SCIENCE CHINA
Information Sciences

• **Supplementary File** •

Distribution-dependent concentration inequalities for tighter generalization bounds

Xinxing WU[1] & Junping ZHANG[2*]

[1]*Department of Computer, Shanghai Technical Institute of Electronics and Information, Shanghai* 201411*, China;*
[2]*Shanghai Key Laboratory of Intelligent Information Processing,*
School of Computer Science, Fudan University, Shanghai 200433*, China*

Appendix A More Related work

Here, we give the much more detailed related work as the supplement of the paper.

Although Hoeffding's and McDiarmid's inequalities have achieved great success in learning theory, [1–4] have noted their limitations in applications due to the fact that these inequalities are distribution independent and cannot provide generalization bounds for unbounded loss functions.

To address these issues, researchers have studied more general conditions under which concentration inequalities exist. Specifically, assuming that the function is bounded on one side, [5] gave an extension of Hoeffding's inequality to unbounded random variables with bounded mathematical expectation. [2,6] proved two extensions of McDiarmid's inequality to strongly and weakly difference-bounded functions (see Definitions 2-3 in Section 3) for the study of the generalization error.

[2, 6] assumed that there exist some constant vectors, e.g., b and c, with $b_i \geqslant c_i$ for all $i = 1, 2, \ldots, n$, such that the function f has b bounded differences on a subset set $\mathcal{D}$ of $\mathcal{X}$ and c bounded differences on the complement of the set $\mathcal{D}$. [3] noted that the strong and weak bounded difference conditions proposed by [2, 6] have their limitations in practice and the bounds of the inequalities are uninformative if b is infinite. In order to relax the difference-bounded conditions, [3] introduced the subgaussian diameter and proved an extension of McDiarmid's inequality using the subgaussian diameter. Nevertheless, the approach [3] proposed requires an extra metric on the sample space and that the subgaussian diameter is bounded. Recently, [4] developed more general difference-bounded conditions: the function f has c bounded differences on a high probability set $\mathcal{D}(\subset \mathcal{X})$ and is arbitrary outside of $\mathcal{D}$, the measure of which is controlled by a probability p (this is similar to Assumption 3 in Section 4). Finally, [4] proposed an extension of McDiarmid's inequality:

$$\mathrm{P}\left(|f(X_1, X_2, \ldots, X_n) - \mathrm{E}(f(X_1, X_2, \ldots, X_n|\mathcal{D}))| \geqslant t\right) \leqslant 2 \cdot \left(p + \exp\left(-2((t - p \cdot \overline{c})_+)^2 \Big/ \sum_{i=1}^{n} c_i^2\right)\right) \tag{A1}$$

Here, $\overline{c} = \sum_{i=1}^{n} c_i$, $(t - p \cdot \overline{c})_+ = \max\{t - p \cdot \overline{c}, 0\}$. It is worth pointing out that the above bound in the equation (A1) discussed by [4] can be further tightened.

Different from the boundedness conditions previously discussed, our extensional conditions do not require 1) loss functions to be bounded as in [2], and 2) the extra metrics as in [3]. Roughly speaking, they can be classified into two cases:

■ Being similar to the boundedness conditions given by [4] (see Assumptions 1 and 3 in Section 4). In this case, we will obtain a refined bound.

■ Refining the boundedness and bounded differences (see Assumptions 2 and 4 in Section 4). In this case, we will obtain tighter bound.

Appendix B More Notations and Definitions

In this section, we introduce more notations and definitions as the supplement of the paper.

■ Let I_A denote the indicator function.

■ Let $\mathbb{N}$ be the set of natural numbers, $\mathbb{R}$ be the set of real numbers. Let $\mathbb{N}_n = \{1, 2, \ldots, n\}, n \in \mathbb{N}$.

* Corresponding author (email: jpzhang@fudan.edu.cn)

■ Let $(\Omega, \mathcal{A}, \mathrm{P})$ be a probability space, that is, Ω alone is called the sample space, $\mathcal{A}$ is a σ-algebra on Ω, and P is a probability measure on $(\Omega, \mathcal{A})$. And Ω has the structure $\Omega = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ are the input space and output space respectively. The set $\emptyset$ denotes the empty set.

■ Let $\mathcal{F}$ be the set of all measurable functions $f : \mathcal{X} \longrightarrow \mathcal{Y}$. Assume that $\mathcal{H}$ is a subset of $\mathcal{F}$, i.e., $\mathcal{H} \subset \mathcal{F}$, the set $\mathcal{H}$ is called the hypothesis class.

■ Let $S = \{z_i = (x_i, y_i), i \in \mathbb{N}_n\}$ be a finite set of labeled training samples, and assume that these samples are independent and identically distributed (i.i.d.) according to P. Denote the bold letter as a vector, for example, the bold $\mathbf{z}$ presents a vector $(z_1, z_2, \ldots, z_n)$.

■ Let L be the loss function, $L : \mathcal{Y} \times \mathcal{Y} \longrightarrow [0, +\infty]$, and the loss of f on a sample point $z = (x, y)$ is defined by $Q(f, z) = L(f(x), y)$. We can see that the function Q is nonnegative, but not necessarily bounded. Three well-known examples for this function often used in machine learning domain are the absolute loss $Q(f, z) = |f(x) - y|$, squared loss $Q(f, z) = (f(x) - y)^2$ and log loss $Q(f, z) = -\log p_f(y|x)$ [3, 7]. Here $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and $\{p_f | f \in \mathcal{F}\}$ is a statistical model of conditional densities for $y|x$.

■ Let the symbol $\mathcal{R}_n(\mathcal{H})$ be Rademacher complexity, defined as $\mathrm{E}[\sup_{h \in \mathcal{H}}(1/n) \sum_{i=1}^n \sigma_i h]$, here the random variables $\sigma_i, i = 1, 2, \ldots, n$ are Rademacher variables.

In learning theory, one of the goals is to find a function h in hypothesis space $\mathcal{H}$ that minimizes the following generalization error

$$\mathrm{E}(Q) \triangleq \int_{\Omega} Q(h, z) d\mathrm{P}$$

Generally speaking, the distribution P in the equation $\mathrm{E}(Q)$ is unknown. Rather than minimizing $\mathrm{E}(Q)$, we usually minimize the training error

$$\mathrm{E}_n(Q) \triangleq \frac{1}{n} \cdot \sum_{i=1}^{n} Q(h, z_i)$$

In this paper, we are interested in the uniform estimation of $\mathrm{E}(Q) - \mathrm{E}_n(Q)$.

Definition 1 (Uniformly difference-bounded [2, 6]). Let $g : \prod_{k=1}^n \Omega_k \to \mathbb{R}$ be a function. We say that g is uniformly difference-bounded by $\{c_k, k \in \mathbb{N}_n\}$, if the following holds:

For any $k \in \mathbb{N}_n$, if $\omega, \omega' \in \prod_{k=1}^n \Omega_k$ differ only in the kth coordinate, that is, there exists $\omega_1, \omega_2, \ldots, \omega_n, \omega_k' \in \Omega$, s.t. $\omega = (\omega_1, \omega_2, \ldots, \omega_k, \ldots, \omega_n)$ and $\omega' = (\omega_1, \omega_2, \ldots, \omega_k', \ldots, \omega_n)$, then we have $|g(\omega) - g(\omega')| \leqslant c_k$.

Definition 2 (Strongly difference-bounded [2, 6]). Let $g : \prod_{k=1}^n \Omega_k \to \mathbb{R}$ be a function. We say that g is strongly difference-bounded by $(\{b_k, k \in \mathbb{N}_n\}, \{c_k, k \in \mathbb{N}_n\}, \delta)$, if the following holds:

There exists a "bad" subset $B \subset \prod_{k=1}^n \Omega_k$, where $\delta = \mathrm{P}(B)$. For any $k \in \mathbb{N}_n$, if $\omega, \omega' \in \prod_{k=1}^n \Omega_k$ differ only in the kth coordinate and $\omega \notin B$, then $|g(\omega) - g(\omega')| \leqslant c_k$; if $\omega, \omega' \in \prod_{k=1}^n \Omega_k$ differ only in the kth coordinate, then $|g(\omega) - g(\omega')| \leqslant b_k$.

Definition 3 (Weakly difference-bounded [2, 6]). Let $g : \prod_{k=1}^n \Omega_k \to \mathbb{R}$ be a function. We say that g is weakly difference-bounded by $(\{b_k, k \in \mathbb{N}_n\}, \{c_k, k \in \mathbb{N}_n\}, \delta)$, if the following holds:

For any $j \in \mathbb{N}_n$, we have

$$\mathrm{P}\left((\omega, v) \in \left(\prod_{j=1}^{n} \Omega_j\right) \times \Omega_j | \quad |g(\omega) - g(\omega')| > c_j\right) \leqslant \delta \tag{B1}$$

where $\omega' \in \prod_{k=1}^n \Omega_k$, $\omega_j' = v$ and $\omega_i' = \omega_i$ for $i \neq j$.

