A Formally Verified Proof of the Central Limit Theorem

Luke Serafin

December 27, 2015

Abstract

This thesis describes the results of a collaborative effort to formalize the proof of the central limit theorem of probability theory. That project was carried out in the Isabelle proof assistant, and builds upon and extends the libraries for mathematical analysis, in particular measure-theoretic probability theory. The formalization introduces the notion of weak convergence (or convergence in distribution) required to state the central limit theorem, and uses characteristic functions (Fourier transforms) to demonstrate convergence to the standard normal distribution under the hypotheses of the central limit theorem. Supporting such reasoning motivated significant changes to the measure-theoretic integration libraries of Isabelle.

Contents

1 Introduction **3**

2 Probabilistic Prelimilaries **6**
2.1 Measure Spaces 6
2.2 Independent Events 9
2.3 Random Variables 10
2.4 Integration 11
2.5 Normal Distributions 13
2.6 Convergence of Random Variables 14
2.7 Convolutions and Characteristic Functions 14

3 Summary of the Proof **15**

4 The Isabelle Interactive Proof Assistant **17**
4.1 Overview of Isabelle 18
4.2 Types and Locales 23
4.3 Limits and Filters 24

5 The Formalized Proof **25**
5.1 Distribution Functions 26
5.2 Weak Convergence 29
5.3 Helly Selection Theorem and Tightness 39
5.3.1 Helly Selection Theorem 39
5.3.2 Tightness of Sequences of Measures 41
5.4 Integration 44
5.4.1 General Remarks on Formalizing Integration 44
5.4.2 The Sine Integral Function 47
5.5 Characteristic Functions 51
5.6 The Lévy Theorems 54
5.6.1 The Lévy Inversion Theorem 55
5.6.2 The Lévy Continuity Theorem 57
5.7 The Central Limit Theorem 58

6 Conclusion: Opportunities for Improvement and Extension **62**

1 Introduction

Consider a toss of a fair coin. If we treat a result of tails as having value zero and a result of heads as having value one, we may treat the coin toss as a random variable, say X.[1] Thus X is supported on $\{0, 1\}$, and

$$\mathbb{P}(X = 0) = \mathbb{P}(X = 1) = \frac{1}{2}.$$

Hence the expected value of X is

$$\mathbb{E}(X) = 0 \cdot \mathbb{P}(X = 0) + 1 \cdot \mathbb{P}(X = 1) = \frac{1}{2}.$$

Now suppose we toss the coin repeatedly, thus generating an infinite sequence $\langle X_n \mid n \in \mathbb{N} \rangle$ of random variables which are pairwise independent and have the same distribution as X. By the strong law of large numbers, the mean $\overline{X}_n = \frac{1}{n} \sum_{i \leq n} X_i$ converges almost surely to $\mathbb{E}(X) = \frac{1}{2}$. But clearly after a finite number of trials there is a nonzero probability that the value of $\overline{X}_n$ will differ from $\mathbb{E}(X)$. In fact, for n odd the probability of deviation is 1, because in that case it is impossible for $\frac{1}{n} \sum_{i \leq n} X_i$ to have the value $\frac{1}{2}$ at any element of the sample space. Nevertheless $|\overline{X}_n - \mathbb{E}(X)|$ must converge to zero, and so the probability of large deviations of the mean $\overline{X}_n$ from the expected value $\mathbb{E}(X)$ is small. Exactly how small is made precise by De Moivre's central limit theorem.

In 1733 De Moivre privately circulated a proof[2] which, in modern terminology, shows that $n^{-1/2}\overline{X}_n$ converges to a normal distribution. This material was later published in the 1738 second edition of his book *The Doctrine of Chances,* the first edition of which was first published in 1712 and is widely regarded as the first textbook on probability theory. De Moivre also considered the case of what we might call a biased coin (an event which has value one with probability p and zero with probability $1 - p$, for some $p \in (0, 1)$), and realized that his convergence theorem continues to hold in that case.

De Moivre's result was generalized by Laplace in the period between about 1776 and 1812 to sums of random variables with various other distributions. For example, in 1776 Laplace proved that $n^{-1/2}\overline{X}_n$ converges to a normal distribution in the case where the X_n's are uniformly distributed. The particular problem Laplace considered in that paper was finding the distribution of the average inclination of a random sample of comets, the distribution for a single comet being assumed uniform between 0° and 90°. Over the next three decades Laplace developed the conceptual and analytical tools to extend this convergence theorem to sums of independent identically distributed random variables with ever more general distributions, and this work culminated in his treatise *Théorie analytique des probabilités.* This included the development of the

[1] An intuitive understanding of probabilistic language suffices for the present section; more precise definitions for many concepts will be discussed in section 2.

[2] This historical information is drawn from [9], and references to original works may be found there.

method of characteristic functions to study the convergence of sums of random variables, a move which firmly established the usefulness of analytic methods in probability theory (in particular Fourier analysis, the characteristic function of a random variable being exactly the Fourier transform of that variable).

Laplace's theorem, which later became known as the central limit theorem (a designation due to Pólya and stemming from its importance both in the theory and applications of probability), states in modern terms that the normalized sum of a sequence of independent and identically distributed random variables with finite, nonzero variance converges to a normal distribution. All of this imprecise language will be made precise later on. In the work of Laplace all the main ingredients of the proof of the central limit theorem are present, though of course the theorem was refined and extended as probability underwent the radical changes necessitated by its move to measure-theoretic foundations in the first half of the twentieth century.

Gauss was one of the first to recognize the importance of the normal distribution to the estimation of measurement errors, and it is notable that the usefulness of the normal distribution in this context is largely a consequence of the central limit theorem, for errors occurring in practice are frequently the result of many independent factors which sum to an overall error in a way which can be regarded as approximated by a sum of independent and identically distributed random variables. The normal distribution also arose with surprising frequency in a wide variety of empirical contexts: from the heights of men and women to the velocities of molecules in a gas. This gave the central limit theorem the character of a natural law, as seen in the following poetic quote from Sir Francis Galton in 1889 [10]:

> I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the "Law of Frequency of Error." The law would have been personified by the Greeks and deified, if they had known of it. It reigns with serenity and in complete self-effacement, amidst the wildest confusion. The huger the mob, and the greater the apparent anarchy, the more perfect is its sway. It is the supreme law of Unreason. Whenever a large sample of chaotic elements are taken in hand and marshaled in the order of their magnitude, an unsuspected and most beautiful form of regularity proves to have been latent all along.

Many more details on the history of the central limit theorem and its proof can be found in [9].

Standards of rigour have evolved a great deal over the course of the history of the central limit theorem, and around the turn of the twentieth century a completely precise notion of proof, developed by Frege, Russell, and many others, finally became available to mathematicians. Actually writing proofs which conform to the precise requirements of this notion did not become the new norm of mathematical practice, however, largely because it is impractical for human mathematicians to work at that level of formal detail. The burden of writing an entirely precise proof in first-order logic (say) simply does not offer

sufficient gain for a human mathematician to undertake it. However, advances in automated computing technology around the middle of the twentieth century quickly progressed to the point where a computer could be programmed to take on the cumbersome burden of verifying all the details of a proof which a human outlined at a high level. This is the domain of interactive theorem proving.

One significant mathematical result to be verified using an interactive proof assistant was the prime number theorem, formalized between 2003 and 2004 at Carnegie Mellon University by Jeremy Avigad, Kevin Donnelly, David Gray, and Paul Raff. Thoughts on that formalization, which was carried out with the Isabelle proof assistant, are recorded in [2]. Though the prime number theorem is traditionally considered a landmark result of analytic number theory, it should be noted that the proof formalized by Avigad and collaborators did not employ complex analysis, but was rather the elementary proof of Selberg [24], using results and methods due to Erdős. Thus the proof of the prime number theorem demonstrated that significant mathematical results could be formalized quite effectively in Isabelle, but did not provide a test of the usefullness of Isabelle for formalizing results based on deep theory. In 2009 John Harrison published a formalization of an analytic proof of the prime number theorem [16].

When the author approached Avigad seeking a research project, Avigad saw an opportunity to carry out in Isabelle a formalization relying on deep analytical theory, and suggested that the author help develop Isabelle's integration libraries by choosing an interesting result to formalize. The author's choice was the central limit theorem, often abbreviated CLT.

A theorem which both played a fundamental role in the development of modern probability theory and has far-reaching applications seemed to us a perfect candidate for formalization, especially because the measure-theoretic libraries of Isabelle are still under active development and we saw an opportunity to contribute to them by formalizing the characteristic function arguments used to prove the CLT. The formalization was completed between 2011 and 2013, and improvements to the proof scripts are ongoing. The formalization was a joint effort between Jeremy Avigad, Johannes Hölzl (Technische Universität München), and the author. A preliminary report [3] was written by all three collaborators and presented at the Vienna Summer of Logic by Hölzl. All the proof scripts from our formalization which have not yet been moved into Isabelle's libraries are found at `https://github.com/avigad/isabelle`.

The fact that our effort to formalize the central limit theorem succeeded in a few months of dedicated formalization effort (interspersed among longer stretches of slower progress) testifies to the maturity of the analysis and measure theory libraries in Isabelle, though of course much remains to be added and many improvements are possible.

Here is the result we verified in Isabelle:

```
theorem (in prob_space) central_limit_theorem:
  fixes
    X :: "nat ⇒ 'a ⇒ real" and
    μ :: "real measure" and
    σ :: real and
```

```
  S :: "nat ⇒ 'a ⇒ real"
assumes
  X_indep: "indep_vars (λi. borel) X UNIV" and
  X_integrable: "⋀n. integrable M (X n)" and
  X_mean_0: "⋀n. expectation (X n) = 0" and
  σ_pos: "σ > 0" and
  X_square_integrable: "⋀n. integrable M (λx. (X n x)^2)" and
  X_variance: "⋀n. variance (X n) = σ^2" and
  X_distrib: "⋀n. distr M borel (X n) = μ"
defines
  "S n ≡ λx. ∑i<n. X i x"
shows
  "weak_conv_m (λn. distr M borel (λx. S n x / sqrt (n * σ^2)))
      (density lborel std_normal_density)"
```

At the time of writing the full proof document is 4499 lines, which is 98 pages after processing with Isabelle's LaTeX facilities. This excludes a large amount of library development which has already been moved to HOL-Probability, but is still dramatically shorter than the proof of the prime number theorem, which excluding previously developed library material and a proof of the law of quadratic reciprocity was still in the neighborhood of 500 pages, or 22,300 lines ([2], p. 8). The fact that the relatively deep measure-theoretic proof of the central limit theorem can be successfully formalized with such comparatively little library augmentation testifies to the maturity of the analysis libraries in Isabelle.

2 Probabilistic Prelimilaries

Readers familiar with the basics of measure-theoretic probability theory may wish to skip to the next section, though this section still serves to establish notation. Readers lacking this background will find a brief introduction here, just sufficient to give an idea of what concepts are behind the proof of the central limit theorem in the next section. Those who wish to learn the measure-theoretic foundations of probability theory should consult a standard work on the subject, such as [6].

2.1 Measure Spaces

We begin with an explication of the idea of a measure. It is intuitively obvious that some sets of points have a definite "size:" A line segment has a length, a circle has an area, a cone has a volume, etc. The notion of a measure is intended to make precise this intuitive notion of the "size" of a set. We shall see later how probabilities are interpreted in terms of measures.

Definition 2.1. Let X be a set, and $\Sigma \subseteq X$ be a collection of subsets of X which contains $\emptyset$ and is closed under complements and countable unions (from which it immediately follows that $X \in \Sigma$ and that Σ is closed under countable

intersections). A *measure* on Σ (often simply called a measure on X when the intended collection Σ is clear from the context) is a function $\mu\colon \Sigma \to [0,\infty]$ with the property that $\mu(\emptyset) = 0$ and which is countably additive, which means that for every pairwise disjoint collection $\{A_n \mid n \in \mathbb{N}\}$ of elements of Σ,

$$\mu\left(\bigcup_{n=0}^{\infty} A_n\right) = \sum_{n=0}^{\infty} \mu(A_n).$$

For $A \in \Sigma$ as in the definition, the value $\mu(A)$ is called the *measure* of A and corresponds to the intuitive notions of size, length, area, volume, etc. Note that the measure of a set may be infinite; indeed, if μ is a measure on $\mathbb{R}$ such that $\mu(n, n+1] = 1$ for $n \in \mathbb{N}$ (as should be the case if μ measures the length of intervals), it is immediate from the definition of a measure that $\mu(\mathbb{R}) = \infty$.

A collection Σ of subsets of a set X satisfying the hypothesis in the above definition (i.e. containing $\emptyset$ and closed under complements and countable unions) is called a σ-algebra. The elements of Σ are called the *measurable sets,* and a set with an associated σ-algebra but no associated measure is called a *measurable space.*

One might ask why a measure should not determine a size for every subset of X; one important reason is that there is no translation-invariant measure μ on $\mathbb{R}$ which assigns intervals the expected length (i.e. $\mu[a,b] = b-a$ for $a < b$) and is defined for all subsets of $\mathbb{R}$. A measure μ on $\mathbb{R}$ is called *translation-invariant* iff for every measurable $A \subseteq \mathbb{R}$ and every $x \in \mathbb{R}$, $\mu(A+x) = \mu(A)$, where $A+x$, the translate of A by x, is $\{a + x \mid a \in A\}$.

Suppose now for contradiction that μ is a translation-invariant measure on $\mathbb{R}$ which measures all subsets of $\mathbb{R}$ (so the measurable space on which μ is defined is $(\mathbb{R}, \mathcal{P}(\mathbb{R}))$, where $\mathcal{P}(\mathbb{R})$ is the powerset of $\mathbb{R}$) and satisfies $\mu[a,b] = b-a$ for $a < b$. Consider the equivalence relation $\sim$ on $\mathbb{R}$ defined by $x \sim y$ iff $|x-y| \in \mathbb{Q}$. Because the rationals are dense in $\mathbb{R}$ it is clear that each equivalence class contains an element of the closed unit interval. Using the axiom of choice, select one element of the intersection of each equivalence class with $[0,1]$ (this use of the axiom of choice is essential; there are models of set theory in which there exists a translation-invariant measure λ defined for all subsets of $\mathbb{R}$ and assigning to intervals the expected length). Denote this collection by $V = \{r_\alpha \mid \alpha \in I\}$, where I is some index set. Enumerate the rationals in $[-1,1]$ as $\{q_n \mid n \in \mathbb{N}\}$, and for each n define $V_n = V + q_n$. Since μ is translation-invariant, all elements of the collection $\{V_n \mid n \in \mathbb{N}\}$ receive the same measure. Let $E = \bigcup_{n\in\mathbb{N}} V_n$; it is clear from the definition of the V_n's that $[0,1] \subseteq E \subseteq [-1,2]$. Furthermore, it is an immediate consequence of countable (in this case finite) additivity that for any measure ν on any space, if A and B are measurable and $A \subseteq B$, then $\nu(A) \leq \nu(B)$ (note $B = A \cup (B \setminus A)$ and this union is disjoint). Hence because μ assigns intervals their expected length, we have $1 \leq \mu(E) \leq 3$. However, by countable additivity

$$\mu(E) = \mu\left(\bigcup_{n=0}^{\infty} V_n\right) = \sum_{n=0}^{\infty} \mu(V_n).$$

Since $\mu(V_n)$ is the same for all n, the sum on the right is infinite if it is not zero, and we have obtained a contradiction. This counterexample is due to Vitali [26].

