A Short Introduction to Concentration Inequalities

Soon Hoe Lim SOON.HOE.LIM@SU.SE
Nordita
KTH Royal Institute of Technology and Stockholm University
Stockholm 106 91, Sweden

Abstract

In this short note, we give a short introduction to concentration inequalities, which are parts of a general phenomenon called concentration of measure. The non-asymptotic nature of these inequalities makes them attractive for applications in statistical physics and data science. The main references for our purpose are Vershynin (2018) and Wainwright (2019). Other useful references for further reading include Rigollet (2015); van Handel (2014); Ledoux (2001); Boucheron et al. (2013); Tao (2012); Bakry et al. (2013); Durrett (2019); Grimmett and Stirzaker (2001); Blum et al. (2020); Mohri et al. (2018); Shalev-Shwartz and Ben-David (2014).

1. Appetizer: Random Variables and Scaling

We are interested in asymptotic and non-asymptotic behavior of a sequence of random variables (RVs) or sequence of functionals of these RVs. Different scaling of such RVs leads to different behavior. We will illustrate this via a series of examples concerning deviations of sample mean of i.i.d. RVs from the true mean.

Take $(\Omega, \mathcal{F}, \mathbb{P})$ as the probability space and denote expectation by $\mathbb{E}$. Let $S_n = X_1 + \cdots + X_n$, where the X_i are independent Rademacher RVs, i.e., $\mathbb{P}[X_i = \pm 1] = 1/2$ for all i. Note that $\mathbb{E}S_n = 0$ and $\mathbb{E}X_1^2 = 1$. Then:

Example 1.1 *(No scale and non-asymptotic)*
By Hoeffding's inequality, we have

$$\mathbb{P}[S_n \geq \delta] \leq e^{-\delta^2/2n} \tag{1}$$

for all $\delta > 0$. This says that the tail of the sequence (S_n) behaves similarly to the tail of the normal distribution. To show Eqn. (1) *directly, let $\delta > 0$ and $\lambda \in \mathbb{R}$ and compute:*

$$\begin{aligned}
\mathbb{P}[S_n \geq \delta] &= \mathbb{P}[e^{\lambda S_n} \geq e^{\lambda\delta}] && (2)\\
&\leq e^{-\lambda\delta}\mathbb{E}e^{\lambda S_n} && \textit{(by Markov's inequality)} \quad (3)\\
&= e^{-\lambda\delta}(\mathbb{E}e^{\lambda X_1})^n && \textit{(by independence of the } X_i\textit{)} \quad (4)\\
&= e^{-\lambda\delta}\cosh^n(\lambda) && (5)\\
&\leq e^{-\lambda\delta+n\lambda^2/2} && \textit{(using } \cosh(x) \leq e^{x^2/2} \textit{ for } x \in \mathbb{R}\textit{)} \quad (6)\\
&\leq e^{-\delta^2/2n} && \textit{(optimizing over } \lambda\textit{)} \quad (7)
\end{aligned}$$

Two-sided version of the inequality follows by applying the above inequality for $-X_i$ instead of X_i to get the same bound for $\mathbb{P}[-S_n \geq \delta]$. Then, $\mathbb{P}[|S_n| \geq \delta] = \mathbb{P}[S_n \geq \delta] + \mathbb{P}[-S_n \geq \delta] \leq 2e^{-\delta^2/2n}$.

Last Updated: December 24, 2020

Example 1.2 *(Scale of "large deviation")*
Replace δ by $n\delta$ in the two-sided version of inequality (1)*:*

$$\mathbb{P}[|S_n| \geq n\delta] \leq 2e^{-n\delta^2/2} = 2e^{-nI(\delta)} \tag{8}$$

for all $\delta > 0$, where $I(\delta) = \delta^2/2$ (rate function). Thus,

$$\lim_{n\to\infty} \frac{1}{n} \ln \mathbb{P}[S_n \geq n\delta] = -I(\delta). \tag{9}$$

for all $\delta > 0$. This is an example of a large deviation statement and it is a special case of Cramér's theorem.

