Exponential Concentration I: Part B

SDS 391P.6, Spring 2026

Pratik Patil

These notes are a work in progress and are provided as-is for instructional purposes only. They are not (yet) at the level of a scholarly document. In particular, the notes draw from various sources and do not (yet) have sufficient references to the original sources. Additionally, almost surely the notes have errors and they are only probably approximately correct. The notes will be updated regularly as the course progresses. Last updated: 2026-02-02.

1 Motivation

In the previous lecture, we proved Hoeffding's inequality for sums of bounded independent random variables. Boundedness is a clean and useful assumption, but it is also restrictive: many basic distributions (Gaussian, Poisson, χ^2, etc.) are unbounded.

A key lesson from the proof is that boundedness itself was not the true driver. What mattered was control of the centered log-moment generating function (log-MGF)

$$\psi_X(\lambda) := \log \mathbb{E} e^{\lambda(X - \mathbb{E}X)}.$$

Hoeffding's lemma showed that if $a \leqslant X \leqslant b$ almost surely, then

$$\psi_X(\lambda) \leqslant \frac{\lambda^2(b-a)^2}{8} \qquad \text{for all } \lambda \in \mathbb{R}.$$

This is exactly the same quadratic functional form as for a Gaussian: if $G \sim \mathcal{N}(0, \sigma^2)$ then $\psi_G(\lambda) = \sigma^2\lambda^2/2$.

This suggests a general idea: if $\psi_X(\lambda)$ is dominated by a quadratic function of λ (globally or locally), then the Laplace method produces exponential concentration. In this lecture we formalize two important classes built around this idea:

- *sub-Gaussian* random variables, which exhibit Gaussian-type tails.
- *sub-exponential* random variables, which exhibit exponential-type tails and lead to a characteristic two-regime (Gaussian/exponential) concentration for sums. The two-regime behavior will be captured by Bernstein's inequality, proved again via the Laplace method.

2 Sub-Gaussian random variables

Throughout this section it is helpful to explicitly separate centering from tail behavior. Given X, define its centered version

$$Z := X - \mathbb{E}X, \qquad \text{so that} \qquad \psi_X(\lambda) = \log \mathbb{E} e^{\lambda Z}.$$

We say that X (equivalently, Z) is σ^2-*sub-Gaussian* if

$$\log \mathbb{E} e^{\lambda Z} \leqslant \frac{\sigma^2\lambda^2}{2} \qquad \text{for all } \lambda \in \mathbb{R}. \tag{1}$$

