Selected Topics in Mathematics of Learning

High-Dimensional Statistics

Lecturer: Marius Yamakou

Winter Semester 2024/25

Department of Data Science, FAU

November 5, 2024

Part II

Concentration bounds

Objectives:

- Understand concentration bounds by examining classical examples and their relevance in probabilistic analysis.
- Define sub-Gaussian random variables and identify their properties, focusing on their tail bounds and sum behavior.

Part II

Concentration bounds

Objectives:

- Understand concentration bounds by examining classical examples and their relevance in probabilistic analysis.
- Define sub-Gaussian random variables and identify their properties, focusing on their tail bounds and sum behavior.
- Apply key concentration inequalities, such as Hoeffding's and Chernoff's inequalities, to bound probabilities involving sums of sub-Gaussian random variables.
- Recognize and use different characterizations of sub-Gaussianity, understanding the equivalences and implications of these definitions.

Part II

Concentration bounds

Objectives:

- Understand concentration bounds by examining classical examples and their relevance in probabilistic analysis.
- Define sub-Gaussian random variables and identify their properties, focusing on their tail bounds and sum behavior.
- Apply key concentration inequalities, such as Hoeffding's and Chernoff's inequalities, to bound probabilities involving sums of sub-Gaussian random variables.
- Recognize and use different characterizations of sub-Gaussianity, understanding the equivalences and implications of these definitions.
- Explore sub-exponential concentration and learn how it extends concepts of concentration to a broader class of random variables with heavier tails than sub-Gaussian random variables.

Outline

Outline

1. Concentration bounds: Classical examples
2. Sub-Gaussian Random variables
    1. Tail bound
    2. Sum of sub-Gaussian RVs
    3. Hoeffding
    4. Chernoff

Outline

1. Concentration Bounds

Definition:

- A concentration bound is a type of inequality that provides an upper bound on the probability that a random variable deviates significantly from a central value (often its mean or median). It is crucial in high-dimensional settings where traditional low-dimensional intuition may fail due to the curse of dimensionality.

1. Concentration Bounds

Definition:

- A concentration bound is a type of inequality that provides an upper bound on the probability that a random variable deviates significantly from a central value (often its mean or median). It is crucial in high-dimensional settings where traditional low-dimensional intuition may fail due to the curse of dimensionality.
- Mathematically, for a random variable X with mean μ. A concentration inequality typically takes the form:

$$\mathbb{P}(|X - \mu| \geq t) \leq \varphi(t),$$

where $t > 0$ and $\varphi(t)$ is a function that decays as t increases.

1.1 Purpose of concentration bounds

Why are concentration bounds useful?

- Provide probabilistic guarantees, which are crucial in uncertain environments like machine learning, high-dimensional data analysis, and statistical estimation.

1.1 Purpose of concentration bounds

Why are concentration bounds useful?

- Provide probabilistic guarantees, which are crucial in uncertain environments like machine learning, high-dimensional data analysis, and statistical estimation.
- They help estimate how closely a random variable concentrates around its mean, which is essential in high-dimensional settings where traditional low-dimensional intuition may fail due to the curse of dimensionality.

1.1 Purpose of concentration bounds

Why are concentration bounds useful?

- Provide probabilistic guarantees, which are crucial in uncertain environments like machine learning, high-dimensional data analysis, and statistical estimation.
- They help estimate how closely a random variable concentrates around its mean, which is essential in high-dimensional settings where traditional low-dimensional intuition may fail due to the curse of dimensionality.
- Concentration bounds are typically non-asymptotic, providing probabilistic guarantees that hold for a finite number of observations or trials instead of relying on limits as the sample size $n \to \infty$. This is useful in practical settings where the sample size is fixed or small, allowing analysts to make probabilistic statements about deviations without relying on large-sample approximations.

1.2 Intuition behind concentration bounds

What do concentration bounds tell us?

- They provide insights into how likely certain outcomes are, especially deviations from expected values.

1.2 Intuition behind concentration bounds

What do concentration bounds tell us?

- They provide insights into how likely certain outcomes are, especially deviations from expected values.
- Example: If X is a random variable with mean μ:

$$\mathbb{P}(|X - \mu| \geq t) \leq \exp(-Ct^2), \quad t > 0$$

implies that large deviations from μ are exponentially unlikely.

1.3 Concentration bounds and the Law of Large Numbers

Relation to the Law of Large Numbers (LLN):

- LLN states that as the number of samples n increases, the sample average converges to the expected value.

1.3 Concentration bounds and the Law of Large Numbers

Relation to the Law of Large Numbers (LLN):

- LLN states that as the number of samples n increases, the sample average converges to the expected value.
- Concentration bounds strengthen this by providing explicit probabilities for deviations:

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq \epsilon\right) \leq \exp(-Cn\epsilon^2).$$

- This tells us how the sample mean is likely to deviate from the true mean, even for finite n.

1.4 Example: Estimating sample mean

Illustrating Concentration with Sample Mean:

- Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with mean μ.

1.4 Example: Estimating sample mean

Illustrating Concentration with Sample Mean:

- Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with mean μ.
- By applying a concentration bound, we can bound the probability that the sample mean $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$ deviates from μ:

$$\mathbb{P}(|\hat{\mu} - \mu| \geq \epsilon) \leq \exp(-Cn\epsilon^2).$$

1.4 Example: Estimating sample mean

Illustrating Concentration with Sample Mean:

- Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with mean μ.
- By applying a concentration bound, we can bound the probability that the sample mean $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$ deviates from μ:

$$\mathbb{P}(|\hat{\mu} - \mu| \geq \epsilon) \leq \exp(-Cn\epsilon^2).$$

- This indicates that as n increases, the probability of a large deviation becomes exponentially smaller.

1.5 Understanding the “looseness" or “tightness" of concentration inequalities

- Concentration inequalities provide bounds on how a random variable deviates from a typical (mean or median) value.

1.5 Understanding the “looseness" or “tightness" of concentration inequalities

- Concentration inequalities provide bounds on how a random variable deviates from a typical (mean or median) value.
- The “looseness" of a concentration inequality refers to how far the bound provided by the inequality is from the actual probability of a given event.

1.5 Understanding the "looseness" or "tightness" of concentration inequalities

- Concentration inequalities provide bounds on how a random variable deviates from a typical (mean or median) value.
- The "looseness" of a concentration inequality refers to how far the bound provided by the inequality is from the actual probability of a given event.
- If a concentration inequality is loose (i.e., not tight), it means:
  - The inequality gives an upper bound that is much higher than the true probability.
  - The result is *less precise* and potentially less useful in applications.

1.5 Understanding the “looseness" or “tightness" of concentration inequalities

- Concentration inequalities provide bounds on how a random variable deviates from a typical (mean or median) value.
- The “looseness" of a concentration inequality refers to how far the bound provided by the inequality is from the actual probability of a given event.
- If a concentration inequality is loose (i.e., not tight), it means:
  - The inequality gives an upper bound that is much higher than the true probability.
  - The result is *less precise* and potentially less useful in applications.

Causes of Looseness

- *Minimal assumptions and higher moments ignored*: Many concentration inequalities (e.g., Markov’s and Chebyshev’s) only use basic properties like the mean or variance, and ignore other distributional features like skewness or kurtosis.

1.5 Understanding the “looseness" or “tightness" of concentration inequalities

- Concentration inequalities provide bounds on how a random variable deviates from a typical (mean or median) value.
- The “looseness" of a concentration inequality refers to how far the bound provided by the inequality is from the actual probability of a given event.
- If a concentration inequality is loose (i.e., not tight), it means:
  - The inequality gives an upper bound that is much higher than the true probability.
  - The result is *less precise* and potentially less useful in applications.

Causes of Looseness

- *Minimal assumptions and higher moments ignored*: Many concentration inequalities (e.g., Markov’s and Chebyshev’s) only use basic properties like the mean or variance, and ignore other distributional features like skewness or kurtosis.
- *Lack of Tail Behavior Information*: Inequalities that do not leverage information about the tail behavior of the distribution tend to be looser.

1.6 Common Concentration Inequalities: Markov's Inequality

Statement: For any non-negative random variable X and $a > 0$:

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

1.6 Common Concentration Inequalities: Markov's Inequality

Statement: For any non-negative random variable X and $a > 0$:

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

Derivation:

- Let $X \geq 0$. Using the definition of expectation: $\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X \geq t)\, dt$.
- Splitting the integral at a: $\mathbb{E}[X] \geq \int_a^\infty \mathbb{P}(X \geq a)\, dt = \mathbb{P}(X \geq a) \cdot a$.
- Rearranging gives Markov's inequality: $\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$.

1.6 Common Concentration Inequalities: Markov's Inequality

Statement: For any non-negative random variable X and $a > 0$:

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

Derivation:

- Let $X \geq 0$. Using the definition of expectation: $\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X \geq t)\, dt$.
- Splitting the integral at a: $\mathbb{E}[X] \geq \int_a^\infty \mathbb{P}(X \geq a)\, dt = \mathbb{P}(X \geq a) \cdot a$.
- Rearranging gives Markov's inequality: $\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$.

Example: Suppose X represents the number of customers in a queue with $\mathbb{E}[X] = 5$:

$$\mathbb{P}(X \geq 15) \leq \frac{5}{15} = \frac{1}{3}.$$

1.6 Common Concentration Inequalities: Markov's Inequality

Statement: For any non-negative random variable X and $a > 0$:

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

Derivation:

- Let $X \geq 0$. Using the definition of expectation: $\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X \geq t)\, dt$.
- Splitting the integral at a: $\mathbb{E}[X] \geq \int_a^\infty \mathbb{P}(X \geq a)\, dt = \mathbb{P}(X \geq a) \cdot a$.
- Rearranging gives Markov's inequality: $\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$.

Example: Suppose X represents the number of customers in a queue with $\mathbb{E}[X] = 5$:

$$\mathbb{P}(X \geq 15) \leq \frac{5}{15} = \frac{1}{3}.$$

Use Case: Markov's inequality is useful when only the mean of a random variable is known. It is generally loose as it uses only the mean, providing an often-loose bound.