Example 1.3 *(Scale of "large deviation" $\implies$ Strong Law of Large Numbers (SLNN))*
Eqn. (8) *implies:*

$$\sum_{n\geq 1} \mathbb{P}\left[\left|\frac{S_n}{n}\right| \geq \delta\right] \leq 2\sum_{n\geq 1} e^{-n\delta^2/2} < \infty \tag{10}$$

for all $\delta > 0$. Therefore, by the first Borel-Cantelli lemma,

$$\mathbb{P}\left[\left|\frac{S_n}{n}\right| \geq \delta \quad i.o.\right] = 0 \tag{11}$$

for all $\delta > 0$. Thus, $S_n/n \to 0$ almost surely (a.s.) as $n \to \infty$. This is precisely the SLLN.

Example 1.4 *(Scale of the central limit theorem (CLT))*
Replace δ by $\sqrt{n}\delta$ in the two-sided version of inequality (1)*:*

$$\mathbb{P}[|S_n| \geq \sqrt{n}\delta] \leq 2e^{-\delta^2/2} \tag{12}$$

for all $\delta > 0$. This is not quite the CLT, which says that:

$$\lim_{n\to\infty} \mathbb{P}[|S_n| \geq \sqrt{n}\delta] = \frac{2}{\sqrt{2\pi}} \int_\delta^\infty e^{-x^2/2} dx \leq \frac{2}{\sqrt{2\pi}\delta} e^{-\delta^2/2} \tag{13}$$

for all $\delta > 0$. However, the key observation here is that the right hand side in Eqn. (12) *is independent of n.*

A tighter bound is provided by Berry-Esseen: for any $\delta > 0$,

$$\mathbb{P}[|S_n| \geq \sqrt{n}\delta] \leq \frac{2}{\sqrt{2\pi}\delta} e^{-\delta^2/2} + \frac{2C}{\sqrt{n}}, \tag{14}$$

for some constant $C > 0$.

How will the above results change if the X_i are other RVs? What if the i.i.d. assumption is relaxed? What if we have a sequence of random processes instead? We will address these questions later.

Remark 1.1 *On the other hand, the Weak Law of Large Numbers (WLNN) says that $S_n/n \to 0$ in probability, as $n \to \infty$. One may wonder if there is a scaling a_n such that $S_n/a_n \to 0$ in probability but not almost surely as $n \to \infty$. The answer is provided by the law of iterated logarithm (LIL), a fine intermediate result that shows what happens in between the scales of LLN and CLT. The LIL tells us exactly how large the fluctuations suffered by the sequence $(S_n/\sqrt{n})$ are on its route towards the normal distribution. The scaling of LIL is given by $a_n = \sqrt{2n \log\log n}$ and one has:*

$$\limsup_{n\to\infty} \frac{S_n}{a_n} \to 1 \quad \textit{and} \quad \liminf_{n\to\infty} \frac{S_n}{a_n} \to -1, \quad \textit{a.s.}. \tag{15}$$

So, $S_n/a_n \to 1$ a.s. (or with probability one) but $S_n/a_n \to 0$ in probability, as $n \to \infty$.

In particular, LIL implies the following guarantee on the deviations of the sample mean as an estimator of the true mean: with probability one (w.p.1), there is a $n_0 \in \mathbb{N}$ such that $|S_n/n| \leq (2 \log \log n/n)^{1/2}$ for every $n \geq n_0$.

