Concentration Inequalities for General Functions of Heavy-Tailed Random Variables

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

Abstract

Concentration inequalities play an essential role in the study of machine learning and high dimensional statistics. In this paper, we obtain unbounded analogues of the popular bounded difference inequality for functions of independent random variables with heavy-tailed distributions. The main results provide a general framework applicable to all heavy-tailed distributions with finite variance. To illustrate the strength of our results, we present applications to sub-exponential tails, sub-Weibull tails, and heavier polynomially decaying tails. Applied to some standard problems in statistical learning theory (vector valued concentration, Rademacher complexity, and algorithmic stability), we show that these inequalities allow an extension of existing results to heavy-tailed distributions up to finite variance.

1. Introduction

Concentration inequalities are fundamental in empirical science and serve as a critical toolkit for the study of both natural and artificial learning systems (Boucheron et al., 2005; 2013). They have been investigated over several decades and applied in numerous fields, including convex geometry, functional analysis, probability theory, information theory, communications and coding theory, learning theory, and theoretical computer science (Raginsky & Sason, 2015; 2018).

The bounded difference inequality, also known as McDiarmid's inequality (McDiarmid, 1998), is among the most renowned concentration inequalities. It has been extensively used within statistical learning theory (Bousquet & Elisseeff, 2002; Bartlett & Mendelson, 2002) as a potent instrument. The bounded difference inequality encapsulates the probabilities that a function of independent random variables varies from its mean in terms of the sum of conditional ranges. It surpasses the general Hoeffding-type and Bernstein-type inequalities that rely on the summation of independent random variables (Vershynin, 2018; Wainwright, 2019) by characterizing a general function; this renders it more adaptable and effective for estimating complex statistics beyond the sum (Maurer, 2019; Maurer & Pontil, 2018; 2019).

However, a limitation of the classical bounded difference inequality (McDiarmid, 1998) is its requirement that conditional ranges be uniformly bounded, a constraint that significantly reduces its applicability. Many scenarios demand the use of its unbounded variants. For example, extensive research has focused on establishing generalization bounds in unbounded situations, for which the classic inequality is not readily applicable (Cortes et al., 2021; Kontorovich, 2014; Meir & Zhang, 2003). It is possible for conditional ranges to be infinite, yet the conditional versions (obtained by fixing all but one argument of the function) may exhibit certain decaying tails (Maurer & Pontil, 2021). In such relaxed conditions, one might anticipate the existence of bounded difference-type inequalities. Some studies (Meir & Zhang, 2003; Kontorovich, 2014; Kutin, 2002; Maurer & Pontil, 2021) have successfully extended the classical inequality to cover the unbounded case. Specifically, Kutin (2002); Meir & Zhang (2003); Kontorovich (2014) have provided concentration inequalities for sub-Gaussian decay tails, while Maurer & Pontil (2021) have further contributed inequalities for heavier sub-exponential decay tails.

Unsatisfactory, both sub-Gaussian and sub-exponential distributions are relatively light-tailed distributions (Vladimirova et al., 2020; Wainwright, 2019). A distinctive difference between heavy-tailed distributions and sub-Gaussian and sub-exponential distributions is the moment generating function (MGF). The MGF exists for sub-Gaussian and sub-exponential distributions (Vershynin, 2018), while it does not exist for heavy-tailed distributions (Foss et al., 2011). Thus, the technique used in (Meir & Zhang, 2003; Kontorovich, 2014; Kutin, 2002; Maurer & Pontil, 2021) to find upper bounds for the MGF clearly fails for heavy-tailed distributions. Meanwhile, in many applications, such as probability theory (Wong

[*]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.

et al., 2020), high-dimensional statistics (Kuchibhotla & Chakrabortty, 2018; Guédon et al., 2014), stochastic optimization (Gurbuzbalaban et al., 2021), and signal processing (Bakhshizadeh et al., 2020), the assumption of light-tailed sub-Gaussian and sub-exponential distributions seems inappropriate. Therefore, concentration inequalities for heavy-tailed random variables are necessary (Foss et al., 2011).

This paper amis to provide bounded difference-type concentration inequalities, where the centered conditional versions have heavy-tailed distributions. We prove these inequalities using the entropy method (Boucheron et al., 2003; 2013; Ledoux, 2001; Maurer, 2012; Raginsky & Sason, 2018) and the truncation technique for random variables (Bakhshizadeh et al., 2023; Nagaev, 1979; Hahn & Klass, 1997; Klass & Nowicki, 2007; Hitczenko & Montgomery-Smith, 2001). We demonstrate that the techniques used in this paper can be applied to deduce concentration inequalities for all distributions with finite variance.

Our main results, Theorem 3.1 and Theorem 3.2 in Section 3.1, provide general frameworks that are applicable to all finite variance distributions. To illustrate the strength of our frameworks, we present applications to sub-exponential, sub-Weibull, and polynomially decaying tails, respectively. These results are detailed in Section 3.2. Moreover, we obtain refined concentration results in Section 3.3 based on an asymptotic argument. To demonstrate the application of these derived concentration inequalities, we apply them to some standard problems in statistical learning: vector-valued concentration, Rademacher complexity, and algorithmic stability. These applications follow the work (Maurer & Pontil, 2021). Based on the bounded difference inequality, Rademacher complexity and algorithmic stability are two fundamental tools for deducing generalization bounds for various learning problems. However, when deriving high probability bounds, these tools typically require boundedness of loss functions. As a result, our inequalities facilitate the extension of these results to heavy-tailed distributions with finite variance.

This paper is organized as follows. We first introduce the preliminaries relevant to our discussion in Section 2. Section 3 presents the main results, detailing the concentration inequalities and their discussions. Section 4 is devoted to applications: vector valued concentration, Rademacher complexity and generalization, and algorithmic stability and generalization. We conclude this paper in Section 5. All proofs are postponed to the Appendix.

2. Preliminaries

Conventionally, we use uppercase letters to present random variables and vector of random variables, and use lowercase letters to present scalars and vector of scalars. Let $X = (X_1, ..., X_n)$ be a vector of independent random variables with values in a space $\mathcal{X}$, and the vector $X' = (X'_1, ..., X'_n)$ is independent and identically distributed (i.i.d.) to X. 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 its expectation, i.e.,

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

To proceed, we need the following definition to characterize the fluctuations of f in the k-th variable X_k, when the other variables $(x_i : i \neq k)$ are given.

Definition 2.1. 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 variable

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

Then $f_k(X)$ is a random-variable-valued-function $f_k(X) : x \in \mathcal{X}^n \to f_k(X)(x)$, which does not depend on the k-th coordinate of x. And

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

The expectation in $f_k(X)(x)$ and $f_k(X)(X)$ is taken on the k-th random variable of f since we are interested in its centered conditional version. Also, consider the summation case $f(x) = \sum_{i=1}^n x_i$, then $f_k(X)(x) = X_k - \mathbb{E}[X_k]$ is independent of x.

We then introduce some notations relevant to the heavy-tailed random variable. The following definition describes the tail property of a random variable.

Definition 2.2. Let $h : \mathbb{R} \to \mathbb{R}^{\geq 0}$ be an increasing and continuous function. We say h captures the right tail of random variable Z if

$$\mathbb{P}(Z > t) \leq \exp(-h(t)), \quad \forall t > 0.$$

The function $h(t)$ is generic. We will describe our main results in terms of this generic function and will be interested in some concrete tails: sub-exponential tail ($h(t) = ct$ for some fixed coefficient c), sub-Weibull tail ($h(t) = c_\theta t^{\frac{1}{\theta}}$ for some $\theta \geq 1$), and polynomially decaying tail ($h(t) = c \log t$ such that $c > 2$). In the sequel, for any random variable Z, we use the notation Z^τ to present its truncated version, i.e.,

$$Z^\tau = Z\mathbb{I}(Z \leq \tau), \quad \tau > 0.$$

3. Results

In this section, we show concentration inequalities for general functions of heavy-tailed random variables. We first give general frameworks in Section 3.1 and then apply them to some popular heavy-tailed distributions in Section 3.2. Finally, we will further provide refined results in Section 3.3 based on an asymptotic argument.

3.1. Main Results

Let $f_k^\tau(X)(x) = f_k(X)(x)\mathbb{I}(f_k(X)(x) \leq \tau)$. The first result is a general framework.

Theorem 3.1. *Suppose that the right tail of $f_k(X)(x)$ for all $x \in \mathcal{X}^n$ and any $k = 1, ..., n$ is captured by $h(t)$ as defined in Definition 2.2. Let $\eta \in (0, 1]$ and $\beta = \frac{\eta h(\tau)}{\tau}$, we define*

$$\Lambda(\tau,\eta) \triangleq \sup_{x\in\mathcal{X}^n} \mathbb{E}\left[(f_k^\tau(X)(x))^2\,\mathbb{I}(f_k^\tau(X)(x) \leq 0)\right] + \mathbb{E}\left[(f_k^\tau(X)(x))^2\, e^{\beta f_k^\tau(X)(x)}\mathbb{I}(0 < f_k^\tau(X)(x))\right].$$

Further, we define $t(\eta) \triangleq \sup\left\{t \geq 0 : t \leq \frac{\eta h(t)}{t} n\Lambda(t,\eta)\right\}$. Then, (1) if $t \geq t(\eta)$, we have

$$\mathbb{P}\left(f(X) - \mathbb{E}[f(X')] > t\right) \leq \exp\left(-c_t \eta h(t)\right) + n\exp(-h(t)),$$

where $c_t = \left(1 - \frac{1}{2t}\frac{\eta h(t)}{t} n\Lambda(t,\eta)\right) \in [\frac{1}{2}, 1)$;

(2) if $0 \leq t < t(\eta)$, we have

$$\mathbb{P}\left(f(X) - \mathbb{E}[f(X')] > t\right) \leq \exp\left(-\frac{t^2}{2n\Lambda(t(\eta),\eta)}\right) + n\exp\left(-\frac{t(\eta)^2}{\eta n\Lambda(t(\eta),\eta)}\right).$$

We give some remarks.

(1) Theorem 3.1 assumes that $f_k(X)(x)$ has a uniform tail captured by $h(t)$ for all x simultaneously, which follows the work (Maurer & Pontil, 2021). Specifically, for $f_k(X)(x)$ being sub-exponential, Theorem 4 in (Maurer & Pontil, 2021) shows

$$\mathbb{P}\left(f(X) - \mathbb{E}f(X') > t\right) \leq \exp\left(\frac{-t^2}{4e^2A + 2eBt}\right), \quad (1)$$

where $A = \sup_{x\in\mathcal{X}^n}\sum_{k=1}^n \|f_k(X)(x)\|_{\psi_1}^2$, $B = \max_k \sup_{x\in\mathcal{X}^n}\|f_k(X)(x)\|_{\psi_1}$ and $\|\cdot\|_{\psi_1}$ is the sub-exponential norm. This inequality is built on the supremum $\sup_{x\in\mathcal{X}^n}$ of $f_k(X)(x)$, which means assuming a uniform tail on $f_k(X)(x)$ for all x simultaneously. Since this paper aims to extend the sub-Gaussian and sub-exponential distributions in (Maurer & Pontil, 2021) to heavy-tailed ones, we adhere to similar assumptions for consistency. Similarly, we assume that $\Lambda(\tau,\eta)$ is constant for all $x \in \mathcal{X}^n$.

(2) Theorem 3.1 is an unbounded version of the bounded difference inequality (McDiarmid, 1998), featuring a mixture of two tails, a sub-Gaussian tail for small deviations, which is expected from the central limit theorem, and a heavy tail of magnitude $O(\exp(-ch(t)))$ for large deviations, where $c > 0$ is a constant, which is expected from the right tail of $f_k(X)(x)$. The transition between the two different regimes occurs at a threshold $t(\eta)$. Given that $h(t)$ is a generic function, the result is general enough to derive concrete concentration inequalities for distributions with finite variance.

(3) To derive concrete concentration inequalities, one needs to get an upper bound for $\Lambda(\tau,\eta)$, denoted by $\bar{\Lambda}(\tau,\eta)$. It is noteworthy that $\Lambda(\tau,\eta)$ can be substituted by $\bar{\Lambda}(\tau,\eta)$ for all values of τ, including $\tau = t$ and $\tau = t(\eta)$. One should then set $\tau = t$ and identify the region $t \leq \frac{\eta h(t)}{t} n\bar{\Lambda}(t,\eta)$ for $t \geq 0$, and subsequently find its supremum $t(\eta)$. Now, applying inequalities from Theorem 3.1 and replacing $\Lambda(t(\eta),\eta)$ with $\bar{\Lambda}(t(\eta),\eta)$ yields concrete concentration inequalities.

(4) The justification for substituting $\Lambda(t,\eta)$ with an upper bound $\bar{\Lambda}(t,\eta)$ is that the proof of Theorem 3.1 accommodates it. The proof holds by using an upper bound of $\Lambda(t,\eta)$. In other words, if $\Lambda(t,\eta) \leq \bar{\Lambda}(t,\eta)$, Theorem 3.1 remains valid by replacing $\Lambda(t,\eta)$ with $\bar{\Lambda}(t,\eta)$ in the definition of $t(\eta)$ and the coefficients appeared in the inequalities.

(5) Considering a simplified case where $\Lambda(\tau,\eta) \leq c$, the inequality is established by replacing $\Lambda(\tau,\eta)$ with c. In Section 3.3, we will show that $\Lambda(\tau,\eta)$ converges to the variance $\sup_{x\in\mathcal{X}^n}\mathbb{E}[(f_k(X)(x))^2]$ if τ grows to infinity. In this case, the inequality is constructed by substituting $\Lambda(\tau,\eta)$ with this variance.

(6) A limitation of Theorem 3.1 is its inability to provide the optimal rate for Gaussian variables. For a better discussion, we mention concentration results for Lipschitz functions (Vershynin, 2018). If $X_1, ..., X_n$ are independent random variables, each bounded on $[a, b]$, and let $f : \mathbb{R}^n \to \mathbb{R}$ is convex and L-Lipschitz with respect to the Euclidean norm, then for all $t \geq 0$

$$\mathbb{P}\left(|f(X) - \mathbb{E}[f(X)]| > t\right) \leq 2\exp\left(-\frac{t^2}{2L^2(b-a)^2}\right).$$

This result is truly remarkable since the concentration holds for a quantity independent of n. Note that the convexity assumption cannot be dropped in general; see (Ledoux & Talagrand, 2013), pp17. However, if X_i are distributed normally, we no longer need the convexity assumption, resulting in the following concentration: let $X_1, ..., X_n$ be independent random variables each distributed $\mathcal{N}(0, 1)$, and let $f : \mathbb{R}^n \to \mathbb{R}$ be L-Lipschitz with respect to the Euclidean

norm, then for all $t \geq 0$

$$\mathbb{P}\left(|f(X) - \mathbb{E}[f(X)]| > t\right) \leq 2\exp\left(-\frac{t^2}{2L^2}\right).$$

This bound illustrates that the concentration properties of Lipschitz functions of Gaussian variables exhibit a particularly attractive form of dimension-free concentration. Looking back at our inequalities. When $h(t) = t^2$ (sub-Gaussian tail), $\Lambda(\tau, \eta)$ is bounded by a constant, and $t(\eta)$ is infinity. Thus for any $t \geq 0$ we have the following inequality

$$\begin{aligned}&\mathbb{P}\left(f(X) - \mathbb{E}[f(X')] > t\right)\\ \leq &\exp\left(-\frac{t^2}{2n\Lambda(t(\eta), \eta)}\right) + n\exp\left(-\frac{t(\eta)^2}{\eta n\Lambda(t(\eta), \eta)}\right).\end{aligned}$$

The analysis in Section 3.3 demonstrates that if $t(\eta)$ grows to infinity, $\Lambda(t(\eta), \eta)$ converges to the variance $\sup_{x\in\mathcal{X}^n}\mathbb{E}[(f_k(X)(x))^2]$. In this context, we have

$$\begin{aligned}&\mathbb{P}\left(f(X) - \mathbb{E}[f(X')] > t\right)\\ \leq &2\exp\left(-\frac{t^2}{2n\sup_{x\in\mathcal{X}^n}\mathbb{E}[(f_k(X)(x))^2]}\right),\end{aligned}$$

which introduces a dimension-dependent bound. By comparison, our bound gives a weaker result, thereby not providing the optimal rate for Gaussian variables. However, our inequality, as formulated in Theorem 3.1, does not require the Lipschitz condition, offering broader applicability. Moreover, the dependency on the dimension n is inherited from the drawback of the bounded difference-type inequality.