1.6 Common Concentration Inequalities: Markov's Inequality

Let X be a random variable that takes values: $X = \begin{cases} 0 & \text{with probability } 0.9, \\ 10 & \text{with probability } 0.1. \end{cases}$

1.6 Common Concentration Inequalities: Markov's Inequality

Let X be a random variable that takes values: $X = \begin{cases} 0 & \text{with probability } 0.9, \\ 10 & \text{with probability } 0.1. \end{cases}$

The mean of X is: $E[X] = 0.9 \cdot 0 + 0.1 \cdot 10 = 1$.

Using Markov's inequality, for any $a > 0$, $P(X \geq a) \leq \frac{E[X]}{a} = \frac{1}{a}$.
Let's apply this for $a = 5$:

$$P(X \geq 5) \leq \frac{1}{5} = 0.2.$$

1.6 Common Concentration Inequalities: Markov's Inequality

Let X be a random variable that takes values: $X = \begin{cases} 0 & \text{with probability } 0.9, \\ 10 & \text{with probability } 0.1. \end{cases}$

The mean of X is: $E[X] = 0.9 \cdot 0 + 0.1 \cdot 10 = 1$.

Using Markov's inequality, for any $a > 0$, $P(X \geq a) \leq \dfrac{E[X]}{a} = \dfrac{1}{a}$.
Let's apply this for $a = 5$:

$$P(X \geq 5) \leq \frac{1}{5} = 0.2.$$

However, we can directly calculate $P(X \geq 5)$:

$$P(X \geq 5) = P(X = 10) = 0.1.$$

1.6 Common Concentration Inequalities: Markov's Inequality

Let X be a random variable that takes values: $X = \begin{cases} 0 & \text{with probability } 0.9, \\ 10 & \text{with probability } 0.1. \end{cases}$

The mean of X is: $E[X] = 0.9 \cdot 0 + 0.1 \cdot 10 = 1$.

Using Markov's inequality, for any $a > 0$, $P(X \geq a) \leq \frac{E[X]}{a} = \frac{1}{a}$.
Let's apply this for $a = 5$:

$$P(X \geq 5) \leq \frac{1}{5} = 0.2.$$

However, we can directly calculate $P(X \geq 5)$:

$$P(X \geq 5) = P(X = 10) = 0.1.$$

Conclusion: The bound provided by Markov's inequality (0.2) is much larger than the actual probability (0.1), demonstrating that Markov's inequality can be loose or "not tight" in this case.

1.6 Common Concentration Inequalities: Chebyshev's Inequality

Statement: For a random variable X with mean μ and variance σ^2:

$$\mathbb{P}(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}, \quad \text{for any } k > 0.$$

1.6 Common Concentration Inequalities: Chebyshev's Inequality

Statement: For a random variable X with mean μ and variance σ^2:

$$\mathbb{P}(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}, \quad \text{for any } k > 0.$$

Derivation:

- Apply Markov's inequality to $Y = (X - \mu)^2$:

$$\mathbb{P}((X - \mu)^2 \geq k^2\sigma^2) \leq \frac{\mathbb{E}[(X - \mu)^2]}{k^2\sigma^2} = \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}.$$

- This gives Chebyshev's inequality:

$$\mathbb{P}(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

1.6 Common Concentration Inequalities: Chebyshev's Inequality

Statement: For a random variable X with mean μ and variance σ^2:

$$\mathbb{P}(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}, \quad \text{for any } k > 0.$$

Derivation:

- Apply Markov's inequality to $Y = (X - \mu)^2$:

$$\mathbb{P}((X - \mu)^2 \geq k^2\sigma^2) \leq \frac{\mathbb{E}[(X - \mu)^2]}{k^2\sigma^2} = \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}.$$

- This gives Chebyshev's inequality:

$$\mathbb{P}(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

Example: For X with $\mu = 10$ and $\sigma^2 = 4$:

$$\mathbb{P}(|X - 10| \geq 4) \leq \frac{1}{4} = 0.25.$$

1.6 Common Concentration Inequalities: Chebyshev's Inequality

Statement: For a random variable X with mean μ and variance σ^2:

$$\mathbb{P}(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}, \quad \text{for any } k > 0.$$

Derivation:

- Apply Markov's inequality to $Y = (X - \mu)^2$:

$$\mathbb{P}((X - \mu)^2 \geq k^2\sigma^2) \leq \frac{\mathbb{E}[(X - \mu)^2]}{k^2\sigma^2} = \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}.$$

- This gives Chebyshev's inequality:

$$\mathbb{P}(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

Example: For X with $\mu = 10$ and $\sigma^2 = 4$:

$$\mathbb{P}(|X - 10| \geq 4) \leq \frac{1}{4} = 0.25.$$

Use Case: Useful when only mean and variance are known, providing bounds for all distributions with finite variance. It still can be quite loose because it ignores other properties like skewness or kurtosis.

1.7 Key takeaways on the concept of concentration bounds so far

- Concentration bounds provide a way to quantify the deviation of random variables from their mean or median.

1.7 Key takeaways on the concept of concentration bounds so far

- Concentration bounds provide a way to quantify the deviation of random variables from their mean or median.
- They are critical for understanding behavior in random processes, deriving error bounds, and analyzing algorithms in machine learning.

1.7 Key takeaways on the concept of concentration bounds so far

- Concentration bounds provide a way to quantify the deviation of random variables from their mean or median.
- They are critical for understanding behavior in random processes, deriving error bounds, and analyzing algorithms in machine learning.
- Concentration bounds strengthen the law of large numbers by giving specific probabilities for how close a sample average is to its expectation.

1.7 Key takeaways on the concept of concentration bounds so far

- Concentration bounds provide a way to quantify the deviation of random variables from their mean or median.
- They are critical for understanding behavior in random processes, deriving error bounds, and analyzing algorithms in machine learning.
- Concentration bounds strengthen the law of large numbers by giving specific probabilities for how close a sample average is to its expectation.
- Concentration bounds can be relatively loose (or tight).

1.7 Key takeaways on the concept of concentration bounds so far

- Concentration bounds provide a way to quantify the deviation of random variables from their mean or median.
- They are critical for understanding behavior in random processes, deriving error bounds, and analyzing algorithms in machine learning.
- Concentration bounds strengthen the law of large numbers by giving specific probabilities for how close a sample average is to its expectation.
- Concentration bounds can be relatively loose (or tight).
- Concentration bounds are typically non-asymptotic, including non-asymptotic versions of CLT.

1.8 Some Motivation

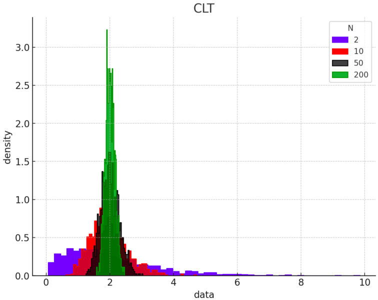
Lemma (Asymptotic CLT)

Let $X_1, X_2, \ldots$ be i.i.d. random variables with mean μ and variance σ^2. Consider the sum:

$$S_N = X_1 + \cdots + X_N$$

and $Z_N = \frac{S_N - \mathbb{E}S_N}{\sqrt{Var(S_N)}}$. *Then, as* $N \to \infty$,

$$Z_N \to \mathcal{N}(0,1) \quad \textit{in distribution.}$$



Visual Insight: As we add more i.i.d. random variables, the distribution of their normalized sum approaches a normal distribution, even if the original variables are not normally distributed.

1.8 Some Motivation

- Consider independent Bernoulli random variables $X_1, \ldots, X_n \sim Ber(1/2)$, each representing a simple coin flip with a success probability of $1/2$. Define $S_n = \sum_{i=1}^{n} X_i$, the total number of successes in n trials.

1.8 Some Motivation

- Consider independent Bernoulli random variables $X_1, \ldots, X_n \sim Ber(1/2)$, each representing a simple coin flip with a success probability of $1/2$. Define $S_n = \sum_{i=1}^{n} X_i$, the total number of successes in n trials.
- By the Central Limit Theorem (CLT), the standardized sum S_n converges in distribution to a normal distribution:

$$Z_n := \frac{S_n - n/2}{\sqrt{n/4}} \xrightarrow{D} \mathcal{N}(0, 1)$$

  This result implies that for large n, S_n behaves approximately like a normal variable centered at $n/2$ with variance $n/4$.

1.8 Some Motivation

- Consider independent Bernoulli random variables $X_1, \ldots, X_n \sim Ber(1/2)$, each representing a simple coin flip with a success probability of $1/2$. Define $S_n = \sum_{i=1}^n X_i$, the total number of successes in n trials.
- By the Central Limit Theorem (CLT), the standardized sum S_n converges in distribution to a normal distribution:

$$Z_n := \frac{S_n - n/2}{\sqrt{n/4}} \xrightarrow{D} \mathcal{N}(0,1)$$

  This result implies that for large n, S_n behaves approximately like a normal variable centered at $n/2$ with variance $n/4$.
- Using this approximation, we can bound the probability that S_n deviates from its mean. Let $G \sim \mathcal{N}(0,1)$, so:

$$\mathbb{P}\left(Z_n > t\right) = \mathbb{P}\left(\frac{S_n - n/2}{\sqrt{n/4}} > t\right) = \mathbb{P}\left(S_n > \frac{n}{2} + \sqrt{\frac{n}{4}}t\right).$$

1.8 Some Motivation

- Using CLT,

$$\mathbb{P}\left(S_n > \frac{n}{2} + \sqrt{\frac{n}{4}}t\right) \overset{CLT}{\approx} \mathbb{P}(G > t).$$

We have

$M_G(\lambda) := \mathbb{E}[e^{\lambda G}] = \int_{-\infty}^{\infty} e^{\lambda g} \cdot f_G(g) dg = \int_{-\infty}^{\infty} e^{\lambda g} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{g^2}{2}} \overset{?}{=} e^{\frac{\lambda^2}{2}},$

and we also have

1.8 Some Motivation

- Using CLT,

$$\mathbb{P}\left(S_n > \frac{n}{2} + \sqrt{\frac{n}{4}}t\right) \overset{CLT}{\approx} \mathbb{P}(G > t).$$

We have

$M_G(\lambda) := \mathbb{E}[e^{\lambda G}] = \int_{-\infty}^{\infty} e^{\lambda g} \cdot f_G(g)dg = \int_{-\infty}^{\infty} e^{\lambda g} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{g^2}{2}} \overset{?}{=} e^{\frac{\lambda^2}{2}},$

and we also have

$\mathbb{P}(G > t) = \mathbb{P}(e^{\lambda G} > e^{\lambda t}) \overset{Markov}{\leq} \frac{\mathbb{E}[e^{\lambda G}]}{e^{\lambda t}} = \frac{e^{\frac{\lambda^2}{2}}}{e^{\lambda t}} = e^{\frac{\lambda^2}{2} - \lambda t}$