The standard solution to the nonexistence problem noted in the preceding paragraph is to restrict which sets are assigned measures—hence the σ-algebra Σ in the definition of a measure. This allows "bad" sets such as V from the counterexample to be excluded from receiving a measure, and is essential to a useful theory of measure.

Since measures extend the notion of length of intervals, it is natural to suppose that all intervals should be measurable (let us say open intervals, for definiteness). If all open intervals are measurable, then all sets in the σ-algebra generated by the open intervals—the intersection of all σ-algebrae containing all the open intervals, an object which is easily verified to be a σ-algebra—must also be meaurable. The σ-algebra generated by all open intervals in $\mathbb{R}$ is called the *Borel* σ-algebra on $\mathbb{R}$, and more generally for any topological space the associated Borel σ-algebra is the σ-algebra generated by the open sets. Most measures encountered in our formalization are Borel measures, that is, measures defined on the Borel σ-algebra.

A note regarding limits: For a sequence $\langle a_n \mid n \in \mathbb{N}\rangle$ of real numbers or real-valued functions, the notation $a_n \to a$ is used to indicate the sequence converges (in a sense which should be clear from the context) to the limit a. $a_n \uparrow a$ and $a_n \downarrow a$ indicate the convergence is monotone in the obvious direction. For $\langle A_n \mid n \in \mathbb{N}\rangle$, $A_n \uparrow A$ indicates $\bigcup_{n\in\mathbb{N}} A_n = A$ and $A_n \subseteq A_{n+1}$ for each n, while $A_n \downarrow A$ indicates $\bigcap_{n\in\mathbb{N}} A_n = A$ and $A_{n+1} \subseteq A_n$ for each n. It is an easy consequence of countable additivity that if each A_n and A are measurable subsets of some measure space (X, Σ, μ) and $A_n \uparrow A$, then $\mu(A_n) \uparrow \mu(A)$, and similarly if $A_n \downarrow A$ then $\mu(A_n) \downarrow A$. These are called the upward and downward continuity of the measure μ, respectively.

There is a unique Borel measure λ on $\mathbb{R}$, called Lebesgue measure, which assigns to each interval the expected length: $\lambda[a, b] = b - a$ for $a < b$. As expected, this measure has the property that $\lambda\{x\} = 0$ for any single real x: Note that $[x - 1/n, x + 1/n] \downarrow \{x\}$, and so $\lambda\{x\} = \lim_{n\to\infty} 2/n = 0$. A set with measure zero is called *null*. In general, if μ is a measure on a measurable space (X, Σ), $x \in X$, and $\mu\{x\} = 0$, x is called a continuity point of μ. It is possible in general for there to be singletons of positive measure; if $\mu\{x\} > 0$, x is called an atom of μ. A measure is called continuous just in case it has no atoms.

How does all this talk of measures relate to probability? Well, in probability theory, one wishes to assign probabilities to events. The probability that a fair coin turns up heads should be $\frac{1}{2}$, the probability that a randomly selected element of the unit interval $[0, 1]$ is between $\frac{1}{3}$ and $\frac{2}{3}$ should be $\frac{1}{3}$, etc. How can these events be modelled formally? Let Ω be the set of all possible states of the world (or simply the set of all possible worlds, if one is of a philosophical bent); an event is simply a collection of such states (namely the collection where the event occurs). Thus for the toss of a fair coin we may take $\Omega = \{H, T\}$, where H is a world where the coin turns up heads, and T a world where it turns up tails.

The event that the coin turns up heads is simply $\{H\}$. Similarly, for randomly selecting an element of the unit interval we may take $\Omega = [0, 1]$, $\omega \in \Omega$ being a world where the selected element is ω. The event that the selected number is between $\frac{1}{3}$ and $\frac{2}{3}$ is then simply $(\frac{1}{3}, \frac{2}{3})$. The space Ω is called the *sample space* in probability theory.

It is intuitively clear that the probability of the impossible event $\emptyset$ is zero, and that the probability of a union of disjoint events should be the sum of their probabilities. It is therefore reasonable to suppose probability determines a measure, $\mathbb{P}$, on some collection of events (which we might consider "observable"). Vacuously $\emptyset$ is observable, and it is clear that if an event E is observable then so should be its complement, and that if events $\{E_n \mid n \in \mathbb{N}\}$ are observable then so should be their union. Thus the collection of observable events should be a σ-algebra. It should be noted that in probability theory, the term "event" is generally reserved for what we have termed "observable event," and we shall follow this convention in the sequel.

Measures determined by probabilities of events are naturally called *probability measures.* An important feature of such measures, and indeed the sole criterion in the mathematical definition of a probability measure, is that $\mathbb{P}(X) = 1$. Probabilities cannot be arbitrarily large, and it is assumed to be certain that something in the sample space is the true state of the world (so the sample space is exhaustive). A measure μ is called *finite* iff $\mu(X) < \infty$, and modulo the zero measure the theory of finite measures is the same as the theory of probability measures (any nonzero finite measure μ can be scaled by $\frac{1}{\mu(X)}$ to obtain a probability measure).

2.2 Independent Events

Consider now tossing two coins successively. It is intuitively clear that the outcome of the first toss does not influence in any way the outcome of the second. Taking the sample space as $\Omega = \{HH, HT, TH, TT\}$ and assigning each singleton event (e.g. $\{HH\}$, the event that both tosses come up heads) equal probability (so 1/4), the event that the first coin comes up heads is $E_1 = \{HH, HT\}$, while the event that the second comes up heads is $E_2 = \{HH, TH\}$. Note that $\mathbb{P}(E_1) = \mathbb{P}(E_2) = 1/2$, and $\mathbb{P}(E_1 \cap E_2) = \mathbb{P}(\{HH\}) = 1/4 = \mathbb{P}(E_1)\mathbb{P}(E_2)$. It turns out that this multiplicative rule for computing probabilities of intersections is a useful formal explication of the intuitive notion of independence (this can be explained in terms of conditional probabilities; we refer the reader to any elementary account of probability).

Definition 2.2. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and $E_1, E_2 \in \mathcal{F}$. E_1 and E_2 are *independent* (written $E_1 \perp\!\!\!\perp E_2$) iff $\mathbb{P}(E_1 \cap E_2) = \mathbb{P}(E_1)\mathbb{P}(E_2)$.

For a larger collection $E_1, \ldots, E_n$ of events, we want independence to entail

$$\mathbb{P}\left(\bigcap_{i=1}^{n} E_i\right) = \prod_{i=1}^{n} \mathbb{P}(E_i).$$

Pairwise independence is too weak for this—in tossing a coin twice, the events $\{HH, HT\}$, $\{HH, TH\}$, and $\{HT, TH\}$ are pairwise independent each with probability $1/2$, but their intersection is impossible (empty), and does not have probability $1/8$. We instead use

Definition 2.3. Events $E_1, \ldots, E_n$ are *independent* iff

$$\mathbb{P}\left(\bigcap_{i=1}^{n} E_i\right) = \prod_{i=1}^{n} \mathbb{P}(E_i).$$

This does not generalize directly to an infinite collection of events; the definition employed in that case is

Definition 2.4. A collection $\mathcal{E} \subseteq \mathcal{F}$ of events is *independent* iff every finite subcollection $\{E_1, \ldots, E_n\} \subseteq \mathcal{E}$ is independent.

We spoke earlier of the σ-algebra of events being the collection of "observable" events, and indeed σ-algebrae are useful for keeping track of information acquired by observation. Roughly, obtaining information about the state of the world can make more events observable. For example, if a coin is flipped twice but the result of the second flip is hidden, the states HH, HT and the states TH, TT are indistinguishable, and a reasonable σ-algebra of observable events is $\mathcal{F} = \{\emptyset, \{HH, HT\}, \{TH, TT\}, \{HH, HT, TH, TT\}\}$. If, however, the result of the second flip is revealed, singleton events such as $\{HH\}$ become observable, and the σ-algebra of observable events should accordingly be expanded to the full powerset of the sample space $\{HH, HT, TH, TT\}$.

If observations are independent, the refined σ-algebrae they give rise to should also be independent; this is made precise by extending the notion of independence to σ-algebrae:

Definition 2.5. A collection $\{\mathcal{F}_\alpha \mid \alpha \in I\}$ of σ-algebrae (I an index set of arbitrary cardinality) is *independent* iff for each choice of precisely one set E_α from each σ-algebra $\mathcal{F}_\alpha$, the collection $\{E_\alpha \mid \alpha \in I\}$ of events is independent.

For $\mathcal{F}$, $\mathcal{G}$ σ-algebrae, the notation $\mathcal{F} \perp\!\!\!\perp \mathcal{G}$ means of course that $\mathcal{F}$ and $\mathcal{G}$ are independent.

2.3 Random Variables

Often one wishes to talk about real-valued statistics associated to sample points rather than the sample points themselves; examples include the waiting time for a train, the length of a manufactured screw, or the average height of the Danish population. These can be thought of as functions from the sample space Ω to the real line $\mathbb{R}$ (or perhaps the extended reals $[-\infty, \infty]$; the bus may *never* arrive). In order to be accessible to probability theory, the "interesting" events associated with such statistics should be measurable; in particular, for X such a statistic and $a < b$, the preimage $X^{-1}[(a, b)] = \{\omega \in \Omega \colon a < X(\omega) < b\}$ should be measurable. This is handled by the notion of a measurable function:

Definition 2.6. Let (X_1, Σ_1), (X_2, Σ_2) be measurable spaces. A function $f\colon X_1 \to X_2$ is *measurable* (with respect to (Σ_1, Σ_2)) iff for every $E \in \Sigma_2$, $f^{-1}[E] \in \Sigma_1$.

Note the similarity to the definition of continuity from topology. For $(\Omega, \mathcal{F}, \mathbb{P})$ a probability space, a *random variable* is simply a measurable function $X\colon \Omega \to \mathbb{R}$, where $\mathbb{R}$ is assumed to be equipped with the σ-algebra of Borel sets (sometimes the codomain is taken instead as $[-\infty, \infty]$). This makes precise the notion of a probabilistically accessible real-valued statistic on a sample space.

It should be noted that when talking about probabilities and events in the context of random variables, evaluation is often left implicit. For example $\mathbb{P}(X > 0)$ is an abusive notation for $\mathbb{P}\{\omega \in \Omega \mid X(\omega) > 0\}$.

A random variable X induces a probability measure μ_X on the real line via $\mu_X(A) = \mathbb{P}(X \in A)$; in this case we say X is distributed as μ_X and write $X \sim \mu_X$.

Events can be viewed as random variables via their indicator functions (called characteristic functions outside probability, the term characteristic function in probability denoting a Fourier transform). In general, for $A \subseteq X$, the indicator function $\mathbb{1}_A$ is the function which has value one at elements of A and zero elsewhere. In case A is a measurable set in a probability space, the indicator function of A is a random variable.

Random variables make precise the notion of an "observation" alluded to while discussing independence of σ-algebrae in the preceding section: one can observe the clock while waiting for the train to arrive, or measure the average height of a random sample drawn from the Danish population, etc. The amount of information that knowing the value of a random variable makes available is determined by the σ-algebra generated by the random variable, which we now define.

Definition 2.7. Let X be a random variable defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The *$\sigma$-algebra generated by X*, $\sigma(X)$, is the σ-algebra generated by the collection of preimages of Borel sets (or equivalently open intervals) under X, i.e. $\{X^{-1}[A] \mid A \in \mathcal{B}\}$, where $\mathcal{B}$ is the collection of Borel subsets of $\mathbb{R}$.

Note that $\sigma(X)$ is a σ-algebra on the real line.

Now we can define the notion of independence for random variables:

Definition 2.8. A collection $\{X_\alpha \mid \alpha \in I\}$ of random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ (I an index set of arbitrary cardinality) is *independent* iff the collection $\{\sigma(X_\alpha) \mid \alpha \in I\}$ is independent.

Thus random variables X, Y are independent (written $X \perp\!\!\!\perp Y$) iff the information obtained by learning each of their values is independent.

2.4 Integration

Frequently in applications of probability one is interested in the average value of a random variable, called its expected value. In computing this average,

sample points are weighted by their probabilities. For example, if a weighted coin turns up heads with probability 2/3 and tails with probability 1/3, and we bet a dollar that it will come up heads (and so gain one dollar if it does and lose one dollar if it doesn't), our expected gain is $(2/3)1 + (1/3)(-1) = 1/3$. For a continuous random variable X (such as the waiting time for a train or the length of a screw), this weighted sum takes the form of a weighted integral, called the integral with respect to a probability measure μ (the measure on the domain of the random variable) and denoted $\int X\, d\mu$. This integral is over the entire space; sometimes one wishes to take an integral over a subset A of the space (effectively regarding the random variable X as zero outside A); this is $\int_A X\, d\mu = \int X \mathbb{1}_A\, d\mu$.

Taking such weighted integrals is by no means limited to random variables; if (X, Σ, μ) is any measure space and $f\colon X \to \mathbb{R}$, we may consider $\int f\, d\mu$, and $\int_A f\, d\mu$ for any $A \subseteq X$. This general integral with respect to a measure is called the Lebesgue integral; it simultaneously generalizes weighted discrete sums, Riemann integrals, and much more.

Note that the integral of a function may fail to exist; for example, the integral[3] $\int_{[0,\infty)} x\, \lambda(dx)$ is infinite (which may be regarded as having value ∞ or being undefined, depending on context), while the integral $\int x\, \lambda(dx)$ does not exist. Functions f for which $\int f\, d\mu < \infty$ (which implicitly assumes this integral exists) are called *μ-integrable*, or simply integrable if the measure is clear from the context.

For a full account of Lebesgue integration see [6]. Here we shall briefly outline some properties of integration which will be used later.First, the integral is linear in full generality:

$$\int (cf + g)\, d\mu = c \int f\, d\mu + \int g\, d\mu$$

for any integrable $f, g\colon X \to \mathbb{R}$ and any $c \in \mathbb{R}$. Second, the monotone convergence theorem holds: If $\langle f_n \mid n \in \mathbb{N} \rangle$ are measurable functions, $f_n\colon X \to [0, \infty]$ and $f_n \leq f_{n+1}$ for each n, then the pointwise limit f is measurable, and furthermore

$$\int f_n\, d\mu \uparrow \int f\, d\mu.$$

Here, the value ∞ is allowed for the integrals. Finally, the dominated convergence theorem holds: If $\int g\, d\mu < \infty$, and $\langle f_n \mid n \in \mathbb{N} \rangle$ is a sequence of measurable functions $f_n\colon X \to \mathbb{R}$, and $|f_n| \leq g$ for each n, then the pointwise limit f is integrable, and furthermore

$$\int f_n\, d\mu \to \int f d\mu.$$

In the special case where g is a constant ($\int g\, d\mu$ exists in this case if its domain is a finite measure space, such as a probability space), this result is called the bounded convergence theorem.

