

for General Functions of Independent Random Variables with Heavy Tails

#### **Abstract**

- ► The concentration of measure -> an essential role in high dimensional statistics and machine learning
- In this paper:
  - ► Unbounded analogues of bounded difference inequality for functions of independent random variables with heavy-tailed distributions
  - Provide a general framework
    - ▶ for all heavy-tailed distributions with finite variance
    - ▶ moment inequality for sums of independent random variables → for general functions
  - ► Applications:
    - ▶ bounded, Bernstein's moment condition, weak-exponential, and polynomial-moment random variables
    - ▶ vector valued concentration, Rademacher complexity, and algorithmic stability

#### Introduction

- Concentration and moment inequalities -> essential toolkit for natural and artificial learning systems
- ► The bounded difference inequality compared to the general Hoeffding and Bernstein-type inequality
  - not only for sums
  - but also, for general functions of independent random variables

# Bounded Difference Inequality (McDiarmid's Inequality)

- $X = (X_1, ..., X_n)$  the sequence of independent random variables
- $|f(x)-f(x')| \le c_k$  whenever x and x' differ only in the k-th coordinate

$$\mathbb{P}(f(X) - \mathbb{E}[f(X)] > t) \le \exp\left(\frac{-2t^2}{\sum_k c_k^2}\right) \quad \forall t > 0$$



## Limitations of McDiarmid's Inequality

- Requires the conditional ranges to be uniformly bounded
- > => Cause limitations on applicability to unbounded loss functions
- But concentration properties of unbounded functions are important:
  - ▶ Generalization bounds in unbounded settings
  - ► PAC-Bayes learning

#### Previous Works

- Combes (2015):
  - Extension for functions with bounded differences on a high probability set and no restriction outside this set
  - ▶ Entail complex statement, the conditions are too restrictive in practice
- Warnke (2016):
  - ightharpoonup relaxing the worst-case changes  $c_k$  to typical changes (used in combinatorial applications)
  - ▶ Results realm on bounded random variables
- Maurer and Pontil (2021):
  - ▶ Study the heavier sub-exponential distributions than sub-gaussian cases
  - ▶ Still relatively light-tailed

## Heavy-tailed vs. Sub-exponential

- ▶ A distinctive difference is the moment generating function (MGF)
- ► Sub-exponential distributions:
  - ▶ MGF exists in a neighborhood around zero
- ► Heavy-tailed distributions:
  - ► MGF does not exist
- The technique to find upper bounds for the MGF, fails for heavy-tailed distributions

## Way to Moment Inequalities

- ▶ In random combinatorics, statistics and empirical process theory:
  - ▶ Search for moment inequalities dealing with heavy-tailed random variables
- ▶ Lack of relevant results in bounded difference-type moment inequalities
- ▶ On the other hand, there is a growing demand
- ▶ Instead of focus on concentration inequalities -> moment inequalities

#### **Notations**

- $\blacktriangleright \ X = (X_1, \dots, X_n)$  a vector of independent random variables with values in a space  $\mathcal X$
- lacksquare  $X' = (X'_1, \dots, X'_n)$  independent and identically distributed to X
- $ightharpoonup f\colon \mathcal{X}^n \mapsto \mathbb{R} \text{ be a function}$
- $L_p \text{ norm: } \|Z\|_p = (\mathbb{E}[|Z|^p])^{1/p}$



## k-th Centered Conditional version of f

$$f_k(X)(x) = f(x_1, \dots, x_{k-1}, X_k, x_{k+1}, \dots, x_n) - \mathbb{E}[f(x_1, \dots, x_{k-1}, X'_k, x_{k+1}, \dots, x_n)]$$

- $f_k(X): x \in \mathcal{X}^n \to f_k(X)(x)$ : random-variable-valued-function
- **>** =>

