## 6 Convolution, Smoothing, and Weak Convergence

### 6.1 Convolution and Smoothing

Definition 6.1. Let $\mu, v$ be finite Borel measures on $\mathbb{R}$. The convolution $\mu * v=v * \mu$ is the unique finite Borel measure on $\mathbb{R}$ such that for every bounded continuous function $f: \mathbb{R} \rightarrow \mathbb{R}$,

$$
\begin{equation*}
\int f d(\mu * v)=\iint f(x+y) d \mu(x) d v(y)=\iint f(x+y) d v(y) d \mu(x) . \tag{6.1}
\end{equation*}
$$

This notion is perhaps most easily understood in the language of induced measures (sec. 2.3). If $\mu \times v$ is the product measure on $\mathbb{R} \times \mathbb{R}=\mathbb{R}^{2}$, and $T: \mathbb{R}^{2} \rightarrow \mathbb{R}$ the mapping $T(x, y)=x+y$, then $T$ induces a Borel measure $(\mu \times v) \circ T^{-1}$ on $\mathbb{R}$ : this is the convolution $\mu * v$. From this point of view the equations (6.1) are just a transparent reformulation of Fubini's theorem, because by definition of $T$

$$
\iint f(x+y) d \mu(x) d v(y)=\int f \circ T d(\mu \times v) .
$$

In the special case where $\mu$ and $v$ are both probability measures, this equation identifies the convolution $\mu * v$ as the distribution of $X+Y$, where $X, Y$ are independent random variables with marginal distributions $\mu$ and $v$, respectively.

Recall that if $g: \mathbb{R} \rightarrow \mathbb{R}_{+}$is a nonnegative, integrable function relative to Lebesgue measure $\lambda$ then $g$ determines a finite Borel measure $v=v_{g}$ on $\mathbb{R}$ by $v(B)=\int_{\mathbb{R}} \mathbf{1}_{B} g d \lambda$. What happens when this measure is convolved with another finite Borel measure $\mu$ ? By the translation invariance of Lebesgue measure and Tonelli's theorem, for any nonnegative, continuous, bounded function $f$,

$$
\begin{aligned}
\int f d(\mu \times v) & =\iint f(x+y) d v(y) d \mu(x) \\
& =\int\left(\int f(x+y) g(y) d \lambda(y)\right) d \mu(x) \\
& =\int\left(\int f(y) g(y-x) d \lambda(y)\right) d \mu(x) \\
& =\int f(y)\left(\int g(y-x) d \mu(x)\right) d \lambda(y) \\
& =\int f(y)(g * \mu)(y) d \lambda(y)
\end{aligned}
$$

where $g * \mu$ is defined to be the (integrable) function

$$
\begin{equation*}
g * \mu(y)=\int g(y-x) d \mu(x) \tag{6.2}
\end{equation*}
$$

This proves the following result.

Proposition 6.2. If $\mu, v$ are finite Borel measures such that $v$ has a density $g$ with respect to Lebesgue measure, then the convolution $\mu * v$ has density $g * \mu$, as defined by (6.2).

The definition (6.2) can also be used for functions $g$ that are not necessarily nonnegative, and also for functions $g$ that are bounded but not necessarily integrable. When the measure $\mu$ in (6.2) is a probability measure, the integral in (6.2) can be interpreted as an expectation: in particular, if $Y$ is a random variable with distribution $\mu$ then

$$
g * \mu(x)=\int g(y-x) d \mu(x)=E g(x-Y)
$$

Proposition 6.3. If g is a bounded, continuously differentiable function whose derivative $g^{\prime}$ is also bounded then for any finite Borel measure $\mu$ the convolution $g * \mu$ is bounded and continuously differentiable, with derivative

$$
\begin{equation*}
(g * \mu)^{\prime}(x)=\left(g^{\prime}\right) * \mu(x)=\int g^{\prime}(x-y) d \mu(y) \tag{6.3}
\end{equation*}
$$

Proof. This is a routine exercise in the use of the dominated convergence theorem. To show that the difference quotients $(g(x+\varepsilon)-g(x)) / \varepsilon$ are dominated, use the mean value theorem of calculus and the hypothesis that the derivative $g^{\prime}$ is a bounded function.