[3] Note that for $f\colon X \to \mathbb{R}$ and (X, Σ, μ) a measure space, $\int f(x)\, \mu(dx)$ denotes the integral of f with respect to μ.

For X a random variable defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, the integral $\int X \, d\mathbb{P}$ is called the *expected value*, *expectation*, or *mean* of X and denoted $\mathbb{E}(X)$, while $\mathbb{E}((X - \mathbb{E}(X))^2)$ is called the *variance* of X. The mean is the "center of mass" of the distribution of X, while its variance indicates how "spread out" the mass of its distribution is.

2.5 Normal Distributions

As mentioned in the introduction, the familiar bell-shaped curve of the normal distribution was discovered to arise frequently in empirical investigations involving errors or deviance from the mean, and the central limit theorem largely explains this frequent appearance. To proceed further with reasoning about the central limit theorem, it will of course be necessary to define precisely what such a distribution is. For this we need the notion of the density of a probability measure.

Definition 2.9. Let $f \colon \mathbb{R} \to \mathbb{R}$ be a function. A Borel measure μ on $\mathbb{R}$ is said to have *density* f (with respect to Lebesgue measure λ) iff for every Borel set A, $\mu(A) = \int_A f \, d\lambda$.

Not all measures have densities; it is straightforward to verify that a probability measure such that $\mu\{0\} = 1$—a unit mass at zero—has no density. However, normal distributions do have densities, and are most conveniently described in terms of those densities.

Definition 2.10. Let $\mu \in \mathbb{R}$, $\sigma > 0$. The *normal distribution* with mean μ and variance σ^2, denoted $\mathcal{N}_{\mu,\sigma^2}$, is the Borel measure with density

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

The reason for labelling the variance of a normal distribution with σ^2 is that σ, the square root of the variance, is a quantity of importance in statistics, called the standard deviation. In general, the square root of the variance of a distribution is called its standard deviation, but that quantity will not be of importance to us. Variance is a more convenient quantity for our purposes, because the variance of the sum of two independent random variables is the sum of their variances.

The normal distribution $\mathcal{N}_{0,1}$ is called the *standard normal distribution*. Note that for any normal distribution $\mathcal{N}_{\mu,\sigma^2}$, $\frac{\mathcal{N}_{\mu,\sigma^2} - \mu}{\sigma^2}$ is a standard normal distribution. In general, any random variable with finite mean and finite, nonzero variance can be "normalized" to a random variable with mean zero and unit variance by subtracting the mean and dividing by the variance; this will be of importance for ensuring convergence to a standard normal distribution in the central limit theorem.

2.6 Convergence of Random Variables

Especially in the context of the central limit theorem, one is interested in the convergence of a sequence $\langle X_n \mid n \in \mathbb{N} \rangle$ of random variables to a given random variable X. Pointwise convergence is generally uninteresting because it often fails on a negligible set of sample points (negligible in the sense that it has measure zero, and so may be safely ignored for most purposes in probability theory); the notion of convergence almost surely (pointwise convergence except on a set of measure zero; called convergence almost everywhere in general probability theory) fixes this problem, but is still a very strong condition. A weaker condition is that for every $\varepsilon > 0$ $\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0$. Intuitively, this says that the sequence $\langle X_n \mid n \in \mathbb{N} \rangle$ eventually becomes very close to X outside a very small (though not necessarily measure zero) set. However, this notion of convergence, called convergence in probability (convergence in measure in general measure theory) is still too strong for the central limit theorem.

For the convergence of the central limit theorem, we want to say that the distribution of the normalized sum of independent random variables with finite, nonzero variance "resembles more and more" the standard normal distribution. This is made precise by the notion of convergence in distribution, or weak convergence (weak* convergence in the sense of functional analysis), which will be defined and discussed in the course of our treatment of the formalization of the CLT. Weak convergence of a sequence $\langle X_n \mid n \in \mathbb{N} \rangle$ of random variables to a weak limit X is denoted $X_n \Rightarrow X$, and this makes sense also for the distributions of the random variables ($\mu_n \Rightarrow \mu$, where $X_n \sim \mu_n$ and $X \sim \mu$).

2.7 Convolutions and Characteristic Functions

Definition 2.11. Let μ, ν be Borel measures on $\mathbb{R}$. The *convolution* $\mu * \nu$ is defined by

$$(\mu * \nu)(B) = \int \nu(B - x)\mu(dx),$$

where $B \subseteq \mathbb{R}$ is a Borel set and $B - x = \{y - x \mid y \in B\}$ (which is easily verified to also be a Borel set).

It turns out that the distribution of the sum of two independent random variables is the convolution of their distributions. More precisely, if X and Y are random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and $X \sim \mu$ and $Y \sim \nu$, then for any Borel set $B \subseteq \mathbb{R}$, $\mathbb{P}(X + Y \in B) = (\mu * \nu)(B)$. If μ, ν have densities f, g, respectively, then $\mu * \nu$ has density $f * g$, the convolution of the functions f and g, which is defined by

$$(f * g)(x) = \int g(x - t)f(t)\, dt.$$

For proofs of these remarks and more information about convolutions, see [6] or any standard treatment of measure theory.

The notion of a Fourier transform is central to harmonic analysis (see [20] or any standard text on the subject); the *Fourier transform* $\hat{\mu}$ of a measure μ on $\mathbb{R}$ with the Borel σ-algebra is given by

$$\hat{\mu}(x) = \int e^{itx}\, \mu(dt).$$

In case μ has density f, $\hat{\mu}$ is also denoted $\hat{f}$ and called the Fourier transform of f. Note that

$$\hat{f}(x) = \int e^{ixf(t)}\, dt.$$

In a probabilistic context, the Fourier transform of a random variable or distribution is called its *characteristic function*; these will be discussed in more detail in section 5.5.

The importance of characteristic functions for the proof of the CLT is largely the fact that weak convergence of a sequence of random variables (or distributions) is equivalent to pointwise convergence of their characteristic functions; this is very useful because pointwise convergence is much easier to work with analytically. Also, for any measures μ, ν on $\mathbb{R}$ with the Borel σ-algebra, the Fourier transform of $\mu * \nu$ is simply $\hat{\mu}\hat{\nu}$, the pointwise product of $\hat{\mu}$ and $\hat{\nu}$. Thus if X, Y are independent random variables with characteristic functions φ, ψ, respectively, the characteristic function of $X + Y$ is simply $\varphi\psi$. This is helpful because pointwise products are in general much easier to work with than convolutions. In particular, the n-fold convolution which would arise when computing the distribution of a sum of n independent identically distributed random variables (as occurs in the statement of the central limit theorem) is replaced by the nth power of their characteristic functions, a much more approachable object. All this will be discussed more thoroughly in section 5.5.

3 Summary of the Proof

Before finally diving into the details of the formalization, we pause to give a quick overview of how the proof will succeed. Our model proof for the central limit theorem was that found in Billingsley's classic text *Probability and Measure* [6], and we refer the reader seeking additional details to that excellent work.

As noted in the preliminaries, if X and Y are independent random variables with distributions μ and ν, respectively, the distribution of their sum is the convolution of their distributions: $X + Y \sim \mu * \nu$. Thus if $\langle X_k \mid k \leq n \rangle$ is a sequence of independent random variables all with distribution μ, the sum $\sum_{k \leq n} X_k$ is distributed as the n-fold convolution of μ with itself. However, this n-fold convolution is a technically inconvenient object to work with, and to study the asymptotic distribution of $\sum_{k \leq n} X_k$ as $n \to \infty$, it turns out to be technically advantageous to take Fourier transforms of the random variables, or to put this in probabilistic language, to study their characteristic functions. The advantage of this method of characteristic functions is twofold: first, if X and Y

are independent then the characteristic function of $X + Y$ is simply the product of the characteristic function of X and that of Y; and secondly, a sequence $\langle X_n \mid n \in \mathbb{N} \rangle$ converges in distribution to a random variable X if and only if the corresponding characteristic functions converge pointwise. This is a significant advantage because products are far easier to work with than convolutions, as is pointwise convergence easier than convergence in distribution. This shift of focus from random variables to their characteristic functions is justified by the Lévy inversion theorem, which entails that two random variables with the same characteristic function have the same distribution.

We can now express the idea of the proof of the central limit theorem: The characteristic function of the average of n independent identically distributed random variables with unit variance is shown to converge pointwise to the characteristic function of the standard normal distribution. The general result for averages of variables with finite nonzero variance can then be deduced by normalizing general variables to have zero mean and unit variance. Proving the case of zero mean and unit variance is reasonably straightforward, though it requires several delicate estimates based on Taylor series and complex variables, all of which require rather tedious formalization. The bulk of the work in our formalization, however, was in supporting the use of characteristic functions to study the distributions of sums of independent random variables, by proving the Lévy inversion and continuity theorems.

The Lévy inversion theorem is the result which justifies using characteristic functions to study probability distributions, for it demonstrates that if two distributions have the same characteristic function, they are in fact the same distribution. The proof of this theorem employs something akin to kernel methods from harmonic analysis, using the function[4]

$$\mathrm{Si}(t) = \int_0^t \frac{\sin x}{x}\, dx$$

to "concentrate" near points of interest. This requires, among other things, proving that $\lim_{t\to\infty} \mathrm{Si}(t) = \pi/2$, which we accomplished using Fubini's theorem (see Billingsley [6], pp. 235–236) and required tedious verification of many "obvious" facts about integrals (including the validity of integration by substitution). Deriving uniqueness from the Lévy continuity theorem required using the fact that the complement of a countable subset of $\mathbb{R}$ is dense in $\mathbb{R}$ and the π-λ theorem from measure theory (the latter having already been formalized by Hölzl).

Verification of the Lévy continuity theorem required proving the portmanteau theorem, more calculations with integrals (and another use of Fubini's theorem), and use of the theory of tightness of sequences of probability measures. The portmanteau theorem establishes that convergence in distribution is equivalent to various definitions of weak* convergence in the sense of functional analysis; we verified equivalence to two such definitions, as done in Billingsley

[4]The notation $\int_a^b f(x)\, dx$, familiar from calculus, denotes simply the Lebesgue integral of f over the interval $[a, b]$ with respect to Lebesgue measure.

[6]. The proof of the portmanteau theorem uses Skorohod's theorem, which states that if a sequence $\langle \mu_n \mid n \in \mathbb{N} \rangle$ of probability measures converges in distribution to a probability measure μ, then there exists a sequence of random variables $\langle X_n \mid n \in \mathbb{N} \rangle$ and a random variable X, all defined on a common probability space, such that $X_n \sim \mu_n$, $X \sim \mu$, and $X_n \to X$ pointwise. The proof of this result requires only elementary analysis.

The theory of tightness of sequences of measures gives an analogue in the space of probability measures of the Weierstrass theorem that every bounded sequence of reals has a convergent subsequence. Tightness is the requisite analogue of boundedness: Roughly a sequence of probability measures is called tight iff no mass "escapes to infinity." The sequence $\langle \mu_n \mid n \in \mathbb{N} \rangle$, where μ_n is a unit mass at n, gives an example of how mass can "escape to infinity." The key result regarding tightness of a sequence $\langle \mu_n \mid n \in \mathbb{N} \rangle$ of probability measures is that it is equivalent to the condition that for every subsequence $\langle \mu_{n_k} \mid k \in \mathbb{N} \rangle$ there exists a subsubsequence $\langle \mu_{n_{k_j}} \mid j \in \mathbb{N} \rangle$ which converges in distribution to some probability measure. A corollary of this result is that if a sequence $\langle \mu_n \mid n \in \mathbb{N} \rangle$ is tight, and it can be shown that every subsequence which has a weak limit must converge in distribution to a given probability measure μ, then in fact $\mu_n \Rightarrow \mu$.

The proof of the main result concerning tight sequences of measures requires the Helly selection theorem, which is of importance also in functional analysis. This is another analogue of Weierstrass theorem, this time giving that if $\langle F_n \mid n \in \mathbb{N} \rangle$ is any sequence of distribution functions[5] then it has a subsequence $\langle F_{n_k} \mid k \in \mathbb{N} \rangle$ which converges vaguely to some nondecreasing, right-continuous function F (which may not be a distribution function because its limit at ∞ may be less than 1). Vague convergence is defined the same way as weak convergence, and differs only in not requiring that the limit function have limit 1 at ∞. The Helly selection theorem is proven using the method of diagonal subsequences to obtain values for the requisite function F at rationals, and extending to all reals by right-continuity.[6] The elementary analytical arguments involved in this were straightforward but tedious to verify.

4 The Isabelle Interactive Proof Assistant

Before giving details of our formalization, we pause for a quick overview of some relevant features of the Isabelle interactive proof assistant that we employed. Readers familiar with this system may wish to skip to section 5, though some of the material on axiomatic type classes and locales in section 4.2 and on the implementation of general limits using a definition involving filters in section 4.3 may still be of interest.

[5] The distribution function F of a Borel probability measure μ on $\mathbb{R}$ is given by $F(x) = \mu(-\infty, x]$; these functions are treated in section 5.1.

[6] This summary is not quite accurate, but provides reasonable intuition. The actual proof is given in section 5.3.

4.1 Overview of Isabelle

When the author first decided to undertake an interactive formalization project, he knew nothing about the available proof assistants, and chose Isabelle with Avigad's advice. Other possibilities considered were Coq and HOL-Light, a variant of the HOL system. The Coq assistant [5] had the best support for algebraic reasoning (in the manner of abstract algebra), and an attractive implementation of dependent type theory (which allows types to depend on parameters, e.g. the type of integers modulo n depends on the natural number n). However, Coq is intended to formalize proofs in constructive logic, and constructive analysis is in general very different from classical analysis. This could be overcome by adding the law of the excluded middle (or something entailing it, such as the axiom of choice) as an axiom. That solution is slightly inelegant, but the main reason for deciding against using Coq was the fact that it had far less powerful automated tools than our other options. HOL-Light [17] arguably had the most extensive support for analysis (the Jordan curve theorem [15] and other significant results had already been formalized in this system, and it was the system of choice for the Flyspeck project [14], a now-completed effort to formalize Tom Hales' proof of the Kepler conjecture), but its user-interface was more difficult than that of the others (the user must directly type ML code). Isabelle had both good analysis libraries and a good user interface, and furthermore was the system chosen by Hölzl and collaborators for the development of measure-theory libraries, making it a clearly optimal choice for the formalization of the central limit theorem.

The Isabelle system was initially developed by Larry Paulson at Cambridge University and Tobias Nipkow at Technische Universität München, and has grown to include a large number of additional developers and library contributers. Isabelle is generic in the sense that it provides a small, trusted core reasoner (known as "Isabelle/Pure") and allows extension in any direction over that. It employs an LCF-style architecture [11] to ensure extensions preserve soundness. The extension of Pure which has received the most attention by library developers is Isabelle/HOL, where HOL stands for higher-order logic. Our formalization was carried out in Isabelle/HOL-Probability, an extension of the HOL main library incorporating a significant amount of measure-theoretic probability theory.