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

#### Main Theorem!

- $ightharpoonup X_1, ..., X_n$  independent random variables with values in a space  $\mathcal{X}$
- $ightharpoonup f\colon \mathcal{X}^n \mapsto \mathbb{R}$  be a measurable function
- $S = f(X_1, ..., X_{i-1}, X_i, X_{i+1}, ..., X_n)$  and  $S_i = f(X_1, ..., X_{i-1}, X_i', X_{i+1}, ..., X_n)$
- ▶ Assume,  $|S S_i| \le F_i(x_1, ..., x_{i-1}, X_i, X_i', x_{i+1}, ..., x_n)$  for  $F_i: \mathcal{X}^{n+1} \mapsto \mathbb{R}$
- ▶ Then, for all  $p \ge 1$ :

$$||f(X_1,...,X_n) - \mathbb{E}f(X_1,...,X_n)||_p \le 2 \left\| \sum_{i=1}^n \sup_{x \in \mathcal{X}^n} f_i(X)(x) \right\|_p$$



#### Main Weakness

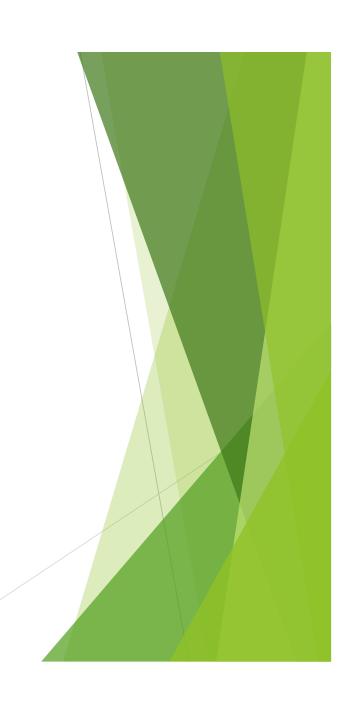
- Major weakness: Supremum in the sum
- Reasons:
  - 1. Supremum means that some sort of boundedness remains necessary
  - 2. If this sum of random variables satisfies a central limit theorem,  $f(X_1, ..., X_n)$  may not
  - 3. Most of applications require uniform boundedness of the conditional variances
- ► These happens due to appearance of a corresponding variance proxy in the inequality!

#### Another Version of Main Theorem!

lacksquare If we assume  $|S-S_i| \leq F_i(X_i,X_i')$  , Then,

$$\|f(X_1,...,X_n) - \mathbb{E}f(X_1,...,X_n)\|_p \le \left\|\sum_{i=1}^n \epsilon_i F_i(X_i,X_i')
ight\|_p$$

lacktriangle This assumption is like a Lipschitz condition if  $F_i$  is a distance function



#### How to use the theorem?

- lacksquare Given the p-th moment-bound of the sum variables  $\|\sum_i \sup_{x\in\mathcal{X}^n} f_i(X)(x)\|_p$
- lacksquare With the theorem we have a moment inequality  $\|f-\mathbb{E}f\|_p$
- lacksquare Using Markov's inequality to transfer it to a tail inequality  $\mathbb{P}(|f-\mathbb{E}f|\geq t)$

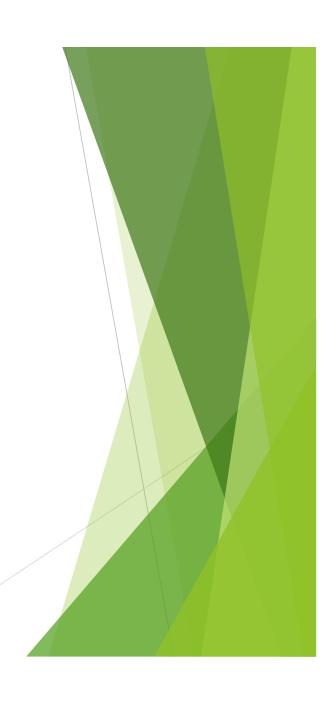
#### Random Bounded Variables

- For all i and  $x \in \mathcal{X}^n$ ,
  - $|f_i(X)(x)| \le b$
  - $\blacktriangleright \ \mathbb{E}[\left(f_i(X)(x)\right)^2] \le \sigma_i^2$
- For all  $p \ge 1$ :

$$\|f(X_1,...,X_n) - \mathbb{E}f(X_1,...,X_n)\|_p \leq 2\left(\sqrt{6p}\left(\sum_{i=1}^n \sigma_i^2
ight)^{1/2} + 10pb
ight)$$