For any ω and ω' differing only in the kth coordinate, moreover, $|g(\omega) - g(\omega')| \leqslant b_k$.

Remark 1. The equation (B1) means that if we construct $\omega' \in \prod_{k=1}^n \Omega_k$ by replacing the kth entry of $\omega \in \prod_{k=1}^n \Omega_k$ with v, then $|g(\omega) - g(\omega')| \leqslant c_k$ holds for all but a δ fraction of the choices.

For the discussion of later sections and to be self-contained, we provide the original forms of Hoeffding's and McDiarmid's inequalities as follows:

Hoeffding's inequality [8] Let $X_1, X_2, \ldots, X_n$ be independent random variables on a probability space $(\Omega, \mathcal{A}, \mathrm{P})$, s.t. $X_i \in [a_i, b_i], i = 1, 2, \ldots, n$. Set $S_n = \sum_{i=1}^n X_i$. Then, for all $t \geqslant 0$ we have

$$\mathrm{P}(|S_n - \mathrm{E}(S_n)| \geqslant t) \leqslant 2 \cdot \exp\left(-2t^2 \Big/ \left(\sum_{i=1}^{n} (a_i - b_i)^2\right)\right) \tag{B2}$$

McDiarmid's inequality [9] Let $X_1, X_2, \ldots, X_n$ be independent random variables on a probability space $(\Omega, \mathcal{A}, \mathrm{P})$. Then, for all $t \geqslant 0$ we have

$$\mathrm{P}(|f(X_1, X_2, \ldots, X_n) - \mathrm{E}(f(X_1, X_2, \ldots, X_n))| \geqslant t) \leqslant 2 \cdot \exp\left(-2t^2 \Big/ \left(\sum_{i=1}^{n} c_i^2\right)\right) \tag{B3}$$

where the function f is a real-valued function of the sequence $X_1, X_2, \ldots, X_n$, s.t. $|f(x) - f(x')| \leqslant c_i$, whenever x and x' differ only in the ith coordinate, $i = 1, 2, \ldots, n$. The uniformly difference bounded function f is uniformly difference-bounded.

Appendix C Addressing the Limitations of Previous Concentration Inequalities

In this section, we analyze the conditions of the inequalities in previous works, and show that they have limitations in some examples. At the end of this section, we discuss several general assumptions for the boundedness conditions of concentration inequalities. These are as the supplement of the paper.

Appendix C.1 Limitations in two cases

Here, we analyze two examples to illustrate the limitations of previous concentration inequalities.

Example 1. Let $\omega \in \mathbb{N}$, set $X(\omega) \triangleq \omega \cdot I_{(n_0,+\infty)}(\omega)$. For all $\omega = k, k \in [1,+\infty)$, if we set $\mathrm{P}(\omega = k) \triangleq 6/\left(\pi^2 k^2\right)$, then we have the identity $\sum_{k=1}^{\infty} \mathrm{P}(\omega = k) = \sum_{k=1}^{\infty} \left(6/\pi^2 k^2\right) = 1$. By the above definition of $X(\omega)$, it follows that $\mathrm{E}(X \cdot I_{[1,n_0]}) = \sum_{k=1}^{n_0} \left(0 \cdot 6/\pi^2 k^2\right) = 0$ and $\mathrm{E}(X \cdot I_{(n_0,\infty)}) = \sum_{k=n_0}^{\infty} \left(k \cdot 6/\pi^2 k^2\right) = \left(6/\pi^2\right) \sum_{k=n_0}^{\infty} (1/k) = \infty$.

Remark 2. In this example, the random variable $X(\omega)$ is unbounded, and thus does not satisfy the condition of Hoeffding's inequality. If the random variable $X(\omega)$ is limited in a certain range (for example, $\omega \in [1, n_0]$), the condition of Hoeffding's inequality will be satisfied.

Example 2. Let $\omega \in \mathbb{N}$, set $f(X_1(\omega), X_2(\omega), \ldots, X_n(\omega)) \triangleq M^n I_{\omega=M}(\omega)$. Here, we set $X_i(\omega) = \omega, \omega \in \mathbb{N}_M$, M is a constant, $i \in \mathbb{N}_n$.

For all $\omega = k, k \in \mathbb{N}_M$, we assume that $\mathrm{P}(\omega = k) \triangleq 1/M$ where M presents the cardinality of $\mathbb{N}_M$. By the above definition of $f(X_1(\omega), X_2(\omega), \ldots, X_n(\omega))$, it follows that $\mathrm{E}(f(X_1, X_2, \ldots, X_n) \cdot I_{[1,M-1]}) = \sum_{k=1}^{M-1} (0 \cdot 1/M) = 0$ and $\mathrm{E}(f(X_1, X_2, \ldots, X_n) \cdot I_M) = M^n \cdot 1/M^n = 1$.

Remark 3. In this example, we know that f is not uniformly differences-bounded, failing to the condition of McDiarmid's inequality. But if the sample space is limited in a certain range (for example, $\omega \in [1, M-1]$), the condition of McDiarmid's inequality is satisfied. In addition, we observe that f is neither weakly differences-bounded nor strongly differences-bounded. For such a case, we will introduce some new assumptions in the next part.

Appendix C.2 Assumptions

The aforementioned examples 1 and 2 do not satisfy various existing definitions of the boundedness and bounded differences. A possible reason behind is that these existing definitions either are sort of restrictive or neglect the robustness of learning theory in some cases. Therefore, four assumptions below are proposed to alleviate these issues.

Assumption 1 (p_i bounded)**.** Let X_i be the independent random variable on a probability $(\Omega_i, \mathcal{A}_i, \mathrm{P}_i), i \in \mathbb{N}_n$, s.t. $\mathrm{P}_i(a_i \leqslant X_i(\omega) \leqslant b_i) = p_i, i \in \mathbb{N}_n$. If this is true, then we say that X_i is p_i bounded by the pair $(a_i, b_i), i \in \mathbb{N}_n$.

Assumption 2 ((p_{ij}, k) hierarchy-bounded)**.** Let X_i be the independent random variable on a probability $(\Omega_i, \mathcal{A}_i, \mathrm{P}_i)$, $i \in \mathbb{N}_n$, s.t. there exists an integer $k > 1$, we have $\mathrm{P}_i(a_{ij} \leqslant X_i(\omega) \leqslant b_{ij}) = p_{ij}, j \in \mathbb{N}_k, i \in \mathbb{N}_n$ and $\mathrm{P}_i(\bigcup_{j=1}^{k}(a_{ij} \leqslant X_i(\omega) \leqslant b_{ij})) = 1, i \in \mathbb{N}_n$. If this is true, then X_i is (p_{ij}, k) hierarchy-bounded by the pair $(a_{ij}, b_{ij}), j \in \mathbb{N}_k, i \in \mathbb{N}_n$.

Assumption 3 (p difference-bounded)**.** [1)] Let $g : \prod_{i=1}^{n} \Omega_i \to \mathbb{R}$ be a function, s.t. for any $\ell \in \mathbb{N}_n$, $\exists A \subset \prod_{i=1}^{n} \Omega_i$, for any $\omega \in A$ and $\omega' \in A$ differ only in the ℓth coordinate, we have $|g(\omega) - g(\omega')| \leqslant c_\ell$ and $\mathrm{P}(A) = p$. If this is true, then g is p difference-bounded by $\{c_\ell, \ell \in \mathbb{N}_n\}$.

Assumption 4 ((p_j, k) hierarchy-difference-bounded)**.** Let $g : \prod_{i=1}^{n} \Omega_i \to \mathbb{R}$ be a function, s.t. for any $\ell \in \mathbb{N}_n$, there exists an integer $k > 1$, $A_j \subset \prod_{i=1}^{n} \Omega_i, j \in \mathbb{N}_k$, $\bigcup_{j=1}^{k} A_j = \prod_{i=1}^{n} \Omega_i$ and $A_i \cap A_j = \emptyset, i \neq j, i, i, j \in \mathbb{N}_n$, for any $\omega \in A_j$ and $\omega' \in A_j$ differ only in the ℓth coordinate, we have $|g(\omega) - g(\omega')| \leqslant c_{\ell j}$ and $\mathrm{P}(A_j) = p_j$. If this is true, then g is (p_j, k) hierarchy-difference-bounded by $\{c_{\ell j}, \ell \in \mathbb{N}_n\}, j \in \mathbb{N}_k$.

