

Lecture 11: Concentration of sums of independent random variables



School of Mathematical Sciences, Xiamen University

1. Gaussian, or normal, distribution $\mathcal{N}(\mu, \sigma^2)$

- Standard normal random variable $X \sim \mathcal{N}(0, 1)$:

(i) Tail:

$$\mathbb{P}\{|X| \geq t\} \leq 2 \exp(-t^2/2), \quad \forall t \geq 0.$$

(ii) Moment:

$$\|X\|_p := (\mathbb{E}|X|^p)^{1/p} = \mathcal{O}(\sqrt{p}) \quad \text{as } p \rightarrow \infty.$$

(iii) Moment Generation function (MGF):

$$\mathbb{E} \exp(\lambda X) = \exp(\lambda^2/2) \quad \text{for all } \lambda \in \mathbb{R}.$$

(iv) MGF of square:

$$\mathbb{E} \exp(cX^2) \leq 2 \quad \text{for some } c > 0.$$

- The sum of independent normal random variables is also normal.

2. Sub-gaussian distributions

Theorem 1 (Sub-gaussian properties)

For a random variable X , the following properties are equivalent.

- *Tail: $\mathbb{P}\{|X| \geq t\} \leq 2 \exp(-t^2/K_1^2)$ for all $t \geq 0$.*
- *Moment: $\|X\|_p := (\mathbb{E}|X|^p)^{1/p} \leq K_2\sqrt{p}$ for all $p \geq 1$.*
- *MGF of square: $\mathbb{E} \exp(X^2/K_3^2) \leq 2$.*

Moreover, if $\mathbb{E}X = 0$ then these properties are also equivalent to the following one:

- *MGF: $\mathbb{E} \exp(\lambda X) \leq \exp(\lambda^2 K_4^2)$ for all $\lambda \in \mathbb{R}$.*

Remark. The parameters $K_i > 0$ appearing in these properties can be different. However, they may differ from each other by at most an absolute constant factor. This means that there exists an **absolute constant C** such that property 1 implies property 2 with parameter $K_2 \leq CK_1$, and similarly for every other pair of properties.

Definition 2

Random variables that satisfy one of the first three properties (and thus all of them) in Theorem 1 are called *sub-gaussian*.

- The best K_3 is called the sub-gaussian norm of X , and is usually denoted $\|X\|_{\psi_2}$, that is

$$\|X\|_{\psi_2} := \inf\{t > 0 : \mathbb{E} \exp(X^2/t^2) \leq 2\}.$$

- Sub-gaussian random variable examples.
 - (i) Normal random variables $X \sim \mathcal{N}(\mu, \sigma^2)$.
 - (ii) Bernoulli random variable $X = 0, 1$ with probabilities $1/2$ each.
 - (iii) More generally, any bounded random variable X .
- Not sub-gaussian random variable examples.
Poisson, exponential, Pareto and Cauchy distributions.

Theorem 3 (Sums of sub-gaussians)

Let X_1, \dots, X_N be independent, mean zero, sub-gaussian random variables. Then $\sum_{i=1}^N X_i$ is sub-gaussian, and

$$\left\| \sum_{i=1}^N X_i \right\|_{\psi_2}^2 \leq C \sum_{i=1}^N \|X_i\|_{\psi_2}^2,$$

where C is an absolute constant.

Proof. Let us bound the MGF of the sum for any $\lambda \in \mathbb{R}$:

$$\begin{aligned} \mathbb{E} \exp \left(\lambda \sum_{i=1}^N X_i \right) &= \prod_{i=1}^N \mathbb{E} \exp(\lambda X_i) \quad (\text{using independence}) \\ &\leq \prod_{i=1}^N \exp(C\lambda^2 \|X_i\|_{\psi_2}^2) \quad (\text{by last property in Theorem 1}) \\ &= \exp(\lambda^2 K^2) \quad \text{where } K^2 := C \sum_{i=1}^N \|X_i\|_{\psi_2}^2. \quad \square \end{aligned}$$

2.1. Hoeffding's inequality

- We rewrite Theorem 3 as a concentration inequality by using the first property in Theorem 1.