By Markov's inequality:

$$\mathbb{P}(|f(X_1, ..., X_n) - \mathbb{E}f(X_1, ..., X_n)| \ge t) \le \exp\left(-\min\left\{\frac{t^2}{96e^2 \sum_{i=1}^n \sigma_i^2}, \frac{t}{40eb}\right\}\right)$$



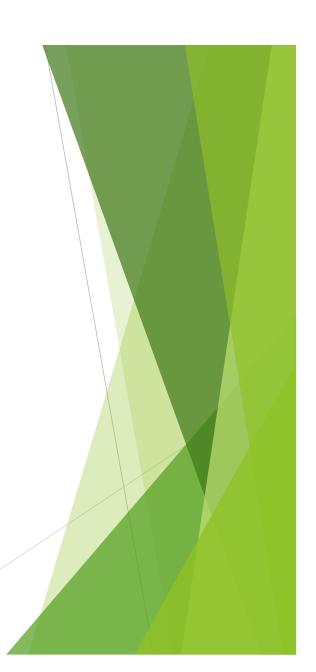
#### Bernstein's Moment Condition

- For all i and  $x \in \mathcal{X}^n$ ,
  - $ightharpoonup f_i(X)(x)$  satisfies Bernstein's moment condition
  - $\blacktriangleright \ \mathbb{E}[\left(f_i(X)(x)\right)^2] \le \sigma_i^2$
- For all  $p \ge 2$ :

$$\|f(X_1,...,X_n) - \mathbb{E}f(X_1,...,X_n)\|_p \le 2\left(4\sqrt{p}\left(\sum_{i=1}^n \sigma_i^2\right)^{1/2} + 8pb\right)$$

By Markov's inequality:

$$\mathbb{P}(|f(X_1, ..., X_n) - \mathbb{E}f(X_1, ..., X_n)| \ge t) \le \exp\left(-\min\left\{\frac{t^2}{256e^2\sum_{i=1}^n \sigma_i^2}, \frac{t}{32eb}\right\}\right)$$



## Weak-Exponential Random Variables

- $\qquad \qquad \psi_{\alpha} \colon \mathbb{R}_{+} \to \mathbb{R}_{+} \implies \psi_{\alpha}(x) = \exp(x^{\alpha}) 1$
- ▶ Orlicz norm:  $||X||_{\psi_{\alpha}} = \inf\{\lambda > 0: \mathbb{E}\psi_{\alpha}\left(\frac{|X|}{\lambda}\right) \le 1\}$
- ▶ Random variable X is weak-exponential if for  $t \ge 0$ :

$$\mathbb{P}(|X| \ge t) \le 2 \exp\left(-\left(\frac{t}{\|X\|_{\psi_{lpha}}}\right)^{lpha}\right)$$

- ► Also, known as sub-Weibull variables
- $\alpha = 2 \Rightarrow \text{sub-Gaussian}$
- $\alpha = 1 \Rightarrow \text{sub-Exponential}$
- $\alpha = \infty$  => bounded variables

## Weak-Exponential variables $0 < \alpha \le 1$

- For all i and  $x \in \mathcal{X}^n$ ,
  - $||f_i(X)(x)||_{\psi_\alpha} \le b$
  - $\triangleright \mathbb{E}[\left(f_i(X)(x)\right)^2] \le \sigma_i^2$
- If  $0 < \alpha \le 1$ :

$$||f(X_1,...,X_n) - \mathbb{E}f(X_1,...,X_n)||_p \le 2\left(\sqrt{6p}\left(\sum_{i=1}^n \sigma_i^2\right)^{1/2} + C_{\alpha}K_{\alpha}b(\log(n+1))^{1/\alpha}p^{1/\alpha}\right)$$

