Basic Tail and Concentration Bounds

Yu Zhang

Department of Statistics and Finance, USTC

2022/9/5

1. From Markov to Chernof
2. Sub-Gaussian Variables and Hoeffding Bounds
3. Sub-exponential Variables and Bernstein Bounds
4. Introduction of Martingale-based Methods
5. Concentration Bounds for Martingale Difference Sequences
6. Lipschitz Functions of Gaussian Variables

1. From Markov to Chernof
2. Sub-Gaussian Variables and Hoeffding Bounds
3. Sub-exponential Variables and Bernstein Bounds
4. Introduction of Martingale-based Methods
5. Concentration Bounds for Martingale Difference Sequences
6. Lipschitz Functions of Gaussian Variables

Markov's inequality

- Markov's inequality (X: non-negative and a finite mean):

$$P(X \geq t) \leq \frac{E[X]}{t}, \forall t > 0. \tag{2.1}$$

- Chebyshev's inequality ($Y = (X - \mu)^2$):

$$P(|X - \mu| \geq t) \leq \frac{var(X)}{t^2}, \forall t > 0. \tag{2.2}$$

- Extensions of Markov's inequality (X has a central moment of order k):

$$P(|X - \mu| \geq t) \leq \frac{E|X - \mu|^k}{t^k}, \forall t > 0. \tag{2.3}$$

Chernoff bound

Suppose that the random variable X has a moment generating function in a neighborhood of zero, meaning that there is some constant $b > 0$ such that the function $\phi(\lambda) = E[e^{\lambda(X-\mu)}]$ exists for all $\lambda \leq |b|$. In this case, for any $\lambda \in [0, b]$, then apply Markov's inequality:

$$P((X-\mu) \geq t) = P(e^{\lambda(X-\mu)} \geq e^{\lambda t}) \leq \frac{E[e^{\lambda(X-\mu)}]}{e^{\lambda t}}. \tag{2.4}$$

Optimizing our choice of λ so as to obtain the tightest result yields the Chernoff bound:

$$logP(X-\mu \geq t) \leq inf_{\lambda \in [0,b]}(logE[e^{\lambda(X-\mu)}] - \lambda t). \tag{2.5}$$

1 From Markov to Chernof

2 Sub-Gaussian Variables and Hoeffding Bounds

3 Sub-exponential Variables and Bernstein Bounds

4 Introduction of Martingale-based Methods

5 Concentration Bounds for Martingale Difference Sequences

6 Lipschitz Functions of Gaussian Variables

Gaussian tail bounds

Example 2.1: Gaussian tail bounds

Let $X \sim N(\mu, \sigma^2)$,then we find that X has the moment generating function

$$E[e^{\lambda X}] = e^{\mu\lambda + \frac{\sigma^2\lambda^2}{2}}, \forall \lambda \in R. \tag{2.6}$$

Substituting this expression into the optimization problem defining the optimized Chernoff bound (2.5), we obtain

$$\inf_{\lambda \geq 0}(log E[e^{\lambda(X-\mu)}] - \lambda t) = \inf_{\lambda \geq 0}(\frac{\lambda^2\sigma^2}{2} - \lambda t) = -\frac{t^2}{2\sigma^2}$$

We conclude that any $N(\mu, \sigma^2)$ random variable satisfies the upper deviation inequality

$$P(X \geq \mu + t) \leq e^{-\frac{t^2}{2\sigma^2}}, \forall t > 0. \tag{2.7}$$

Sub-Gaussian

Definition 2.2 (Sub-Gaussian) A random variable X with mean μ is sub-Gaussian if there is a positive number σ such that

$$E[e^{\lambda(X-\mu)}] \leq e^{\frac{\sigma^2\lambda^2}{2}}, \forall \lambda \in R. \tag{2.8}$$

The constant σ is reffered to as the sub-Gaussian parameter.

Note: Any Gaussian variable with variance σ^2 is sub-Gaussian with parameter σ.