1.8 Some Motivation

- Using CLT,

$$\mathbb{P}\left(S_n > \frac{n}{2} + \sqrt{\frac{n}{4}}t\right) \overset{CLT}{\approx} \mathbb{P}(G > t).$$

We have
$M_G(\lambda) := \mathbb{E}[e^{\lambda G}] = \int_{-\infty}^{\infty} e^{\lambda g} \cdot f_G(g)dg = \int_{-\infty}^{\infty} e^{\lambda g} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{g^2}{2}} \overset{?}{=} e^{\frac{\lambda^2}{2}}$,
and we also have
$\mathbb{P}(G > t) = \mathbb{P}(e^{\lambda G} > e^{\lambda t}) \overset{Markov}{\leq} \frac{\mathbb{E}[e^{\lambda G}]}{e^{\lambda t}} = \frac{e^{\frac{\lambda^2}{2}}}{e^{\lambda t}} = e^{\frac{\lambda^2}{2} - \lambda t}$
To get the best bound, we choose $\lambda = t$ which minimizes the upper bound, and we get $\mathbb{P}(G > t) \leq e^{-\frac{t^2}{2}}$, which yields:

$$\mathbb{P}\left(S_n > \frac{n}{2} + \sqrt{\frac{n}{4}}t\right) \overset{CLT}{\approx} \mathbb{P}(G > t) \lesssim \frac{1}{2} e^{-\frac{t^2}{2}}$$

1.8 Some Motivation

- Using CLT,

$$\mathbb{P}\left(S_n > \frac{n}{2} + \sqrt{\frac{n}{4}}t\right) \overset{CLT}{\approx} \mathbb{P}(G > t).$$

We have

$M_G(\lambda) := \mathbb{E}[e^{\lambda G}] = \int_{-\infty}^{\infty} e^{\lambda g} \cdot f_G(g)dg = \int_{-\infty}^{\infty} e^{\lambda g} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{g^2}{2}} \overset{?}{=} e^{\frac{\lambda^2}{2}}$,

and we also have

$\mathbb{P}(G > t) = \mathbb{P}(e^{\lambda G} > e^{\lambda t}) \overset{Markov}{\leq} \frac{\mathbb{E}[e^{\lambda G}]}{e^{\lambda t}} = \frac{e^{\frac{\lambda^2}{2}}}{e^{\lambda t}} = e^{\frac{\lambda^2}{2} - \lambda t}$

To get the best bound, we choose $\lambda = t$ which minimizes the upper bound, and we get $\mathbb{P}(G > t) \leq e^{-\frac{t^2}{2}}$, which yields:

$$\mathbb{P}\left(S_n > \frac{n}{2} + \sqrt{\frac{n}{4}}t\right) \overset{CLT}{\approx} \mathbb{P}(G > t) \lesssim \frac{1}{2} e^{-\frac{t^2}{2}}$$

where the $\frac{1}{2}$ factor accounts for the symmetric nature of the normal distribution (since $\mathbb{P}(G > 0) = \frac{1}{2}$). This provides an exponential decay rate for the tail probability, highlighting the rarity of large deviations.

1.8 Some Motivation

- Setting $t = \alpha\sqrt{n}$ yields:

$$\mathbb{P}\left(S_n > \frac{n}{2}(1+\alpha)\right) \lesssim \frac{1}{2}e^{-\frac{n\alpha^2}{2}}$$

Here, we see that deviations on the order of $\alpha\sqrt{n}$ in S_n become exponentially unlikely as n grows.

1.8 Some Motivation

- Setting $t = \alpha\sqrt{n}$ yields:

$$\mathbb{P}\left(S_n > \frac{n}{2}(1+\alpha)\right) \lesssim \frac{1}{2}e^{-\frac{n\alpha^2}{2}}$$

  Here, we see that deviations on the order of $\alpha\sqrt{n}$ in S_n become exponentially unlikely as n grows.

- Problem: Although the CLT gives a general approximation, it may not always be tight, especially for finite n or large α. Improving this bound requires more refined probabilistic techniques.

1.8.1 Berry-Esseen Central Limit Theorem

Theorem (Berry-Esseen Central Limit Theorem)

Under the assumptions of the Central Limit Theorem, with $\delta = \frac{\mathbb{E}|X_1 - \mu|^3}{\sigma^3}$*, we have:* $|\mathbb{P}(Z_n > t) - \mathbb{P}(G > t)| \leq \frac{\delta}{\sqrt{n}}$*, where* Z_n *is the standardized sum, and* $G \sim \mathcal{N}(0, 1)$ *is the standard normal distribution.*

1.8.1 Berry-Esseen Central Limit Theorem

Theorem (Berry-Esseen Central Limit Theorem)

Under the assumptions of the Central Limit Theorem, with $\delta = \frac{\mathbb{E}|X_1 - \mu|^3}{\sigma^3}$, *we have:* $|\mathbb{P}(Z_n > t) - \mathbb{P}(G > t)| \leq \frac{\delta}{\sqrt{n}}$, *where* Z_n *is the standardized sum, and* $G \sim \mathcal{N}(0, 1)$ *is the standard normal distribution.*

- Interpretation: The bound quantifies the rate at which the distribution of Z_n converges to the normal distribution.

1.8.1 Berry-Esseen Central Limit Theorem

Theorem (Berry-Esseen Central Limit Theorem)

Under the assumptions of the Central Limit Theorem, with $\delta = \frac{\mathbb{E}|X_1 - \mu|^3}{\sigma^3}$*, we have:* $|\mathbb{P}(Z_n > t) - \mathbb{P}(G > t)| \leq \frac{\delta}{\sqrt{n}}$*, where* Z_n *is the standardized sum, and* $G \sim \mathcal{N}(0,1)$ *is the standard normal distribution.*

- Interpretation: The bound quantifies the rate at which the distribution of Z_n converges to the normal distribution.
- Tight Bound Example: In the case of a Bernoulli random variable, since each $X_i \sim \text{Ber}(1/2)$, the probability of each outcome is $1/2$, and the sum $S_n = \sum_{i=1}^{n} X_i$ follows a binomial distribution: $S_n \sim \text{Binomial}(n, 1/2)$. For the event $S_n = n/2$ (exactly half successes), the probability is: $\mathbb{P}(S_n = n/2) = \binom{n}{n/2} \left(\frac{1}{2}\right)^n = \frac{1}{2^n}\binom{n}{n/2}$.

1.8.1 Berry-Esseen Central Limit Theorem

For large n, we can approximate $\binom{n}{n/2}$ using Stirling's approximation, which states that for large k, $k! \approx \sqrt{2\pi k}\left(\frac{k}{e}\right)^k$.

1.8.1 Berry-Esseen Central Limit Theorem

For large n, we can approximate $\binom{n}{n/2}$ using Stirling's approximation, which states that for large k, $k! \approx \sqrt{2\pi k}\left(\frac{k}{e}\right)^k$.

Applying Stirling's approximation to $\binom{n}{n/2} = \frac{n!}{(n/2)!(n/2)!}$ yields $\binom{n}{n/2} \approx \frac{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n}{\left(2\pi \cdot \frac{n}{2}\right)\left(\frac{n}{2e}\right)^n} = \frac{2^n}{\sqrt{\pi n}}$.

1.8.1 Berry-Esseen Central Limit Theorem

For large n, we can approximate $\binom{n}{n/2}$ using Stirling's approximation, which states that for large k, $k! \approx \sqrt{2\pi k}\left(\frac{k}{e}\right)^k$.

Applying Stirling's approximation to $\binom{n}{n/2} = \frac{n!}{(n/2)!(n/2)!}$ yields $\binom{n}{n/2} \approx \frac{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n}{\left(2\pi \cdot \frac{n}{2}\right)\left(\frac{n}{2e}\right)^n} = \frac{2^n}{\sqrt{\pi n}}$.

Thus, $\mathbb{P}(S_n = n/2) = \frac{1}{2^n}\binom{n}{n/2} \approx \frac{1}{2^n} \cdot \frac{2^n}{\sqrt{\pi n}} = \frac{1}{\sqrt{\pi n}}$.

1.8.1 Berry-Esseen Central Limit Theorem

For large n, we can approximate $\binom{n}{n/2}$ using Stirling's approximation, which states that for large k, $k! \approx \sqrt{2\pi k}\left(\frac{k}{e}\right)^k$.

Applying Stirling's approximation to $\binom{n}{n/2} = \frac{n!}{(n/2)!(n/2)!}$ yields $\binom{n}{n/2} \approx \frac{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n}{\left(2\pi\cdot\frac{n}{2}\right)\left(\frac{n}{2e}\right)^n} = \frac{2^n}{\sqrt{\pi n}}$.

Thus, $\mathbb{P}(S_n = n/2) = \frac{1}{2^n}\binom{n}{n/2} \approx \frac{1}{2^n}\cdot\frac{2^n}{\sqrt{\pi n}} = \frac{1}{\sqrt{\pi n}}$.

For simplicity, this is often expressed as $\mathbb{P}(S_n = n/2) \approx \frac{1}{\sqrt{n}}$, showing that the $\frac{1}{\sqrt{n}}$ rate is optimal for certain distributions.

1.8.1 Berry-Esseen Central Limit Theorem

- Implication: The approximation error is $O\left(\frac{1}{\sqrt{n}}\right)$, which is relatively large compared to an exponential bound like $O(e^{-\frac{n\alpha^2}{2}})$ that we aim to achieve in concentration inequalities.

1.8.1 Berry-Esseen Central Limit Theorem

- Implication: The approximation error is $O\left(\frac{1}{\sqrt{n}}\right)$, which is relatively large compared to an exponential bound like $O(e^{-\frac{n\alpha^2}{2}})$ that we aim to achieve in concentration inequalities.
- Solution: To achieve a tighter bound, we can use concentration inequalities directly, which often yield exponential decay rates.