(7) Another limitation of Theorem 3.1 is its inability to handle distributions with infinite variance. Our proof requires bounding the term $\Lambda(\tau, \eta)$, which excludes the infinite variance. To address scenarios with infinite variance, additional methodologies may be necessary. Here, we would like to highlight the significance of the infinite variance setting. Distributions with infinite variance are characterized by heavier tails and have broader applicability. We believe that Φ-entropies and moment inequalities (Boucheron et al., 2013) could potentially serve as powerful tools to move to the infinite variance setting.

Theorem 3.1 provides the best possible result implied by our analysis. We can also obtain a single inequality by adding the inequalities for each regime.

Theorem 3.2. *Let $f_k(X)(x)$, η, β and $\Lambda(\tau, \eta)$ be as in Theorem 3.1. Then we have*

$$\begin{aligned}&\mathbb{P}\left(f(X) - \mathbb{E}[f(X')] > t\right) \leq \exp\left(-\frac{t^2}{2n\Lambda(t, \eta)}\right)\\ &+ \exp\left(-\max\left\{c_t, \frac{1}{2}\right\}\eta h(t)\right) + n\exp(-h(t)),\end{aligned}$$

where $c_t = 1 - \frac{\eta h(t)}{2t^2}n\Lambda(t, \eta)$.

3.2. Heavy-tailed Distributions

To illustrate the strength of our main results, we apply them to some popular distributions, including sub-exponential, sub-Weibull, and heavier polynomially distributions.

The first result assumes that the right tail of $f_k(X)(x)$ is sub-exponential.

Theorem 3.3. *Suppose that the right tail of $f_k(X)(x)$ for all $x \in \mathcal{X}^n$ and any $k = 1, ..., n$ is captured by $h(t) = ct$ for some fixed coefficient c. Assume $\sup_{x\in\mathcal{X}^n}\mathbb{E}[(f_k(X)(x))^2\mathbb{I}(f_k(X)(x) \leq 0)] = \sigma_-^2 < \infty$. Let $\eta \in (0, 1)$, we have*

$$\Lambda(\tau, \eta) \leq \frac{2}{(1-\eta)^3c^2} + \sigma_-^2 = \bar{\Lambda}(\tau, \eta).$$

Let $t(\eta) = \eta cn\bar{\Lambda}(\tau, \eta)$. Then, (1) if $t \geq t(\eta)$, we have

$$\begin{aligned}&\mathbb{P}\left(f(X) - \mathbb{E}[f(X')] > t\right)\\ \leq &\exp\left(-c_t\eta ct\right) + n\exp(-ct),\end{aligned}$$

where $c_t = 1 - \frac{n\eta c\bar{\Lambda}(\tau,\eta)}{2t}$; (2) if $0 \leq t < t(\eta)$, we have

$$\begin{aligned}&\mathbb{P}\left(f(X) - \mathbb{E}[f(X')] > t\right)\\ \leq &\exp\left(-\frac{t^2}{2n\bar{\Lambda}(\tau, \eta)}\right) + n\exp\left(-\frac{t(\eta)^2}{\eta n\bar{\Lambda}(\tau, \eta)}\right).\end{aligned}$$

We give some remarks.

(1) Both the bound $\bar{\Lambda}(\tau, \eta)$ and the inequalities exclude the truncated parameter τ, highlighting the benign property of sub-exponential random variables. We explore the sub-exponential tail since it is relatively simple and intuitive. Theorem 3.3 effectively illustrates how to apply Theorem 3.1 to concrete tails.

(2) We present a user-friendly corollary.

Corollary 3.4. *Let $f_k(X)(x)$ and $h(t)$ be as in Theorem 3.3. Assume $\sup_{x\in\mathcal{X}^n}\mathbb{E}[(f_k(X)(x))^2\mathbb{I}(f_k(X)(x) \leq 0)] = \sigma_-^2 < \infty$. Let $\alpha = \frac{16}{c^2} + \sigma_-^2$ and $t(\eta) = \frac{cn}{2}\alpha$. Then, if $t \geq t(\eta)$, we have*

$$\mathbb{P}\left(f(X) - \mathbb{E}[f(X')] > t\right) \leq 2\exp\left(-\frac{1}{4}ct\right);$$

if $0 \leq t < t(\eta)$, when $n \geq \frac{8\log n}{3c^2\alpha}$, we have

$$\mathbb{P}\left(f(X) - \mathbb{E}[f(X')] > t\right) \leq 2\exp\left(-\frac{t^2}{2n\alpha}\right).$$

Given n grows much faster than $\log n$, the condition $n \geq \frac{8\log n}{3c^2\alpha}$ is readily met. Indeed, it is easy to verify that $\sigma_-^2 \leq 2c^{-2}$ for $h(t) = ct$. Plugging this bound into α, we see that the condition $n \geq \frac{8\log n}{3\times 18}$ is always satisfied.

(3) For bounded difference-type inequalities of sub-exponential random variables, a related result is Theorem 4 in (Maurer & Pontil, 2021), as shown in (1). As a comparison, the proof techniques used are different. Our inequalities also allow a fine-grained analysis where the right tail of $f_k(X)(x)$ is assumed to be captured by $h(t) = c_k t$ for any $k = 1, ..., n$. In comparison, our proof employs truncation techniques, which are not utilized in (Maurer & Pontil, 2021). We then examine the similarities between the two bounds. Specifically, given the equivalent properties of sub-exponential variables regarding their tails and moments (refer to Proposition 2.7.1 in (Vershynin, 2018)), both the two bounds exhibit a sub-Gaussian tail governed by the variance proxy $\frac{n}{c^2}$ for small deviations, and a sub-exponential tail governed by the scale-proxy $\frac{1}{c}$ for large deviations. Moreover, the transition between the two different regimes occurs at a similar threshold $\frac{n}{c}$.

(4) If f is a sum, i.e., $f(x) = \sum_{i=1}^n x_i$, we recover Bernstein's inequality for sub-exponential random variables, referring to Theorem 2.8.1 in (Vershynin, 2018) and Proposition 2.9 in (Wainwright, 2019).

The next result assumes that the right tail of $f_k(X)(x)$ is sub-Weibull.

Theorem 3.5. *Suppose that the right tail of $f_k(X)(x)$ for all $x \in \mathcal{X}^n$ and any $k = 1, ..., n$ is captured by $h(t) = c_\theta t^{\frac{1}{\theta}}$ for some $\theta \geq 1$. Assume $\sup_{x\in\mathcal{X}^n} \mathbb{E}[(f_k(X)(x))^2 \mathbb{I}(f_k(X)(x) \leq 0)] = \sigma_-^2 < \infty$. Let $\eta \in (0, 1)$ and $\beta = \frac{\eta c_\theta \tau^{\frac{1}{\theta}}}{\tau}$, we have*

$$\Lambda(\tau, \eta) \leq \frac{\Gamma(2\theta+1)}{((1-\eta)c_\theta)^{2\theta}} + \frac{\eta c_\theta \tau^{\frac{1}{\theta}} \Gamma(3\theta+1)}{3\tau((1-\eta)c_\theta)^{3\theta}} + \sigma_-^2 = \bar{\Lambda}(\tau, \eta).$$

Let $t(\eta) = \left(\eta c_\theta n \bar{\Lambda}(t, \eta)\right)^{\frac{\theta}{2\theta-1}}$. Then, one can plug $t(\eta)$ and $\bar{\Lambda}(t(\eta), \eta)$ into Theorem 3.1 to get the concentration inequalities. Furthermore, using the bound in Theorem 3.2 we get

$$\mathbb{P}\left(f(X) - \mathbb{E}[f(X')] > t\right) \leq \exp\left(-\frac{t^2}{2n\bar{\Lambda}(t,\eta)}\right) + \exp\left(-\max\left\{c_t, \frac{1}{2}\right\}\eta c_\theta t^{\frac{1}{\theta}}\right) + n\exp(-c_\theta t^{\frac{1}{\theta}}),$$

where $c_t = 1 - \frac{\eta c_\theta t^{\frac{1}{\theta}}}{2t^2} n\bar{\Lambda}(t, \eta)$.

We give some remarks.

(1) Sub-Weibull distributions are parameterized by a positive tail index θ and reduced to sub-Gaussian distributions for $\theta = 1/2$ and to sub-exponential distributions for $\theta = 1$. A higher tail parameter θ indicates a heavier tail. The MGF of sub-Weibull distributions becomes unbounded when $\theta > 1$, posing serious challenges in deducing concentration inequalities for this distribution. For more details of sub-Weibull distributions, please refer to (Vladimirova et al., 2020; Kuchibhotla & Chakrabortty, 2018; Bong & Kuchibhotla, 2023).

(2) In Theorem 3.5, the upper bound of $\Lambda(\tau, \eta)$ includes the truncated parameter τ, contrasting sharply with the sub-exponential case.

(3) We provide a user-friendly corollary.

Corollary 3.6. *Let $f_k(X)(x)$ and $h(t)$ be as in Theorem 3.5. Assume $\sup_{x\in\mathcal{X}^n} \mathbb{E}[(f_k(X)(x))^2 \mathbb{I}(f_k(X)(x) \leq 0)] = \sigma_-^2 < \infty$. Let $\alpha = \frac{\Gamma(2\theta+1)2^{2\theta}}{c_\theta^{2\theta}} + \frac{c_\theta c_1^{1/\theta}\Gamma(3\theta+1)2^{3\theta}}{6c_1 c_\theta^{3\theta}} + \sigma_-^2$ and $t(\eta) = (\frac{1}{2}c_\theta n\alpha)^{\frac{\theta}{2\theta-1}}$, where c_1 is a positive constant. Then, if $t \geq t(\eta)$, when $n^{\frac{1}{2\theta-1}} \geq \frac{4}{3c_\theta}(\frac{1}{2}c_\theta\alpha)^{\frac{1}{1-2\theta}} \log n$, we have*

$$\mathbb{P}\left(f(X) - \mathbb{E}[f(X')] > t\right) \leq 2\exp\left(-\frac{1}{4}c_\theta t^{\frac{1}{\theta}}\right);$$

if $0 \leq t < t(\eta)$, when $n^{\frac{1}{2\theta-1}} \geq \frac{8}{3}\alpha(\frac{1}{2}c_\theta\alpha)^{\frac{2\theta}{1-2\theta}} \log n$, we have

$$\mathbb{P}\left(f(X) - \mathbb{E}[f(X')] > t\right) \leq 2\exp\left(-\frac{t^2}{2n\alpha}\right).$$

Given $n^{\frac{1}{2\theta-1}}$ grows faster than $\log n$, the conditions $n^{\frac{1}{2\theta-1}} \geq \frac{4}{3c_\theta}(\frac{1}{2}c_\theta\alpha)^{\frac{1}{1-2\theta}} \log n$ and $n^{\frac{1}{2\theta-1}} \geq \frac{8}{3}\alpha(\frac{1}{2}c_\theta\alpha)^{\frac{2\theta}{1-2\theta}} \log n$ are easily met. It is evident that the bounds delineate two distinct regimes: the sub-Gaussian and the sub-Weibull tails.

(4) For bounded difference-type inequalities of sub-Weibull random variables, we have not found results comparable to Theorem 3.5 in the literature. Concentration inequalities for independent sub-Weibull random variables are primarily devoted to the sum, referring to (Kuchibhotla & Chakrabortty, 2018; Zhang & Wei, 2021; Bakhshizadeh et al., 2023; Bong & Kuchibhotla, 2023). Compared with these bounds, our inequalities exhibit a similar threshold $n^{\frac{\theta}{2\theta-1}}$ between the two different regimes (sub-Gaussian and sub-Weibull tail), as seen in Theorem 1.(c) in (Zhang & Wei, 2021), Theorem 3.1 in (Kuchibhotla & Chakrabortty, 2018), Corollary 2 in (Bakhshizadeh et al., 2023), and Theorem 2.3 in (Bong & Kuchibhotla, 2023).

In the last, we assume that the right tail of $f_k(X)(x)$ is polynomially decaying.

Theorem 3.7. *Suppose that the right tail of $f_k(X)(x)$ for all $x \in \mathcal{X}^n$ any $k = 1, ..., n$ is captured by $h(t) = c \log t$ such that $c > 2$. Assume $\sup_{x\in\mathcal{X}^n} \mathbb{E}[(f_k(X)(x))^2 \mathbb{I}(f_k(X)(x) \leq 0)] = \sigma_-^2 < \infty$.*

Let $\eta \in (0,1)$ and $\beta = \frac{\eta c \log\tau}{\tau}$, we have (1) if $\eta \neq 1 - \frac{2}{c}$

$$\Lambda(\tau,\eta) \leq \frac{2+\eta c\big(-\frac{1}{2-(1-\eta)c}\big)}{2-(1-\eta)c}\Big(\tau^{2-(1-\eta)c}-1\Big) + \frac{\tau^{2-(1-\eta)c}\eta c\log\tau}{2-(1-\eta)c} + \tau^{c\eta\tau^{-1}} + \sigma_-^2 = \bar{\Lambda}(\tau,\eta).$$

(2) if $\eta = 1 - \frac{2}{c}$,

$$\Lambda(\tau,\eta) \leq 2\log\tau + \frac{(c-2)(\log\tau)^2}{2} + \tau^{c\eta\tau^{-1}} + \sigma_-^2 = \bar{\Lambda}(\tau,\eta).$$

Then, one can plug $\bar{\Lambda}(\tau,\eta)$ into Theorem 3.1 to get the concentration inequalities. Furthermore, using the bound in Theorem 3.2 we get

$$\mathbb{P}\left(f(X) - \mathbb{E}[f(X')] > t\right) \leq \exp\left(-\frac{t^2}{2n\bar{\Lambda}(t,\eta)}\right) + \frac{n}{t^{\eta c\max\{c_t,0.5\}}} + \frac{n}{t^c},$$

where $c_t = 1 - \frac{\eta c\log t}{2t^2} n\bar{\Lambda}(t,\eta)$.

We give some remarks.

(1) The polynomially decaying tails are heavier than the sub-Weibull ones. Similar to discussions regarding the sub-exponential and sub-Weibull cases, a user-friendly corollary can be derived by following the same methodology.