Sub-Gaussian

If r.v. X is sub-Gaussian with parameter σ ,then it satisfies the upper deviation inequality (2.7). By the symmetry, $-X$ also is sub-Gaussian with parameter σ, so that we have the lower deviation inequality $P[X \leq \mu - t] \leq e^{-\frac{t^2}{2\sigma^2}} \forall t \geq 0$. Thus the sub-Gaussian variable satisfies the concentration inequality

$$P(|X - \mu| \geq t) \leq 2e^{-\frac{t^2}{2\sigma^2}}, t \in R. \tag{2.9}$$

Examples

Example 2.3: Rademacher variables

A Rademacher random variable ε takes the values $\{-1, +1\}$ equiprobably. By taking expectations and using the power-series expansion for the exponential, we obtain

$$\mathbb{E}\left[e^{\lambda\varepsilon}\right] = \frac{1}{2}\left\{e^{-\lambda} + e^{\lambda}\right\} = \frac{1}{2}\left\{\sum_{k=0}^{\infty}\frac{(-\lambda)^k}{k!} + \sum_{k=0}^{\infty}\frac{(\lambda)^k}{k!}\right\}$$

$$= \sum_{k=0}^{\infty}\frac{\lambda^{2k}}{(2k)!} \leq 1 + \sum_{k=1}^{\infty}\frac{\lambda^{2k}}{2^k k!} = e^{\lambda^2/2}$$

which shows that ε is sub-Gaussian with parameter $\sigma = 1$.

Examples

Example 2.4: Bounded random variables

Let X be zero-mean, and supported on some interval $[a, b]$. Letting X' be an independent copy, for any $\lambda \in \mathbb{R}$, we have

$$\mathbb{E}_X\left[e^{\lambda X}\right] = \mathbb{E}_X\left[e^{\lambda(x-\mathbb{E}_{X'}[x'])}\right] \leq \mathbb{E}_{X,X'}\left[e^{\lambda(X-X')}\right],$$

Letting ε be an independent Rademacher variable, note that the distribution of $(X - X')$ is the same as that of $\varepsilon(X - X')$:

$$P(\varepsilon(X - X') \leq t) = \frac{1}{2}P((X - X') \leq t) + \frac{1}{2}P((X' - X) \leq t) \quad = P((X - X$$

So that we have

$$\mathbb{E}_{X,X'}\left[e^{\lambda(x-X')}\right] = \mathbb{E}_{X,X'}\left[\mathbb{E}_{\varepsilon}\left[e^{\lambda\varepsilon(X-X')}\right]\right] \overset{\text{(i)}}{\leq} \mathbb{E}_{X,X'}\left[e^{\frac{\lambda^2(x-x')^2}{2}}\right],$$

Examples

Since$|X - X'| \leq b - a$,we are guaranteed that

$$\mathbb{E}_{X,X'}\left[e^{\frac{\lambda^2(x-x')^2}{2}}\right] \leq e^{\frac{\lambda^2(b-a)^2}{2}}$$

Thus X is sub-Gaussian with parameter at most $\sigma = b - a$.
In Exercise 2.4, we work through a more involved argument to show that X is sub-Gaussian with parameter at most $\sigma = \frac{b-a}{2}$.

Hoeffding bound

The property of sub-Gaussianity is preserved by linear operations. If an independent sequence $\{X_k\}_{k=1}^n$ of random variables, such that X_k has mean μ_k, and is sub-Gaussian with parameters σ_k.

$$\mathbb{E}\left[e^{\lambda \sum_{k=1}^n (X_k - \mu_k)}\right] = \prod_{k=1}^n \mathbb{E}\left[e^{\lambda(X_k-\mu_k)}\right] \leq \prod_{k=1}^n e^{\lambda^2\sigma_k^2/2}, \quad \forall \lambda \in \mathbb{R}.$$

Then the variable $\sum_{k=1}^n (X_k - \mu_k)$ is sub-Gaussian with the parameter $\sqrt{\sum_{k=1}^n \sigma_k^2}$.

Hoeffding bound

Proposition 2.5 (Hoeffding bound)

Suppose that the variables $X_i, i = 1, \ldots, n$, are independent, and X_i has mean μ_i and sub-Gaussian parameter σ_i. Then for all $t \geq 0$, we have

$$\mathbb{P}\left[\sum_{i=1}^{n}(X_i - \mu_i) \geq t\right] \leq \exp\left\{-\frac{t^2}{2\sum_{i=1}^{n}\sigma_i^2}\right\}. \tag{2.10}$$

Note: If $X_i \in [a, b], i = 1, \ldots, n$,then it is sub-Gaussian with parameter $\sigma = \frac{b-a}{2}$, so that we obtain the bound

$$\mathbb{P}\left[\sum_{i=1}^{n}(X_i - \mu_i) \geq t\right] \leq \exp\left\{-\frac{2t^2}{n(b-a)^2}\right\}. \tag{2.11}$$

Sub-Gaussian properties

Theorem 2.6 (Equivalent characterizations of sub-Gaussian variables)

Given any zero-mean random variable X, the following properties are equivalent:

(I) There is a constant $\sigma \geq 0$ such that

$$\mathbb{E}\left[e^{\lambda X}\right] \leq e^{\frac{\lambda^2 \sigma^2}{2}} \quad \text{for all } \lambda \in \mathbb{R}$$

(II) There is a constant $c \geq 0$ and Gaussian random variable $Z \sim \mathcal{N}\left(0, \tau^2\right)$ such that

$$\mathbb{P}[|X| \geq s] \leq c\mathbb{P}[|Z| \geq s] \quad \text{for all } s \geq 0.$$

Sub-Gaussian properties

(III) There is a constant $\theta \geq 0$ such that

$$\mathbb{E}\left[X^{2k}\right] \leq \frac{(2k)!}{2^k k!}\theta^{2k} \quad \text{for all } k = 1, 2, \ldots.$$

(IV) There is a constant $\sigma \geq 0$ such that

$$\mathbb{E}\left[e^{\frac{\lambda X^2}{2\sigma^2}}\right] \leq \frac{1}{\sqrt{1-\lambda}} \quad \text{for all } \lambda \in [0, 1)$$

See Appendix A (Section 2.4) for the proof of these equivalences.

1. From Markov to Chernof
2. Sub-Gaussian Variables and Hoeffding Bounds
3. Sub-exponential Variables and Bernstein Bounds
4. Introduction of Martingale-based Methods
5. Concentration Bounds for Martingale Difference Sequences
6. Lipschitz Functions of Gaussian Variables

Sub-exponential variables

Definition 2.7 (sub-exponential)
A random variable X with mean $\mu = \mathbb{E}[X]$ is sub-exponential if there are non-negative parameters (v, α) such that

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq e^{\frac{v^2\lambda^2}{2}} \quad \text{for all } |\lambda| < \frac{1}{\alpha} \tag{2.13}$$

Note: Any sub-Gaussian variable is also sub-exponential . However, the converse statement is not true.

Example

Example 2.8 (Sub-exponential but not sub-Gaussian)
Let $Z \sim \mathcal{N}(0,1)$, and consider the random variable $X = Z^2$. For $\lambda < \frac{1}{2}$, we have

$$\mathbb{E}\left[e^{\lambda(X-1)}\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{\lambda\left(z^2-1\right)} e^{-z^2/2} dz$$
$$= \frac{e^{-\lambda}}{\sqrt{1-2\lambda}}$$

For $\lambda > \frac{1}{2}$, the moment generating function is infinite, which reveals that X is not sub-Gaussian.

Example

Following some calculus, we find that

$$\frac{e^{-\lambda}}{\sqrt{1-2\lambda}} \leq e^{2\lambda^2} = e^{4\lambda^2/2}, \qquad \text{for all } |\lambda| < \frac{1}{4},$$

which shows that X is sub-exponential with parameters $(v, \alpha) = (2, 4)$.

Sub-exponential tail bound

Proposition 2.9 (Sub-exponential tail bound)
Suppose that X is sub-exponential with parameters (v, α). Then

$$\mathbb{P}[X-\mu \geq t] \leq \begin{cases} e^{-\frac{t^2}{2v^2}} & \text{if } 0 \leq t \leq \frac{v^2}{\alpha} \\ e^{-\frac{t}{2\alpha}} & \text{for } t > \frac{v^2}{\alpha} \end{cases}$$

Proof: Without loss of generality that $\mu = 0$. Combining it with the definition (2.13) of a sub-exponential variable yields the upper bound

$$\mathbb{P}[X \geq t] \leq e^{-\lambda t}\mathbb{E}\left[e^{\lambda X}\right] \leq \exp\underbrace{\left(-\lambda t + \frac{\lambda^2 v^2}{2}\right)}_{g(\lambda,t)}, \quad \text{valid for all } \lambda \in \left[0, \alpha^{-1}\right).$$

Sub-exponential tail bound

For each fixed $t \geq 0$, the quantity $g^*(t) := \inf_{\lambda \in [0, \alpha^{-1})} g(\lambda, t)$. Note that the unconstrained minimum of the function $g(\cdot, t)$ occurs at $\lambda^* = \frac{t}{v^2}$.

(1) $\frac{t}{v^2} < \frac{1}{\alpha}$ ($0 \leq t < \frac{v^2}{\alpha}$), then $g^*(t) = -\frac{t^2}{2v^2}$ over this interval.