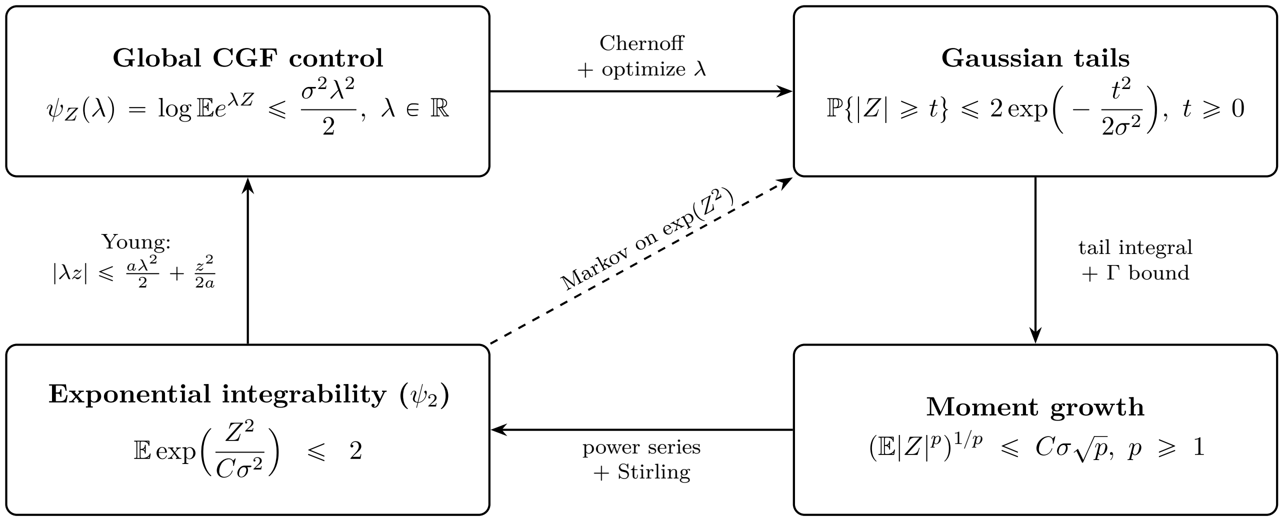

Figure 1: Equivalent (up to universal constants) characterizations of sub-Gaussianity for $Z := X - \mathbb{E}X$. Arrow labels indicate standard proof routes.

The parameter σ^2 is called a *variance proxy*. By the Chernoff bound, (1) immediately yields Gaussian tails:

$$\mathbb{P}\{|Z| \geqslant t\} \leqslant 2 \exp\Big(-\frac{t^2}{2\sigma^2}\Big).$$

This definition is also extremely convenient for sums: if $X_1, \dots, X_n$ are independent and $Z_i := X_i - \mathbb{E}X_i$, then

$$\log \mathbb{E} e^{\lambda \sum_{i=1}^n Z_i} = \sum_{i=1}^n \log \mathbb{E} e^{\lambda Z_i},$$

so quadratic log-MGF bounds tensorize immediately.

2.1 Equivalent ways to recognize sub-Gaussianity

The log-MGF definition is calculus-friendly, but not always the most intuitive. A useful fact is that sub-Gaussianity can be recognized in several equivalent ways: Gaussian tails, Gaussian-like moment growth, and exponential-square integrability are all the same phenomenon up to universal constants. Figure 1 summarizes these equivalences and typical proof routes. (Proving the equivalences carefully is on Homework 2.)

The ψ_2 (Orlicz) norm. It is often helpful to compress these equivalences into a single quantity. Define the *sub-Gaussian norm* (also called the ψ_2 norm) by

$$\|X\|_{\psi_2} := \inf\Big\{K > 0 : \ \mathbb{E} \exp\Big(\frac{X^2}{K^2}\Big) \leqslant 2\Big\}. \tag{2}$$

Then X is sub-Gaussian iff $\|X\|_{\psi_2} < \infty$. Moreover, up to absolute constants, the ψ_2 norm controls (and is controlled by) each of: Gaussian tails, moment growth, and quadratic log-MGF bounds (Figure 1). In practice we freely use the following consequences: there exist universal constants $c, C > 0$ such that

$$\mathbb{P}\{|X| \geqslant t\} \leqslant 2 \exp\Big(-c\,\frac{t^2}{\|X\|_{\psi_2}^2}\Big), \qquad \|X\|_{L^p} \leqslant C\, \|X\|_{\psi_2} \sqrt{p} \quad (p \geqslant 1),$$

and, if $\mathbb{E}X = 0$,

$$\log \mathbb{E}e^{\lambda X} \leqslant C\lambda^2 \|X\|_{\psi_2}^2 \qquad (\lambda \in \mathbb{R}).$$

(Also, centering is harmless: $\|X - \mathbb{E}X\|_{\psi_2} \lesssim \|X\|_{\psi_2}$.)

Examples and non-examples.

- If $X \sim \mathcal{N}(0, \sigma^2)$ then X is σ^2-sub-Gaussian and $\|X\|_{\psi_2} \asymp \sigma$.
- If $X \in [a, b]$ almost surely then $X - \mathbb{E}X$ is sub-Gaussian with proxy $\asymp (b-a)^2$ (Hoeffding's lemma).
- A Rademacher random variable $\varepsilon \in \{-1, +1\}$ is sub-Gaussian.
- The exponential distribution is *not* sub-Gaussian: its tails behave like $\exp(-ct)$, not $\exp(-ct^2)$.

2.2 Sums of independent sub-Gaussians

Here is the basic closure property that replaces "bounded implies Hoeffding" by "sub-Gaussian implies Hoeffding".