In simple form, higher-order logic is a conservative extension of first-order logic incorporating quantification over predicates and functions, higher-order predicates and functions (e.g. a predicate $T(R)$ which holds iff R is a transitive binary relation), and quantification over these. This can be augmented by a definite description operator (`THE` in Isabelle), and the extension remains conservative roughly because definite descriptions can be eliminated via Russell's well-known interpretation [23]. The indefinite description operator `SOME` included in Isabelle/HOL is a more radical departure, for its presence entails the axiom of choice (which can be easily stated in higher-order logic), and thus breaks conservativity of the HOL extension. However, the axiom of choice (or at least some weak variant of it) is essential to the usual development of mathe-

matical analysis, and so the indefinite description operator is a welcome addition for our purposes.

One might reasonably ask why higher-order logic rather than set theory is used as the basis of our formalization. Pragmatically, the Isabelle support for higher-order logic is far better than that for set theory, but why is this? A number of advantages exist, including functions as primitives, type-checking and type inference, and the additional resources provided by the type system for pruning the search space of automated procedures. However, all these advantages could in principle be obtained in a system based on set theory, say by augmenting ordinary set theory with "weak types." Further discussion of the tradeoffs between higher-order logic and set theory in the context of automated deduction, and possibilities for combining the two approaches, are found in [12].

We shall now briefly describe some of the Isabelle infrastructure and tools we used while formalizing the central limit theorem. Readers seeking to gain some working knowledge of Isabelle are advised to consult Nipkow's tutorial introduction [21] and the wealth of resources available at `http://isabelle.in.tum.de/`. A high-level overview can also be found at that website.

We mentioned earlier that one of our reasons for choosing Isabelle was its friendly user-interface, and this is provided by the Isar proof-scripting language [27], developed by Markus Wenzel as part of his doctoral thesis at Technische Universität München. The goal was to provide a more human-readable paradigm for proof scripts, and we certainly agree that Wenzel achieved his goal. The old style for proof scripts consisted in repeatedly applying tactics (using the `apply` method) to refine a goal into subgoals until all these are proved. Figure 1 gives an example of a tactic-style proof from our formalization, while figure 2 gives an example of an Isar proof. These styles can also be combined, as seen in figure 3. Clearly Isar offers a great improvement in readability, and hence in ease of maintenance, for proof scripts.

Nevertheless, often it is easier to hack through a tactic-style proof than to carefully structure an Isar proof, and so a significant amount of our formalization is still written in tactic style. Most of the tactic-style proofs involve fiddly calculations in one way or another, and we see both an opportunity to clean up our own proof scripts (rewriting long tactic-style proofs in Isar is one of the main goals of our continued development of the proof scripts) and to improve the usability of Isabelle for doing fiddly calculations without resorting to a tactic-style proof. It seems likely advances in Isar, Isabelle's libraries, and Isabelle's automated tools could all benefit this latter goal.

To help the user prove theorems more efficiently, Isabelle provides support for proof automation. The basic tools for this are the equational-reasoning simplifier `simp` (though more functionality is being continually built into this tool), the higher-order tableau prover `auto` (which uses extensive heuristic reasoning that can be influenced by the user), and the first-order sequent prover `blast`, though many variants and alternatives are available. In addition, the `sledgehammer` tool [22] provides a link to an extensible suite of external provers. Proofs generated by `sledgehammer` tools can be inserted into Isabelle proof scripts via proof

```
lemma ex_18_4_2_ubdd_integral:
  fixes x
  assumes pos: "0 < x"
  shows "LBINT u=0..∞. |exp (-u * x) * sin x| = |sin x| / x"
 apply (subst interval_integral_FTC_nonneg [where F = "λu. 1/x * (1 -
exp (-u * x)) * |sin x|"
    and A = 0 and B = "abs (sin x) / x"])
  apply force
  apply (rule ex_18_4_2_deriv)
  apply auto

  apply (subst zero_ereal_def)+
  apply (simp_all add: ereal_tendsto_simps)

  apply (rule filterlim_mono [of _ "nhds 0" "at 0"], auto)
  prefer 2
  apply (rule at_le, simp)
  apply (subst divide_real_def)
  apply (rule tendsto_mult_left_zero)+
  apply (subgoal_tac "0 = 1 - 1")
  apply (erule ssubst)
  apply (rule tendsto_diff, auto)
  apply (subgoal_tac "1 = exp 0")
  apply (erule ssubst)
  apply (rule tendsto_compose[OF tendsto_exp])
  apply (subst isCont_def [symmetric], auto)
  apply (rule tendsto_minus_cancel, auto)
  apply (rule tendsto_mult_left_zero, rule tendsto_ident_at)

  apply (subst divide_real_def)+
  apply (subgoal_tac "abs (sin x) * inverse x = 1 * abs (sin x) *
inverse x")
  apply (erule ssubst)
  apply (rule tendsto_mult)+
  apply auto
  apply (rule tendsto_eq_intros)
  apply (rule tendsto_const)
  apply (rule filterlim_compose[OF exp_at_bot])
  unfolding filterlim_uminus_at_bot
  apply simp
  apply (subst mult.commute)
  apply (rule filterlim_tendsto_pos_mult_at_top[OF tendsto_const pos
filterlim_ident])
  apply simp
  done
```

Figure 1: A tactic-style proof.

```
lemma (in pair_sigma_finite) Fubini_integrable:
  fixes f :: "_ ⇒ _::{banach, second_countable_topology}"
  assumes f[measurable]: "f ∈ borel_measurable (M1 ⨂M M2)"
    and integ1: "integrable M1 (λx. ∫ y. norm (f (x, y)) ∂M2)"
    and integ2: "AE x in M1. integrable M2 (λy. f (x, y))"
  shows "integrable (M1 ⨂M M2) f"
proof (rule integrableI_bounded)
  have "(∫+ p. norm (f p) ∂(M1 ⨂M M2)) = (∫+ x. (∫+ y. norm (f (x,
y)) ∂M2) ∂M1)"
    by (simp add: M2.nn_integral_fst [symmetric])
  also have "... = (∫+ x. |∫ y. norm (f (x, y)) ∂M2| ∂M1)"
    apply (intro nn_integral_cong_AE)
    using integ2
  proof eventually_elim
    fix x assume "integrable M2 (λy. f (x, y))"
    then have f: "integrable M2 (λy. norm (f (x, y)))"
      by simp
    then have "(∫+y. ereal (norm (f (x, y))) ∂M2) = ereal (LINT y|M2.
norm (f (x, y)))"
      by (rule nn_integral_eq_integral) simp
    also have "... = ereal |LINT y|M2. norm (f (x, y))|"
      using f by (auto intro!: abs_of_nonneg[symmetric]
integral_nonneg_AE)
    finally show "(∫+y. ereal (norm (f (x, y))) ∂M2) = ereal |LINT
y|M2. norm (f (x, y))|" .
  qed
  also have "... < ∞"
    using integ1 by (simp add: integrable_iff_bounded
integral_nonneg_AE)
  finally show "(∫+ p. norm (f p) ∂(M1 ⨂M M2)) < ∞" .
qed fact
```

Figure 2: An Isar proof.

```
lemma Billingsley_ex_17_5:
  shows "set_integrable lborel (einterval (-∞) ∞) (λx. inverse (1 +
x^2))"
    "LBINT x=-∞..∞. inverse (1 + x^2) = pi"
proof -
  have 1: "⋀x. - (pi / 2) < x ⟹ x * 2 < pi ⟹ (tan
has_real_derivative 1 + (tan x)^2) (at x)"
    apply (subst tan_sec)
    using pi_half cos_is_zero
    apply (metis cos_gt_zero_pi less_divide_eq_numeral1(1)
less_numeral_extra(3))
    using DERIV_tan
    by (metis cos_gt_zero_pi less_divide_eq_numeral1(1) power2_less_0
power_inverse
      power_zero_numeral)
  have 2: "⋀x. - (pi / 2) < x ⟹ x * 2 < pi ⟹ isCont (λx. 1 + (tan
x)^2) x"
    apply (rule isCont_add, force)
    apply (subst power2_eq_square)
    apply (subst isCont_mult)
    apply (rule isCont_tan)

    using pi_half cos_is_zero
    apply (metis cos_gt_zero_pi less_divide_eq_numeral1(1)
less_numeral_extra(3))
    by simp
  show "LBINT x=-∞..∞. inverse (1 + x^2) = pi"
    apply (subst interval_integral_substitution_nonneg[of "-pi/2" "pi/2"
tan "λx. 1 + (tan x)^2"])
    apply (auto intro: derivative_eq_intros simp add:
ereal_tendsto_simps filterlim_tan_at_left add_nonneg_eq_0_iff)
    apply (erule (1) 1)
    apply (erule (1) 2)
    apply (subst minus_divide_left)+
    by (rule filterlim_tan_at_right)
  show "set_integrable lborel (einterval (-∞) ∞) (λx. inverse (1 +
x^2))"
    apply (subst interval_integral_substitution_nonneg[of "-pi/2" "pi/2"
tan "λx. 1 + (tan x)^2"])
    apply (auto intro: derivative_eq_intros simp add:
ereal_tendsto_simps filterlim_tan_at_left add_nonneg_eq_0_iff)
    apply (erule (1) 1)
    apply (erule (1) 2)
    apply (subst minus_divide_left)+
    by (rule filterlim_tan_at_right)
qed
```

Figure 3: A mixed-style proof.

text automatically generated by `sledgehammer`; typically these will employ the Isabelle-native SMT solver[7] `metis`. The tools `simp` and `auto` and their variants can be influenced by the user by declaring results to be simplification rules `[simp]` or introduction rules `[intro]`, which enables their use by `simp` or `auto`, respectively. All these tools were extensively employed in our formalization.

4.2 Types and Locales

The existence of the type `nat` of natural numbers must be guaranteed by an axiom of infinity (because otherwise higher-order logic has models where all types are finite), but everything else can be built up from there. Integers (type `int`) are defined as equivalence classes of pairs of natural numbers, and rationals (type `rat`) as equivalence classes of pairs of integers, both in the usual manner (see for example [8]). Real numbers (type `real`) are defined as equivalence classes of Cauchy sequences of rational numbers (see [25]; perhaps a more common method of construction in analysis texts is that via Dedekind cuts). Extended reals (type `ereal`) are defined as the sum type of $\{-\infty\}$, $\mathbb{R}$, and $\{\infty\}$, with the expected ordering and operations. Complex numbers are defined as pairs of real numbers.

All this is supported by Isabelle's system of polymorphic type constructors, which we briefly describe. Full detail can be found in the Isabelle/HOL documentation available at `http://isabelle.in.tum.de/documentation.html`. The fact that a variable x has type α can be indicated with the annotation $x::\alpha$, though Isabelle's polymorphic type inference system often eliminates the need for such annotations. If α and β are types, the type $\alpha \Rightarrow \beta$ is the type of functions with domain α and codomain β (note that this is more commonly denoted $\alpha \to \beta$ in the literature on higher-order logic), while $\alpha \times \beta$ is the type of pairs with first element of type α and second of type β. If α is a type, so is α `set`, the type of sets of elements of type α. These could be interpreted as predicates (functions of type $\alpha \Rightarrow$ `bool`), where `bool` is the type with precisely two elements `true`, `false`), but it is harmless and notationally useful to take these as separate types.

Types and type variables (such as α in α `set`) can belong to axiomatic type classes, for example the annotation α`::linordered_field` indicates that α is a type with the structure of an ordered field (and so α possesses binary operations $+$ and $\cdot$, unary operations $-$ and $^{-1}$, constants 0 and 1, and a binary relation $\leq$ which satisfy the axioms of a linearly ordered field). See [13] for further details. Types can also be organized into locales, which allows them to be structured into hierarchies (as is familiar in an informal setting from algebra). For example, if the locale `field` is a sublocale of the locale `ring`, then any theorem proved for rings can be used for fields. A type can be shown to instantiate a locale with the reserved words `instance` and `instance proof`. Further information

[7] SMT stands for Satisfiability Modulo Theories; an SMT solver demonstrates satisfiability of a normal form in the presence of assumptions from a background theory. A quick overview can be found in [7].

on locales can be found in [4]. Syntax for locales will arise as we examine the proof scripts developed for the central limit theorem.

4.3 Limits and Filters

Because it is interesting and different than what is found in standard concrete presentations of analysis, we describe the flexible method of filterlimits used in the Isabelle limit libraries. Full detail can be found in [19].

Definition 4.1. Let X be a set. A *filter* over X is a nonempty set $\mathcal{F} \subseteq \mathcal{P}(X)$ such that

1. If $A \subseteq B$ and $A \in \mathcal{F}$, then $B \in \mathcal{F}$.
2. If $A, B \in \mathcal{F}$, then $A \cap B \in \mathcal{F}$.

Filters can be thought of as "large" sets, in some vague sense. $\mathcal{P}(X)$ is the *trivial* filter, and often must be excluded (say by the condition that $\emptyset \notin \mathcal{F}$). In Isabelle, predicates are employed instead of sets in the definition of a filter.

We now turn to some examples of filters. Please note that none of the notation used here is standard. The verification that these actually are filters is elementary, and the interested reader may regard it as an exercise.

1. If $A \subseteq X$, the filter $\mathfrak{P}_A = \{B \subseteq X \mid A \subseteq B\}$ is called the *principal filter* over A. A filter which is not equal to $\mathfrak{P}_A$ for any $A \subseteq X$ is called *nonprincipal*. In case $A = \{a\}$ is a singleton, $\mathfrak{P}_{\{a\}}$ is denoted $\mathfrak{P}_a$ and called the principal filter over a.
2. For X infinite, the smallest nonprincipal filter on X is the *cofinite filter* $\{A \subseteq X \mid |X \setminus A| < \infty\}$.
3. If τ is a topology on X,

   $$\mathcal{N}_x^\tau = \{A \subseteq X \mid \exists U \in \tau.\ x \in U \subseteq A\}$$

   is the *neighborhood filter* on X, while

   $$\mathcal{N}_{(x)}^\tau = \{B \subseteq X \mid \exists A \in \mathcal{N}_x^\tau.\ B = A \setminus \{x\}\}$$

   is the *punctured neighborhood filter*. The superscript τ can be omitted when there is no danger of ambiguity.
4. If $\leq$ is a linear order on X, there are two natural filters on X "going to infinity" in the two possible directions. These are

   $$\mathcal{U} = \{A \subseteq X \mid \forall x \in A.\ x \leq y \Longrightarrow y \in A\},$$

   the filter of *upper sets*, and

   $$\mathcal{L} = \{A \subseteq X \mid \forall x \in A.\ y \leq x \Longrightarrow y \in A\},$$

   the filter of *lower sets*.

5. We can also define left and right "punctured half-neighborhood filters" for a linear order $(X, \leq)$:

$$\mathcal{H}_x^+ = \{A \subseteq X \mid \exists y > x.\ \forall z \in X.\ x \leq z \leq y \implies z \in A\}$$

is the right punctured half-neighborhood filter of x, while

$$\mathcal{H}_x^- = \{A \subseteq X \mid \exists y < x.\ \forall z \in X.\ y \leq z \leq x \implies z \in A\}$$

is the left punctured half-neighborhood filter.