Under Assumptions 1 and 3, we study if Hoeffding's and McDiarmid's inequalities still hold. Under Assumptions 2 and 4, we investigate if the convergence bounds of Hoeffding's and McDiarmid's inequalities can be better.

Actually, it is not difficult to see from Example 2 that it does not always concentrate around its expectation. For instance, $\mathrm{P}(f(X_1, X_2, \ldots, X_n) = 1) = 1/M^n$ will be close to 0 as n tends to infinity. Therefore, Hoeffding's and McDiarmid's inequalities for such functions in Example 2 do not hold. Based on these assumptions, we will present several similar Hoeffding's and McDiarmid's inequalities (see Theorems 1, 2 and Corollaries 1, 2 in Section 5), which can deal with unbounded and hierarchy-bounded functions.

At the end of this section, we compare four proposed bounded conditions with the previous three definitions in Section 2. Since the sums of random variables can be regarded as a special case of a multivariate random function, we will only discuss the relationships between the p difference-bounded, (p_j, k) hierarchy-difference-bounded, uniformly differences-bounded, strongly differences-bounded and weakly differences-bounded conditions. Next, a proposition is given on the pairwise comparisons among the four conditions.

Proposition 1.

$$
\begin{aligned}
& (p_j, k) \quad \text{hierarchy} - \text{difference} - \text{bounded} \\
\Longrightarrow\ & \text{uniformly} \quad \text{differences} - \text{bounded} \\
\Longrightarrow\ & \text{strongly} \quad \text{differences} - \text{bounded} \\
\Longrightarrow\ & \text{weakly} \quad \text{differences} - \text{bounded} \\
\Longrightarrow\ & p \quad \text{difference} - \text{bounded}
\end{aligned}
$$

Here the symbol "$\Longrightarrow$" means that the item is strictly stronger on the left than on the right.

Proof. 1) The p difference-bounded condition and the weakly differences-bounded condition

First, if the p difference-bounded condition is satisfied, we set $A' = (\prod_{i=1}^{n} \Omega_i) \backslash A$, there exists a set $A_0 \subset A'$, for any $\omega \in A_0$ and $\omega' \in A_0$ differ only in the ℓth coordinate. Then, for any c'_ℓ, the equality $\mathrm{P}(|g(\omega) - g(\omega')| > c'_\ell) > 0$ holds.

So, the p difference-bounded condition is not stronger than the weakly differences-bounded condition.

1) This assumption is similar to the assumption in [4].

For the other hand of this comparison, it is obvious to see that the p difference-bounded condition can be inferred from the weakly differences-bounded condition.

Therefore, the p difference-bounded condition is strictly weaker than the weakly differences-bounded condition. That is, weakly differences-bounded $\Longrightarrow$ p difference-bounded.

2) The (p_j, k) hierarchy-difference-bounded condition and the uniformly differences-bounded condition

It is obvious that if the uniformly differences-bounded condition is true, it doesn't guarantee the truth of the (p_j, k) hierarchy-difference-bounded condition.

We now turn to the other hand of this comparison.

We assume that the (p_j, k) hierarchy-difference-bounded condition is satisfied. According to the Assumption 1, it follows that, for any $\ell \in \mathbb{N}_n$, there exists a set $A_j \subset \prod_{i=1}^n \Omega_i, j \in \mathbb{N}_k$, s.t. $\bigcup_{j=1}^k A_j = \prod_{i=1}^n \Omega_i$ and $A_i \cap A_j = \emptyset, i \neq j, i, j \in \mathbb{N}_n$. For any $\omega \in A_j$ and $\omega' \in A_j$ differ only in the ℓth coordinate, we have $|g(\omega) - g(\omega')| \leqslant c_{\ell j}$ and $\mathrm{P}(A_j) = p_j$.

Take $c_\ell = \max_{j \in \mathbb{N}_n} c_{\ell j}$, for any $\omega \in \bigcup_{j=1}^k A_j$ and $\omega' \in \bigcup_{j=1}^k A_j$ differ only in the ℓth coordinate, we have $|g(\omega) - g(\omega')| \leqslant c_\ell$.

Therefore, the uniformly differences-bounded condition is strictly weaker than the (p_j, k) hierarchy-difference-bounded condition. That is, (p_j, k) hierarchy-difference-bounded $\Longrightarrow$ uniformly differences-bounded.

3) The weakly differences-bounded condition and the strongly differences-bounded condition

It is obvious that the weakly differences-bounded condition is strictly weaker than the strongly differences-bounded condition. That is, strongly differences-bounded $\Longrightarrow$ weakly differences-bounded.

4) The strongly differences-bounded condition and the uniformly differences-bounded condition

If we set $B = \emptyset$ in Definition 2, then we have the uniformly differences-bounded condition. Therefore, it is obvious that the strongly differences-bounded condition is strictly weaker than the uniformly differences-bounded condition. That is, uniformly differences-bounded $\Longrightarrow$ strongly differences-bounded.

The proof of Proposition 1 is completed.

Remark 4. From Proposition 1, it tells that the p difference-bounded condition is the weakest and the (p_j, k) hierarchy-difference-bounded condition is the strongest. The comparisons between p difference-bounded and (p_j, k) hierarchy-difference-bounded can be directly analyzed as follows:

First, in Assumption 3, we set $A' = (\prod_{i=1}^n \Omega_i) \backslash A$.

If for any $\omega \in A'$ and $\omega' \in A'$ differ only in the ℓth coordinate, we assume that there exists a set $A_0 \subset A'$, for any $\omega \in A_0$ and $\omega' \in A_0$ differ only in the ℓth coordinate, and for any c'_ℓ, the equality $\mathrm{P}(|g(\omega) - g(\omega')| > c'_\ell) > 0$ holds. Then we have that the (p_j, k) hierarchy-difference-bounded condition doesn't hold.

So, the p difference-bounded condition is not stronger than the (p_j, k) hierarchy-difference-bounded condition.

Now, we turn to the other hand of this comparison.

If the (p_j, k) hierarchy-difference-bounded condition holds, then we have that there exits a constant $j_0 \in \mathbb{N}_k$, for any $\omega \in A_{j_0}$ and $\omega' \in A_{j_0}$ differ only in the ℓth coordinate. Then we have $|g(\omega) - g(\omega')| \leqslant c_{lj_0}$ and $\mathrm{P}(A_{j_0}) = p_{j_0}$.

Therefore, the p difference-bounded condition is strictly weaker than the (p_j, k) hierarchy-difference-bounded condition. That is, (p_j, k) hierarchy-difference-bounded $\Longrightarrow$ p difference-bounded.

Appendix D Detailed Proofs for Extensions of Hoeffding's and McDiarmid's Inequalities

In this section, we will give the detailed proofs for extensions to Hoeffding's and McDiarmid's inequalities.

Essentially, Hoeffding's inequality (B2) in Section 3 can be proved by combining the properties of convex functions, Taylor expansion, the monotonicity of probability measures, the exponential Markov inequality and the independence of random variables. Meanwhile, McDiarmid's inequality (B3) in Section 3 can be proved by constructing the martingale difference sequences in combination with a similar proof of Hoeffding's inequality. To extend these two concentration inequalities, we prove Theorems 1 and 2 using conditional mathematical expectation. We assume that X_i is the independent random variable on a probability space $(\Omega_i, \mathcal{A}_i, \mathrm{P}_i), i \in \mathbb{N}_n$, and give the following condition:

Condition 1 (Partition of product space). Let $\cup_{j=1}^k A_{ij} = \Omega_i, A_{ij'} \cap A_{ij''} = \emptyset, j' \neq j'', j'', j', j \in \mathbb{N}_k, i \in \mathbb{N}_n$. Set $\Psi_{n,k} = \{\prod_{i=1}^n A_{ij} | A_{ij} \in \{A_{i1}, A_{i2}, \ldots, A_{ik}\}\}$, the set $\Psi_{n,k}$ is called a partition of $\prod_{i=1}^n \Omega_i$.

We first provide two lemmas which will be used later. For avoiding distraction, the detailed proofs of lemmas are deferred to Appendix.

