TWO FRIENDLY PROOFS OF THE BERRY–ESSEEN THEOREM

ROMAN VERSHYNIN

Abstract. A gem of classical probability, the Berry–Esseen theorem provides a non-asymptotic form of the central limit theorem. This note gives a friendly and intuitive exposition of two different proofs of the Berry–Esseen theorem for nonidentically distributed random variables: the classical Fourier-analytic proof and a proof by Stein's method following E. Bolthausen. Both proofs are self-contained; the reader may read either proof without reading the other. The exposition is suitable for a basic graduate course in probability.

The central limit theorem describes the eventual Gaussian behavior of sums of independent random variables. The Berry–Esseen theorem makes this approximation quantitative for every finite sum.

Theorem 0.1 (Berry–Esseen theorem [1, 4])**.** *Let $X_1, \dots, X_n$ be independent mean-zero random variables whose sum $S_n \coloneqq X_1 + \cdots + X_n$ satisfies $\mathrm{Var}(S_n) = 1$. Let G be a standard normal random variable. Then*

$$\sup_{a \in \mathbb{R}} \Big| \mathbb{P}\{S_n \le a\} - \mathbb{P}\{G \le a\} \Big| \le C \sum_{k=1}^{n} \mathbb{E}|X_k|^3,$$

provided the right-hand side is finite, where C is an absolute constant.

To see how this implies the classical central limit theorem, let $Y_1, Y_2, \dots$ be independent mean-zero random variables with unit variances and uniformly bounded third moments. Applying Theorem 0.1 to $X_k = Y_k/\sqrt{n}$ gives

$$\sup_{a \in \mathbb{R}} \left| \mathbb{P}\Big\{ \frac{Y_1 + \cdots + Y_n}{\sqrt{n}} \le a \Big\} - \mathbb{P}\{G \le a\} \right| = O\left(\frac{1}{\sqrt{n}} \right). \tag{0.1}$$

Thus the normalized sums converge in distribution to G. More precisely, (0.1) gives uniform convergence of the distribution functions, the optimal rate $O(n^{-1/2})$, and a bound valid for every finite n.

We prove the theorem twice. Part I gives the classical Fourier-analytic argument, including an intuitive proof of Esseen's smoothing inequality. Part II gives a proof by Stein's method, following an extremely concise argument of Bolthausen [2]. The two parts are logically independent: neither proof invokes a definition, lemma, or estimate from the other. A reader interested in just one method may therefore go directly to the corresponding part.

There are several existing textbook proofs of the Berry–Esseen theorem. The classical Fourier-analytic proof is covered, for example, in Feller [6, Ch. XVI, Sec. 5, Thm. 2]; for an introduction to Stein's method, see Ross [7] and Chen, Goldstein, and Shao [3].

Partially supported by the NSF Grant DMS 2451011 and the U.S. Air Force Grant FA9550-25-1-0294.

Part I. The Fourier-analytic proof

This part is self-contained. It begins with the proof strategy, recalls the needed Fourier analysis, and develops its own smoothing and characteristic-function estimates.

1. The proof strategy

Here is a bird's-eye, non-rigorous sketch of the proof of Theorem 0.1.

1.1. Smoothing. We begin by writing

$$\mathbb{P}\{S_n \le a\} = \mathbb{E}\, \mathbf{1}_{\{S_n \le a\}}. \tag{1.1}$$

The difficulty is that the function $x \mapsto \mathbf{1}_{\{x \le a\}}$ is too rough: it has a discontinuous drop from 1 to 0 at the point a. To mitigate this issue, we smooth the indicator function: approximate it with a function $f(x)$ that decreases from 1 to 0 gradually. With such approximation, (1.1) becomes

$$\mathbb{P}\{S_n \le a\} \approx \mathbb{E}\, f(S_n).$$

A rigorous version of this smoothing step will be given in Lemma 3.1.

1.2. Fourier transform. Next, we replace $f(x)$ by an even simpler function – a complex exponential $e^{2\pi i t x}$. This can be done using the Fourier transform inversion formula (recalled in (2.2) below):

$$\mathbb{E}\, f(S_n) = \mathbb{E} \int_{\mathbb{R}} e^{2\pi i t S_n}\, \widehat{f}(t)\, dt = \int_{\mathbb{R}} \mathbb{E}\left[e^{2\pi i t S_n}\right]\, \widehat{f}(t)\, dt.$$

Applying the same method to a standard normal random variable G, subtracting the two expressions, and using Jensen's inequality, we obtain

$$\left|\mathbb{P}\{S_n \le a\} - \mathbb{P}\{G \le a\}\right| \lesssim \int_{\mathbb{R}} \left|\mathbb{E}\, e^{2\pi i t S_n} - \mathbb{E}\, e^{2\pi i t G}\right| |\widehat{f}(t)|\, dt. \tag{1.2}$$

This reduces the problem to comparing the *characteristic functions* $\mathbb{E}\, e^{itS_n}$ and $\mathbb{E}\, e^{itG}$. For the standard normal random variable G, a simple computation yields

$$\mathbb{E}\, e^{itG} = \exp(-t^2/2), \quad t \in \mathbb{R}. \tag{1.3}$$

1.3. Taylor approximation. For S_n, independence of the random variables X_k gives

$$\mathbb{E}\, e^{itS_n} = \mathbb{E}\left[e^{itX_1}\right] \cdots \mathbb{E}\left[e^{itX_n}\right],$$

which reduces the problem to computing the characteristic function $\mathbb{E}\, e^{itX_k}$ for each random variable X_k separately. By assumption,

$$\mathbb{E}\, X_k = 0 \quad \text{and} \quad \mathbb{E}\, X_k^2 =: \sigma_k^2 < \infty.$$

Using the Taylor approximation $e^z \approx 1 + z + z^2/2$ and taking expectations, we get

$$\begin{aligned}\mathbb{E}\, e^{itX_k} &\approx \mathbb{E}\big(1 + itX_k - t^2 X_k^2/2\big) \\ &= 1 - \sigma_k^2 t^2/2 \approx \exp\big(-\sigma_k^2 t^2/2\big).\end{aligned}$$

Multiplying over k gives

$$\mathbb{E}\, e^{itS_n} \approx \exp\big(-\sigma_1^2 t^2/2\big) \cdots \exp\big(-\sigma_n^2 t^2/2\big) = \exp\big(-t^2/2\big), \tag{1.4}$$

since $\sigma_1^2 + \cdots + \sigma_n^2 = 1$ by assumption. Comparing to (1.3), we see that

$$\mathbb{E}\, e^{itS_n} \approx \mathbb{E}\, e^{itG}.$$

Substituting this approximation into (1.2) completes the heuristic "proof".

1.4. **A challenge and a fix.** The main issue with this heuristic argument lies in the quality of the Taylor approximation. We cannot expect

$$\mathbb{E}\, e^{itX_k} \approx \exp(-\sigma_k^2 t^2/2) \tag{1.5}$$

uniformly over all $t \in \mathbb{R}$. For example, if X_k has the Rademacher distribution, its characteristic function equals $\cos(t)$, which does not even decay to zero as $|t| \to \infty$.

There is an elegant fix for this issue: choose the smoothing function f whose Fourier transform $\widehat{f}$ is supported on a *compact interval* centered at the origin. The integrand in (1.2) would vanish outside that interval, and so it suffices to establish the approximation (1.5) only for t close to the origin. This idea leads to *Esseen's smoothing inequality* (Theorem 3.3).