Corollary 6.4. If g is a $k$-times continuously differentiable function (i.e., a $C^{k}$ function) on $\mathbb{R}$ with compact support then for any finite Borel measure $\mu$ the convolution $g * \mu$ is also of class $C^{k}$, and for each $j \leq k$ the $j$ th derivative of $g * \mu$ is $g^{(j)} * \mu$.

Proposition 6.5. If g is a $C^{k}$ function on $\mathbb{R}$ with compact support and ifv is the uniform distribution on a finite interval $[a, b]$ then the convolution $g * v$ is of class $C^{k+1}$, and

$$
\begin{equation*}
(g * v)(j+1)(x)=(b-a)^{-1}\left(g^{(j)}(x-a)-g^{(j)}(x-b)\right) \tag{6.4}
\end{equation*}
$$

Proof. By induction it suffices to show this for $k=j=0$, that is, that convolving a continuous function of compact support with the uniform distribution produces a continuously differentiable function. For this, evaluate the differences

$$
\begin{aligned}
(b-a)(g * v(x+\varepsilon)-g * v(x)) & =\int_{a}^{b} g(x+\varepsilon-t) d t-\int_{a}^{b} g(x-t) d t \\
& =-\int_{x-b}^{x+\varepsilon-b} g(y) d y+\int_{x-a}^{x+\varepsilon-a} g(y) d y .
\end{aligned}
$$

Now divide by $\varepsilon$ and take the limit as $\varepsilon \rightarrow 0$, using the fundamental theorem of calculus; this gives

$$
(g * v)^{\prime}(x)=(b-a)^{-1}(g(x-a)-g(x-b)) .
$$

Corollary 6.6. There exists an even, $C^{\infty}$ probability density $\psi(x)$ on $\mathbb{R}$ with support $[-1,1]$.
Proof. Let $U_{1}, U_{2}, \ldots$ be independent, identically distributed uniform-[-1,1] random variables and set

$$
Y=\sum_{n=1}^{\infty} U_{n} / 2^{n} .
$$

The random variable $Y$ is certainly between -1 and 1 , and its distribution is symmetric about 0 , so if it has a density $\psi$ the density must be an even function with support $[-1,1]$.

To show that $Y$ does have a density, we will first look at the distributions of the partial sums

$$
Y_{m}=\sum_{n=1}^{m} U_{n} / 2^{n} .
$$

Brute force calculation (exercise) shows that $Y_{2}$ has a continuous density with support $\left[-\frac{3}{4}, \frac{3}{4}\right]$ whose graph is a (piecewise linear) trapezoid with vertices

$$
(-3 / 4,0),(-1 / 4,1),(+1 / 4,1),(+3 / 4,0)
$$

Now $Y_{3}=Y_{2}+U_{3} / 8$, so its distribution is the convolution of the distribution of $Y_{2}$ with the uniform distribution on $\left[-\frac{1}{8}, \frac{1}{8}\right]$; hence, by Proposition 6.5, the distribution of $Y_{3}$ has a $C^{1}$ density. By an easy induction argument, for any $m \geq 3$ the distribution of $Y_{m}$ has a $C^{m-2}$ density.

Finally, consider the distribution of $Y$. Since $Y=Y_{m}+\sum_{n \geq m+1} U_{n} / 2^{n}$, its distribution is the convolution of the distribution of $Y_{m}$ with that of $\sum_{n \geq m+1} U_{n} / 2^{n}$. Since $Y_{m}$ has a density of class $C^{m-2}$, Corollary 6.6 implies that $Y$ has a density of class $C^{m-2}$. But since $m$ can be taken arbitrarily large, it follows that the density of $Y$ is $C^{\infty}$.

Proposition 6.7. Let $g: \mathbb{R} \rightarrow \mathbb{R}$ be a bounded, continuous function, and let $\varphi$ be a continuous probability density with support $[-1,1]$. For each $1>\varepsilon>0$, let $v_{\varepsilon}$ be the Borel probability measure with density $\varepsilon^{-1} \varphi(x / \varepsilon)$. Then for each $x \in \mathbb{R}$,

$$
\begin{equation*}
\lim _{\varepsilon \rightarrow 0} g * v_{\varepsilon}(x)=g(x), \tag{6.5}
\end{equation*}
$$

and the convergence holds uniformly for $x$ in any compact interval.
Proof. For any $L>0$ the function $g$ is uniformly continuous on $[-L-1, L+1]$, so for any $\delta>0$ there exists $\varepsilon>0$ so small that $|g(x)-g(y)|<\delta$ for any $x, y \in[-L-1, L+1]$ such that $|x-y|<\varepsilon$. But for $x \in[-L, L]$,