As shown in the last sentence of Remark 5, the applicability of the LIL is limited since it does not tell us how large n should be for a given deviation, i.e., we would like to have a workable expression for n_0. This limitation can be lifted by exploiting concentration inequalities, such as the Hoeffding's inequality in Example 1, that can be seen as non-asymptotic versions of the CLT. These inequalities not only confirm our intuition that sample mean should concentrate tightly about the true mean but also quantify how a RV X fluctuates around its mean μ by providing bounds for the tails of $X - \mu$, such as

$$\mathbb{P}[|X - \mu| > t] \leq \text{something small.} \tag{16}$$

Moreover, unlike the classical limit theorems such inequalities are *non-asymptotic* (i.e., they hold for all fixed N as opposed to $N \to \infty$). The non-asymptotic nature of such inequalities makes them attractive for applications in statistical physics (where N often corresponds to **system size**[1]) and data science (where N often corresponds to **sample size**). Concentration inequalities are parts of a general phenomenon called *concentration of measure* that we will explore more later. On the other hand, there are anti-concentration phenomena, as illustrated by the following example, that we will not study here.

Example 1.5 *(Anti-concentration of Gaussian distribution) Let $X \sim N(0, \sigma^2)$, then*

$$\mathbb{P}[|X| \leq t] \in \left(\frac{2}{3} \frac{t}{\sigma}, \frac{4}{5} \frac{t}{\sigma} \right). \tag{17}$$

For completeness, we state the classical limit theorems (SLLN, WLLN and CLT) and a non-asymptotic bound (Berry-Esseen) for general i.i.d. RVs (not just Rademacher as in the above examples) without proof in the following. Complete proof can be found in standard probability textbooks Durrett (2019); Grimmett and Stirzaker (2001).

Let $S_n = X_1 + \cdots + X_n$, where the X_i are i.i.d. RVs with mean μ, variance σ^2, characteristic function ϕ and CDF F. Then:

Theorem 1.1 *(Classical limit theorems for sum of i.i.d. RVs)*
(a) (SLLN)

$$S_n/n \to \mu \quad a.s.. \tag{18}$$

(b) (WLLN) The following three statements are equivalent.
(i) ϕ is differentiable at 0, and $\phi'(0) = i\mu$.
(ii) As $t \to \infty$, we have $t[1 - F(t) + F(-t)] \to 0$ and $\int_{-t}^{t} xF(dx) \to \mu$.
(iii) $S_n/n \to \mu$, in probability, as $n \to \infty$.
(c) (CLT) As $n \to \infty$,

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \to N(0, 1), \quad \text{in distribution.} \tag{19}$$

Theorem 1.2 *(Berry-Esseen theorem) Assume that $\mathbb{E}|X_i|^3 < \infty$. Then:*

$$\sup_{t \in \mathbb{R}} \left| \mathbb{P}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq t \right) - \Phi(t) \right| \leq \frac{C\mathbb{E}|X_i|^3}{\sigma^3\sqrt{n}}, \tag{20}$$

where $0.41 \leq C \leq 0.4748$ is a universal constant and $\Phi(t)$ is the CDF of $N(0, 1)$ RV.

1. See also: http://www.cgogolin.de/downloads/meascons.beamer.pdf

If we assume more, such as the RV being bounded, or having a bounded moment generating function (MGF), we can prove many "tail" inequalities, such as Hoeffding's and others (Bennett's, Bernstein's etc.).

Remark 1.2 *(Looking ahead) One way to prove CLT is to use a stability argument (Lindeberg's swapping trick – a perturbation in a (normalized) sum by a random variable with matching first and second moments does not change the (normalized) sum distribution in the limit) – see Tao (2012). It turns out that stability argument lies at the heart of key results in random matrix theory and statistical learning theory, as we shall see later.*

2. Concentration Inequalities

For what class of RVs does a concentration inequality like Hoeffding's hold? It turns out that such RVs must have sub-Gaussian tails.

Definition 2.1 *A RV X with $\mu = \mathbb{E}X$ is sub-Gaussian if there exists $\sigma > 0$ such that $\mathbb{E}e^{\lambda(X-\mu)} \leq e^{\sigma^2\lambda^2/2}$ for all $\lambda \in \mathbb{R}$. In this case, σ is called the sub-Gaussian parameter.*

A sub-Gaussian RV with parameter σ satisfies the concentration inequality:

$$\mathbb{P}[|X - \mu| \geq t] \leq 2e^{-t^2/2\sigma^2}, \tag{21}$$

for all $t \in \mathbb{R}$ (check this!).

