Basic tail and concentration bounds

Kaveh S. Nobari

Lectures in High-Dimensional Statistics

Department of Mathematics and Statistics
Lancaster University

Contents

Motivation

It is often of interest to obtain bounds on the tails of a random variable, or two-sided inequalities, which guarantee that the random variable is close to its mean or median. These slides follow the structure of chapter 2 of Wainwright (2019) to shed light on the elementary techniques for obtaining deviation and concentration inequalities.

One way of controlling a tail probability $P[X \geq t]$ is by controlling the moments of the random variable X, where by controlling higher-order moments of the variable X, we can obtain sharper bounds on tail probabilities. This motivates the "Classical bounds" section of the notes.

We then extend the derivation of bounds to more general functions of the random variables in the "Martingale-based methods" section using martingale decompositions, as opposed to limiting the techniques to deriving bounds on the sum of independent random variables.

Finally, the seminar is concluded with a classical result on the concentration properties of Lipschitz functions of Gaussian variables.

The most elementary tail bound is Markov's inequality:

Markov's inequality

Given a non-negative random variable X with finite mean - i.e. $\mathbb{E}[X] < \infty$, we have

$$P[X \geq t] \leq \frac{\mathbb{E}[X]}{t}, \quad \forall t > 0$$

It is immediately obvious that Markov's inequality requires only the existence of the first moment. If the random variable X also has finite variance - i.e. $\text{var}(X) < \infty$, we have Chebyshev's inequality:

Chebyshev's inequality

For a random variable X that has a finite mean and variance, we have

$$P[|X - \mu| \geq t] \leq \frac{\text{var}(X)}{t^2}, \quad \forall t > 0$$

Proof: Chebyshev's inequality follows from Markov's inequality, by considering the variable $(X - \mathbb{E}[X])^2$ and the constant t^2. By substituting these in the Markov inequality, we get

$$P[(X - \mathbb{E}[X])^2 \geq t^2] \leq \frac{\mathbb{E}[(X - \mathbb{E}[X])^2]}{t^2} \tag{1}$$

$$P[|X - \mathbb{E}[X]| \geq t] \leq \frac{\mathbb{E}[(X - \mathbb{E}[X])^2]}{t^2} \tag{2}$$

Since, $\mathbb{E}[X] = \mu$ and $\text{var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$, we get

$$P[|X - \mu| \geq t] \leq \frac{\text{var}(X)}{t^2}$$

The earlier results can be generalised as follows:

Extensions of Markov's inequality

Whenever a variable X has a central moment of order k, an application of Markov's inequality to the random variable $|X - \mu|^k$ yields:

$$P[|X - \mu| \geq t] \leq \frac{\mathbb{E}[|X - \mu|]^k}{t^k}, \quad \forall t > 0.$$

and this is not limited to polynomials $|X - \mu|^k$:
Suppose X has a mgf in a neighbourhood of zero,such that there is a constant $b > 0$ that the functions $\rho(\lambda) = \mathbb{E}[\exp(\lambda(X - \mu))]$ exists for all $\lambda < |b|$. Thus, for any $\lambda \in [0, b]$, we may apply Markov's inequality to the random variable $Y = \exp(\lambda(X - \mu))$, obtaining the upper bound:

$$P[(X - \mu) \geq t] = P[\exp(\lambda(X - \mu)) \geq \exp(\lambda t)] \leq \frac{\mathbb{E}[\exp(\lambda(X - \mu))]}{\exp(\lambda t)}$$

By taking the log of both side of the latter inequality, we get:

$$\log P[(X - \mu) \geq t] \leq \log \mathbb{E}[\exp(\lambda(X - \mu))] - \lambda t$$

Optimising, our choice of λ, we can obtain the tightest results that yields the Chernoff bound:

Chernoff bound

$$\log P[(X - \mu) \geq t] \leq \inf_{\lambda \in [0,b]} \{\log \mathbb{E}[\exp(\lambda(X - \mu))] - \lambda t\}.$$

Evidently, the form of the tail bound obtained using the Chernoff approach depends on the growth rate of the mgf. Naturally, in the study of the tail bounds the random variables are then classified in terms of their mgfs. The simplest type of behaviour is known as sub-Gaussian, which shall be motivated by deriving tail bounds for a Gaussian variables, say, X, such that $X \sim N(\mu, \sigma^2)$, with density

$$f(X) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(X-\mu)^2}{2\sigma^2}\right)$$

and thus, the mgf

$$\mathbb{E}[\exp(\lambda X)] = \exp\left(\mu\lambda + \frac{\sigma^2\lambda^2}{2}\right), \quad \forall\lambda \in \mathbb{R}$$

Example (Gaussian tail bounds)

