.

Tsybakov, A. B. *Introduction à l'estimation non paramétrique*, volume 41. Springer Science & Business Media, 2003.

Vapnik, V. and Chervonenkis, A. Theory of pattern recognition, 1974.

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.

von Luxburg, U. and Bousquet, O. Distance-based classification with lipschitz functions. *The Journal of Machine Learning Research*, 5(Jun):669–695, 2004.

Wasserman, L. *All of nonparametric statistics*. Springer Science & Business Media, 2006.

Yuan, X. and Li, P. Exponential generalization bounds with near-optimal rates for l_q-stable algorithms. In *International Conference on Learning Representations*, 2023.

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

A. Auxiliary Lemmas

Lemma A.1 (Theorem 1.3.1 in (De la Pena & Giné, 2012)). *Let $a_1, ..., a_n$ a finite non-random sequence, $\{\epsilon_i\}_{i=1}^n$ be a sequence of independent Rademacher variables and $1 < p < q < \infty$. Then,*

$$\left\| \sum_{i=1}^n \epsilon_i a_i \right\|_q \leq \left(\frac{q-1}{p-1} \right)^{1/2} \left\| \sum_{i=1}^n \epsilon_i a_i \right\|_p .$$

Lemma A.2 (Theorem 2 of (Latała, 1997)). *Let $X_1, ..., X_n$ be a sequence of independent symmetric random variables, and $p \geq 2$. Then,*

$$\frac{e-1}{2e^2} \|(X_i)\|_p \leq \|X_1 + ... + X_n\| \leq e\|(X_i)\|_p,$$

where $\|(X_i)\|_p := \inf\{t > 0 : \sum_{i=1}^n \log \psi_p(X_i/t) \leq p\}$ with $\psi_p(X) := \mathbb{E}|1 + X|^p$.

Lemma A.3 (Example 3.2 and 3.3 of (Latała, 1997)). *Assume X be a symmetric random variable satisfying $\mathbb{P}(|X| \geq t) = e^{-N(t)}$. For any $t \geq 0$, we have*

(a) If $N(t)$ is concave, then $\log \psi_p(e^{-2}tX) \leq pM_{p,X}(t) := \max\{(t^p\|X\|_p^p), (pt^2\|X\|_2^2)\}$.

(b) For convex $N(t)$, denote the convex conjugate function $N^*(t) := \sup_{s>0}\{ts - N(s)\}$ and

$$M_{p,X}(t) := \begin{cases} p^{-1}N^*(p|t|), & \text{if } p|t| \geq 2 \\ pt^2, & \text{if } p|t| < 2. \end{cases}$$

Then $\log \psi_p(tX/4) \leq pM_{p,X}(t)$.

Lemma A.4 (Marcinkiewicz-Zygmund's inequality (Ren & Liang, 2001)). *Let $X_1, ..., X_n$ be independent centered random variables with a finite p-th moment for $p \geq 2$. Then,*

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

B. Proofs of Section 2.2

In this section, we provide proofs of results in Section 2.2. To proceed, we state some technical lemmas.

Lemma B.1. *Let $h : \mathbb{R} \to \mathbb{R}$ be a convex functions, $\epsilon_1, ..., \epsilon_n$ a sequence of independent Rademacher variables and $a_1, ..., a_n$, $b_1, ..., b_n$ two sequences of nonnegative 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 of Lemma B.1. 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 $0 < s < t$

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

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 B.2. *Let $h : \mathbb{R} \to \mathbb{R}$ be a convex function and $g = f(X_1, ..., X_{i-1}, X_i, X_{i+1}, ..., X_n)$, 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 $g_i = f(X_1, ..., X_{i-1}, X_i', X_{i+1}, ..., X_n)$, where $(X_1', ..., X_n')$ is an independent copy of $(X_1, ..., X_n)$. Assume moreover that $|g - g_i| \leq H_i(X_i, X_i')$ for some functions $H_i : \mathcal{X}^2 \to \mathbb{R}$, $i = 1, ..., n$. Then,*

$$\mathbb{E}h(g - \mathbb{E}g) \leq \mathbb{E}h\left(\sum_{i=1}^{n} \epsilon_i H_i(X_i, X_i')\right),$$

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 of Lemma B.2. We will use induction with respect to n. For $n = 0$ the statement is obvious, since $\mathbb{E}h(g - \mathbb{E}g) = \mathbb{E}h\left(\sum_{i=1}^{n} \epsilon_i H_i(X_i, X_i')\right) = h(0)$. Assume that the lemma is true for $n - 1$, then

$$\begin{aligned}
&\mathbb{E}h(g - \mathbb{E}g) = \mathbb{E}h(g - \mathbb{E}_{X_n'} g_n + \mathbb{E}_{X_n} g - \mathbb{E}g) \\
&\leq \mathbb{E}h(g - g_n + \mathbb{E}_{X_n} g - \mathbb{E}g) \\
&= \mathbb{E}h(g_n - g + \mathbb{E}_{X_n} g - \mathbb{E}g) \\
&= \mathbb{E}h(\epsilon_n |g - g_n| + \mathbb{E}_{X_n} g - \mathbb{E}g) \\
&\leq \mathbb{E}h(\epsilon_n H_n(X_n, X_n') + \mathbb{E}_{X_n} g - \mathbb{E}g),
\end{aligned}$$

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

$$|Z - Z_i| = |\mathbb{E}_{X_n} g - \mathbb{E}_{X_n} g_i| \leq \mathbb{E}_{X_n} |g - g_i| \leq H_i(X_i, X_i'),$$

and thus for fixed X_n, X_n' and ϵ_n, we can apply the induction assumption to the function $t \to h(\epsilon_n H_n(X_n, X_n') + t)$ instead of h and $\mathbb{E}_{X_n} g$ instead of g, to obtain

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

The proof is complete. □

Next lemma provides a moment inequality for the sum of independent subweibull random variables. The proof follows the technique of (Zhang & Wei, 2022).