Clearly, any Gaussian RV with positive variance σ^2 is sub-Gaussian with parameter σ. From the computation in Example 1, we see that Rademacher RVs (and, more generally, Bernoulli RVs) are sub-Gaussian. Moreover:

Proposition 2.1 *Any bounded RV is sub-Gaussian.*

Proof The key idea of the proof is to use a symmetrization argument.

WLOG, let $\mu = \mathbb{E}_X X = 0$ and take the support of X to be the interval $[a, b]$.

Let X' be an independent copy of X and ϵ be an independent Rademacher RV. Recall $E_\epsilon[e^{\lambda\epsilon}] \leq e^{\lambda^2/2}$. Then:

$$\begin{aligned}
\mathbb{E}_X e^{\lambda X} &= \mathbb{E}_X[e^{\lambda(X - \mathbb{E}_{X'}X')}] && (22)\\
&\leq E_{X,X'}[e^{\lambda(X-X')}] \qquad \text{(by Jensen's inequality)} && (23)\\
&= \mathbb{E}_{X,X'}[\mathbb{E}_\epsilon[e^{\epsilon\lambda(X-X')}]] && (24)\\
&\leq E_{X,X'}e^{\lambda^2(X-X')^2/2} \leq e^{\lambda^2(b-a)^2/2}. && (25)
\end{aligned}$$

Therefore, X is sub-Gaussian with parameter at most $\sigma = b-a$ (one can sharpen this to $\sigma = (b-a)/2$: try this!). ■

Working with bounded RVs, which we know are sub-Gaussian by now, is a standard trick (the truncation method Tao (2012)) in proving results such as LLN and CLT.

On the other hand, Poisson, exponential, Pareto and Cauchy distributions are not sub-Gaussian (check this!).

We have a generalization of (1) to sum of independent sub-Gaussian RVs.

Proposition 2.2 *(Hoeffding bound for sum of independent sub-Gaussian RVs) Let $S_n = X_1 + \cdots + X_n$, where the X_i are independent mean zero sub-Gaussian RVs with parameter σ_i. Then, for all $t \geq 0$,*

$$\mathbb{P}[S_n \geq t] \leq \exp\left(-\frac{t^2}{2\|\boldsymbol{\sigma}\|_2^2}\right), \tag{26}$$

where $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_n)$ and $\|\boldsymbol{\sigma}\|_2^2 = \sum_{i=1}^n \sigma_i^2$.

Proof Show that S_n is sub-Gaussian with parameter $\|\boldsymbol{\sigma}\|_2$ (try this!) ■

It turns out that there are relations among the tails, MGF and the L^p norms for sub-Gaussian RVs.

Proposition 2.3 *(Sub-Gaussian properties) Let X be a RV. Then the following are equivalent (the parameters $K_i > 0$ below differ from each other by at most an absolute constant factor).*

(i) The tails of X satisfy:

$$\mathbb{P}[|X| \geq t] \leq 2\exp(-t^2/K_1^2), \quad \text{for all } t \geq 0. \tag{27}$$

(ii) The moments of X satisfy

$$\|X\|_{L^p} = (\mathbb{E}|X|^p)^{1/p} \leq K_2\sqrt{p}, \quad \text{for all } p \geq 1. \tag{28}$$

(iii) The MGF of X^2 satisfies:

$$\mathbb{E}\exp(\lambda^2 X^2) \leq \exp(K_3^2\lambda^2), \quad \text{for all } \lambda \text{ such that } |\lambda| \leq 1/K_3 \tag{29}$$

(iv) the MGF of X^2 is bounded at some point, i.e.,

$$\mathbb{E}\exp(X^2/K_4^2) \leq C \tag{30}$$

where $C = 2$ (in general, C can be any constant > 1).