Let $X \sim N(\mu, \sigma^2)$ be a Gaussian r.v., which has mgf

$$\mathbb{E}[\exp(\lambda X)] = \exp\left(\mu\lambda + \frac{\sigma^2\lambda^2}{2}\right), \quad \forall \lambda \in \mathbb{R}$$

substituting this into the optimising problem of the Chernoff bound, we get

$$\inf_{\lambda \geq 0}\{\log \mathbb{E}[\exp(\lambda(X-\mu)) - \lambda t]\} = \inf_{\lambda \geq 0}\left\{\frac{\sigma^2\lambda^2}{2} - \lambda t\right\} = -\frac{t^2}{2\sigma^2} \tag{3}$$

Therefore, we can conclude that any $N(\mu, \sigma^2)$ r.v. satisfies the upper deviation inequality

$$P[X \geq \mu + t] \leq \exp\left(-\frac{t^2}{2\sigma^2}\right) \tag{4}$$

Proof of equation (3).

To solve the optimisation problem below

$$\inf_{\lambda \geq 0} \left\{ \frac{\sigma^2 \lambda^2}{2} - \lambda t \right\}$$

we take derivatives to find the optimum of this quadratic function, - i.e.

$$\frac{\partial}{\partial \lambda} \left(\frac{\sigma^2 \lambda^2}{2} - \lambda t \right) = 0,$$

which leads to $\lambda_{opt} = \frac{t}{\sigma^2}$. Substituting λ_{opt} with λ in the above equation yields relationship (3). □

Definition (Sub-Gaussianity)

A r.v. X with mean $\mu = \mathbb{E}[X]$ is sub-Gaussian if there is a positive number σ, such that

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

where the constant σ is referred to as the sub-Gaussian parameter. Moreover, by the symmetry of the definition, the variable $-X$ is sub-Gaussian iff X is sub-Gaussian, so that we also have lower deviation inequality $P[X \leq \mu - t] \leq \exp\left(-\frac{t^2}{2\sigma^2}\right), \quad \forall t \geq 0$. Thus, we conclude that any sub-Gaussian variable satisfies the concentration inequality

$$P[|X - \mu| \geq t] \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right), \quad \forall t \in \mathbb{R}.$$

We may have scenarios in which sub-Gaussian variables are non-Gaussian.

Example (Rademacher variables)

A Rademacher r.v. ε takes the values $[-1, +1]$ equiprobably -i.e $P[\varepsilon = -1] = P[\varepsilon = +1] = \frac{1}{2}$. Thus, the mgf of ε is as follows

$$\mathbb{E}[\exp(\lambda\varepsilon)] = \sum_{i\in\{-1,+1\}} \exp(\lambda\varepsilon_i)p(\varepsilon = i) = \frac{1}{2}[\exp(-\lambda) + \exp(\lambda)]$$

where the Maclaurin-series expansion of the terms $\exp(-\lambda)$ and $\exp(\lambda)$ leads gives us

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

with the sub-Gaussian parameter $\sigma = 1$.

Some preliminaries: The exponential function $g(z) = \exp(z)$ is convex; thus, Jensen's inequality for convex functions applies as follows

$$g(\mathbb{E}[z]) \leq \mathbb{E}[g(z)]$$

A r.v. Z' is an independent copy of Z, if it has a same the same distribution as Z, and where Z and Z' are independent.
Given the above definitions, we provide a simple example of symmetrization argument, in which first an independent copy of X, X' is introduced and the problem is symmetrized using a Rademacher variable.

Symmetrization argument

Let X be a r.v. with mean zero - i.e. $\mu = \mathbb{E}_X[X] = 0$, with a support on the interval $[a, b]$, and let X' be an independent copy of X, for any $\lambda \in \mathbb{R}$, we have

$$\mathbb{E}_X[\exp(\lambda X)] = \mathbb{E}_X[\exp(\lambda(X - \mathbb{E}_{X'}[X']))]$$

since $\mathbb{E}_X[X] = \mathbb{E}_{X'}[X'] = 0$. Using Jensen's inequality, we further establish that

$$\mathbb{E}_X[\exp(\lambda(X - \mathbb{E}[X']))] \leq \mathbb{E}_{X,X'}[\exp(\lambda(X - X'))]$$

Further, note that $\varepsilon(X - X')$ and $(X - X')$ possess the same distribution, where ε is a Rademacher r.v., so that

$$\mathbb{E}_{X,X'}[\exp(\lambda(X - X'))] = \mathbb{E}_{X,X'}[\mathbb{E}_\varepsilon[\exp(\lambda\varepsilon(X - X'))]],$$

,

where from the earlier example, we know that

$$\mathbb{E}_{X,X'}[\mathbb{E}_{\varepsilon}[\exp(\lambda\varepsilon(X-X'))]] \leq \mathbb{E}_{X,X'}\left[\exp\left(\frac{\lambda^2(X-X')^2}{2}\right)\right]$$

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

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

thus, we have shown that X is sub-Gaussian with sub-Gaussian parameter $\sigma = b - a$

Quiz:

1) Two independent sub-Gaussian variables X_1 and X_2 possess the sub-Gaussian parameters σ_1 and σ_2 respectively. What is the sub-Gaussian parameter of $X_1 + X_2$?
2) Now once again consider the sub-Gaussian tail bound (4). How is this result extended to the variable $X_1 + X_2$?