Lemma 1. Assume that X_i is $(\mathrm{P}_i(A_{ij}), k)$ hierarchy-bounded by the pair (a_{ij}, b_{ij}), and Condition 1 holds. Let $\psi_{n,k} = \{(j_1, j_2, \ldots, j_n) | j_r \in \mathbb{N}_k, r = 1, 2, \ldots, n\}$. Set $S_n = \sum_{i=1}^n X_i$ and $\widetilde{A}_{\mathbf{j}} = \prod_{i=1}^n A_{ij} \in \Psi_{n,k}$, $j \in \mathbb{N}_k$, $\mathbf{j} \in \psi_{n,k}$. Then, for any $t > 0$, we have

$$\mathrm{P}\left(\left(\left|S_n - \sum_{\mathbf{j} \in \psi_{\mathbf{n},\mathbf{k}}} \mathrm{E}(S_n | \widetilde{A}_{\mathbf{j}}) I_{\widetilde{A}_{\mathbf{j}}}\right| \geqslant t\right) \cap \widetilde{A}_{\mathbf{j_0}}\right) \leqslant 2 \cdot \mathrm{P}(\widetilde{A}_{\mathbf{j_0}}) \cdot \exp\left(-2t^2 \Big/ \sum_{i=1}^n (a_{i\mathbf{j_0}} - b_{i\mathbf{j_0}})^2\right) \tag{D1}$$

where $\mathbf{j_0} = (j_{10}, j_{20}, \ldots, j_{n0}) \in \psi_{n,k}$ is a constant vector, and we agreed that $a_{i\mathbf{j_0}} = a_{ij_{i0}}$ and $b_{i\mathbf{j_0}} = b_{ij_{i0}}$ $i \in \mathbb{N}_n$.

Proof. From the assumptions, for any $j \in \mathbb{N}_k, i \in \mathbb{N}_n, \omega \in A_{ij}$, we have $a_{ij} \leqslant X_i(\omega) \leqslant b_{ij}$ and $\cup_{j=1}^k A_{ij} = \Omega_i$. By definition of conditional mathematical expectation and the additivity and monotonicity of the probability measure, for any $t \geqslant 0$ and $s > 0$, by taking a constant vector $\mathbf{j_0}$, we have

$$
\begin{aligned}
&\mathrm{P}\left(\left(S_n-\sum_{\mathbf{j}\in\psi_{\mathbf{n},\mathbf{k}}}\mathrm{E}(S_n|\widetilde{A}_{\mathbf{j}})I_{\widetilde{A}_{\mathbf{j}}}\geqslant t\right)\cap\widetilde{A}_{\mathbf{j0}}\right)\\
=&\int_{\left(S_n-\sum\limits_{\mathbf{j}\in\psi_{\mathbf{n},\mathbf{k}}}\mathrm{E}(S_n|\widetilde{A}_{\mathbf{j}})I_{\widetilde{A}_{\mathbf{j}}}\geqslant t\right)\cap\widetilde{A}_{\mathbf{j0}}} I_\Omega\mathrm{dP}\\
=&\int_{\left(\exp\left(s\cdot\left(S_n-\sum\limits_{\mathbf{j}\in\psi_{\mathbf{n},\mathbf{k}}}\mathrm{E}(S_n|\widetilde{A}_{\mathbf{j}})I_{\widetilde{A}_{\mathbf{j}}}\right)\right)\geqslant\exp(st)\right)\cap\widetilde{A}_{\mathbf{j0}}} I_\Omega\mathrm{dP}\\
\leqslant&\int_{\left(\exp\left(s\cdot\left(S_n-\sum\limits_{\mathbf{j}\in\psi_{\mathbf{n},\mathbf{k}}}\mathrm{E}(S_n|\widetilde{A}_{\mathbf{j}})I_{\widetilde{A}_{\mathbf{j}}}\right)\right)\geqslant\exp(st)\right)\cap\widetilde{A}_{\mathbf{j0}}} \frac{\exp\left(s\cdot\left(S_n-\sum\limits_{\mathbf{j}\in\psi_{\mathbf{n},\mathbf{k}}}\mathrm{E}(S_n|\widetilde{A}_{\mathbf{j}})I_{\widetilde{A}_{\mathbf{j}}}\right)\right)}{\exp(st)}\mathrm{dP}\\
\leqslant&\int_{\widetilde{A}_{\mathbf{j0}}}\frac{\exp\left(s\cdot\left(S_n-\sum\limits_{\mathbf{j}\in\psi_{\mathbf{n},\mathbf{k}}}\mathrm{E}(S_n|\widetilde{A}_{\mathbf{j}})I_{\widetilde{A}_{\mathbf{j}}}\right)\right)}{\exp(st)}\mathrm{dP}\\
=&\exp(-st)\int_{\prod\limits_{i=1}^n A_{ij_{i0}}}\prod_{i=1}^n\exp\left(s\cdot\left(X_i-\sum_{\mathbf{j}\in\psi_{\mathbf{n},\mathbf{k}}}\mathrm{E}(X_i|\widetilde{A}_{\mathbf{j}})I_{\widetilde{A}_{\mathbf{j}}}\right)\right)\mathrm{dP}\\
=&\exp(-st)\prod_{i=1}^n\int_{A_{ij_{i0}}}\exp\left(s\cdot\left(X_i-\sum_{\mathbf{j}\in\psi_{\mathbf{n},\mathbf{k}}}\mathrm{E}(X_i|\widetilde{A}_{\mathbf{j}})I_{\widetilde{A}_{\mathbf{j}}}\right)\right)\mathrm{dP}\\
\leqslant&\exp(-st)\prod_{i=1}^n\int_{A_{ij_{i0}}}\left(\frac{b_{ij_{i0}}-\left(X_i-\sum\limits_{\mathbf{j}\in\psi_{\mathbf{n},\mathbf{k}}}\mathrm{E}(X_i|\widetilde{A}_{\mathbf{j}})I_{\widetilde{A}_{\mathbf{j}}}\right)}{b_{ij_{i0}}-a_{ij_{i0}}}\cdot\exp\left(s\cdot a_{ij_{i0}}\right)\right.\\
&\left.+\frac{X_i-\sum\limits_{\mathbf{j}\in\psi_{\mathbf{n},\mathbf{k}}}\mathrm{E}(X_i|\widetilde{A}_{\mathbf{j}})I_{\widetilde{A}_{\mathbf{j}}}-a_{ij_{i0}}}{b_{ij_{i0}}-a_{ij_{i0}}}\cdot\exp\left(s\cdot b_{ij_{i0}}\right)\right)\mathrm{dP}_i\\
=&\exp(-st)\prod_{i=1}^n\int_{A_{ij_{i0}}}\left(\frac{b_{ij_{i0}}}{b_{ij_{i0}}-a_{ij_{i0}}}\cdot\exp\left(s\cdot a_{ij_{i0}}\right)+\frac{-a_{ij_{i0}}}{b_{ij_{i0}}-a_{ij_{i0}}}\cdot\left(s\cdot b_{ij_{i0}}\right)\right)\mathrm{dP}_i\\
\leqslant&\exp(-st)\cdot\mathrm{P}\left(\widetilde{A}_{\mathbf{j0}}\right)\cdot\prod_{i=1}^n\exp\left(\frac{s^2(b_{i\mathbf{j0}}-a_{i\mathbf{j0}})^2}{8}\right)
\end{aligned}
\tag{D2}
$$

Take $s=4t\big/\sum_{i=1}^n(b_{i\mathbf{j0}}-a_{i\mathbf{j0}})^2$, the inequality (D2) gets the minimum value

$$\mathrm{P}(\widetilde{A}_{\mathbf{j0}})\cdot\exp\cdot\left(-2t^2\Big/\sum_{i=1}^n(a_{i\mathbf{j0}}-b_{i\mathbf{j0}})^2\right)$$

Finally, we have

$$\mathrm{P}\left(\left(S_n-\sum_{\mathbf{j}\in\psi_{\mathbf{n},\mathbf{k}}}\mathrm{E}(S_n|\widetilde{A}_{\mathbf{j}})I_{\widetilde{A}_{\mathbf{j}}}\geqslant t\right)\cap\widetilde{A}_{\mathbf{j0}}\right)\leqslant\mathrm{P}(\widetilde{A}_{\mathbf{j0}})\cdot\exp\left(-2t^2\Big/\sum_{i=1}^n(a_{i\mathbf{j0}}-b_{i\mathbf{j0}})^2\right)$$

The proof of Lemma 1 is completed.

Similarly, we have the following Lemma 2.