$$
\left(g * v_{\varepsilon}\right)(x)-g(x)=\int_{-\varepsilon}^{\varepsilon}(g(x-z)-g(x)) v_{\varepsilon}(d z)
$$

so

$$
\left|\left(g * v_{\varepsilon}\right)(x)-g(x)\right| \leq \delta .
$$

Corollary 6.8. The $C^{\infty}$ functions are dense in the space of continuous functions with compact support, that is, for any continuous $f: \mathbb{R} \rightarrow \mathbb{R}$ with compact support and any $\delta>0$ there exists a $C^{\infty}$ function $g$ with compact support such that

$$
\begin{equation*}
\|f-g\|_{\infty}:=\sup _{x \in \mathbb{R}}|f(x)-g(x)|<\delta . \tag{6.6}
\end{equation*}
$$

Proof. Set $g=f * v_{\varepsilon}$ where $v_{\varepsilon}$ has density $\varepsilon^{-1} \varphi(x / \varepsilon)$ and $\varphi$ is a $C^{\infty}$ probability density with support $[-1,1]$.

### 6.2 Weak Convergence

Definition 6.9. A sequence of Borel probability measures $\mu_{n}$ on $\mathbb{R}^{k}$ converges weakly to a Borel probability measure $\mu$ on $\mathbb{R}^{k}$ if for every continuous function $f: \mathbb{R}^{k} \rightarrow \mathbb{R}$ with compact support,

$$
\begin{equation*}
\lim _{n \rightarrow \infty} \int f d \mu_{n}=\int f d \mu \tag{6.7}
\end{equation*}
$$

A sequence of $k$-dimensional random vectors $X_{n}$ is said to converge in distribution ${ }^{4}$ if their distributions $\mu_{n}$ convergence weakly to a probability distribution $\mu$, i.e., if for every continuous, compactly supported function $f: \mathbb{R}^{k} \rightarrow \mathbb{R}$,

$$
\begin{equation*}
\lim _{n \rightarrow \infty} E f\left(X_{n}\right)=\int f d \mu \tag{6.8}
\end{equation*}
$$

Lemma 6.10. If (6.7) holds for all $C^{\infty}$ functions with compact support then it holds for all continuous functions with compact support.

Proof. This is an easy consequence of Corollary 6.8.

## Exercises.

Exercise 6.11. Let $\mu_{n}$ and $\mu$ be Borel probability measures on $\mathbb{R}$ with cumulative distribution functions $F_{n}$ and $F$ i.e., for every real number $x$

$$
\mu_{n}(-\infty, x]=F_{n}(x) \quad \text { and } \quad \mu(-\infty, x]=F(x) .
$$

Prove that the following conditions are equivalent:
(a) $\mu_{n} \Rightarrow \mu$;
(b) $\lim _{n \rightarrow \infty} F_{n}(x)=F(x)$ for every continuity point $x$ of $F$; and
(c) on some probability space $(\Omega, \mathscr{F}, P)$ there are random variables $X_{n}, X$ such that

[^0](i) $X_{n}$ has distribution $\mu_{n}$;
(ii) $X$ has distribution $\mu$; and
(iii) $X_{n} \rightarrow X$ almost surely.

HinT: $(\mathrm{a}) \Longrightarrow(\mathrm{b}) \Longrightarrow(\mathrm{c}) \Longrightarrow(\mathrm{a})$.
Exercise 6.12. Show that a sequence $X_{n}$ of integer-valued random variables converges in distribution to a probability measure $\mu$ if and only if $\mu$ is supported by the integers and for every $k \in \mathbb{Z}$,

$$
\lim _{n \rightarrow \infty} P\left\{X_{n}=k\right\}=\mu(\{k\}) .
$$

Exercise 6.13. Let $X_{n} \sim \operatorname{Binomial}-\left(n, p_{n}\right)$. Show that if $n p_{n} \rightarrow \lambda>0$ then $X_{n} \Rightarrow$ Poisson distribution with mean $\lambda$, that is, for every $k=0,1,2, \ldots$,

$$
\lim _{n \rightarrow \infty} P\left(X_{n}=k\right)=\frac{\lambda^{k}}{k!} e^{-\lambda}
$$