Theorem 4 (Hoeffding's inequality)

Let X_1, \dots, X_N be independent, mean zero, sub-gaussian random variables. Then, for every $t \geq 0$ we have

$$\mathbb{P} \left\{ \left| \sum_{i=1}^N X_i \right| \geq t \right\} \leq 2 \exp \left(- \frac{Ct^2}{\sum_{i=1}^N \|X_i\|_{\psi_2}^2} \right).$$

Remark 5

Hoeffding's inequality controls how far and with what probability a sum of independent random variables can deviate from its mean, which is zero.

3. Sub-exponential distributions

- The square X^2 of a normal random variable $X \sim \mathcal{N}(0, 1)$ is not sub-gaussian.

Theorem 6 (Sub-exponential properties)

For a random variable X , the following properties are equivalent.

- *Tail:* $\mathbb{P}\{|X| \geq t\} \leq 2 \exp(-t/K_1)$ for all $t \geq 0$.
- *Moment:* $\|X\|_p := (\mathbb{E}|X|^p)^{1/p} \leq pK_2$ for all $p \geq 1$.
- *MGF of square:* $\mathbb{E} \exp(|X|/K_3) \leq 2$.

Moreover, if $\mathbb{E}X = 0$ then these properties imply the following one:

- *MGF:* $\mathbb{E} \exp(\lambda X) \leq \exp(\lambda^2 K_4^2)$ for $|\lambda| \leq 1/K_4$.

Definition 7

Random variables that satisfy one of the first three properties (and thus all of them) in Theorem 6 are called *sub-exponential*.

- The best K_3 is called the sub-exponential norm of X , and is usually denoted $\|X\|_{\psi_1}$, that is

$$\|X\|_{\psi_1} := \inf\{t > 0 : \mathbb{E} \exp(|X|/t) \leq 2\}.$$

- All squares of sub-gaussian random variables are sub-exponential random variables. (Conversely?) We have

$$\|X^2\|_{\psi_1} = \|X\|_{\psi_2}^2.$$

3.1 Bernstein's inequality

Theorem 8 (Bernstein's inequality)

Let X_1, \dots, X_N be independent, mean zero, sub-exponential random variables. Then, for every $t \geq 0$ we have

$$\mathbb{P} \left\{ \left| \sum_{i=1}^N X_i \right| \geq t \right\} \leq 2 \exp \left[-C \min \left(\frac{t^2}{\sum_{i=1}^N \|X_i\|_{\psi_1}^2}, \frac{t}{\max_i \|X_i\|_{\psi_1}} \right) \right].$$

Proof. Choose $\lambda \geq 0$ and use Markov's inequality to get ($S = \sum_{i=1}^N X_i$)

$$\mathbb{P}\{S \geq t\} = \mathbb{P}\{\exp(\lambda S) \geq \exp(\lambda t)\} \leq e^{-\lambda t} \mathbb{E} \exp(\lambda S).$$

Then by independence, we have

$$\mathbb{P}\{S \geq t\} \leq e^{-\lambda t} \prod_{i=1}^N \mathbb{E} \exp(\lambda X_i).$$

If we choose λ small enough so that $0 < \lambda \leq \frac{C}{\max_i \|X_i\|_{\psi_1}}$, then by the last property in Theorem 6 we have

$$\mathbb{E} \exp(\lambda X_i) \leq \exp\left(C\lambda^2 \|X_i\|_{\psi_1}^2\right).$$

Hence,

$$\mathbb{P}\{S \geq t\} \leq \exp\left(-\lambda t + C\lambda^2 \sigma^2\right), \quad \sigma^2 = \sum_{i=1}^N \|X_i\|_{\psi_1}^2.$$

The remaining part is left as an exercise. □

- Why does Bernstein's inequality have a mixture of two tails?
 - (i) The sub-exponential tail should of course be there. Indeed, even if the entire sum consisted of a single term X_i , the best bound we could hope for would be of the form $\exp(-Ct/\|X_i\|_{\psi_1})$.
 - (ii) The sub-gaussian term could be explained by the central limit theorem, which states that the sum should become approximately normal as the number of terms N increases to infinity.

