Lecture A.1: Basic Concentration Inequalities

Readings: Bach, *Learning Theory From First Principles* (Ch. 1), Vershynin, *High-Dimensional Probability* (Ch. 2), Wainwright, *High-Dimensional Statistics* (Ch. 2)

Topics: concentration of measure, sample mean deviations under different scalings, Hoeffding's inequality, sub-Gaussian random variables, sub-exponential random variables, Bernstein-type inequalities.

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, data science, and machine learning.

1 Appetizer: Random Variables and Scaling

We are interested in the asymptotic and non-asymptotic behavior of sequences of random variables, or sequences of functionals of these random variables. Different scalings of such random variables lead to different limiting behavior. We illustrate this through examples concerning deviations of the sample mean of i.i.d. random variables from the true mean.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and denote expectation by $\mathbb{E}$. Let

$$S_n = X_1 + \cdots + X_n,$$

where the X_i are independent Rademacher random variables, i.e.

$$\mathbb{P}[X_i = \pm 1] = \frac{1}{2}.$$

Note that $\mathbb{E}S_n = 0$ and $\mathbb{E}X_1^2 = 1$.

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 $\delta > 0$.

To show (1) directly, let $\delta > 0$ and $\lambda \in \mathbb{R}$. Then

$$\begin{aligned}
\mathbb{P}[S_n \geq \delta] &= \mathbb{P}[e^{\lambda S_n} \geq e^{\lambda\delta}] \\
&\leq e^{-\lambda\delta}\mathbb{E}e^{\lambda S_n} && \text{(Markov's inequality)} \\
&= e^{-\lambda\delta}\left(\mathbb{E}e^{\lambda X_1}\right)^n && \text{(independence)} \\
&= e^{-\lambda\delta}\cosh^n(\lambda) \\
&\leq e^{-\lambda\delta+n\lambda^2/2} && \text{(using } \cosh(x) \leq e^{x^2/2} \text{ for } x \in \mathbb{R}) \\
&\leq e^{-\delta^2/(2n)} && \text{(optimizing over } \lambda).
\end{aligned}$$

A two-sided version follows by applying the same inequality to $-X_i$:

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

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

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

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

Thus

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}[S_n \geq n\delta] = -I(\delta),$$

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

Example 1.3 (Large deviation scale implies the strong law)**.** From (2),

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

for all $\delta > 0$.

Therefore, by the first Borel–Cantelli lemma,

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

for all $\delta > 0$. Hence $S_n/n \to 0$ almost surely as $n \to \infty$, which is precisely the strong law of large numbers.

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

$$\mathbb{P}[|S_n| \geq \sqrt{n}\,\delta] \leq 2e^{-\delta^2/2} \tag{3}$$

for all $\delta > 0$.

This is not quite the central limit theorem, which says that

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

for all $\delta > 0$.

However, the key point is that the right-hand side in (3) is independent of n.

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

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

for some constant $C > 0$.

How do these results change if the X_i are other random variables? What if the i.i.d. assumption is relaxed? What if we have a sequence of random processes instead? We will return to these

questions later.

Remark 1.5. The weak law of large numbers says that $S_n/n \to 0$ in probability as $n \to \infty$. One may wonder whether there is an intermediate scale a_n such that $S_n/a_n \to 0$ in probability but not almost surely. The answer is provided by the law of the iterated logarithm (LIL), which lies between the scales of the law of large numbers and the central limit theorem.

The LIL scaling is

$$a_n = \sqrt{2n \log\log n},$$

and one has

$$\limsup_{n\to\infty} \frac{S_n}{a_n} = 1, \qquad \liminf_{n\to\infty} \frac{S_n}{a_n} = -1 \qquad \text{a.s.}$$

Thus S_n/a_n does not converge to 0 almost surely, even though this scale is larger than the CLT scale $\sqrt{n}$.