Lemma B.3. *Suppose $X_1, X_2, ..., X_n$ are independent subweibull random variables with mean zero. For any vector $a = (a_1, ..., a_n) \in \mathbb{R}^n$, let $b = (a_1 \|X_1\|_{\psi_\alpha}, ..., a_n \|X_n\|_{\psi_\alpha}) \in \mathbb{R}^n$. Then for $p \geq 2$,*

- *if $0 < \alpha \leq 1$, let $c_\alpha = 2\sqrt{2}\left((\log 2)^{1/\alpha} + e^3 \Gamma^{1/2}(\frac{2}{\alpha} + 1) + e^3 3^{\frac{2-\alpha}{3\alpha}} \sup_{p \geq 2} p^{\frac{-1}{\alpha}} \Gamma^{1/p}(\frac{p}{\alpha} + 1)\right)$,*

$$\left\|\sum_{i=1}^{n} a_i X_i\right\|_p \leq c_\alpha \left(\sqrt{p}\|b\|_2 + p^{1/\alpha}\|b\|_\infty\right);$$

- *if $\alpha > 1$, let $1/\alpha^* + 1/\alpha = 1$ and $c_\alpha' = 8e + 2(\log 2)^{1/\alpha}$,*

$$\left\|\sum_{i=1}^{n} a_i X_i\right\|_p \leq c_\alpha' \left(\sqrt{p}\|b\|_2 + p^{1/\alpha}\|b\|_{\alpha^*}\right).$$

Proof of Lemma B.3. Without loss of generality, we assume $\|X_i\|_{\psi_\alpha} = 1$. Define $Y_i = (|X_i| - (\log 2)^{1/\alpha})_+$, then it is easy to check that $\mathbb{P}(|X_i| \geq t) \leq 2e^{-t^\alpha}$, which also implies that $\mathbb{P}(Y_i \geq t) \leq e^{-t^\alpha}$. According to the symmetrization inequality, Proposition 6.3 of (Ledoux & Talagrand, 1991), we have

$$\|\sum_{i=1}^{n} a_i X_i\|_p \leq 2\|\sum_{i=1}^{n} \epsilon_i a_i X_i\|_p = 2\|\sum_{i=1}^{n} \epsilon_i a_i |X_i|\|_p,$$

where $\{\epsilon_i\}_{i=1}^n$ are independent Rademacher random variables and we have used that $\epsilon_i X_i$ and $\epsilon_i|X_i|$ are identically distributed. By triangle inequality,

$$2\|\sum_{i=1}^n \epsilon_i a_i |X_i|\|_p \le 2\|\sum_{i=1}^n \epsilon_i a_i (Y_i + (\log 2)^{1/\alpha})\|_p \le 2\|\sum_{i=1}^n \epsilon_i a_i Y_i\|_p + 2(\log 2)^{1/\alpha}\|\sum_{i=1}^n \epsilon_i a_i\|_p.$$

Next, we bound the second term of the above upper bound. By Khinchin-Kahane inequality, Lemma A.1, we have

$$\begin{aligned}\|\sum_{i=1}^n \epsilon_i a_i\|_p &\le \big(\frac{p-1}{2-1}\big)^{1/2}\|\sum_{i=1}^n \epsilon_i a_i\|_2 \le \sqrt{p}\|\sum_{i=1}^n \epsilon_i a_i\|_2 = \sqrt{p}(\mathbb{E}(\sum_{i=1}^n \epsilon_i a_i)^2)^{1/2} \\ &= \sqrt{p}\big(\mathbb{E}(\sum_{i=1}^n \epsilon_i^2 a_i^2 + 2\sum_{1\le i<j\le n} \epsilon_i\epsilon_j a_i a_j)\big)^{1/2} = \sqrt{p}\big(\sum_{i=1}^n a_i^2\big)^{1/2} = \sqrt{p}\|a\|_2.\end{aligned}$$

Let $\{Z_i\}_{i=1}^n$ be independent symmetric random variables satisfying $\mathbb{P}(|Z_i| \ge t) = \exp(-t^\alpha)$ for all $t \ge 0$, we have

$$\|\sum_{i=1}^n \epsilon_i a_i Y_i\|_p = \|\sum_{i=1}^n \epsilon_i a_i Z_i\|_p = \|\sum_{i=1}^n a_i Z_i\|_p,$$

since $\epsilon_i Z_i$ and Z_i have the same distribution due to symmetry. Combining the above inequalities together, we reach

$$\|\sum_{i=1}^n a_i X_i\|_p \le 2(\log 2)^{1/\alpha}\sqrt{p}\|a\|_2 + 2\|\sum_{i=1}^n a_i Z_i\|_p.$$

In the case of $0 < \alpha \le 1$, $N(t) = t^\alpha$ is concave. Then Lemma A.2 and Lemma A.3 (a) gives for $p \ge 2$

$$\begin{aligned}\|\sum_{i=1}^n a_i Z_i\|_p &\le e\inf\{t>0 : \sum_{i=1}^n \log\psi_p(e^{-2}(\frac{a_i e^2}{t})Z_i) \le p\} \le e\inf\{t>0 : \sum_{i=1}^n pM_{p,Z_i}(\frac{a_i e^2}{t}) \le p\} \\ &= e\inf\{t>0 : \sum_{i=1}^n (\max\{(\frac{a_i e^2}{t})^p\|Z_i\|_p^p, p(\frac{a_i e^2}{t})^2\|Z_i\|_2^2\}) \le p\} \\ &\le e\inf\{t>0 : \sum_{i=1}^n (\frac{a_i e^2}{t})^p\|Z_i\|_p^p + \sum_{i=1}^n p(\frac{a_i e^2}{t})^2\|Z_i\|_2^2 \le p\} \\ &\le e\inf\{t>0 : 2p\Gamma(\frac{p}{\alpha}+1)\frac{e^{2p}}{t^p}\|a\|_p^p \le p\} + e\inf\{t>0 : 2p^2\Gamma(\frac{2}{\alpha}+1)\frac{e^4}{t^2}\|a\|_2^2\}] \le p\},\end{aligned}$$

where we have used $\|Z_i\|_p^p = p\Gamma(\frac{p}{\alpha}+1)$ and set $p = 2$ sometimes. Thus,

$$\|\sum_{i=1}^n a_i Z_i\|_p \le \sqrt{2}e^3(\Gamma^{1/p}(\frac{p}{\alpha}+1)\|a\|_p + \sqrt{p}\Gamma^{1/2}(\frac{2}{\alpha}+1)\|a\|_2).$$