Hint: First show that if $\lambda_{n \rightarrow \lambda}$ then $\left(1-\lambda_{n} / n\right)^{n} \longrightarrow e^{-\lambda}$.
Definition 6.14. A sequence of Borel probability measures on $\mathbb{R}^{k}$ is said to be tight if for every $\varepsilon>0$ there is a compact subset $K \subset \mathbb{R}^{k}$ such that

$$
\inf _{n} \mu_{n}(K) \geq 1-\varepsilon .
$$

Exercise 6.15. Show that if $\mu_{n} \Longrightarrow \mu$ then the sequence $\mu_{n}$ is tight.
Theorem 6.16. (Helly's Selection Principle) If $\left\{\mu_{n}\right\}_{n \geq 1}$ is a tight sequence of Borel probability measures on $\mathbb{R}$ then there is a subsequence $\mu_{n_{k}}$ that converges weakly to a probability measure $\mu$.

NOTE: Different subsequences could, in principle, have different weak limits. (Exercise: FInd an example.) However, if one can show by other means that there is only one possible subsequential limit $\mu$, then it will follow that $\mu_{n} \Rightarrow \mu$.

Proof. Let $F_{n}$ be the cumulative distribution function of $\mu_{n}$. By the Bolzano-Weierstrass theorem, for every rational $q$ there is a subsequence $F_{k}$ such that $\lim _{k} F_{k}(q)$ exists. Since the rationals are countable, Cantor's diagonal method ensures that there is a subsequence $F_{k}$ such that for every rational $q$,

$$
\lim _{k \rightarrow \infty} F_{k}(q):=G(q)
$$

exists. Let $F$ be the right-continuous extension of $G$, that is, for every $x \in \mathbb{R}$ set

$$
F(x)=\inf _{q>x} G(q) .
$$

Claim: $F$ is the cumulative distribution function of a Borel probability measure on $\mathbb{R}$.

Proof of Claim: By construction $F$ is right-continuous and non-decreasing, and clearly $0 \leq F(x) \leq F(1)$ for every $x \in \mathbb{R}$. Hence, it suffices to show that $\lim _{x \rightarrow-\infty} F(x)=0$ and $\lim _{x \rightarrow \infty} F(x)=1$ (see section 1.8).

This is where the assumption of tightness comes in. Fix $\varepsilon>0$; then by tightness there exists a bounded interval $[-A, A]$ such that every $\mu_{n}$ assigns mass at least $1-\varepsilon$ to $[-A, A]$. But this implies that $F_{n}(A) \geq 1-\varepsilon$ and $F_{n}(-A-1)<\varepsilon$; consequently,

$$
G(q) \geq 1-\varepsilon \text { for all } q>A \quad \text { and } G(q) \leq \varepsilon \text { for all } q<-A-1
$$

It now follows that $F(A) \geq 1-\varepsilon$ and $F(-A-1) \leq \varepsilon$. Since $\varepsilon>0$ is arbitrary, it follows that $\lim _{x \rightarrow-\infty} F(x)=0$ and $\lim _{x \rightarrow \infty} F(x)=1$.

### 6.3 The Normal Distribution

Definition 6.17. The normal (or Gaussian) distribution with mean $\mu \in \mathbb{R}$ and variance $\sigma^{2}>0$ is the Borel probability measure on $\mathbb{R}$ with density

$$
\begin{equation*}
\varphi_{\mu, \sigma^{2}}(x)=\frac{1}{\sqrt{2 \pi} \sigma} \exp \left\{-(x-\mu)^{2} / 2 \sigma^{2}\right\} \tag{6.9}
\end{equation*}
$$

The Gaussian distribution with mean 0 and variance 1 is called the standard normal distribution. It is an easy calculus exercise to check that (i) if $Z$ has the standard normal distribution then $a Z+b$ has the normal distribution with mean $b$ and variance $a^{2}$, and (ii) that if $X$ has the normal distribution with mean $b$ and variance $a^{2}$ then $(X-b) / a$ has the standard normal distribution.