Theorem 2.1 (Sub-Gaussian Hoeffding inequality)**.** *Let $X_1, \dots, X_n$ be independent, mean-zero random variables. Assume each X_i is sub-Gaussian and set $K_i := \|X_i\|_{\psi_2}$. Then for every $t \geqslant 0$,*

$$\mathbb{P}\left\{ \left| \sum_{i=1}^n X_i \right| \geqslant t \right\} \;\leqslant\; 2\exp\left(-c\,\frac{t^2}{\sum_{i=1}^n K_i^2} \right),$$

where $c > 0$ is an absolute constant. Equivalently, $\left\| \sum_i X_i \right\|_{\psi_2}^2 \;\leqslant\; C \sum_i \|X_i\|_{\psi_2}^2$ for an absolute constant C.

Proof sketch. Independence factorizes the MGF: $\mathbb{E}e^{\lambda \sum_i X_i} = \prod_i \mathbb{E}e^{\lambda X_i}$. The sub-Gaussian MGF bound gives $\mathbb{E}e^{\lambda \sum_i X_i} \leqslant \exp(C\lambda^2 \sum_i K_i^2)$. Apply Chernoff and optimize in λ.

As a consistency check, if $X_i \in [a_i, b_i]$ almost surely, then Hoeffding's lemma gives $K_i^2 \lesssim (b_i - a_i)^2$, and Theorem 2.1 recovers Hoeffding's inequality from the previous lecture (up to constants).

3 Sub-exponential random variables

Sub-Gaussian tails are strong, and they rule out many random variables that still exhibit meaningful exponential concentration for sums. A basic example is G^2 where $G \sim \mathcal{N}(0, 1)$:

$$\mathbb{P}\{G^2 \geqslant t\} = \mathbb{P}\{|G| \geqslant \sqrt{t}\} \approx e^{-t/2},$$

which decays exponentially in t (not Gaussian in t). Thus G^2 is not sub-Gaussian, but it is still light-tailed in a different sense.

This motivates a broader class: random variables with (at least) exponential tails.

The ψ_1 norm. We say that X is *sub-exponential* if

$$\|X\|_{\psi_1} \;:=\; \inf\left\{ K > 0 :\; \mathbb{E}\exp\left(\frac{|X|}{K} \right) \leqslant 2 \right\} \;<\; \infty.$$

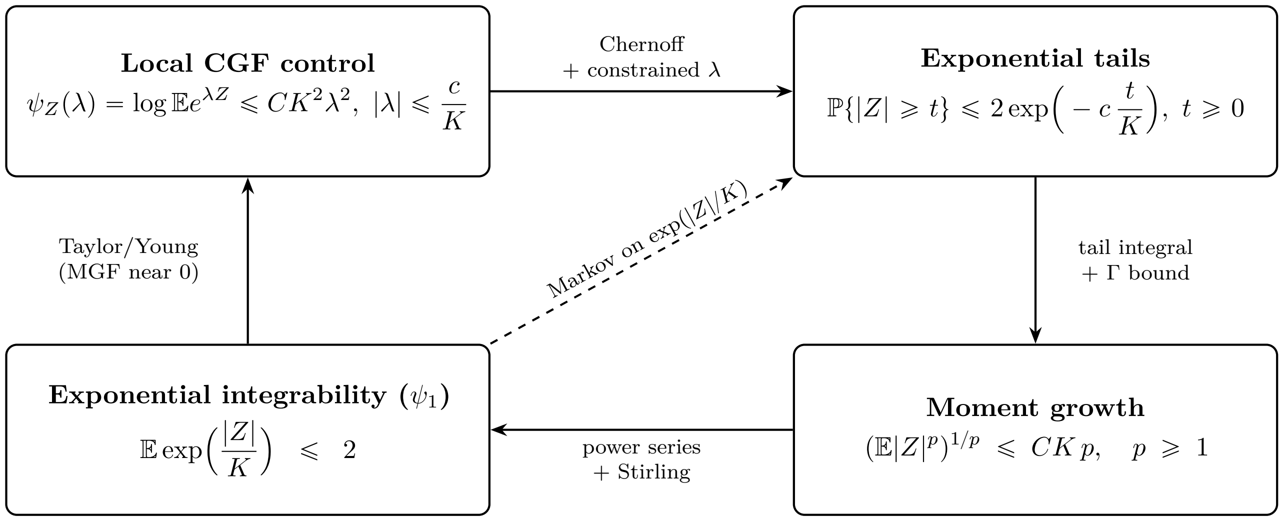

Figure 2: Equivalent (up to universal constants) characterizations of sub-exponentiality for $Z := X - \mathbb{E}X$. Here K is a scale parameter (comparable to $\|Z\|_{\psi_1}$). Arrow labels indicate standard proof routes.