By homogeneity, we can assume that $\sqrt{p}\|a\|_2 + p^{1/\alpha}\|a\|_\infty = 1$. Then $\|a\|_2 \le p^{-1/2}$ and $\|a\|_\infty \le p^{-1/\alpha}$. Therefore, for $p \ge 2$,

$$\begin{aligned}\|a\|_p &\le (\sum_{i=1}^n |a_i|^2\|a\|_\infty^{p-2})^{1/p} \le (p^{-1-\frac{(p-2)}{\alpha}})^{1/p} = (p^{-p/\alpha}p^{(2-\alpha)/\alpha})^{1/p} \le 3^{\frac{2-\alpha}{3\alpha}}p^{-1/\alpha} \\ &= 3^{\frac{2-\alpha}{3\alpha}}p^{-1/\alpha}(\sqrt{p}\|a\|_2 + p^{1/\alpha}\|a\|_\infty),\end{aligned}$$

where we used $p^{1/p} \le 3^{1/3}$ for any $p \ge 2$, $p \in \mathbb{N}$. Therefore, for $p \ge 2$,

$$\begin{aligned}&\|\sum_{i=1}^n a_i X_i\|_p \le 2(\log 2)^{1/\alpha}\sqrt{p}\|a\|_2 + 2\sqrt{2}e^3(\Gamma^{1/p}(\frac{p}{\alpha}+1)\|a\|_p + \sqrt{p}\Gamma^{1/2}(\frac{2}{\alpha}+1)\|a\|_2) \\ &\le 2\sqrt{2}((\log 2)^{1/\alpha} + e^3\Gamma^{1/2}(\frac{2}{\alpha}+1) + e^3 3^{\frac{2-\alpha}{3\alpha}}p^{\frac{-1}{\alpha}}\Gamma^{1/p}(\frac{p}{\alpha}+1))\sqrt{p}\|a\|_2) + 2\sqrt{2}e^3 3^{\frac{2-\alpha}{3\alpha}}\Gamma^{1/p}(\frac{p}{\alpha}+1)\|a\|_\infty.\end{aligned}$$

Let $c_\alpha = 2\sqrt{2}((\log 2)^{1/\alpha} + e^3\Gamma^{1/2}(\frac{2}{\alpha}+1) + e^3 3^{\frac{2-\alpha}{3\alpha}} \sup_{p\geq 2} p^{\frac{-1}{\alpha}}\Gamma^{1/p}(\frac{p}{\alpha}+1))$, we have

$$\|\sum_{i=1}^n a_i X_i\|_p \leq c_\alpha(\sqrt{p}\|a\|_2 + p^{1/\alpha}\|a\|_\infty).$$

In the case of $\alpha > 1$, $N(t) = t^\alpha$ is convex with $N^*(t) = \alpha^{-\frac{1}{\alpha-1}}(1-\alpha^{-1})t^{\frac{\alpha}{\alpha-1}}$. Then Lemma A.2 and Lemma A.3 (b) gives for $p \geq 2$

$$\begin{aligned}
\|\sum_{i=1}^n a_i Z_i\|_p &\leq e\inf\{t>0 : \sum_{i=1}^n \log\psi_p(\frac{4a_i}{t}Z_i/4) \leq p\} + e\inf\{t>0 : \sum_{i=1}^n pM_{p,Z_i}(\frac{4a_i}{t}) \leq p\} \\
&\leq e\inf\{t>0 : \sum_{i=1}^n p^{-1}N^*(p|\frac{4a_i}{t}|) \leq 1\} + e\inf\{t>0 : \sum_{i=1}^n p(\frac{4a_i}{t})^2 \leq 1\} \\
&= 4e(\sqrt{p}\|a\|_2 + (p/\alpha)^{1/\alpha}(1-\alpha^{-1})^{1/\alpha^*}\|a\|_{\alpha^*}),
\end{aligned}$$

where α^* is mentioned in the statement. Therefore, for $p \geq 2$,

$$\|\sum_{i=1}^n a_i X_i\|_p \leq (8e + 2(\log 2)^{1/\alpha})\sqrt{p}\|a\|_2 + 8e(1/\alpha)^{1/\alpha}(1-\alpha^{-1})^{1/\alpha^*}p^{1/\alpha}\|a\|_{\alpha^*}.$$

Since $8e + 2(\log 2)^{1/\alpha} \geq 8e(1/\alpha)^{1/\alpha}(1-\alpha^{-1})^{1/\alpha^*}$, let $c'_\alpha = 8e + 2(\log 2)^{1/\alpha}$, we have

$$\|\sum_{i=1}^n a_i X_i\|_p \leq c'_\alpha(\sqrt{p}\|a\|_2 + p^{1/\alpha}\|a\|_{\alpha^*}).$$

Replacing a with b, the proof is complete. □

We now give proofs of Theorem 2.1 and Theorem 2.2.

Proof. Using Lemma B.2 with $h(t) = |t|^p$, for $p \geq 2$,

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

Then, using Lemma B.3 and setting $a_i = 1$ for all $i = 1,...,n$, we have if $0 < \alpha \leq 1$,

$$\|f(X_1,...,X_n) - \mathbb{E}f(X_1,...,X_n)\|_p \leq c_\alpha\left(\sqrt{p}\left(\sum_{i=1}^n \|H_i(X_i,X'_i)\|_{\psi_\alpha}^2\right)^{\frac{1}{2}} + p^{1/\alpha}\max_{1\leq i\leq n}\|H_i(X_i,X'_i)\|_{\psi_\alpha}\right);$$

while if $\alpha > 1$,

$$\|f(X_1,...,X_n) - \mathbb{E}f(X_1,...,X_n)\|_p \leq c'_\alpha\left(\sqrt{p}\left(\sum_{i=1}^n \|H_i(X_i,X'_i)\|_{\psi_\alpha}^2\right)^{\frac{1}{2}} + p^{1/\alpha}\|(\|H_i(X_i,X'_i)\|_{\psi_\alpha})\|_{\alpha^*}\right),$$

which completes the proof of Theorem 2.1.

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

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

Let $\exp(-p) = \mathbb{E}|f(X_1,...,X_n) - \mathbb{E}f(X_1,...,X_n)|^p/t^p$, we get

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

Further, let $\delta = \exp(-p)$, we have $p = \log(1/\delta)$ and $0 < \delta < 1/e^2$. Putting the above results together, the proof of Theorem 2.2 is complete. □

C. Proofs of Section 2.3

In this section, we provide proofs of results in Section 2.3.

Proof of Lemma 2.9. Given any samples $S = \{Z_1, ..., Z_n\} \in \mathcal{Z}^n$ and $S^i = \{Z_1, ..., Z_{i-1}, Z_i', Z_{i+1}, ..., Z_n\} \in \mathcal{Z}^n$, according to Lemma 7 in (Bousquet & Elisseeff, 2002), for all $i \in [n]$,

$$\mathbb{E}[R(\mathcal{A}, S) - R_n(\mathcal{A}, S)] = \mathbb{E}_{S, Z_i'}[\ell(\mathcal{A}_S, Z_i') - \ell(\mathcal{A}_{S^i}, Z_i')].$$

For fixed $i \in [n]$ and Z_i^{i-1}, Z_{i+1}^n, define

$$V_i(Z_i, Z_i') = \ell(\mathcal{A}_{Z_1^n}, Z_i') - \ell(\mathcal{A}_{Z_1^{i-1}, Z_i', Z_{i+1}^n}, Z_i').$$