(2) Due to the presence of terms $\tau^{2-(1-\eta)c}$ and $\frac{1}{2-(1-\eta)c}$, it is necessary for $2-(1-\eta)c < 0$, i.e., $c > \frac{2}{1-\eta} > 2$, to guarantee that the upper bound of $\Lambda(\tau,\eta)$ is finite. Under this condition, even if τ closes to infinity the bound of $\Lambda(\tau,\eta)$ remains bounded, and so is the concentration results. Proving concentration inequalities for independent variables with polynomially decaying tails when $c \leq 2$ remains an open problem.

(3) For bounded difference-type inequalities of polynomially decaying variables, we have not found results comparable to Theorem 3.7 in the literature. The bounds in Theorem 3.7 also delineate two distinct regimes: the sub-Gaussian tail and the polynomially decaying tail. A related result is Corollary 3 in (Bakhshizadeh et al., 2023), but it applies only to the sum of independent variables.

3.3. Refined Results

In this section, we give refined concentration results by considering enough large values of the truncated parameter τ, which is motivated by Bakhshizadeh et al. (2023). Our first result studies the sub-Weibull tail.

Theorem 3.8. *Suppose that the right tail of $f_k(X)(x)$ for all $x \in \mathcal{X}^n$ and any $k = 1,...,n$ is captured by $h(t) = c_\theta t^{\frac{1}{\theta}}$ for some $\theta > 1$. Let $\sup_{x\in\mathcal{X}^n}\mathbb{E}[(f_k(X)(x))^2] = \sigma^2$. Then for any $\eta \in (0,1)$ and $\delta > 0$, we have $t(\eta) = \big(\eta c_\theta n(\sigma^2+\delta)\big)^{\frac{\theta}{2\theta-1}}$, and there exists a positive constant c_δ such that for any $t \geq c_\delta$,*

$$\mathbb{P}\left(f(X) - \mathbb{E}[f(X')] > t\right) \leq \begin{cases} \exp\left(-c_t\eta c_\theta t^{\frac{1}{\theta}}\right) + n\exp\left(-c_\theta t^{\frac{1}{\theta}}\right) & \text{if } t > t(\eta), \\ \exp\left(-\frac{t^2}{2n(\sigma^2+\delta)}\right) + n\exp\left(-\frac{t(\eta)^2}{\eta n(\sigma^2+\delta)}\right) & \text{if } t \leq t(\eta), \end{cases}$$

where $c_t = \left(1 - \frac{1}{2t}\frac{\eta c_\theta t^{\frac{1}{\theta}}}{t} n(\sigma^2+\delta)\right) \in [\frac{1}{2},1)$.

We give some remarks.

(1) We focus on the concentration behavior for large values of the truncation parameter τ. The proof demonstrates that if τ grows to infinity, $\Lambda(\tau,\eta)$ converges to the variance $\sup_{x\in\mathcal{X}^n}\mathbb{E}[(f_k(X)(x))^2]$:

$$\lim_{\tau\to\infty}\mathbb{E}[(f_k^\tau(X)(x))^2\mathbb{I}(f_k^\tau(X)(x)\leq 0) + (f_k^\tau(X)(x))^2 e^{\beta f_k^\tau(X)(x)}\mathbb{I}(0 < f_k^\tau(X)(x) \leq \tau)] = \mathbb{E}\left[(f_k(X)(x))^2\right]$$

whenever $\beta \leq \frac{\eta h(\tau)}{\tau}$. This implies that if τ is very large, we get a tighter bound of $\Lambda(\tau,\eta)$. Both the bounds $\exp(-\frac{t^2}{2n(\sigma^2+\delta)}) + n\exp(-\frac{t(\eta)^2}{\eta n(\sigma^2+\delta)})$ and $\exp(-c_t\eta c_\theta t^{\frac{1}{\theta}}) + n\exp(-c_\theta t^{\frac{1}{\theta}})$ are very possible to be sharp, thus we conclude that a tighter bound for $\Lambda(\tau,\eta)$ leads to an accurate concentration result.

(2) Since the bound of $\Lambda(\tau,\eta)$ in Theorem 3.3 does not involve the parameter τ, the results for sub-exponential tails do not require a refinement.

The second result studies the polynomially decaying tail, with discussions that can follow Theorem 3.8.

Theorem 3.9. *Suppose that the right tail of $f_k(X)$ for all $x \in \mathcal{X}^n$ and any $k = 1,...,n$ is captured by $h(t) = c\log t$ for $c > 2$ and $\eta < 1 - \frac{2}{c}$. Let $\sup_{x\in\mathcal{X}^n}\mathbb{E}[(f_k(X)(x))^2] = \sigma^2$. Then for any $\eta \in (0,1)$ and $\delta > 0$, we have $t(\eta) = \frac{\eta c\log t(\eta)}{t(\eta)} n(\sigma^2+\delta)$, and there exists a positive constant c_δ such that for any $t \geq c_\delta$,*

$$\mathbb{P}\left(f(X) - \mathbb{E}[f(X')] > t\right) \leq \begin{cases} \exp\left(-c_t\eta c\log t\right) + n\exp\left(-c\log t\right) & \text{if } t > t(\eta), \\ \exp\left(-\frac{t^2}{2n(\sigma^2+\delta)}\right) + n\exp\left(-\frac{t(\eta)^2}{\eta n(\sigma^2+\delta)}\right) & \text{if } t \leq t(\eta), \end{cases}$$

where $c_t = \left(1 - \frac{1}{2t}\frac{\eta c\log t}{t} n(\sigma^2+\delta)\right) \in [\frac{1}{2},1)$.

We give a remark to discuss the proof technique of our results in Section 3.

The proof of our results in Section 3 mainly follows Bakhshizadeh et al. (2023), which studies the concentration for the sums of i.i.d. random variables with heavy-tailed distributions, while we study the concentration for general functions of heavy-tailed random variables. Specifically, our proof combines the entropy method used in (Maurer, 2012) and the truncation technique of random variables used in (Bakhshizadeh et al., 2023). We apply the entropy method to derive our concentration inequalities for functions of independent random variables and employ the truncation technique to manage heavy-tailed random variables. The subadditivity of the entropy method facilitates the tensorization of the total entropy, while the truncation technique allows for the continued use of the MGF on truncated variables.

4. Applications

To illustrate the application of these inequalities we give their use in vector valued concentration and two different methods to prove generalization bounds: Rademacher complexity and algorithmic stability. For conciseness, we focus mainly on applications of the two user-friendly corollaries. Applications of other tails can often be substituted by the reader by following the same pattern.

4.1. Vector Valued Concentration

Define the L_p-norm of a real-valued random variable Z as $\|Z\|_p = (\mathbb{E}[|Z|^p])^{1/p}$ for $p \geq 1$. We study concentration of vectors in a normed space $(\mathcal{X}, \|\cdot\|)$.

Theorem 4.1. *Suppose the X_i are i.i.d. random variables with values in a normed space $(\mathcal{X}, \|\cdot\|)$ and the right tails of the $\|X_i\|$ are captured by ct for some fixed coefficient c. Let $\alpha = \frac{16}{(c'c)^2} + 4\mathbb{E}[\|X_1\|^2]$ and $t(\eta) = \frac{c'cn}{2}\alpha$, where c' is an absolute positive constant. (i) Then if $t \geq t(\eta)$*

$$\mathbb{P}\left(\left\|\sum_{i=1}^n X_i\right\| - \mathbb{E}\left\|\sum_{i=1}^n X_i\right\| > t\right) \leq 2\exp\left(-\frac{1}{4}c'ct\right);$$

if $0 \leq t < t(\eta)$ and when $n \geq \frac{8\log n}{3(c'c)^2\alpha}$

$$\mathbb{P}\left(\left\|\sum_{i=1}^n X_i\right\| - \mathbb{E}\left\|\sum_{i=1}^n X_i\right\| > t\right) \leq 2\exp\left(-\frac{t^2}{2n\alpha}\right).$$

(ii) If $\mathcal{X}$ is a Hilbert space, then if $t \geq t(\eta)$

$$\mathbb{P}\left(\left\|\sum_{i=1}^n (X_i - \mathbb{E}X_1')\right\| > t + \sqrt{n}\|\|X_1\|\|_2\right) \leq 2\exp\left(-\frac{1}{4}c'ct\right);$$

if $0 \leq t < t(\eta)$ and when $n \geq \frac{8\log n}{3(c'c)^2\alpha}$

$$\mathbb{P}\left(\left\|\sum_{i=1}^n (X_i - \mathbb{E}X_1')\right\| > t + \sqrt{n}\|\|X_1\|\|_2\right) \leq 2\exp\left(-\frac{t^2}{2n\alpha}\right).$$

We compare Theorem 4.1 with the result of (Latała, 1997). Example 3.3 in (Latała, 1997) demonstrates that if X_i are independent symmetric random variables with logarithmically convex tails, i.e., $\mathbb{P}(|X_i| \geq t) = e^{-N(t)}$ for $t \geq 0$, where $N : \mathbb{R}_+ \to \mathbb{R}_+$ is a concave function, then for $p \geq 2$, it holds that $\|\sum_{i=1}^n X_i\|_p \leq c\left(p\sum \mathbb{E}X_i^2\right)^{1/2} + \left(\sum \mathbb{E}|X_i|^p\right)^{1/p}$, where c is a constant. Considering the case where $N(t) = t$, by homogeneity, we derive

$$\left\|\sum_{i=1}^n X_i\right\|_p \leq c'(\sqrt{p}\sqrt{n} + p),$$

where c' is a constant, see Corollary 1.2 in (Bogucki, 2015). Using Markov's inequality to transfer the moment to probability, we obtain

$$\mathbb{P}\left(\sum_{i=1}^n X_i \geq t\right) \leq \exp\left(-\min\left\{\frac{t^2}{4e^2c'^2n}, \frac{t}{2ec'}\right\}\right).$$

This bound exhibits a mixture of two tails, a sub-Gaussian tail $\exp\left(-\frac{t^2}{4e^2c'^2n}\right)$ and a sub-exponential tail $\exp\left(-\frac{t}{2ec'}\right)$. By comparison, our upper bound of $\mathbb{P}\left(\|\sum_{i=1}^n X_i\| - \mathbb{E}\|\sum_{i=1}^n X_i\| > t\right)$ similarly exhibits a mixture of a sub-Gaussian tail $2\exp\left(-\frac{t^2}{2n\alpha}\right)$ and a sub-exponential tail $2\exp\left(-\frac{1}{4}c'ct\right)$. The exact relationship between the constants in our bounds compared to those in the result of (Latała, 1997) remains to be clarified. However, given that our Theorem 4.1 is devised for vector-valued variables, the bound from (Latała, 1997) can be considered a one-dimensional instance of our results.

The second result studies the sub-Weibull tail.

Theorem 4.2. *Suppose the X_i are i.i.d. random variables with values in a normed space $(\mathcal{X}, \|\cdot\|)$ and the right tails of $\|X_i\|$ are captured by $h(t) = c_\theta t^{\frac{1}{\theta}}$ for some $\theta \geq 1$. Let $\alpha = \frac{\Gamma(2\theta+1)2^{2\theta}}{(c_\theta')^{2\theta}} + \frac{c_\theta' c_1^{1/\theta}\Gamma(3\theta+1)2^{3\theta}}{6c_1(c_\theta')^{3\theta}} + 4\mathbb{E}[\|X_1\|^2]$ and $t(\eta) = \left(\frac{1}{2}c_\theta' n\alpha\right)^{\frac{\theta}{2\theta-1}}$, where c_1 and c_θ' are two positive constants. (i) Then if $t \geq t(\eta)$ and when $n^{\frac{1}{2\theta-1}} \geq \frac{4}{3c_\theta'}\left(\frac{1}{2}c_\theta'\alpha\right)^{\frac{1}{1-2\theta}}\log n$*

$$\mathbb{P}\left(\left\|\sum_{i=1}^n X_i\right\| - \mathbb{E}\left\|\sum_{i=1}^n X_i\right\| > t\right) \leq 2\exp\left(-\frac{1}{4}c_\theta' t^{\frac{1}{\theta}}\right);$$

if $t \geq t(\eta)$ *and when* $n^{\frac{1}{2\theta-1}} \geq \frac{4}{3c'_\theta}(\frac{1}{2}c'_\theta\alpha)^{\frac{1}{1-2\theta}}\log n$,

$$\mathbb{P}\left(\left\|\sum_{i=1}^n X_i\right\| - \mathbb{E}\left\|\sum_{i=1}^n X_i\right\| > t\right) \leq 2\exp\left(-\frac{t^2}{2n\alpha}\right).$$

(ii) If $\mathcal{X}$ *is a Hilbert space, then if* $t \geq t(\eta)$ *and when* $n^{\frac{1}{2\theta-1}} \geq \frac{4}{3c'_\theta}(\frac{1}{2}c'_\theta\alpha)^{\frac{1}{1-2\theta}}\log n$

$$\begin{aligned}&\mathbb{P}\left(\left\|\sum_{i=1}^n (X_i - \mathbb{E}[X'_1])\right\| > t + \sqrt{n}|||X_1|||_2\right)\\ \leq& 2\exp\left(-\frac{1}{4}c'_\theta t^{\frac{1}{\theta}}\right);\end{aligned}$$

if $0 \leq t < t(\eta)$ *and when* $n^{\frac{1}{2\theta-1}} \geq \frac{8}{3}\alpha(\frac{1}{2}c'_\theta\alpha)^{\frac{2\theta}{1-2\theta}}\log n$

$$\begin{aligned}&\mathbb{P}\left(\left\|\sum_{i=1}^n (X_i - \mathbb{E}[X'_1])\right\| > t + \sqrt{n}|||X_1|||_2\right)\\ \leq& 2\exp\left(-\frac{t^2}{2n\alpha}\right).\end{aligned}$$

It is straightforward to obtain an upper bound for the second order moment $|||X_1|||_2$ using standard analysis. Vector valued concentration inequalities have broad applications in learning theory, which will also be used to prove generalization bounds in Section 4.2.

4.2. Rademacher Complexity

Rademacher complexity is a modern concept of complexity that is dependent on the distribution. Let's consider a class of functions $\mathcal{G}$, where each function $g : \mathcal{X} \to \mathbb{R}$. The Rademacher complexity of $\mathcal{G}$ is defined as:

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

where $\epsilon_1, ..., \epsilon_n$ are independent Rademacher variables such that $\mathbb{P}(\epsilon_i = 1) = \mathbb{P}(\epsilon_i = -1) = 1/2$. Together with the symmetrization argument, it leads to a uniform bound $\mathbb{E}\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_i g(X_i) - \mathbb{E}[g(X'_i)]\right] \leq 2\mathcal{R}(\mathcal{G})$. Based on the bounded difference inequality, the classical method in statistical learning shows that $\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_i g(X_i) - \mathbb{E}[g(X'_i)]$ is sharply concentrated around its mean $\mathbb{E}[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_i g(X_i) - \mathbb{E}[g(X'_i)]]$. This leads to the following generalization bound:

$$\begin{aligned}&\mathbb{P}\left(\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_i g(X_i) - \mathbb{E}[g(X'_i)] > 2\mathcal{R}(\mathcal{G}) + t\right)\\ \leq& \exp\left(-2nt^2\right), \quad t \geq 0,\end{aligned}$$

where the function g is assumed to satisfy $g : \mathcal{X} \to [0, 1]$. While fundamental, this approach necessitates that $g(X_i)$ be bounded random variables due to using the bounded difference inequality. However, deriving bounds on the Rademacher complexity does not necessarily require boundedness, and Lipschitz properties are more commonly assumed.