1.8.1 Berry-Esseen Central Limit Theorem

- Implication: The approximation error is $O\left(\frac{1}{\sqrt{n}}\right)$, which is relatively large compared to an exponential bound like $O(e^{-\frac{n\alpha^2}{2}})$ that we aim to achieve in concentration inequalities.
- Solution: To achieve a tighter bound, we can use concentration inequalities directly, which often yield exponential decay rates.
- Method: Chernoff Bound: For any $\lambda > 0$,

$$\mathbb{P}(Z_n > t) = \mathbb{P}(e^{\lambda Z_n} > e^{\lambda t}) \stackrel{Markov}{\leq} \frac{\mathbb{E}[e^{\lambda Z_n}]}{e^{\lambda t}}, \quad t \in \mathbb{R},$$

which relies on analyzing the moment-generating function (MGF) of Z_n.

1.8.1 Berry-Esseen Central Limit Theorem

- Implication: The approximation error is $O\left(\frac{1}{\sqrt{n}}\right)$, which is relatively large compared to an exponential bound like $O(e^{-\frac{n\alpha^2}{2}})$ that we aim to achieve in concentration inequalities.
- Solution: To achieve a tighter bound, we can use concentration inequalities directly, which often yield exponential decay rates.
- Method: Chernoff Bound: For any $\lambda > 0$,

$$\mathbb{P}(Z_n > t) = \mathbb{P}(e^{\lambda Z_n} > e^{\lambda t}) \overset{Markov}{\leq} \frac{\mathbb{E}[e^{\lambda Z_n}]}{e^{\lambda t}}, \quad t \in \mathbb{R},$$

  which relies on analyzing the moment-generating function (MGF) of Z_n.
- NOTE: The moment-generating function (MGF) of Z_n given as $M_{Z_n}(\lambda) = \mathbb{E}[e^{\lambda Z_n}]$, is the expected value of $e^{\lambda Z_n}$, where λ is a real parameter. The MGF, when it exists, is a useful tool for deriving the moments of Z_n by differentiating $M_{Z_n}(\lambda)$ with respect to λ and evaluating at $\lambda = 0$.

2. Sub-Gaussian Random variables

Definition (Sub-Gaussian Random Variable)

A random variable X with mean $\mu = \mathbb{E}[X]$ is called *sub-Gaussian* with parameter $\sigma > 0$ if:

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq e^{\sigma^2\lambda^2/2}, \quad \forall \lambda \in \mathbb{R}.$$

In this case, we say X is σ-sub-Gaussian and has **variance proxy** σ^2.

2. Sub-Gaussian Random variables

Definition (Sub-Gaussian Random Variable)

A random variable X with mean $\mu = \mathbb{E}[X]$ is called *sub-Gaussian* with parameter $\sigma > 0$ if:

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq e^{\sigma^2\lambda^2/2}, \quad \forall \lambda \in \mathbb{R}.$$

In this case, we say X is σ-sub-Gaussian and has **variance proxy** σ^2.

- **Standard Gaussian:** If $X \sim \mathcal{N}(0,1)$, then X is sub-Gaussian with equality.

2. Sub-Gaussian Random variables

Definition (Sub-Gaussian Random Variable)

A random variable X with mean $\mu = \mathbb{E}[X]$ is called *sub-Gaussian* with parameter $\sigma > 0$ if:

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq e^{\sigma^2\lambda^2/2}, \quad \forall \lambda \in \mathbb{R}.$$

In this case, we say X is σ-sub-Gaussian and has **variance proxy** σ^2.

- **Standard Gaussian:** If $X \sim \mathcal{N}(0,1)$, then X is sub-Gaussian with equality.
- **Rademacher Random Variable:** A Rademacher random variable R takes values $+1$ or -1 with probability $\frac{1}{2}$ each, making R a 1-sub-Gaussian random variable.

2. Sub-Gaussian Random variables

Definition (Sub-Gaussian Random Variable)

A random variable X with mean $\mu = \mathbb{E}[X]$ is called *sub-Gaussian* with parameter $\sigma > 0$ if:

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq e^{\sigma^2\lambda^2/2}, \quad \forall \lambda \in \mathbb{R}.$$

In this case, we say X is σ-sub-Gaussian and has **variance proxy** σ^2.

- **Standard Gaussian:** If $X \sim \mathcal{N}(0,1)$, then X is sub-Gaussian with equality.
- **Rademacher Random Variable:** A Rademacher random variable R takes values $+1$ or -1 with probability $\frac{1}{2}$ each, making R a 1-sub-Gaussian random variable.
- **Bounded Random Variable:** If $|X - \mathbb{E}[X]| \leq M$ almost surely, then X is M-sub-Gaussian.

2 Sub-Gaussian Random Variables

To prove that if $X \sim \mathcal{N}(0,1)$, then X is 1-sub-Gaussian with equality, we need to verify that:

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] = e^{\sigma^2\lambda^2/2}$$

where $\mu = \mathbb{E}[X] = 0$ and $\sigma = 1$.

2 Sub-Gaussian Random Variables

To prove that if $X \sim \mathcal{N}(0,1)$, then X is 1-sub-Gaussian with equality, we need to verify that:

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] = e^{\sigma^2\lambda^2/2}$$

where $\mu = \mathbb{E}[X] = 0$ and $\sigma = 1$.

- Step 1: Set up the moment generating function:

For a random variable $X \sim \mathcal{N}(0,1)$, the moment generating function $\mathbb{E}\left[e^{\lambda X}\right]$ can be computed directly. Since X is standard normal, we have that $\mathbb{E}[X] = 0$, and $\mathrm{Var}(X) = 1$. We want to calculate $\mathbb{E}\left[e^{\lambda X}\right]$.

2 Sub-Gaussian Random Variables

To prove that if $X \sim \mathcal{N}(0,1)$, then X is 1-sub-Gaussian with equality, we need to verify that:

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] = e^{\sigma^2\lambda^2/2}$$

where $\mu = \mathbb{E}[X] = 0$ and $\sigma = 1$.

- Step 1: Set up the moment generating function:

For a random variable $X \sim \mathcal{N}(0,1)$, the moment generating function $\mathbb{E}\left[e^{\lambda X}\right]$ can be computed directly. Since X is standard normal, we have that $\mathbb{E}[X] = 0$, and $\mathrm{Var}(X) = 1$. We want to calculate $\mathbb{E}\left[e^{\lambda X}\right]$.

- Step 2: Calculate $\mathbb{E}\left[e^{\lambda X}\right]$ for $X \sim \mathcal{N}(0,1)$

By definition of the expectation for continuous random variables:

$$\mathbb{E}\left[e^{\lambda X}\right] = \int_{-\infty}^{\infty} e^{\lambda x} f_X(x)\, dx$$

where $f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ is the PDF of X, since $X \sim \mathcal{N}(0,1)$.

2 Sub-Gaussian Random Variables

Thus: $\mathbb{E}\left[e^{\lambda X}\right] = \int_{-\infty}^{\infty} e^{\lambda x} \cdot \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx.$

2 Sub-Gaussian Random Variables

Thus: $\mathbb{E}\left[e^{\lambda X}\right] = \int_{-\infty}^{\infty} e^{\lambda x} \cdot \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx.$

- Step 3: Simplify the Integral

Combine the exponential terms: $\mathbb{E}\left[e^{\lambda X}\right] = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\lambda x - \frac{x^2}{2}}\, dx.$

Rewrite the exponent: $\lambda x - \frac{x^2}{2} = -\frac{1}{2}\left(x^2 - 2\lambda x\right).$

Complete the square for $x^2 - 2\lambda x$: $x^2 - 2\lambda x = (x-\lambda)^2 - \lambda^2.$

Substitute back: $\mathbb{E}\left[e^{\lambda X}\right] = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left((x-\lambda)^2 - \lambda^2\right)}\, dx.$

Separate the terms: $\mathbb{E}\left[e^{\lambda X}\right] = e^{\frac{\lambda^2}{2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\lambda)^2}\, dx.$

2 Sub-Gaussian Random Variables

Thus: $\mathbb{E}\left[e^{\lambda X}\right] = \int_{-\infty}^{\infty} e^{\lambda x} \cdot \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx.$

- Step 3: Simplify the Integral

Combine the exponential terms: $\mathbb{E}\left[e^{\lambda X}\right] = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\lambda x - \frac{x^2}{2}}\, dx.$
Rewrite the exponent: $\lambda x - \frac{x^2}{2} = -\frac{1}{2}\left(x^2 - 2\lambda x\right).$
Complete the square for $x^2 - 2\lambda x$: $x^2 - 2\lambda x = (x-\lambda)^2 - \lambda^2.$
Substitute back: $\mathbb{E}\left[e^{\lambda X}\right] = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left((x-\lambda)^2 - \lambda^2\right)}\, dx.$
Separate the terms: $\mathbb{E}\left[e^{\lambda X}\right] = e^{\frac{\lambda^2}{2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\lambda)^2}\, dx.$

- Step 4: Recognize the Gaussian Integral

The integral $\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\lambda)^2}\, dx$ is the integral of a normal distribution with mean λ and variance 1, which equals 1. So: $\mathbb{E}\left[e^{\lambda X}\right] = e^{\frac{\lambda^2}{2}}.$

2 Sub-Gaussian Random Variables

Thus: $\mathbb{E}\left[e^{\lambda X}\right] = \int_{-\infty}^{\infty} e^{\lambda x} \cdot \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx.$

- Step 3: Simplify the Integral

Combine the exponential terms: $\mathbb{E}\left[e^{\lambda X}\right] = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\lambda x - \frac{x^2}{2}}\, dx.$

Rewrite the exponent: $\lambda x - \frac{x^2}{2} = -\frac{1}{2}\left(x^2 - 2\lambda x\right).$

Complete the square for $x^2 - 2\lambda x$: $x^2 - 2\lambda x = (x-\lambda)^2 - \lambda^2.$

Substitute back: $\mathbb{E}\left[e^{\lambda X}\right] = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left((x-\lambda)^2 - \lambda^2\right)}\, dx.$

Separate the terms: $\mathbb{E}\left[e^{\lambda X}\right] = e^{\frac{\lambda^2}{2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\lambda)^2}\, dx.$

- Step 4: Recognize the Gaussian Integral

The integral $\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\lambda)^2}\, dx$ is the integral of a normal distribution with mean λ and variance 1, which equals 1. So: $\mathbb{E}\left[e^{\lambda X}\right] = e^{\frac{\lambda^2}{2}}.$

- Step 5: Conclude that X is 1-sub-Gaussian with equality

This shows that $\mathbb{E}\left[e^{\lambda X}\right] = e^{\frac{\lambda^2}{2}} = e^{\sigma^2\lambda^2/2}$ with $\sigma = 1$. Therefore, X is 1-sub-Gaussian with equality.