The Two-Dimensional Standard Normal Distribution. How does one check that equation (6.12) actually defines a probability density, i.e., that $\varphi_{\mu, \sigma^{2}}$ integrates to 1 ? It is enough to do this for $\mu=0$ and $\sigma^{2}=1$, because then the general case will follow by a
linear substitution in the integral. By Fubini and a change to polar coordinates,

$$
\begin{align*}
\left(\int_{-\infty}^{\infty} e^{-x^{2} / 2} d x\right)^{2} & =\left(\int_{-\infty}^{\infty} e^{-x^{2} / 2} d x\right)\left(\int_{-\infty}^{\infty} e^{-y^{2} / 2} d y\right)  \tag{6.10}\\
& =\iint_{\mathbb{R}^{2}} e^{-x^{2} / 2} e^{-y^{2} / 2} d x d y \\
& =\iint_{\mathbb{R}^{2}} e^{-r^{2} / 2} r d r d \theta \\
& =(2 \pi) \int_{0}^{\infty} r e^{-r^{2} / 2} d r \\
& =(2 \pi) \int_{0}^{\infty} e^{-s / 2} d s / 2 \\
& =(2 \pi)
\end{align*}
$$

This shows that $1 / \sqrt{2 \pi}$ is the right normalizing factor to make $e^{-x^{2} / 2}$ a probability density.
This derivation also reveals something else of importance about the two-dimensional standard normal distribution. By definition, the two-dimensional standard normal distribution is the distribution of the random vector $(X, Y)$ where $X$ and $Y$ are independent standard normal random variables. For any two-dimensional Borel set $B$,

$$
\begin{aligned}
P((X, Y) \in B) & =\frac{1}{2 \pi} \iint_{B} e^{-x^{2} / 2} e^{-y^{2} / 2} d x d y \\
& =\frac{1}{2 \pi} \iint_{B} e^{-r^{2} / 2} r d r
\end{aligned}
$$

For sets $B$ of the form $B=A_{1} \times A_{2}$ this follows by independence and Fubini; it then follows for all Borel sets $B$ by the usual tricks. What is interesting about this formula, though, is that the two-dimensional density $e^{-r^{2} / 2} /(2 \pi)$ is a function only of $r=\sqrt{x^{2}+y^{2}}$; thus, the two-dimensional standard normal distribution is invariant by rotations about the origin. This proves the following proposition.

Proposition 6.18. If $X$ and $Y$ are independent standard normal random variables, then for any $\theta \in \mathbb{R}$ so are

$$
\begin{aligned}
& X^{\prime}=X \cos \theta-Y \sin \theta \quad \text { and } \\
& Y^{\prime}=X \sin \theta+Y \cos \theta
\end{aligned}
$$

Corollary 6.19. If $X$ and $Y$ are independent, normally distributed random variables then any non-trivial linear combination $a X+b Y$ is normally distributed.

Proof. It suffices to consider the case where $X, Y$ are independent standard normal random variables, and where the scalars $a, b$ satisfy $a^{2}+b^{2}=1$. (This follows from our remarks on scaling and translation rules for normals, immediately following Definition 6.22 above.) If $a^{2}+b^{2}=1$ then there is some $\theta \in \mathbb{R}$ such that $a=\cos \theta$ and $b=\sin \theta$.

Another way to state Corollary 6.19 is that the family of one-dimensional normal distributions is closed under convolution. It is this property that ultimately lies behind the central limit theorem.

There is one other interesting feature in the chain of equations (6.10) that we haven't yet commented on. This is the substitution $s=r^{2}$ in the penultimate equation. Observe that when we integrated out the variable $\theta$ in the preceding equality, we were in essence determining the probability density of the random variable $R^{2}=X^{2}+Y^{2}$ : in particular, if $X, Y$ are independent standard normals and $R^{2}=X^{2}+Y^{2}$ then

$$
P\left(R^{2} \in B\right)=\int_{B} r e^{-r^{2} / 2} d r \quad \text { for any Borel set } B
$$

and so by the substitution $s=r^{2}$,

$$
P(R \in B)=\int_{B} e^{-s / 2} d s / 2 \quad \text { for any Borel set } B .
$$

Thus, $R$ has the exponential distribution with mean 2 . Now an exponential random variable is very easy to simulate: if $U$ is uniformly distributed on $[0,1]$ then $-\log U$ is exponentially distributed with mean 1 , and so $-2 \log U$ is exponentially distributed with mean 2 . Consequently, if $U, V$ are independent random variables with the uniform distribution on $[0,1]$, then the transformation

$$
R=-2 \log U \quad \text { and } \quad \Theta=2 \pi V
$$

gives a two-dimensional standard normal $(R, \Theta)$ in polar coordinates. For this one can recover the rectangular coordinates

$$
X=R \cos \Theta, Y=R \sin \Theta
$$

This gives an efficient and easy-to-code ${ }^{5}$ way to simulate standard normals using the output of a standard random number generator. Random number generators provide streams of (pseudo-random) uniforms; one can use these in pairs to produce pairs of standard normals at the computational cost of computing one log, one sin, and one cos. By comparison, the computational cost of obtaining a standard normal by the quantile transform is enormous, because computing the inverse $\Phi^{-1}$ of the standard normal distribution function is prohibitively expensive.