The totally Lipschitz stable condition implies that

$$|V_i(Z_i, Z_i')| \leq \gamma d(Z_i, Z_i').$$

This gives

$$\mathbb{E}[R(\mathcal{A}, S) - R_n(\mathcal{A}, S)] \leq \gamma \mathbb{E} d(Z_i, Z_i'). \tag{5}$$

We now consider two cases separately. In the case of $\alpha > 1$, (5) gives

$$\begin{aligned}
&\exp\left(\left(\frac{\mathbb{E}[R(\mathcal{A}, S) - R_n(\mathcal{A}, S)]}{\gamma \|d(Z_i, Z_i')\|_{\psi_\alpha}}\right)^\alpha\right) \leq \exp\left(\left(\frac{|\mathbb{E}[R(\mathcal{A}, S) - R_n(\mathcal{A}, S)]|}{\gamma \|d(Z_i, Z_i')\|_{\psi_\alpha}}\right)^\alpha\right) \\
\leq &\exp\left(\left(\frac{|\gamma \mathbb{E} d(Z_i, Z_i')|}{\gamma \|d(Z_i, Z_i')\|_{\psi_\alpha}}\right)^\alpha\right) \leq \mathbb{E} \exp\left(\left(\frac{\gamma |d(Z_i, Z_i')|}{\gamma \|d(Z_i, Z_i')\|_{\psi_\alpha}}\right)^\alpha\right) \leq 2,
\end{aligned}$$

where the third inequality follows from the Jensen's inequality and the last inequality uses the definition $\mathbb{E} \exp\left(\left(\frac{|d(Z_i, Z_i')|}{\|d(Z_i, Z_i')\|_{\psi_\alpha}}\right)^\alpha\right) \leq 2$. Thus, taking logarithms yields the following inequality

$$\mathbb{E}[R(\mathcal{A}, S) - R_n(\mathcal{A}, S)] \leq (\log 2)^{1/\alpha} \gamma \|d(Z_i, Z_i')\|_{\psi_\alpha} = (\log 2)^{1/\alpha} \gamma \Delta_\alpha(\mathcal{Z}).$$

In the case of $0 < \alpha \leq 1$,

$$\begin{aligned}
&\mathbb{E}[d(Z_i, Z_i')] \leq \int_0^\infty \mathbb{P}(|d(Z_i, Z_i')| > x) dx = \int_0^\infty \mathbb{P}\left(e^{\left(\frac{|d(Z_i, Z_i')|}{\|d(Z_i, Z_i')\|_{\psi_\alpha}}\right)^\alpha} > e^{\left(\frac{x}{\|d(Z_i, Z_i')\|_{\psi_\alpha}}\right)^\alpha}\right) dx \\
&\leq \int_0^\infty \frac{\mathbb{E}[e^{(\frac{|d(Z_i, Z_i')|}{\|d(Z_i, Z_i')\|_{\psi_\alpha}})^\alpha}]}{e^{(\frac{x}{\|d(Z_i, Z_i')\|_{\psi_\alpha}})^\alpha}} dx \leq 2 \int_0^\infty e^{-(\frac{x}{\|d(Z_i, Z_i')\|_{\psi_\alpha}})^\alpha} dx \\
&= 2\|d(Z_i, Z_i')\|_{\psi_\alpha} \frac{1}{\alpha} \int_0^\infty e^{-u} u^{\frac{1}{\alpha}-1} du = 2\|d(Z_i, Z_i')\|_{\psi_\alpha} \frac{1}{\alpha} \Gamma(\frac{1}{\alpha}) = 2\|d(Z_i, Z_i')\|_{\psi_\alpha} \Gamma(\frac{1}{\alpha} + 1).
\end{aligned}$$

Thus, we get

$$\mathbb{E}[R(\mathcal{A}, S) - R_n(\mathcal{A}, S)] \leq 2\Gamma(\frac{1}{\alpha} + 1) \gamma \|d(Z_i, Z_i')\|_{\psi_\alpha} = 2\Gamma(\frac{1}{\alpha} + 1) \gamma \Delta_\alpha(\mathcal{Z}).$$

The proof is complete. □

D. Proofs of Section 2.4

In this section, we provide proofs of results in Section 2.4.

We first introduce the following lemma which translates a moment bound into a high probability bound.

Lemma D.1. *Let Z be a random variable with*

$$\|Z\|_p \leq \sqrt{p}a + p^{1/\alpha}b$$

for some $a, b > 0$ and for any $p \geq 2$. Then for any $\delta \in (0, 1/e^2)$ we have, with probability at least $1-\delta$,

$$|Z| \leq e\Big(a\sqrt{\log\left(\frac{1}{\delta}\right)} + b\log^{1/\alpha}\left(\frac{1}{\delta}\right)\Big),$$

where e is the base of the natural logarithm.

Proof of Lemma D.1. By Markov's inequality for any $\delta \in (0, 1/e^2)$,

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

Picking $p = \log(\frac{1}{\delta}) \geq 2$, so that $|Z| \leq e\Big(a\sqrt{\log\left(\frac{1}{\delta}\right)} + b\log^{1/\alpha}\left(\frac{1}{\delta}\right)\Big)$. □

We then give the proof of Lemma 2.15.

Proof of Lemma 2.15. The proof follows the technique of (Bousquet et al., 2020). Suppose that $n = 2^k$, otherwise, we can add extra functions equal to zero, increasing the number of terms by at most two times. We use the notation $[n] = \{1, ..., n\}$. Consider a sequence of partitions $\mathcal{B}_0, ..., \mathcal{B}_k$ with $\mathcal{B}_0 = \{\{i\} : i \in [n]\}$, $\mathcal{B}_k = \{[n]\}$, and to get $\mathcal{B}_l$ from $\mathcal{B}_{l+1}$ we split each subset in $\mathcal{B}_{l+1}$ into two equal parts. We have

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

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

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

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

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

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

and the total sum of interest satisfies by the triangle inequality

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

Since $\mathbb{E}(\mathbb{E}[g_i|Z_i]) = 0$, by applying Lemma B.3 we have

- if $0 < \alpha \leq 1$, let $c_\alpha = 2\sqrt{2}((\log 2)^{1/\alpha} + e^3\Gamma^{1/2}(\frac{2}{\alpha} + 1) + e^3 3^{\frac{2-\alpha}{3\alpha}} \sup_{p\geq 2} p^{\frac{-1}{\alpha}}\Gamma^{1/p}(\frac{p}{\alpha} + 1))$,

$$\left\|\sum_{i=1}^n \mathbb{E}[g_i|Z_i]\right\|_p \leq c_\alpha\left(\sqrt{pn}M + p^{1/\alpha}M\right);$$