We will demonstrate that the boundedness can be relaxed by heavy-tailed distributions for uniformly Lipschitz function classes up to finite variance distributions. We illustrate this viewpoint with the sub-exponential and sub-Weibull tails.

Theorem 4.3. *Let* $X = (X_1, ..., X_n)$ *be a vector of i.i.d. random variables with values in a Banach space* $(\mathcal{X}, \|\cdot\|)$ *and let* $\mathcal{G}$ *be a class of function* $g : \mathcal{X} \to \mathbb{R}$ *such that* $g(x) - g(y) \leq L\|x - y\|$ *for all* $g \in \mathcal{G}$ *and that* $x, y \in \mathcal{X}$. *Suppose the right tails of the* $\|X_i\|$ *are captured by* ct *for some fixed coefficient* c. *Let* $\alpha = \frac{16}{(\frac{n}{L}c'c)^2} + \frac{4L^2}{n^2}\mathbb{E}[\|X_1\|^2]$ *and* $t(\eta) = \frac{c'cn^2}{2L}\alpha$, *where* c' *is an absolute positive constant. Then, (i) if* $t \geq t(\eta)$

$$\begin{aligned}&\mathbb{P}\Big(\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_i g(X_i) - \mathbb{E}[g(X'_i)] > 2\mathcal{R}(\mathcal{G}) + t\\ &\qquad + \frac{2L}{\sqrt{n}}|||X_1|||_2\Big) \leq 2\exp\left(-\frac{n}{4L}c'ct\right);\end{aligned}$$

(ii) if $0 \leq t < t(\eta)$ *and when* $n \geq \frac{8\log n}{3(\frac{n}{L}c'c)^2\alpha}$

$$\begin{aligned}&\mathbb{P}\Big(\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_i g(X_i) - \mathbb{E}[g(X'_i)] > 2\mathcal{R}(\mathcal{G}) + t\\ &\qquad + \frac{2L}{\sqrt{n}}|||X_1|||_2\Big) \leq 2\exp\left(-\frac{t^2}{2n\alpha}\right).\end{aligned}$$

The second result considers the sub-Weibull tail.

Theorem 4.4. *Let* $X = (X_1, ..., X_n)$ *be a vector of i.i.d. random variables with values in a Banach space* $(\mathcal{X}, \|\cdot\|)$ *and let* $\mathcal{G}$ *be a class of function* $g : \mathcal{X} \to \mathbb{R}$ *such that* $g(x) - g(y) \leq L\|x - y\|$ *for all* $g \in \mathcal{G}$ *and that* $x, y \in \mathcal{X}$. *Suppose the right tails of the* $\|X_i\|$ *are captured by* $h(t) = c_\theta t^{\frac{1}{\theta}}$ *for some* $\theta \geq 1$. *Let* $\alpha = \frac{\Gamma(2\theta+1)2^{2\theta}}{((\frac{n}{L})^{\frac{1}{\theta}}c'_\theta)^{2\theta}} + \frac{(\frac{n}{L})^{\frac{1}{\theta}}c'_\theta c_1^{1/\theta}\Gamma(3\theta+1)2^{3\theta}}{6c_1((\frac{n}{L})^{\frac{1}{\theta}}c'_\theta)^{3\theta}} + \frac{4L^2}{n^2}\mathbb{E}[\|X_1\|^2]$ *and* $t(\eta) = (\frac{1}{2}(\frac{n}{L})^{\frac{1}{\theta}}c'_\theta n\alpha)^{\frac{\theta}{2\theta-1}}$, *where* c_1 *and* c'_θ *are two positive constants. Then (i) if* $t \geq t(\eta)$ *and when* $n^{\frac{1}{2\theta-1}} \geq \frac{4}{3(\frac{n}{L})^{\frac{1}{\theta}}c'_\theta}(\frac{1}{2}(\frac{n}{L})^{\frac{1}{\theta}}c'_\theta\alpha)^{\frac{1}{1-2\theta}}\log n$

$$\begin{aligned}&\mathbb{P}\Big(\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_i g(X_i) - \mathbb{E}[g(X'_i)] > 2\mathcal{R}(\mathcal{G}) + t\\ &\qquad + \frac{2L}{\sqrt{n}}|||X_1|||_2\Big) \leq 2\exp\left(-\frac{1}{4}(\frac{n}{L})^{\frac{1}{\theta}}c'_\theta t^{\frac{1}{\theta}}\right);\end{aligned}$$

(ii) if $0 \leq t < t(\eta)$ *and when* $n^{\frac{1}{2\theta-1}} \geq$

$\frac{8}{3}\alpha(\frac{1}{2}(\frac{n}{L})^{\frac{1}{\theta}}c'_\theta\alpha)^{\frac{2\theta}{1-2\theta}}\log n$

$$\mathbb{P}\Big(\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_i g(X_i)-\mathbb{E}[g(X'_i)]>2\mathcal{R}(\mathcal{G})+t +\frac{2L}{\sqrt{n}}\|\|X_1\|\|_2\Big)\le 2\exp\left(-\frac{t^2}{2n\alpha}\right).$$

4.3. Algorithmic Stability

Algorithmic stability is gaining increasing attention in the generalization analysis of machine learning algorithms as this approach is beneficial to give dimension-free generalization bounds. Based on the bounded difference inequality and specific measures on the algorithmic stability, algorithmic stability demonstrates the sharp concentration of $f(X)-\mathbb{E}[f(X')]$ around its mean, yielding stability-based generalization bounds (Bousquet & Elisseeff, 2002). However, this method necessitates boundedness. Here, we review two related works that extend classical stability theory to unbounded cases. Suppose $(\mathcal{X},d,\mu)$ constitutes a metric probability space, and $X,X'\sim\mu$ are independent and identically distributed random variables with values in $\mathcal{X}$. (Kontorovich, 2014) examines the sub-Gaussian tail of $d(X,X')$. (Maurer & Pontil, 2021) further extend the approach of (Kontorovich, 2014) from sub-Gaussian to sub-exponential distributions. They operate with sub-Gaussian and sub-exponential norms defined respectively as $\|d(X,X')\|_{\psi_2}$ and $\|d(X,X')\|_{\psi_1}$ for independent $X',X\sim\mu$.

Our results build upon the methods of (Maurer & Pontil, 2021; Kontorovich, 2014), extending them to encompass all heavy-tailed distributions with finite variance. We illustrate this extension using the sub-exponential and sub-Weibull tails.

Theorem 4.5. *Let $X=(X_1,...,X_n)$ be a vector of i.i.d. random variables with values in $\mathcal{X}$ and let $f:\mathcal{X}^n\to\mathbb{R}$ have Lipschitz constant L with respect to the metric ρ on $\mathcal{X}^n$ defined by $\rho(x,y)=\sum_i d(x_i,y_i)$. Suppose the right tails of the $d(X_i,X'_i)$ are captured by ct for some fixed coefficient c. Let $\alpha=\frac{16}{(\frac{1}{L}c'c)^2}+L^2\mathbb{E}[(d(X_1,X'_1))^2]$ and $t(\eta)=\frac{c'cn}{2L}\alpha$, where c' is an absolute positive constant. Then, we have (i) if $t\ge t(\eta)$*

$$\mathbb{P}\left(f(X)-\mathbb{E}[f(X')]>t\right)\le 2\exp\left(-\frac{1}{4L}c'ct\right);$$

(ii) if $0\le t<t(\eta)$ and when $n\ge\frac{8\log n}{3(\frac{1}{L}c'c)^2\alpha}$

$$\mathbb{P}\left(f(X)-\mathbb{E}[f(X')]>t\right)\le 2\exp\left(-\frac{t^2}{2n\alpha}\right).$$

The second result considers the sub-Weibull tail.

Theorem 4.6. *Let $X=(X_1,...,X_n)$ be a vector of i.i.d. random variables with values in $\mathcal{X}$ and let $f:\mathcal{X}^n\to\mathbb{R}$ have Lipschitz constant L with respect to the metric ρ on $\mathcal{X}^n$ defined by $\rho(x,y)=\sum_i d(x_i,y_i)$. Suppose the right tails of the $d(X_i,X'_i)$ are captured by $h(t)=c_\theta t^{\frac{1}{\theta}}$ for some $\theta\ge 1$. Let $\alpha=\frac{\Gamma(2\theta+1)2^{2\theta}}{((\frac{1}{L})^{\frac{1}{\theta}}c'_\theta)^{2\theta}}+\frac{(\frac{1}{L})^{\frac{1}{\theta}}c'_\theta c_1^{1/\theta}\Gamma(3\theta+1)2^{3\theta}}{6c_1((\frac{1}{L})^{\frac{1}{\theta}}c'_\theta)^{3\theta}}+L^2\mathbb{E}[(d(X_1,X'_1))^2]$ and $t(\eta)=(\frac{1}{2}(\frac{1}{L})^{\frac{1}{\theta}}c'_\theta n\alpha)^{\frac{\theta}{2\theta-1}}$, where c_1 and c'_θ are two positive constants. Then we have (i) if $t\ge t(\eta)$ and when $n^{\frac{1}{2\theta-1}}\ge\frac{4}{3(\frac{1}{L})^{\frac{1}{\theta}}c'_\theta}(\frac{1}{2}(\frac{1}{L})^{\frac{1}{\theta}}c'_\theta\alpha)^{\frac{1}{1-2\theta}}\log n$*

$$\mathbb{P}\left(f(X)-\mathbb{E}[f(X')]>t\right)\le 2\exp\left(-\frac{1}{4}(\frac{1}{L})^{\frac{1}{\theta}}c'_\theta t^{\frac{1}{\theta}}\right);$$

(ii) if $0\le t<t(\eta)$ and when $n^{\frac{1}{2\theta-1}}\ge\frac{8}{3}\alpha(\frac{1}{2}(\frac{1}{L})^{\frac{1}{\theta}}c'_\theta\alpha)^{\frac{2\theta}{1-2\theta}}\log n$

$$\mathbb{P}\left(f(X)-\mathbb{E}[f(X')]>t\right)\le 2\exp\left(-\frac{t^2}{2n\alpha}\right).$$

Our results can be equally substituted to establish generalization bounds using the concept of total Lipschitz stability, just as in (Maurer & Pontil, 2021; Kontorovich, 2014).

5. Conclusion

In this paper, we presented bounded difference-type concentration inequalities for functions of heavy-tailed independent random variables. The results provided a probabilistic toolbox that can be employed to derive bounded difference-type concentration inequalities for a very large number of heavy-tailed distributions, which holds for all distributions with finite variance. We illustrated our concentration inequalities to several popular distributions. Applications to statistical learning theory are also provided.

It would be interesting to show more applications of these inequalities in future work.

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 paper presents work whose goal is to advance the field of statistical learning theory, and it does not present any foreseeable societal consequence.

References

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

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

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.

Bogucki, R. Suprema of canonical weibull processes. *Statistics & Probability Letters*, 107:253–263, 2015.

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.

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

Boucheron, S., Bousquet, O., Lugosi, G., and Massart, P. Moment inequalities for functions of independent random variables. *The Annals of Statistics*, 33(1):514–560, 2005.

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. *Journal of Machine Learning Research*, 2:499–526, 2002.

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

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

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

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

Hahn, M. G. and Klass, M. J. Approximation of partial sums of arbitrary iid random variables and the precision of the usual exponential upper bound. *The Annals of Probability*, 25(3):1451–1470, 1997.

Hitczenko, P. and Montgomery-Smith, S. Measuring the magnitude of sums of independent random variables. *The Annals of probability*, pp. 447–466, 2001.

Klass, M. and Nowicki, K. Uniformly accurate quantile bounds via the truncated moment generating function: the symmetric case. *Electronic Journal of Probability*, 12:1276–1298, 2007.

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. *arXiv preprint arXiv:1804.02605*, 2018.

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.

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

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

Ledoux, M. and Talagrand, M. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

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

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

Maurer, A. and Pontil, M. Empirical bounds for functions with weak interactions. In *Conference On Learning Theory*, pp. 987–1010, 2018.

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.

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.

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

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

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

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

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.

Wong, K. C., Li, Z., and Tewari, A. Lasso guarantees for β-mixing heavy-tailed time series. *The Annals of Statistics*, 48(2):1124–1142, 2020.

Zhang, H. and Wei, H. Sharper sub-weibull concentrations. *arXiv preprint arXiv:2102.02450*, 2021.

A. Proofs of Section 3

The following section collects a set of tools.

A.1. Some necessary tools

We introduce some necessary tools of the entropy method. 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 $\mathbb{E}_Z[Y] = \mathbb{E}[Ye^Z]/\mathbb{E}[e^Z]$. Besides, we have the following fluctuation representation of the entropy.

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

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

If $f : \mathcal{X}^n \to \mathbb{R}$ and X and the 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 $x \in \mathcal{X}^n$. The following lemma shows the sub-additivity of entropy, which states that the total entropy is no greater than the thermal average of the sum of the conditional entropies.

Lemma A.2. *(Maurer, 2012) The sub-additivity of entropy is*

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

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

Lemma A.3. *(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 $t \geq 0$,

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

A.2. Proof of Theorem 3.1

Proof. To begin, we define two events $A : f(X) - \mathbb{E}[f(X')] > t$ and $B : \exists k, f_k(X)(x) > \tau$, and we define $\bar{B}$ as the complement event of B. Then we have

$$\mathbb{P}(A) \leq P(A\bar{B}) + P(B),$$

that is,

$$\mathbb{P}(f(X) - \mathbb{E}[f(X')] > t) \leq \mathbb{P}(f(X) - \mathbb{E}[f(X')] > t \text{ and } \bar{B}) + \mathbb{P}(\exists k, f_k(X)(x) > \tau). \tag{2}$$

Firstly, let's focus on $\mathbb{P}(f(X) - \mathbb{E}[f(X')] > t \text{ and } \bar{B})$. Using the Markov's inequality, we obtain

$$\mathbb{P}(f(X) - \mathbb{E}[f(X')] > t \text{ and } \bar{B}) \leq \exp\left(\ln \mathbb{E}\left[e^{\beta(f(X) - \mathbb{E}[f(X')]) 1_{\bar{B}}} \right] - \beta t \right), \tag{3}$$

where the value of $1_{\bar{B}}$ is 1 if $\bar{B}$ holds true, and 0 otherwise. Together with Lemma A.3 and then with Lemma A.2, the bound in (3) implies

$$\begin{aligned}
&\mathbb{P}(f(X) - \mathbb{E}[f(X')] > t \text{ and } \bar{B}) \\
\leq &\exp\left(\beta \int_0^\beta \frac{S(\gamma(f(X))1_{\bar{B}})d\gamma}{\gamma^2} - \beta t\right) \\
\leq &\exp\left(\beta \int_0^\beta \frac{\mathbb{E}_{\gamma(f(X))1_{\bar{B}}}\left[\sum_k S(\gamma(f_k(X)(X))1_{\bar{B}})\right]d\gamma}{\gamma^2} - \beta t\right) \\
\leq &\exp\left(\beta \int_0^\beta \frac{\sum_k \sup_{x\in\mathcal{X}^n} S(\gamma f_k^\tau(X)(x))d\gamma}{\gamma^2} - \beta t\right),
\end{aligned} \tag{4}$$

where $f_k^\tau(X)(x) = f_k(X)(x)\mathbb{I}(f_k(X)(x) \leq \tau)$, the first inequality uses Lemma A.2, and the second uses Lemma A.2.