In Lemma 4.1, we will prove (1.5) in a neighborhood of the origin. Multiplying over k, we will deduce a rigorous form of (1.4) in Lemma 4.3. Plugging it into the smoothing inequality will complete the proof of Theorem 0.1.

Now let's do this step by step.

2. Background on Fourier transform

First, let's recall some basic facts about Fourier transform that we will use in the proof.

The *Fourier transform* of an integrable function $f : \mathbb{R} \to \mathbb{C}$ is the bounded and continuous function $\widehat{f} : \mathbb{R} \to \mathbb{C}$ defined as

$$\widehat{f}(t) = \int_{\mathbb{R}} e^{-2\pi i t x} f(x)\, dx. \tag{2.1}$$

If f is a *Schwartz function* (that is, f and all its derivatives decay faster than any polynomial at infinity), then $\widehat{f}$ is also a Schwartz function.

Any Schwartz function f can be reconstructed from its Fourier transform using the *Fourier inversion formula*:

$$f(x) = \int_{\mathbb{R}} e^{2\pi i t x} \widehat{f}(t)\, dt. \tag{2.2}$$

Comparing with (2.1), we see that applying the Fourier transform twice yields $f(-x)$. In particular, the Fourier transform is an involution on even functions.

The Fourier transform preserves the L^2 norm: for a Schwartz function f, the *Plancherel identity* says that

$$\int_{\mathbb{R}} |f(x)|^2\, dx = \int_{\mathbb{R}} |\widehat{f}(t)|^2\, dt. \tag{2.3}$$

The *convolution* of two integrable functions f and g is the integrable function $f * g$ defined as

$$(f * g)(x) = \int_{\mathbb{R}} f(y) g(x-y)\, dy.$$

The *convolution theorem* states that

$$\widehat{f * g} = \widehat{f} \cdot \widehat{g} \quad \text{pointwise.} \tag{2.4}$$

3. Smoothing

We can write the cumulative distribution function of a random variable X as

$$\mathbb{P}\{X \le a\} = \mathbb{E}\,\mathbf{1}_{(-\infty,0]}(X-a),$$

where $\mathbf{1}_{(-\infty,0]}(x)$ is the indicator of $(-\infty,0]$. We now smooth this indicator, replacing it by a function that transitions gradually from 1 to 0, with most of the transition occurring in an ε-neighborhood of the origin.

Lemma 3.1 (Smoothing the discrepancy)**.** *Let φ be a probability density function, which is also a Schwartz function. Consider a smoothed version of the indicator function of $(-\infty,0]$:*

$$f = \mathbf{1}_{(-\infty,0]} * \varphi.$$

Let X and Y be random variables. Assume that the probability density function of Y is bounded by M. Then we have for any $\varepsilon > 0$:

$$\sup_{a\in\mathbb{R}} \Big| \mathbb{P}\{X \le a\} - \mathbb{P}\{Y \le a\} \Big| \le 2 \sup_{a\in\mathbb{R}} \Big| \mathbb{E}\, f\Big(\frac{X-a}{\varepsilon}\Big) - \mathbb{E}\, f\Big(\frac{Y-a}{\varepsilon}\Big) \Big| + C_\varphi M \varepsilon. \tag{3.1}$$

Here C_φ depends only on the choice of the smoothing function φ.

Proof. Step 1. Regularity of the discrepancy function. By rescaling, it is enough to consider $\varepsilon = 1$. Our task is to bound the discrepancy function

$$\Delta(a) \coloneqq F_X(a) - F_Y(a),$$

where $F_X(a) \coloneqq \mathbb{P}\{X \le a\}$ and $F_Y(a) \coloneqq \mathbb{P}\{Y \le a\}$ are the cumulative distribution functions.

The function F_X increases, while F_Y cannot increase too fast because its derivative is bounded by M. Putting this together, we see that Δ cannot decrease too fast:

$$\Delta(\bar a + t) \ge \Delta(\bar a) - Mt \quad \text{for any } \bar a \in \mathbb{R},\ t \ge 0. \tag{3.2}$$

In particular, once Δ achieves its maximum, Δ will have to remain close to its maximum for a while. To make this precise, assume without loss of generality that

$$\bar\Delta \coloneqq \sup_{a\in\mathbb{R}} |\Delta(a)| = \sup_{a\in\mathbb{R}} \Delta(a). \tag{3.3}$$

(Otherwise replace X, Y by $-X, -Y$.) Choose a point $\bar a$ where $\Delta(\bar a) \ge 0.9\bar\Delta$. Then (3.2) gives

$$\Delta(\bar a + t) \ge 0.9\bar\Delta - Mt \ge \bar\Delta/2$$

as long as $0 \le t \le \bar\Delta/3M$. In other words, we found an interval of length $\bar\Delta/3M$ on which Δ is bounded below by $\bar\Delta/2$.

Step 2. The smoothed discrepancy function. Our task is to compare the discrepancy function $\Delta(a)$ to its smoothed version

$$\Delta_f(a) \coloneqq \mathbb{E}\, f(X-a) - \mathbb{E}\, f(Y-a).$$

To express Δ_f in terms of Δ, note that

$$\mathbb{E}\, f(X-a) = \mathbb{E} \int_{\mathbb{R}} \mathbf{1}_{(-\infty,0]}(X-a-y)\varphi(y)\,dy = \int_{\mathbb{R}} F_X(a+y)\varphi(y)\,dy$$

by the Fubini theorem. Express $\mathbb{E}\, f(Y-a)$ similarly, subtract, and obtain

$$\Delta_f(a) = \int_{-\infty}^{\infty} \Delta(a+y)\varphi(y)\,dy. \tag{3.4}$$

Step 3. Integrating. In Step 1, we found an interval $[a_0 - T, a_0 + T]$ with $T = \bar{\Delta}/6M$ and on which Δ is bounded below by $\bar{\Delta}/2$. Let's decompose (3.4):

$$\Delta_f(a_0) = \underbrace{\int_{|y|\le T} \Delta(a_0+y)\varphi(y)\,dy}_{I_1} + \underbrace{\int_{|y|>T} \Delta(a_0+y)\varphi(y)\,dy}_{I_2}.$$

To bound I_2, recall that (3.3) implies that Δ is bounded below by $-\bar{\Delta}$ everywhere. Moreover, φ is a Schwartz function, so $\int_{|y|>T} \varphi(y)\,dy \le C_\varphi/T$. Thus

$$I_2 \ge -\bar{\Delta} \cdot \frac{C_\varphi}{T}.$$

To bound I_1, recall that Δ is bounded below by $\bar{\Delta}/2$ in the range of integration. Moreover, φ is a probability density function, so its total integral equals 1. Thus

$$I_1 \ge \frac{\bar{\Delta}}{2} \cdot \Big(1 - \frac{C_\varphi}{T}\Big).$$

Adding the two bounds, we conclude that

$$\Delta_f(a_0) \ge \frac{\bar{\Delta}}{2} - \frac{3C_\varphi \bar{\Delta}}{2T} \ge \frac{\bar{\Delta}}{2} - 9C_\varphi M.$$

Rearranging the terms yields $\bar{\Delta} \le 2\Delta_f(a_0) + 18C_\varphi M$, which completes the proof. □

In Lemma 3.1, we replaced the indicator function $\mathbf{1}_{(-\infty,0]}$ by a smooth function f. Now, we further replace f with a very particular choice: the complex exponential.