In particular, the LIL implies the following guarantee on the deviations of the sample mean as an estimator of the true mean: with probability one, there exists $n_0 \in \mathbb{N}$ such that

$$\left|\frac{S_n}{n}\right| \le \left(\frac{2\log\log n}{n}\right)^{1/2}$$

for every $n \ge n_0$.

As seen in the last sentence above, the applicability of the LIL is limited since it does not tell us how large n should be for a given deviation. We would like to have a workable expression for such a threshold. This limitation can be lifted by exploiting concentration inequalities, such as Hoeffding's inequality in Example 1, which may be viewed as non-asymptotic versions of the CLT.

These inequalities not only confirm our intuition that the sample mean should concentrate tightly about the true mean, but also quantify how a random variable X fluctuates around its mean μ by providing bounds for the tails of $X - \mu$, such as

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

Unlike the classical limit theorems, such inequalities are non-asymptotic: they hold for every fixed sample size N, not only as $N \to \infty$. This makes them especially useful in statistical physics, where N often corresponds to system size, and in data science, where N often corresponds to sample size. Concentration inequalities are part of a broader phenomenon called concentration of measure. On the other hand, there are also anti-concentration phenomena.

Example 1.6 (Anti-concentration of the Gaussian distribution)**.** Let $X \sim N(0, \sigma^2)$. Then

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

for sufficiently small $t > 0$.

For completeness, we now state the classical limit theorems and Berry–Esseen for general i.i.d. random variables, without proof.

Let

$$S_n = X_1 + \cdots + X_n,$$

where the X_i are i.i.d. random variables with mean μ, variance σ^2, characteristic function ϕ, and cumulative distribution function F.

Theorem 1.7 (Classical limit theorems for sums of i.i.d. random variables)**.** *We have the following results.*

1. ***Strong law of large numbers (SLLN).***

$$\frac{S_n}{n} \to \mu \qquad a.s.$$

2. ***Weak law of large numbers (WLLN).*** *The following three statements are equivalent:*

   (i) ϕ *is differentiable at* 0*, and* $\phi'(0) = i\mu$*.*

   (ii) As $t \to \infty$*, one has*

$$t[1 - F(t) + F(-t)] \to 0 \quad \text{and} \quad \int_{-t}^{t} x\, F(dx) \to \mu.$$

   (iii)

$$\frac{S_n}{n} \to \mu \qquad \text{in probability as } n \to \infty.$$

3. ***Central limit theorem (CLT).*** *As* $n \to \infty$*,*

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \Rightarrow N(0,1),$$

   where $\Rightarrow$ *denotes convergence in distribution.*

Theorem 1.8 (Berry–Esseen theorem)**.** *Assume that* $\mathbb{E}|X_i|^3 < \infty$*. Then*

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

where $0.41 \le C \le 0.4748$ *is a universal constant and* Φ *is the CDF of a standard normal random variable.*

If we assume more, such as boundedness or bounded moment generating functions, we can prove stronger tail inequalities, including Hoeffding, Bennett, and Bernstein inequalities.

Remark 1.9 (Looking ahead)**.** One way to prove the CLT is by a stability argument, namely Lindeberg's swapping trick: a perturbation in a normalized sum by a random variable with matching first and second moments does not change the limiting distribution. Stability arguments of this type also play an important role in random matrix theory and statistical learning theory.

2 Concentration Inequalities

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

Definition 2.1. A random variable X with mean $\mu = \mathbb{E}X$ is called *sub-Gaussian* if there exists $\sigma > 0$ such that

$$\mathbb{E}e^{\lambda(X-\mu)} \leq e^{\sigma^2\lambda^2/2} \qquad \text{for all } \lambda \in \mathbb{R}.$$

In this case, σ is called a sub-Gaussian parameter.

A sub-Gaussian random variable with parameter σ satisfies

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

for all $t \geq 0$.

Clearly, any Gaussian random variable with variance σ^2 is sub-Gaussian with parameter σ. From Example 1, Rademacher random variables are sub-Gaussian. More generally:

Proposition 2.2. *Any bounded random variable is sub-Gaussian.*

Proof. The key idea is a symmetrization argument.