A property P is said to hold *eventually* with respect to a filter $\mathcal{F}$ iff $\{x \mid Px\} \in \mathcal{F}$. For example, if $\mathfrak{P}_a$ is the principal filter over a, a property holds eventually with respect to $\mathfrak{P}_a$ iff it holds at a. A property holds eventually with respect to the cofinite filter iff it holds at all but finitely many points, and it holds eventually with respect to the neighborhood filter $\mathcal{N}_x$ of x iff it holds in some neighborhood of x. A property which holds eventually with respect to $\mathcal{U}$ may be thought of as holding "at ∞," and analogously for $\mathcal{L}$.

A filter $\mathcal{G}$ is *finer than* $\mathcal{F}$ iff fewer properties hold eventually for $\mathcal{G}$ than for $\mathcal{F}$, equivalently iff $\mathcal{G} \subseteq \mathcal{F}$. For example, if (X, τ) is a topological space and $x \in X$, we have that $\mathcal{N}_x$ is finer than $\mathcal{N}_{(x)}$, which is in turn finer than $\mathfrak{P}_x$.

Convergence is now easily defined: A function f converges to a filter $\mathcal{G}$ with respect to a filter $\mathcal{F}$ iff $f(\mathcal{F})$ is finer than $\mathcal{G}$, where $f(\mathcal{F}) = \{A \mid f^{-1}[A] \in \mathcal{F}\}$ (Hölzl calls this the *filtermap* operation). In case $\mathcal{G}$ is $\mathcal{N}_{(x)}$ for some x in a topological space, we say that f converges to x with respect to the filter $\mathcal{F}$. The extra filter $\mathcal{G}$ is useful for saying such things as that a function f converges to $\mathcal{U}$, or in intuitive terms $f \to \infty$; this is easily expressible with the two-filter definition of convergence but would be lost if the filter to which a function converges were always assumed to be a puncturedneighborhood filter. Also, limits from the left and right can be expressed in terms of $\mathcal{H}_x^-$ and $\mathcal{H}_x^+$ rather than requiring separate definitions. This filterlimit framework gives rise to an amazing amount of flexibility which was very useful when formalizing the CLT; further examples and explanation are found in [19].

5 The Formalized Proof

Having given an overview of the proof we formalized and the system in which it was carried out, we turn now to the details. Along the way, we shall try to point out best practices, pitfalls, and opportunities for improvement which we encountered in the course of our formalization. Full proof scripts are often included, but not generally intended to be read in full detail. A quick skim of these scripts should give a flavour of how they work, and we always include an informal proof first.

Why include formal proofs at all? It is a common practice when presenting a formalization to omit all formal proofs and just indicate the informal proofs of formalized statements. This implicitly assumes that the proof scripts

are far less readable than their informal counterparts, which is certainly true to an extent, but it is our hope that the readability of proof scripts has improved to the point that the reader may benefit by at least skimming them. At the very least this will give side-by-side comparisons of formal to informal mathematics, a comparison sometimes difficult to make when reading papers on formalization. If scripts are particularly long or difficult to read, we omit them (of course the full formalization is available in the project git repository `https://github.com/avigad/isabelle`).

5.1 Distribution Functions

Often it is more convenient to work with a real-valued function determining a measure on $\mathbb{R}$ than directly with the measure, and an obvious way to accomplish this for finite measures is to study the distribution function of the measure, which when evaluated at an argument gives the amount of mass below that argument.

Definition 5.1. Let μ be a finite measure on $\mathbb{R}$. The (cumulative) distribution function F_μ is defined by $F_\mu(x) = \mu(-\infty, x]$.

In Isabelle, this is rendered as

```
definition
  cdf :: "real measure ⇒ real ⇒ real"
where
  "cdf M ≡ λx. measure M {..x}"

lemma cdf_def2: "cdf M x = measure M {..x}"
```

The lemma gives what is intuitively the definition; the actual definition uses lambda abstraction. If $\tau(x)$ is a term, $\lambda x.\, \tau(x)$ is the function which, when evaluated at input a, gives $\tau(a)$. This is a standard notation in higher-order logic; for more details we refer the reader to any standard textbook treatment of the subject, for example [1].

Before proving that the distribution function of a measure uniquely determines that measure, let us note some general properties of distribution functions. For convenience we assume the measures we are working with are probability measures; other nonzero finite measures can be normalized to probability measures, and the zero measure is trivial.

Theorem 5.2. The distribution function F_μ of a finite measure μ is nondecreasing and right-continuous and satisfies $\lim_{x\to-\infty} F_\mu(x) = 0$ and $\lim_{x\to\infty} F_\mu(x) = 1$.

F_μ nondecreasing follows from monotonicity of μ; right-continuity and $\lim_{x\to-\infty} F_\mu(x) = 0$ follow from continuity of μ from above as if $x_n \downarrow x$ then $(-\infty, x_n] \downarrow (-\infty, x]$ and $(-\infty, -n] \downarrow \emptyset$. The fact that $\lim_{x\to\infty} F_\mu(x) = 1$ follows from continuity of μ from below as $(-\infty, n] \uparrow \mathbb{R}$.

In Isabelle this expands to

```
lemma cdf_nondecreasing [rule_format]: "(∀x y. x ≤ y ⟶ cdf M x ≤ cdf
M y)"
unfolding cdf_def by (auto intro!: finite_measure_mono)

lemma cdf_is_right_cont: "continuous (at_right a) (cdf M)"
unfolding continuous_within
proof (rule tendsto_at_right_sequentially[where b="a + 1"])
  fix f :: "nat ⇒ real" and x assume f: "decseq f" "f ----> a"
  then have "(λn. cdf M (f n)) ----> measure M (⋂i. {.. f i})"
    unfolding cdf_def
    apply (intro finite_Lim_measure_decseq)
    using ‘decseq f‘ apply (auto simp: decseq_def)
    done
  also have "(⋂i. {.. f i}) = {.. a}"
    using decseq_le[OF f] by (auto intro: order_trans LIMSEQ_le_const[OF f(2)])
  finally show "(λn. cdf M (f n)) ----> cdf M a"
    by (simp add: cdf_def)
qed simp

lemma cdf_lim_at_bot: "(cdf M ---> 0) at_bot"
proof -
  have 1: "((%x :: real. - cdf M (- x)) ---> 0) at_top"
    apply (rule tendsto_at_topI_sequentially_real)
    apply (auto simp add: mono_def cdf_nondecreasing cdf_lim_neg_infty)
    using cdf_lim_neg_infty by (metis minus_zero tendsto_minus_cancel_left)
  from filterlim_compose [OF 1, OF filterlim_uminus_at_top_at_bot]
  show ?thesis
    by (metis "1" filterlim_at_bot_mirror minus_zero tendsto_minus_cancel_left)
qed

lemma cdf_lim_at_top_prob: "(cdf M ---> 1) at_top"
by (subst prob_space [symmetric], rule cdf_lim_at_top)
```

The annotation `[rule_format]` indicates that the universally quantified variables may be freely instantiated by Isabelle's automated tools, and that the implication may be used to refine a subgoal (replacing the consequent with the antecedent).

In turn, any function with the properties listed in the preceding theorem is the distribution function of a probability measure on $\mathbb{R}$:

Theorem 5.3. Suppose $F\colon \mathbb{R} \to \mathbb{R}$ is nondecreasing, right-continuous, and satisfies $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$. Then there exists a Borel probability measure μ on $\mathbb{R}$ such that $F = F_\mu$.

The requisite measure μ is constructed by defining $\mu(a, b] = F(b) - F(a)$ and extending this to the Borel σ-algebra using the Carathéodory extension theorem, which Johannes Hölzl had already formalized. In Isabelle this is

```
lemma real_distribution_interval_measure:
  fixes F :: "real ⇒ real"
  assumes nondecF : "⋀ x y. x ≤ y ⟹ F x ≤ F y" and
    right_cont_F : "⋀a. continuous (at_right a) F" and
    lim_F_at_bot : "(F ---> 0) at_bot" and
    lim_F_at_top : "(F ---> 1) at_top"
  shows "real_distribution (interval_measure F)"
proof -
  let ?F = "interval_measure F"
  interpret prob_space ?F
  proof
    have "ereal (1 - 0) = (SUP i::nat. ereal (F (real i) - F (- real i)))"
      by (intro LIMSEQ_unique[OF _ LIMSEQ_SUP] lim_ereal[THEN iffD2] tendsto_intros
         lim_F_at_bot[THEN filterlim_compose] lim_F_at_top[THEN filterlim_compose]
         lim_F_at_bot[THEN filterlim_compose] filterlim_real_sequentially
         filterlim_uminus_at_top[THEN iffD1])
         (auto simp: incseq_def intro!: diff_mono nondecF)
    also have "... = (SUP i::nat. emeasure ?F {- real i<..real i})"
      by (subst emeasure_interval_measure_Ioc) (simp_all add: nondecF right_cont_F)
    also have "... = emeasure ?F (⋃i::nat. {- real i<..real i})"
      by (rule SUP_emeasure_incseq) (auto simp: incseq_def)
    also have "(⋃i. {- real (i::nat)<..real i}) = space ?F"
      by (simp add: UN_Ioc_eq_UNIV)
    finally show "emeasure ?F (space ?F) = 1"
      by (simp add: one_ereal_def)
  qed
  show ?thesis
    proof qed simp_all
qed
```

Here `real_distribution` is a locale for Borel probability measures, and `interval_measure` is a function defined by Hölzl which generates a measure from a nondecreasing, right-continuous function. Note the calculation paradigm `have...also have...finally show`; this is Isar syntax for chaining equations and is very convenient for writing calculations in Isar as opposed to a tactic style.

A method of proof similar to that used to prove all nondecreasing right-continuous functions with the appropriate limits at $\pm\infty$ are distribution functions also gives that the distribution function of a probability measure is unique in the sense that if $F_\mu = F_\nu$ then $\mu = \nu$, because $F_\mu = F_\nu$ implies $\mu(a, b] = \nu(a, b]$ for every a, b, and the half-open intervals are a π-system generating the Borel sets on $\mathbb{R}$, so $\mu = \nu$ by Dynkin's uniqueness lemma[8]. In Isabelle this is

```
lemma cdf_unique:
  fixes M1 M2
  assumes "real_distribution M1" and "real_distribution M2"
  assumes "cdf M1 = cdf M2"
  shows "M1 = M2"
```

[8] This is an easy consequence of the π-λ theorem; see [6], p. 42.

```
proof (rule measure_eqI_generator_eq[where Ω=UNIV])
  fix X assume "X ∈ range (λ(a, b). {a<..b::real})"
  then obtain a b where Xeq: "X = {a<..b}" by auto
  then show "emeasure M1 X = emeasure M2 X"
    by (cases "a ≤ b")
       (simp_all add: assms(1,2)[THEN real_distribution.emeasure_Ioc] assms(3))
next
  show "(⋃i. {- real (i::nat)<..real i}) = UNIV"
    by (rule UN_Ioc_eq_UNIV)
qed (auto simp: real_distribution.emeasure_Ioc[OF assms(1)]
  assms(1,2)[THEN real_distribution.events_eq_borel] borel_sigma_sets_Ioc
  Int_stable_def)
```

5.2 Weak Convergence

We are finally ready to give the definition of weak convergence. The fundamental notion is for distribution functions, and it immediately lifts to measures (distributions) and random variables.

Definition 5.4. A sequence $\langle F_n \mid n \in \mathbb{N} \rangle$ of distribution functions *converges weakly* to a distribution function F iff

$$\lim_{n \to \infty} F_n(x) = F(x)$$

for every $x \in \mathbb{R}$ such that F is continuous at x. A sequence $\langle \mu_n \mid n \in \mathbb{N} \rangle$ of probability measures converges weakly to a probability measure μ iff the corresponding distribution functions converge weakly, and a sequence $\langle X_n \mid n \in \mathbb{N} \rangle$ of random variables converges weakly to a random variable X iff the corresponding distributions converge weakly.

To see why convergence is allowed to fail at continuity points of F, note that intuitively a sequence $\langle \mu_n \mid n \in \mathbb{N} \rangle$ of unit masses at a_n should converge to a unit mass μ at a iff $a_n \to a$. The distribution functions F_n of μ_n are two-valued, taking either the value zero or one, and so if $a_n > a$ for infinitely many n, $F_n(a)$ does not converge to $F(a)$ (since $F_n(a) = 0$ if $a_n > a$, while $F(a) = 1$). This example is taken from Billingsley [6], and further explanation and examples can be found in sections 14 and 25 of that book.

In Isabelle we have:

```
definition
  weak_conv :: "(nat ⇒ (real ⇒ real)) ⇒ (real ⇒ real) ⇒ bool"
where
  "weak_conv F_seq F ≡ ∀x. isCont F x ⟶ (λn. F_seq n x) ----> F x"

definition
  weak_conv_m :: "(nat ⇒ real measure) ⇒ real measure ⇒ bool"
where
  "weak_conv_m M_seq M ≡ weak_conv (λn. cdf (M_seq n)) (cdf M)"
```

```
abbreviation (in prob_space)
  "weak_conv_r X_seq X ≡ weak_conv_m (λn. distr M borel (X_seq n)) (distr
M borel X)"
```

One technical result we needed for working with weak convergence was the fact that a nondecreasing function $f\colon \mathbb{R} \to \mathbb{R}$ has just countably many discontinuities. For this we shall use the concept of the oscillation of a function.

Definition 5.5. Let $f\colon \mathbb{R} \to \mathbb{R}$, and $x \in \mathbb{R}$. The *oscillation* of f at x is defined by

$$\operatorname{osc}_x f = \limsup_{t \to x} f(t) - \liminf_{t \to x} f(t).$$

Note that the oscillation is always nonnegative, and f is continuous at x iff $\operatorname{osc}_x f = 0$.

Lemma 5.6. Suppose $f\colon \mathbb{R} \to \mathbb{R}$ is nondecreasing. Then the discontinuity set $D = \{x \in \mathbb{R} \mid \operatorname{osc}_x f > 0\}$ is countable.