(2) $\frac{t}{v^2} \geq \frac{1}{\alpha}$ ($t \geq \frac{v^2}{\alpha}$)The function $g(\cdot, t)$ is monotonically decreasing in the interval $[0, \lambda^*)$, the constrained minimum is achieved at the boundary point α^{-1}, and we have

$$g^*(t) = g\left(\alpha^{-1}, t\right) = -\frac{t}{\alpha} + \frac{1}{2\alpha}\frac{v^2}{\alpha} \overset{(i)}{\leq} -\frac{t}{2\alpha},$$

where inequality (i) uses the fact that $\frac{v^2}{\alpha} \leq t$.

Bernstein's condition

Given a random variable X with mean μ and variance σ^2, we say that Bernstein's condition with parameter b holds if

$$\left|\mathbb{E}\left[(X-\mu)^k\right]\right| \leq \frac{1}{2}k!\sigma^2 b^{k-2} \quad \text{for } k = 2, 3, 4, \ldots \tag{2.15}$$

- When X satisfies the Bernstein condition, then it is sub-exponential with parameters $(\sqrt{2}\sigma, 2b)$.
- Even for bounded variables, our next result will show that the Bernstein condition can be used to obtain tail bounds that may be tighter than the Hoeffding bound.

Bernstein's condition

Proof:

By the power-series expansion of the exponential, we have

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] = 1 + \frac{\lambda^2\sigma^2}{2} + \sum_{k=3}^{\infty} \lambda^k \frac{\mathbb{E}\left[(X-\mu)^k\right]}{k!}$$
$$\overset{\text{(i)}}{\leq} 1 + \frac{\lambda^2\sigma^2}{2} + \frac{\lambda^2\sigma^2}{2}\sum_{k=3}^{\infty}(|\lambda|b)^{k-2}$$

where the inequality (i) makes use of the Bernstein condition (2.15). For any $|\lambda| < 1/b$, we can sum the geometric series so as to obtain

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq 1 + \frac{\lambda^2\sigma^2/2}{1-b|\lambda|} \overset{\text{(ii)}}{\leq} e^{\frac{\lambda^2\sigma^2/2}{1-b|\lambda|}}, \tag{2.16}$$

where inequality (ii) follows from the bound $1 + t \leq e^t$.

Bernstein's condition

Consequently, we conclude that

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq e^{\frac{\lambda^2(\sqrt{2}\sigma)^2}{2}} \quad \text{for all } |\lambda| < \frac{1}{2b},$$

showing that X is sub-exponential with parameters $(\sqrt{2}\sigma, 2b)$.

Bernstein-type bound

Proposition 2.10 **(Bernstein-type bound)**
For any random variable satisfying the Bernstein condition (2.15), we have

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \le e^{\frac{\lambda^2\sigma^2/2}{1-b|\lambda|}} \quad \text{for all } |\lambda| < \frac{1}{b},$$

and, moreover, the concentration inequality

$$\mathbb{P}[|X-\mu| \ge t] \le 2e^{-\frac{t^2}{2(\sigma^2+bt)}} \quad \text{for all } t \ge 0$$

The sum fo sub-exponential variables

Consider an independent sequence $\{X_k\}_{k=1}^n$ of random variables, such that X_k has mean μ_k, and is sub-exponential with parameters (v_k, α_k). We compute the moment generating function

$$\mathbb{E}\left[e^{\lambda \sum_{k=1}^n (X_k - \mu_k)}\right] = \prod_{k=1}^n \mathbb{E}\left[e^{\lambda (X_k - \mu_k)}\right] \leq \prod_{k=1}^n e^{\lambda^2 v_k^2/2}$$

valid for all $|\lambda| < \left(\max_{k=1,\dots,n} \alpha_k\right)^{-1}$. Thus, we conclude that the variable $\sum_{k=1}^n (X_k - \mu_k)$ is sub-exponential with the parameters (v_*, α_*), where

$$\alpha_* := \max_{k=1,\dots,n} \alpha_k \text{ and } v_* := \sqrt{\sum_{k=1}^n v_k^2}.$$

The sum fo sub-exponential variables

Using the same argument as in Proposition 2.9, this observation leads directly to the upper tail bound

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}(X_k-\mu_k)\geq t\right]\leq\begin{cases}e^{-\frac{m^2}{2\left(v_*^2/n\right)}} & \text{for } 0\leq t\leq\frac{v_*^2}{n\alpha_*}\\ e^{-\frac{nt}{2a_*}} & \text{for } t>\frac{v_*^2}{n\alpha_*}\end{cases}\tag{2.18}$$

Example

Example 2.11 (χ^2-variables)