Lemma 3.2 (Smoothed discrepancy via characteristic functions)**.** *There exists a probability density function φ, which is also a Schwartz function, with the following property. Consider a smoothed version of the indicator function of $(-\infty, 0]$:*

$$f = \mathbf{1}_{(-\infty,0]} * \varphi.$$

Let X and Y be random variables. Then we have for any $\varepsilon > 0$:

$$\sup_{a\in\mathbb{R}} \Big| \mathbb{E}\, f\Big(\frac{X-a}{\varepsilon}\Big) - \mathbb{E}\, f\Big(\frac{Y-a}{\varepsilon}\Big) \Big| \le \int_{-1/\varepsilon}^{1/\varepsilon} \Big| \frac{\mathbb{E}\, e^{itX} - \mathbb{E}\, e^{itY}}{t} \Big|\, dt.$$

Proof. Step 1. Choosing a smoothing function. By translation and dilation, we can assume that $a = 0$ and $\varepsilon = 1$, so it suffices to prove the following version of the conclusion:

$$|\mathbb{E}\, f(X) - \mathbb{E}\, f(Y)| \le \int_{-1}^{1} \Big| \frac{\mathbb{E}\, e^{itX} - \mathbb{E}\, e^{itY}}{t} \Big|\, dt.$$

Choose any probability density function φ, which is also a Schwartz function, and whose Fourier transform is supported in the interval $\big[-1/(2\pi), 1/(2\pi)\big]$.

(Why does such φ exist? Take any even Schwartz function ψ supported in $[-1/4\pi, 1/4\pi]$ and satisfying $\int_{\mathbb{R}} \psi(x)^2\,dx = 1$, and set

$$\varphi(t) := \widehat{\psi}(t)^2.$$

Since ψ is even and Schwartz, $\widehat{\psi}$ is real-valued and Schwartz, so φ is nonnegative and Schwartz. Moreover, Plancherel identity (2.3) gives $\int_{\mathbb{R}} \widehat{\psi}(t)^2\,dt = \int_{\mathbb{R}} \psi(x)^2\,dx = 1$, showing that φ is a probability density function. Finally, we have $\widehat{\varphi} = \psi * \psi$: to see this, first apply the convolution theorem (2.4) and then the Fourier inversion formula (2.2), noting that ψ is even. Thus, $\widehat{\varphi}$ is supported on $\big[-1/(2\pi), 1/(2\pi)\big]$.)

Step 2. Bounding the Fourier transform. Since the function f approaches 1 at $-\infty$, it is not integrable. To fix this, let us consider the integrable function

$$f_M = \mathbf{1}_{(-M,0]} * \varphi.$$

Since the indicators $\mathbf{1}_{(-M,0]}$ increase to $\mathbf{1}_{(-\infty,0]}$ pointwise as $M \to \infty$, the monotone convergence theorem implies that the functions f_M increase to f pointwise. Another application of the monotone convergence theorem yields

$$\mathbb{E}\, f_M(X) \to \mathbb{E}\, f(X), \quad \mathbb{E}\, f_M(Y) \to \mathbb{E}\, f(Y) \quad \text{as } M \to \infty. \tag{3.5}$$

We claim that

$$\widehat{f_M} \text{ is supported on } \Big[-\frac{1}{2\pi}, \frac{1}{2\pi}\Big] \quad \text{and} \quad \big|\widehat{f_M}(t)\big| \le \frac{1}{\pi|t|}, \quad t \in \mathbb{R}. \tag{3.6}$$

Indeed, the convolution theorem (2.4) shows that $\widehat{f_M} = \widehat{\mathbf{1}_{(-M,0]}} \cdot \widehat{\varphi}$. Now,

$$\widehat{\mathbf{1}_{(-M,0]}}(t) = \int_{-M}^{0} e^{-2\pi i t x}\, dx = \frac{e^{2\pi i t M} - 1}{2\pi i t}, \quad \text{so} \quad \Big|\widehat{\mathbf{1}_{(-M,0]}}(t)\Big| \le \frac{1}{\pi|t|}.$$

Moreover, by construction, $|\widehat{\varphi}(t)|$ vanishes outside $\big[-1/(2\pi), 1/(2\pi)\big]$. It is bounded by $\int_{\mathbb{R}} |\varphi(x)|\, dx = 1$ since φ is a probability density function. Combining these bounds proves the claim.

Step 3. Integrating. Using the inverse Fourier transform formula, we can write

$$\mathbb{E}\, f_M(X) = \mathbb{E} \int_{\mathbb{R}} e^{2\pi i t X} \widehat{f_M}(t)\, dt = \int_{\mathbb{R}} \mathbb{E}\left[e^{2\pi i t X}\right] \widehat{f_M}(t)\, dt$$

by Fubini theorem (note that f_M is Schwartz by construction, so $\widehat{f_M}$ is Schwartz, too). Write the same for $\mathbb{E}\, f_M(Y)$, subtract and use Jensen's inequality to get

$$\begin{aligned}
|\mathbb{E}\, f_M(X) - \mathbb{E}\, f_M(Y)| &\le \int_{\mathbb{R}} \big|\mathbb{E}\, e^{2\pi i t X} - \mathbb{E}\, e^{2\pi i t Y}\big| \big|\widehat{f_M}(t)\big|\, dt \\
&\le \int_{-1/(2\pi)}^{1/(2\pi)} \big|\mathbb{E}\, e^{2\pi i t X} - \mathbb{E}\, e^{2\pi i t Y}\big| \cdot \frac{1}{\pi|t|}\, dt \quad \text{(using (3.6))}.
\end{aligned}$$

Make the change of variable $s = 2\pi t$, take limit as $M \to \infty$ using (3.5), and the proof is complete. $\square$

Combining Lemmas 3.1 and 3.2, we immediately obtain

Theorem 3.3 (Esseen's smoothing inequality [5])**.** *Let X and Y be random variables. Assume that the probability density function of Y is bounded by M. Then we have for any $\varepsilon > 0$:*

$$\sup_{a \in \mathbb{R}} \Big|\mathbb{P}\{X \le a\} - \mathbb{P}\{Y \le a\}\Big| \le 2 \int_{-1/\varepsilon}^{1/\varepsilon} \left|\frac{\mathbb{E}\, e^{itX} - \mathbb{E}\, e^{itY}}{t}\right| dt + CM\varepsilon.$$

Here C is an absolute constant.

4. Characteristic functions

Esseen's smoothing inequality (Theorem 3.3) reduces the problem of approximating a cumulative distribution function $\mathbb{P}\{X \le a\}$ to approximating the characteristic function $\mathbb{E}\, e^{itX}$. So, what can we say about the characteristic function?