Remark 9 (Bernstein's inequality for bounded random variables)

Suppose further the random variables X_i are uniformly bounded, which is a stronger assumption than being sub-gaussian. If $K > 0$ is such that $|X_i| \leq K$ almost surely for all i , then, for every $t \geq 0$, we have

$$\mathbb{P} \left\{ \left| \sum_{i=1}^N X_i \right| \geq t \right\} \leq 2 \exp \left(-\frac{t^2/2}{\sigma^2 + CKt} \right),$$

where $\sigma^2 = \sum_{i=1}^N \mathbb{E}X_i^2$ is the variance of the sum.

- Note that $\sigma^2 + CKt \leq 2 \max(\sigma^2, CKt)$. So we can state the probability bound as

$$2 \exp \left[-C \min \left(\frac{t^2}{\sigma^2}, \frac{t}{K} \right) \right].$$

Just as before, here we also have a mixture of two tails, sub-gaussian and sub-exponential.

The sub-gaussian tail is a bit *sharper* than in Theorem 8, since it depends on the variances rather than sub-exponential norms of X_i .

The sub-exponential tail, on the other hand, is *weaker*, since it depends on the sup-norms rather than the sub-exponential norms of X_i .

- **More on concentration.** 

4. Sub-gaussian random vectors

- **Definition.** Consider a random vector \mathbf{X} taking values in \mathbb{R}^n . We call \mathbf{X} a sub-gaussian random vector if all one-dimensional marginals of \mathbf{X} , i.e., the random variables $\langle \mathbf{X}, \mathbf{x} \rangle$ for all $\mathbf{x} \in \mathbb{R}^n$, are sub-gaussian.
- The sub-gaussian norm of \mathbf{X} is defined as

$$\|\mathbf{X}\|_{\psi_2} := \sup_{\mathbf{x} \in \mathbb{R}^n, \|\mathbf{x}\|_2=1} \|\langle \mathbf{X}, \mathbf{x} \rangle\|_{\psi_2}.$$

- Sub-gaussian random vector examples

- (i) The standard normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ (why?)
- (ii) The uniform distribution on the centered Euclidean sphere of radius \sqrt{n}
- (iii) The uniform distribution on the cube $\{-1, 1\}^n$
- (iv) A random vector $\mathbf{X} = (X_1, \dots, X_n)$ with independent and sub-gaussian coordinates is sub-gaussian, $\|\mathbf{X}\|_{\psi_2} \leq C \max_i \|X_i\|_{\psi_2}$.

5. Johnson-Lindenstrauss Lemma

- Concentration inequalities like Hoeffding's and Bernstein's are successfully used in the analysis of algorithms.
- Let us give one example for the problem of dimension reduction. Suppose we have some data that is represented as a set of N points in \mathbb{R}^n . We would like to compress the data by representing it in a lower dimensional space \mathbb{R}^m instead of \mathbb{R}^n with $m \ll n$. By how much can we reduce the dimension without losing the important features of the data?
- The basic result in this direction is the Johnson-Lindenstrauss Lemma. It states that a remarkably simple dimension reduction method works - a random linear map from \mathbb{R}^n to \mathbb{R}^m with $m \sim \log N$. The logarithmic function grows very slowly, so we can usually reduce the dimension dramatically.
- What exactly is a random linear map? We consider an $m \times n$ matrix \mathbf{A} whose rows are independent, mean zero, *isotropic* ($\mathbb{E} [\mathbf{X} \mathbf{X}^\top] = \mathbf{I}_n$) and sub-gaussian random vectors in \mathbb{R}^n .