Let $Y = \sum_{k=1}^n Z_k^2$ where $Z_k \sim \mathcal{N}(0,1)$ are i.i.d. variates. As discussed in Example 2.8, the variable Z_k^2 is sub-exponential with parameters $(2,4)$. Consequently, since the variables $\{Z_k\}_{k=1}^n$ are independent, the χ^2-variate Y is sub-exponential with parameters $(v,\alpha) = (2\sqrt{n}, 4)$, and the preceding discussion yields the two-sided tail bound

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{k=1}^n Z_k^2 - 1\right| \geq t\right] \leq 2e^{-nt^2/8}, \quad \text{for all } t \in (0,1).$$

Sub-exponential properties

Theorem 2.13 **(Equivalent characterizations of sub-exponential variables)**
For a zeromean random variable X, the following statements are equivalent:
(I) There are non-negative numbers (v, α) such that

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

(II) There is a positive number $c_0 > 0$ such that $\mathbb{E}\left[e^{\lambda X}\right] < \infty$ for all $|\lambda| \leq c_0$.

Sub-exponential properties

(III) There are constants $c_1, c_2 > 0$ such that

$$\mathbb{P}[|X| \geq t] \leq c_1 e^{-c_2 t} \quad \text{for all } t > 0.$$

(IV) The quantity $\gamma := \sup_{k \geq 2} \left[\frac{\mathbb{E}[X^k]}{k!} \right]^{1/k}$ is finite.

See Appendix B (Section 2.5) for the proof of this claim.

1 From Markov to Chernof

2 Sub-Gaussian Variables and Hoeffding Bounds

3 Sub-exponential Variables and Bernstein Bounds

4 Introduction of Martingale-based Methods

5 Concentration Bounds for Martingale Difference Sequences

6 Lipschitz Functions of Gaussian Variables

Backgroud

- Let $\{X_k\}_{k=1}^n$ be a sequence of independent random variables, and consider the random variable $f(X) = f(X_1 \ldots, X_n)$, for some function $f : \mathbb{R}^n \to \mathbb{R}$
- Suppose that our goal is to obtain bounds on the deviations of f from its mean. In order to do so, we consider the sequence of random variables given by $Y_0 = \mathbb{E}[f(X)]$, $Y_n = f(X)$, and

$$Y_k = \mathbb{E}\left[f(X) \mid X_1, \ldots, X_k\right] \quad \text{for } k = 1, \ldots, n-1,$$

- Based on this intuition, the martingale approach to tail bounds is based on the telescoping decompositio

$$f(X) - \mathbb{E}[f(X)] = Y_n - Y_0 = \sum_{k=1}^n \underbrace{(Y_k - Y_{k-1})}_{D_k},$$

Martingale

Definition 2.15 (Martingale)

Given a sequence $\{Y_k\}_{k=1}^{\infty}$ of random variables adapted to a filtration $\{\mathcal{F}_k\}_{k=1}^{\infty}$, the pair $\{(Y_k, \mathcal{F}_k)\}_{k=1}^{\infty}$ is a martingale if, for all $k \geq 1$,

$$\mathbb{E}\left[|Y_k|\right] < \infty \quad \text{and} \quad \mathbb{E}\left[Y_{k+1} \mid \mathcal{F}_k\right] = Y_k$$

- Let $\{\mathcal{F}_k\}_{k=1}^{\infty}$ be a sequence of σ-fields that are nested, meaning that $\mathcal{F}_k \subseteq \mathcal{F}_{k+1}$ for all $k \geq 1$; such a sequence is known as a filtration.
- If a sequence is martingale with respect to itself (i.e., with $\mathcal{F}_k = \sigma(Y_1, \ldots, Y_k)$), then we say simply that $\{Y_k\}_{k=1}^{\infty}$ forms a martingale sequence.

Examples

Example2.17 (Doob construction)

Given a sequence of independent random variables $\{X_k\}_{k=1}^n$, recall the sequence $Y_k = \mathbb{E}[f(X) \mid X_1, \ldots, X_k]$ previously defined, and suppose that $\mathbb{E}[|f(X)|] < \infty$.

Indeed, in terms of the shorthand $X_1^k = (X_1, X_2, \ldots, X_k)$,we have

$$\mathbb{E}[|Y_k|] = \mathbb{E}\left[\left|\mathbb{E}\left[f(X) \mid X_1^k\right]\right|\right] \leq \mathbb{E}[|f(X)|] < \infty,$$

where the bound follows from Jensen's inequality.

Turning to the second property, we have

$$\mathbb{E}\left[Y_{k+1} \mid X_1^k\right] = \mathbb{E}\left[\mathbb{E}\left[f(X) \mid X_1^{k+1}\right] \mid X_1^k\right] \stackrel{\text{(i)}}{=} \mathbb{E}\left[f(X) \mid X_1^k\right] = Y_k,$$

where we have used the tower property of conditional expectation in step (i).

