Basic tail and concentration bounds (Part I)

Kaveh S. Nobari

Reading Sessions 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) and chapters 1 and 2 of Vershynin (2018) to shed light on elementary techniques for obtaining deviation and concentration inequalities.

Concentration inequalities generally take the form of

$$P[|X - \mu| \geq t] \leq \textit{something small}$$

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.

Motivation

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, we conclude this 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

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

Proof: For $t \geq 0$, we can express any real number x, as

$$x = x\mathbb{1}\{x \geq t\} + x\mathbb{1}\{x < t\}$$

Substituting the random variable X for x and taking expectations from both sides, we obtain

$$\begin{aligned}\mathbb{E}[X] &= \mathbb{E}[X\mathbb{1}\{X \geq t\}] + \mathbb{E}[X\mathbb{1}\{X < t\}] \\ &\geq \mathbb{E}[t\mathbb{1}\{X \geq t\}] + 0 \\ &\geq tP[X \geq t]\end{aligned}$$

Thus, dividing both sides by t yields Markov's inequality.

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

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

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

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

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

Example: Toss a fair coin N times. What is the probability that we get at least $\frac{3N}{4}$ heads?

Let us express the outcome of the coin toss as a random variable X that takes values in $\{1, 0\}$, that signify head and tail respectively. We know that $P[X = 1] = P[X = 0] = 1/2$, since it is a fair coin; thus, $X \sim binom(1, 0.5)$. Now we are interested in the sum $S_N = X_1 + X_2 + \cdots + X_N$, where S_N follows a $S_N \sim binom(N, 0.5)$ with $\mathbb{E}[S_N] = Np$ and $Var(S_N) = Np(1-p)$. We are interested in the probability

$$P\left[S_n \geq \frac{3N}{4}\right].$$

We know that

$$\mathbb{E}[S_N] = \frac{N}{2} \quad \text{and} \quad Var(S_N) = \frac{N}{4}$$

Thus, from Chebyshev's inequality, we can show that

$$P\left[\left|S_n - \frac{N}{2}\right| \geq t\right] \leq \frac{N}{4t}$$

Thus, substituting $N/4$ for t, we get

$$P\left[\left|S_n - \frac{N}{2}\right| \geq \frac{N}{4}\right] \leq \frac{4}{N}$$

So the probability converges to zero at least linearly in N. 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 moment generating function 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

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

1. Motivation

2. Classical bounds
   - From Markov to Chernoff
   - Sub-Gaussian variables and Hoeffding bounds
   - Sub-exponential variables and Bernstein bounds

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

$$\begin{aligned}\mathbb{E}[\exp(\lambda X)] &= \int_{\mathbb{R}} \exp(\lambda x) f(x) dx \\ &= \int_{\mathbb{R}} \exp(\lambda x) \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(X-\mu)^2}{2\sigma^2}\right) dx \\ &= \exp\left(\mu\lambda + \frac{\sigma^2\lambda^2}{2}\right), \quad \forall \lambda \in \mathbb{R}\end{aligned}$$

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

$$\inf_{\lambda \geq 0}\{\log \mathbb{E}[\exp(\lambda(X - \mu)) - \lambda t]\} = \inf_{\lambda \geq 0}\left\{\frac{\sigma^2\lambda^2}{2} - \lambda t\right\} = -\frac{t^2}{2\sigma^2} \tag{1}$$

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

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

Proof of equation (1).

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 (1). □

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, when the above condition is combined with Chernoff bound, as in the Gaussian example, with $x \sim N(\mu, \sigma^2)$, it then satisfies the upper deviation inequality

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

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

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

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

Some preliminaries: By definition, a function g is convex if

$$g(\lambda x + (1-\lambda)y) \leq \lambda g(x) + (1-\lambda)g(y), \quad \forall \lambda \in [0,1]$$

For a convex function $g : \mathbb{R} \rightarrow \mathbb{R}$, Jensen's inequality 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 (2). 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

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

Exercise 2.4 of Wainwright (2019): Consider a r.v. X with mean $\mu = \mathbb{E}[X]$, and such that for scalars $b > a$, $X \in [a, b]$ almost surely.

1) Defining the function $\psi(\lambda) = \log \mathbb{E}[\exp(\lambda X)]$, show that $\psi(0) = 0$ and $\psi'(0) = \mu$.
2) Show that $\psi''(\lambda) = \mathbb{E}_\lambda[X^2] - (\mathbb{E}_\lambda[X])^2$, where we define $\mathbb{E}_\lambda[f(X)] := \frac{\mathbb{E}[f(X)\exp(\lambda X)]}{\mathbb{E}[\exp(\lambda X)]}$. Use this fact to obtain an upper bound on $\sup_{\lambda \in \mathbb{R}} |\psi''(\lambda)|$.
3) Use parts (a) and (b) to establish that X is sub-Gaussian with parameter at most $\sigma = \frac{b-a}{2}$.

Solution:

Part 1:

$$\psi(0) = \log \mathbb{E}[\exp(0)] = \log 1 = 0$$

For $\psi'(0)$, we know that the derivative of the m.g.f equals μ, so

$$\psi'(0) = \frac{\mu}{\mathbb{E}[\exp(0X)]} = \mu.$$

Part 2:

The identity for $\psi''(\lambda)$ follows from the chain rule. For the upper bound, observe that we can define a new distribution Q_λ by taking $\exp(\lambda X)/\mathbb{E}[\exp(\lambda X)]$ to be its Radon–Nikodym derivative (density) with respect to the distribution of X. Hence establishing a bound on $\psi''(\lambda)$ is equivalent to bounding the supremum over variances of random variables $X_\lambda \sim Q_\lambda$.
Taking, $m := \frac{1}{2}(a+b)$, using that the mean minimises the mean squared error, and using that $X_\lambda \in [a,b]$ a.s. for all λ,

$$\sup_\lambda Var(X_\lambda) = \sup_\lambda \mathbb{E}[(X - \mathbb{E}[X_\lambda])^2] \leq \sup_\lambda \mathbb{E}[(X-m)^2] \leq (b-m)^2 = \frac{(b-a)^2}{4}.$$

Part 2:
Taking a Maclaurin expansion of $\psi(\lambda)$

$$\psi(\lambda) = \psi(0) + \lambda\psi'(0) + \frac{\lambda^2}{2}\psi''(\varepsilon),$$

for some $\varepsilon \in (0,\lambda)$. Substituting the results from 1 and 2 ,

$$\psi(\lambda) \leq \lambda\mu + \frac{\lambda^2}{2}\frac{(b-a)^2}{5}$$

as desired.

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

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

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, 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 Mill'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

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

Equivalent characterizations of the sub-Gaussian variables (III-IV)

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

$$\mathbb{E}\left[X^{2k}\right] \leq \frac{(2k)!}{2^k k!}\theta^{2k}, \quad \forall k = 1, 2, \cdots$$

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

$$\mathbb{E}\left[\exp\left(\frac{\lambda X^2}{2\sigma^2}\right)\right] \leq \frac{1}{\sqrt{1-\lambda}}, \quad \forall \lambda \in [0, 1).$$

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

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

References

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

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