Up to absolute constants, $\|X\|_{\psi_1}$ is equivalent to each of:

$$\mathbb{P}\{|X| \geqslant t\} \leqslant 2\exp\Big(-c\,\frac{t}{\|X\|_{\psi_1}}\Big), \qquad \|X\|_{L^p} \leqslant C\,\|X\|_{\psi_1}\,p \quad (p \geqslant 1).$$

(Centering is again harmless: $\|X - \mathbb{E}X\|_{\psi_1} \lesssim \|X\|_{\psi_1}$.)

3.1 Equivalent ways to recognize sub-exponentiality

Just as for sub-Gaussian variables, there are several interchangeable ways to express that a centered random variable $Z = X - \mathbb{E}X$ is sub-exponential: exponential tails, linear moment growth, exponential integrability of $|Z|$, and a log-MGF bound. The key distinction from the sub-Gaussian case is that the log-MGF control is typically only valid *locally* (near $\lambda = 0$), reflecting the fact that many sub-exponential variables have MGFs that blow up at a finite radius (e.g. exponential random variables).

Figure 2 summarizes these equivalences and standard proof routes.

Local log-MGF control. A particularly useful consequence is the following: if $\mathbb{E}Z = 0$ and $\|Z\|_{\psi_1} \leqslant K$, then

$$\log \mathbb{E} e^{\lambda Z} \ \leqslant\ C\lambda^2 K^2 \qquad \text{for all } |\lambda| \leqslant \frac{c}{K}, \tag{3}$$

with absolute constants $c, C > 0$. This *restricted* quadratic control is exactly what drives Bernstein's inequality.

3.2 Sub-Gaussian versus sub-exponential

Sub-exponential is strictly weaker than sub-Gaussian:

$$X \text{ sub-Gaussian} \quad \Rightarrow \quad X \text{ sub-exponential}.$$

Two particularly important closure facts are:

- If X is sub-Gaussian then X^2 is sub-exponential, and $\|X^2\|_{\psi_1} \asymp \|X\|_{\psi_2}^2$.
- If X, Y are sub-Gaussian then XY is sub-exponential and $\|XY\|_{\psi_1} \lesssim \|X\|_{\psi_2}\|Y\|_{\psi_2}$. (A quick proof uses $|xy| \leqslant x^2/2 + y^2/2$ and exponential-square integrability.)

These facts explain why sub-exponential concentration appears naturally in problems involving norms and quadratic forms: squares and products show up everywhere.

3.3 Sums of independent sub-exponentials

For bounded variables, Hoeffding gives purely Gaussian tails. For sub-exponential variables, purely Gaussian tails are not available at all scales: even a single summand may have only $\exp(-ct)$ tails. The correct behavior is a *two-regime* inequality: Gaussian for moderate deviations and exponential for very large deviations.

Theorem 3.1 (Bernstein inequality for sub-exponential sums)**.** *Let $X_1, \dots, X_n$ be independent, mean-zero sub-exponential random variables. Set*

$$K_i := \|X_i\|_{\psi_1}, \qquad V := \sum_{i=1}^n K_i^2, \qquad B := \max_{1 \leqslant i \leqslant n} K_i.$$

Then for all $t \geqslant 0$,

$$\mathbb{P}\left\{\left|\sum_{i=1}^n X_i\right| \geqslant t\right\} \;\leqslant\; 2\exp\left[-c\min\left(\frac{t^2}{V},\; \frac{t}{B}\right)\right],$$

where $c > 0$ is an absolute constant.