Martingale difference sequence

Definition (Martingale difference sequence)

A closely related notion is that of martingale difference sequence, meaning an adapted sequence $\{(D_k, \mathcal{F}_k)\}_{k=1}^{\infty}$ such that, for all $k \geq 1$,

$$\mathbb{E}\left[|D_k|\right] < \infty \quad \text{and} \quad \mathbb{E}\left[D_{k+1} \mid \mathcal{F}_k\right] = 0.$$

In particular, given a martingale $\{(Y_k, \mathcal{F}_k)\}_{k=0}^{\infty}$, let us define $D_k = Y_k - Y_{k-1}$ for $k \geq 1$.

1. From Markov to Chernof
2. Sub-Gaussian Variables and Hoeffding Bounds
3. Sub-exponential Variables and Bernstein Bounds
4. Introduction of Martingale-based Methods
5. Concentration Bounds for Martingale Difference Sequences
6. Lipschitz Functions of Gaussian Variables

A general Bernstein-type bound

Theorem 2.19

Let $\{(D_k, \mathcal{F}_k)\}_{k=1}^{\infty}$ be a martingale difference sequence, and suppose that $\mathbb{E}\left[e^{\lambda D_k} \mid \mathcal{F}_{k-1}\right] \leq e^{\lambda^2 \nu_k^2/2}$ almost surely for any $|\lambda| < 1/\alpha_k$. Then the following hold:

(a) The sum $\sum_{k=1}^n D_k$ is sub-exponential with parameters $\left(\sqrt{\sum_{k=1}^n \nu_k^2}, \alpha_*\right)$ where $\alpha_* := \max_{k=1,\ldots,n} \alpha_k$.

(b) The sum satisfies the concentration inequality

$$\mathbb{P}\left[\left|\sum_{k=1}^n D_k\right| \geq t\right] \leq \begin{cases} 2e^{-\frac{t^2}{2\sum_{k=1}^n \nu_k^2}} & \text{if } 0 \leq t \leq \frac{\sum_{k=1}^n \nu_k^2}{2} \\ 2e^{-\frac{t}{2\alpha_*}} & \text{if } t > \frac{\sum_{k=1}^n \nu_k^2}{\alpha_*}. \end{cases} \tag{2.28}$$

A general Bernstein-type bound

Proof:

For any scalar λ such that $|\lambda| < \frac{1}{\alpha_s}$, conditioning on $\mathcal{F}_{n-1}$ and applying iterated expectation yields

$$\mathbb{E}\left[e^{\lambda\left(\sum_{k=1}^{n} D_k\right)}\right] = \mathbb{E}\left[e^{\lambda\left(\sum_{k=1}^{n-1} D_k\right)}\mathbb{E}\left[e^{\lambda D_n} \mid \mathcal{F}_{n-1}\right]\right]$$
$$\leq \mathbb{E}\left[e^{\lambda \sum_{k=1}^{n-1} D_k}\right] e^{\lambda^2 V_n^2/2},$$

where the inequality follows from the stated assumption on D_n.

Iterating this procedure , $\mathbb{E}\left[e^{\lambda \sum_{k=1}^{n} D_k}\right] \leq e^{\lambda^2 \sum_{k=1}^{n} \nu_k^2/2}$, valid for all $|\lambda| < \frac{1}{\alpha_*}$.

By definition, we conclude that $\sum_{k=1}^{n} D_k$ is sub-exponential with parameters $\left(\sqrt{\sum_{k=1}^{n} v_k^2}, \alpha_*\right)$

The tail bound (2.28) follows by applying Proposition 2.9.

Azuma-Hoeffding

Corollary 2.20 (Azuma-Hoeffding)

Let $(\{(D_k, \mathcal{F}_k)\}_{k=1}^{\infty})$ be a martingale difference sequence for which there are constants $\{(a_k, b_k)\}_{k=1}^{n}$ such that $D_k \in [a_k, b_k]$ almost surely for all $k = 1, \dots, n$. Then, for all $t \geq 0$,

$$\mathbb{P}\left[\left|\sum_{k=1}^{n} D_k\right| \geq t\right] \leq 2e^{-\frac{2t^2}{\sum_{k=1}^{n}(b_k - a_k)^2}}.$$