Moreover, if $\mathbb{E}X = 0$ then (i)-(iv) are also equivalent to:
(v) The MGF of X satisfies:

$$\mathbb{E}\exp(\lambda X) \leq \exp(K_5^2\lambda^2) \quad \text{for all } \lambda \in \mathbb{R}. \tag{31}$$

Proof See Proposition 2.5.2 in Vershynin (2018). ■

Although the class of sub-Gaussian distributions is natural and quite large, it leaves out some important distributions whose tails are heavier than Gaussians. For instance, in applications one is often interested in functionals like X_1^2 or $\|X\|_2^2 := \sum_{i=1}^N X_i^2$ (Euclidean norm), where the X_i are independent sub-Gaussian RVs. However, the X_i^2 are not sub-Gaussian. In fact,

$$\mathbb{P}[X_i^2 > t] = \mathbb{P}[|X_i| > \sqrt{t}] \sim \exp(-ct) \tag{32}$$

for some constant $c > 0$. This motivates us to look at distributions that have at least an exponential tail decay and prove an analog of Hoeffding's inequality for them. More generally, one can look at distributions that have tail decay of the form $\exp(-ct^\alpha)$ for $\alpha > 0$ or heavy tail (i.e., the MGF $\mathbb{E}e^{\lambda X}$ is infinite for all $\lambda > 0$) but we will not pursue the general theory here.

Definition 2.2 *A RV with $\mu = \mathbb{E}X$ is sub-exponential if there exists (ν, α), $\nu, \alpha \geq 0$, such that $\mathbb{E}e^{\lambda(X-\mu)} \leq e^{\nu^2\lambda^2/2}$ for all $|\lambda| \leq 1/\alpha$.*

In the above we adopt the definition in Wainwright (2019), which is slightly more general than the one considered in Vershynin (2018) (setting $\mu = 0$, $K_5 = \nu/\sqrt{2} = \alpha$ in Proposition 2.7.1 there gives us other useful sub-exponential properties).

Proposition 2.4 *(Sub-exponential properties) Let X be a zero mean RV. Then the following are equivalent.*

(i) The tails of X satisfy: there exists constants $c_1, c_2 > 0$ such that

$$\mathbb{P}[|X| \geq t] \leq c_1 e^{-c_2 t}, \quad \text{for all} t > 0. \tag{33}$$

(ii) There exists $c_0 > 0$ such that $\mathbb{E}[e^{\lambda X}] < \infty$ for all $|\lambda| \leq c_0$.

(iii) $\gamma := \sup_{k\geq 2}(\mathbb{E}[X^k]/k!)^{1/k}$ is finite.

(iv) There exists (ν, α), $\nu, \alpha \geq 0$, such that $\mathbb{E}e^{\lambda X} \leq e^{\nu^2\lambda^2/2}$ for all $|\lambda| < 1/\alpha$.

Proof See Wainwright (2019) or Vershynin (2018). ■

Clearly, any sub-Gaussian RVs with positive parameter σ are sub-exponential (with $\nu = \sigma > 0$ and $\alpha = 0$). Products (such as squares) of sub-Gaussian RVs are sub-exponential. Other examples include the exponential and Poisson distributions (check this!).

Example 2.1 *(Sub-exponential but not sub-Gaussian) Let $Z \sim N(0, 1)$ and consider $X = Z^2$. Then $\mathbb{E}X = VarZ = 1$ and*

$$\mathbb{E}\exp(\lambda(X-1)) = \begin{cases} \frac{e^{-\lambda}}{\sqrt{1-2\lambda}} & \text{if } \lambda < 1/2 \\ \infty & \text{otherwise.} \end{cases}$$

This shows that X is not sub-Gaussian. Since $\frac{e^{-\lambda}}{\sqrt{1-2\lambda}} \leq e^{2\lambda^2}$ for $|\lambda| < 1/4$, X is sub-exponential with $(\nu, \alpha) = (2, 4)$.