2.1 Tail Bound for Sub-Gaussian Random Variables

Theorem (Tail Bound for Sub-Gaussian Random Variables)

If a random variable X with finite mean μ is σ-sub-Gaussian, then for any $t > 0$, $\mathbb{P}(|X - \mu| \geq t) \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right)$.

2.1 Tail Bound for Sub-Gaussian Random Variables

Theorem (Tail Bound for Sub-Gaussian Random Variables)

If a random variable X with finite mean μ is σ-sub-Gaussian, then for any $t > 0$, $\mathbb{P}(|X - \mu| \geq t) \leq 2 \exp\left(-\frac{t^2}{2\sigma^2}\right)$.

- **Intuition**: This inequality tells us that the probability of large deviations from the mean μ decreases exponentially for sub-Gaussian random variables, giving a strong concentration around μ.

2.1 Tail Bound for Sub-Gaussian Random Variables

Theorem (Tail Bound for Sub-Gaussian Random Variables)

If a random variable X with finite mean μ is σ-sub-Gaussian, then for any $t > 0$, $\mathbb{P}(|X - \mu| \geq t) \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right)$.

- **Intuition**: This inequality tells us that the probability of large deviations from the mean μ decreases exponentially for sub-Gaussian random variables, giving a strong concentration around μ.
- **Significance**: Sub-Gaussian tail bounds are widely used in high-dimensional data and concentration inequalities.

Proof (sketched below and the details of the proof left as an exercise (?)):

1. Start by applying Markov's inequality to the exponential moment $\mathbb{E}[\exp(\lambda(X - \mu))]$.
2. Use the sub-Gaussian property to bound this moment for λ in terms of σ.
3. Conclude by optimizing over λ and applying symmetry to achieve the final bound.

2.2 Sum of Sub-Gaussian Random Variables

Proposition (Sum of sub-Gaussian random variables is sub-Gaussian)

If $X_1, \ldots, X_n$ are independent sub-Gaussian random variables with variance proxies $\sigma_1^2, \ldots, \sigma_n^2$, then $Z = \sum_{i=1}^n X_i$ is sub-Gaussian with variance proxy $\sum_{i=1}^n \sigma_i^2$.

2.2 Sum of Sub-Gaussian Random Variables

Proposition (Sum of sub-Gaussian random variables is sub-Gaussian)

If $X_1, \ldots, X_n$ are independent sub-Gaussian random variables with variance proxies $\sigma_1^2, \ldots, \sigma_n^2$, then $Z = \sum_{i=1}^n X_i$ is sub-Gaussian with variance proxy $\sum_{i=1}^n \sigma_i^2$.

Proof:

- Since X_i is sub-Gaussian, there exists σ_i^2 such that $\forall \lambda \in \mathbb{R}$, we have $\mathbb{E}[\exp(\lambda X_i)] \leq \exp\left(\frac{\lambda^2 \sigma_i^2}{2}\right)$.

2.2 Sum of Sub-Gaussian Random Variables

Proposition (Sum of sub-Gaussian random variables is sub-Gaussian)

If $X_1, \dots, X_n$ are independent sub-Gaussian random variables with variance proxies $\sigma_1^2, \dots, \sigma_n^2$, then $Z = \sum_{i=1}^n X_i$ is sub-Gaussian with variance proxy $\sum_{i=1}^n \sigma_i^2$.

Proof:

- Since X_i is sub-Gaussian, there exists σ_i^2 such that $\forall \lambda \in \mathbb{R}$, we have $\mathbb{E}[\exp(\lambda X_i)] \leq \exp\left(\frac{\lambda^2 \sigma_i^2}{2}\right)$.
- For the sum $Z = \sum_{i=1}^n X_i$, we apply the MGF of the sum of independent variables X_i:

$$\mathbb{E}[\exp(\lambda Z)] = \mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^n X_i\right)\right] = \mathbb{E}\prod_{i=1}^n \exp(\lambda X_i) \overset{Indep.}{=} \prod_{i=1}^n \mathbb{E}[\exp(\lambda X_i)].$$

2.2 Sum of Sub-Gaussian Random Variables

Proposition (Sum of sub-Gaussian random variables is sub-Gaussian)

If $X_1, \dots, X_n$ are independent sub-Gaussian random variables with variance proxies $\sigma_1^2, \dots, \sigma_n^2$, then $Z = \sum_{i=1}^n X_i$ is sub-Gaussian with variance proxy $\sum_{i=1}^n \sigma_i^2$.

Proof:

- Since X_i is sub-Gaussian, there exists σ_i^2 such that $\forall \lambda \in \mathbb{R}$, we have $\mathbb{E}[\exp(\lambda X_i)] \leq \exp\left(\frac{\lambda^2 \sigma_i^2}{2}\right)$.
- For the sum $Z = \sum_{i=1}^n X_i$, we apply the MGF of the sum of independent variables X_i:

$$\mathbb{E}[\exp(\lambda Z)] = \mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^n X_i\right)\right] = \mathbb{E} \prod_{i=1}^n \exp(\lambda X_i) \overset{Indep.}{=} \prod_{i=1}^n \mathbb{E}[\exp(\lambda X_i)].$$

- Using the sub-Gaussian property of each X_i, this becomes:

$$\mathbb{E}[\exp(\lambda Z)] = \prod_{i=1}^n \mathbb{E}[\exp(\lambda X_i)] \leq \prod_{i=1}^n \exp\left(\frac{\lambda^2 \sigma_i^2}{2}\right) = \exp\left(\frac{\lambda^2}{2} \sum_{i=1}^n \sigma_i^2\right).$$

2.2 Sum of Sub-Gaussian Random Variables

Proposition (Sum of sub-Gaussian random variables is sub-Gaussian)

If $X_1, \ldots, X_n$ are independent sub-Gaussian random variables with variance proxies $\sigma_1^2, \ldots, \sigma_n^2$, then $Z = \sum_{i=1}^n X_i$ is sub-Gaussian with variance proxy $\sum_{i=1}^n \sigma_i^2$.

Proof:

- Since X_i is sub-Gaussian, there exists σ_i^2 such that $\forall \lambda \in \mathbb{R}$, we have $\mathbb{E}[\exp(\lambda X_i)] \leq \exp\left(\frac{\lambda^2 \sigma_i^2}{2}\right)$.
- For the sum $Z = \sum_{i=1}^n X_i$, we apply the MGF of the sum of independent variables X_i:

$$\mathbb{E}[\exp(\lambda Z)] = \mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^n X_i\right)\right] = \mathbb{E}\prod_{i=1}^n \exp(\lambda X_i) \overset{Indep.}{=} \prod_{i=1}^n \mathbb{E}[\exp(\lambda X_i)].$$

- Using the sub-Gaussian property of each X_i, this becomes:

$$\mathbb{E}[\exp(\lambda Z)] = \prod_{i=1}^n \mathbb{E}[\exp(\lambda X_i)] \leq \prod_{i=1}^n \exp\left(\frac{\lambda^2 \sigma_i^2}{2}\right) = \exp\left(\frac{\lambda^2}{2} \sum_{i=1}^n \sigma_i^2\right).$$

- Therefore, Z is sub-Gaussian with variance proxy $\sum_{i=1}^n \sigma_i^2$.

Theorem (Hoeffding's Inequality for Sub-Gaussian Sums)

Let $X_1, \ldots, X_n$ be independent sub-Gaussian random variables with variance proxies $\sigma_1^2, \ldots, \sigma_n^2$. Then, for the sum $Z = \sum_{i=1}^n X_i$, we have the following tail bound: $\mathbb{P}\Big(|Z - \mathbb{E}[Z]| \geq t\Big) \leq 2\exp\left(-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2}\right)$ *,* $\forall t > 0.$

Proof:

- Step 1: Moment-Generating Function of Sub-Gaussian Variables: Each X_i is sub-Gaussian, meaning that there exists σ_i^2 such that: $\mathbb{E}\left[e^{\lambda(X_i - \mathbb{E}[X_i])}\right] \leq e^{\frac{\lambda^2\sigma_i^2}{2}}, \quad \forall \lambda \in \mathbb{R}.$
  This inequality characterizes the concentration properties of each X_i and allows us to control the tail behavior of $Z = \sum_{i=1}^n X_i$.

Theorem (Hoeffding's Inequality for Sub-Gaussian Sums)

Let $X_1, \ldots, X_n$ be independent sub-Gaussian random variables with variance proxies $\sigma_1^2, \ldots, \sigma_n^2$. Then, for the sum $Z = \sum_{i=1}^n X_i$, we have the following tail bound: $\mathbb{P}\Big(|Z - \mathbb{E}[Z]| \geq t\Big) \leq 2\exp\left(-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2}\right)$ *,*
$\forall t > 0.$

Proof:

- Step 1: Moment-Generating Function of Sub-Gaussian Variables:
  Each X_i is sub-Gaussian, meaning that there exists σ_i^2 such that:
  $\mathbb{E}\left[e^{\lambda(X_i - \mathbb{E}[X_i])}\right] \leq e^{\frac{\lambda^2\sigma_i^2}{2}}, \quad \forall \lambda \in \mathbb{R}.$
  This inequality characterizes the concentration properties of each X_i and allows us to control the tail behavior of $Z = \sum_{i=1}^n X_i$.
- Step 2: Moment-Generating Function of the Sum Z:
  Since $Z = \sum_{i=1}^n X_i$, and by independence of X_i, we have:

2.3 Hoeffding's Inequality for Sub-Gaussian Random Variables

$$\mathbb{E}\left[e^{\lambda(Z-\mathbb{E}[Z])}\right] = \mathbb{E}\left[e^{\lambda\sum_{i=1}^{n}(X_i-\mathbb{E}[X_i])}\right] = \prod_{i=1}^{n}\mathbb{E}\left[e^{\lambda(X_i-\mathbb{E}[X_i])}\right].$$