- if $\alpha > 1$, let $1/\alpha^* + 1/\alpha = 1$ and $c'_\alpha = 8e + 2(\log 2)^{1/\alpha}$,

$$\left\| \sum_{i=1}^{n} \mathbb{E}[g_i | Z_i] \right\|_p \leq c'_\alpha \left(\sqrt{pn} M + p^{1/\alpha} n^{\frac{1}{\alpha^*}} M \right).$$

Furthermore, observe that

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

that is, the expectation is taken w.r.t. the variables Z_j , $j \in B^{l+1}(i)\setminus B^l(i)$. It is also not hard to see that the function g_i^l preserves the differences property, just like the the function g_i. Therefore, if we apply Theorem 2.1 conditioned on $Z_i, Z_{[n]\setminus B^{l+1}(i)}$, we obtain a uniform bound

- if $0 < \alpha \leq 1$, let $c_\alpha = 2\sqrt{2}((\log 2)^{1/\alpha} + e^3 \Gamma^{1/2}(\frac{2}{\alpha} + 1) + e^3 3^{\frac{2-\alpha}{3\alpha}} \sup_{p \geq 2} p^{\frac{-1}{\alpha}} \Gamma^{1/p}(\frac{p}{\alpha} + 1))$,

$$\|g_i^l - g_i^{l+1}\|_p (Z_i, Z_{[n]\setminus B^{l+1}(i)}) \leq c_\alpha \left(\sqrt{p 2^l} \beta + p^{1/\alpha} \beta \right);$$

- if $\alpha > 1$, let $1/\alpha^* + 1/\alpha = 1$ and $c'_\alpha = 8e + 2(\log 2)^{1/\alpha}$,

$$\|g_i^l - g_i^{l+1}\|_p (Z_i, Z_{[n]\setminus B^{l+1}(i)}) \leq c'_\alpha \left(\sqrt{p 2^l} \beta + p^{1/\alpha} (2^l)^{\frac{1}{\alpha^*}} \beta \right),$$

as there are 2^l indices in $B^{l+1}(i)\setminus B^l(i)$. It follows from Lemma 2 of (Bousquet et al., 2020) that

- if $0 < \alpha \leq 1$,

$$\|g_i^l - g_i^{l+1}\|_p \leq \|g_i^l - g_i^{l+1}\|_p (Z_i, Z_{[n]\setminus B^{l+1}(i)}) \leq c_\alpha \left(\sqrt{p 2^l} \beta + p^{1/\alpha} \beta \right);$$

- if $\alpha > 1$,

$$\|g_i^l - g_i^{l+1}\|_p \leq \|g_i^l - g_i^{l+1}\|_p (Z_i, Z_{[n]\setminus B^{l+1}(i)}) \leq c'_\alpha \left(\sqrt{p 2^l} \beta + p^{1/\alpha} (2^l)^{\frac{1}{\alpha^*}} \beta \right).$$

Let us take a look at the sum $\sum_{i \in B^l} g_i^l - g_i^{l+1}$ for $B^l \in \mathcal{B}_l$. Since $g_i^l - g_i^{l+1}$ for $i \in B^l$ depends only on Z_i, $Z_{[n]\setminus B^l}$, the terms are independent and zero mean conditioned on $Z_{[n]\setminus B^l}$. Applying Lemma A.4, we have for any $p \geq 2$,

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

Integrating with respect to $(Z_{[n]\setminus B^l})$ and using the bound of $\|g_i^l - g_i^{l+1}\|_p$, we have if $0 < \alpha \leq 1$,

$$\left\| \sum_{i \in B^l} g_i^l - g_i^{l+1} \right\|_p \leq c_\alpha 3\sqrt{2p2^l} \left(\sqrt{p 2^l} \beta + p^{1/\alpha} \beta \right);$$

while if $\alpha > 1$,

$$\left\| \sum_{i \in B^l} g_i^l - g_i^{l+1} \right\|_p \leq c'_\alpha 3\sqrt{2p2^l} \left(\sqrt{p 2^l} \beta + p^{1/\alpha} (2^l)^{\frac{1}{\alpha^*}} \beta \right),$$

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

$$\| \sum_{i \in [n]} g_i^l - g_i^{l+1} \|_p \leq \sum_{B^l \in B_l} \| \sum_{i \in B^l} g_i^l - g_i^{l+1} \|_p \leq 2^{k-l} \times \| \sum_{i \in B^l} g_i^l - g_i^{l+1} \|_p,$$

which means if $0 < \alpha \leq 1$,

$$\| \sum_{i\in[n]} g_i^l - g_i^{l+1} \|_p \leq c_\alpha 3\sqrt{2p2^l} 2^{k-l} \left(\sqrt{p2^l}\beta + p^{1/\alpha}\beta \right);$$

while if $\alpha > 1$,

$$\| \sum_{i\in[n]} g_i^l - g_i^{l+1} \|_p \leq c'_\alpha 3\sqrt{2p2^l} 2^{k-l} \left(\sqrt{p2^l}\beta + p^{1/\alpha}(2^l)^{\frac{1}{\alpha^*}}\beta \right).$$

Recall, that $2^k < 2n$ due to the possible extension of the sample. If $0 < \alpha \leq 1$,

$$\begin{aligned}\sum_{l=0}^{k-1} \| \sum_{i\in[n]} g_i^l - g_i^{l+1} \|_p &\leq c_\alpha 3\sqrt{2} \left(\sum_{l=0}^{k-1} p2n\beta + \sum_{l=0}^{k-1} 2^{k-\frac{l}{2}} p^{1/\alpha+1/2}\beta \right) \\ &\leq c_\alpha 3\sqrt{2} \left(p2n\beta\lceil \log_2 n \rceil + 2np^{1/\alpha+1/2}\beta \right);\end{aligned}$$

while if $\alpha > 1$,

$$\begin{aligned}\sum_{l=0}^{k-1} \| \sum_{i\in[n]} g_i^l - g_i^{l+1} \|_p &\leq c'_\alpha 3\sqrt{2} \left(\sum_{l=0}^{k-1} 2np\beta + \sum_{l=0}^{k-1} 2^{k-\frac{l}{2}+\frac{l}{\alpha^*}} p^{1/\alpha+1/2}\beta \right) \\ &\leq c'_\alpha 3\sqrt{2} \left(2np\beta\lceil \log_2 n \rceil + \sum_{l=0}^{k-1} 2^{k-\frac{l}{2}+\frac{l}{\alpha^*}} p^{1/\alpha+1/2}\beta \right) \\ &\leq c'_\alpha 3\sqrt{2} \left(2np\beta\lceil \log_2 n \rceil + 3\times 2^{(\frac{1}{\alpha^*}+\frac{1}{2})k} p^{1/\alpha+1/2}\beta \right) \\ &\leq c'_\alpha 3\sqrt{2} \left(2np\beta\lceil \log_2 n \rceil + 3\times (2n)^{(\frac{1}{\alpha^*}+\frac{1}{2})} p^{1/\alpha+1/2}\beta \right),\end{aligned}$$