Lemma 4.1 (The characteristic function of a random variable)**.** *Let X be a random variable with*

$$\mathbb{E}\, X = 0, \quad \mathbb{E}\, X^2 = \sigma^2, \quad \mathbb{E}|X|^3 = \rho^3 < \infty. \tag{4.1}$$

Then

$$\mathbb{E}\, e^{itX} = \exp\Big(-\frac{\sigma^2 t^2}{2} + O(\rho^3 t^3)\Big) \quad \textit{whenever} \quad |t| \le \frac{1}{\rho} \tag{4.2}$$

and

$$\big|\mathbb{E}\, e^{itX}\big| \le \exp\Big(-\frac{\sigma^2 t^2}{2} + O(\rho^3 t^3)\Big) \quad \textit{for any} \quad t \in \mathbb{R}. \tag{4.3}$$

In the statement and proof of this lemma, we use the $O(\cdot)$ notation to hide factors that are bounded by absolute constants. Precisely, $O(a)$ stands for θa where θ is some quantity that satisfies $|\theta| \le C$, where C is an absolute constant. The quantity θ and the constant C may change from line to line.

Proof. Step 1. Approximating the exponential function. To prove (4.2), write a Taylor approximation of the exponential function:

$$e^{ix} = 1 + ix - \frac{x^2}{2} + \theta_0 x^3 \quad \text{for some } \theta_0 = \theta_0(x) \text{ satisfying } |\theta_0| \le \frac{1}{6}.$$

(This holds since the third derivative of e^{ix} is bounded by 1 in modulus.) Substitute $x = tX$ and take expectation to get

$$\mathbb{E}\, e^{itX} = 1 + it\,\mathbb{E}\, X - \frac{t^2\,\mathbb{E}\, X^2}{2} + t^3\,\mathbb{E}[\theta_0 X^3]$$

for some random variable θ_0 satisfying $|\theta_0| \le \frac{1}{6}$ pointwise. Using the assumptions (4.1), we get

$$\mathbb{E}\, e^{itX} = 1 - \frac{\sigma^2 t^2}{2} + \theta \rho^3 t^3 \quad \text{for some } \theta = \theta(x) \text{ satisfying } |\theta| \le \frac{1}{6}. \tag{4.4}$$

For convenience, let us rewrite this as

$$\mathbb{E}\, e^{itX} = 1 - \frac{a}{2} + \theta b, \quad \text{where} \quad a = \sigma^2 t^2 \text{ and } b = \rho^3 t^3, \tag{4.5}$$

and note that

$$a^2 \le |b| \le 1. \tag{4.6}$$

(The second inequality follows from the assumption $|t| \le \frac{1}{\rho}$. To check the first inequality, note that $\sigma \le \rho$ by Jensen's inequality, so $a^2 = \sigma^4 t^4 \le \rho^4 t^4 = |b|^{4/3} \le |b|$.)

Step 2. Linearizing the logarithmic function. Now write a Taylor approximation of the logarithmic function:

$$\ln(1+x) = x + O(x^2) \quad \text{whenever} \quad |x| \le \frac{2}{3}.$$

We can use this for $x \coloneqq -\frac{a}{2} + \theta b$, since (4.6) and (4.4) guarantee that $|x| \le \frac{1}{2} + \frac{1}{6} = \frac{2}{3}$. We get

$$\ln\Big(1 - \frac{a}{2} + \theta b\Big) = -\frac{a}{2} + \theta b + O\Big(\big(-\frac{a}{2} + \theta b\big)^2\Big).$$

Expanding the square and letting the $O(\cdot)$ notation absorb the factors bounded by absolute constants, we conclude that

$$\ln\Big(1-\frac{a}{2}+\theta b\Big) = -\frac{a}{2}+O(b)+O(a^2)+O(ab)+O(b^2) = -\frac{a}{2}+O(b),$$

where the last step follows from (4.6). Recalling (4.5), we see that we proved that

$$\ln\big(\mathbb{E}\, e^{itX}\big) = -\frac{\sigma^2 t^2}{2}+O(\rho^3 t^3),$$

as claimed.

Step 3. Deducing the bound (4.3). In the range $|t|\le \frac{1}{\rho}$, the bound (4.3) follows from (4.2). And if $|t|>\frac{1}{\rho}$, the bound is nearly trivial. Indeed, since $|e^{ix}|=1$ holds pointwise, we have

$$|\mathbb{E}\, e^{itX}|\le 1.$$

On the other hand, $\sigma\le\rho$ and $\rho|t|>1$ yield $\sigma^2t^2/2\le\rho^2t^2/2\le\rho^3|t|^3$, so

$$\exp\Big(-\frac{\sigma^2t^2}{2}+\rho^3|t|^3\Big)\ge \exp(0)=1,$$

and (4.3) follows. □

Remark 4.2 (No approximation everywhere). One might ask whether the bound (4.2) could hold for all $t\in\mathbb{R}$, in which case we won't need the separate bound (4.3). This is false in general. For instance, a characteristic function $\mathbb{E}\, e^{itX}$ may have compact support; then the left-hand side of (4.2) vanishes for large $|t|$, while the right-hand side remains positive.

Our next goal is to approximate the characteristic function of a sum of independent random variables. To do this, we will multiply the bounds from Lemma 4.1 to obtain:

Lemma 4.3 (The characteristic function of a sum)**.** *There exist absolute constants $C, c>0$ so that the following holds. Let $X_1,\ldots,X_n$ be independent random variables that satisfy*

$$\mathbb{E}\, X_k=0,\quad \mathbb{E}\, X_k^2=\sigma_k^2,\quad \mathbb{E}|X_k|^3=\rho_k^3<\infty.$$

Assume that

$$\sum_{k=1}^n \sigma_k^2 = 1 \quad \textit{and let} \quad \rho^3 := \sum_{k=1}^n \rho_k^3.$$

Then the sum $S_n := X_1+\cdots+X_n$ satisfies

$$\Big|\mathbb{E}\, e^{itS_n}-e^{-t^2/2}\Big| \le C\rho^3|t|^3e^{-t^2/4} \quad \textit{whenever} \quad |t|\le\frac{c}{\rho^3}.$$

Proof. Step 1. Assume that $|t|\le\frac{1}{\rho}$. Then $|t|\le\frac{1}{\rho_k}$ for each k, which allows us to apply (4.2) and get

$$\mathbb{E}\, e^{itX_k} = \exp\Big(-\frac{\sigma_k^2t^2}{2}+\theta_k\rho_k^3t^3\Big) \quad \text{for some } \theta_k=\theta_k(t) \text{ satisfying } |\theta_k|\le C.$$

By independence, this yields

$$\mathbb{E}\, e^{itS_n} = \prod_{k=1}^n \mathbb{E}\, e^{itX_k} = \exp\Big(-\frac{t^2}{2}+\theta\rho^3t^3\Big) \quad \text{for some } \theta=\theta(t) \text{ satisfying } |\theta|\le C.$$

Therefore

$$\Big|\mathbb{E}\, e^{itS_n}-e^{-t^2/2}\Big| = e^{-t^2/2}\Big|e^{\theta\rho^3t^3}-1\Big| \le C_1\rho^3|t|^3e^{-t^2/2},$$

and we are done. (The last bound follows once we apply the Taylor approximation of the exponential function $|e^x - 1| \le |x|e^{|x|}$ for $x := \theta\rho^3 t^3$ and note that $|x| \le C$ by assumption on t.)