Lemma 2. Let the function f be a map from $\mathcal{X}^n$ to $\mathbb{R}$. Assume that f is $\mathrm{P}(A)$ difference-bounded by $\{c_m, m\in\mathbb{N}_n\}$. Set $A=\prod_{i=1}^n A_i$. Then, for any $t>0$, we have

$$\mathrm{P}\left((f(X_1,\ldots,X_i,\ldots,X_n)-\mathrm{E}(f(X_1,\ldots,X_i,\ldots,X_n)|A)\geqslant t)\cap A\right)\leqslant\mathrm{P}(A)\cdot\exp\left(-2t^2\Big/\sum_{i=1}^n c_i^2\right)\tag{D3}$$

Proof. Let $V_i=\mathrm{E}(f|X_1,X_2,\ldots,X_i)-\mathrm{E}(f|X_1,X_2,\ldots,X_{i-1}), i\in\mathbb{N}_n$, $\hat{V}_i=V_i\cdot I_{\prod_{\ell=1}^i A_\ell}, i\in\mathbb{N}_n$. Assume that $\mathrm{E}(f|X_0)=\mathrm{E}(f|A)$, then we have

$$\sum_{i=1}^n V_i=\mathrm{E}(f|X_1,X_2,\ldots,X_n)-\mathrm{E}(f|X_0)=f(X_1,X_2,\ldots,X_n)-\mathrm{E}(f|A)$$

and

$$\sum_{i=1}^n\hat{V}_i=\mathrm{E}(f|X_1,X_2,\ldots,X_n)\cdot I_A-\mathrm{E}(f|X_0)\cdot I_A=f(X_1,X_2,\ldots,X_n)\cdot I_A-\mathrm{E}(f|A)\cdot I_A$$

Notice that

$$\begin{aligned}
&\mathrm{E}\left(\sum_{i=1}^{n}\hat{V}_i|\hat{V}_1,\hat{V}_2,\ldots,\hat{V}_{n-1}\right)\\
=\ &\mathrm{E}\left(\left(\sum_{i=1}^{n-1}\hat{V}_i+\hat{V}_n\right)|\hat{V}_1,\hat{V}_2,\ldots,\hat{V}_{n-1}\right)\\
=\ &\sum_{i=1}^{n-1}\hat{V}_i+\mathrm{E}\left(\hat{V}_n|\hat{V}_1,\hat{V}_2,\ldots,\hat{V}_{n-1}\right)
\end{aligned}$$

$$\begin{aligned}
=\ &\sum_{i=1}^{n-1}\hat{V}_i+\mathrm{E}\left(\mathrm{E}\left(f|X_1,X_2,\ldots,X_n\right)\cdot I_A-\mathrm{E}\left(f|X_1,X_2,\ldots,X_{n-1}\right)\cdot I_A|\hat{V}_1,\hat{V}_2,\ldots,\hat{V}_{n-1}\right)\\
=\ &\sum_{i=1}^{n-1}\hat{V}_i
\end{aligned}$$

therefore, it follows that $\sum\limits_{i=1}^{n}\hat{V}_i$ is a martingale.

Let

$$L_i=\inf_{x\in\{X_i(\omega),\omega\in A_i\}}\left\{\mathrm{E}(f|X_1,X_2,\ldots,X_{i-1},x)\cdot I_{\prod_{\ell=1}^{i}A_\ell}-\mathrm{E}(f|X_1,X_2,\ldots,X_{i-1})\cdot I_{\prod_{\ell=1}^{i}A_\ell}\right\}$$

and

$$U_i=\sup_{x\in\{X_i(\omega),\omega\in A_i\}}\left\{\mathrm{E}(f|X_1,X_2,\ldots,X_{i-1},x)\cdot I_{\prod_{\ell=1}^{i}A_\ell}-\mathrm{E}(f|X_1,X_2,\ldots,X_{i-1})\cdot I_{\prod_{\ell=1}^{i}A_\ell}\right\}$$

where $i\in\mathbb{N}_n$.

Note that $L_i\leqslant\widetilde{V}_i\leqslant U_i$ and

$$\begin{aligned}
&U_i-L_i\\
\leqslant\ &\sup_{x\in\{X_i(\omega),\omega\in A_i\}}\left\{\mathrm{E}(f|X_1,X_2,\ldots,X_{i-1},x)\cdot I_{\prod_{\ell=1}^{i}A_\ell}-\mathrm{E}(f|X_1,X_2,\ldots,X_{i-1})\cdot I_{\prod_{\ell=1}^{i}A_\ell}\right\}\\
&-\inf_{x\in\{X_i(\omega),\omega\in A_i\}}\left\{\mathrm{E}(f|X_1,X_2,\ldots,X_{i-1},x)\cdot I_{\prod_{\ell=1}^{i}A_\ell}-\mathrm{E}(f|X_1,X_2,\ldots,X_{i-1})\cdot I_{\prod_{\ell=1}^{i}A_\ell}\right\}\\
=\ &\sup_{x\in\{X_i(\omega),\omega\in A_i\}}\left\{\mathrm{E}(f|X_1,X_2,\ldots,X_{i-1},x)\cdot I_{\prod_{\ell=1}^{i}A_\ell}\right\}
\end{aligned}$$

$$\begin{aligned}
&-\inf_{x\in\{X_i(\omega),\omega\in A_i\}}\left\{\mathrm{E}(f|X_1,X_2,\ldots,X_{i-1},x)\cdot I_{\prod_{\ell=1}^{i}A_\ell}\right\}\\
=\ &\sup_{x,y\in\{X_i(\omega),\omega\in A_i\}}(\int f(X_1,X_2,\ldots,X_{i-1},x)\cdot I_{\prod_{\ell=1}^{i}A_\ell}-\int f(X_1,X_2,\ldots,X_{i-1},y)\cdot I_{\prod_{\ell=1}^{i}A_\ell})\mathrm{d}\prod_{\ell=i+1}^{n}\mathrm{P}_\ell\\
&\left(\text{or}\,\mathrm{P}_{X_1,\ldots,X_i}(x_{i+1},\ldots,x_n)\right)\\
\leqslant\ &c_i
\end{aligned}$$

Then, we have

$$\begin{aligned}
&\mathrm{P}((f(X_1,X_2,\ldots,X_i,\ldots,X_n)-\mathrm{E}(f(X_1,X_2,\ldots,X_i,\ldots,X_n)|A)\geqslant t)\cap A)\\
=\ &\int_{(f(X_1,X_2\ldots,X_i,\ldots,X_n)-\mathrm{E}(f(X_1,X_2\ldots,X_i,\ldots,X_n)|A)\geqslant t)\cap A}\cdot I_\Omega\mathrm{dP}\\
=\ &\int_{\left(\sum\limits_{i=1}^{n}V_i\geqslant t\right)\cap A}\cdot I_\Omega\mathrm{dP}\\
=\ &\int_{\left(\sum\limits_{i=1}^{n}V_i\geqslant t\right)\cap A}\cdot I_A\mathrm{dP}\\
\leqslant\ &\int_A\frac{\exp\left(s\sum\limits_{i=1}^{n}\widetilde{V}_i\right)}{\exp\left(st\right)}\mathrm{dP}
\end{aligned}$$

$$
\begin{aligned}
&\leqslant \exp(-st)\cdot\int_A\prod_{i=1}^{n}\exp\left(s\widetilde{V}_i\right)\mathrm{dP}\\
&= \exp(-st)\cdot\mathrm{E}\left(\left(\prod_{i=1}^{n}\exp\left(s\widetilde{V}_i\right)\right)\cdot I_A\right)\\
&= \exp(-st)\cdot\mathrm{EE}\left(\left(\left(\prod_{i=1}^{n}\exp\left(s\widetilde{V}_i\right)\right)\cdot I_A\right)|\widetilde{V}_1,\widetilde{V}_2,\ldots,\widetilde{V}_{n-1}\right)\\
&= \exp(-st)\cdot\mathrm{E}\left(\left(\prod_{i=1}^{n-1}\exp\left(s\widetilde{V}_i\right)\right)\cdot I_A\right)\cdot\mathrm{E}\left(\left(\exp\left(s\widetilde{V}_n\right)\cdot I_A\right)|\widetilde{V}_1,\widetilde{V}_2,\ldots,\widetilde{V}_{n-1}\right)\\
&= \exp(-st)\cdot\mathrm{E}\left(\left(\prod_{i=1}^{n-1}\exp\left(s\widetilde{V}_i\right)\right)\cdot I_A\right)\cdot\mathrm{E}\left(I_{A_n}|\widetilde{V}_1,\widetilde{V}_2,\ldots,\widetilde{V}_{n-1}\right)\cdot\exp\left(\frac{s^2c_n^2}{8}\right)\\
&= \exp(-st)\cdot\mathrm{E}\left(\left(\prod_{i=1}^{k}\exp\left(s\widetilde{V}_i\right)\right)\cdot I_A\right)\cdot\prod_{\ell=k}^{n}\mathrm{E}\left(I_{A_l}|\widetilde{V}_1,\widetilde{V}_2,\ldots,\widetilde{V}_{\ell-1}\right)\cdot\exp\left(\frac{s^2c_l^2}{8}\right)\\
&\quad(k\in\mathbb{N}_n)\\
&= \exp(-st)\cdot\prod_{\ell=1}^{n}\mathrm{E}\left(I_{A_l}|\widetilde{V}_1,\widetilde{V}_2,\ldots,\widetilde{V}_{\ell-1}\right)\cdot\exp\left(\frac{s^2c_l^2}{8}\right)\\
&= \exp(-st)\cdot\mathrm{P}(A)\cdot\prod_{i=1}^{n}\exp\left(\frac{s^2c_i^2}{8}\right)
\end{aligned}
\tag{D4}
$$

Take $s=4t/\sum_{i=1}^{n}c_i^2$, the inequality (2) gets the minimum value

$$\mathrm{P}(A)\cdot\exp\left(-2t^2/\sum_{i=1}^{n}c_i^2\right)$$

Finally, we have

$$
\begin{aligned}
&\mathrm{P}(|\{f(X_1,X_2,\ldots,X_i,\ldots,X_n)-\mathrm{E}(f(X_1,X_2,\ldots,X_i,\ldots,X_n)|A)|\leqslant -t\}\cap A)\\
&\leqslant 2\cdot\mathrm{P}(A)\cdot\exp\left(\frac{-2t^2}{\sum\limits_{i=1}^{n}c_i^2}\right)
\end{aligned}
$$

The proof of Lemma 2 is completed.