The answers to the above quiz, can be generalised to the variables $X_1, \cdots, X_n$ with mean μ_i and sub-Gaussian parameters σ_i for $i = 1, \cdots, n$ leading to the Hoeffding bound

Hoeffding bounds

Suppose that the variables $X_1, \cdots, X_n$ each with mean $\mu_1, \cdots, \mu_n$ and sub-Gaussian parameter $\sigma_1, \cdots, \sigma_n$ are independent. Then we have

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

To prove the equivalent characterizations of sub-Gaussian variables, it is of interest to first answer Exercise 2.2 of Wainwright (2019) which introduces Mills ratio.

Exercise 2.2 of Wainwright (2019): Let $\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-z^2}{2}\right)$ be the density function of a standard normal $Z \sim N(0,1)$ variate.

1) Show that $\phi'(z) + z\phi(z) = 0$
2) Use part 1 to show that

$$\phi(z)\left(\frac{1}{z} - \frac{1}{z^3}\right) \leq P[Z \geq z] \leq \phi(z)\left(\frac{1}{z} - \frac{1}{z^3} + \frac{3}{z^5}\right), \quad \forall z > 0$$

Solution:

Part 1:

$$\phi'(z) = -\frac{z}{\sqrt{2\pi}} \exp\left(\frac{-z^2}{2}\right) \quad \text{and} \quad z\phi(z) = \frac{z}{\sqrt{2\pi}} \exp\left(\frac{-z^2}{2}\right)$$

thus,

$$\phi'(z) + z\phi(z) = 0$$

Part 2:

Note that $P[Z \geq z] = \int_z^\infty \phi(t)dt$. Furthermore, from part 1, we know that $\phi(z) = \frac{-\phi'(z)}{z}$. By substituting $\frac{-\phi'(z)}{z}$ into the earlier integral, we get

$$\int_z^\infty \phi(t)dt = \int_z^\infty \frac{-\phi'(t)}{t}dt = \left[\frac{-\phi'(t)}{t}\right]_z^\infty - \int_z^\infty \frac{\phi(t)}{t^2}dt$$

We know that $\lim_{t\to\infty} \frac{-\phi'(t)}{t} = 0$, therefore, we may apply the the above expression can be written as

$$\frac{\phi(z)}{z} - \int_z^\infty \frac{-\phi'(t)}{t^3}dt$$

using the substitution derived from Miller's ratio. Using integration by parts yet again, we obtain

$$\begin{aligned}
\frac{\phi(z)}{z} - \int_z^\infty \frac{-\phi'(t)}{t^3}\,dt &= \frac{\phi(z)}{z} + \left[\frac{\phi(t)}{t^3}\right]_z^\infty - \int_z^\infty \frac{-3\phi(t)}{t^4}\,dt \\
&= \frac{\phi(z)}{z} + \frac{\phi(z)}{z^3} + \int_z^\infty \frac{3\phi(t)}{t^4}\,dt \\
P[Z \geq z] &= \phi(z)\left(\frac{1}{z} + \frac{1}{z^3}\right) + \underbrace{\int_z^\infty \frac{3\phi(t)}{t^4}\,dt}_{\geq 0} \\
&\geq \phi(z)\left(\frac{1}{z} + \frac{1}{z^3}\right)
\end{aligned}$$

Applying the same procedure again will prove the upper inequality. This is left as an exercise to the reader.

Equivalent characterizations of the sub-Gaussian variables (I-II)

(I) From the definition of sub-Gaussian variables, a r.v. with $\mu = \mathbb{E}[X] = 0$ is sub-Gaussian for $\sigma \geq 0$,

$$\mathbb{E}[\exp(\lambda X)] \leq \exp\left(\frac{\sigma^2\lambda^2}{2}\right), \quad \forall \lambda \in \mathbb{R}$$

(II) There is a constant $c \geq 0$ and Gaussian r.v. $Z \sim N(0, \tau^2)$, such that

$$P[|X| \geq s] \leq cP[|Z| \geq s], \quad \forall s \geq 0.$$

The notion of sub-Gaussianity is rather restrictive. We thus now introduce sub-exponential variables, which impose milder conditions on the mgf.