Next, we need to bound the conditional entropy $S(\gamma f_k^\tau(X)(x))$ for any $x \in \mathcal{X}^n$. According to Lemma A.1, we derive

$$\begin{aligned}
&\mathbb{E}_{sf_k^\tau(X)(x)}\left[\left(f_k^\tau(X)(x) - \mathbb{E}_{sf_k^\tau(X)(x)}\left[f_k^\tau(X)(x)\right]\right)^2\right] \leq \mathbb{E}_{sf_k^\tau(X)(x)}\left[\left(f_k^\tau(X)(x)\right)^2\right] = \frac{\mathbb{E}\left[(f_k^\tau(X)(x))^2 e^{sf_k^\tau(X)(x)}\right]}{\mathbb{E}\left[e^{sf_k^\tau(X)(x)}\right]} \\
=&\frac{\mathbb{E}\left[(f_k^\tau(X)(x))^2 e^{sf_k^\tau(X)(x)}\mathbb{I}(f_k^\tau(X)(x) \leq 0)\right]}{\mathbb{E}\left[e^{sf_k^\tau(X)(x)}\right]} + \frac{\mathbb{E}\left[(f_k^\tau(X)(x))^2 e^{sf_k^\tau(X)(x)}\mathbb{I}(f_k^\tau(X)(x) > 0)\right]}{\mathbb{E}\left[e^{sf_k^\tau(X)(x)}\right]},
\end{aligned} \tag{5}$$

where the first inequality follows from the variational property of the variance and the first identity follows from the definition of the entropy. Since $(f_k^\tau(X)(x))^2\mathbb{I}(f_k^\tau(X)(x) \leq 0)$ is a nonincreasing function and $e^{sf_k^\tau(X)(x)}$ is a nondecreasing function of $f_k^\tau(X)(x)$, Harris' inequality (Theorem 2.15 in (Boucheron et al., 2013)) implies that

$$\mathbb{E}\left[(f_k^\tau(X)(x))^2 e^{sf_k^\tau(X)(x)}\mathbb{I}(f_k^\tau(X)(x) \leq 0)\right] \leq \mathbb{E}\left[(f_k^\tau(X)(x))^2\mathbb{I}(f_k^\tau(X)(x) \leq 0)\right]\mathbb{E}\left[e^{sf_k^\tau(X)(x)}\right]. \tag{6}$$

And since $\mathbb{E}[e^{sf_k^\tau(X)(x)}] = \mathbb{E}[e^{sf_k^\tau(X)(x)}\mathbb{I}(f_k^\tau(X)(x) \leq 0)] + \mathbb{E}[e^{sf_k^\tau(X)(x)}\mathbb{I}(f_k^\tau(X)(x) > 0)]$, $e^{sf_k^\tau(X)(x)} > 0$ when $f_k^\tau(X)(x) \leq 0$, and $e^{sf_k^\tau(X)(x)} \geq 1$ when $f_k^\tau(X)(x) > 0$, this implies that

$$\begin{aligned}
&\frac{\mathbb{E}\left[(f_k^\tau(X)(x))^2 e^{sf_k^\tau(X)(x)}\mathbb{I}(f_k^\tau(X)(x) > 0)\right]}{\mathbb{E}\left[e^{sf_k^\tau(X)(x)}\right]} \leq \frac{\mathbb{E}\left[(f_k^\tau(X)(x))^2 e^{sf_k^\tau(X)(x)}\mathbb{I}(f_k^\tau(X)(x) > 0)\right]}{\mathbb{E}\left[e^{sf_k^\tau(X)(x)}\mathbb{I}(f_k^\tau(X)(x) > 0)\right]} \\
\leq &\mathbb{E}\left[(f_k^\tau(X)(x))^2 e^{sf_k^\tau(X)(x)}\mathbb{I}(f_k^\tau(X)(x) > 0)\right].
\end{aligned} \tag{7}$$

Combining (5), (6) and (7) we obtain

$$\mathbb{E}_{sf_k^\tau(X)(x)}\left[\left(f_k^\tau(X)(x)\right)^2\right] \leq \mathbb{E}\left[(f_k^\tau(X)(x))^2\mathbb{I}(f_k^\tau(X)(x) \leq 0)\right] + \mathbb{E}\left[(f_k^\tau(X)(x))^2 e^{sf_k^\tau(X)(x)}\mathbb{I}(f_k^\tau(X)(x) > 0)\right]. \tag{8}$$

In Lemma A.1 and Lemma A.3, we know that $s \leq \gamma \leq \beta$, and (8) becomes

$$\mathbb{E}_{sf_k^\tau(X)(x)}\left[\left(f_k^\tau(X)(x)\right)^2\right] \leq \mathbb{E}\left[\left(f_k^\tau(X)(x)\right)^2\mathbb{I}(f_k^\tau(X)(x) \leq 0)\right] + \mathbb{E}\left[\left(f_k^\tau(X)(x)\right)^2 e^{\beta f_k^\tau(X)(x)}\mathbb{I}(0 < f_k^\tau(X)(x))\right].$$

Define $\Lambda(\tau,\eta) \triangleq \sup_{x\in\mathcal{X}^n}\mathbb{E}[(f_k^\tau(X)(x))^2\mathbb{I}(f_k^\tau(X)(x) \leq 0)] + \mathbb{E}[(f_k^\tau(X)(x))^2 e^{\beta f_k^\tau(X)(x)}\mathbb{I}(f_k^\tau(X)(x) > 0)]$. Together with Lemma A.1 this gives the following entropy bound

$$S(\gamma f_k^\tau(X)(x)) \leq \int_0^\gamma \left(\int_t^\gamma \Lambda(\tau,\eta)ds\right)dt = \frac{\gamma^2}{2}\Lambda(\tau,\eta).$$

Plugging the upper bound of $S(\gamma f_k^\tau(X)(x))$ into (4) then gives

$$\mathbb{P}(f(X) - \mathbb{E}[f(X')] > t \text{ and } \bar{B}) \leq \exp\left(\beta \int_0^\beta \frac{\sum_k \frac{\gamma^2}{2}\Lambda(\tau,\eta)d\gamma}{\gamma^2} - \beta t\right) \leq \exp\left(\frac{n\beta^2}{2}\Lambda(\tau,\eta) - \beta t\right).$$

Combining this entropy bound and (2), we obtain

$$\begin{aligned}&\mathbb{P}\left(f(X)-\mathbb{E}[f(X')]>t\right)\\ \leq&\exp\left(\frac{n\beta^2}{2}\Lambda(\tau,\eta)-\beta t\right)+n\mathbb{P}(f_k(X)(x)>\tau)\\ \leq&\exp\left(\frac{n\beta^2}{2}\Lambda(\tau,\eta)-\beta t\right)+n\exp(-h(\tau)).\end{aligned}\tag{9}$$

In the following, we need to choose values for free parameters β and τ to get the best bound. We first consider $t>t(\eta)$. In this case, we select $\tau=t$ and $\beta=\frac{\eta h(t)}{t}=\frac{\eta h(\tau)}{\tau}$, then we get

$$\begin{aligned}&\mathbb{P}\left(f(X)-\mathbb{E}[f(X')]>t\right)\\ \leq&\exp\left(-\left(1-\frac{n\beta}{2t}\Lambda(\tau,\eta)\right)\beta t\right)+n\exp(-h(\tau))\\ =&\exp\left(-\left(1-\frac{n\eta h(t)}{2t^2}\Lambda(t,\eta)\right)\eta h(t)\right)+n\exp(-h(t)).\end{aligned}$$

Since $t>t(\eta)$, we have $t>\frac{\eta h(t)}{t}n\Lambda(t,\eta)$, which implies $1-\frac{n\eta h(t)}{2t^2}\Lambda(t,\eta)\in[\frac{1}{2},1)$. We then consider $t\leq t(\eta)$. In this case, we select $\tau=t(\eta)$ and $\beta=\frac{t}{n\Lambda(\tau,\eta)}\leq\frac{t(\eta)}{n\Lambda(\tau,\eta)}=\frac{\eta h(\tau)}{\tau}$, then we get

$$\begin{aligned}&\mathbb{P}\left(f(X)-\mathbb{E}[f(X')]>t\right)\\ \leq&\exp\left(-\frac{t^2}{2n\Lambda(\tau,\eta)}\right)+n\exp(-h(\tau))\\ =&\exp\left(-\frac{t^2}{2n\Lambda(\tau,\eta)}\right)+n\exp\left(-\frac{t(\eta)^2}{\eta n\Lambda(\tau,\eta)}\right)\\ =&\exp\left(-\frac{t^2}{2n\Lambda(t(\eta),\eta)}\right)+n\exp\left(-\frac{t(\eta)^2}{\eta n\Lambda(t(\eta),\eta)}\right),\end{aligned}$$

where the first inequality using the fact $\beta=\frac{t}{n\Lambda(\tau,\eta)}$, and where the first identity using $\frac{t(\eta)}{n\Lambda(\tau,\eta)}=\frac{\eta h(\tau)}{\tau}$ and $\tau=t(\eta)$. Combining the two cases, the proof is complete. □

A.3. Proof of Theorem 3.2

Proof. Let $\tau=t$. Using (9) and the fact that $\Lambda(\tau,\eta)$ is increasing in η we have

$$\mathbb{P}\left(f(X)-\mathbb{E}[f(X')]>t\right)\leq\exp\left(\frac{n\beta^2}{2}\Lambda(\tau,\eta)-\beta t\right)+n\exp(-h(\tau)),\quad\forall\beta\leq\frac{\eta h(t)}{t}.$$

To prove Theorem 3.2, we need to ensure that the following bound holds for $\beta\leq\frac{\eta h(t)}{t}$

$$\exp\left(\frac{n\beta^2}{2}\Lambda(\tau,\eta)-\beta t\right)\leq\exp\left(-\frac{t^2}{2n\Lambda(t,\eta)}\right)+\exp\left(-\max\left\{c_t,\frac{1}{2}\right\}\eta h(t)\right).$$

We consider two cases. When $c_t\geq\frac{1}{2}$, by selecting $\beta=\frac{\eta h(t)}{t}$, we obtain

$$\begin{aligned}&\exp\left(\frac{n\beta^2}{2}\Lambda(\tau,\eta)-\beta t\right)=\exp\left(-\left(1-\frac{n\beta}{2t}\Lambda(\tau,\eta)\right)\beta t\right)\\ =&\exp\left(-\left(1-\frac{n\eta h(t)}{2t^2}\Lambda(t,\eta)\right)\eta h(t)\right)=\exp\left(-c_t\eta h(t)\right)=\exp\left(-\max\left\{c_t,\frac{1}{2}\right\}\eta h(t)\right).\end{aligned}$$

When $c_t<\frac{1}{2}$, we get $\frac{t}{n\Lambda(t,\eta)}<\frac{\eta h(t)}{t}$. Selecting $\beta=\frac{t}{n\Lambda(t,\eta)}$, we then get

$$\exp\left(\frac{n\beta^2}{2}\Lambda(\tau,\eta)-\beta t\right)=\exp\left(-\frac{t^2}{2n\Lambda(t,\eta)}\right).$$

Combining the two cases, the proof is complete. □

A.4. Proof of Theorem 3.3

Proof. Given Theorem 3.1, we need to bound $\Lambda(\tau,\eta)$. Firstly, we have

$$\sup_{x\in\mathcal{X}^n}\mathbb{E}\left[\left(f_k^\tau(X)(x)\right)^2\mathbb{I}(f_k^\tau(X)(x)\le 0)\right]\le\sup_{x\in\mathcal{X}^n}\mathbb{E}\left[\left(f_k(X)(x)\right)^2\mathbb{I}(f_k(X)(x)\le 0)\right]=\sigma_-^2.$$

Then, we derive that for any $x\in\mathcal{X}^n$

$$\begin{aligned}
&\mathbb{E}\left[\left(f_k^\tau(X)(x)\right)^2 e^{\beta f_k^\tau(X)(x)}\mathbb{I}(0<f_k^\tau(X)(x)\le\tau)\right]\\
=&\int_0^\infty\mathbb{P}\left(\left(f_k^\tau(X)(x)\right)^2 e^{\beta f_k^\tau(X)(x)}\mathbb{I}(0<f_k^\tau(X)(x)\le\tau)>u\right)du\\
=&\int_0^\infty\mathbb{P}\left(\left(f_k^\tau(X)(x)\right)^2 e^{\beta f_k^\tau(X)(x)}>t^2e^{\beta t},0<f_k^\tau(X)(x)\le\tau\right)(2t+\beta t^2)e^{\beta t}dt\\
=&\int_0^\infty\mathbb{P}\left(|f_k^\tau(X)(x)|>t,0<f_k^\tau(X)(x)<\tau\right)(2t+\beta t^2)e^{\beta t}dt\\
=&\int_0^\tau\mathbb{P}\left(f_k^\tau(X)(x)>t\right)(2t+\beta t^2)e^{\beta t}dt\\
\le&\int_0^\tau\exp(-ct)(2t+\beta t^2)e^{\beta t}dt\\
=&\int_0^\tau\exp(-(1-\beta c^{-1})ct)(2t+\beta t^2)dt\\
\le&\int_0^\infty\exp(-u)\left(\frac{2u}{((1-\beta c^{-1})c)^2}+\frac{\beta u^2}{((1-\beta c^{-1})c)^3}\right)du\\
=&\frac{\beta\Gamma(3)}{((1-\beta c^{-1})c)^3}+\frac{2\Gamma(2)}{((1-\beta c^{-1})c)^2}=\frac{2}{(1-\beta c^{-1})^3c^2},
\end{aligned}$$

where $\Gamma(z)=\int_0^\infty t^{z-1}e^{-t}\mathrm{dt}$. Combining the two bounds, we obtain

$$\Lambda(\tau,\eta)\le\frac{2}{(1-\beta c^{-1})^3c^2}+\sigma_-^2.$$

Since $\beta=\eta c$, the above bound implies that $\Lambda(\tau,\eta)\le\frac{2}{(1-\eta)^3c^2}+\sigma_-^2$. Substituting this bound into Theorem 3.1, the proof is complete. □

A.5. Proof of Corollary 3.4

Proof. In Theorem 3.3, choosing $\eta=\frac{1}{2}$ gives $\bar{\Lambda}(\tau,\eta)=\frac{16}{c^2}+\sigma_-^2\triangleq\alpha$ and $t(\eta)=\frac{cn}{2}(\frac{16}{c^2}+\sigma_-^2)$. In the first case $t\ge t(\eta)$, we know $c_t\in[\frac{1}{2},1)$. Selecting the worst-case value $c_t=\frac{1}{2}$ we obtain

$$\mathbb{P}(f(X)-\mathbb{E}[f(X')]>t)\le\exp\left(-\frac{1}{4}ct\right)+n\exp(-ct).$$