Denote $\mathcal{B}$ as the $\sigma-$algebra $\sigma(\Psi_{n,k})$, and $\mathcal{B}\subset\prod_{i=1}^{n}\mathcal{A}_i$.

Theorem 1. Under the assumptions of Lemma 1, for any $t>0$, we have

$$\mathrm{P}(|S_n-\mathrm{E}(S_n|\mathcal{B})|\geqslant t)\leqslant 2\cdot\sum_{\mathbf{j}\in\psi_{\mathbf{n},\mathbf{k}}}\mathrm{P}(\widetilde{A}_{\mathbf{j}})\cdot\exp\left(-2t^2\Big/\sum_{i=1}^{n}(a_{i\mathbf{j}}-b_{i\mathbf{j}})^2\right)\tag{D5}$$

where we agreed that $a_{i\mathbf{j}}=a_{ij_i}$ and $b_{i\mathbf{j}}=b_{ij_i}$, $i\in\mathbb{N}_n$.

Proof. From the assumptions, we have

$$\bigcup_{B\in\Psi_{n,k}}B=\bigcup_{\mathbf{j}\in\psi_{\mathbf{n},\mathbf{k}}}\widetilde{A}_{\mathbf{j}}=\prod_{i=1}^{n}\Omega_i\tag{D6}$$

By the equation (D6) and definition of conditional mathematical expectation and the additivity and monotonicity of the probability measure, for any $t\geqslant 0$, we have

$$\mathrm{P}(S_n-\mathrm{E}(S_n|\mathcal{B})\geqslant t)=\sum_{\mathbf{j}'\in\psi_{\mathbf{n},\mathbf{k}}}\mathrm{P}(\{S_n-\sum_{\mathbf{j}\in\psi_{\mathbf{n},\mathbf{k}}}\mathrm{E}(S_n|\widetilde{A}_{\mathbf{j}})I_{\widetilde{A}_{\mathbf{j}}}\geqslant t\}\cap\widetilde{A}_{\mathbf{j}'})$$

By Lemma 1, it is easy to show that Theorem 1 holds.

Theorem 2. Let the function f be a map from $\mathcal{X}^n$ to $\mathbb{R}$. Assume that f is $(\mathrm{P}_i(A_{ij}),k)$ hierarchy-difference-bounded by $\{c_{mj},m\in\mathbb{N}_n\},j\in\mathbb{N}_k$, and Condition 1 holds. Let $\psi_{n,k}=\{(j_1,j_2,\ldots,j_n)|j_r\in\mathbb{N}_k,r=1,2,\ldots,n\}$. Set $\widetilde{A}_{\mathbf{j}}=\prod_{i=1}^{n}A_{ij}\in\Psi_{n,k}$, $j\in\mathbb{N}_k$, $\mathbf{j}\in\psi_{n,k}$. For any $t>0$, we have

$$\mathrm{P}(|f(X_1,\ldots,X_n)-\mathrm{E}(f(X_1,\ldots,X_n)|\mathcal{B})|\geqslant t)\leqslant 2\cdot\sum_{\mathbf{j}\in\psi_{\mathbf{n},\mathbf{k}}}\mathrm{P}(\widetilde{A}_{\mathbf{j}})\cdot\exp\left(-2t^2\Big/\sum_{i=1}^{n}c_{m\mathbf{j}}^2\right)\tag{D7}$$

where we agreed that $c_{i\mathbf{j}}=c_{mj_m},m\in\mathbb{N}_n$.

Proof. By combining the Lemma 2 and the methods employed in Theorem 1, Theorem 2 can be easily proved.

From the above Theorems, we have the following corollaries:

Corollary 1. Assume that X_i is $\mathrm{P}_i(A_i)$ bounded by the pair $(a_i,b_i),i\in\mathbb{N}_n$. Set $S_n=\sum_{i=1}^{n}X_i$ and $A=\prod_{i=1}^{n}A_i$. Then, for any $t>0$, we have

$$\mathrm{P}(|S_n-\mathrm{E}(S_n|A)|\geqslant t)\leqslant 2\cdot\mathrm{P}(A)\cdot\exp\left(-2t^2\Big/\sum_{i=1}^{n}(a_i-b_i)^2\right)+1-\mathrm{P}(A)\tag{D8}$$

Proof. By the additivity and monotonicity of the probability measure, for any $t \geqslant 0$, we have

$$\begin{aligned} & \mathrm{P}(|S_n - \mathrm{E}(S_n|A)| \geqslant t) \\ = \ & \mathrm{P}(\{|S_n - \mathrm{E}(S_n|A)| \geqslant t\} \cap (A \cup A^c)) \\ = \ & \mathrm{P}(\{|S_n - \mathrm{E}(S_n|A)| \geqslant t\} \cap A) + \mathrm{P}(|\{S_n - \mathrm{E}(S_n|A)| \geqslant t\} \cap A^c) \\ \leqslant \ & \mathrm{P}(\{|S_n - \mathrm{E}(S_n|A)| \geqslant t\} \cap A) + \mathrm{P}(A^c) \\ = \ & \mathrm{P}(\{|S_n - \mathrm{E}(S_n|A)| \geqslant t\} \cap A) + 1 - \mathrm{P}(A) \end{aligned} \tag{D9}$$

By Lemma 1, we substitute the inequality (D1) into the inequality (D9), which completes the proof of Corollary 1.

Corollary 2. Under the assumptions of Lemma 2. Then, for any $t > 0$, we have

$$= \mathrm{P}(|f(X_1, \ldots, X_n) - \mathrm{E}(f(X_1, \ldots, X_n)|A)| \geqslant t) \leqslant 2 \cdot \mathrm{P}(A) \cdot \exp\left(-2t^2 \Big/ \sum_{i=1}^{n} c_i^2\right) + 1 - \mathrm{P}(A) \tag{D10}$$

Proof. By the additivity and monotonicity of the probability measure, we have, for any $t \geqslant 0$

$$\begin{aligned} & \mathrm{P}(|f(X_1, \ldots, X_i, \ldots, X_n) - \mathrm{E}(f(X_1, \ldots, X_i, \ldots, X_n)|A)| \geqslant t) \\ = \ & \mathrm{P}(|\{f(X_1, \ldots, X_i, \ldots, X_n) - \mathrm{E}(f(X_1, \ldots, X_i, \ldots, X_n)|A)| \geqslant t\} \cap (A \cup A^c)) \\ = \ & \mathrm{P}(|\{f(X_1, \ldots, X_i, \ldots, X_n) - \mathrm{E}(f(X_1, \ldots, X_i, \ldots, X_n)|A)| \geqslant t\} \cap A) \\ & + \mathrm{P}(|\{f(X_1, \ldots, X_i, \ldots, X_n) - \mathrm{E}(f(X_1, \ldots, X_i, \ldots, X_n)|A)| \geqslant t\} \cap A^c) \\ = \ & \mathrm{P}(|\{f(X_1, \ldots, X_i, \ldots, X_n) - \mathrm{E}(f(X_1, \ldots, X_i, \ldots, X_n)|A)| \geqslant t\} \cap A) + (1 - \mathrm{P}(A)) \end{aligned} \tag{D11}$$

By Lemma 2, we substitute the inequality (D3) into the inequality (D11), the proof is completed.

Remark 5. From the above theorems and corollaries, we can conclude that:

■ Under the four assumptions proposed by us, the random variable or multivariate random function does not concentrate around its mathematical expectation. Theorems 1 and 2 imply that the random variable or multivariate random function should concentrate around its conditional expectation. And to some extent, Corollaries 1 and 2 also imply such concentration in these two cases: 1) $1 - \mathrm{P}(A)$ is close to 0 as n tends to ∞, see Example 3, or 2) $A = \Omega$.