Proposition 2.5 *(Sub-exponential tail bound) Let X be a sub-exponential RV with parameters (ν, α). Then,*

$$\mathbb{P}[X - \mu \geq t] \leq \begin{cases} \exp(-t^2/2\nu^2) & \text{if } 0 \leq t \leq \nu^2/\alpha \\ \exp(-t/2\alpha) & \text{if } t > \nu^2/\alpha. \end{cases}$$

Proof WLOG, assume $\mu = 0$. We have, proceeding as in the Chernoff-type approach:

$$\mathbb{P}[X \geq t] \leq e^{-\lambda t}\mathbb{E}[e^{\lambda X}] \leq e^{-\lambda t + \lambda^2\nu^2/2} =: e^{g(\lambda,t)} \tag{34}$$

for all $\lambda \in [0, 1/\alpha)$.

It remains to compute, for a fixed $t \geq 0$, $g^*(t) := \inf_{\lambda\in[0,1/\alpha)} g(\lambda, t)$. There is an unconstrained minimum at $\lambda^* = t/\nu^2$.

Now, if $t \in [0, \nu^2/\alpha)$, then the unconstrained minimum is also the constrained minimum and so $g^*(t) = -t^2/2\nu^2$. Otherwise, we may assume that $t \geq \nu^2/\alpha$. Since $g(\cdot, t)$ is monotonically decreasing in $[0, \lambda^*)$, the constrained minimum is achieved at the boundary point $\lambda^+ = 1/\alpha$. So, $g^*(t) = -t/\alpha + \nu^2/2\alpha^2 \leq -t/2\alpha$. The proof is done. ■

An alternative way to verify sub-exponential property is by controlling the polynomial moments of the RV.

Definition 2.3 *Given a RV X, wa say that Bernstein's condition with parameter b holds if*

$$|\mathbb{E}[(X-\mu)^k]| \le \frac{1}{2}k!\sigma^2 b^{k-2}, \quad \text{for } k=2,3,4,\ldots. \tag{35}$$

For example, if X is a bounded RV and, in particular, if $|X-\mu| \le b$, then X satisfies the above Bernstein's condition. In fact, any RV satisfying Bernstein's condition is sub-exponential. As we will see, this condition implies that the tail bounds are tighter than Hoeffding's.

Lemma 2.1 *If X is a RV satisfying Bernstein's condition* (35)*, then X is sub-exponential with parameter $(\sqrt{2}\sigma, 2b)$, where $\sigma^2 = \mathbb{E}[(X-\mu)^2]$.*

Proof

$$E[e^{\lambda(X-\mu)}] = 1 + \frac{\lambda^2\sigma^2}{2} + \sum_{k=3}^{\infty}\lambda^k\frac{\mathbb{E}[(X-\mu)^k]}{k!} \le 1 + \frac{\lambda^2\sigma^2}{2} + \frac{\lambda^2\sigma^2}{2}\sum_{k=3}^{\infty}(|\lambda|b)^{k-2}. \tag{36}$$

For $|\lambda| < \frac{1}{b}$, $\mathbb{E}[e^{\lambda(X-\mu)}] \le 1 + \frac{\lambda^2\sigma^2}{2(1-b|\lambda|)} \le \exp(\frac{\lambda^2\sigma^2}{2(1-b|\lambda|)})$, where we have used $1+t \le e^t$ for $t \ge 0$. Therefore, for $|\lambda| < \frac{1}{2b}$, we have $\mathbb{E}[e^{\lambda(X-\mu)}] \le \exp(\frac{\lambda^2}{2}\frac{(\sqrt{2}\sigma)^2}{2})$. ■

The following proposition gives tail bounds on RVs satisfying the Bernstein's condition (e.g., sub-exponential RVs).