Proof. We begin with showing that a finite interval contains at most countably many discontinuities of f. Suppose $a < b$, and let $D_k^{a,b}$ be the set of discontinuities of f in the interval $[a, b]$ with oscillation at least $\frac{1}{k}$. Because f is nondecreasing, $f(a) \leq f(x) \leq f(b)$ for every $x \in [a, b]$, and from the definition of oscillation it is immediate that for any $u > 0$, if $\operatorname{osc}_x f \geq u$, then $f(b) \geq f(a)+u$. Hence

$$|D_k^{a,b}| \leq \left\lfloor \frac{f(b) - f(a)}{k} \right\rfloor,$$

and in particular $D_k^{a,b}$ is finite. Thus we have that $D^{a,b} = \bigcup_{k \in \mathbb{N}} D_k^{a,b}$ is countable for every a, b (being a countable union of finite sets), and hence so is $D = \bigcup_{n \in \mathbb{N}} D^{-n,n}$. □

Rather than formalizing this directly, we formalized an argument which works for nondecreasing functions only defined on a subset $A \subseteq \mathbb{R}$, and concluded the above result from that. We omit the rather long proof of the more general result.

```
lemma mono_on_ctble_discont:
  fixes f :: "real ⇒ real"
  fixes A :: "real set"
  assumes "mono_on f A"
  shows "countable {a∈A. ¬ continuous (at a within A) f}"

lemma mono_ctble_discont:
  fixes f :: "real ⇒ real"
  assumes "mono f"
  shows "countable {a. ¬ isCont f a}"
using assms mono_on_ctble_discont [of f UNIV] unfolding mono_on_def mono_def
by auto
```

Another general lemma which we needed to formalize is the fact that nondecreasing functions on $\mathbb{R}$ are Borel measurable. This is intuitively because the

preimage of a semi-infinite interval $(-\infty, x)$ under a monotone function is an interval, and the semi-infinite intervals generate the Borel σ-algebra on $\mathbb{R}$ (note they are a subbase for the open subsets of $\mathbb{R}$). In Isabelle we formalized something slightly more general, allowing application of the theorem when a function is only defined (or only of interest) on a subset of $\mathbb{R}$. This is useful in particular in the proof of Skorohod's theorem.

```
lemma borel_measurable_mono_on_fnc:
  fixes f :: "real ⇒ real" and A :: "real set"
  assumes "mono_on f A"
  shows "f ∈ borel_measurable (restrict_space borel A)"
apply (rule measurable_restrict_countable[OF mono_on_ctble_discont[OF assms]])
  apply (auto intro!: image_eqI[where x="{x}" for x] simp: sets_restrict_space)
  apply (auto simp add: sets_restrict_restrict_space continuous_on_eq_continuous_within
              cong: measurable_cong_sets
              intro!: borel_measurable_continuous_on_restrict intro: continuous_within_subset)
  done
```

Though weak convergence has the advantage of being a flexible notion which applies much more generally than such notions as almost sure convergence or convergence in probability, it is often difficult to work with directly. The following theorem, known as Skorohod's theorem, allows one to replace weak convergence with pointwise convergence under appropriate conditions.

Theorem 5.7. Let $\langle \mu_n \mid n \in \mathbb{N} \rangle$, μ be probability measures on the $\mathbb{R}$, and suppose $\mu_n \Rightarrow \mu$. Then there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and random variables $\langle Y_n \mid n \in \mathbb{N} \rangle$, Y on Ω such that $Y_n \sim \mu_n$ for each n, $Y \sim \mu$, and $Y_n \to Y$ pointwise.

Proof. The probability space will be simply the unit interval $(0, 1)$ with Lebesgue measure. For each n, let F_n be the distribution function of μ_n, and F be the distribution function of μ. Let Y_n be the pseudoinverse of the nondecreasing function F_n, that is $Y_n(\omega) = \inf\{x \in \mathbb{R} \mid \omega \leq F_n(x)\}$ for $\omega \in (0, 1)$, and similarly $Y(\omega) = \inf\{x \in \mathbb{R} \mid \omega \leq F(x)\}$. Thus we have that for any $\omega \in (0, 1)$, $x \in \mathbb{R}$, and $n \in \mathbb{N}$, $\omega \leq F_n(x)$ iff $Y_n(\omega) \leq x$, and similarly $\omega \leq F(x)$ iff $Y(\omega) \leq x$. Consequently

$$\mathbb{P}(Y_n \leq x) = \mathbb{P}\{\omega \in (0, 1) \mid \omega \leq F_n(x)\} = F_n(x),$$

and so F_n is the distribution function of Y_n. By the same reasoning F is the distribution function of Y.

The idea of the proof that $Y_n \to Y$ pointwise is that we know $F_n \Rightarrow F$, and because Y_n is the pseudoinverse of F_n and Y is the pseudoinverse of F, this is sufficient for $Y_n(\omega)$ to converge to $Y(\omega)$ for almost all $\omega \in (0, 1)$. Note that this is certainly true in case F_n, F are continuous (in which case Y_n, Y are literal inverses of F_n, F).

Let $\omega \in (0, 1)$ and $\varepsilon > 0$. Choose x such that $Y(\omega) - \varepsilon < x < Y(\omega)$ and $\mu\{x\} = 0$, so x is a continuity point of F and $F_n(x) \to F(x)$. It is immediate from the definition of Y that $F(x) < \omega$, and so for sufficiently large

n we have $F_n(x) < \omega$, which in turn gives $Y(\omega) - \varepsilon < x < Y_n(\omega)$. Hence $\liminf_{n\to\infty} Y_n(\omega) \geq Y(\omega)$.

Now let $\omega' \in (0,1)$ and $\varepsilon > 0$, and suppose $\omega \in (0,1)$ and $\omega < \omega'$. Choose y a continuity point of F (so $\mu\{y\} = 0$) such that $Y(\omega') < y < Y(\omega') + \varepsilon$. From the definition of Y we have $\omega < \omega' \leq F(Y(\omega')) \leq F(y)$, and thus for sufficiently large n, $\omega \leq F_n(y)$, which gives $Y_n(\omega) \leq y < Y(\omega') + \varepsilon$ and hence $\limsup_{n\to\infty} Y_n(\omega) \leq Y(\omega')$. Since this holds for arbitrary $\omega < \omega'$, we have that $Y_n(\omega) \to Y(\omega)$ if Y is continuous at ω.

It is immediate from the definition of Y and the fact that F is nondecreasing that Y is nondecreasing, and hence has at most countably many discontinuities, a set of Lebesgue measure zero. Thus we may redefine Y_n, Y to be zero at all the discontinuities of Y without affecting the distributions of these random variables, and obtain $Y_n(\omega) \to Y(\omega)$ for every $\omega \in (0,1)$. □

This proof presented a number of technical challenges for formalization. For one, we needed to choose a continuity point of an arbitrary probability measure in an arbitrary open interval. For this we needed to know that the set of continuity points of an arbitrary finite measure is dense. To see that, recall that there are at most countably many atoms, and note that the complement of a countable set is dense. Rather than formalizing directly the fact that the complement of a countable set is dense, we simply proved that an interval is uncountable and so the result of subtracting the set of atoms is still uncountable, and hence nonempty.

Here is our formalization of the fact that a finite measure has countably many atoms:

```
lemma countable_atoms: "countable {x. measure M {x} > 0}"
proof -
  { fix B i
    assume finB: "finite B" and
      subB: "B ⊆ {x. inverse (real (Suc i)) < Sigma_Algebra.measure M {x}}"
    have "measure M B = (∑x∈B. measure M {x})"
      by (rule measure_eq_setsum_singleton [OF finB], auto)
    also have "... ≥ (∑x∈B. inverse (real (Suc i)))" (is "?lhs ≥ ?rhs")
      using subB by (intro setsum_mono, auto)
    also (xtrans) have "?rhs = card B * inverse (real (Suc i))"
      by simp
    finally have "measure M B ≥ card B * inverse (real (Suc i))" .
  } note * = this
  have "measure M (space M) < real (Suc (nat (ceiling (measure M (space M)))))"
    by (auto simp del: zero_le_ceiling
              simp add: measure_nonneg ceiling_nonneg intro!: less_add_one)
       linarith
  then obtain X :: nat where X: "measure M (space M) < X" ..

  { fix i :: nat
  have "finite {x. inverse (real (Suc i)) < Sigma_Algebra.measure M {x}}"
    apply (rule ccontr)
```

```
    apply (drule infinite_arbitrarily_large [of _ "X * Suc i"])
    apply clarify
    apply (drule *, assumption)
    apply (drule leD, erule notE, erule ssubst)
    apply (subst of_nat_mult)
    apply (subst mult.assoc)
    apply (subst right_inverse)
    apply simp_all
    by (rule order_le_less_trans [OF bounded_measure X])
  } note ** = this
  have "{x. measure M {x} > 0} = (⋃i :: nat. {x. measure M {x} > inverse
(Suc i)})"
    apply (auto intro: reals_Archimedean)
    using ex_inverse_of_nat_Suc_less apply auto
    by (metis inverse_positive_iff_positive less_trans of_nat_0_less_iff of_nat_Suc
zero_less_Suc)
  thus "countable {x. measure M {x} > 0}"
    apply (elim ssubst)
    apply (rule countable_UN, auto)
    apply (rule countable_finite)
    using ** by auto
qed
```

And here is our formal statement of Skorohod's theorem; the proof is very long, and we do not include it.

```
theorem Skorohod:
  fixes
    μ :: "nat ⇒ real measure" and
    M :: "real measure"
  assumes
    "⋀n. real_distribution (μ n)" and
    "real_distribution M" and
    "weak_conv_m μ M"
  shows "∃ (Ω :: real measure) (Y_seq :: nat ⇒ real ⇒ real) (Y :: real
⇒ real).
    prob_space Ω ∧
    (∀n. Y_seq n ∈ measurable Ω borel) ∧
    (∀n. distr Ω borel (Y_seq n) = μ n) ∧
    Y ∈ measurable Ω lborel ∧
    distr Ω borel Y = M ∧
    (∀x ∈ space Ω. (λn. Y_seq n x) ----> Y x)"
```

Another important fact which will be needed later is that a sequence $\langle \mu_n \mid n \in \mathbb{N} \rangle$ of probability measures converges weakly to a probability measure μ if and only if it converges to μ in the weak* topology of functional analysis. Weak* convergence of a sequence of probability measures can be defined in a variety of ways; two common ones are

$$\lim_{n \to \infty} \int f \, d\mu_n = \int f \, d\mu$$

for every bounded continuous real-valued function f, and

$$\lim_{n\to\infty} \mu_n(A) = \mu(A)$$

for every set A such that $\mu(\partial A) = 0$ (where ∂A is the boundary of A). The equivalence of weak convergence of probability measures with these two definitions of weak* convergence is called the portmanteau theorem.

Theorem 5.8. Let $\langle \mu_n \mid n \in \mathbb{N} \rangle$ be a sequence of measures on $\mathbb{R}$. The following are equivalent:

1. $\mu_n \Rightarrow \mu$.
2. For each bounded continuous $f\colon \mathbb{R} \to \mathbb{R}$, $\int f\, d\mu_n \to \int f\, d\mu$.
3. If $\mu(\partial A) = 0$, then $\mu_n(A) \to \mu(A)$.

Proof. We prove (1) $\iff$ (2) and (1) $\iff$ (3).

(1) $\implies$ (2): Assume $\mu_n \Rightarrow \mu$. Let $Y_n \sim \mu_n$ and $Y \sim \mu$ be random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that $Y_n \to Y$ pointwise, the existence of such random variables being guaranteed by Skorohod's theorem. Let $f\colon \mathbb{R} \to \mathbb{R}$ be a bounded continuous function. Then $f \circ Y_n \to f \circ Y$ pointwise, and so changing variables using Y_n, Y and invoking the bounded convergence theorem yields

$$\int f\, d\mu_n = \int f \circ Y_n\, d\mathbb{P} \to \int f \circ Y\, d\mathbb{P} = \int f\, d\mu.$$

(2) $\implies$ (1): For each n, let F_n be the distribution function of μ_n, and let F be the distribution function of μ. For $x < y$, define f by

$$f(t) = \begin{cases} 1 & \text{if } t < x \\ \frac{y-t}{y-x} & \text{if } x \le t \le y \\ 0 & \text{if } y < t \end{cases}$$

Note that f is continuous and bounded. For each n, $F_n(x) \le \int f\, d\mu_n$, and $\int f\, d\mu \le F(y)$, so by (2) we have $\limsup_{n\to\infty} F_n(x) \le F(y)$. and so because $y > x$ was arbitrary and F is right continuous, in fact $\limsup_{n\to\infty} F_n(x) \le F(x)$. A similar argument gives that $F(u) \le \liminf_{n\to\infty} F_n(x)$ for $u < x$, and so $\sup_{u<x} F(u) \le \liminf_{n\to\infty} F_n(x)$. If F is continuous at x, then $\sup_{u<x} F(u) = F(x)$, and so we obtain $F_n(x) \to F(x)$, which gives $\mu_n \Rightarrow \mu$.

(1) $\iff$ (3): The proof of (1) $\implies$ (2) goes through if the hypothesis that f is continuous is weakened to the hypothesis that f is measurable and the discontinuity set of f has μ-measure zero: merely weaken $f \circ Y_n \to f \circ Y$ pointwise to convergence almost surely (with respect to μ), and the rest of the proof is exactly the same. In particular, if $A \subseteq \mathbb{R}$ is measurable and $f = \mathbb{1}_A$, then f is bounded and has discontinuity set ∂A, so if $\mu(\partial A) = 0$, then $\int \mathbb{1}_A\, d\mu_n \to \int \mathbb{1}_A\, d\mu$, which is to say, $\mu_n(A) \to \mu(A)$. The converse is an immediate consequence of the definition of weak convergence and the fact that $\partial(-\infty, x] = \{x\}$. □

Rather than formalize this as a single result, we formalized various implications. First, we have the fact that $\mu_n \Rightarrow \mu$ implies $\int f\, d\mu_n \to \int f\, d\mu$ for f measurable and bounded with a discontinuity set of μ-measure zero.

```
theorem weak_conv_imp_bdd_ae_continuous_conv:
  fixes
    M_seq :: "nat ⇒ real measure" and
    M :: "real measure" and
    f :: "real ⇒ 'a::{banach, second_countable_topology}"
  assumes
    distr_M_seq: "⋀n. real_distribution (M_seq n)" and
    distr_M: "real_distribution M" and
    wcM: "weak_conv_m M_seq M" and
    discont_null: "M ({x. ¬ isCont f x}) = 0" and
    f_bdd: "⋀x. norm (f x) ≤ B" and
    [measurable]: "f ∈ borel_measurable borel"
  shows
    "(λ n. integral^L (M_seq n) f) ----> integral^L M f"
```

This of course immediately gives (1) $\Longrightarrow$ (2) from the informal theorem. The annotation `[measurable]` indicates that this fact should be used by a specialized tool for checking measurability of sets and functions, called `measurable` and implemented by Hölzl.

```
theorem weak_conv_imp_integral_bdd_continuous_conv:
  fixes
    M_seq :: "nat ⇒ real measure" and
    M :: "real measure" and
    f :: "real ⇒ 'a::{banach, second_countable_topology}"
  assumes
    "⋀n. real_distribution (M_seq n)" and
    "real_distribution M" and
    "weak_conv_m M_seq M" and
    "⋀x. isCont f x" and
    "⋀x. norm (f x) ≤ B"
  shows
    "(λ n. integral^L (M_seq n) f) ----> integral^L M f"
using assms apply (intro weak_conv_imp_bdd_ae_continuous_conv, auto)
  apply (rule borel_measurable_continuous_on1)
by (rule continuous_at_imp_continuous_on, auto)
```

As in the informal treatment, (1) $\Longrightarrow$ (3) can now be obtained in a straightforward manner.

```
theorem weak_conv_imp_continuity_set_conv:
  fixes
    M_seq :: "nat ⇒ real measure" and
    M :: "real measure" and
    f :: "real ⇒ real"
  assumes
    "⋀n. real_distribution (M_seq n)" "real_distribution M" and
    "weak_conv_m M_seq M" and
```

```
    [measurable]: "A ∈ sets borel" and
    "M (frontier A) = 0"
  shows
    "(λ n. (measure (M_seq n) A)) ----> measure M A"
proof -
  interpret M: real_distribution M by fact
  interpret M_seq: real_distribution "M_seq n" for n by fact

  have "(λn. (∫ x. indicator A x ∂M_seq n) :: real) ----> (∫ x. indicator
A x ∂M)"
    by (intro weak_conv_imp_bdd_ae_continuous_conv[where B=1])
       (auto intro: assms simp: isCont_indicator)
  then show ?thesis
    by simp
qed
```