## The n-Dimensional Normal Distribution

The standard $n$-dimensional normal distribution is defined to be the joint distribution of $\mathbf{X}=\left(X_{1}, X_{2}, \ldots, X_{n}\right)$ where the entries $X_{i}$ are independent one-dimensional

[^1]standard normal random variables. By the same argument as in the two-dimensional case, one sees that the distribution of $\mathbf{X}$ has a density: in particular, for any $B \in \mathscr{B}_{\mathbb{R}^{n}}$,
\[

$$
\begin{equation*}
P(\mathbf{X} \in B)=\frac{1}{(2 \pi)^{n / 2}} \iint \cdots \int_{B} e^{-r^{2} / 2} d x_{1} d x_{2} \cdots d x_{n} . \tag{6.11}
\end{equation*}
$$

\]

As in the two-dimensional case, the density depends only on $R$, and hence it is invariant under rotations.

Exercise 6.20. Use the rotational invariance of the $n$-dimensional normal distribution to prove that if $X_{1}, X_{2}, \ldots, X_{n}$ are independent standard normals then the sample mean and the sample variance are independent.

The rotational symmetry of the $n$-dimensional normal distribution can be used to show that there is an invariant probability measure on the rotation group $O_{n}$, called the Haar measure. Let $\mathbf{X}_{1}, \mathbf{X}_{2}, \ldots, \mathbf{X}_{n}$ be independent random vectors, each with the standard $n$-dimensional normal distribution. Construct a random $n \times n$ matrix $M$ by using the vectors $\mathbf{X}_{1}, \mathbf{X}_{2}, \ldots, \mathbf{X}_{n}$ as its columns. Then construct a random orthogonal matrix $U$ by applying the Gram-Schmidt orthogonalization algorithm to the columns of $M$.

Exercise 6.21. Show that if $\mathbf{X}_{1}, \mathbf{X}_{2}, \ldots, \mathbf{X}_{n}$ are independent $n$-dimensional standard normal random vectors then with probability one the linear subspace of $\mathbb{R}^{n}$ spanned by these vectors is $\mathbb{R}^{n}$. Therefore, the Gram-Schmidt algorithm will produce an orthogonal matrix $U$.

Fact: The random matrix $U$ has a distribution that is invariant by rotations, that is, for any non-random $n \times n$ orthogonal matrix $A$ the distribution of $A U$ is the same as that of $U$.

Proof. Because the $n$-dimensional standard normal distribution is invariant by rotations, the random matrix $A M$ has the same distribution as does $M$. But the GramSchmidt algorithm is also equivariant under rotation: for any linearly independent vectors $\mathbf{x}_{1}, \mathbf{x}_{2}, \ldots, \mathbf{x}_{n}$ the orthonormal basis produced by Gram-Schmidt when applied to $A \mathbf{x}_{1}, A \mathbf{x}_{2}, \ldots, A \mathbf{x}_{n}$ is the rotation by $A$ of the orthonormal basis produced by GramSchmidt when applied to $\mathbf{x}_{1}, \mathbf{x}_{2}, \ldots, \mathbf{x}_{n}$.

### 6.4 The Central Limit Theorem

Theorem 6.22. (Central Limit Theorem) Let $\xi_{1}, \xi_{2}, \ldots$ be independent, identically distributed random variables with mean zero and variance 1 . Then

$$
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_{i} \Longrightarrow Z
$$

where $Z$ is a standard normal random variable.