Proof sketch. Apply Chernoff to $\sum_i X_i$ and use independence to factor the MGF. The local log-MGF bound (3) gives

$$\log \mathbb{E} e^{\lambda \sum_i X_i} \leqslant C\lambda^2 V \quad \text{as long as} \quad |\lambda| \leqslant c/B.$$

Optimizing $-\lambda t + C\lambda^2 V$ gives the "Gaussian" choice $\lambda^\star \asymp t/V$, but we must respect $|\lambda| \leqslant c/B$. If $\lambda^\star \leqslant c/B$ we obtain an exponent of order t^2/V; otherwise we hit the constraint and obtain an exponent of order t/B.

Figure 3 visualizes the minimum of these two exponents.

4 Application: norm concentration for Gaussian vectors

Let $g = (g_1, \dots, g_d) \sim \mathcal{N}(0, I_d)$. Then $\|g\|_2^2 = \sum_{i=1}^d g_i^2$ and $\mathbb{E}\|g\|_2^2 = d$. The summands $g_i^2 - 1$ are i.i.d. mean-zero and sub-exponential (in fact, g_i is sub-Gaussian, hence g_i^2 is sub-exponential).

Applying Bernstein's inequality (Theorem 3.1) to

$$\|g\|_2^2 - d = \sum_{i=1}^d (g_i^2 - 1)$$

yields

$$\mathbb{P}\Big\{\,\big|\|g\|_2^2 - d\big| \geqslant t\,\Big\} \;\leqslant\; 2\exp\left[-c\min\left(\frac{t^2}{d}, t\right)\right]. \tag{4}$$

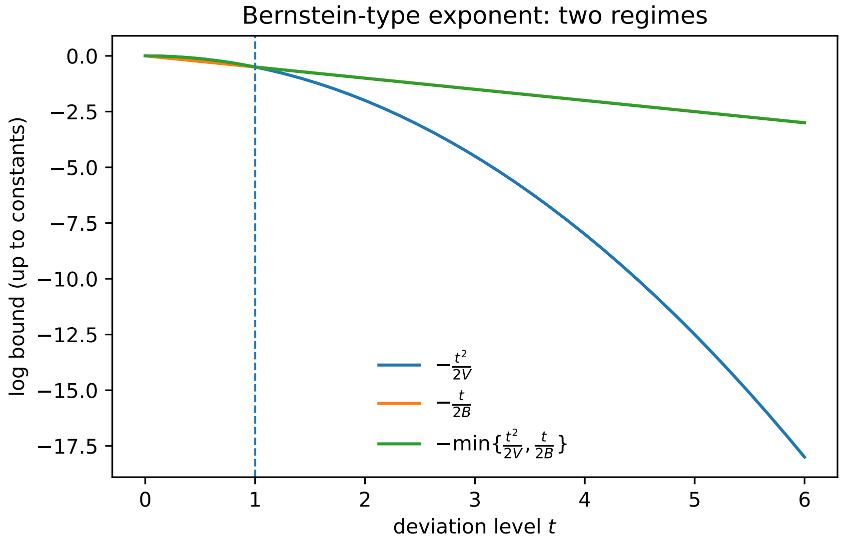

Figure 3: Bernstein-type exponent has two regimes: $-\min\{t^2/(2V),\, t/(2B)\}$ behaves quadratically near 0 (Gaussian regime) and linearly for large t (exponential regime).

In particular, taking $t = \varepsilon d$ for $\varepsilon \in (0,1)$ gives

$$\mathbb{P}\Big\{ \big|\|g\|_2^2 - d\big| \geqslant \varepsilon d \Big\} \;\leqslant\; 2e^{-c\varepsilon^2 d}.$$

Thus $\|g\|_2^2$ concentrates sharply around d on a relative scale ε. A short algebraic step (using $|\sqrt{u} - \sqrt{v}| \leqslant |u-v|/(\sqrt{u}+\sqrt{v})$) converts (4) into concentration of $\|g\|_2$ around $\sqrt{d}$.

This "thin shell" phenomenon is one of the basic geometric consequences of exponential concentration, and it will reappear repeatedly in high-dimensional geometry and randomized algorithms.