Definition

Sub-exponentiality A r.v. X with mean $\mu = \mathbb{E}[X]$ is sub-exponential if there are non-negative parameters (ν, α), such that

$$\mathbb{E}[\exp(\lambda(X - \mu))] \leq \exp\left(\frac{\nu^2\lambda^2}{2}\right), \quad \forall|\lambda| < \frac{1}{\alpha}$$

It is immediately obvious that any sub-Gaussian variable is also sub-exponential, where the former is a special case of the latter, with $\nu = \sigma$ and $\alpha = 0$. However, the converse is not true.

An example of a case where a variable is sub-exponential but not sub-Gaussian is as follows

Example (sub-exponential but not sub-Gaussian)

Let $Z \sim N(0,1)$, and consider the r.v. $X = Z^2$, such that $Z \sim \chi_1^2$. Therefore, the mean $\mu = \mathbb{E}[\chi_1^2] = 1$. For $\lambda < \frac{1}{2}$, we have the mgf as follows

$$\begin{aligned}\mathbb{E}[\exp(\lambda(X-1))] &= \int_{-\infty}^{+\infty} \exp(\lambda(Z^2-1)) f(z) dz \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp(\lambda(Z^2-1)) \exp\left(\frac{-Z^2}{2}\right) dz \\ &= \frac{\exp(-\lambda)}{\sqrt{1-2\lambda}}.\end{aligned}$$

for $\lambda \geq \frac{1}{2}$ the mgf is infinite, which reveals that X is not sub-Gaussian.

To obtain the tail-bounds of sub-exponential variables, we refer to the Chernoff-type approach - i.e.

$$P[X - \mu \geq t] = P[\exp(\lambda(X - \mu)) \geq \exp(t\lambda)] \leq \frac{\mathbb{E}[\exp(\lambda(X - \mu))]}{\exp(\lambda t)}$$

where from the definition of sub-exponential variables, we get the upper bound

$$P[X - \mu \geq t] \leq \frac{\mathbb{E}[\exp(\lambda(X - \mu))]}{\exp(\lambda t)} \leq \exp\left(\frac{\lambda^2 \nu^2}{2} - \lambda t\right), \quad \forall \lambda \in \left[0, \frac{1}{\alpha}\right),$$

where the Chernoff optimisation problem is

$$\log P[X - \mu \geq t] \leq \inf_{\lambda \in [0, \alpha^{-1}]} \left\{ \frac{\lambda^2 \nu^2}{2} - \lambda t \right\}$$

where using the same unconstrained optimisation approach as for sub-Gaussian variables, we'd obtain $\lambda_{opt} = \frac{t}{\nu^2}$, which yields the minimum $-\frac{t^2}{2\nu^2}$.

Recall the constraint $0 \leq \lambda < \frac{1}{\alpha}$. This implies that the unconstrained optimal λ_{opt} must be between $0 \leq \frac{t}{\nu^2} < \frac{1}{\alpha}$, which implies that in the interval $0 \leq t < \frac{\nu^2}{\alpha^2}$, the unconstrained optimum corresponds to the constrained optimum.
Otherwise for $t \geq \frac{\nu^2}{\alpha^2}$, considering that the function $g(., t) = \frac{\lambda^2 \nu^2}{2} - \lambda t$ is monotonically decreasing, in the interval $[0, \lambda^*)$, the constrained minimum is obtained at the boundary - i.e. $\lambda^{\#} = \frac{1}{\alpha}$, which leads to the minimum

$$g^*(t) = g(\lambda^{\#}, t) = -\frac{t}{\alpha} + \frac{1}{2\alpha}\frac{\nu^2}{\alpha} \leq -\frac{t}{2\alpha}$$

where this inequality used the fact that $\frac{\nu^2}{\alpha} \leq t$.
The results above lead to the sub-exponential tail bounds as follows

Sub-exponential tail bounds

Suppose X is sub-exponential with parameters (ν, α). Then

$$P[X - \mu \geq t] \leq \begin{cases} \exp\left(-\frac{t}{2\nu^2}\right) & 0 \leq t \leq \frac{\nu^2}{\alpha}, \\ \exp\left(-\frac{t}{2\alpha}\right) & t > \frac{\nu^2}{\alpha}. \end{cases}$$

The sub-exponential property can be verified by computing or bounding the mgf, which may not be practical in many different settings. One other approach is based on the control of the polynomial moments of X, which leads to the Bernstein condition

Bernstein condition

Given a r.v. X with mean $\mu = \mathbb{E}[X]$ and variance $\sigma^2 = \mathbb{E}[X^2] - \mu^2$, the Bernstein condition with parameter b holds if

$$|\mathbb{E}[(X - \mu)^k]| \leq \frac{1}{2} k! \sigma^2 b^{k-2}, \quad k \geq 2$$