Without loss of generality, assume $\mu = \mathbb{E}_X X = 0$ and that the support of X is contained in the interval $[a, b]$.

Let X' be an independent copy of X and let ϵ be an independent Rademacher random variable. Since

$$\mathbb{E}_\epsilon[e^{\lambda\epsilon}] \leq e^{\lambda^2/2},$$

we have

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

Therefore X is sub-Gaussian with parameter at most $\sigma = b - a$. □

Working with bounded random variables, which are therefore sub-Gaussian, is a standard trick in proving laws of large numbers and central limit theorems.

On the other hand, Poisson, exponential, Pareto, and Cauchy distributions are not all sub-Gaussian; some have heavier tails.

Proposition 2.3 (Hoeffding bound for sums of independent sub-Gaussian random variables)**.** *Let*

$$S_n = X_1 + \cdots + X_n,$$

where the X_i are independent mean-zero sub-Gaussian random variables with parameters σ_i. Then for all $t \geq 0$,

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

where $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_n)$ *and*

$$\|\boldsymbol{\sigma}\|_2^2 = \sum_{i=1}^{n} \sigma_i^2.$$

Proof. One shows that S_n is itself sub-Gaussian with parameter $\|\boldsymbol{\sigma}\|_2$, and then applies the usual Chernoff bound. □

There are useful equivalences between the tail behavior, the moment generating function, and the L^p growth of a sub-Gaussian random variable.

Proposition 2.4 (Sub-Gaussian properties)**.** *Let X be a random variable. The following are equivalent, up to absolute constants in the parameters:*

(i) The tails satisfy

$$\mathbb{P}[|X| \geq t] \leq 2\exp(-t^2/K_1^2), \qquad t \geq 0.$$

(ii) The moments satisfy

$$\|X\|_{L^p} = (\mathbb{E}|X|^p)^{1/p} \leq K_2\sqrt{p}, \qquad p \geq 1.$$

(iii) The MGF of X^2 satisfies

$$\mathbb{E}\exp(\lambda^2 X^2) \leq \exp(K_3^2\lambda^2), \qquad |\lambda| \leq 1/K_3.$$

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

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

for some constant $C > 1$.

(v) If additionally $\mathbb{E}X = 0$, then the preceding are also equivalent to

$$\mathbb{E}\exp(\lambda X) \leq \exp(K_5^2\lambda^2) \qquad \textit{for all } \lambda \in \mathbb{R}.$$

Proof. See Proposition 2.5.2 in Vershynin. □

Although the class of sub-Gaussian distributions is natural and large, it excludes important random variables with heavier tails. In many applications one is interested in functionals such as X_1^2 or

$$\|X\|_2^2 = \sum_{i=1}^{N} X_i^2,$$

where the X_i are independent sub-Gaussian random variables. However, the X_i^2 are typically not sub-Gaussian. Indeed,

$$\mathbb{P}[X_i^2 > t] = \mathbb{P}[|X_i| > \sqrt{t}] \sim e^{-ct}$$

for some constant $c > 0$, which suggests an exponential tail rather than a Gaussian one.

This motivates the study of sub-exponential random variables.

Definition 2.5. A random variable X with mean $\mu = \mathbb{E}X$ is called *sub-exponential* if there

exist parameters (ν, α) with $\nu, \alpha \geq 0$ such that

$$\mathbb{E}e^{\lambda(X-\mu)} \leq e^{\nu^2\lambda^2/2} \qquad \text{for all } |\lambda| \leq 1/\alpha.$$

This definition follows Wainwright and is slightly more general than the one used in Vershynin.