5 Application: Johnson–Lindenstrauss via Bernstein

A second canonical application is dimension reduction. Suppose we want to embed a finite set of points in $\mathbb{R}^d$ into a much smaller $\mathbb{R}^m$ while approximately preserving Euclidean distances. Random linear maps succeed because they preserve norms of *fixed* vectors with high probability, and Bernstein's inequality lets us union bound over a finite set.

Concentration for a fixed vector. Let $A \in \mathbb{R}^{m\times d}$ have independent rows $a_1, \dots, a_m$. For a fixed $x \in \mathbb{R}^d$,

$$\frac{1}{m}\|Ax\|_2^2 = \frac{1}{m}\sum_{r=1}^{m}\langle a_r, x\rangle^2.$$

If each row a_r is an *isotropic sub-Gaussian* vector (so $\mathbb{E}\langle a_r, x\rangle^2 = \|x\|_2^2$ and $\langle a_r, x\rangle$ is sub-Gaussian with scale $\lesssim \|x\|_2$), then $\langle a_r, x\rangle^2 - \|x\|_2^2$ is mean-zero and sub-exponential with scale $\lesssim \|x\|_2^2$. Bernstein's inequality then yields, for $\varepsilon \in (0,1)$,

$$\mathbb{P}\left\{ \left| \frac{1}{m}\|Ax\|_2^2 - \|x\|_2^2 \right| \geqslant \varepsilon\|x\|_2^2 \right\} \;\leqslant\; 2\exp(-c\varepsilon^2 m).$$

This is the basic "one-vector" guarantee.

Union bound over a finite set. Let $\mathcal{V} \subset \mathbb{R}^d$ be a finite set with $|\mathcal{V}| = N$. Apply the one-vector bound to all difference vectors $x = u - v$ with $u, v \in \mathcal{V}$, and union bound over the $\binom{N}{2}$ pairs. Choosing $m \gtrsim \varepsilon^{-2} \log N$ makes the failure probability small.

A clean statement is the Johnson–Lindenstrauss lemma:

Theorem 5.1 (Johnson–Lindenstrauss lemma (one standard form))**.** *Fix $\varepsilon \in (0,1)$ and a finite set $\mathcal{V} \subset \mathbb{R}^d$ with $|\mathcal{V}| = N$. Let $A \in \mathbb{R}^{m \times d}$ have i.i.d. $\mathcal{N}(0,1)$ entries, and define*

$$\Phi(x) := \frac{1}{\sqrt{m}} Ax \in \mathbb{R}^m.$$

If $m \geqslant C\varepsilon^{-2} \log N$, then with probability at least $1 - 2e^{-c\varepsilon^2 m}$, we have for all $u, v \in \mathcal{V}$,

$$(1-\varepsilon)\|u-v\|_2^2 \ \leqslant\ \|\Phi(u) - \Phi(v)\|_2^2 \ \leqslant\ (1+\varepsilon)\|u-v\|_2^2.$$

Here $c, C > 0$ are absolute constants.

The proof is precisely the two-step plan above: concentration for a fixed vector (via sub-exponential Bernstein), followed by a union bound over finitely many vectors.

6 Look ahead

So far, we have used the Laplace method plus independence to control the log-MGF of a *sum*. A major theme of modern concentration is that similar exponential tail bounds hold for *nonlinear* functions of independent inputs, $Z = f(X_1, \ldots, X_n)$, when f has a suitable notion of low sensitivity.

In the previous unit, variance bounds were driven by tensorization of variance and Efron–Stein-type arguments. In the exponential world, the analogous device is *tensorization of entropy* and related functional inequalities. These ideas lead to concentration inequalities for Lipschitz functions, bounded differences, and beyond, often with near-Gaussian tails.

Next time we will start building these tools!

Source material

Parts of this lecture are based on references: Vershynin (2018); Tropp (2023); van Handel (2016), in addition to the author's accumulated experience working on related topics.

References

Tropp, J. A. (2023). Probability in high dimensions. Caltech CMS Lecture Notes 2021-01.

van Handel, R. (2016). Probability in high dimension. Lecture Notes (Princeton University). APC 550.

Vershynin, R. (2018). *High-dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press.