Further, we derive

$$n\exp(-ct)=\exp(\log n-ct)\le\exp\left(-\frac{1}{4}ct\right)$$

whenever $\log n\le\frac{3}{4}ct$. Since $t\ge t(\eta)$, we have

$$\frac{3}{4}ct\ge 6n+\frac{3c^2n\sigma_-^2}{8}>\log n.$$

Therefore, we conclude $\mathbb{P}(f(X)-\mathbb{E}[f(X')]>t)\le 2\exp(-\frac{1}{4}ct)$ when $t>t(\eta)$. In the second case $t<t(\eta)$, we have

$$\mathbb{P}(f(X)-\mathbb{E}[f(X')]>t)\le\exp\left(-\frac{t^2}{2n\alpha}\right)+n\exp\left(-\frac{2t(\eta)^2}{n\alpha}\right)=\exp\left(-\frac{t^2}{2n\alpha}\right)+\exp\left(-\frac{2t(\eta)^2}{n\alpha}+\log n\right).$$

Due to $t < t(\eta)$, we derive

$$\frac{t^2}{2n\alpha} - \frac{2t(\eta)^2}{n\alpha} + \log n = \frac{t^2 - t(\eta)^2 - \frac{3}{4}c^2n^2\alpha^2 + 2n\alpha\log n}{2n\alpha} < \frac{-\frac{3}{4}c^2n^2\alpha^2 + 2n\alpha\log n}{2n\alpha} \leq 0$$

whenever $n \geq \frac{8\log n}{3c^2\alpha}$. In this case, the first term dominates the second, and thus we get

$$\mathbb{P}\left(f(X) - \mathbb{E}[f(X')] > t\right) \leq 2\exp\left(-\frac{t^2}{2n\alpha}\right).$$

The proof is complete. □

A.6. Proof of Theorem 3.5

Proof. Also, given Theorem 3.1, we need to bound $\Lambda(\tau, \eta)$. Firstly, we have

$$\sup_{x\in\mathcal{X}^n} \mathbb{E}\left[\left(f_k^\tau(X)(x)\right)^2 \mathbb{I}(f_k^\tau(X)(x) \leq 0)\right] \leq \sup_{x\in\mathcal{X}^n} \mathbb{E}\left[\left(f_k(X)(x)\right)^2 \mathbb{I}(f_k(X)(x) \leq 0)\right] = \sigma_-^2.$$

Then, following the proof of Theorem 3.3, we derive that for any $x \in \mathcal{X}^n$

$$\begin{aligned}
&\mathbb{E}\left[\left(f_k^\tau(X)(x)\right)^2 e^{\beta f_k^\tau(X)(x)} \mathbb{I}(0 < f_k^\tau(X)(x) \leq \tau)\right] \\
=&\int_0^\tau \mathbb{P}\left(f_k^\tau(X)(x) > t\right)(2t + \beta t^2)e^{\beta t}dt \\
\leq&\int_0^\tau \exp(-c_\theta t^{\frac{1}{\theta}})(2t + \beta t^2)e^{\beta t}dt \\
=&\int_0^\tau \exp(-(1 - \beta c_\theta^{-1}t^{1-\frac{1}{\theta}})c_\theta t^{\frac{1}{\theta}})(2t + \beta t^2)dt \\
\leq&\int_0^\infty \exp(-(1 - \beta c_\theta^{-1}\tau^{1-\frac{1}{\theta}})c_\theta t^{\frac{1}{\theta}})(2t + \beta t^2)dt \\
\leq&\int_0^\infty \exp(-u)\left(\frac{2\theta u^{2\theta-1}}{((1 - \beta c_\theta^{-1}\tau^{1-\frac{1}{\theta}})c_\theta)^{2\theta}} + \frac{\beta\theta u^{3\theta-1}}{((1 - \beta c_\theta^{-1}\tau^{1-\frac{1}{\theta}})c_\theta)^{3\theta}}\right)du \\
=&\frac{2\theta\Gamma(2\theta)}{((1 - \beta c_\theta^{-1}\tau^{1-\frac{1}{\theta}})c_\theta)^{2\theta}} + \frac{\beta\theta\Gamma(3\theta)}{((1 - \beta c_\theta^{-1}\tau^{1-\frac{1}{\theta}})c_\theta)^{3\theta}} \\
=&\frac{\Gamma(2\theta+1)}{((1 - \beta c_\theta^{-1}\tau^{1-\frac{1}{\theta}})c_\theta)^{2\theta}} + \frac{\beta\Gamma(3\theta+1)}{3((1 - \beta c_\theta^{-1}\tau^{1-\frac{1}{\theta}})c_\theta))^{3\theta}}.
\end{aligned}$$

Combining the two bounds gives

$$\Lambda(\tau, \eta) \leq \frac{\Gamma(2\theta+1)}{((1 - \beta c_\theta^{-1}\tau^{1-\frac{1}{\theta}})c_\theta)^{2\theta}} + \frac{\beta\Gamma(3\theta+1)}{3((1 - \beta c_\theta^{-1}\tau^{1-\frac{1}{\theta}})c_\theta)^{3\theta}} + \sigma_-^2.$$

Since $\beta = \frac{\eta c_\theta \tau^{\frac{1}{\theta}}}{\tau}$, the above bound implies that $\Lambda(\tau, \eta) \leq \frac{\Gamma(2\theta+1)}{((1-\eta)c_\theta)^{2\theta}} + \frac{\eta c_\theta \tau^{\frac{1}{\theta}}\Gamma(3\theta+1)}{3\tau((1-\eta)c_\theta)^{3\theta}} + \sigma_-^2$. Substituting this bound into Theorem 3.1, the proof is complete. □

A.7. Proof of Corollary 3.6

Proof. According to the bounds in Theorem 3.5, for all $\tau \geq c_1$ where c_1 is a positive constant depending only on the distribution of $f_k(X)(x)$, we have

$$\bar{\Lambda}(\tau, \eta) \leq \frac{\Gamma(2\theta+1)}{((1-\eta)c_\theta)^{2\theta}} + \frac{\eta c_\theta c_1^{1/\theta}\Gamma(3\theta+1)}{3c_1((1-\eta)c_\theta)^{3\theta}} + \sigma_-^2 \triangleq \alpha$$

and

$$t(\eta) = (\eta c_\theta n\alpha)^{\frac{\theta}{2\theta-1}}.$$

Then, choosing $\eta = \frac{1}{2}$ gives $\alpha = \frac{\Gamma(2\theta+1)2^{2\theta}}{c_\theta^{2\theta}} + \frac{c_\theta c_1^{1/\theta}\Gamma(3\theta+1)2^{3\theta}}{6c_1 c_\theta^{3\theta}} + \sigma_-^2$ and $t(\eta) = (\frac{1}{2}c_\theta n\alpha)^{\frac{\theta}{2\theta-1}}$. Similarly, in the first case $t \geq t(\eta)$, selecting the worst-case value $c_t = \frac{1}{2}$ we get

$$\mathbb{P}\left(f(X) - \mathbb{E}[f(X')] > t\right) \leq 2\exp\left(-\frac{1}{4}c_\theta t^{\frac{1}{\theta}}\right)$$

whenever $\log n \leq \frac{3}{4}c_\theta t^{\frac{1}{\theta}}$.

We offer two ways to parse this bound. Firstly, since $t \geq t(\eta)$, we have

$$\frac{3}{4}c_\theta t^{\frac{1}{\theta}} \geq \frac{3}{4}c_\theta(\frac{1}{2}c_\theta n\alpha)^{\frac{1}{2\theta-1}} \geq \log n$$

whenever $n^{\frac{1}{2\theta-1}} \geq \frac{4}{3c_\theta}(\frac{1}{2}c_\theta\alpha)^{\frac{1}{1-2\theta}}\log n$. Given $n^{\frac{1}{2\theta-1}}$ grows faster than $\log n$, the condition $n^{\frac{1}{2\theta-1}} \geq \frac{4}{3c_\theta}(\frac{1}{2}c_\theta\alpha)^{\frac{1}{1-2\theta}}\log n$ can be easily satisfied, and thus we get $\mathbb{P}\left(f(X) - \mathbb{E}[f(X')] > t\right) \leq 2\exp(-\frac{1}{4}c_\theta t^{\frac{1}{\theta}})$.

Secondly, for all $t \geq c_2 n^{\frac{\theta}{2\theta-1}}$, where c_2 is a positive constant, we have

$$\frac{3}{4}c_\theta t^{\frac{1}{\theta}} \geq \frac{3}{4}c_\theta c_2^{1/\theta} n^{\frac{1}{2\theta-1}}.$$

Given $n^{\frac{1}{2\theta-1}}$ grows faster than $\log n$, by choosing large c_2, we can ensure $\log n \leq \frac{3}{4}c_\theta t^{\frac{1}{\theta}}$ holds for all integer n. Therefore, there exists a positive constant $c_2 > 0$, such that for every $t \geq c_2 n^{\frac{\theta}{2\theta-1}}$, we have $\mathbb{P}\left(f(X) - \mathbb{E}[f(X')] > t\right) \leq 2\exp(-\frac{1}{4}c_\theta t^{\frac{1}{\theta}})$.

In the second case $t < t(\eta)$, we have

$$\mathbb{P}(f(X) - \mathbb{E}[f(X')] > t) \leq \exp\left(-\frac{t^2}{2n\alpha}\right) + n\exp\left(-\frac{2t(\eta)^2}{n\alpha}\right) = \exp\left(-\frac{t^2}{2n\alpha}\right) + \exp\left(-\frac{2t(\eta)^2}{n\alpha} + \log n\right).$$

Similar to the sub-exponential case, due to $t < t(\eta)$, we derive

$$\frac{t^2}{2n\alpha} - \frac{2t(\eta)^2}{n\alpha} + \log n = \frac{t^2 - t(\eta)^2 - \frac{3}{4}(\frac{1}{2}c_\theta n\alpha)^{\frac{2\theta}{2\theta-1}} + 2n\alpha\log n}{2n\alpha} < \frac{-\frac{3}{4}(\frac{1}{2}c_\theta n\alpha)^{\frac{2\theta}{2\theta-1}} + 2n\alpha\log n}{2n\alpha} \leq 0$$

whenever $n^{\frac{1}{2\theta-1}} \geq \frac{8}{3}\alpha(\frac{1}{2}c_\theta\alpha)^{\frac{2\theta}{1-2\theta}}\log n$. In this case, the first term $\exp(-\frac{t^2}{2n\alpha})$ also dominates the second, and thus we get $\mathbb{P}\left(f(X) - \mathbb{E}[f(X')] > t\right) \leq 2\exp(-\frac{t^2}{2n\alpha})$. The proof is complete. □

A.8. Proof of Theorem 3.7

Proof. Again, given Theorem 3.1, we need to bound $\Lambda(\tau, \eta)$. Firstly, we have

$$\sup_{x\in\mathcal{X}^n}\mathbb{E}\left[(f_k^\tau(X)(x))^2\,\mathbb{I}(f_k^\tau(X)(x) \leq 0)\right] \leq \sup_{x\in\mathcal{X}^n}\mathbb{E}\left[(f_k(X)(x))^2\,\mathbb{I}(f_k(X)(x) \leq 0)\right] = \sigma_-^2.$$

Then, for any $x \in \mathcal{X}^n$, we can decompose $\mathbb{E}[f_k^\tau(X)^2 e^{\beta f_k^\tau(X)}\mathbb{I}(f_k^\tau(X) > 0)]$ as follows

$$\begin{aligned}
&\mathbb{E}\left[(f_k^\tau(X)(x))^2\, e^{\beta f_k^\tau(X)(x)}\mathbb{I}(0 < f_k^\tau(X)(x) \leq \tau)\right] \\
=&\mathbb{E}\left[(f_k^\tau(X)(x))^2\, e^{\beta f_k^\tau(X)(x)}\mathbb{I}(0 < f_k^\tau(X)(x) \leq 1)\right] + \mathbb{E}\left[(f_k^\tau(X)(x))^2\, e^{\beta f_k^\tau(X)(x)}\mathbb{I}(1 < f_k^\tau(X)(x) \leq \tau)\right].
\end{aligned}$$

Note that $\beta = \eta\tau^{-1}(c\log\tau)$. When $2-(1-\eta)c \neq 0$, i.e., $\eta \neq 1-\frac{2}{c}$, we derive

$$
\begin{aligned}
&\mathbb{E}\left[\left(f_k^\tau(X)(x)\right)^2 e^{\beta f_k^\tau(X)(x)}\mathbb{I}(1 < f_k^\tau(X)(x) \leq \tau)\right] \\
=&\int_1^\tau \mathbb{P}\left(f_k^\tau(X)(x) > t\right)(2t+\beta t^2)e^{\beta t}dt \\
\leq&\int_1^\tau \exp(-c\log t)(2t+\beta t^2)e^{\beta t}dt \\
=&\int_1^\tau \exp(-(1-\beta t(c\log t)^{-1})c\log t)(2t+\beta t^2)dt \\
=&\int_1^\tau t^{1-(1-\beta t(c\log t)^{-1})c}\left(2+\beta t\right)dt \\
\leq&\int_1^\tau t^{1-(1-\beta \tau(c\log \tau)^{-1})c}\left(2+\beta t\right)dt \\
=&\int_1^\tau t^{1-(1-\eta)c}\left(2+\eta c\frac{\log\tau}{\tau}t\right)dt \\
\leq&\int_1^\tau t^{1-(1-\eta)c}\left(2+\eta c\frac{\log t}{t}t\right)dt \\
=&\frac{t^{2-(1-\eta)c}}{2-(1-\eta)c}\left(2+\eta c\Big(-\frac{1}{2-(1-\eta)c}+\log t\Big)\right)\Bigg|_1^\tau \\
=&\frac{\tau^{2-(1-\eta)c}}{2-(1-\eta)c}\left(2+\eta c\Big(-\frac{1}{2-(1-\eta)c}+\log \tau\Big)\right)-\frac{\left(2+\eta c\left(-\frac{1}{2-(1-\eta)c}\right)\right)}{2-(1-\eta)c} \\
=&\frac{2+\eta c\left(-\frac{1}{2-(1-\eta)c}\right)}{2-(1-\eta)c}(\tau^{2-(1-\eta)c}-1)+\frac{\tau^{2-(1-\eta)c}\eta c\log\tau}{2-(1-\eta)c},
\end{aligned}
$$

where we have used the fact that $\int t^a dt = \frac{t^{a+1}}{a+1}$, $\int t^a\log t dt = (\frac{\log t}{k+1}-\frac{1}{(k+1)^2})t^{a+1}$ and $\frac{\log t}{t} \geq \frac{\log\tau}{\tau}$. When $2-(1-\eta)c = 0$, i.e., $\eta = 1-\frac{2}{c}$, we derive

$$
\begin{aligned}
&\mathbb{E}\left[\left(f_k^\tau(X)(x)\right)^2 e^{\beta f_k^\tau(X)(x)}\mathbb{I}(1 < f_k^\tau(X)(x) \leq \tau)\right] \\
\leq&\int_1^\tau t^{1-(1-\eta)c}\left(2+\eta c\frac{\log t}{t}t\right)dt = \int_1^\tau t^{-1}\left(2+c(1-\frac{2}{c})\frac{\log t}{t}t\right)dt = 2\log\tau+\frac{(c-2)(\log\tau)^2}{2}.
\end{aligned}
$$

For $\mathbb{E}[(f_k^\tau(X)(x))^2 e^{\beta f_k^\tau(X)(x)}\mathbb{I}(0 < f_k^\tau(X)(x) \leq 1)]$, we derive that

$$
\mathbb{E}\left[\left(f_k^\tau(X)(x)\right)^2 e^{\beta f_k^\tau(X)(x)}\mathbb{I}(0 < f_k^\tau(X)(x) \leq 1)\right] \leq \mathbb{E}[e^\beta] = e^{\eta\tau^{-1}(c\log\tau)} = \tau^{c\eta\tau^{-1}}.
$$