■ The original inequalities (B2) and (B3) can be viewed as special cases of the extensions (D8) and (D10): if A increases up to Ω, then the inequality (D8) reduces to the inequality (B2) in Section 3. If A increases up to Ω, then the inequality (D10) reduces to the inequality (B3) in Section 3. Here A equals to $\prod_{i=1}^{n} A_i$, and Ω equals to $\prod_{i=1}^{n} \Omega_i$. In addition, the bounds of the extensions (the equalities (D8) and (D10)) improve the bounds of the equation (A1) given by [4]: 1) the bound of the equation (A1) is trivial if the item p in the equation (A1) is larger than $1/2$, whereas $\mathrm{P}(A)$ in our extensions has no such limitations, 2) the factor of the item 'exp' in the equation (A1) is always 1, whereas our factor is a probability $\mathrm{P}(A)$.

■ The bounds of the extensions (the equalities (D5) and (D7)) are tighter than the equalities (B2) and (B3): let $a'_{i\mathbf{j}} = a_{ij_i} = \min_{j \in \{j_1, \ldots, j_n\}} a_{ij}$ and $b'_{i\mathbf{j}} = b_{ij_i} = \max_{j \in \{j_1, \ldots, j_n\}} b_{ij}, i \in \mathbb{N}_n$ where $a_{ij_i} \leqslant b_{ij_i}, i \in \mathbb{N}_n$. Then we have

$$\sum_{\mathbf{j} \in \psi_{n,k}} P(\widetilde{A}_j) \cdot \exp\left(-2t^2 / \sum_{i=1}^{n} (a_{i\mathbf{j}} - b_{i\mathbf{j}})^2\right) \leqslant \exp\left(-2t^2 \Big/ \sum_{i=1}^{n} (a'_{i\mathbf{j}} - b'_{i\mathbf{j}})^2\right) \tag{D12}$$

According to definition of conditional expectation, $\mathcal{B} = \{\emptyset, \Omega\}$ implies $\mathrm{E}(X|\mathcal{B}) = \mathrm{E}(X)$. By substituting the inequality (D12) into the inequality (D5), we thus get

$$\mathrm{P}(|S_n - \mathrm{E}(S_n)| \geqslant t) \leqslant 2 \cdot \exp\left(-2t^2 \Big/ (\sum_{i=1}^{n} (a'_{i\mathbf{j}} - b'_{i\mathbf{j}})^2)\right) \tag{D13}$$

The inequality (D13) means that, if we take $a'_{i\mathbf{j}} = a_{ij_i}$ and $b'_{i\mathbf{j}} = b_{ij_i}$, $i \in \mathbb{N}_n$ in Theorem 1, then the inequality (D5) will reduce to the inequality (B2) in Section 3. Thus, we conclude that if the bound of the random variables is refined on a sample space Ω, then a tighter bound by Theorem 1 will be obtained.

Similarly, if we take $c'_{m\mathbf{j}_m} = c_{mj_m} = \max_{j \in \{j_1, \ldots, j_n\}} b_{mj}$, $m \in \mathbb{N}_n$ in Theorem 2, then the inequality (D7) will reduce to the inequality (B3) in Section 3, leading to the conclusion that if the bounded differences of the function f is refined on a sample space Ω, then a tighter bound by Theorem 2 will be reached.

Remark 6. For unbounded random variables, there are also some Bernstein-like results [10,11]. These results all require that the moment (for example, variance) exists or is uniformly bounded, and this limits their extension to some applications, whereas our results have no restrictions on the moment based on Corollary 1 or 2.

Appendix E More Applications in Statistical Learning Theory

In the previous section, we have proposed more extensions and compared them with existing bounds as the supplement of the paper (Due to the page length, we only give three examples in the paper. Here, Example 5 is new).

Now we discuss these extensions to applications in learning theory through four examples. We show that our extensions are slightly faster than the existing results in some special cases and artificially bounding the unbounded loss function may not discover the overfit. For practical application, we introduce a theorem to guarantee generalization.

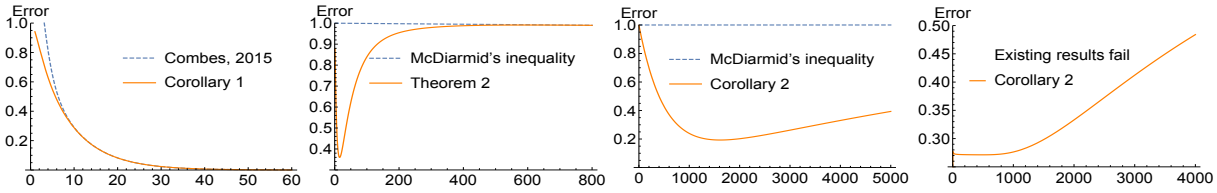

Figure E1 Here the horizontal axis means the sample complexity. From left to right, 1) The error probabilities from Corollary 1 and [20] (take $t = 0.5$). 2) The error probability from Theorem 2 is slightly faster than the error probability from McDiarmid's inequality (take $t = 25$). 3) The error probability from Corollary 2 is slightly faster than that from McDiarmid's inequality (take $t = 3$). 4) The relation between the sample complexity and the error probability from Corollary 2, and the existing results fail (take $\phi(n) = n^{1001/1000}$ and $t = 50$).

Example 3. [4] gave an example as follows:

Let $\Omega = \{0,1\}^n$, X_i follows a Bernoulli distribution $Bern(1,p)$, $i = 1,\ldots,n$, $A = \Omega\backslash\{(0,\ldots,0),(1,\ldots,1)\}$, there exists a constant $B \geqslant 0$. Set f a piecewise function: if $X_i = 0, i = 1,2,\ldots,n$, $f(X_1,X_2,\ldots,X_n) = B$; if $X_i = 1, i = 1,2,\ldots,n$, $f(X_1,X_2,\ldots,X_n) = -B$; otherwise, $f(X_1,X_2,\ldots,X_n) = (1/n)\sum_{i=1}^n 2(X_i - 1)$. Then, [4] obtained the generalization bound as follows

$$\mathrm{P}(f(X) \geqslant t) \leqslant 2^{-n} + \exp\left(-\frac{n}{2}((t - 2^{1-n})_+)^2\right)$$

From Corollary 1, we have the following generalization bound

$$\mathrm{P}(f(X) \geqslant t) \leqslant 2^{-n} + (1 - 2^{-n}) \cdot \exp\left(-\frac{n}{2}t^2\right)$$

From Figure E1.(1) it can be seen that the error probability from Corollary 1 is slightly faster than the error probability from [4].

Example 4. We assume that $\Omega = \{0,1,\ldots,97,1000,10000\}^n$, X_i follows a multinomial distribution $Mult(100,\mathbf{p})$, $\mathbf{p} = (1/100,1/100,\ldots,1/100)$, $i = 1,2,\ldots,n$. Let $f(X_1,X_2,\ldots,X_n) = \sum_{i=1}^n X_i$. By the assumptions, we have

$$\mathrm{P}(|f(X_1,X_2,\ldots,X_i,\ldots,X_n) - f(X_1,X_2,\ldots,X_i',\ldots,X_n)| \leqslant 97/n) = (1 - 2/100)^n,$$

$$\mathrm{P}(97/n \geqslant |f(X_1,X_2,\ldots,X_i,\ldots,X_n) - f(X_1,X_2,\ldots,X_i',\ldots,X_n)| \leqslant 1000/n) = (1 - 1/100)^n - (1 - 2/100)^n$$

and

$$\mathrm{P}(1000/n \geqslant |f(X_1,X_2,\ldots,X_i,\ldots,X_n) - f(X_1,X_2,\ldots,X_i',\ldots,X_n)| \leqslant 10000/n) = 1 - (1 - 1/100)^n.$$

Then, from the equation $(B3)$, we have

$$\mathrm{P}(f(X_1,X_2,\ldots,X_n) - \mathrm{E}(f(X_1,X_2,\ldots,X_n)|\mathcal{B}) \geqslant t) \leqslant \exp\left(-2t^2 \Big/ \sum_{i=1}^n (10000/n)^2\right)$$

From Theorem 2 we have

$$\begin{aligned}
&\mathrm{P}(f(X_1,X_2,\ldots,X_n) - \mathrm{E}(f(X_1,\ldots,X_n)|\mathcal{B}) \geqslant t)\\
\leqslant\ &(1 - 2/100)^n \cdot \exp\left(-2t^2 \Big/ \sum_{i=1}^n (97/n)^2\right)\\
&+((1 - 1/100)^n - (1 - 2/100)^n) \cdot \exp\left(-2t^2 \Big/ \sum_{i=1}^n (1000/n)^2\right)\\
&+(1 - (1 - 1/100)^n) \cdot \exp\left(-2t^2 \Big/ \sum_{i=1}^n (10000/n)^2\right)
\end{aligned}$$

where the values of $(1-2/100)^n$, $((1-1/100)^n - (1-2/100)^n)$ and $(1-(1-1/100)^n)$ can be seen as the weights obtained according to the proportion of samples.