One sufficient condition for the Bernstein condition to hold is that X is bounded. When X satisfies the Bernstein condition, then it is sub-exponential with parameters σ^2 and b. The Maclaurin-series expansion of the mgf can be expressed as follows

$$\begin{aligned}
\mathbb{E}[\exp(\lambda(X-\mu))] &= \mathbb{E}\left\{\sum_{i=0}^{\infty}\frac{f^{(i)}(0)}{i!}\left[\lambda(X-\mu)\right]^i\right\} \\
&= \sum_{i=0}^{\infty}\mathbb{E}\left\{\frac{f^{(i)}(0)}{i!}\left[\lambda(X-\mu)\right]^i\right\} \\
&= 1+\lambda\mathbb{E}[(X-\mu)]+\frac{\lambda^2\mathbb{E}[(X-\mu)]^2}{2}+\sum_{i=3}^{\infty}\frac{\lambda^i\mathbb{E}[(X-\mu)^i]}{i!} \\
&= 1+\frac{\lambda^2\sigma^2}{2}+\sum_{i=3}^{\infty}\frac{\lambda^i\mathbb{E}[(X-\mu)^i]}{i!}
\end{aligned}$$

Note that from the definition of Bernstein condition, we have

$$\frac{|\mathbb{E}[(X-\mu)^i]|}{i!} \leq \frac{1}{2}\sigma^2 b^{i-2}$$

Therefore,

$$\mathbb{E}[\exp(\lambda(X-\mu))] \leq 1 + \frac{\lambda^2\sigma^2}{2} + \frac{\lambda^2\sigma^2}{2}\sum_{i=3}^{\infty}(|\lambda|b)^{i-2}$$

For $|\lambda| < \frac{1}{b}$, we sum the geometric series,

$$\sum_{i=3}^{\infty}(|\lambda|b)^{i-2} = \frac{1}{1-|\lambda|b}$$

which leads to the following inequality

$$\mathbb{E}[\exp(\lambda(X-\mu))] \leq 1 + \frac{\lambda^2\sigma^2}{2} + \frac{\lambda^2\sigma^2}{2}\frac{1}{1-|\lambda|b}$$

Noting that

$$\exp\left(\frac{\lambda^2\sigma^2/2}{1-|\lambda|b}\right) = 1 + \frac{\lambda^2\sigma^2/2}{1-|\lambda|b} + \cdots$$
$$\geq 1 + \frac{\lambda^2\sigma^2/2}{1-|\lambda|b}$$

leading to Bernstein-type bound.

Bernstein-type bound

For any r.v. satisfying the Bernstein condition, we have

$$E[\exp(\lambda(X-\mu))] \leq \exp\left(\frac{\lambda^2\sigma^2/2}{1-|\lambda|b}\right), \quad \forall|\lambda| < \frac{1}{b}$$

As with the sub-Gaussian property, the sub-exponential property is preserved under summation for independent r.v.s. Consider the independent

sequence $X_1, \cdots, X_n$, with means $\mu_1, \cdots, \mu_n$ and sub-exponential parameters $(\nu_1, \alpha_1), \cdots, (\nu_n, \alpha_n)$. The mgf can be calculated as follows

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

for all $|\lambda| < (\max_{i=1,\cdots,n})^{-1}$. Hence, the variable $\sum_{i=1}^{n}(X_i - \mu_i)$ is sub-exponential with parameters (ν^*, α^*), where

$$\alpha^* := max_{i=1,\cdots,n}\alpha_i, \quad \text{and} \quad \nu^* := \sqrt{\sum_{i=1}^{n}\nu_i^2}$$

which using a Chernoff-type approach as before, leads to upper tail bound

$$P\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_i) \geq t\right] \leq \begin{cases} \exp\left(-\frac{nt^2}{2(\nu^{*2}/n)}\right), & 0 \leq t \leq \frac{\nu^{*2}}{n\alpha^*} \\ \exp\left(-\frac{nt}{2\alpha^*}\right), & t \geq \frac{\nu^{*2}}{n\alpha^*} \end{cases}$$

Let us extend the techniques considered for independent r.v.s to more general functions of the variables. One classical approach is based on martingale decomposition. Consider the independent r.vs $X_1, \cdots, X_n$ and consider a function $f(X) = f(X_1, \cdots, X_n)$ with the mapping $f : \mathbb{R}^n \to \mathbb{R}$. Suppose our goal is to obtain bounds on the deviations of f from its mean. To achieve this, let us consider the sequence of r.v.s given by $Y_0 = \mathbb{E}[f(X)]$, $Y_n = f(X)$, and

$$Y_k = \mathbb{E}[f(X) \mid X_1, \cdots, X_k] \quad k = 1, \cdots, n-1,$$

where Y_0 is a constant and the variables $Y_1, \cdots, Y_n$ tend to exhibit more fluctuations as they move along the sequence. Based on this intuition the martingale approach is based on the telescoping decomposition

$$f(X) - \mathbb{E}[X] = Y_n - Y_0 = \sum_{i=1}^{n} \underbrace{Y_i - Y_{i-1}}_{D_i}$$