where we have used the fact $\sum_{l=0}^{k-1} 2^{k-\frac{l}{2}+\frac{l}{\alpha^*}} \leq 2^k \times 3 \times 2^{(\frac{1}{\alpha^*}-\frac{1}{2})k}$. Plugging the above bound together we get if $0 < \alpha \leq 1$,

$$\left\| \sum_{i=1}^n g_i(Z) \right\|_p \leq c_\alpha 3\sqrt{2} \left(p2n\beta\lceil \log_2 n \rceil + 2np^{1/\alpha+1/2}\beta \right) + c_\alpha \left(\sqrt{pn}M + p^{1/\alpha}M \right);$$

while if $\alpha > 1$,

$$\left\| \sum_{i=1}^n g_i(Z) \right\|_p \leq c'_\alpha 3\sqrt{2} \left(2np\beta\lceil \log_2 n \rceil + 3\times (2n)^{(\frac{1}{\alpha^*}+\frac{1}{2})} p^{1/\alpha+1/2}\beta \right) + c'_\alpha \left(\sqrt{pn}M + p^{1/\alpha} n^{\frac{1}{\alpha^*}} M \right),$$

The proof is complete. □

We now give the proof of Theorem 2.17.

Proof of Theorem 2.17. Given any samples $S = \{Z_1, ..., Z_n\} \in \mathcal{Z}^n$ and $S^i = \{Z_1, ..., Z_{i-1}, Z'_i, Z_{i+1}, ..., Z_n\} \in \mathcal{Z}^n$, let us consider $h_i(S) := R(\mathcal{A}, S) - \ell(\mathcal{A}_S, Z_i)$, $g_i(S) = \mathbb{E}_{Z'_i}[R(\mathcal{A}, S^i) - \ell(\mathcal{A}_{S^i}, Z_i)]$. It is clear that

$$\|R(\mathcal{A}, S) - R_n(\mathcal{A}, S)\|_p \leq \frac{1}{n} \left\| \sum_{i=1}^n h_i(S) \right\|_p \leq \frac{1}{n} \left(\left\| \sum_{i=1}^n g_i(S) \right\|_p + \left\| \sum_{i=1}^n h_i(S) - g_i(S) \right\|_p \right).$$

We firset focus on $\|\sum_{i=1}^n g_i(S)\|_p$. By definition it holds that $\mathbb{E}_{Z_i}[g_i(S)] = 0$. Based on the triangle inequality we can show that

$$\begin{aligned}\left\| \mathbb{E}_{S\setminus\{Z_i\}}[g_i(S)] \right\|_{\psi_\alpha} &\leq \|g_i(S)\|_{\psi_\alpha} = \|\mathbb{E}_{Z'_i}[\mathbb{E}_{Z''_i}[\ell(\mathcal{A}_{S^i}, Z''_i)] - \ell(\mathcal{A}_{S^i}, Z_i)]\|_{\psi_\alpha} \\ &= \|\mathbb{E}_{Z'_i}[\mathbb{E}_{Z''_i}[\ell(\mathcal{A}_{S^i}, Z''_i) - \ell(\mathcal{A}_{S^i}, Z_i)]]\|_{\psi_\alpha} \leq \gamma \|d(Z''_i, Z_i)\|_{\psi_\alpha} = \gamma\Delta_\alpha(\mathcal{Z}),\end{aligned}$$

where in the last inequality we have used the total Lipschitz stability assumption on the algorithm $\mathcal{A}$. Next we further show that g_i has a difference property with respect to all variables in S except Z_i. Indeed, for each $j \neq i$ it can be verified that

$$\begin{aligned} g_i(S) - g_i(S^j) &\leq \mathbb{E}_{Z_i'}[R(\mathcal{A}, S^i) - R(\mathcal{A}, (S^i)^j)] + \mathbb{E}_{Z_i'}[\ell(\mathcal{A}_{(S^i)^j}, Z_i) - \ell(\mathcal{A}_{S^i}, Z_i)] \\ &= \mathbb{E}_{Z_i'}[\mathbb{E}_Z[\ell(\mathcal{A}_{S^i}, Z) - \ell(\mathcal{A}_{(S^i)^j}, Z)] + \mathbb{E}_{Z_i'}[\ell(\mathcal{A}_{(S^i)^j}, Z_i) - \ell(\mathcal{A}_{S^i}, Z_i)] \\ &\leq 2\gamma d(Z_j, Z_j'), \end{aligned}$$

where in the last inequality we have used the total Lipschitz stability assumption on the algorithm $\mathcal{A}$, and it is clear that $\|2\gamma d(Z_j, Z_j')\|_{\psi_\alpha} \leq 2\gamma\Delta_\alpha(\mathcal{Z})$. Therefore, $\{g_i\}$ satisfy the conditions of Lemma 2.15 and thus

- if $0 < \alpha \leq 1$,

$$\Big\|\sum_{i=1}^n g_i(S)\Big\|_p \leq c_\alpha 6\sqrt{2}\left(p2n\gamma\Delta_\alpha(\mathcal{Z})\lceil\log_2 n\rceil + 2np^{1/\alpha+1/2}\gamma\Delta_\alpha(\mathcal{Z})\right) + 2c_\alpha\left(\sqrt{pn}\gamma\Delta_\alpha(\mathcal{Z}) + p^{1/\alpha}\gamma\Delta_\alpha(\mathcal{Z})\right);$$

- if $\alpha > 1$,

$$\Big\|\sum_{i=1}^n g_i(S)\Big\|_p \leq c_\alpha' 6\sqrt{2}\gamma\Delta_\alpha(\mathcal{Z})\left(2np\lceil\log_2 n\rceil + 3\times(2n)^{(\frac{1}{\alpha^*}+\frac{1}{2})}p^{1/\alpha+1/2}\right) + 2c_\alpha'\gamma\Delta_\alpha(\mathcal{Z})\left(\sqrt{pn} + p^{1/\alpha}n^{\frac{1}{\alpha^*}}\right).$$