Step 2. Assume that $\frac{1}{\rho} < |t| \le \frac{c}{\rho^3}$. Arguing similarly to Step 1, but applying (4.3) instead, we obtain

$$\left|\mathbb{E}\, e^{itS_n}\right| \le \exp\Big(-\frac{t^2}{2} + \theta\rho^3 t^3\Big) \quad \text{for some } \theta = \theta(t) \text{ satisfying } |\theta| \le C.$$

Choosing the absolute constant $c > 0$ in the assumption on t small enough, we can make sure that $\theta\rho^3 t^3 \le t^2/4$. This gives

$$\left|\mathbb{E}\, e^{itS_n}\right| \le e^{-t^2/4}.$$

Hence, by triangle inequality, we conclude that

$$\left|\mathbb{E}\, e^{itS_n} - e^{-t^2/2}\right| \le e^{-t^2/4} + e^{-t^2/2} \le 2\rho^3|t|^3 e^{-t^2/4},$$

since $\rho^3|t|^3 \ge 1$ by assumption. The lemma is proved. □

5. Proof of Theorem 0.1

Now we are ready to prove the Berry–Esseen theorem, Theorem 0.1. Set

$$\rho^3 := \sum_{k=1}^n \mathbb{E}|X_k|^3.$$

Apply Esseen's smoothing inequality (Theorem 3.3) with $X = S_n$, $Y = G$, and $1/\varepsilon = c/\rho^3$. Since the density of G is bounded by an absolute constant and its characteristic function equals $e^{-t^2/2}$, we obtain

$$\sup_{a\in\mathbb{R}}\Big|\mathbb{P}\{S_n \le a\} - \mathbb{P}\{G \le a\}\Big| \le 2\int_{-c/\rho^3}^{c/\rho^3}\left|\frac{\mathbb{E}\, e^{itS_n} - e^{-t^2/2}}{t}\right| dt + C_1\rho^3.$$

Now substitute the bound on the characteristic function of S_n given by Lemma 4.3. We get

$$\sup_{a\in\mathbb{R}}\Big|\mathbb{P}\{S_n \le a\} - \mathbb{P}\{G \le a\}\Big| \le C_2\rho^3\int_{-\infty}^{\infty} t^2 e^{-t^2/4}\, dt + C_1\rho^3 \le C_3\rho^3.$$

The Berry–Esseen theorem is proved. □

Part II. The proof by Stein's method

This part is also self-contained. It restarts from the proof strategy and develops its own smoothing lemma, Stein equation, and inductive argument. Nothing from Part I is used.

6. The proof strategy

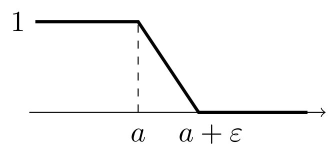
We begin with a bird's-eye, non-rigorous sketch of the proof of Theorem 0.1.

6.1. **Smoothing.** We begin by writing the cumulative distribution function of S_n as

$$\mathbb{P}\{S_n \le a\} = \mathbb{E}\,\mathbf{1}_{\{S_n \le a\}}. \tag{6.1}$$

The indicator function $x \mapsto \mathbf{1}_{\{x \le a\}}$ is difficult to handle analytically because it jumps from 1 to 0 at a. To soften this jump, fix $\varepsilon > 0$ and define the following *ε-smoothing function at a*:

$$h_a(x) \coloneqq \begin{cases} 1, & x \le a, \\ 1 - \dfrac{x-a}{\varepsilon}, & a < x < a + \varepsilon, \\ 0, & x \ge a + \varepsilon. \end{cases} \tag{6.2}$$



This function decreases linearly from 1 to 0 over the interval $[a, a+\varepsilon]$. With this smoothing function, (6.1) becomes

$$\mathbb{P}\{S_n \le a\} \approx \mathbb{E}\,h_a(S_n).$$

Lemma 7.1 makes this approximation rigorous and records the error introduced by smoothing.

6.2. **Gaussian integration by parts.** Let $\varphi(x) = (2\pi)^{-1/2}e^{-x^2/2}$ denote the standard normal density. Since $\varphi'(x) = -x\varphi(x)$, integration by parts gives

$$\mathbb{E}\,f'(G) = \int_{\mathbb{R}} f'(x)\varphi(x)\,dx = \int_{\mathbb{R}} x f(x)\varphi(x)\,dx = \mathbb{E}\,G f(G)$$

for every continuously differentiable function f such that f and f' are bounded. Thus, if S_n were Gaussian, the difference

$$\mathbb{E}\,f'(S_n) - \mathbb{E}\,S_n f(S_n) \tag{6.3}$$

would vanish. Stein's method turns this observation around: if (6.3) is small for suitable functions f, then S_n is approximately Gaussian.

6.3. **Stein's ODE.** Consider the following ordinary differential equation, which we call Stein's ODE:

$$f'(x) - x f(x) = h(x) - \mathbb{E}\,h(G), \qquad \text{where } G \sim N(0,1), \tag{6.4}$$

where $h = h_a$ is the ε-smoothing function (6.2). Suppose for the moment that this ODE has a well-behaved solution f; we will verify this in Sections 8 and 9. Substitute $x = S_n$ and take expectations:

$$\mathbb{E}\,f'(S_n) - \mathbb{E}\,S_n f(S_n) = \mathbb{E}\,h(S_n) - \mathbb{E}\,h(G). \tag{6.5}$$

Together with smoothing, this identity suggests that

$$\Big|\mathbb{P}\{S_n \le a\} - \mathbb{P}\{G \le a\}\Big| \approx \Big|\mathbb{E}\,h(S_n) - \mathbb{E}\,h(G)\Big| = \Big|\mathbb{E}\,S_n f(S_n) - \mathbb{E}\,f'(S_n)\Big|.$$

Our new task is therefore to show that

$$\mathbb{E}\,S_n f(S_n) \approx \mathbb{E}\,f'(S_n). \tag{6.6}$$

We have reduced the problem to estimating a quantity intrinsic to the sum S_n.

6.4. **Expansion.** Let us unpack the key quantity

$$\mathbb{E}\, S_n f(S_n) = \sum_{k=1}^{n} \mathbb{E}\, X_k f(S_n) \tag{6.7}$$

and estimate its terms separately, beginning with $\mathbb{E}\, X_n f(S_n)$. The factors X_n and $f(S_n)$ are not independent because S_n contains X_n. Let us isolate this summand:

$$S_n = S_{n-1} + X_n$$

The fundamental theorem of calculus gives

$$f(S_n) = f(S_{n-1}) + X_n \int_0^1 f'(S_{n-1} + tX_n)\, dt.$$

6.5. **Approximation.** Multiply both sides by X_n and take expectations:

$$\begin{aligned}
\mathbb{E}\, X_n f(S_n) &= \underbrace{\mathbb{E}\, X_n f(S_{n-1})}_{=0 \text{ by independence}} + \mathbb{E}\, X_n^2 \int_0^1 \underbrace{f'(S_{n-1} + tX_n)}_{\approx f'(S_{n-1}) \text{ if } f'' \text{ is controlled}} \, dt & (6.8)\\
&\approx \mathbb{E}\, X_n^2 f'(S_{n-1}) & (6.9)\\
&= \sigma_n^2\, \mathbb{E}\, f'(S_{n-1}) \quad (\text{by independence, denoting } \sigma_k^2 = \mathbb{E}\, X_k^2) &\\
&\approx \sigma_n^2\, \mathbb{E}\, f'(S_n) \quad (\text{if } f'' \text{ is controlled}). & (6.10)
\end{aligned}$$

The same argument applies to every term, giving

$$\mathbb{E}\, X_k f(S_n) \approx \sigma_k^2\, \mathbb{E}\, f'(S_n), \quad k = 1, \dots, n.$$

Summing these estimates and using (6.7), we obtain

$$\mathbb{E}\, S_n f(S_n) \approx \sum_{k=1}^{n} \sigma_k^2\, \mathbb{E}\, f'(S_n) = \mathbb{E}\, f'(S_n)$$

because the variances $\sigma_k^2 = \mathrm{Var}(X_k)$ sum to 1. This is precisely (6.6), and it completes the heuristic "proof".