Thus, $f(X) - \mathbb{E}[f(X)]$ is expressed as the sum of increments $D_1, \cdots, D_n$. This is a specific example of a martingale sequence, most commonly referred to as Doob martingale, whereas $D_1, \cdots, D_n$ is a martingale difference sequence (MDS hereafter).
We now provide a general definition of a martingale sequence by first defining a filtration, as follows

Filtration

Let $\{\mathcal{F}_i\}_{i=1}^{\infty}$ be a sequence of σ-fields that are nested, meaning that $\mathcal{F}_m \subseteq \mathcal{F}_n$ for $n \geq m$. Such a sequence is known as a filtration.

In the Doob martingale described earlier, the σ-field $\sigma(X_1, \cdots, X_m)$ is spanned by the first m variables $X_1, \cdots, X_m$ and plays the role of $\mathcal{F}_m$. Let $\{Y_i\}_{i=1}^{\infty}$ be a sequence of r.vs such that Y_i is measurable wrt to the σ-field $\mathcal{F}_i$. We say that $\{Y_i\}_{i=1}^{\infty}$ is adapted to the filtration $\{\mathcal{F}_i\}_{i=1}^{\infty}$.

Martingale

Given a sequence $\{Y_i\}_{i=1}^{\infty}$ of r.v.s adapted to a filtration $\{\mathcal{F}_i\}_{i=1}^{\infty}$, the pair $\{(Y_i, \mathcal{F}_i)\}_{i=1}^{\infty}$ is a martingale if, for all $i \geq 1$

$$\mathbb{E}[|Y_i|] < \infty \quad \text{and} \quad \mathbb{E}[Y_{i+1} \mid \mathcal{F}_i] = Y_i.$$

Example (Partial sums as martingales)

Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of i.i.d r.v.s with mean μ, and define the partial sums $S_m := \sum_{i=1}^{m} X_i$. Define $\mathcal{F}_m = \sigma(X_1, \cdots, X_m)$, the r.v. S_m is measurable wrt to $\mathcal{F}_m$, and, we have

$$\begin{aligned}
\mathbb{E}[S_{m+1} \mid \mathcal{F}_m] &= \mathbb{E}[X_{m+1} + S_m \mid X_1, \cdots, X_m] \\
&= \mathbb{E}[X_{m+1} \mid X_1, \cdots, X_m] + \mathbb{E}[S_m \mid X_1, \cdots, X_m] \\
&= \mathbb{E}[X_{m+1}] + S_m = \mu + S_m.
\end{aligned}$$

A closely related concept is that of the martingale difference sequence, which is an adapted sequence $\{D_i, \mathcal{F}_i\}_{i=1}^{\infty}$ such that, for all $i \geq 1$,

$$\mathbb{E}[|D_i|] < \infty \quad \text{and} \quad \mathbb{E}[D_{i+1} \mid \mathcal{F}_i] = 0.$$

Difference sequences arise naturally from martingales. Given a martingale $\{(Y_i, \mathcal{F}_i)\}_{i=0}^{\infty}$, define $D_i = Y_i - Y_{i-1}$ for $i \geq 1$. We then have

$$\begin{aligned}\mathbb{E}[D_{i+1} \mid \mathcal{F}_i] &= \mathbb{E}[Y_{i+1} - Y_i \mid \mathcal{F}_i] \\ &= \mathbb{E}[Y_{i+1} \mid \mathcal{F}_i] - Y_i \\ &= Y_i - Y_i = 0\end{aligned}$$

using the martingale property and the fact that Y_i is measurable wrt to $\mathcal{F}_i$. Thus, for any martingale sequence $\{Y_i\}_{i=0}^{n}$, we have the telescoping decomposition.

Telescoping decomposition

Let $\{D_i\}_{i=1}^{\infty}$ be a MDS. Then for any martingale sequence $\{Y_i\}_{i=0}^{\infty}$, we have the telescoping decomposition

$$Y_n - Y_0 = \sum_{i=1}^{n} D_i$$

Example (Doob construction)

Consider the sequence on independent r.v.s $X_1, \cdots, X_n$, recall the sequence $Y_k = \mathbb{E}[f(X) \mid X_1, \cdots, X_k]$ previously defined, and suppose that $\mathbb{E}[|f(X)|] < \infty$. We claim that $Y_0, \cdots, Y_n$ is a martingale w.r.t to $X_1, \cdots, X_n$. We have

$$\mathbb{E}[|Y_k|] = \mathbb{E}[|\mathbb{E}[f(X) \mid X_1, \cdots, X_k]|].$$

From Jensen's inequality, we have

$$\mathbb{E}[|\mathbb{E}[f(X) \mid X_1, \cdots, X_k]|] \leq \mathbb{E}[|f(X)|] < \infty.$$