Then, we focus on $\|\sum_{i=1}^n h_i(S) - g_i(S)\|_p$. It can be verified that

$$\begin{aligned} \Big\|\sum_{i=1}^n h_i(S) - g_i(S)\Big\|_p &\leq \Big\|\sum_{i=1}^n \mathbb{E}_{Z_i'}[R(\mathcal{A}, S) - R(\mathcal{A}, S^i)]\Big\|_p + \Big\|\sum_{i=1}^n \mathbb{E}_{Z_i'}[\ell(\mathcal{A}_S, Z_i) - \ell(\mathcal{A}_{S^i}, Z_i)]\Big\|_p \\ &= \|\sum_{i=1}^n \mathbb{E}_{Z_i'}[\mathbb{E}_Z[\ell(\mathcal{A}_S, Z) - \ell(\mathcal{A}_{S^i}, Z)]\|_p + \|\sum_{i=1}^n \mathbb{E}_{Z_i'}[\ell(\mathcal{A}_S, Z_i) - \ell(\mathcal{A}_{S^i}, Z_i)]\|_p \\ &\leq \|\sum_{i=1}^n \gamma d(Z_i, Z_i')\|_p + \|\sum_{i=1}^n \gamma d(Z_i, Z_i')\|_p \leq 2\gamma\sum_{i=1}^n \|d(Z_i, Z_i')\|_p, \end{aligned}$$

and in the second inequality we have used the total Lipschitz stability assumption. By Markov's inequality, we obtain

$$\begin{aligned} &\mathbb{E}[|d(Z_i, Z_i')|^p] \leq \int_0^\infty P(|d(Z_i, Z_i')|^p > t)dt \leq \int_0^\infty P(|d(Z_i, Z_i')| > t^{1/p})dt \\ &\leq \int_0^\infty \frac{\mathbb{E}[e^{(|d(Z_i,Z_i')|/\|d(Z_j,Z_j')\|_{\psi_\alpha})^\alpha}]}{e^{(t^{1/p}/\|d(Z_j,Z_j')\|_{\psi_\alpha})^\alpha}}dt \leq 2\int_0^\infty e^{-(t^{1/p}/\|d(Z_j,Z_j')\|_{\psi_\alpha})^\alpha}dt = 2\int_0^\infty e^{-(t/\|d(Z_j,Z_j')\|_{\psi_\alpha}^p)^{\alpha/p}}dt \\ &= 2\|d(Z_j, Z_j')\|_{\psi_\alpha}^p \frac{p}{\alpha}\int_0^\infty e^{-u}u^{\frac{p}{\alpha}-1}du = 2\|d(Z_j, Z_j')\|_{\psi_\alpha}^p \frac{p}{\alpha}\Gamma(\frac{p}{\alpha}) = 2\|d(Z_j, Z_j')\|_{\psi_\alpha}^p \Gamma(\frac{p}{\alpha}+1). \end{aligned}$$

Taking the k-th root of the expression above yields

$$\sum_{i=1}^n \|d(Z_i, Z_i')\|_p \leq n\Delta_\alpha(\mathcal{Z})(2\Gamma(\frac{p}{\alpha}+1))^{\frac{1}{p}}.$$

Plugging these bounds together, we get

- if $0 < \alpha \leq 1$,

$$\begin{aligned} &\|R(\mathcal{A}, S) - R_n(\mathcal{A}, S)\|_p \\ \leq &c_\alpha 6\sqrt{2}\gamma\Delta_\alpha(\mathcal{Z})\left(p2\lceil\log_2 n\rceil + 2p^{1/\alpha+1/2}\right) + \frac{2c_\alpha}{n}\gamma\Delta_\alpha(\mathcal{Z})\left(\sqrt{pn} + p^{1/\alpha}\right) + 2\gamma\Delta_\alpha(\mathcal{Z})(2\Gamma(\frac{p}{\alpha}+1))^{\frac{1}{p}}; \end{aligned}$$

- if $\alpha > 1$,

$$\|R(\mathcal{A},S) - R_n(\mathcal{A},S)\|_p \leq c'_\alpha 6\sqrt{2}\gamma\Delta_\alpha(\mathcal{Z})\left(2p\lceil\log_2 n\rceil + 3\times 2^{(\frac{1}{\alpha^*}+\frac{1}{2})}\frac{1}{n^{\frac{1}{2}-\frac{1}{\alpha^*}}}p^{1/\alpha+1/2}\right)$$
$$+\frac{2c'_\alpha}{n}\gamma\Delta_\alpha(\mathcal{Z})\left(\sqrt{pn}+p^{1/\alpha}n^{\frac{1}{\alpha^*}}\right)+2\gamma\Delta_\alpha(\mathcal{Z})(2\Gamma(\frac{p}{\alpha}+1))^{\frac{1}{p}}.$$

By Lemma D.1, picking $p = \log(\frac{1}{\delta})$ we have that for any $\delta \in (0, 1/e^2)$, with probability at least $1-\delta$

- if $0 < \alpha \leq 1$,

$$|R(\mathcal{A},S) - R_n(\mathcal{A},S)| \leq c_\alpha 6\sqrt{2}\gamma\Delta_\alpha(\mathcal{Z})\left(2\lceil\log_2 n\rceil\log(\frac{1}{\delta})+2\log^{1/\alpha+1/2}(\frac{1}{\delta})\right)$$
$$+\frac{2c_\alpha}{n}\gamma\Delta_\alpha(\mathcal{Z})\left(\sqrt{n}\log^{\frac{1}{2}}(\frac{1}{\delta})+\log^{1/\alpha}(\frac{1}{\delta})\right)+2\gamma\Delta_\alpha(\mathcal{Z})(2\Gamma(\log(\frac{1}{\delta})\frac{1}{\alpha}+1))^{\frac{1}{\log(\frac{1}{\delta})}};$$

- if $\alpha > 1$,

$$|R(\mathcal{A},S) - R_n(\mathcal{A},S)| \leq c'_\alpha 6\sqrt{2}\gamma\Delta_\alpha(\mathcal{Z})\left(2\lceil\log_2 n\rceil\log(\frac{1}{\delta})+3\times 2^{(\frac{1}{\alpha^*}+\frac{1}{2})}\frac{1}{n^{\frac{1}{2}-\frac{1}{\alpha^*}}}\log^{1/\alpha+1/2}(\frac{1}{\delta})\right)$$
$$+\frac{2c'_\alpha}{n}\gamma\Delta_\alpha(\mathcal{Z})\left(\sqrt{n}\log^{\frac{1}{2}}(\frac{1}{\delta})+\log^{1/\alpha}(\frac{1}{\delta})n^{\frac{1}{\alpha^*}}\right)+2\gamma\Delta_\alpha(\mathcal{Z})(2\Gamma(\log(\frac{1}{\delta})\frac{1}{\alpha}+1))^{\frac{1}{\log(\frac{1}{\delta})}}.$$

The proof is complete. □

E. Proofs of Section 2.6

To prove Theorem 2.20, we introduce the following technical lemma.