6.6. **How to make this rigorous?** Apart from smoothing, which we handle in Section 7, the heuristic argument relies on one unverified approximation:

$$f'(S_{n-1} + y) \approx f'(S_{n-1}). \tag{6.11}$$

We used this approximation in (6.8) for $y = tX_n$ and in (6.10) for $y = X_n$. To verify this approximation, it is enough that y is small (which is intuitively true: X_n is just one term of a large sum S_n, so we expect it to be relatively small on average) and that f'' is controlled.

Here, f'' is the second derivative of a solution to Stein's ODE (6.4). We solve the ODE explicitly in Section 8 and study its solution in Section 9. This leads to Lemma 10.2, a rigorous version of (6.11).

We now carry out these steps.

7. Smoothing

We can write the cumulative distribution function of a random variable X as

$$\mathbb{P}\{X \le a\} = \mathbb{E}\,\mathbf{1}_{(-\infty,a]}(X).$$

We approximate the indicator function $\mathbf{1}_{(-\infty,a]}$ by the ε-smoothing function h_a defined in (6.2). For the moment, think of ε as a fixed small quantity. We will eventually choose $\varepsilon = C\sum_{k=1}^n \mathbb{E}|X_k|^3$, which gives the error in the Berry–Esseen theorem. We first record the error introduced by smoothing.

Lemma 7.1 (Smoothing)**.** *Let h_a be the ε-smoothing function at a, and let X and Y be random variables. Assume that the probability density function of Y is bounded by M. Then*

$$\sup_{a\in\mathbb{R}}\Big|\mathbb{P}\{X \le a\} - \mathbb{P}\{Y \le a\}\Big| \le \sup_{a\in\mathbb{R}}\Big|\mathbb{E}\,h_a(X) - \mathbb{E}\,h_a(Y)\Big| + M\varepsilon.$$

Proof. Note the pointwise inequality

$$\mathbf{1}_{(-\infty,a]} \le h_a \le \mathbf{1}_{(-\infty,a+\varepsilon]}. \tag{7.1}$$

Then

$$\begin{aligned}\mathbb{P}\{X \le a\} - \mathbb{P}\{Y \le a\} &\le \mathbb{P}\{X \le a\} - \mathbb{P}\{Y \le a+\varepsilon\} + M\varepsilon\\ &= \mathbb{E}\,\mathbf{1}_{(-\infty,a]}(X) - \mathbb{E}\,\mathbf{1}_{(-\infty,a+\varepsilon]}(Y) + M\varepsilon\\ &\le \mathbb{E}\,h_a(X) - \mathbb{E}\,h_a(Y) + M\varepsilon.\end{aligned}$$

For the reverse bound, apply (7.1) with $a-\varepsilon$ in place of a. The same argument gives

$$\mathbb{P}\{Y \le a\} - \mathbb{P}\{X \le a\} \le \mathbb{E}\,h_{a-\varepsilon}(Y) - \mathbb{E}\,h_{a-\varepsilon}(X) + M\varepsilon.$$

Combining the two bounds completes the proof. $\square$

8. Solving Stein's ODE

Fix a bounded continuous function $h : \mathbb{R} \to \mathbb{R}$ and consider Stein's ODE (6.4), restated here:

$$f'(x) - xf(x) = h(x) - \mathbb{E}\,h(G), \qquad \text{where } G \sim N(0,1). \tag{8.1}$$

We solve this ODE by the method of integrating factors. Multiplying both sides by $e^{-x^2/2}$ gives

$$\Big(e^{-x^2/2}f(x)\Big)' = e^{-x^2/2}\Big(h(x) - \mathbb{E}\,h(G)\Big).$$

The fundamental theorem of calculus now gives

$$e^{-x^2/2}f(x) = \int_{-\infty}^x e^{-y^2/2}\Big(h(y) - \mathbb{E}\,h(G)\Big)\,dy + C.$$

Setting $C = 0$, we obtain the canonical solution

$$f(x) = e^{x^2/2}\int_{-\infty}^x e^{-y^2/2}\Big(h(y) - \mathbb{E}\,h(G)\Big)\,dy. \tag{8.2}$$

Any other solution differs from this one by $Ce^{x^2/2}$, so the canonical solution is the unique bounded solution. Moreover, the total integral is zero:

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-y^2/2}\Big(h(y) - \mathbb{E}\,h(G)\Big)\,dy = \mathbb{E}\,h(G) - \mathbb{E}\,h(G) = 0,$$

and therefore we may rewrite (8.2) as

$$f(x) = -e^{x^2/2} \int_x^\infty e^{-y^2/2} \Big(h(y) - \mathbb{E}\, h(G) \Big) \, dy. \tag{8.3}$$

9. Properties of the solution of Stein's ODE

The two formulas (8.2) and (8.3) give three simple bounds on the canonical solution of Stein's ODE.

Lemma 9.1 (Bounds on the solution of Stein's ODE)**.** *Let $h : \mathbb{R} \to [0,1]$ be continuous. Then the canonical solution (8.2) of Stein's ODE (8.1) satisfies, for every $x \in \mathbb{R}$,*

$$|f(x)| \le 2; \qquad |xf(x)| \le 1; \qquad |f'(x)| \le 2. \tag{9.1}$$

Proof. First, assume that $x > 0$. Since $0 \le h \le 1$ everywhere, (8.3) gives

$$|f(x)| \le e^{x^2/2} \int_x^\infty e^{-y^2/2} \, dy = \int_0^\infty e^{-xt - t^2/2} \, dt$$

by the change of variables $y = x + t$. Dropping the term $-xt$ gives

$$|f(x)| \le \int_0^\infty e^{-t^2/2} \, dt = \sqrt{\frac{\pi}{2}} \le 2,$$

proving the first bound in (9.1) for $x > 0$. Dropping the term $-t^2/2$ instead gives

$$|f(x)| \le \int_0^\infty e^{-xt} \, dt = \frac{1}{x},$$

proving the second bound in (9.1) for $x > 0$. For $x \le 0$, the same argument applies using (8.2) in place of (8.3).

Finally, Stein's ODE (8.1) and the triangle inequality give

$$|f'(x)| \le |xf(x)| + |h(x) - \mathbb{E}\, h(G)|.$$

The first term on the right is bounded by 1, as proved above. The second is also bounded by 1 because the range of h is $[0,1]$. Hence $|f'(x)| \le 2$. □

We return to (6.11), where we need to show that f' varies slowly. For this, we examine f''.