Proposition 2.6 (Sub-exponential properties)**.** *Let X be a mean-zero random variable. The following are equivalent:*

(i) There exist constants $c_1, c_2 > 0$ such that

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

(ii) There exists $c_0 > 0$ such that

$$\mathbb{E}[e^{\lambda X}] < \infty \qquad \text{for all } |\lambda| \leq c_0.$$

(iii)

$$\gamma := \sup_{k \geq 2} \left(\frac{\mathbb{E}[X^k]}{k!} \right)^{1/k} < \infty.$$

(iv) There exist (ν, α) with $\nu, \alpha \geq 0$ such that

$$\mathbb{E}e^{\lambda X} \leq e^{\nu^2\lambda^2/2} \qquad \text{for all } |\lambda| < 1/\alpha.$$

Proof. See Wainwright or Vershynin. □

Clearly, every sub-Gaussian random variable is sub-exponential. Products of sub-Gaussian random variables, such as squares, are often sub-exponential. Exponential and Poisson distributions also fit naturally into this class.

Example 2.7 (Sub-exponential but not sub-Gaussian)**.** Let $Z \sim N(0,1)$ and set $X = Z^2$. Then $\mathbb{E}X = 1$ and

$$\mathbb{E}\exp(\lambda(X-1)) = \begin{cases} \dfrac{e^{-\lambda}}{\sqrt{1-2\lambda}}, & \lambda < 1/2, \\ \infty, & \lambda \geq 1/2. \end{cases}$$

Thus X is not sub-Gaussian. On the other hand, one can check that

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

so X is sub-exponential with parameters, for example, $(\nu, \alpha) = (2, 4)$.

Proposition 2.8 (Sub-exponential tail bound)**.** *Let X be a sub-exponential random variable with parameters (ν, α). Then*

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

Proof. Without loss of generality, take $\mu = 0$. By the Chernoff bound,

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

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

Now minimize $g(\lambda, t)$ over $\lambda \in [0, 1/\alpha)$. The unconstrained minimizer is $\lambda^* = t/\nu^2$.

If $0 \leq t < \nu^2/\alpha$, then $\lambda^* < 1/\alpha$, so the unconstrained minimum is valid and gives

$$g^*(t) = -\frac{t^2}{2\nu^2}.$$

If $t \geq \nu^2/\alpha$, the constrained minimum is achieved at the boundary $\lambda = 1/\alpha$, yielding

$$g^*(t) = -\frac{t}{\alpha} + \frac{\nu^2}{2\alpha^2} \leq -\frac{t}{2\alpha}.$$

This proves the result. □

An alternative way to verify sub-exponential behavior is via polynomial moments.

Definition 2.9. A random variable X is said to satisfy *Bernstein's condition* with parameter b if

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

where $\mu = \mathbb{E}X$ and $\sigma^2 = \mathbb{E}[(X-\mu)^2]$.

For example, if $|X - \mu| \leq b$, then X satisfies Bernstein's condition. In fact, any random variable satisfying Bernstein's condition is sub-exponential. This leads to tail bounds sharper than Hoeffding's in some regimes.

Lemma 2.10. *If X satisfies Bernstein's condition* (4)*, then X is sub-exponential with parameters $(\sqrt{2}\sigma, 2b)$.*

Proof. Expand the exponential:

$$\begin{aligned}\mathbb{E}[e^{\lambda(X-\mu)}] &= 1 + \frac{\lambda^2\sigma^2}{2} + \sum_{k=3}^{\infty} \lambda^k \frac{\mathbb{E}[(X-\mu)^k]}{k!} \\ &\leq 1 + \frac{\lambda^2\sigma^2}{2} + \frac{\lambda^2\sigma^2}{2}\sum_{k=3}^{\infty}(|\lambda|b)^{k-2}.\end{aligned}$$

For $|\lambda| < 1/b$,

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

In particular, if $|\lambda| < 1/(2b)$, then

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq \exp\left(\frac{(\sqrt{2}\sigma)^2\lambda^2}{2}\right).$$

