SCIENCE CHINA
Information Sciences

• HIGHLIGHT •

April 2018, Vol. 61 048105:1–048105:3
https://doi.org/10.1007/s11432-017-9225-2

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*

Received 9 April 2017/Revised 16 August 2017/Accepted 21 August 2017/Published online 16 March 2018

Citation Wu X X, Zhang J P. Distribution-dependent concentration inequalities for tighter generalization bounds. Sci China Inf Sci, 2018, 61(4): 048105, https://doi.org/10.1007/s11432-017-9225-2

Concentration inequalities play a crucial role in statistical learning theory because they are useful for deriving the generalization capacity of learning models. Generally, they can be used to estimate the deviations between empirical risk and expectation risk [1]. Some important learning theories such as Rademacher complexity have been developed by applying concentration inequalities to bound such deviations [2–4].

Two commonly used concentration inequalities in learning theory are Hoeffding's and McDiarmid's inequalities. These two inequalities however have two major limitations: (1) they cannot deal with unbounded functions; (2) their bounds are weak for functions with a larger constant on a small exceptional set. If we generalize these inequalities to any distribution, the estimation of the deviations is likely to be loose. In the latter case, the bounds given by them will become less tight because the large constant dominates this bound. To address these issues, Refs. [5, 6] proved two extensions of McDiarmid's inequality for strongly and weakly difference-bounded functions and used them to study the generalization capacity. Ref. [7] proved an extension of McDiarmid's inequality with the subgaussian diameter. Recently, Ref. [8] proposed an extension of McDiarmid's inequality for functions with bounded differences on a high probability set and no restriction outside this set. Ref. [9] extended McDiarmid's inequality by relaxing the Lipschitz condition since the approach only needs Lipschitz-bounds for changing one variable.

However, both the strong and weak bounded difference conditions proposed by [5, 6] have their shortcomings in practice. The approach proposed by [7] requires an extra metric on the sample space and a bounded subgaussian diameter. The weaker Lipschitz condition given by [9] is only useful for bounded functions. Meanwhile, the bound discussed by [8] can be further tightened. After exploring the assumptions of Hoeffding's and McDiarmid's inequalities, we propose some extensions to these two inequalities to treat the cases of probabilistic boundedness and bounded differences. Our results improve the bound in [8] and the bounds in the original inequalities, and can also handle unbounded functions without introducing extra metrics.

Motivation. The unbounded functions often occur in the regression and classification. Since Hoeffding's and McDiarmid's inequalities provide bounds independent of the distribution, we expect that our proposed distribution-dependent bounds will be tighter for each specific case.

Notations. Let $\mathbb{N}_n = \{1, 2, \ldots, n\}$, $n \in \mathbb{N}$. Set 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, $\mathcal{H}$ is called the hypothesis class, σ_i, $i = 1, 2, \ldots, n$ are Rademacher variables. Denote $\mathcal{B}$ as the σ-algebra $\sigma(\Psi_{n,k})$, and $\mathcal{B} \subset \prod_{i=1}^{n}\mathcal{A}_i$. The set $\emptyset$ denotes the

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

empty set.

Contribution. We prove some distribution-dependent extensions of Hoeffding's and McDiarmid's inequalities and obtain tighter bounds. Firstly, we introduce four assumptions.

Assumption 1 (p_i bounded)**.** Let X_i be the independent random variable on a probability $(\Omega_i, \mathcal{A}_i, P_i),\ i \in \mathbb{N}_n$, s.t. $P_i(a_i \leqslant X_i(\omega) \leqslant b_i) = p_i,\ i \in \mathbb{N}_n$. 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 $(\Omega_i, \mathcal{A}_i, P_i),\ i \in \mathbb{N}_n$, s.t. there exists an integer $k > 1$, we have $P_i(a_{ij} \leqslant X_i(\omega) \leqslant b_{ij}) = p_{ij},\ j \in \mathbb{N}_k,\ i \in \mathbb{N}_n$ and $P_i(\bigcup_{j=1}^{k}(a_{ij} \leqslant X_i(\omega) \leqslant b_{ij})) = 1,\ i \in \mathbb{N}_n$. 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)**.** 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 ℓ-th coordinate, we have $|g(\omega) - g(\omega')| \leqslant c_\ell$ and $P(A) = p$. 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 $A_i \cap A_j = \emptyset,\ i \neq j,\ i, j \in \mathbb{N}_n$, for any $\omega \in A_j$ and $\omega' \in A_j$ differ only in the ℓ-th coordinate, we have $|g(\omega) - g(\omega')| \leqslant c_{\ell j}$ and $P(A_j) = p_j$. If this is true, then g is (p_j, k) hierarchy-difference-bounded by $\{c_{\ell j}, \ell \in \mathbb{N}_n\},\ j \in \mathbb{N}_k$.

We assume that X_i is the independent random variable on a probability space $(\Omega_i, \mathcal{A}_i, P_i),\ i \in \mathbb{N}_n$, 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 $\prod_{i=1}^{n}\Omega_i$.