We now specialize to the functions needed in the proof. Let $h = h_a$ be the ε-smoothing function at a defined in (6.2). This function is differentiable except at a and $a + \varepsilon$. Differentiating Stein's ODE (8.1) gives

$$f''(x) = f(x) + xf'(x) + h'(x)$$

whenever $x \ne a, a + \varepsilon$. By (9.1), $|f(x)| \le 2$ and $|f'(x)| \le 2$, while $h'(x) = -\frac{1}{\varepsilon} \mathbf{1}_{[a,a+\varepsilon]}(x)$ wherever the derivative exists. Thus

$$|f''(x)| \le 2 + 2|x| + \frac{1}{\varepsilon} \mathbf{1}_{[a,a+\varepsilon]}(x). \tag{9.2}$$

A bound on f'' controls the increments of f'. Indeed, the fundamental theorem of calculus gives[1]

$$f'(x+y) - f'(x) = \int_x^{x+y} f''(t) \, dt = y \int_0^1 f''(x + sy) \, ds.$$

[1] More precisely, the smoothing function $h = h_a$ is Lipschitz and f is continuously differentiable. Hence, by (8.1), $f' = xf + h - \mathbb{E}\, h(G)$ is locally absolutely continuous. It follows that $f'' = f + xf' + h'$ almost everywhere, and the fundamental theorem of calculus applies to f'.

Thus,

$$|f'(x+y)-f'(x)| \le |y| \int_0^1 |f''(x+sy)|\,ds.$$

Substituting $x+sy$ for x in (9.2) and using $|x+sy| \le |x|+|y|$, we obtain

$$|f'(x+y)-f'(x)| \le |y| \left(2+2|x|+2|y| + \frac{1}{\varepsilon} \int_0^1 \mathbf{1}_{[a,a+\varepsilon]}(x+sy)\,ds \right). \tag{9.3}$$

This simplifies to the following deterministic bound.

Lemma 9.2 (A deterministic increment bound)**.** *Let $h = h_a$ be the ε-smoothing function at a, defined in* (6.2)*. Then the canonical solution* (8.3) *of Stein's ODE* (8.1) *satisfies for all $x, y \in \mathbb{R}$:*

$$|f'(x+y)-f'(x)| \le 4|y| \left(1+|x| + \frac{1}{\varepsilon} \int_0^1 \mathbf{1}_{[a,a+\varepsilon]}(x+sy)\,ds \right). \tag{9.4}$$

Proof. If $|y| \le 1$, the conclusion follows from (9.3) after replacing $2|y|$ by 2. If $|y| > 1$, the left-hand side of (9.4) is bounded by 4 by (9.1), while its right-hand side is at least $4|y| > 4$. □

10. Preparing the induction

We will soon average the deterministic increment bound over the partial sum S_{n-1}. But before that, let us introduce the error term in Berry-Esseen theorem we want to bound:

Definition 10.1 (The error in the Berry–Esseen theorem)**.** *Let $\delta_n(\rho)$ denote the infimum of all $\delta > 0$ such that*

$$\sup_{a\in\mathbb{R}} \Big| \mathbb{P}\{S_n \le a\} - \mathbb{P}\{G \le a\} \Big| \le \delta$$

whenever $S_n = X_1 + \cdots + X_n$ is a sum of independent random variables satisfying

$$\mathbb{E}\, X_k = 0, \qquad \sum_{k=1}^n \mathbb{E}\, X_k^2 = 1, \qquad \sum_{k=1}^n \mathbb{E}|X_k|^3 \le \rho^3. \tag{10.1}$$

Our goal is to prove that[2]

$$\delta_n(\rho) \le C\rho^3 \tag{10.2}$$

for every n and ρ, where C is an absolute constant.

We may restrict ourselves to the case where no single summand accounts for more than half of the total variance:

$$\mathbb{E}\, X_k^2 \le \frac{1}{2} \qquad \text{for every } k = 1, \ldots, n. \tag{10.3}$$

Indeed, if $\mathbb{E}\, X_k^2 > \frac{1}{2}$ for some k, then

$$\sum_{j=1}^n \mathbb{E}|X_j|^3 \ge \mathbb{E}|X_k|^3 \ge \big(\mathbb{E}\, X_k^2\big)^{3/2} > 2^{-3/2} > \frac{1}{3}.$$

Since $\delta_n(\rho) \le 1$ holds trivially, the desired bound (10.2) is then immediate as long as we choose $C \ge 3$. Thus, it remains to consider the case (10.3).

We are now ready for a rigorous version of the key approximation (6.11).

[2]Here and throughout the rest of the proof, the notation $a \lesssim b$ means $a \le Cb$ for an absolute constant C.

Lemma 10.2 (An averaged increment bound)**.** *Let $h = h_a$ be the ε-smoothing function at a defined in (6.2), and let f be the canonical solution (8.2) of Stein's ODE. Suppose that*

$$S_n = X_1 + \ldots + X_n, \quad n \geq 2,$$

is a sum of independent random variables satisfying (10.1) and (10.3). Then, for every fixed $y \in \mathbb{R}$,

$$\mathbb{E}\Big|f'(S_{n-1} + y) - f'(S_{n-1})\Big| \leq C|y| \left(1 + \frac{\delta_{n-1}(2\rho)}{\varepsilon}\right).$$

Proof. Apply Lemma 9.2 with $x = S_{n-1}$ and take expectations:

$$\mathbb{E}\Big|f'(S_{n-1} + y) - f'(S_{n-1})\Big| \lesssim |y| \left(1 + \mathbb{E}|S_{n-1}| + \frac{1}{\varepsilon}\int_0^1 \mathbb{P}\big\{S_{n-1} + sy \in [a, a+\varepsilon]\big\}\, ds\right) \quad (10.4)$$

By Cauchy–Schwarz,

$$\mathbb{E}|S_{n-1}| \leq \sqrt{\mathrm{Var}(S_{n-1})} \leq \sqrt{\mathrm{Var}(S_n)} = 1. \quad (10.5)$$

It remains to bound $\mathbb{P}\big\{S_{n-1} + sy \in [a, a+\varepsilon]\big\}$. Set

$$\sigma^2 \coloneqq \mathrm{Var}(S_{n-1}) = \sum_{k=1}^{n-1} \mathbb{E}\, X_k^2 \geq \frac{1}{2}$$

where the last inequality follows from (10.1) and (10.3). Thus, S_{n-1}/σ is a sum of $n-1$ independent mean-zero random variables X_k/σ. Their second moments sum to 1, and the sum of their third moments is at most $(\rho/\sigma)^3$. Comparing the distribution functions at the two endpoints of the interval introduces two error terms. Therefore,

$$\begin{aligned}
\mathbb{P}\big\{S_{n-1} + sy \in [a, a+\varepsilon]\big\} &= \mathbb{P}\big\{S_{n-1}/\sigma \in [b, b+\varepsilon/\sigma]\big\} \quad (\text{where } b = (a - sy)/\sigma) \\
&\leq \mathbb{P}\big\{G \in [b, b+\varepsilon/\sigma]\big\} + 2\delta_{n-1}(\rho/\sigma) \quad (\text{by Definition 10.1}) \\
&\lesssim \varepsilon + \delta_{n-1}(2\rho)
\end{aligned}$$

because $\sigma \geq 1/2$ and the density of G is bounded by an absolute constant. Substituting this estimate and (10.5) into (10.4) gives

$$\mathbb{E}\Big|f'(S_{n-1} + y) - f'(S_{n-1})\Big| \lesssim |y| \left(1 + \frac{1}{\varepsilon}\big(\varepsilon + \delta_{n-1}(2\rho)\big)\right),$$

which proves the lemma. □

The appearance of δ_{n-1} is the key. We have used the quantity we want to estimate, but with one fewer summand, so induction on n becomes possible.