Proposition 2.6 *(Bernstein-type bound) Let X be a RV satisfying the Bernstein condition* (35)*. Then, for $|\lambda| < 1/b$,*

$$\mathbb{E}[e^{\lambda(X-\mu)}] \le \exp\left(\frac{\lambda^2\sigma^2}{2(1-b|\lambda|)}\right). \tag{37}$$

Moreover, we have the concentration inequality:

$$\mathbb{P}[|X-\mu| \ge t] \le 2\exp\left(-\frac{t^2}{2(\sigma^2+bt)}\right) \tag{38}$$

for all $t \ge 0$.

Proof We have already shown the first statement in proof of the previous lemma. The second statement can be shown by using Hoeffding-type approach in Example 1 and (37) (try it!). ■

Sub-exponential property is preserved under summation for independent RVs. In particular:

Example 2.2 *Let X_i, $i = 1,\ldots,N$, be independent RVs, $\mathbb{E}X_i = \mu_i$ and sub-exponential with parameter (ν_k, α_k). Then, $\tilde{S}_N := \sum_{i=1}^{N}(X_i - \mu_i)$ is sub-exponential with parameter (ν_*, α_*), where $\alpha_* = \max_i \alpha_i$ and $\nu_*^2 = \sum_i \nu_i^2$. Moreover, for $t \ge 0$,*

$$\mathbb{P}\left[\frac{\tilde{S}_N}{N} \ge t\right] \le \exp\left(-\frac{1}{2}\min\left(\frac{N^2t^2}{\nu_*^2}, \frac{Nt}{\alpha_*}\right)\right). \tag{39}$$

Check all these! Also, one can derive a Bernstein-type bound for $\tilde{S}_N/N$ (do this!).

Then, setting $\mu_i = 0$, you should be able to obtain:

$$\mathbb{P}[S_N/\sqrt{N} \ge t] \le \begin{cases} 2\exp(-ct^2) & \text{if } t \le C\sqrt{N} \\ 2\exp(-t\sqrt{N}) & \text{if } t \ge C\sqrt{N}, \end{cases} \tag{40}$$

for some constants $c, C > 0$.

Therefore, in the small deviation regime where $t \leq C\sqrt{N}$*, we have a sub-Gaussian tail bound as if the sum had a normal distribution with constant variance. Note that this domain widens as* N *increases and the CLT becomes more powerful. For large deviations where* $t \geq C\sqrt{N}$*, the sum has a heavier, sub-exponential tail bound. In short, Bernstein's inequality for a sum of i.i.d. sub-exponential RVs gives a mixture of two tails:* **sub-Gaussian for small deviations and sub-exponential for large deviations**.

The Bernstein-type bound can be strengthened by Bennett's inequality. We will not cover Bennett's here – see Exercise 2.7 in Wainwright (2019) for details.

References

Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and Geometry of Markov Diffusion Operators*, volume 348. Springer Science & Business Media, 2013.

Avrim Blum, John Hopcroft, and Ravindran Kannan. *Foundations of Data Science*. Cambridge University Press, 2020.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Rick Durrett. *Probability: Theory and Examples*, volume 49. Cambridge University Press, 2019.

G. Grimmett and D. Stirzaker. *Probability and Random Processes*. OUP Oxford, 2001. ISBN 9780198572220.

Michel Ledoux. *The Concentration of Measure Phenomenon*. Number 89. American Mathematical Soc., 2001.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT press, 2018.

Philippe Rigollet. 18. s997: High dimensional statistics. *Lecture Notes, Cambridge, MA, USA: MIT Open-CourseWare*, 2015.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

Terence Tao. *Topics in Random Matrix Theory*, volume 132. American Mathematical Soc., 2012.

Ramon van Handel. Probability in high dimension. Technical report, Princeton University, 2014.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47. Cambridge University Press, 2018.

Martin J Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, volume 48. Cambridge University Press, 2019.