Lemma 1. Assume that X_i is $(P_i(A_{ij}), k)$ 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 $\widetilde{A}_{\boldsymbol{j}} = \prod_{i=1}^{n}A_{ij} \in \Psi_{n,k},\ j \in \mathbb{N}_k,\ \boldsymbol{j} \in \psi_{n,k}$. Then, for any $t > 0$, we have

$$P\left(\left(\left|S_n - \sum_{\boldsymbol{j}\in\psi_{n,k}}\mathrm{E}(S_n|\widetilde{A}_{\boldsymbol{j}})I_{\widetilde{A}_{\boldsymbol{j}}}\right| \geqslant t\right) \cap \widetilde{A}_{\boldsymbol{j}_0}\right) \leqslant 2 \cdot P(\widetilde{A}_{\boldsymbol{j}_0}) \cdot \exp\left(-2t^2 \Big/ \sum_{i=1}^{n}(a_{i\boldsymbol{j}_0} - b_{i\boldsymbol{j}_0})^2\right),$$

where $\boldsymbol{j}_0 = (j_{10}, j_{20}, \ldots, j_{n0}) \in \psi_{n,k}$ is a constant vector, $a_{i\boldsymbol{j}_0} = a_{ij_{i0}}$ and $b_{i\boldsymbol{j}_0} = b_{ij_{i0}},\ i \in \mathbb{N}_n$.

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 $P(A)$ 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

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

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

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

where $a_{i\boldsymbol{j}} = a_{ij_i}$ and $b_{i\boldsymbol{j}} = b_{ij_i},\ i \in \mathbb{N}_n$.

Theorem 2. Let the function f be a map from $\mathcal{X}^n$ to $\mathbb{R}$. Assume that f is $(P_i(A_{ij}), k)$ 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 $\widetilde{A}_{\boldsymbol{j}} = \prod_{i=1}^{n}A_{ij} \in \Psi_{n,k},\ j \in \mathbb{N}_k,\ \boldsymbol{j} \in \psi_{n,k}$. For any $t > 0$, we have

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

where $c_{i\boldsymbol{j}} = c_{mj_m},\ m \in \mathbb{N}_n$.

From the above Theorems, it follows:

Corollary 1. Assume that X_i is $P_i(A_i)$ 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

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

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

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

Remark 1. In Theorems 1, 2 and Corollaries 1, 2, we obtain four inequalities for the cases of probabilistic boundedness and bounded differences.

Then, we see that these above extensions are distribution dependent, and consequently yield better estimation for some examples (see Example 1), and can describe the evolution of the error probability with the sample complexity (that is, the number of samples) while the existing results are trivial or failure (see Examples 2 and 3).

Example 1. Let $\Omega = \{0,1\}^n$, X_i follows a Bernoulli distribution $\mathrm{Bern}(1,p)$, $i = 1,\ldots,n$, $A = \Omega\backslash\{(0,\ldots,0),(1,\ldots,1)\}$, 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, Ref. [8] 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:

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

Example 2. We assume that $\Omega = \{0,1,\ldots,98,\infty\}^n$, X_i follows a multinomial distribution $\mathrm{Mult}(100,\boldsymbol{p})$, $\boldsymbol{p} = (101/10000,101/10000,\ldots,101/10000,1/10000)$, $i = 1,2,\ldots,n$. 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

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

Example 3. 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

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

Finally, we introduce Theorem 3 to guarantee the generalization in practical application.

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

$$\begin{aligned}&\mathrm{E}(g) - \mathrm{E}_n(g)\\ &\quad\leqslant\sqrt{\frac{2\cdot\ln\left(P(A)/(\delta-(1-P(A)))\right)}{n}}\\ &\qquad+2\cdot\mathcal{R}_n(G)\cdot I_A.\end{aligned}$$

Conclusion. In this article, we review the conditions and limitations of Hoeffding's and McDiarmid's inequalities and propose four new assumptions. Based on our proposed assumptions, we obtain several extensions of Hoeffding's and McDiarmid's inequalities. Through three examples, we also discuss the potential applications of our extensional results in learning theory. For practical application, we introduce a theorem to guarantee the generalization.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant No. 61673118), Shanghai Pujiang Program (Grant No. 16PJD009), and Shanghai Talents Development Funds (Grant No. 201629).

Supporting information Appendixes A–E. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Vapnik V. The Nature of Statistical Learning Theory. New York: Springer-Verlag, 1995
2 Koltchinskii V. Rademacher penalties and structural risk minimization. IEEE Trans Inf Theory, 2001, 47: 1902–1914
3 Bartlett P L, Mendelson S. Rademacher and Gaussian complexities: risk bounds and structural results. J Mach Learn Res, 2003, 3: 463–482
4 Gao W, Zhou Z H. Dropout rademacher complexity of deep neural networks. Sci China Inf Sci, 2016, 59: 072104
5 Kutin S. Extensions to McDiarmid's Inequality When Differences are Bounded With High Probability. Technical Report TR-2002-04. 2002
6 Kutin S, Niyogi P. Almost-everywhere algorithmic stability and generalization error. 2002. http://arxiv.org/pdf/1301.0579v1.pdf
7 Kontorovich A. Concentration in unbounded metric spaces and algorithmic stability. ICML, 2014. arXiv:1309.1007
8 Combes R. An extension of McDiarmid's inequality. 2015. http://arxiv.org/pdf/1511.05240v1.pdf
9 Warnke L. On the method of typical bounded differences. Comb Probab Comput, 2015, 25: 269–299