2.3 Hoeffding's Inequality for Sub-Gaussian Random Variables

$\mathbb{E}\left[e^{\lambda(Z-\mathbb{E}[Z])}\right] = \mathbb{E}\left[e^{\lambda\sum_{i=1}^{n}(X_i-\mathbb{E}[X_i])}\right] = \prod_{i=1}^{n}\mathbb{E}\left[e^{\lambda(X_i-\mathbb{E}[X_i])}\right].$

By substituting the sub-Gaussian bound from step 1, we get:

$\mathbb{E}\left[e^{\lambda(Z-\mathbb{E}[Z])}\right] \leq \prod_{i=1}^{n} e^{\frac{\lambda^2\sigma_i^2}{2}} = e^{\frac{\lambda^2}{2}\sum_{i=1}^{n}\sigma_i^2}.$

$\mathbb{E}\left[e^{\lambda(Z-\mathbb{E}[Z])}\right] = \mathbb{E}\left[e^{\lambda \sum_{i=1}^{n}(X_i-\mathbb{E}[X_i])}\right] = \prod_{i=1}^{n} \mathbb{E}\left[e^{\lambda(X_i-\mathbb{E}[X_i])}\right].$

By substituting the sub-Gaussian bound from step 1, we get:

$\mathbb{E}\left[e^{\lambda(Z-\mathbb{E}[Z])}\right] \leq \prod_{i=1}^{n} e^{\frac{\lambda^2\sigma_i^2}{2}} = e^{\frac{\lambda^2}{2}\sum_{i=1}^{n}\sigma_i^2}.$

- Step 3: Applying Chernoff's Bound:
  To bound $\mathbb{P}(Z-\mathbb{E}[Z] \geq t)$, we use Markov's inequality:

$$\mathbb{P}(Z-\mathbb{E}[Z] \geq t) = \mathbb{P}\left(e^{\lambda(Z-\mathbb{E}[Z])} \geq e^{\lambda t}\right) \leq \frac{\mathbb{E}\left[e^{\lambda(Z-\mathbb{E}[Z])}\right]}{e^{\lambda t}}.$$

  By substituting the result from step 2, we get:

$$\mathbb{P}(Z-\mathbb{E}[Z] \geq t) \leq \frac{e^{\frac{\lambda^2}{2}\sum_{i=1}^{n}\sigma_i^2}}{e^{\lambda t}} = e^{\frac{\lambda^2}{2}\sum_{i=1}^{n}\sigma_i^2-\lambda t}.$$

$\mathbb{E}\left[e^{\lambda(Z-\mathbb{E}[Z])}\right] = \mathbb{E}\left[e^{\lambda\sum_{i=1}^{n}(X_i-\mathbb{E}[X_i])}\right] = \prod_{i=1}^{n}\mathbb{E}\left[e^{\lambda(X_i-\mathbb{E}[X_i])}\right].$

By substituting the sub-Gaussian bound from step 1, we get:

$\mathbb{E}\left[e^{\lambda(Z-\mathbb{E}[Z])}\right] \leq \prod_{i=1}^{n} e^{\frac{\lambda^2\sigma_i^2}{2}} = e^{\frac{\lambda^2}{2}\sum_{i=1}^{n}\sigma_i^2}.$

- Step 3: Applying Chernoff's Bound:
  To bound $\mathbb{P}(Z-\mathbb{E}[Z]\geq t)$, we use Markov's inequality:

$$\mathbb{P}(Z-\mathbb{E}[Z]\geq t) = \mathbb{P}\left(e^{\lambda(Z-\mathbb{E}[Z])}\geq e^{\lambda t}\right) \leq \frac{\mathbb{E}\left[e^{\lambda(Z-\mathbb{E}[Z])}\right]}{e^{\lambda t}}.$$

  By substituting the result from step 2, we get:

$$\mathbb{P}(Z-\mathbb{E}[Z]\geq t) \leq \frac{e^{\frac{\lambda^2}{2}\sum_{i=1}^{n}\sigma_i^2}}{e^{\lambda t}} = e^{\frac{\lambda^2}{2}\sum_{i=1}^{n}\sigma_i^2-\lambda t}.$$

- Step 4: Optimizing over λ:
  To get the tightest bound, we minimize the exponent by choosing λ that minimizes the power of the expo function, i.e, $\frac{\lambda^2}{2}\sum_{i=1}^{n}\sigma_i^2-\lambda t$.

2.3 Hoeffding's Inequality for Sub-Gaussian Random Variables

That is, we solve for λ in $\frac{d}{d\lambda}\left[\frac{\lambda^2}{2}\sum_{i=1}^n \sigma_i^2 - \lambda t\right] = \lambda \sum_{i=1}^n \sigma_i^2 - t$=0.

This gives $\lambda = \frac{t}{\sum_{i=1}^n \sigma_i^2}$. Substituting this value of λ back, we obtain:

$$\mathbb{P}(Z - \mathbb{E}[Z] \geq t) \leq \exp\left(-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2}\right).$$

2.3 Hoeffding's Inequality for Sub-Gaussian Random Variables

That is, we solve for λ in $\frac{d}{d\lambda}\left[\frac{\lambda^2}{2}\sum_{i=1}^n \sigma_i^2 - \lambda t\right] = \lambda \sum_{i=1}^n \sigma_i^2 - t = 0$.

This gives $\lambda = \frac{t}{\sum_{i=1}^n \sigma_i^2}$. Substituting this value of λ back, we obtain:

$$\mathbb{P}(Z - \mathbb{E}[Z] \geq t) \leq \exp\left(-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2}\right).$$

- Step 5: Bounding the Lower Tail:
  A similar argument shows that:
  $$\mathbb{P}(Z - \mathbb{E}[Z] \leq -t) \leq \exp\left(-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2}\right).$$

2.3 Hoeffding's Inequality for Sub-Gaussian Random Variables

That is, we solve for λ in $\frac{d}{d\lambda}\left[\frac{\lambda^2}{2}\sum_{i=1}^n \sigma_i^2 - \lambda t\right] = \lambda \sum_{i=1}^n \sigma_i^2 - t$=0.

This gives $\lambda = \frac{t}{\sum_{i=1}^n \sigma_i^2}$. Substituting this value of λ back, we obtain:

$$\mathbb{P}(Z - \mathbb{E}[Z] \geq t) \leq \exp\left(-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2}\right).$$

- Step 5: Bounding the Lower Tail:
  A similar argument shows that:
  $$\mathbb{P}(Z - \mathbb{E}[Z] \leq -t) \leq \exp\left(-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2}\right).$$
- Step 6: Combining Upper and Lower Tail Bounds:
  By the union bound, i.e.,
  $\mathbb{P}(|Z - \mathbb{E}[Z]| \geq t) \leq \mathbb{P}(Z - \mathbb{E}[Z] \leq -t) + \mathbb{P}(Z - \mathbb{E}[Z] \geq t)$, we conclude:
  $\mathbb{P}(|Z - \mathbb{E}[Z]| \geq t) \leq 2\exp\left(-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2}\right)$. End of proof. □

That is, we solve for λ in $\frac{d}{d\lambda}\left[\frac{\lambda^2}{2}\sum_{i=1}^n \sigma_i^2 - \lambda t\right] = \lambda\sum_{i=1}^n \sigma_i^2 - t$=0.

This gives $\lambda = \frac{t}{\sum_{i=1}^n \sigma_i^2}$. Substituting this value of λ back, we obtain:

$$\mathbb{P}(Z - \mathbb{E}[Z] \geq t) \leq \exp\left(-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2}\right).$$

- Step 5: Bounding the Lower Tail:
  A similar argument shows that:
  $$\mathbb{P}(Z - \mathbb{E}[Z] \leq -t) \leq \exp\left(-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2}\right).$$
- Step 6: Combining Upper and Lower Tail Bounds:
  By the union bound, i.e.,
  $\mathbb{P}(|Z - \mathbb{E}[Z]| \geq t) \leq \mathbb{P}(Z - \mathbb{E}[Z] \leq -t) + \mathbb{P}(Z - \mathbb{E}[Z] \geq t)$, we conclude:
  $\mathbb{P}(|Z - \mathbb{E}[Z]| \geq t) \leq 2\exp\left(-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2}\right)$. End of proof. $\square$

Please note that Hoeffding's inequality is a two-sided concentration inequality—it provides control over both the probability that the sum deviates above its expectation and the probability that it deviates below it.

2.4 Chernoff's Inequality

Lemma (Chernoff's Inequality)

Let X_i be independent Bernoulli random variables with parameters p_i. Define the sum $S_N = \sum_{i=1}^{N} X_i$ and let $\mu = \mathbb{E}[S_N]$ be its mean. Then, for any $t > \mu$, we have: $\mathbb{P}(S_N \geq t) \leq \exp(-\mu) \left(\frac{\exp(1)\mu}{t} \right)^t$.

Proof:

Step 1: To bound $\mathbb{P}(S_N \geq t)$, we use Markov's inequality on an exponential function of S_N. For any $\lambda > 0$, we have:

$$\mathbb{P}(S_N \geq t) = (e^{\lambda S_N} \geq e^{\lambda t}) \leq \frac{\mathbb{E}[e^{\lambda S_N}]}{e^{\lambda t}}.$$

2.4 Chernoff's Inequality

Lemma (Chernoff's Inequality)

Let X_i be independent Bernoulli random variables with parameters p_i. Define the sum $S_N = \sum_{i=1}^{N} X_i$ and let $\mu = \mathbb{E}[S_N]$ be its mean. Then, for any $t > \mu$, we have: $\mathbb{P}(S_N \geq t) \leq \exp(-\mu) \left(\frac{\exp(1)\mu}{t}\right)^t$.

Proof:

Step 1: To bound $\mathbb{P}(S_N \geq t)$, we use Markov's inequality on an exponential function of S_N. For any $\lambda > 0$, we have:

$$\mathbb{P}(S_N \geq t) = (e^{\lambda S_N} \geq e^{\lambda t}) \leq \frac{\mathbb{E}[e^{\lambda S_N}]}{e^{\lambda t}}.$$

Step 2: We calcuate the Moment Generating Function $\mathbb{E}[e^{\lambda S_N}]$. Since X_i are independent, we can write:

$$\mathbb{E}[e^{\lambda S_N}] = \mathbb{E}\left[e^{\lambda \sum_{i=1}^{N} X_i}\right] = \prod_{i=1}^{N} \mathbb{E}[e^{\lambda X_i}].$$

2.4 Chernoff's Inequality

Step 3: We evaluate $\mathbb{E}[e^{\lambda X_i}]$ for a Bernoulli Random Variable:
For a Bernoulli random variable X_i with success probability p_i, we have:

$$\mathbb{E}[e^{\lambda X_i}] = p_i e^{\lambda} + (1 - p_i) = 1 + p_i(e^{\lambda} - 1).$$

Thus, $\mathbb{E}[e^{\lambda S_N}] = \prod_{i=1}^{N} \left(1 + p_i(e^{\lambda} - 1)\right)$.