A Convenient Notation

Given vectors $x, x' \in \mathbb{R}^n$ and an index $k \in \{1, 2, \ldots, n\}$, we define a new vector $x^{\backslash k} \in \mathbb{R}^n$ via

$$x_j^{\backslash k} := \begin{cases} x_j & \text{if } j \neq k, \\ x_k' & \text{if } j = k \end{cases} \tag{2.31}$$

With this notation, we say that $f : \mathbb{R}^n \to \mathbb{R}$ satisfies the bounded difference inequality with parameters $(L_1, \ldots, L_n)$ if, for each index $k = 1, 2, \ldots, n$,

$$\left| f(x) - f\left(x^{\backslash k}\right) \right| \leq L_k \quad \text{for all } x, x' \in \mathbb{R}^n. \tag{2.32}$$

Bounded Differences Inequality

Corollary 2.21 **(Bounded differences inequality)**
Suppose that f satisfies the bounded difference property (2.32) with parameters $(L_1, \ldots, L_n)$ and that the random vector $X = (X_1, X_2, \ldots, X_n)$ has independent components. Then

$$\mathbb{P}[|f(X) - \mathbb{E}[f(X)]| \geq t] \leq 2e^{-\frac{2t^2}{\sum_{k=1}^{n} L_k^2}} \quad \text{for all } t \geq 0 \tag{2.33}$$

Bounded Differences Inequality

Sketch of Proof:

- Recalling the Doob martingale introduced in Example 2.17, consider the associated martingale difference sequence

$$D_k = \mathbb{E}\left[f(X) \mid X_1, \ldots, X_k\right] - \mathbb{E}\left[f(X) \mid X_1, \ldots, X_{k-1}\right].$$

  and

$$f(X) - \mathbb{E}[f(X)] = Y_n - Y_0 = \sum_{k=1}^{n} \underbrace{(Y_k - Y_{k-1})}_{D_k},$$

- Claim that D_k lies in an interval of length at most L_k almost surely.
- The claim follows as a corollary of the Azuma-Hoeffding inequality.

Examples

Example 2.22: (Classical Hoeffding from bounded differences)
Let $X_i \in [a, b]$, the function $f(x_1, \ldots, x_n) = \sum_{i=1}^{n} (x_i - \mu_i)$, where $\mu_i = \mathbb{E}[X_i]$ is the mean of the i th random variable.
For any index $k \in \{1, \ldots, n\}$, we have

$$\begin{aligned} \left| f(x) - f\left(x^{\backslash k}\right) \right| &= \left| (x_k - \mu_k) - \left(x_k' - \mu_k\right) \right| \\ &= \left| x_k - x_k' \right| \le b - a, \end{aligned}$$

Then f satisfies the bounded difference inequality in each coordinate with parameter $L = b - a$. Combining Corollary 2.21,we have

$$\mathbb{P}\left[\left| \sum_{i=1}^{n} (X_i - \mu_i) \right| \ge t \right] \le 2e^{-\frac{2t^2}{n(b-a)^2}}$$

which is the classical Hoeffding bound for independent random

Examples

Example 2.23: (U-Statistics) Let $g : \mathbb{R}^2 \to \mathbb{R}$ be a symmetric function of its arguments. Given an i.i.d. sequence $X_k, k \geq 1$, of random variables, the quantity

$$U := \frac{1}{\binom{n}{2}} \sum_{j<k} g(X_j, X_k)$$

is known as a pairwise U-statistic. If g is bounded (say $\|g\|_\infty \leq b$), then Corollary 2.21 can be used to establish the concentration of U around its mean.

U-Statistics

Viewing U as a function $f(x) = f(x_1, \ldots, x_n)$, for any given coordinate k, we have

$$\left|f(x) - f\left(x^{\backslash k}\right)\right| \leq \frac{1}{\binom{n}{2}} \sum_{j \neq k} \left|g(x_j, x_k) - g\left(x_j, x_k'\right)\right|$$
$$\leq \frac{(n-1)(2b)}{\binom{n}{2}} = \frac{4b}{n}$$

so that the bounded differences property holds with parameter $L_k = \frac{4b}{n}$ in each coordinate. Thus, we conclude that

$$\mathbb{P}[|U - \mathbb{E}[U]| \geq t] \leq 2e^{-\frac{nt^2}{8b^2}}$$

1. From Markov to Chernof
2. Sub-Gaussian Variables and Hoeffding Bounds
3. Sub-exponential Variables and Bernstein Bounds
4. Introduction of Martingale-based Methods
5. Concentration Bounds for Martingale Difference Sequences
6. **Lipschitz Functions of Gaussian Variables**