▶ By Markov's inequality:

$$\mathbb{P}\left(|f(X) - \mathbb{E}f(X')| \ge t\right) \le \exp\left(-\min\left\{\frac{t^2}{96e^2\sum_{i=1}^n \sigma_i^2}, \frac{t^\alpha}{(4eC_\alpha K_\alpha)^\alpha \log(n+1)b^\alpha}\right\}\right)$$

### Weak-Exponential variables $\alpha > 1$

- For all i and  $x \in \mathcal{X}^n$ ,
  - $||f_i(X)(x)||_{\psi_\alpha} \le b$
  - $\triangleright \mathbb{E}[\left(f_i(X)(x)\right)^2] \le \sigma_i^2$
- If  $\alpha > 1$ :

$$\|f(X_1,...,X_n) - \mathbb{E}f(X_1,...,X_n)\|_p \le 2\left(\sqrt{6p}\left(\sum_{i=1}^n \sigma_i^2\right)^{1/2} + C_{\alpha}b(\log(n+1))^{1/\alpha}p\right)$$

By Markov's inequality:

$$\mathbb{P}\left(|f(X) - \mathbb{E}f(X')| \ge t\right) \le \exp\left(-\min\left\{\frac{t^2}{96e^2\sum_{i=1}^n \sigma_i^2}, \frac{t}{4eC_\alpha(\log(n+1))^{1/\alpha}b}\right\}\right)$$

## Polynomial-Moment Random Variables

- For all i and  $x \in \mathcal{X}^n$ ,
  - $ightharpoonup \mathbb{E} |f_i(X)(x)|^p \le b_i, \quad p \ge 2$
  - $\mathbb{E}[\left(f_i(X)(x)\right)^2] \le \sigma_i^2$
- For all  $p \ge 2$ :

$$\|f(X_1,...,X_n) - \mathbb{E}f(X_1,...,X_n)\|_p \leq 2\sqrt{2\kappa(2+ heta)p}\left(\sum_{i=1}^n \sigma^2
ight)^{rac{1}{2}} + 2p\kappa\sqrt{1+rac{1}{ heta}}\left(\sum_{i=1}^n b_i
ight)^{rac{1}{p}}$$

▶ By Markov's inequality:

$$\mathbb{P}\left(|f(X) - \mathbb{E}f(X')| \ge t\right) \le \exp\left(-\frac{t^2}{16e^2(2\kappa(2+\theta))\sum_{i=1}^n \sigma_i^2}\right) + \frac{(4p\kappa\sqrt{1+\frac{1}{\theta}})^p\sum_{i=1}^n b_i}{t^p}$$

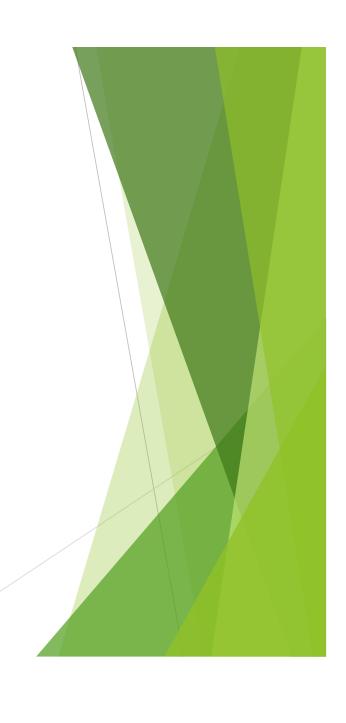
#### Without Bounded Variance

- For all i and  $x \in \mathcal{X}^n$ ,
- For all  $p \ge 2$ :

$$||f(X_1,...,X_n) - \mathbb{E}f(X_1,...,X_n)||_p \le 6\sqrt{2np}b^{1/p}$$

▶ By Markov's inequality:

$$\mathbb{P}(|f(X_1,...,X_n) - \mathbb{E}f(X_1,...,X_n)| \ge t) \le \frac{(6\sqrt{2np}b^{1/p})^p}{t^p}$$