Putting these terms together gives: (1.) If $\eta \neq 1-\frac{2}{c}$, we have $\Lambda(\tau,\eta) \leq \frac{2+\eta c\left(-\frac{1}{2-(1-\eta)c}\right)}{2-(1-\eta)c}(\tau^{2-(1-\eta)c}-1)+\frac{\tau^{2-(1-\eta)c}\eta c\log\tau}{2-(1-\eta)c}+\tau^{c\eta\tau^{-1}}+\sigma_-^2$. (2.) If $\eta = 1-\frac{2}{c}$, we have $\Lambda(\tau,\eta) \leq 2\log\tau+\frac{(c-2)(\log\tau)^2}{2}+\tau^{c\eta\tau^{-1}}+\sigma_-^2$. Substituting these bounds into Theorem 3.1, the proof is complete. □

A.9. Proof of Theorem 3.8

Proof. The proof uses the Lebesgue Dominated Convergence Theorem. It's essential to bound $\Lambda(\tau,\eta) \triangleq \sup_{x\in\mathcal{X}^n}\mathbb{E}[(f_k^\tau(X)(x))^2\mathbb{I}(f_k^\tau(X)(x) \leq 0)] + \mathbb{E}[(f_k^\tau(X)(x))^2 e^{\beta f_k^\tau(X)(x)}\mathbb{I}(0 < f_k^\tau(X)(x) \leq \tau)]$. Firstly, we have $(f_k^\tau(X)(x))^2 \leq (f_k(X)(x))^2$. If τ grows to infinity, $f_k^\tau(X)(x)$ converges to $f_k(X)(x)$ almost surely. Thus, using the Lebesgue Dominated Convergence Theorem, we obtain

$$
\lim_{\tau\to\infty}\mathbb{E}[(f_k^\tau(X)(x))^2\mathbb{I}(f_k^\tau(X)(x) \leq 0)] = \mathbb{E}[(f_k(X)(x))^2\mathbb{I}(f_k(X)(x) \leq 0)].
$$

Next, we examine $\mathbb{E}[(f_k^\tau(X)(x))^2 e^{\beta f_k^\tau(X)(x)} \mathbb{I}(0 < f_k^\tau(X)(x) \leq \tau)]$. From the proof of Theorem 3.1, we know that $\beta \leq \frac{\eta h(\tau)}{\tau}$, where $\eta < 1$. When $h(t) = c_\theta t^{\frac{1}{\theta}}$ such that $\theta > 1$, we have $\frac{1}{\theta} - 1 < 0$ and $\lim_{\tau\to\infty} \beta \leq \lim_{\tau\to\infty} \frac{\eta c_\theta \tau^{\frac{1}{\theta}}}{\tau} = 0$, which further implies that $(f_k^\tau(X)(x))^2 e^{\beta f_k^\tau(X)(x)}$ converges to $(f_k(X)(x))^2$ almost surely.

To proceed, we introduce the following dominated variable

$$(f_k(X)(x))^2 e^{\eta c_\theta (f_k(X)(x))^{\frac{1}{\theta}}} \mathbb{I}(0 < f_k(X)(x)). \tag{10}$$

Consequently, we derive

$$e^{\beta f_k^\tau(X)(x)} \leq \exp\left(\frac{\eta c_\theta \tau^{\frac{1}{\theta}}}{\tau} f_k^\tau(X)(x)\right) \leq \exp\left(\frac{\eta c_\theta (f_k^\tau(X)(x))^{\frac{1}{\theta}}}{f_k^\tau(X)(x)} f_k^\tau(X)(x)\right) \leq \exp\left(\eta c_\theta (f_k(X)(x))^{\frac{1}{\theta}}\right),$$

which means that

$$(f_k^\tau(X)(x))^2 e^{\beta f_k^\tau(X)(x)} \mathbb{I}(0 < f_k^\tau(X)(x) \leq \tau) \leq (f_k(X)(x))^2 e^{\eta c_\theta (f_k(X)(x))^{\frac{1}{\theta}}} \mathbb{I}(0 < f_k(X)(x)).$$

The following step is to prove the integrability of $(f_k(X)(x))^2 e^{\eta c_\theta (f_k(X)(x))^{\frac{1}{\theta}}} \mathbb{I}(0 < f_k(X)(x))$, leading to

$$\begin{aligned}
&\mathbb{E}\left[(f_k(X)(x))^2 e^{\eta c_\theta (f_k(X)(x))^{\frac{1}{\theta}}} \mathbb{I}(0 < f_k(X)(x))\right] \\
&= \int_0^\infty \mathbb{P}\left((f_k(X)(x))^2 e^{\eta c_\theta (f_k(X)(x))^{\frac{1}{\theta}}} \mathbb{I}(0 < f_k(X)(x)) > u\right) du \\
&= \int_0^\infty \mathbb{P}\left((f_k(X)(x))^2 e^{\eta c_\theta (f_k(X)(x))^{\frac{1}{\theta}}} > t^2 e^{\eta c_\theta t^{\frac{1}{\theta}}}, f_k(X)(x) > 0\right) dt^2 e^{\eta c_\theta t^{\frac{1}{\theta}}} \\
&= \int_0^\infty \mathbb{P}\left(f_k(X)(x) > t\right) \left(2t + t^2 \frac{1}{\theta} \eta c_\theta t^{\frac{1}{\theta}-1}\right) e^{\eta c_\theta t^{\frac{1}{\theta}}} dt \\
&\leq \int_0^\infty \exp\left(-c_\theta t^{\frac{1}{\theta}}\right) \left(2t + t^2 \frac{1}{\theta} \eta c_\theta t^{\frac{1}{\theta}-1}\right) e^{\eta c_\theta t^{\frac{1}{\theta}}} dt \\
&\leq \int_0^\infty \exp\left(-(1-\eta) c_\theta t^{\frac{1}{\theta}}\right) \left(2t + \frac{1}{\theta} \eta c_\theta t^{\frac{1}{\theta}+1}\right) dt < \infty,
\end{aligned}$$

where the last inequality follows from that $\eta < 1$ and that the exponential term converges faster than the polynomial. Thus, $(f_k(X)(x))^2 e^{\eta c_\theta (f_k(X)(x))^{\frac{1}{\theta}}} \mathbb{I}(0 < f_k(X)(x))$ is integrable. By the Lebesgue Dominated Convergence Theorem, we obtain

$$\lim_{\tau\to\infty} \mathbb{E}\left[(f_k^\tau(X)(x))^2 e^{\beta f_k^\tau(X)(x)} \mathbb{I}(0 < f_k^\tau(X)(x) \leq \tau)\right] = \mathbb{E}[(f_k(X)(x))^2 \mathbb{I}(f_k(X)(x) > 0)].$$

Combining the two bounds, we obtain

$$\begin{aligned}
&\lim_{\tau\to\infty} \sup_{x\in\mathcal{X}^n} \mathbb{E}\left[(f_k^\tau(X)(x))^2 e^{\beta f_k^\tau(X)(x)} \mathbb{I}(0 < f_k^\tau(X)(x) \leq \tau) + (f_k^\tau(X)(x))^2 \mathbb{I}(f_k^\tau(X)(x) \leq 0)\right] \\
&= \sup_{x\in\mathcal{X}^n} \mathbb{E}\left[(f_k(X)(x))^2\right] = \sigma^2.
\end{aligned}$$

Thus, for any given $\delta > 0$, we will obtain a constant $c_\delta > 0$ such that for any $\tau > c_\delta$

$$\sup_{x\in\mathcal{X}^n} \mathbb{E}\left[(f_k^\tau(X)(x))^2 e^{\beta f_k^\tau(X)(x)} \mathbb{I}(0 < f_k^\tau(X)(x) \leq \tau) + (f_k^\tau(X)(x))^2 \mathbb{I}(f_k^\tau(X)(x) \leq 0)\right] \leq \sigma^2 + \delta.$$

This inequality also implies that for all $t > c_\delta$, Theorem 3.8 is proved by plugging $\Lambda(\tau, \eta) = \sigma^2 + \delta$ into Theorem 3.1. The proof is complete. □

A.10. Proof of Theorem 3.9

Proof. The proof is similar to the proof of Theorem 3.8. In (10), we instead need to introduce the following dominated variable

$$(f_k(X)(x))^2 e^{\eta c \log f_k(X)(x)} \mathbb{I}(0 < f_k(X)(x)) = (f_k(X)(x))^{2+\eta c} \mathbb{I}(0 < f_k(X)(x)).$$

In this case, we derive

$$e^{\beta f_k^\tau(X)(x)} \leq \exp\Big(\frac{\eta c \log \tau}{\tau} f_k^\tau(X)(x)\Big) \leq \exp\Big(\frac{\eta c \log f_k^\tau(X)(x)}{f_k^\tau(X)(x)} f_k^\tau(X)(x)\Big) \leq \exp\Big(\eta c \log f_k(X)(x)\Big),$$

which means that

$$(f_k^\tau(X)(x))^2 e^{\beta f_k^\tau(X)(x)} \mathbb{I}(0 < f_k^\tau(X)(x) \leq \tau) \leq (f_k(X)(x))^2 e^{\eta c \log f_k(X)(x)} \mathbb{I}(0 < f_k(X)(x)).$$

The following step is also to prove the integrability of $(f_k(X)(x))^{2+\eta c} \mathbb{I}(0 < f_k(X)(x))$, leading to

$$\begin{aligned}
&\mathbb{E}\left[(f_k(X)(x))^{2+\eta c} \mathbb{I}(0 < f_k(X)(x))\right] \\
&= \int_0^\infty \mathbb{P}\left((f_k(X)(x))^{2+\eta c} > t^{2+\eta c}, f_k(X)(x) > 0\right)(2+\eta c) t^{1+\eta c} dt \\
&= \int_0^\infty \mathbb{P}\left(f_k(X)(x) > t\right)(2+\eta c) t^{1+\eta c} dt \\
&\leq \int_0^\infty \exp(-c \log t)(2+\eta c) t^{1+\eta c} dt \\
&= \int_0^\infty t^{(1+\eta c - c)}(2+\eta c) dt < \infty,
\end{aligned}$$

where the last inequality follows from that $1 + \eta c - c < 0$ since $c > 2$ and $\eta < 1 - \frac{2}{c}$. Thus, by the Lebesgue Dominated Convergence Theorem, for $c > 2$ and $\eta < 1 - \frac{2}{c}$, we obtain

$$\begin{aligned}
&\lim_{\tau \to \infty} \sup_{x \in \mathcal{X}^n} \mathbb{E}\left[(f_k^\tau(X)(x))^2 e^{\beta f_k^\tau(X)(x)} \mathbb{I}(0 < f_k^\tau(X)(x) \leq \tau) + (f_k^\tau(X)(x))^2 \mathbb{I}(f_k^\tau(X)(x) \leq 0)\right] \\
&= \sup_{x \in \mathcal{X}^n} \mathbb{E}\left[(f_k(X)(x))^2\right] = \sigma^2.
\end{aligned}$$

Then, for any given $\delta > 0$, we will obtain a constant $c_\delta > 0$ such that for any $\tau > c_\delta$

$$\sup_{x \in \mathcal{X}^n} \mathbb{E}\left[(f_k^\tau(X)(x))^2 e^{\beta f_k^\tau(X)(x)} \mathbb{I}(0 < f_k^\tau(X)(x) \leq \tau) + (f_k^\tau(X)(x))^2 \mathbb{I}(f_k^\tau(X)(x) \leq 0)\right] \leq \sigma^2 + \delta.$$

This inequality also implies that for all $t > c_\delta$, Theorem 3.9 is proved by plugging $\Lambda(\tau, \eta) = \sigma^2 + \delta$ into Theorem 3.1. The proof is complete. □

B. Proofs of Section 4

Lemma B.1 (Lemma 6 in (Maurer & Pontil, 2021)). *Let X, X' be iid with values in $\mathcal{X}$, $\phi : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ measurable. Then*

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

B.1. Proof of Theorem 4.1

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

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

By Lemma B.1, we get $\|\mathbb{E}[\|X_k - X_k'\||X]\|_p \leq 2\|\|X_k\|\|_p$ and thus $\|f_k(X)(x)\|_p \leq 2\|\|X_k\|\|_p$. Given that the $\|X_k\|$ have the tail ct, the tail of $f_k(X)(x)$ can be expressed as $c'ct$, where c' is an absolute positive constant and comes from the equivalent properties of sub-exponential random variables on their tails and moments, see Proposition 2.7.1 in (Vershynin, 2018). We now provide the proof. For all $p \geq 1$, by Markov's inequality,

$$\mathbb{P}(f_k(X)(x) > t) \leq \frac{\mathbb{E}[|f_k(X)(x)|^P]}{t^p}.$$

Setting t ssuch that $\exp(-p) = \frac{\mathbb{E}[|f_k(X)(x)|^P]}{t^p}$, we obtain

$$\mathbb{P}(f_k(X)(x) > e\|f_k(X)(x)\|_p) \leq \exp(-p),$$

implying that

$$\mathbb{P}(f_k(X)(x) > 2e\|\|X_k\|\|_p) \leq \exp(-p).$$

Note that if the $\|X_k\|$ have the tail ct, $\|\|X_k\|\|_p \leq \frac{1}{c'c}p$, where c' is an absolute positive constant, see Proposition 2.7.1 in (Vershynin, 2018). Thus, solving p, we get

$$\mathbb{P}(f_k(X)(x) > t) \leq \exp\left(-\frac{c'c}{2e}t\right), \tag{11}$$

meaning that the tail of $f_k(X)(x)$ can be expressed as $c'ct$, where c' is an absolute positive constant.

Furthermore, we have

$$\sigma_-^2 = \sup_{x\in\mathcal{X}^n} \mathbb{E}[(f_k(X)(x))^2 \mathbb{I}(f_k(X)(x) \leq 0)] \leq \mathbb{E}[(\mathbb{E}[\|X_k - X_k'\| \, |X_k])^2] \leq \mathbb{E}[\|X_k - X_k'\|^2] \leq 4\mathbb{E}[\|X_1\|^2],$$

where the second inequality uses Lemma B.1 and the last inequality uses the i.i.d. assumption.

Plugging these bounds into Corollary 3.4, we get $\alpha = \frac{16}{(c'c)^2} + 4\mathbb{E}[\|X_1\|^2]$ and $t(\eta) = \frac{c'cn}{2}\alpha$. Hence, if $t \geq t(\eta)$ we have

$$\mathbb{P}\left(\left\|\sum_{i=1}^n X_i\right\| - \mathbb{E}\left\|\sum_{i=1}^n X_i\right\| > t\right) \leq 2\exp\left(-\frac{1}{4}c'ct\right);$$

if $0 \leq t < t(\eta)$, when $n \geq \frac{8\log n}{3(c'c)^2\alpha}$, we have

$$\mathbb{P}\left(\left\|\sum_{i=1}^n X_i\right\| - \mathbb{E}\left\|\sum_{i=1}^n X_i\right\| > t\right) \leq 2\exp\left(-\frac{t^2}{2n\alpha}\right).$$

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

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

Then we have

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

Similarly, by Lemma B.1, we have $\|f_k(X)(x)\|_p \leq 2\|\|X_k\|\|_p$, and the tail of $f_k(X)(x)$ can be expressed as $c'ct$, where c' is an absolute positive constant.