The converse is also easily obtained, as in the informal development.

```
theorem continuity_set_conv_imp_weak_conv:
  fixes
    M_seq :: "nat ⇒ real measure" and
    M :: "real measure" and
    f :: "real ⇒ real"
  assumes
    real_dist_Mn [simp]: "⋀n. real_distribution (M_seq n)" and
    real_dist_M [simp]: "real_distribution M" and
    *: "⋀A. A ∈ sets borel ⟹ M (frontier A) = 0 ⟹
        (λ n. (measure (M_seq n) A)) ----> measure M A"
  shows
    "weak_conv_m M_seq M"
proof -
  interpret real_distribution M by simp
  show ?thesis
   unfolding weak_conv_m_def weak_conv_def cdf_def2 apply auto
   by (rule *, auto simp add: frontier_real_Iic isCont_cdf emeasure_eq_measure)
qed
```

The informal proof of (2) $\Longrightarrow$ (1) uses an approximation of step functions by continuous functions, specifically for $x < y$ we defined

$$f(t) = \begin{cases} 1 & \text{if } t < x \\ \frac{y-t}{y-x} & \text{if } x \leq t \leq y \\ 0 & \text{if } y < t \end{cases}.$$

This is formalized in Isabelle by a general definition which is available to all further development.

```
definition
  cts_step :: "real ⇒ real ⇒ real ⇒ real"
where
  "cts_step a b x ≡
    if x ≤ a then 1
    else (if x ≥ b then 0 else (b - x) / (b - a))"
```

The fact that these functions are in fact uniformly continuous is included as a remark after the proof of the portmanteau theorem on p. 335 of [6], and is formalized in a natural manner:

```
lemma cts_step_uniformly_continuous:
  fixes a b
  assumes [arith]: "a < b"
  shows "uniformly_continuous_on UNIV (cts_step a b)"
unfolding uniformly_continuous_on_def
proof (clarsimp)
  fix e :: real
  assume [arith]: "0 < e"
  let ?d = "min (e * (b - a)) (b - a)"
  have "?d > 0" by (auto simp add: field_simps)
  {
    fix x x'
    assume 1: "|x' - x| < e * (b - a)" and 2: "|x' - x| < b - a" and "x ≤
x'"
    hence "|cts_step a b x' - cts_step a b x| < e"
      unfolding cts_step_def apply auto
      apply (auto simp add: field_simps)[2]
      by (subst diff_divide_distrib [symmetric], simp add: field_simps)
  } note * = this
  have "∀x x'. dist x' x < ?d ⟶ dist (cts_step a b x') (cts_step a b x)
< e"
  proof (clarsimp simp add: dist_real_def)
    fix x x'
    assume "|x' - x| < e * (b - a)" and "|x' - x| < b - a"
    thus "|cts_step a b x' - cts_step a b x| < e"
      apply (case_tac "x ≤ x'")
      apply (rule *, auto)
      apply (subst abs_minus_commute)
      by (rule *, auto)
  qed
  with `?d > 0` show
    "∃d > 0. ∀x x'. dist x' x < d ⟶ dist (cts_step a b x') (cts_step a
b x) < e"
    by blast
qed
```

To formalize (2) $\Longrightarrow$ (1) from the informal proof of the portmanteau theorem, we need that the continuous step functions are integrable and satisfy

$$F_\mu(x) \leq \int f_{xy}\, d\mu \leq F_\mu(y),$$

where F_μ is the distribution function of μ and f_{xy} is the continuous step from x to y (we assume $x < y$).

```
lemma (in real_distribution) integrable_cts_step: "a < b ⟹ integrable M
(cts_step a b)"
apply (rule integrable_const_bound [of _ 1])
  apply (force simp add: cts_step_def)
  apply (rule measurable_finite_borel)
  apply (rule borel_measurable_continuous_on1)
  apply (rule uniformly_continuous_imp_continuous)
by (rule cts_step_uniformly_continuous)
```

```
lemma (in real_distribution) cdf_cts_step:
  fixes
    x y :: real
  assumes
    "x < y"
  shows
    "cdf M x ≤ integral^L M (cts_step x y)" and
    "integral^L M (cts_step x y) ≤ cdf M y"
unfolding cdf_def
proof -
  have "prob {..x} = integral^L M (indicator {..x})"
    by simp
  thus "prob {..x} ≤ expectation (cts_step x y)"
    apply (elim ssubst)
    apply (rule integral_mono)
    apply simp
    apply (auto intro!: integrable_cts_step assms) []
    apply (auto simp add: cts_step_def indicator_def field_simps)
    done
next
  have "prob {..y} = integral^L M (indicator {..y})"
    by simp
  thus "expectation (cts_step x y) ≤ prob {..y}"
    apply (elim ssubst)
    apply (rule integral_mono)
    apply (rule integrable_cts_step, rule assms)
    unfolding cts_step_def indicator_def using `x < y`
    by (auto simp add: field_simps)
qed
```

The special case of (2) $\implies$ (1) for continuous step functions can now be formalized; we omit the proof.

```
theorem integral_cts_step_conv_imp_weak_conv:
  fixes
    M_seq :: "nat ⇒ real measure" and
    M :: "real measure"
  assumes
    distr_M_seq: "⋀n. real_distribution (M_seq n)" and
    distr_M: "real_distribution M" and
```

```
    integral_conv: "⋀x y. x < y ⟹
         (λn. integral^L (M_seq n) (cts_step x y)) ----> integral^L M (cts_step
x y)"
  shows
    "weak_conv_m M_seq M"
```

This can then be weakened to something slightly stronger than (2) $\implies$ (1):

```
theorem integral_bdd_continuous_conv_imp_weak_conv:
  fixes
    M_seq :: "nat ⇒ real measure" and
    M :: "real measure"
  assumes
    "⋀n. real_distribution (M_seq n)" and
    "real_distribution M" and
    "⋀f B. (⋀x. isCont f x) ⟹ (⋀x. abs (f x) ≤ B) ⟹
         (λn. integral^L (M_seq n) f::real) ----> integral^L M f"
  shows
    "weak_conv_m M_seq M"
apply (rule integral_cts_step_conv_imp_weak_conv [OF assms])
  apply (rule continuous_on_interior)
  apply (rule uniformly_continuous_imp_continuous)
  apply (rule cts_step_uniformly_continuous, auto)
  apply (subgoal_tac "abs(cts_step x y xa) ≤ 1")
  apply assumption
unfolding cts_step_def by auto
```

5.3 Helly Selection Theorem and Tightness

As described in section 3, for the proof of the Lévy continuity theorem (which allows us to study weak convergence through characteristic functions), an analogue of the Bolzano-Weierstrass theorem in the space of probability measures with the topology of weak convergence is needed. In order to prove this, we need the Helly selection theorem, which is itself an analogue of the Bolzano-Weierstrass theorem, and a result of independent interest in functional analysis.

5.3.1 Helly Selection Theorem

We begin with some definitions.

Definition 5.9. A function $F\colon \mathbb{R} \to \mathbb{R}$ is a *subdistribution function* iff F is non-decreasing, right-continuous, and satisfies $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) \leq 1$.

Thus a subdistribution function is the natural analogue of a distribution function in the case of a measure which has total mass at most 1 (a *subprobability measure*). The correspondence between subdistribution functions and subprobability measures can be proved in an analogous manner to the proof of

the same correspondence between distribution functions and probability measures.

We now come to the subdistribution function analogue of weak convergence:

Definition 5.10. A sequence $\langle F_n \mid n \in \mathbb{N} \rangle$ of subdistribution functions *converges vaguely* to a subdistribution function F iff $\lim_{n\to\infty} F_n(x) = F(x)$ for all $x \in \mathbb{R}$ such that F is continuous at x. This is denoted $F_n \Rightarrow F$, just as with weak convergence. The two notions can be easily seen to coincide in case the functions F_n, F are all distribution functions.

These definitions permit a natural statement of the Helly selection theorem.

Theorem 5.11. Let $\langle F_n \mid n \in \mathbb{N} \rangle$ be a sequence of distribution functions. Then there exists a subsequence $\langle F_{n_k} \mid n \in \mathbb{N} \rangle$ and a subdistribution function F such that $F_{n_k} \Rightarrow F$.

The proof requires an application of the diagonal method.

Lemma 5.12. Suppose, for each $n \in \mathbb{N}$, that $\langle a_{n,k} \mid k \in \mathbb{N} \rangle$ is a bounded sequence of real numbers. Then there is a uniform subsequence (a single increasing sequence $\langle n_k \mid k \in \mathbb{N} \rangle$ of natural numbers) such that $\lim_{k\to\infty} a_{i,n_k}$ exists for every $i \in \mathbb{N}$.

Proof. Invoking the Bolzano-Weierstrass theorem, select $\langle n_{0,k} \mid k \in \mathbb{N} \rangle$ such that $\langle a_{0,n_{0,k}} \mid k \in \mathbb{N} \rangle$ converges. Now inductively choose $\langle n_{i+1,k} \mid k \in \mathbb{N} \rangle$ to be a subsequence of $\langle n_{i,k} \mid k \in \mathbb{N} \rangle$ such that $\langle a_{i+1,k} \mid k \in \mathbb{N} \rangle$ converges. It is now easy to see that the subsequence $n_k = n_{k,k}$ is such that $\lim_{k\to\infty} a_{i,n_k}$ exists for every $i \in \mathbb{N}$, as required. □

We were fortunate that Fabian Immler at Technische Universität München had already formalized a diagonal subsequence library as part of his effort to formalize the theory of differential equations.

We are finally ready for an informal presentation of the proof of the Helly selection theorem.

Helly selection theorem. Using the diagonal method 5.12, obtain a subsequence $\langle n_k \mid k \in \mathbb{N} \rangle$ such that $\lim_{k\to\infty} F_{n_k}(q)$ exists for each rational q. Call the value of this limit $G(q)$. Define $F(x) = \inf\{G(q) \mid q \in \mathbb{Q},\ x < q\}$ and note F is nondecreasing because each F_n is.

Let $x \in \mathbb{R}$ and $\varepsilon > 0$. By the definition of F there is $q \in \mathbb{Q}$, $q > x$ such that $G(q) < F(x) + \varepsilon$. F is right-continuous because for $x \le y < q$, $F(y) \le G(q) < F(x) + \varepsilon$.

Suppose now that F is continuous at x. Choose $y < x$ such that $F(x) < F(y)+\varepsilon$, and rationals p, q such that $y < p < x < q$ and $G(p) < F(x)+\varepsilon$. Thus it follows that $F(x)-\varepsilon < G(p) \le G(q) < F(x)+\varepsilon$ and that $F_n(p) \le F_n(x) \le F_n(q)$ for each $n \in \mathbb{N}$. Consequently we have that

$$\left|\liminf_{k\to\infty} F_{n_k}(x) - F(x)\right|, \left|\limsup_{k\to\infty} F_{n_k}(x) - F(x)\right| < \varepsilon,$$

and hence $\lim_{k\to\infty} F_{n_k}(x) = F(x)$, which establishes that $F_{n_k} \Rightarrow F$. □

The formal statement of the Helly selection theorem follows; we omit the very long proof, but note that the first hurdle was to establish a bijection between the naturals and the rationals! Fortunately this could be handled with library support for reasoning about countable sets; the bijection did not need to be constructed explicitly. It should be noted that we did not formalize the definitions of subdistribution functions and vague convergence, as these are not needed anywhere else in our formal development. The predicate `rcont_inc` indicates a function which is nondecreasing and right continuous.

```
theorem Helly_selection:
  fixes f :: "nat ⇒ real ⇒ real"
  assumes rcont_inc: "⋀n. rcont_inc (f n)"
  and unif_bdd: "∀n x. |f n x| ≤ M"
  shows "∃s. subseq s ∧ (∃F. rcont_inc F ∧ (∀x. |F x| ≤ M) ∧
    (∀x.  continuous (at x) F ⟶ (λn. f (s n) x) ----> F x))"
```

5.3.2 Tightness of Sequences of Measures

The applications of Helly's theorem in which we are interested concern tightness of sequences of measures; this is an analogue of boundedness of ordinary sequences, and prevents any mass in a sequence $\langle \mu_n \mid n \in \mathbb{N} \rangle$ from "escaping to infinity" in the (weak) limit. An example of a sequence of probability measure which is not tight is $\langle \mu_n \mid n \in \mathbb{N} \rangle$, where for each $n \in \mathbb{N}$, μ_n is a unit mass at n (it is intuitively clear that all the mass of this sequence "escapes to infinity.")

Definition 5.13. Let $\langle \mu_n \mid n \in \mathbb{N} \rangle$ be a sequence of Borel probability measures on $\mathbb{R}$. This sequence is called *tight* iff for each $\varepsilon > 0$ there exist $a < b$ such that $\mu_n(a, b] < 1 - \varepsilon$ for every $n \in \mathbb{N}$.

It should be clear how the tightness condition keeps mass from getting too far away and ultimately escaping to infinity. In terms of distribution functions, the tightness condition is equivalent to the requirement that for each $\varepsilon > 0$ there exist $x, y \in \mathbb{R}$ such that $F_n(x) < \varepsilon$ and $F_n(y) > 1 - \varepsilon$ for every $n \in \mathbb{N}$.

We can now state the promised analogue of the Bolzano-Weierstrass theorem for tight sequences of probability measures.

Theorem 5.14. A sequence $\langle \mu_n \mid n \in \mathbb{N} \rangle$ of probability measures is tight if and only if for every subsequence $\langle \mu_{n_k} \mid k \in \mathbb{N} \rangle$ there exists a subsubsequence $\langle \mu_{n_{k_j}} \mid j \in \mathbb{N} \rangle$ which converges weakly to some probability measure μ.

Proof. ($\Longrightarrow$): Suppose for contradiction that $\langle \mu_n \mid n \in \mathbb{N} \rangle$ is not tight. Choose $\varepsilon > 0$ such that for every choice of $a < b$, $\mu_n(a, b] \leq 1 - \varepsilon$ for some n. Using this, choose n_k for each $k \in \mathbb{N}$ such that $\mu_{n_k}(-k, k] \leq 1 - \varepsilon$. Let $\langle \mu_{n_{k_j}} \mid j \in \mathbb{N} \rangle$ be a subsequence of $\langle \mu_{n_k} \mid k \in \mathbb{N} \rangle$, and suppose for contradiction that $\mu_{n_{k_j}} \Rightarrow \mu$ for some probability measure μ. Choose $a < b$ nonatoms such that $\mu(a, b] > 1 - \varepsilon$ (which is possible as there are only countably many atoms). There exists j_0 such that $(a, b] \subseteq (-k_j, k_j]$ for $j \geq j_0$, and thus we have

$$1 - \varepsilon \geq \mu_{n_{k_j}}(-k_j, k_j] \geq \mu_{n_{k_j}}(a, b] \to \mu(a, b]$$

as $j \to \infty$. This implies that $\mu(a,b] \leq 1-\varepsilon$, contradicting the earlier established fact that $\mu(a,b] > 1-\varepsilon$.