11. Proof of Theorem 0.1

We now combine smoothing, the averaged increment bound, and induction on the number of summands. For each $a \in \mathbb{R}$, let h_a be the ε-smoothing function at a defined in (6.2), and let f_a be the corresponding canonical solution (8.3) of Stein's ODE. Then

$$\begin{aligned}
\sup_{a\in\mathbb{R}}\Big|\mathbb{P}\{S_n \leq a\} - \mathbb{P}\{G \leq a\}\Big| &\leq \sup_{a\in\mathbb{R}}\Big|\mathbb{E}\, h_a(S_n) - \mathbb{E}\, h_a(G)\Big| + C\varepsilon \\
&= \sup_{a\in\mathbb{R}}\Big|\mathbb{E}\, S_n f_a(S_n) - \mathbb{E}\, f_a'(S_n)\Big| + C\varepsilon. \quad (11.1)
\end{aligned}$$

The first step is Lemma 7.1; the second is Stein's ODE, as in (6.5).

Fix $a \in \mathbb{R}$ and suppress the subscript in f_a. We again unpack the key quantity

$$\mathbb{E}\, S_n f(S_n) = \sum_{k=1}^n \mathbb{E}\, X_k f(S_n). \tag{11.2}$$

Set

$$\sigma_k^2 \coloneqq \mathbb{E}\, X_k^2.$$

Expanding $f(S_n)$ around S_{n-1} as in the heuristic argument (6.8)–(6.10), we obtain

$$\Big|\mathbb{E}\, X_n f(S_n) - \sigma_n^2\, \mathbb{E}\, f'(S_n)\Big| \le A + B, \tag{11.3}$$

where

$$A = \mathbb{E}\, X_n^2 \int_0^1 \Big| f'(S_{n-1} + tX_n) - f'(S_{n-1})\Big|\, dt; \qquad B = \sigma_n^2\, \mathbb{E}\Big| f'(S_n) - f'(S_{n-1})\Big|.$$

(A is the approximation error in (6.9), and B is the approximation error in (6.10).)

To bound A, condition on X_n and write $\mathbb{E}_n[\cdot] \coloneqq \mathbb{E}[\cdot \mid X_n]$. The law of total expectation gives

$$A = \mathbb{E}\left[X_n^2 \int_0^1 \underbrace{\mathbb{E}_n \Big| f'(S_{n-1} + tX_n) - f'(S_{n-1})\Big|}_{\lesssim |X_n|(1+\delta_{n-1}(2\rho)/\varepsilon) \text{ by Lemma 10.2}}\, dt\right] \lesssim \underbrace{\mathbb{E}|X_n|^3}_{=:\rho_n^3} \left(1 + \frac{\delta_{n-1}(2\rho)}{\varepsilon}\right).$$

Similarly,

$$B = \sigma_n^2\, \mathbb{E}\left[\underbrace{\mathbb{E}_n \Big| f'(S_{n-1} + X_n) - f'(S_{n-1})\Big|}_{\lesssim |X_n|(1+\delta_{n-1}(2\rho)/\varepsilon) \text{ by Lemma 10.2}}\right] \lesssim \underbrace{\sigma_n^2\, \mathbb{E}|X_n|}_{\le \rho_n^3} \left(1 + \frac{\delta_{n-1}(2\rho)}{\varepsilon}\right).$$

Here we used $\sigma_n \le \rho_n$ and $\mathbb{E}|X_n| \le \rho_n$. Combining the bounds for A and B controls (11.3). The same computation applies to every k: relabel the summands so that X_k is last. Thus,

$$\Big|\mathbb{E}\, X_k f(S_n) - \sigma_k^2\, \mathbb{E}\, f'(S_n)\Big| \lesssim \rho_k^3 \left(1 + \frac{\delta_{n-1}(2\rho)}{\varepsilon}\right).$$

Summing these estimates in (11.2) and using $\sum_{k=1}^n \sigma_k^2 = 1$, we obtain

$$\Big|\mathbb{E}\, S_n f(S_n) - \mathbb{E}\, f'(S_n)\Big| \lesssim \rho^3 \left(1 + \frac{\delta_{n-1}(2\rho)}{\varepsilon}\right) \quad \text{for any } \rho \text{ such that } \sum_{k=1}^n \rho_k^3 \le \rho^3.$$

This bound is uniform in $a \in \mathbb{R}$. Substituting it into (11.1) and restoring the subscript a, we obtain

$$\sup_{a \in \mathbb{R}} \Big| \mathbb{P}\{S_n \le a\} - \mathbb{P}\{G \le a\}\Big| \lesssim \rho^3 \left(1 + \frac{\delta_{n-1}(2\rho)}{\varepsilon}\right) + \varepsilon.$$

By Definition 10.1,

$$\delta_n(\rho) \le C\rho^3 \left(1 + \frac{\delta_{n-1}(2\rho)}{\varepsilon}\right) + C\varepsilon.$$

Choosing $\varepsilon \coloneqq 10C\rho^3$, we get

$$\delta_n(\rho) \le C'\rho^3 + \frac{\delta_{n-1}(2\rho)}{10},$$

where $C' \ge 1$ is an absolute constant. This recurrence closes the induction. Assume that

$$\delta_{n-1}(\rho) \le 5C'\rho^3 \quad \text{for any } \rho.$$

Then

$$\delta_n(\rho) \le C'\rho^3 + \frac{5C' \cdot 8\rho^3}{10} = 5C'\rho^3 \quad \text{for any } \rho,$$

which verifies the induction step. The base case is immediate: $\delta_1(\rho) \le 1$, while the assumptions give

$$\rho^3 \ge \mathbb{E}|X_1|^3 \ge (\mathbb{E}\, X_1^2)^{3/2} = 1.$$

The Berry–Esseen theorem is proved. □

Acknowledgements

I am grateful to Pedro Abdalla and Guangyi Zou, who kindly read an early version of the Fourier-analytic proof and helped eliminate several typos and inaccuracies.

Statement on AI use. The mathematical arguments and core exposition in this paper are entirely the author's own. After the two notes were completed, ChatGPT Pro was used as an "artificial referee" to review the mathematics, polish the writing, and assist in combining the notes.

References

[1] A. C. Berry, *The accuracy of the Gaussian approximation to the sum of independent variates*, Transactions of the AMS 49 (1941), 122–136.

[2] E. Bolthausen, *An estimate of the remainder in a combinatorial central limit theorem*, Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 66 (1984), 379–386.

[3] L. H. Y. Chen, L. Goldstein, and Q.-M. Shao, *Normal approximation by Stein's method*, Probability and its Applications, Springer, 2011.

[4] C.-G. Esseen, *On the Liapunoff limit of error in the theory of probability*, Arkiv för Matematik, Astronomi och Fysik A28 (1942), 1–19.

[5] C.-G. Esseen, *Fourier analysis of distribution functions. A mathematical study of the Laplace–Gaussian law*, Acta Math. 77 (1945), 1–125.

[6] W. Feller, *An introduction to probability theory and its applications*, 2nd ed., Vol. II, John Wiley & Sons, 1971.

[7] N. Ross, *Fundamentals of Stein's method*, Probability Surveys 8 (2011), 210–293.

Department of Mathematics, University of California, Irvine, U.S.A.
Email address: rvershyn@uci.edu