2.4 Chernoff's Inequality

Step 3: We evaluate $\mathbb{E}[e^{\lambda X_i}]$ for a Bernoulli Random Variable:
For a Bernoulli random variable X_i with success probability p_i, we have:

$$\mathbb{E}[e^{\lambda X_i}] = p_i e^{\lambda} + (1 - p_i) = 1 + p_i(e^{\lambda} - 1).$$

Thus, $\mathbb{E}[e^{\lambda S_N}] = \prod_{i=1}^{N} \left(1 + p_i(e^{\lambda} - 1)\right)$.

Step 4: We simplify using the bound $1 + x \leq e^x$:
Applying the inequality $1 + x \leq e^x$, we get:

$$\mathbb{E}[e^{\lambda S_N}] \leq \prod_{i=1}^{N} e^{p_i(e^{\lambda}-1)} = e^{(e^{\lambda}-1)\sum_{i=1}^{N} p_i} = e^{(e^{\lambda}-1)\mu}.$$

2.4 Chernoff's Inequality

Step 3: We evaluate $\mathbb{E}[e^{\lambda X_i}]$ for a Bernoulli Random Variable:
For a Bernoulli random variable X_i with success probability p_i, we have:

$$\mathbb{E}[e^{\lambda X_i}] = p_i e^{\lambda} + (1 - p_i) = 1 + p_i(e^{\lambda} - 1).$$

Thus, $\mathbb{E}[e^{\lambda S_N}] = \prod_{i=1}^{N} \left(1 + p_i(e^{\lambda} - 1)\right)$.

Step 4: We simplify using the bound $1 + x \leq e^x$:
Applying the inequality $1 + x \leq e^x$, we get:

$$\mathbb{E}[e^{\lambda S_N}] \leq \prod_{i=1}^{N} e^{p_i(e^{\lambda}-1)} = e^{(e^{\lambda}-1)\sum_{i=1}^{N} p_i} = e^{(e^{\lambda}-1)\mu}.$$

Step 5: We combine results by substituting back into the probability bound, and we get:

$$\mathbb{P}(S_N \geq t) \leq \frac{\mathbb{E}[e^{\lambda S_N}]}{e^{\lambda t}} \leq \frac{e^{(e^{\lambda}-1)\mu}}{e^{\lambda t}} = e^{(e^{\lambda}-1)\mu - \lambda t}.$$

2.4 Chernoff's Inequality

Step 6: We optimize by minimizing the exponent:
To make this bound as tight as possible, choose λ to minimize $(e^{\lambda} - 1)\mu - \lambda t$. This leads to solving: $\frac{d}{d\lambda}\left((e^{\lambda} - 1)\mu - \lambda t\right) = 0$.
This gives us $\lambda = \ln\left(\frac{t}{\mu}\right)$.

2.4 Chernoff's Inequality

Step 6: We optimize by minimizing the exponent:
To make this bound as tight as possible, choose λ to minimize $(e^{\lambda}-1)\mu-\lambda t$. This leads to solving: $\frac{d}{d\lambda}\big((e^{\lambda}-1)\mu-\lambda t\big)=0$.
This gives us $\lambda=\ln\left(\frac{t}{\mu}\right)$.
Step 7: We substitute this value of λ back into the bound to obtain:

$$\mathbb{P}(S_N \geq t) \leq \exp\left(\mu\left(\frac{t}{\mu}-1-\ln\left(\frac{t}{\mu}\right)\right)\right).$$

2.4 Chernoff's Inequality

Step 6: We optimize by minimizing the exponent:
To make this bound as tight as possible, choose λ to minimize $(e^\lambda - 1)\mu - \lambda t$. This leads to solving: $\frac{d}{d\lambda}\left((e^\lambda - 1)\mu - \lambda t\right) = 0$.
This gives us $\lambda = \ln\left(\frac{t}{\mu}\right)$.
Step 7: We substitute this value of λ back into the bound to obtain:

$$\mathbb{P}(S_N \geq t) \leq \exp\left(\mu\left(\frac{t}{\mu} - 1 - \ln\left(\frac{t}{\mu}\right)\right)\right).$$

Step 8: Finally, we simplify by using the inequality $x - 1 - \ln x \leq x \ln\left(\frac{e}{x}\right)$ for $x = \frac{t}{\mu}$, we get:

$$\mathbb{P}(S_N \geq t) \leq \exp(-\mu)\left(\frac{e\mu}{t}\right)^t,$$

which completes the proof. □

3. Equivalent Characterizations of Sub-Gaussianity

For a random variable X with $\mathbb{E}[X] = 0$ and constants $K_1, K_2, K_3, K_4 > 0$, the following conditions are equivalent, characterizing X as sub-Gaussian:

1. **Tail Bound:** The probability of large deviations is bounded by

$$\mathbb{P}(|X| > t) \leq 2\exp\left(-\frac{t^2}{K_1^2}\right), \quad \forall\, t \geq 0.$$

3. Equivalent Characterizations of Sub-Gaussianity

For a random variable X with $\mathbb{E}[X] = 0$ and constants $K_1, K_2, K_3, K_4 > 0$, the following conditions are equivalent, characterizing X as sub-Gaussian:

1. **Tail Bound:** The probability of large deviations is bounded by

$$\mathbb{P}(|X| > t) \leq 2 \exp\left(-\frac{t^2}{K_1^2}\right), \quad \forall\, t \geq 0.$$

2. **Moment Growth:** The p-th moments of X grow no faster than $\sqrt{p}$, specifically $\left(\mathbb{E}[|X|^p]\right)^{\frac{1}{p}} \leq K_2\sqrt{p}, \quad \forall\, p \geq 1.$

3. Equivalent Characterizations of Sub-Gaussianity

For a random variable X with $\mathbb{E}[X] = 0$ and constants $K_1, K_2, K_3, K_4 > 0$, the following conditions are equivalent, characterizing X as sub-Gaussian:

1. **Tail Bound:** The probability of large deviations is bounded by
$$\mathbb{P}(|X| > t) \leq 2 \exp\left(-\frac{t^2}{K_1^2}\right), \quad \forall\, t \geq 0.$$
2. **Moment Growth:** The p-th moments of X grow no faster than $\sqrt{p}$, specifically $\left(\mathbb{E}[|X|^p]\right)^{\frac{1}{p}} \leq K_2\sqrt{p}, \quad \forall\, p \geq 1.$
3. **MGF of X^2:** For small values of λ, the MGF of X^2 satisfies
$$\mathbb{E}[\exp(\lambda^2 X^2)] \leq \exp(K_3^2 \lambda^2), \quad \forall\, |\lambda| \leq \frac{1}{K_3}.$$

3. Equivalent Characterizations of Sub-Gaussianity

For a random variable X with $\mathbb{E}[X] = 0$ and constants $K_1, K_2, K_3, K_4 > 0$, the following conditions are equivalent, characterizing X as sub-Gaussian:

1. **Tail Bound:** The probability of large deviations is bounded by

$$\mathbb{P}(|X| > t) \leq 2 \exp\left(-\frac{t^2}{K_1^2}\right), \quad \forall\, t \geq 0.$$

2. **Moment Growth:** The p-th moments of X grow no faster than $\sqrt{p}$, specifically $(\mathbb{E}[|X|^p])^{\frac{1}{p}} \leq K_2 \sqrt{p}, \quad \forall\, p \geq 1.$
3. **MGF of X^2:** For small values of λ, the MGF of X^2 satisfies

$$\mathbb{E}[\exp(\lambda^2 X^2)] \leq \exp(K_3^2 \lambda^2), \quad \forall\, |\lambda| \leq \frac{1}{K_3}.$$

4. **Bounded MGF of X^2 at specific Point:** The MGF of X^2 is bounded at a specific value: $\mathbb{E}\left[\exp\left(\frac{X^2}{K_4^2}\right)\right] \leq 2.$

3. Equivalent Characterizations of Sub-Gaussianity

Please Note:

- The phrase *"the following conditions are equivalent"* in the previous slide means that any of the listed conditions (1 to 4) can be used to determine if X is sub-Gaussian.

3. Equivalent Characterizations of Sub-Gaussianity

Please Note:

- The phrase *"the following conditions are equivalent"* in the previous slide means that any of the listed conditions (1 to 4) can be used to determine if X is sub-Gaussian.
- If one of these conditions holds for a given random variable X, then the others also hold. Thus, we can characterize or define a sub-Gaussian variable by checking any one of these properties.

3. Equivalent Characterizations of Sub-Gaussianity

Please Note:

- The phrase *"the following conditions are equivalent"* in the previous slide means that any of the listed conditions (1 to 4) can be used to determine if X is sub-Gaussian.
- If one of these conditions holds for a given random variable X, then the others also hold. Thus, we can characterize or define a sub-Gaussian variable by checking any one of these properties.
- In essence, these conditions provide different but equivalent ways of defining when a random variable has sub-Gaussian characteristics, which implies it doesn't have heavy tails and is somewhat concentrated around its mean.

4. Sub-exponential Concentration

Definition

A zero-mean random variable X is called **sub-exponential** if there exist constants $\nu, \alpha > 0$ such that

$$\mathbb{E}\left[e^{\lambda X}\right] \leq \exp\left(\frac{\nu^2 \lambda^2}{2}\right), \quad \text{for all } |\lambda| < \frac{1}{\alpha}.$$

A general random variable X is sub-exponential if $X - \mathbb{E}[X]$ is sub-exponential.

- **Example:** If $Z \sim \mathcal{N}(0, 1)$, then Z^2 is sub-exponential with parameters $(2, 4)$.