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

Proposition 2.11 (Bernstein-type bound)**.** *Let X satisfy Bernstein's condition* (4)*. Then for $|\lambda| < 1/b$,*

$$\mathbb{E}[e^{\lambda(X-\mu)}] \le \exp\left(\frac{\lambda^2\sigma^2}{2(1-b|\lambda|)}\right). \tag{5}$$

Moreover,

$$\mathbb{P}[|X-\mu| \ge t] \le 2\exp\left(-\frac{t^2}{2(\sigma^2+bt)}\right) \qquad \text{for all } t \ge 0.$$

Proof. The MGF bound (5) was already established in the previous lemma. The concentration inequality follows by combining the Chernoff bound with (5) and optimizing over λ. □

Sub-exponentiality is preserved under summation for independent random variables.

Example 2.12. Let X_i, $i = 1, \dots, N$, be independent random variables with means μ_i, each sub-exponential with parameters (ν_i, α_i). Then

$$\widetilde{S}_N := \sum_{i=1}^{N}(X_i - \mu_i)$$

is sub-exponential with parameters

$$\alpha_* = \max_i \alpha_i, \qquad \nu_*^2 = \sum_{i=1}^{N}\nu_i^2.$$

Moreover, for $t \ge 0$,

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

In particular, if the X_i are i.i.d. and centered, then for suitable constants $c, C > 0$,

$$\mathbb{P}\left[\frac{S_N}{\sqrt{N}} \ge t\right] \le \begin{cases} 2e^{-ct^2}, & t \le C\sqrt{N}, \\ 2e^{-t\sqrt{N}}, & t \ge C\sqrt{N}. \end{cases}$$

Thus, in the small-deviation regime $t \le C\sqrt{N}$, the sum has a sub-Gaussian tail as though it were approximately normal with constant variance. For larger deviations, one sees a heavier, sub-exponential tail. In short, Bernstein's inequality for sums of i.i.d. sub-exponential random variables yields a mixture of two tail behaviors: sub-Gaussian for small deviations and sub-exponential for large deviations.

Remark 2.13. Bernstein's bound can be strengthened by Bennett's inequality. We do not cover Bennett's inequality here; see Exercise 2.7 in Wainwright for details.

3 Summary

- Concentration inequalities provide finite-sample control of deviations.

- Hoeffding's inequality gives Gaussian-type tails for sums of bounded or sub-Gaussian variables.
- Sub-Gaussian random variables are characterized equivalently by their tails, moments, and MGFs.
- Sub-exponential random variables are heavier-tailed, but still enjoy strong concentration for moderate deviations.
- Bernstein's condition leads to Bernstein-type inequalities, which combine variance information with exponential-tail control.
- These tools are foundational in high-dimensional probability, random matrix theory, and statistical learning theory.

4 Exercises

1. Show directly that if X is sub-Gaussian with parameter σ, then
$$\mathbb{P}(|X - \mathbb{E}X| \geq t) \leq 2e^{-t^2/(2\sigma^2)}.$$
2. Improve the bounded-random-variable argument to show that if $X \in [a, b]$ almost surely, then X is sub-Gaussian with parameter $(b - a)/2$.
3. Prove the Hoeffding bound for sums of independent mean-zero sub-Gaussian random variables.
4. Verify that if $Z \sim \mathcal{N}(0, 1)$, then Z^2 is sub-exponential but not sub-Gaussian.
5. Derive the two-regime tail bound for sub-exponential random variables using the Chernoff method.
6. Prove that Bernstein's condition implies the MGF inequality
$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq \exp\left(\frac{\lambda^2\sigma^2}{2(1 - b|\lambda|)}\right).$$
7. Derive Bernstein's concentration inequality from the MGF bound.
8. Let $X_1, \ldots, X_N$ be i.i.d. sub-exponential. Derive a concentration inequality for the sample mean $N^{-1}\sum_{i=1}^{N} X_i$.
9. Compare Hoeffding and Bernstein inequalities in a setting where both apply. When is Bernstein sharper?
10. Read about Bennett's inequality and compare it conceptually with Bernstein's inequality.