Lemma E.1. *Let $X = \{X_1, ..., X_n\}$ be a set of independent subweibull random variables with values in a Banach space $(\mathcal{X}, \|\cdot\|)$ such that $\|\|X_i\|\|_{\psi_\alpha} \leq \infty$. Define $X^i = \{X_1, ..., X_{i-1}, X'_i, X_{i+1}, ..., X_n\}$, where $(X'_1, ..., X'_n)$ is an independent copy of $(X_1, ..., X_n)$. Then for any $0 < \delta < 1/e^2$, with probability at least $1-\delta$,*

1.) if $0 < \alpha \leq 1$, let $c_\alpha = 2\sqrt{2}((\log 2)^{1/\alpha} + e^3\Gamma^{1/2}(\frac{2}{\alpha}+1) + e^3 3^{\frac{2-\alpha}{3\alpha}}\sup_{p\geq 2} p^{\frac{-1}{\alpha}}\Gamma^{1/p}(\frac{p}{\alpha}+1))$, we have

$$\left\|\sum_{i=1}^n (X_i - \mathbb{E}X'_1)\right\| - \mathbb{E}\left\|\sum_{i=1}^n (X_i - \mathbb{E}X'_i)\right\| \leq 2c_\alpha\left(\sqrt{\log(\frac{1}{\delta})}\left(\sum_{i=1}^n \|\|X_i\|\|^2_{\psi_\alpha}\right)^{\frac{1}{2}} + \log^{1/\alpha}(\frac{1}{\delta})\max_{1\leq i\leq n}\|\|X_i\|\|_{\psi_\alpha}\right);$$

2.) if $\alpha \geq 1$, let $1/\alpha^* + 1/\alpha = 1$ and $c'_\alpha = 8e + 2(\log 2)^{1/\alpha}$, we have

$$\left\|\sum_{i=1}^n (X_i - \mathbb{E}X'_1)\right\| - \mathbb{E}\left\|\sum_{i=1}^n (X_i - \mathbb{E}X'_i)\right\| \leq 2c'_\alpha\left(\sqrt{\log(\frac{1}{\delta})}\left(\sum_{i=1}^n \|\|X_i\|\|^2_{\psi_\alpha}\right)^{\frac{1}{2}} + \log^{1/\alpha}(\frac{1}{\delta})\|(\|\|X\|\|_{\psi_\alpha})\|_{\alpha^*}\right).$$

Proof of Lemma E.1. We look at the function $f(x) = \|\sum_{i=1}^n (x_i - \mathbb{E}X'_1)\|$. Then we have

$$|f(X) - f(X^i)| = \left|\left\|\sum_{j\neq i} X_j + X_i - n\mathbb{E}X'_1\right\| - \left\|\sum_{j\neq i} X_j + X'_i - n\mathbb{E}X'_1\right\|\right| \leq \|X_i - X'_i\|.$$

Thus, we have $\|f(X) - f(X^i)\|_{\psi_\alpha} \leq \|\,\|X_i - X'_i\|\,\|_{\psi_\alpha} \leq 2\|\|X_i\|\|_{\psi_\alpha}$.

Plugging these bounds into Theorem 2.2, if $0 < \alpha \leq 1$

$$\left\|\sum_{i=1}^n (X_i - \mathbb{E}X'_1)\right\| - \mathbb{E}\left\|\sum_{i=1}^n (X_i - \mathbb{E}X'_i)\right\| \leq 2c_\alpha\left(\sqrt{\log(\frac{1}{\delta})}\left(\sum_{i=1}^n \|\|X_i\|\|^2_{\psi_\alpha}\right)^{\frac{1}{2}} + \log^{1/\alpha}(\frac{1}{\delta})\max_{1\leq i\leq n}\|\|X_i\|\|_{\psi_\alpha}\right);$$

if $\alpha \geq 1$

$$\left\|\sum_{i=1}^{n}(X_i - \mathbb{E}X_1')\right\| - \mathbb{E}\left\|\sum_{i=1}^{n}(X_i - \mathbb{E}X_i')\right\| \leq 2c_\alpha' \left(\sqrt{\log(\frac{1}{\delta})}\left(\sum_{i=1}^{n}\|\|X_i\|\|_{\psi_\alpha}^2\right)^{\frac{1}{2}} + \log^{1/\alpha}(\frac{1}{\delta})\|(\|\|X\|\|_{\psi_\alpha})\|_{\alpha^*}\right),$$

where $(\|\|X\|\|_{\psi_\alpha}) = (\|\|X_1\|\|_{\psi_\alpha}, ..., \|\|X_n\|\|_{\psi_\alpha}) \in \mathbb{R}^n$. The proof is complete. □

Proof of Theorem 2.20. The vector space

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

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

$$\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X_i')] = \left\|\sum_{i=1}^{n}\bar{X}_i\right\|_{\mathcal{B}}.$$

By Jensen's inequality for $p \geq 1$, we have

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

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

Thus, by Theorem E.1, we have if $0 < \alpha \leq 1$

$$\begin{aligned}
&\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X_i')] - \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X_i')]\right] \\
&= \left\|\sum_{i=1}^{n}\bar{X}_i\right\|_{\mathcal{B}} - \mathbb{E}\left\|\sum_{i=1}^{n}\bar{X}_i\right\|_{\mathcal{B}} \leq \frac{4L}{n}c_\alpha\left(\sqrt{\log(\frac{1}{\delta})}\left(\sum_{i=1}^{n}\|\|X_i\|\|_{\psi_\alpha}^2\right)^{\frac{1}{2}} + \log^{1/\alpha}(\frac{1}{\delta})\max_{1\leq i\leq n}\|\|X_i\|\|_{\psi_\alpha}\right);
\end{aligned}$$

if $\alpha \geq 1$

$$\begin{aligned}
&\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X_i')] - \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X_i')]\right] \\
&= \left\|\sum_{i=1}^{n}\bar{X}_i\right\|_{\mathcal{B}} - \mathbb{E}\left\|\sum_{i=1}^{n}\bar{X}_i\right\|_{\mathcal{B}} \leq \frac{4L}{n}c_\alpha'\left(\sqrt{\log(\frac{1}{\delta})}\left(\sum_{i=1}^{n}\|\|X_i\|\|_{\psi_\alpha}^2\right)^{\frac{1}{2}} + \log^{1/\alpha}(\frac{1}{\delta})\|(\|\|X\|\|_{\psi_\alpha})\|_{\alpha^*}\right),
\end{aligned}$$

where $(\|\|X\|\|_{\psi_\alpha}) = (\|\|X_1\|\|_{\psi_\alpha}, ..., \|\|X_n\|\|_{\psi_\alpha}) \in \mathbb{R}^n$. The proof is complete. □