4. Sub-exponential Concentration

Definition

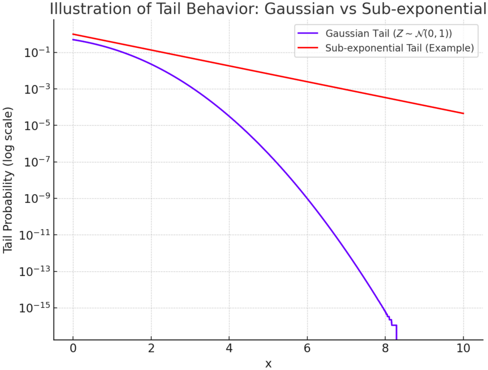
A zero-mean random variable X is called **sub-exponential** if there exist constants $\nu, \alpha > 0$ such that

$$\mathbb{E}\left[e^{\lambda X}\right] \leq \exp\left(\frac{\nu^2\lambda^2}{2}\right), \quad \text{for all } |\lambda| < \frac{1}{\alpha}.$$

A general random variable X is sub-exponential if $X - \mathbb{E}[X]$ is sub-exponential.

- **Example:** If $Z \sim \mathcal{N}(0,1)$, then Z^2 is sub-exponential with parameters $(2,4)$.
- **Tail Behavior:** The tails of $Z^2 - 1$ are heavier than those of a Gaussian distribution, indicating a higher likelihood of larger deviations.

4. Sub-exponential Concentration: Illustration of tail behavior



Graph illustrating the tail behavior of a Gaussian distribution $Z \sim \mathcal{N}(0,1)$ compared to a sub-exponential distribution (e.g., exponential distribution with a heavier tail). The logarithmic scale on the y-axis highlights the difference in tail decay, showing that the sub-exponential tail decreases more slowly, indicating a higher probability of larger deviations.

4. Sub-exponential Concentration Inequality

Proposition (Concentration Bound for Sub-exponential Variables)

Let X be a zero-mean sub-exponential random variable with parameters (ν, α). Then, for any $t > 0$,

$$\mathbb{P}(X > t) \leq \exp\left(-\frac{1}{2}\min\left\{\frac{t^2}{\nu^2}, \frac{t}{\alpha}\right\}\right).$$

4. Sub-exponential Concentration Inequality

Proposition (Concentration Bound for Sub-exponential Variables)

Let X be a zero-mean sub-exponential random variable with parameters (ν, α). Then, for any $t > 0$,

$$\mathbb{P}(X > t) \leq \exp\left(-\frac{1}{2}\min\left\{\frac{t^2}{\nu^2}, \frac{t}{\alpha}\right\}\right).$$

Key Insights:

- This inequality provides a bound on the tail probability for sub-exponential random variables.
- The form of the bound captures both the quadratic decay (when t is small) and linear decay (for large t), reflecting the "sub-exponential" nature of X.

4.1 Maximum of Sub-Gaussian Vectors

Let $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ be a random vector. The $\max$-norm of X is defined as: $\|X\|_{\max} = \max_{i=1,\ldots,n} |X_i|$.

Lemma

Suppose $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ is a random vector with zero-mean, sub-Gaussian coordinates X_i, each having sub-Gaussian parameter $\sigma_i > 0$. Then, for any $\gamma \geq 0$, we have:

$$\mathbb{P}\left(\|X\|_{\max} > \sigma\sqrt{2(1+\gamma)\log(n)}\right) \leq 2n^{-\gamma},$$

where $\sigma = \max_{i=1,\ldots,n} \sigma_i$.

Proof Outline: This bound is derived by applying concentration inequalities for sub-Gaussian random variables to control the maximum of X_i across dimensions.

4.1 Maximum of Sub-Gaussian Vectors

Proof details step-by-step:

- **Step 1. Sub-Gaussian Tail Bound**: Since each coordinate X_i is sub-Gaussian with parameter σ_i, it follows that
$$\mathbb{P}(|X_i| > t) \leq 2\exp\left(-\frac{t^2}{2\sigma_i^2}\right).$$
This individually provides a tail bound for each X_i.

4.1 Maximum of Sub-Gaussian Vectors

Proof details step-by-step:

- **Step 1. Sub-Gaussian Tail Bound**: Since each coordinate X_i is sub-Gaussian with parameter σ_i, it follows that
$$\mathbb{P}(|X_i| > t) \leq 2\exp\left(-\frac{t^2}{2\sigma_i^2}\right).$$
This individually provides a tail bound for each X_i.

- **Step 2. Union Bound**: To bound $\mathbb{P}(\|X\|_{\max} > t)$, we apply the union bound over all n coordinates:
$$\mathbb{P}\left(\|X\|_{\max} > t\right) = \mathbb{P}\left(\max_{i=1,\dots,n} |X_i| > t\right) \leq \sum_{i=1}^{n} \mathbb{P}(|X_i| > t).$$

4.1 Maximum of Sub-Gaussian Vectors

Proof details step-by-step:

- **Step 1. Sub-Gaussian Tail Bound**: Since each coordinate X_i is sub-Gaussian with parameter σ_i, it follows that
$$\mathbb{P}(|X_i| > t) \leq 2\exp\left(-\frac{t^2}{2\sigma_i^2}\right).$$
This individually provides a tail bound for each X_i.
- **Step 2. Union Bound**: To bound $\mathbb{P}(\|X\|_{\max} > t)$, we apply the union bound over all n coordinates:
$$\mathbb{P}\left(\|X\|_{\max} > t\right) = \mathbb{P}\left(\max_{i=1,\ldots,n} |X_i| > t\right) \leq \sum_{i=1}^{n} \mathbb{P}(|X_i| > t).$$
- **Step 3. Applying the Sub-Gaussian Bound**: Substituting the sub-Gaussian tail bound, we get:
$$\mathbb{P}\left(\|X\|_{\max} > t\right) \leq \sum_{i=1}^{n} 2\exp\left(-\frac{t^2}{2\sigma_i^2}\right).$$

4.1 Maximum of Sub-Gaussian Vectors

Since $\sigma = \max_{i=1,\ldots,n} \sigma_i$, we can use σ as an upper bound for each σ_i, leading to: $\mathbb{P}\left(\|X\|_{\max} > t\right) \leq 2n \exp\left(-\frac{t^2}{2\sigma^2}\right)$.

4.1 Maximum of Sub-Gaussian Vectors

Since $\sigma = \max_{i=1,\ldots,n} \sigma_i$, we can use σ as an upper bound for each σ_i, leading to: $\mathbb{P}\left(\|X\|_{\max} > t\right) \leq 2n \exp\left(-\frac{t^2}{2\sigma^2}\right)$.

- **Step 4. Choosing t for Desired Probability**: We want to bound $\mathbb{P}(\|X\|_{\max} > t)$ by $2n^{-\gamma}$. Setting $t = \sigma\sqrt{2(1+\gamma)\log(n)}$, we substitute into the bound:
  $$\mathbb{P}\left(\|X\|_{\max} > \sigma\sqrt{2(1+\gamma)\log(n)}\right) \leq 2n \exp\left(-\frac{(\sigma\sqrt{2(1+\gamma)\log(n)})^2}{2\sigma^2}\right).$$

4.1 Maximum of Sub-Gaussian Vectors

Since $\sigma = \max_{i=1,\ldots,n} \sigma_i$, we can use σ as an upper bound for each σ_i, leading to: $\mathbb{P}\left(\|X\|_{\max} > t\right) \leq 2n \exp\left(-\frac{t^2}{2\sigma^2}\right)$.

- **Step 4. Choosing t for Desired Probability**: We want to bound $\mathbb{P}(\|X\|_{\max} > t)$ by $2n^{-\gamma}$. Setting $t = \sigma\sqrt{2(1+\gamma)\log(n)}$, we substitute into the bound:
  $$\mathbb{P}\left(\|X\|_{\max} > \sigma\sqrt{2(1+\gamma)\log(n)}\right) \leq 2n \exp\left(-\frac{(\sigma\sqrt{2(1+\gamma)\log(n)})^2}{2\sigma^2}\right).$$
- **Step 5. Simplifying the Exponent**:
  $$2n \exp\left(-\frac{2(1+\gamma)\log(n)}{2}\right) = 2n \exp(-(1+\gamma)\log(n)).$$
  Using $\exp(-(1+\gamma)\log(n)) = n^{-(1+\gamma)}$, we have:
  $$\mathbb{P}\left(\|X\|_{\max} > \sigma\sqrt{2(1+\gamma)\log(n)}\right) \leq 2n \cdot n^{-(1+\gamma)} = 2n^{-\gamma}.$$

Thus, we conclude that $\mathbb{P}\left(\|X\|_{\max} > \sigma\sqrt{2(1+\gamma)\log(n)}\right) \leq 2n^{-\gamma}$. $\square$

4.2 Lipschitz Functions of a Standard Gaussian Vector

Theorem

Let $f : \mathbb{R}^n \to \mathbb{R}$ be an L-Lipschitz function with respect to the Euclidean distance, and let $X = (X_1, \ldots, X_n)$ where $X_i \sim \mathcal{N}(0,1)$ are i.i.d. standard Gaussian random variables. Then, for any $t \in \mathbb{R}$,

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

In particular, $f(X)$ is sub-Gaussian with parameter L.

- **Implication:** $f(X)$ is sub-Gaussian, tightly concentrated around $\mathbb{E}[f(X)]$.
- **Remark:** The proof of this result is non-trivial and uses advanced techniques. This is a deep result with substantial applications in high-dimensional probability and statistics.
- **Technical Note:** One-sided tail bounds (e.g., $\mathbb{P}(f(X) - \mathbb{E}[f(X)] \geq t)$) remove the factor of 2.

5. Summary

What is concentration?

- non-asymptotic bound on probability to control deviations
- possible deviations of interest: RV from mean
- deviation between the estimator and true quantity

Important definitions:

- tails of a sub-Gaussian distribution are dominated by the tails of a Gaussian
- distributions with heavy tails are not sub-Gaussian
- tails of distributions with heavy tails might be sub-exponential

5. Summary

Main technique: Let Z be a zero-mean random variable. Then, for $t \in \mathbb{R}$,

$$\mathbb{P}(Z > t) = \mathbb{P}(\exp(\lambda Z) > \exp(\lambda t)) \overset{Markov's}{\leq} \frac{e^{\lambda Z}}{e^{\lambda t}}.$$

- Apply $\exp(\lambda \cdot)$ to both sides.
- Apply Markov's inequality.
- Calculate MGF.
- Minimize over λ.