From the 2^{nd} property of martingales, we have

$$\mathbb{E}[Y_{k+1} \mid X_1^k] = \mathbb{E}[\mathbb{E}[f(X) \mid X_1^{k+1}] \mid X_1^k] = \mathbb{E}[f(X) \mid X_1^k] = Y_k$$

We now turn to the derivation of concentration inequalities for martingales, either

1) as bounds for the difference $Y_n - Y_0$; or
2) as bounds for the sum $\sum_{i=1}^n D_i$ of the associated MDS.

We begin by stating and proving a general Bernstein-type bound for a MDS, based on imposing a sub-exponential condition on the MDS. To do so, we adopt the standard approach of controlling the mgf of $\sum_{i=1}^n D_i$ and then applying the Chernoff bound. Assume that $\mathbb{E}[\exp(\lambda D_i) \mid \mathcal{F}_{i-1}] \leq \exp\left(\frac{\lambda^2 \nu_i^2}{2}\right)$ a.s. for any $|\lambda| < \frac{1}{\alpha_i}$

$$\begin{aligned}
\mathbb{E}[\exp(\lambda \sum_{i=1}^n D_i)] &= \mathbb{E}[\mathbb{E}[\exp(\lambda \sum_{i=1}^n D_i) \mid \mathcal{F}_{n-1}]] \\
&= \mathbb{E}[\mathbb{E}[\exp(\lambda D_n) \exp(\lambda \sum_{i=1}^{n-1} D_i) \mid \mathcal{F}_{n-1}]] \\
&= \mathbb{E}[\exp(\lambda \sum_{i=1}^{n-1} D_i) \mathbb{E}[\exp(\lambda D_n) \mid \mathcal{F}_{n-1}]]
\end{aligned}$$

$$\leq \mathbb{E}[\exp(\lambda \sum_{i=1}^{n-1} D_i)] \exp\left(\frac{\lambda^2 \nu_n^2}{2}\right)$$

we may iterate this procedure again for $\mathbb{E}[\exp(\lambda \sum_{i=1}^{n-1} D_i)]$ and we'd obtain,

$$\begin{aligned}
\mathbb{E}[\exp(\lambda \sum_{i=1}^{n-1} D_i)] &= \mathbb{E}[\mathbb{E}[\exp(\lambda \sum_{i=1}^{n-1} D_i) \mid \mathcal{F}_{n-2}]] \\
&= \mathbb{E}[\mathbb{E}[\exp(\lambda D_{n-1}) \exp(\lambda \sum_{i=1}^{n-2} D_i) \mid \mathcal{F}_{n-2}]] \\
&= \mathbb{E}[\exp(\lambda \sum_{i=1}^{n-2} D_i) \mathbb{E}[\exp(\lambda D_{n-1}) \mid \mathcal{F}_{n-2}]] \\
&\leq \mathbb{E}[\exp(\lambda \sum_{i=1}^{n-2} D_i)] \exp\left(\frac{\lambda^2 \nu_{n-1}^2}{2}\right)
\end{aligned}$$

Continuously iterating this process yields,

$$\mathbb{E}[\exp(\lambda \sum_{i=1}^{n} D_i)] \leq \exp\left(\frac{\lambda^2 \sum_{i=1}^{n} \nu_i^2}{2}\right),$$

valid for all $|\lambda| < \frac{1}{\alpha^*}$. Hence, by definition, it can be concluded that $\sum_{i=1}^{n} D_i$ is sub-exponential with parameters $(\sqrt{\sum_{i=1}^{n} \nu_i^2}, \alpha^*)$. The tail bounds can be derived by using the Chernoff-type approach as before.

Concentration inequalities for MDS

Let $\{(D_i, \mathcal{F}_i)\}_{i=1}^{\infty}$ be a MDS, and suppose that $\mathbb{E}[\exp(\lambda D_i) \mid \mathcal{F}_{i-1}] \leq \frac{\lambda^2 \nu_i^2}{2}$ a.s. for any $|\lambda| < \frac{1}{\alpha}$. Then the following hold