Furthermore, we have

$$\sigma_-^2 = \sup_{x\in\mathcal{X}^n} \mathbb{E}[(f_k(X)(x))^2 \mathbb{I}(f_k(X)(x) \leq 0)] \leq \mathbb{E}[(\mathbb{E}[\|X_k - X_k'\| \, |X_k])^2] \leq \mathbb{E}[\|X_k - X_k'\|^2] \leq 4\mathbb{E}[\|X_1\|^2],$$

where the second inequality uses Lemma B.1 and the last inequality uses the i.i.d. assumption.

Plugging these bounds into Corollary 3.4, we obtain $\alpha = \frac{16}{(c'c)^2} + 4\mathbb{E}[\|X_1\|^2]$ and $t(\eta) = \frac{c'cn}{2}\alpha$. Hence, if $t \geq t(\eta)$, we have

$$\begin{aligned}
&\mathbb{P}\left(\left\|\sum_{i=1}^{n}(X_i - \mathbb{E}X_1')\right\| > t + \mathbb{E}\left\|\sum_{i=1}^{n}(X_i - \mathbb{E}X_i')\right\|\right) \\
\leq &\mathbb{P}\left(\left\|\sum_{i=1}^{n}(X_i - \mathbb{E}X_1')\right\| > t + \sqrt{n}\|\|X_1\|\|_2\right) \leq 2\exp\left(-\frac{1}{4}c'ct\right).
\end{aligned}$$

If $0 \leq t < t(\eta)$, when $n \geq \frac{8\log n}{3(c'c)^2\alpha}$, we have

$$\begin{aligned}
&\mathbb{P}\left(\left\|\sum_{i=1}^{n}(X_i - \mathbb{E}X_1')\right\| > t + \mathbb{E}\left\|\sum_{i=1}^{n}(X_i - \mathbb{E}X_i')\right\|\right) \\
\leq &\mathbb{P}\left(\left\|\sum_{i=1}^{n}(X_i - \mathbb{E}X_1')\right\| > t + \sqrt{n}\|\|X_1\|\|_2\right) \leq 2\exp\left(-\frac{t^2}{2n\alpha}\right).
\end{aligned}$$

The proof is complete. □

B.2. Proof of Theorem 4.2

Proof. The proof follows the proof of Theorem 4.1.

(i) Since the $\|X_k\|$ have the tail $c_\theta t^{\frac{1}{\theta}}$, the tail of $f_k(X)(x)$ can be written as $c'_\theta t^{\frac{1}{\theta}}$, where c'_θ is a positive constant and comes from the equivalent properties of sub-Weibull random variables on its tails and moments, see Theorem 2.1 in (Vladimirova et al., 2020). We now provide the proof. For all real $p \geq 1$, by Markov's inequality,

$$\mathbb{P}(f_k(X)(x) > t) \leq \frac{\mathbb{E}[|f_k(X)(x)|^P]}{t^p}.$$

Setting t such that $\exp(-p) = \frac{\mathbb{E}[|f_k(X)(x)|^P]}{t^p}$, we obtain

$$\mathbb{P}(f_k(X)(x) > e\|f_k(X)(x)\|_p) \leq \exp(-p),$$

implying that

$$\mathbb{P}(f_k(X)(x) > 2e\|\|X_k\|\|_p) \leq \exp(-p).$$

Note that if the $\|X_k\|$ have the tail $c_\theta t^{\frac{1}{\theta}}$, $\|\|X_k\|\|_p \leq \frac{1}{c'_\theta}p^\theta$, where c'_θ is a positive constant, see Theorem 2.1 in (Vladimirova et al., 2020). Thus, solving p, we get

$$\mathbb{P}(f_k(X)(x) > t) \leq \exp\left(-\left(\frac{c'_\theta}{2e}t\right)^{\frac{1}{\theta}}\right), \tag{12}$$

meaning that the tail of $f_k(X)(x)$ can be written as $c'_\theta t^{\frac{1}{\theta}}$, where c'_θ is a positive constant.

Plugging this bound into Corollary 3.6, we get $\alpha = \frac{\Gamma(2\theta+1)2^{2\theta}}{(c'_\theta)^{2\theta}} + \frac{c'_\theta c_1^{1/\theta}\Gamma(3\theta+1)2^{3\theta}}{6c_1(c'_\theta)^{3\theta}} + 4\mathbb{E}[\|X_1\|^2]$ and $t(\eta) = (\frac{1}{2}c'_\theta n\alpha)^{\frac{\theta}{2\theta-1}}$, where c_1 is a positive constant. Hence, if $t \geq t(\eta)$ and when $n^{\frac{1}{2\theta-1}} \geq \frac{4}{3c'_\theta}(\frac{1}{2}c'_\theta\alpha)^{\frac{1}{1-2\theta}}\log n$ we have

$$\mathbb{P}\left(\left\|\sum_{i=1}^{n}X_i\right\| - \mathbb{E}\left\|\sum_{i=1}^{n}X_i\right\| > t\right) \leq 2\exp\left(-\frac{1}{4}c'_\theta t^{\frac{1}{\theta}}\right).$$

If $0 \le t < t(\eta)$ and when $n^{\frac{1}{2\theta-1}} \ge \frac{8}{3}\alpha(\frac{1}{2}c'_\theta\alpha)^{\frac{2\theta}{1-2\theta}}\log n$, we have

$$\mathbb{P}\left(\left\|\sum_{i=1}^n X_i\right\| - \mathbb{E}\left\|\sum_{i=1}^n X_i\right\| > t\right) \le 2\exp\left(-\frac{t^2}{2n\alpha}\right).$$

(ii) This proof follows the same pattern. If $t \ge t(\eta)$ and when $n^{\frac{1}{2\theta-1}} \ge \frac{4}{3c'_\theta}(\frac{1}{2}c'_\theta\alpha)^{\frac{1}{1-2\theta}}\log n$ we have

$$\mathbb{P}\left(\left\|\sum_{i=1}^n (X_i - \mathbb{E}X'_1)\right\| > t + \mathbb{E}\left\|\sum_{i=1}^n (X_i - \mathbb{E}X'_i)\right\|\right) \le \mathbb{P}\left(\left\|\sum_{i=1}^n (X_i - \mathbb{E}X'_1)\right\| > t + \sqrt{n}\|\|X_1\|\|_2\right) \le 2\exp\left(-\frac{1}{4}c'_\theta t^{\frac{1}{\theta}}\right).$$

If $0 \le t < t(\eta)$ and when $n^{\frac{1}{2\theta-1}} \ge \frac{8}{3}\alpha(\frac{1}{2}c'_\theta\alpha)^{\frac{2\theta}{1-2\theta}}\log n$, we have

$$\mathbb{P}\left(\left\|\sum_{i=1}^n (X_i - \mathbb{E}X'_1)\right\| > t + \mathbb{E}\left\|\sum_{i=1}^n (X_i - \mathbb{E}X'_i)\right\|\right) \le \mathbb{P}\left(\left\|\sum_{i=1}^n (X_i - \mathbb{E}X'_1)\right\| > t + \sqrt{n}\|\|X_1\|\|_2\right) \le 2\exp\left(-\frac{t^2}{2n\alpha}\right).$$

The proof is complete. □

B.3. Proof of Theorem 4.3

Proof. The vector space

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

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

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

With Lemma B.1 and the i.i.d. assumption, we have

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

where the first inequality uses the Lipschitz condition. Since the $\|X_i\|$ have the tail ct, the tail of $\|\bar{X}_i\|_\mathcal{B}$ can be written as $\frac{n}{L}c'ct$, where c' is an absolute positive constant and comes from the equivalent properties of sub-exponential random variables on its tails and moments, see (11).

Furthermore, we have

$$\sigma_-^2 \le \frac{4L^2}{n^2}\mathbb{E}[\|X_1\|^2].$$

Let $\alpha = \frac{16}{(\frac{n}{L}c'c)^2} + \frac{4L^2}{n^2}\mathbb{E}[\|X_1\|^2]$ and $t(\eta) = \frac{c'cn^2}{2L}\alpha$. Thus, from Theorem 4.1 (ii), we have

$$\mathbb{P}\left(\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_i g(X_i) - \mathbb{E}[g(X'_i)] - \mathbb{E}\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_i g(X_i) - \mathbb{E}[g(X'_i)]\right] > t + \frac{2L}{\sqrt{n}}\|\|X_1\|\|_2\right)$$
$$\le \begin{cases} 2\exp\left(-\frac{1}{4}\frac{n}{L}c'ct\right) & \text{if } t \ge t(\eta), \\ 2\exp\left(-\frac{t^2}{2n\alpha}\right) & \text{if } 0 \le t < t(\eta) \text{ and when } n \ge \frac{8\log n}{3(\frac{n}{L}c'c)^2\alpha}. \end{cases}$$

The proof is complete. □

B.4. Proof of Theorem 4.4

Proof. The proof follows the proof of Theorem 4.3.

Since the $\|X_i\|$ have the tail $c_\theta t^{\frac{1}{\theta}}$, the tail of $\|\bar{X}_i\|_{\mathcal{B}}$ can be written as $(\frac{n}{L})^{\frac{1}{\theta}} c'_\theta t^{\frac{1}{\theta}}$, where c'_θ is a positive constant and comes from the equivalent properties of sub-Weibull random variables on its tails and moments, referring to (12).

Let $\alpha = \frac{\Gamma(2\theta+1)2^{2\theta}}{((\frac{n}{L})^{\frac{1}{\theta}} c'_\theta)^{2\theta}} + \frac{(\frac{n}{L})^{\frac{1}{\theta}} c'_\theta c_1^{1/\theta} \Gamma(3\theta+1) 2^{3\theta}}{6c_1((\frac{n}{L})^{\frac{1}{\theta}} c'_\theta)^{3\theta}} + \frac{4L^2}{n^2}\mathbb{E}[\|X_1\|^2]$ and $t(\eta) = (\frac{1}{2}(\frac{n}{L})^{\frac{1}{\theta}} c'_\theta n\alpha)^{\frac{\theta}{2\theta-1}}$, where c_1 is a positive constant. Thus, from Theorem 4.1 (ii), we have

$$\begin{aligned}
&\mathbb{P}\left(\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_i g(X_i) - \mathbb{E}[g(X'_i)] - \mathbb{E}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_i g(X_i) - \mathbb{E}[g(X'_i)]\right] > t + \frac{2L}{\sqrt{n}}\|\|X_1\|\|_2\right) \\
&\leq \begin{cases} 2\exp\left(-\frac{1}{4}(\frac{n}{L})^{\frac{1}{\theta}} c'_\theta t^{\frac{1}{\theta}}\right) & \text{if } t \geq t(\eta) \text{ and when } n^{\frac{1}{2\theta-1}} \geq \frac{4}{3(\frac{n}{L})^{\frac{1}{\theta}} c'_\theta}(\frac{1}{2}(\frac{n}{L})^{\frac{1}{\theta}} c'_\theta \alpha)^{\frac{1}{1-2\theta}} \log n, \\ 2\exp\left(-\frac{t^2}{2n\alpha}\right) & \text{if } 0 \leq t < t(\eta) \text{ and when } n^{\frac{1}{2\theta-1}} \geq \frac{8}{3}\alpha(\frac{1}{2}(\frac{n}{L})^{\frac{1}{\theta}} c'_\theta \alpha)^{\frac{2\theta}{1-2\theta}} \log n. \end{cases}
\end{aligned}$$

The proof is complete. □

B.5. Proof of Theorem 4.5

Proof. It is clear that

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

where the first inequality uses the Lipschitz condition and the last inequality uses Lemma B.1. Since the $d(X_k,X'_k)$ have the tail ct, the tail of $f_k(X)(x)$ can be written as $\frac{1}{L}c'ct$, where c' is an absolute positive constant and comes from the equivalent properties of sub-exponential random variables on its tails and moments, see (11). Furthermore, we have

$$\sigma_-^2 = \sup_{x\in\mathcal{X}^n} \mathbb{E}[(f_k(X)(x))^2 \mathbb{I}(f_k(X)(x)\leq 0)] \leq L^2\mathbb{E}[(d(X_1,X'_1))^2].$$

Plugging these bounds into Corollary 3.4, we get $\alpha = \frac{16}{(\frac{1}{L}c'c)^2} + L^2\mathbb{E}[(d(X_1,X'_1))^2]$ and $t(\eta) = \frac{c'cn}{2L}\alpha$. Hence, if $t \geq t(\eta)$ we have

$$\mathbb{P}\left(f(X) - \mathbb{E}[f(X')] > t\right) \leq 2\exp\left(-\frac{1}{4L}c'ct\right);$$

if $0 \leq t < t(\eta)$, when $n \geq \frac{8\log n}{3(\frac{1}{L}c'c)^2\alpha}$, we have

$$\mathbb{P}\left(f(X) - \mathbb{E}[f(X')] > t\right) \leq 2\exp\left(-\frac{t^2}{2n\alpha}\right).$$

The proof is complete. □

B.6. Proof of Theorem 4.6

Proof. The proof follows the proof of Theorem 4.5.

Since the $d(X_k,X'_k)$ have the tail $c_\theta t^{\frac{1}{\theta}}$, the tail of $f_k(X)(x)$ can be written as $(\frac{1}{L})^{\frac{1}{\theta}} c'_\theta t^{\frac{1}{\theta}}$, where c'_θ is a positive constant and comes from the equivalent properties of sub-Weibull random variables on its tails and moments, referring to (12). Plugging this bound into Corollary 3.6, we get $\alpha = \frac{\Gamma(2\theta+1)2^{2\theta}}{((\frac{1}{L})^{\frac{1}{\theta}} c'_\theta)^{2\theta}} + \frac{(\frac{1}{L})^{\frac{1}{\theta}} c'_\theta c_1^{1/\theta}\Gamma(3\theta+1)2^{3\theta}}{6c_1((\frac{1}{L})^{\frac{1}{\theta}} c'_\theta)^{3\theta}} + L^2\mathbb{E}[(d(X_1,X'_1))^2]$ and $t(\eta) =$

$(\frac{1}{2}(\frac{1}{L})^{\frac{1}{\theta}}c'_\theta n\alpha)^{\frac{\theta}{2\theta-1}}$, where c_1 is a positive constant. Hence, if $t \geq t(\eta)$ and when $n^{\frac{1}{2\theta-1}} \geq \frac{4}{3(\frac{1}{L})^{\frac{1}{\theta}}c'_\theta}(\frac{1}{2}(\frac{1}{L})^{\frac{1}{\theta}}c'_\theta\alpha)^{\frac{1}{1-2\theta}}\log n$ we have

$$\mathbb{P}\left(f(X) - \mathbb{E}[f(X')] > t\right) \leq 2\exp\left(-\frac{1}{4}(\frac{1}{L})^{\frac{1}{\theta}}c'_\theta t^{\frac{1}{\theta}}\right).$$

If $0 \leq t < t(\eta)$ and when $n^{\frac{1}{2\theta-1}} \geq \frac{8}{3}\alpha(\frac{1}{2}(\frac{1}{L})^{\frac{1}{\theta}}c'_\theta\alpha)^{\frac{2\theta}{1-2\theta}}\log n$, we have

$$\mathbb{P}\left(f(X) - \mathbb{E}[f(X')] > t\right) \leq 2\exp\left(-\frac{t^2}{2n\alpha}\right).$$

The proof is complete. □