Lipschitz Functions of Gaussian Variables

Let us say that a function $f : \mathbb{R}^n \to \mathbb{R}$ is L**-Lipschitz** with respect to the Euclidean norm $\|\cdot\|_2$ if

$$|f(x) - f(y)| \leq L\|x - y\|_2 \quad \text{for all } x, y \in \mathbb{R}^n.$$

Theorem 2.26

Let $(X_1, \ldots, X_n)$ be a vector of i.i.d. standard Gaussian variables, and let $f : \mathbb{R}^n \to \mathbb{R}$ be L-Lipschitz with respect to the Euclidean norm. Then the variable $f(X) - \mathbb{E}[f(X)]$ is sub-Gaussian with parameter at most L, and hence

$$\mathbb{P}[|f(X) - \mathbb{E}[f(X)]| \geq t] \leq 2e^{-\frac{t^2}{2L^2}} \quad \text{for all } t \geq 0 \tag{2.39}$$

Lipschitz Functions of Gaussian Variables

LEmma 2.27

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable. Then for any convex function $\phi : \mathbb{R} \to \mathbb{R}$, we have

$$\mathbb{E}[\phi(f(X) - \mathbb{E}[f(X)])] \leq \mathbb{E}\left[\phi\left(\frac{\pi}{2}\langle \nabla f(X), Y\rangle\right)\right], \tag{2.40}$$

where $X, Y \sim \mathcal{N}(0, \mathbf{I}_n)$ are standard multivariate Gaussian, and independent.

Lipschitz Functions of Gaussian Variables

Proof: For any fixed $\lambda \in \mathbb{R}$, applying the lemma to the convex function $t \mapsto e^{\lambda t}$ yields

$$\mathbb{E}_X[\exp(\lambda\{f(X) - \mathbb{E}[f(X)]\})] \le \mathbb{E}_{X,Y}\left[\exp\left(\frac{\lambda\pi}{2}\langle Y, \nabla f(X)\rangle\right)\right] = \mathbb{E}_X\left[\exp\left(\frac{\lambda^2\pi^2}{8}\|\nabla f(X)\|_2^2\right)\right],$$

where $\langle Y, \nabla f(x)\rangle$ is a zero-mean Gaussian variable with variance $\|\nabla f(x)\|_2^2$.
Due to the Lipschitz condition on f, we have $\|\nabla f(x)\|_2 \le L$ for all $x \in \mathbb{R}^n$, whence

$$\mathbb{E}[\exp(\lambda\{f(X) - \mathbb{E}[f(X)]\})] \le e^{\frac{1}{8}\lambda^2\pi^2L^2},$$

Lipschitz Functions of Gaussian Variables

Thus $f(X) - \mathbb{E}[f(X)]$ is sub-Gaussian with parameter at most $\frac{\pi L}{2}$. Combined with Proposition 2.5,we can get the tail bound

$$\mathbb{P}[|f(X) - \mathbb{E}[f(X)]| \geq t] \leq 2\exp^{\left(-\frac{2t^2}{\pi^2 L^2}\right)} \quad \text{for all } t \geq 0$$

Example

Example 2.28 (χ^2 concentration)
For a given sequence $\{Z_k\}_{k=1}^n$ of i.i.d. standard normal variates, the random variable $Y := \sum_{k=1}^n Z_k^2$ follows a χ^2-distribution with n degrees of freedom.
Indeed, defining the variable $V = \sqrt{Y}/\sqrt{n}$, we can write $V = \left\|(Z_1, \ldots, Z_n)\right\|_2 / \sqrt{n}$, and since the Euclidean norm is a 1-Lipschitz function, Theorem 2.26 implies that

$$\mathbb{P}[V \geq \mathbb{E}[V] + \delta] \leq e^{-n\delta^2/2} \quad \text{for all } \delta \geq 0.$$

where

$$\mathbb{E}[V] \leq \sqrt{\mathbb{E}\left[V^2\right]} = \left\{\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[Z_k^2\right]\right\}^{1/2} = 1$$

Example

Recalling that $V = \sqrt{Y}/\sqrt{n}$ and putting together the pieces yields

$$\mathbb{P}\left[Y/n \geq (1+\delta)^2\right] \leq e^{-n\delta^2/2} \quad \text{for all } \delta \geq 0.$$

Since $(1+\delta)^2 = 1 + 2\delta + \delta^2 \leq 1 + 3\delta$ for all $\delta \in [0,1]$, we conclude that

$$\mathbb{P}[Y \geq n(1+t)] \leq e^{-nt^2/18} \quad \text{for all } t \in [0,3],$$

where the substitution $t = 3\delta$.