- The sum $\sum_{i=1}^n D_i$ is sub-exponential with parameters $\left(\sqrt{\sum_{i=1}^n \nu_i^2}, \alpha^*\right)$, where $\alpha^* := \max_{i=1,\cdots,n} \alpha_i$.
- The sum satisfies the concentration inequality

$$P\left[|\sum_{i=1}^n D_i| \geq t\right] \leq \begin{cases} 2\exp\left(-\frac{t^2}{2\sum_{i=1}^n \nu_i^2}\right), & 0 \leq t < \frac{\sum_{i=1}^n \nu_i^2}{\alpha^*} \\ 2\exp\left(-\frac{t}{2\alpha^*}\right), & t > \frac{\sum_{i=1}^n \nu_i^2}{\alpha^*} \end{cases}$$

For the concentration inequalities to be useful in practice, we must isolate sufficient easily checkable conditions for the differences D_i to be a.s. sub-exponential (or sub-Gaussian when $\alpha = 0$). As mentioned earlier, bounded r.v.s are sub-Gaussian, which leads to the following corollary

Azuma-Hoeffding

Let $(\{D_i, \mathcal{F}_i\}_{i=1}^n)$ be a MDS for which there are constants $\{a_i, b_i\}_{i=1}^n$ such that $D_i \in [a_i, b_i]$ a.s. for all $k = 1, \cdots, n$. Then for all $t \geq 0$

$$P\left[|\sum_{i=1}^{n} D_i| \geq t\right] \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

Proof: All that needs showing is that the $\mathbb{E}[\exp(\lambda D_i \mid \mathcal{F}_{i-1})] \leq \exp\left(\frac{\lambda^2(b_i - a_i)^2}{8}\right)$ a.s. for each $i = 1, \cdots, n$. But since $D_i \in [a_i, b_i]$ a.s., the conditioned variables $(D_i \mid \mathcal{F}_{i-1})$ also belongs to this interval a.s.

Bounded differences property

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

$$x^{\backslash k} := (x_1, x_2, \cdots, x_{k-1}, x'_k, x_{k+1}, \cdots, x_n)'.$$

We say that $f : \mathbb{R}^n \to \mathbb{R}$ satisfies the bounded difference property with parameters $(L_1, \cdots, L_n)$ if, for each $k = 1, 2, \cdots, n$,

$$|f(x) - f(x^{\backslash k})| \leq L_k \quad \forall x, x' \in \mathbb{R}^n$$

Bounded differences inequality

Suppose that f satisfies the bounded difference property with parameters $(L_1, \cdots, L_n)$ and that the random vector $X = (X_1, X_2, \cdots, X_n)'$ has independent components. Then

$$P[|f(X) - \mathbb{E}[f(X)] \geq t|] \leq 2 \exp\left(-\frac{2t^2}{\sum_{i=1}^{n} L_k^2}\right), \quad \forall t \geq 0$$

Example

Say we have the bounded r.v.s $X_i \in [a, b]$ almost surely, and consider the function $f(x_1, \cdots, x_n) = \sum_{i=1}^{n}(x_i - \mu_i)$, where $\mu_i = \mathbb{E}[X_i]$ is the mean of the i^{th} rv. For any index $l \in \{1, \cdots, n\}$, we have

$$\begin{aligned} |f(x) - f(x^{\backslash k})| &= |(x_k - \mu_k) - (x'_k - \mu_k)| \\ &= |x_k - x'_k| \leq b - a \end{aligned}$$

which shows that f satisfies the bounded difference inequality in each coordinate with parameter $L = b - a$. Consequently, from the bounded inequality it follows

$$P\left[|\sum_{i=1}^{n}(x_i - \mu_i)| \geq t\right] \leq 2\exp\left(-\frac{2t^2}{n(b-a)^2}\right)$$

which is classical Hoeffding bound for independent r.v.s.

L-Lipschitz functions

We say a function $f : \mathbb{R}^n \to \mathbb{R}$ is L-Lipschitz w.r.t the Euclidean norm $\|.\|_2$ if

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

The following guarantees that any such function is sub-Gaussian with parameter at most L:

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

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

The earlier result is of great importance, as it guarantees that any L-Lipschitz function of a standard Gaussian random vector, regardless of

the dimension, exhibits concentration like a scalar Gaussian variable with variance L^2.
Any Lipschitz function is differentiable almost everywhere and the Lipschitz property further guarantees $\|\nabla f(x)\|_2 \leq L$ for all $x \in \mathbb{R}^n$. Therefore, to prove the earlier results, we first begin by providing the following Lemma:

Lemma

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

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

where $X, Y \sim N(0, I_n)$ are standard multivariate and independent.

Proof: For any fixed $\lambda \in \mathbb{R}$ applying the inequality in above Lemma to the convex function $f : t \to \exp(\lambda t)$ yields

$$\mathbb{E}_X[\exp(\lambda\{f(X) - \mathbb{E}[f(X)]\})] \leq \mathbb{E}_{X,Y}\left[\exp\left(\frac{\pi}{2}\langle \nabla f(X), Y\rangle\right)\right]$$

References

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.