## Rademacher Complexity and Generalization

Rademacher complexity:

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

Uniformly for functions in G:

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}rac{1}{n}\sum_{i}g(X_i)-\mathbb{E}[g(X_i')]
ight]\leq 2\mathcal{R}(\mathcal{G})$$

ightharpoonup By using bounded difference inequality and  $g\colon \mathcal{X} \mapsto [0,1]$ , then

$$\mathbb{P}\left(\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i}g(X_i)-\mathbb{E}[g(X_i')]>2\mathcal{R}(\mathcal{G})+t\right)\leq \exp\left(-2nt^2\right)$$



## In Weak-Exponential Setting

- ▶ The boundedness can be relaxed by heavy-tailed distributions
  - ► for uniformly Lipschitz function classes
- ${\bf N}=(X_1,\dots,X_n)$  vector of i.i.d weak-exponential random variables with values in a space  ${\mathcal X}$
- ▶ G is a class of functions  $g: \mathcal{X} \mapsto \mathbb{R}$  such that  $g(x) g(y) \le L||x y||$

If  $0 < \alpha \le 1$ :

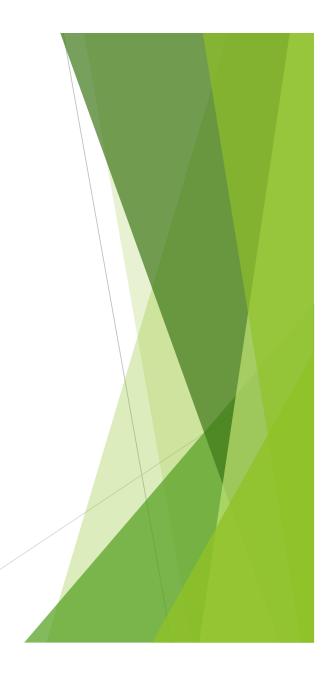
$$\mathbb{P}\left(\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i} g(X_{i}) - \mathbb{E}[g(X_{i}')] \ge t + 2\mathcal{R}(\mathcal{G})\right) \\
\le \exp\left(-\min\left\{\frac{t^{2}}{96e^{2} \frac{16L^{2}}{n} |||X_{1}|||_{2}^{2}}, \frac{t^{\alpha}}{(4eC_{\alpha}K_{\alpha})^{\alpha} \log(n+1)(\frac{4L}{n} |||X_{1}|||_{\psi_{\alpha}})^{\alpha}}\right\}\right)$$

• If  $\alpha \geq 1$ :

$$\mathbb{P}\left(\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i} g(X_{i}) - \mathbb{E}[g(X_{i}')] \ge t + 2\mathcal{R}(\mathcal{G})\right) \\
\le \exp\left(-\min\left\{\frac{t^{2}}{96e^{2} \frac{16L^{2}}{n} |||X_{1}|||_{2}^{2}}, \frac{t}{4eC_{\alpha}(\log(n+1))^{1/\alpha} \frac{4L}{n} |||X_{1}|||_{\psi_{\alpha}}}\right\}\right)$$

#### Conclusion

- Present bounded-difference type concentration and moment inequalities
  - ▶ for general functions of heavy-tailed independent random variables
- Provided a probabilistic toolbox that is general and flexible
  - ▶ to derive bounded difference-type concentration
  - ▶ and moment inequalities for heavy-tailed distributions
- Apply this method to
  - bounded,
  - Bernstein's moment condition,
  - weak-exponential, and
  - polynomial-moment variables
  - ► Rademacher complexity



Thanks for your attention!