From Figure E1.(2) it can be seen that the error probability from Theorem 2 is firstly faster than that from McDiarmid's inequality, and then the distinction of their error probabilities gradually becomes less visible.

Example 5. We assume that $\Omega = \{0,1,\ldots,98,\infty\}^n$, X_i follows a multinomial distribution $Mult(100,\mathbf{p})$, $\mathbf{p} = (101/10000, 101/10000,\ldots,101/10000,1/10000)$, $i = 1,2,\ldots,n$. Let $f(X_1,\ldots,X_n) = \sum_{i=1}^n X_i$. By the assumptions, we have $\mathrm{P}(|f(X_1,\ldots,X_i,\ldots,X_n) - f(X_1,\ldots,X_i',\ldots,X_n)| \leqslant 98/n) = (1 - 1/10000)^n$. Then, from Corollary 2 we have

$$\mathrm{P}(f(X_1,\ldots,X_n) - \mathrm{E}(f(X_1,\ldots,X_n)|\mathcal{B}) \geqslant t) \leqslant (1 - 1/10000)^n \cdot \exp\left(-2t^2 \Big/ \sum_{i=1}^n (98/n)^2\right) + (1 - 1/10000)^n$$

From Figure E1.(3) it can be seen that the error probability decreases as the sample complexity increases when the sample complexity is smaller than about 1500. When the sample complexity is larger than about 1500, meanwhile, the error probability increases with the increase of the sample complexity. However, the result from McDiarmid's inequality is trivial. Examples 1 and 2 in Section 4 can be analyzed similarly.

For simplicity, in the discussions of Examples 3-5 we do not involve loss functions. In the last example, we will introduce a loss function and show how to apply our extensions.

Example 6. We assume that the linear model for regression is $y = h(x) + \epsilon$, where ϵ is a standard Cauchy random variable with the density function $(1/\pi) \cdot 1/(1+x^2)$. The loss function L is defined by the absolute loss $|h(x) - y|, h \in \mathcal{H}$ (denoted by $Q(h, z)$).

It is obvious that the expected value of ϵ does not exist. Therefore, Hoeffding's inequality and McDiarmid's inequality do not hold because Hoeffding's inequality and McDiarmid's inequality are distribution independent. Here, our results will be valid. We can employ Corollary 2 to analyze its generalization bound.

Let the set A_i in Corollary 2 be $-\phi(n) \leqslant \epsilon_i \leqslant \phi(n), i = 1, 2, \ldots, n$. Then, we have

$$\begin{aligned} &\mathrm{P}\left(\frac{1}{n} \cdot \sum_{i=1}^{n} Q(h, z_i) - \mathrm{E}(Q(h, z)|A) \geqslant t\right) \\ \leqslant\ & (1/2 + \arctan(\phi(n))/\pi)^n \cdot \exp\left(-2 \cdot n \cdot t^2/\phi^2(n)\right) + 1 - \left(\frac{1}{2} + \arctan(\phi(n))/\pi\right)^n \end{aligned}$$

From Figure E1.(4) it can be seen that the error probability is smaller than 0.5 when the sample complexity is smaller than about 4500. When the sample complexity is smaller than about 4500, the error probability becomes larger than 0.5, and tends to 1 with the increase of the sample complexity.

Remark 7. Examples 3 and 4 indicate that our extensions are slightly faster than the existing results when the sample complexity is not so large. Examples 5 and 6 show that our extensions describe that the error probability evolves over the sample complexity while the existing results are trivial or failure. Furthermore, from Example 4, we find that the effect of weighting on the proportion of samples is effective when the sample complexity is small, and becomes negligible when the complexity of the sample increases. From examples 5 and 6, we can tell that the generalization analysis on the artificial bound of the unbounded loss function (or the loss function whose expectation does not exist) may not catch the overfit, and for the generalization analysis of these loss functions. However, the classical learning framework may not be enough for the generalization analysis of these loss function. Such a phenomenon has also been observed by [12].

The discussion of generalization bounds in learning theory is mostly under the boundedness conditions of loss functions, since there are lots of handy tools such as Hoeffding's and McDiarmid's inequalities that are for bounded functions [13] are available. However, it is no such a limitation of boundedness conditions of loss functions in practical environment (see Examples 5-6). Suppose that the loss function satisfies the p_i bounded or the p_i difference-bounded condition, we introduce the following theorem to guarantee generalization.

Theorem 3. Let G be a family of functions. For each $g \in G$, assume that g is $\mathrm{P}(A)$ difference-bounded by 1. Then with probability at least $1 - \delta$ $(1 > \delta > 0)$, the following inequality holds for all $g \in G$,

$$\mathrm{E}(g) - \mathrm{E}_n(g) \leqslant 2 \cdot \mathcal{R}_n(G) \cdot I_A + \sqrt{\frac{2 \cdot \ln\left(\mathrm{P}(A)/(\delta - (1 - \mathrm{P}(A)))\right)}{n}} \tag{E1}$$

Proof. Take $Z = \{z_1, \ldots, z_i, \ldots, z_n\}, Z' = \{z_1, \ldots, z_i', \ldots, z_n\}$, $z_1, \ldots, z_i, \ldots, z_n, z_i' \in S, i \in \mathbb{N}_n$. For any $h \in \mathcal{H}$, we have $\mathrm{E}_n^Z[h \cdot I_A(Z)] = \mathrm{E}_n^Z(g \cdot I_A(Z)) = (1/n) \cdot \sum_{i=1}^n g(h, z_i) \cdot I_A(Z)$. By the assumptions, it is easy to see that

$$\begin{aligned} &\sup_{g \in G}((\mathrm{E}(g) - \mathrm{E}_n^Z(g)) \cdot I_A(Z)) - \sup_{g \in G}((\mathrm{E}(g) - \mathrm{E}_n^{Z'}(g)) \cdot I_A(Z')) \\ \leqslant\ & \sup_{g \in G}((\mathrm{E}_n^{Z'}(g) - \mathrm{E}_n^Z(g)) \cdot I_A(Z)) \\ \leqslant\ & \sup_{g \in G} \frac{g(h, z_i) - g(h, z_i')}{n} \cdot I_A(Z) \\ \leqslant\ & 2/n \end{aligned}$$

that is, $\mathrm{E}(g) - \mathrm{E}_n(g)$ is $\mathrm{P}(A)$ difference-bounded by $2/n$.

By Corollary 2, we set $\delta = 2 \cdot \mathrm{P}(A) \cdot \exp\left(-2t^2/\left(\sum_{i=1}^n c_i^2\right)\right) + 1 - \mathrm{P}(A)$. Then, it is straightforward to show that the inequality (E1) holds.

References

1 S. Kutin and P. Niyogi. The interaction of stability and weakness in Adaboost. Technical Report TR-2001-30, Computer Science Department, University of Chicago, 2001.

2 S. Kutin. Extensions to McDiarmid's inequality when differences are bounded with high probability. Technical Report TR-2002-04, Department Computer Science, University of Chicago, 2002.

3 A. Kontorovich. Concentration in unbounded metric spaces and algorithmic stability. ICML, 2014.

4 R. Combes. *An extension of McDiarmid's inequality.* http://arxiv.org/pdf/1511.05240v1.pdf, 2015.

5 V. Bentkus. An extension of the Hoeffding inequality to unbounded random variables. *Lithuanian Mathematical Journal*, 48(48):137–157, 2008.

6 S. Kutin and P. Niyogi. *Almost-everywhere algorithmic stability and generalization error.* http://arxiv.org/pdf/1301.0579v1.pdf, 2002.

7 P. D. Grünwald and N. A. Mehta. *Fast rates with unbounded losses.* http://arxiv.org/pdf/1605.00252v1.pdf, 2016.

8 W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

9 C. McDiarmid. On the method of bounded differences. *Surveys in Combinatorics*, 141:148–188, 1989.
10 V.D. Geer. Empirical processes in m-estimation. *Journal of Nonparametric Statistics*, 10(1):454–455, 2000.
11 E. Giné and R. Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge University Press, Cambridge, 2015.
12 S. Mendelson. Learning without concentration. *Journal of the ACM*, 62(3):21:1–25, 2014.
13 C. Cortes, S. Greenberg, and M. Mohri. Relative deviation learning bounds and generalization with unbounded loss functions. *Comunicazioni Sociali*, 23:234–255, 2013.