Proof. It suffices to prove that for any $C^{\infty}$ functions $f: \mathbb{R} \rightarrow \mathbb{R}$ with compact support,

$$
\begin{equation*}
\lim _{n \rightarrow \infty} E f\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_{i}\right)=E f(Z) \tag{6.12}
\end{equation*}
$$

The proof, due to Lindeberg, depends on the fact that the family of normal densities is closed under convolutions, in particular, if $X$ and $Y$ are independent Gaussian random variables then $X+Y$ is also Gaussian. Consequently, if $\zeta_{1}, \zeta_{2}, \ldots, \zeta_{n}$ are independent standard normal random variables then

$$
Z \stackrel{\mathscr{Q}}{=} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \zeta_{i} .
$$

Without loss of generality we may assume that the underlying probability space is large enough to support not only the random variables $\xi_{i}$ but also an independent sequence of i.i.d. standard Gaussian random variables $\zeta_{i}$. The objective is to show that as $n \rightarrow \infty$,

$$
\begin{equation*}
E f\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_{i}\right)-E f\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \zeta_{i}\right) \longrightarrow 0 \tag{6.13}
\end{equation*}
$$

For notational ease set

$$
\begin{aligned}
\xi_{i}^{\prime} & =\xi_{i} / \sqrt{n} \text { and } \\
\zeta_{i}^{\prime} & =\zeta_{i} / \sqrt{n} ;
\end{aligned}
$$

then relation (6.14) can be re-stated as

$$
E f\left(\sum_{i=1}^{n} \xi_{i}^{\prime}\right)-E f\left(\sum_{i=1}^{n} \zeta_{i}^{\prime}\right) \longrightarrow 0
$$

Lindeberg's strategy for proving (6.14) is to replace the summands $\xi_{i}^{\prime}$ in the first expectation by the corresponding Gaussian summands $\zeta_{i}^{\prime}$, one by one, and to bound at each step the change in the expectation resulting from the replacement of $\xi_{i}^{\prime}$ by $\zeta_{i}^{\prime}$ :

$$
\begin{equation*}
\left|E f\left(\sum_{i=1}^{n} \xi_{i}^{\prime}\right)-E f\left(\sum_{i=1}^{n} \zeta_{i}^{\prime}\right)\right| \leq \sum_{k=1}^{n}\left|E f\left(\sum_{i=1}^{k} \xi_{i}^{\prime}+\sum_{i=k+1}^{n} \zeta_{i}^{\prime}\right)-E f\left(\sum_{i=1}^{k-1} \xi_{i}^{\prime}+\sum_{i=k}^{n} \zeta_{i}^{\prime}\right)\right| \tag{6.14}
\end{equation*}
$$

Since the individual terms $\xi_{i}^{\prime}$ and $\zeta_{i}^{\prime}$ account for only a small fraction of the sums, the differences in the value of $f$ can be approximated by using two-term Taylor series approximations. Furthermore, since $f$ has compact support, the derivatives are uniformly continuous, and so the remainder terms can be estimated uniformly. In particular, for any $\varepsilon>0$ there exist $\delta>0$ and $C<\infty$ such that for any $x, y \in \mathbb{R}$,

$$
\begin{align*}
& \left|f(x+y)-f(x)-f^{\prime}(x) y-f^{\prime \prime}(x) y^{2} / 2\right| \leq \varepsilon y^{2} \quad \text { if }|y| \leq \delta \quad \text { and } \\
& \left|f(x+y)-f(x)-f^{\prime}(x) y-f^{\prime \prime}(x) y^{2} / 2\right| \leq C y^{2} \quad \text { otherwise. } \tag{6.15}
\end{align*}
$$

Consequently, for each $k$,

$$
\begin{align*}
& E f\left(\sum_{i=1}^{k} \xi_{i}^{\prime}+\sum_{i=k+1}^{n} \zeta_{i}^{\prime}\right)-E f\left(\sum_{i=1}^{k-1} \xi_{i}^{\prime}+\sum_{i=k}^{n} \zeta_{i}^{\prime}\right) \\
= & E f^{\prime}\left(\sum_{i=1}^{k-1} \xi_{i}^{\prime}+\sum_{i=k+1}^{n} \zeta_{i}^{\prime}\right)\left(\xi_{k}^{\prime}-\zeta_{k}^{\prime}\right)+\frac{1}{2} E f^{\prime \prime}\left(\sum_{i=1}^{k-1} \xi_{i}^{\prime}+\sum_{i=k+1}^{n} \zeta_{i}^{\prime}\right)\left(\left(\xi_{k}^{\prime}\right)^{2}-\left(\zeta_{k}^{\prime}\right)^{2}\right)+E R_{k}(A)+E R_{k}(B) \tag{6.16}
\end{align*}
$$

where

$$
\begin{aligned}
& R_{k}(A) \leq \varepsilon\left(\xi_{k}^{\prime}\right)^{2}+C\left(\xi_{k}^{\prime}\right)^{2} \mathbf{1}\left\{\left|\xi_{k}^{\prime}\right| \geq \delta\right\} \quad \text { and } \\
& R_{k}(B) \leq \varepsilon\left(\zeta_{k}^{\prime}\right)^{2}+C\left(\zeta_{k}^{\prime}\right)^{2} \mathbf{1}\left\{\left|\zeta_{k}^{\prime}\right| \geq \delta\right\} .
\end{aligned}
$$