Theorem 10 (Johnson-Lindenstrauss Lemma)

Let \mathcal{X} be a set of N points in \mathbb{R}^n and $\varepsilon \in (0, 1)$. Consider an $m \times n$ matrix \mathbf{A} whose rows are independent, mean zero, isotropic and sub-gaussian random vectors \mathbf{X}_i in \mathbb{R}^n . Rescale \mathbf{A} by defining the “Gaussian random projection”

$$\mathbf{P} := \mathbf{A} / \sqrt{m}.$$

Assume that

$$m \geq C\varepsilon^{-2} \log N,$$

where C is an appropriately large constant that depends only on the sub-gaussian norms of the vectors \mathbf{X}_i . Then, with high probability (say, 0.99), the map \mathbf{P} preserves the distances between all points in \mathcal{X} with error ε , that is for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$,

$$(1 - \varepsilon) \|\mathbf{x} - \mathbf{y}\|_2 \leq \|\mathbf{P}\mathbf{x} - \mathbf{P}\mathbf{y}\|_2 \leq (1 + \varepsilon) \|\mathbf{x} - \mathbf{y}\|_2.$$

Examples of \mathbf{A} : Gaussian random matrix, `randn(m,n)` in MATLAB; an $m \times n$ matrix with independent Rademacher entries (± 1 with equal probabilities).

Proof. By linearity of \mathbf{P} , $1 - \varepsilon \geq (1 - \varepsilon)^2$, and $1 + \varepsilon \leq (1 + \varepsilon)^2$, it is sufficient to prove that (with high probability)

$$1 - \varepsilon \leq \|\mathbf{P}\mathbf{z}\|_2^2 \leq 1 + \varepsilon \quad \text{for all } \mathbf{z} \in \mathcal{T}$$

where

$$\mathcal{T} := \left\{ \frac{\mathbf{x} - \mathbf{y}}{\|\mathbf{x} - \mathbf{y}\|_2} : \mathbf{x}, \mathbf{y} \in \mathcal{X} \text{ and } \mathbf{x} \neq \mathbf{y} \right\}.$$

By $\mathbf{P}\mathbf{z} = \mathbf{A}\mathbf{z}/\sqrt{m}$, it is enough to show that (with high probability)

$$\left| \frac{1}{m} \sum_{i=1}^m \langle \mathbf{X}_i, \mathbf{z} \rangle^2 - 1 \right| \leq \varepsilon \quad \text{for all } \mathbf{z} \in \mathcal{T}.$$

We can prove this inequality by combining concentration and a union bound.

In order to use concentration, we first fix $\mathbf{z} \in \mathcal{T}$. By assumption, the random variables $\langle \mathbf{X}_i, \mathbf{z} \rangle^2 - 1$ are independent; they have zero mean (why? Exercise), and they are sub-exponential (why? Exercise). Then Bernstein's inequality gives (why? Exercise)

$$\mathbb{P} \left\{ \left| \frac{1}{m} \sum_{i=1}^m \langle \mathbf{X}_i, \mathbf{z} \rangle^2 - 1 \right| > \varepsilon \right\} \leq 2 \exp(-c\varepsilon^2 m).$$

Finally, we can unfix \mathbf{z} by taking a union bound over all possible $\mathbf{z} \in \mathcal{T}$:

$$\begin{aligned} \mathbb{P} \left\{ \max_{\mathbf{z} \in \mathcal{T}} \left| \frac{1}{m} \sum_{i=1}^m \langle \mathbf{X}_i, \mathbf{z} \rangle^2 - 1 \right| > \varepsilon \right\} &\leq \sum_{\mathbf{z} \in \mathcal{T}} \mathbb{P} \left\{ \left| \frac{1}{m} \sum_{i=1}^m \langle \mathbf{X}_i, \mathbf{z} \rangle^2 - 1 \right| > \varepsilon \right\} \\ &\leq |\mathcal{T}| \cdot 2 \exp(-c\varepsilon^2 m). \end{aligned}$$

By definition of \mathcal{T} , we have $|\mathcal{T}| \leq N^2$. So, if we choose $m \geq C\varepsilon^{-2} \log N$ with appropriately large constant C , we can make

$$|\mathcal{T}| \cdot 2 \exp(-c\varepsilon^2 m) \leq 0.01.$$

The proof is complete. □