The crucial feature of the expansion (6.17) is the independence of the individual terms $\xi_{i}^{\prime}$ and $\zeta_{i}^{\prime}$; this guarantees that the first two expectations on the right side of (6.17) split (as products of expectations), and since $\xi_{k}^{\prime}$ and $\zeta_{k}^{\prime}$ have the same mean and variance, it follows that the first two expectations on the right side are 0 . Consequently, for each $k$,

$$
\begin{aligned}
\mid E f\left(\sum_{i=1}^{k} \xi_{i}^{\prime}+\sum_{i=k+1}^{n} \zeta_{i}^{\prime}\right) & -E f\left(\sum_{i=1}^{k-1} \xi_{i}^{\prime}+\sum_{i=k}^{n} \zeta_{i}^{\prime}\right) \mid \\
& \leq E\left|R_{k}(A)\right|+E\left|R_{k}(B)\right| \\
& \left.\left.\leq \varepsilon E\left(\xi_{k}^{\prime}\right)^{2}+\varepsilon E\left(\zeta_{k}^{\prime}\right)^{2}\right)+C E\left(\xi_{k}^{\prime}\right)^{2} \mathbf{1}\left\{\left|\xi_{k}^{\prime}\right| \geq \delta\right\}+C E\left(\zeta_{k}^{\prime}\right)^{2}\right) \mathbf{1}\left\{\left|\zeta_{k}^{\prime}\right| \geq \delta\right\} \\
& \left.\leq n^{-1} \varepsilon\left(E\left(\xi_{k}\right)^{2}+E\left(\zeta_{k}\right)^{2}\right)+n^{-1} C E\left(\xi_{k}\right)^{2} \mathbf{1}\left\{\left|\xi_{k}\right| \geq \sqrt{n} \delta\right\}+n^{-1} C E\left(\zeta_{k}\right)^{2}\right) \mathbf{1}\left\{\left|\zeta_{k}\right| \geq \sqrt{n} \delta\right\} \\
& \left.\leq 2 \varepsilon n^{-1}+n^{-1} C E\left(\xi_{k}\right)^{2} \mathbf{1}\left\{\left|\xi_{k}\right| \geq \sqrt{n} \delta\right\}+n^{-1} C E\left(\zeta_{k}\right)^{2}\right) \mathbf{1}\left\{\left|\zeta_{k}\right| \geq \sqrt{n} \delta\right\} .
\end{aligned}
$$

Substituting this bound in inequality (6.15) now yields

$$
\left.\left|E f\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_{i}\right)-E f\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \zeta_{i}\right)\right| \leq 2 \varepsilon+C E\left(\xi_{1}\right)^{2} \mathbf{1}\left\{\left|\xi_{1}\right| \geq \sqrt{n} \delta\right\}+C E\left(\zeta_{1}\right)^{2}\right) \mathbf{1}\left\{\left|\zeta_{1}\right| \geq \sqrt{n} \delta\right\} .
$$

Since $E \xi_{1}^{2}=1<\infty$ and $E \zeta_{1}^{2}=1<\infty$, the dominated convergence theorem implies that the last two expectations converge to zero as $n \rightarrow \infty$, and so

$$
\limsup _{n \rightarrow \infty}\left|E f\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_{i}\right)-E f\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \zeta_{i}\right)\right| \leq 2 \varepsilon .
$$

Finally, since $\varepsilon>0$ is arbitrary, the convergence (6.14) must hold.


[^0]:    ${ }^{4}$ The terms vague convergence, weak convergence, and convergence in distribution all mean the same thing. Functional analysts use the term weak-* convergence.

[^1]:    ${ }^{5}$ There are even more efficient methods for producing pseudo-random normals, but they are considerably more complicated and require quite a bit more coding.