($\Longleftarrow$): Let $\langle F_{n_k} \mid k \in \mathbb{N}\rangle$ be the sequence of distribution functions corresponding to $\langle \mu_{n_k} \mid k \in \mathbb{N}\rangle$. Apply the Helly selection theorem to obtain a subsequence $\langle F_{n_{k_j}} \mid j \in \mathbb{N}\rangle$ which converges vaguely to a subdistribution function F. Let μ be the subprobability measure corresponding to F. For $\varepsilon > 0$ use tightness to obtain $a < b$ such that $\mu_n(a,b] > 1-\varepsilon$ for every $n \in \mathbb{N}$. Since F has just countably many discontinuity points, a, b may be chosen to be continuity points of F. Thus $\mu(a,b] > 1-\varepsilon$, and since $\varepsilon > 0$ was arbitrary we see that in fact μ is a probability measure, and hence the vague convergence $F_{n_{k_j}} \Rightarrow F$ is in fact weak convergence, which is to say $\mu_{n_{k_j}} \Rightarrow \mu$ as desired. □

The corresponding formalized proof is again very long, so we provide just the statement (and the formalized definition of tightness).

```
definition tight :: "(nat ⇒ real measure) ⇒ bool"
where "tight μ ≡ (∀n. real_distribution (μ n)) ∧ (∀(ε::real)>0. ∃a b::real.
a < b ∧ (∀n. measure (μ n) {a<..b} > 1 - ε))"

theorem tight_iff_convergent_subsubsequence:
  fixes μ
  assumes "⋀n. real_distribution (μ n)"
  shows "tight μ = (∀s. subseq s ⟶ (∃r. ∃M.  subseq r ∧ real_distribution
M ∧ weak_conv_m (μ ∘ s ∘ r) M))"
```

A corollary of this result will also be useful.

Corollary 5.15. If $\langle \mu_n \mid n \in \mathbb{N}\rangle$ is a tight sequence of probability measures such that each subsequence which has a weak limit in fact has the probability measure μ as its weak limit, then $\mu_n \Rightarrow \mu$.

Proof. By the preceding theorem, for every subsequence $\langle \mu_{n_k} \mid k \in \mathbb{N}\rangle$ there is a subsubsequence $\langle \mu_{n_{k_j}} \mid j \in \mathbb{N}\rangle$ which converges weakly; by hypothesis the weak limit must be μ.

Suppose now for contradiction that μ is not the weak limit of $\langle \mu_n \mid n \in \mathbb{N}\rangle$. Then there exists $x \in \mathbb{R}$ such that $\mu\{x\} = 0$ but $\langle \mu_n(-\infty, x] \mid n \in \mathbb{N}\rangle$ does not converge to $\mu(-\infty, x]$. Thus for some $\varepsilon > 0$ there are infinitely many $i \in \mathbb{N}$ such that $|\mu_i(-\infty,x] - \mu(-\infty,x]| > \varepsilon$. Letting $\langle n_k \mid k \in \mathbb{N}\rangle$ enumerate these i's gives a subsequence $\langle \mu_{n_k} \mid k \in \mathbb{N}\rangle$ no subsequence of which can converge weakly to μ, which contradicts what was established in the first paragraph. □

This is formalized as follows.

```
corollary tight_subseq_weak_converge:
  fixes μ :: "nat ⇒ real measure" and M :: "real measure"
  assumes "⋀n. real_distribution (μ n)" "real_distribution M" and tight:
"tight μ" and
    subseq: "⋀s ν. subseq s ⟹ real_distribution ν ⟹ weak_conv_m (μ ∘
s) ν ⟹ weak_conv_m (μ ∘ s) M"
  shows "weak_conv_m μ M"
```

```
proof (rule ccontr)
  from tight tight_iff_convergent_subsubsequence
  have subsubseq: "∀s. subseq s ⟶ (∃r M. subseq r ∧ real_distribution
M ∧ weak_conv_m (μ ∘ s ∘ r) M)"
    using assms by simp
  {
    fix s assume s: "subseq s"
    with subsubseq subseq have "∃r M. subseq r ∧ real_distribution M ∧ weak_conv_m
(μ ∘ s ∘ r) M"
      by simp
    then guess r .. note r = this
    then guess ν .. note ν = this
    hence subsubseq_conv: "weak_conv_m (μ ∘ (s ∘ r)) ν" by (auto simp add:
o_assoc)
    from s r have sr: "subseq (s ∘ r)" using subseq_o by auto
    with subsubseq_conv subseq ν have "weak_conv_m (μ ∘ (s ∘ r)) M" by auto
    with r have "∃r. subseq r ∧ weak_conv_m (μ ∘ (s ∘ r)) M" by auto
  }
  hence *: "⋀s. subseq s ⟹ ∃r. subseq r ∧ weak_conv_m (μ ∘ (s ∘ r)) M"
by auto
  def f ≡ "λn. cdf (μ n)"
  def F ≡ "cdf M"
  assume CH: "¬ weak_conv_m μ M"
  hence "∃x. isCont F x ∧ ¬((λn. f n x) ----> F x)"
    unfolding weak_conv_m_def weak_conv_def f_def F_def by auto
  then guess x .. note x = this
  hence "∃ε>0. ∃s. subseq s ∧ (∀n. |f (s n) x - F x| ≥ ε)"
    apply (subst (asm) tendsto_iff, auto simp add: not_less)
    apply (subst (asm) eventually_sequentially, auto)
    unfolding dist_real_def apply (simp add: not_less)
    apply (subst subseq_Suc_iff)
    apply (rule_tac x = e in exI, safe)
    proof -
      fix e assume e: "0 < e" "∀N. ∃n≥N. e ≤ |f n x - F x|"
      then obtain n where n: "⋀N. N ≤ n N" "⋀N. e ≤ |f (n N) x - F x|"
by metis
      def s ≡ "rec_nat (n 0) (λ_ i. n (Suc i))"
      then have s[simp]: "s 0 = n 0" "⋀i. s (Suc i) = n (Suc (s i))"
        by simp_all
      { fix i have "s i < s (Suc i)"
          using n(1)[of "Suc (s i)"] n(2)[of 0]  by simp_all }
      moreover { fix i have "e ≤ |f (s i) x - F x|"
        by (cases i) (simp_all add: n) }
      ultimately show "∃s. (∀n. s n < s (Suc n)) ∧ (∀n. e ≤ |f (s n) x
- F x|)"
        by metis
    qed
  then obtain ε s where ε: "ε > 0" and s: "subseq s ∧ (∀n. |f (s n) x -
F x| ≥ ε)" by auto
  hence "⋀r. subseq r ⟹ ¬weak_conv_m (μ ∘ s ∘ r) M"
```

```
    apply (unfold weak_conv_m_def weak_conv_def, auto)
    apply (rule_tac x = x in exI)
    apply (subst tendsto_iff)
    unfolding dist_real_def apply (subst eventually_sequentially)
    using x unfolding F_def apply auto
    apply (subst not_less)
    apply (subgoal_tac "(λn. cdf (μ (s (r n))) x) = (λn. f (s (r n)) x)")
    apply (erule ssubst)
    apply (rule_tac x = ε in exI)
    unfolding f_def by auto
  thus False using subseq * by (metis fun.map_comp s)
qed
```

5.4 Integration

Improving the integration library of Isabelle was one of the primary motivations of the central limit theorem formalization project. A complete rewrite of the library, implementing a change from Lebesgue to Bochner integration, was completed by Hölzl in response to our feedback concerning the usability of the integration library. We begin with a general discussion of the problems and tradeoffs in formalizing integration, and then describe how we used the integration library in the computation of the limit of the sine integral function at infinity.

5.4.1 General Remarks on Formalizing Integration

The proofs of the Lévy inversion and continuity theorems required formal work with integrals, and brought a number of design issues into focus. First, one frequently wishes to integrate a function only over a subset A of the space on which it is defined rather than over the whole space. This can be handled in two ways: The integral over the whole space can be taken as primitive, with the integral of a function f over a set A defined by $\int_A f\,d\mu = \int f \mathbb{1}_A\,d\mu$, or the integral over a set can be taken as primitive with the integral of a function f over the whole space being defined by $\int f\,d\mu = \int_X f\,d\mu$. There is some small advantage to taking the integral over a set as primitive, because this avoids failures of pattern-matching when automated simplifications move indicator functions around or unfold the definition, but this advantage is not major as far as we can tell and could easily be outweighed by complications in proving fundamental lemmata when the domain of integration is an additional parameter. In particular, the Isabelle Bochner integration library takes integration over the entire space as primitive. In any case it is certainly useful to have notation for integration over a set.

One particular type of set occurs particularly frequently as the domain of integration with respect to Lebesgue measure on $\mathbb{R}$, namely a closed interval. In calculus the integral (with respect to Lebesgue meausre) of a function f over a closed interval $[a, b]$ $(a \leq b)$ is denoted $\int_a^b f(x)\,dx$, and it is convenient to have a similar notation for integrals over intervals in Isabelle. A number of

design issues immediately present themselves: What should be the types of a and b? One might assume these should be reals, but frequently integrals are computed over unbounded intervals ($\int_0^\infty e^{-x}\,dx$, $\int_{-\infty}^\infty e^{-x^2}\,dx$), and so there is some advantage to taking a and b to be extended reals to avoid the need for separate lemmata for integrals of the form $\int_a^\infty$, $\int_{-\infty}^b$ and $\int_{-\infty}^\infty$ (this last form being a notational variant of $\int$). However, this advantage is achieved at a cost, because it entails annoying casts between reals and extended reals that we found in practice frequently prevent automated tools from proving apparently obvious facts (which they do in fact obtain when the endpoints are taken to be of type real). This is largely because the extended reals are not as algebraically well-behaved as the reals (they are not a field, for example).

As noted in the preceding paragraph, for interval integrals with respect to Lebesgue measure the interval is generally assumed to be closed (so any continuous function defined on it is uniformly continuous, and other nice properties hold); since Lebesgue measure is continuous, it makes no difference to the value of the integral whether the endpoints are included. However, for general measures this does make a difference if one of the endpoints is an atom, and in particular the interval partition formula, valid for f integrable with respect to Lebesgue measure and $a \le b \le c$:

$$\int_a^b f(x)\,dx + \int_b^c f(x)\,dx = \int_a^c f(x)\,dx$$

fails in general if $\int_a^b$ is defined as $\int_{[a,b]}$. What holds instead for arbitrary measures μ (and functions f integrable over $[a,c]$ with respect to μ) is that

$$\int_{[a,b]} f\,d\mu + \int_{[b,c]} f\,d\mu = \int_{[a,c]} f\,d\mu + f(b)\mu\{b\}.$$

Intuitively this is because the mass of the point b is counted twice. This is obviously less convenient than the ordinary interval partition formula, especially for partitions into large numbers of pieces. The problem can be fixed by using a half-open interval such as $(a,b]$ to ensure pieces of an interval partition are disjoint; such a solution preserves the equality $\int_a^b f\,d\mu = \int_{[a,b]} f\,d\mu$ for continuous measures, and satisfies the intuitive partition formula

$$\int_a^b f\,d\mu + \int_b^c f\,d\mu = \int_a^c f\,d\mu$$

for all a, b, c with $a < b < c$.

A further constellation of design issues arises from considering what should happen if $b < a$ in $\int_a^b f(x)\,dx$. The natural option is to take $\int_a^b f(x)\,dx = -\int_b^a f(x)\,dx$ in this case, but this would require introducing a case split in the definition of $\int_a^b f(x)\,dx$ (e.g. $\int_a^b f(x)\,dx = \int_{[a,b]} f(x)\,dx$ if $a \le b$, otherwise $\int_a^b f(x)\,dx = -\int_{[b,a]} f(x)\,dx$), which then causes headaches for formalization

both for human users and automated tools. $\int_a^b f(x)\,dx$ could also be taken simply to be notation for $\int_{[a,b]} f(x)\,dx$, in which case $b < a$ implies $[a,b] = \emptyset$ and so the integral is zero, but this makes it hard to state results such as the fundamental theorem of calculus or integration by substitution in a natural manner.

Another set of issues concern integrability. Not all functions have integrals, and the notation $\int f\,d\mu$ only makes sense if f is integrable (over X). Isabelle requires that all functions be total, so $\int f\,d\mu$ necessarily has a value no matter what f is. If f is not integrable, the value which $\int f\,d\mu$ receives is arbitrary. A proof that $\int f\,d\mu = c$ is useless unless f is known to be integrable (for otherwise it could be that c just happens to be the default value in the case of integrating f). These are general problems concerning the representation of partial functions, and we shall pause to briefly consider them in full generality.

A partial function with domain X and codomain Y can be thought of as a relation $R \subseteq X \times Y$ such that for every $x \in X$ and $y, z \in Y$, xRy and xRz implies $y = z$ (a [total] function corresponds to such a relation where in addition for every $x \in X$ there exists $y \in Y$ such that xRy, though the higher-order logic used by Isabelle treats functions somewhat differently [not as relations]). Let y^* be some distinguished element of the type of elements of Y, and consider the function $f\colon X \to Y \cup \{y^*\}$ where $f(x) = y$ if there exists $y \in Y$ such that xRy, and otherwise $f(x) = y^*$. The value y^* should be thought of as hidden; the fact that $f(x) = y^*$ if there does not exist $y \in Y$ such that xRy should not be exploited in proofs. In this case the conclusion that $f(x) = y$ is useless unless it is known that there exists $y \in Y$ such that xRy. Since the existence of $y \in Y$ such that xRy means intuitively that f is "defined" at x; let us denote it by Dx. Then xRy is equivalent to Dx and $f(x) = y$. Since $f(x) = y$ is useless without knowing Dx, the conclusions of computations of f for various arguments should either be stated in terms of R or have the auxilliary conclusion that Dx for each argument x of f considered in the computation. xRy is a robust conclusion and functions well as the fundamental notion in terms of which f and D are defined, while it is convenient to have f both to allow statement of results in a manner more similar to mathematical practice, and to allow more convenient computation with values of the partial function (e.g. $f(x_1) + f(x_2)$ is easier to work with than `(THE` y`.` x_1Ry`) + (THE` y`.` x_2Ry`)` or something like that).

The Isabelle libraries we used during the formalization of the central limit theorem employed an integral operator and an integrability predicate; there was no instantiation of the relation R from the preceding paragraph. This could have worked had every use of the integral operator been accompanied by a corresponding proof of integrability, but unfortunately that was not the case, and occasionally we needed to either modify a library proof to obtain integrability as well as the value of an integral, or repeat long arguments from library lemmata with trivial modifications so as to obtain integrability. This is striking, because other partial functions such as limits and derivatives had already been implemented with relations such as `has_derivative` and what one may think of as `has_limit` (though the actual definition of limits uses two filters

as described in section 4.3). The problem of partial functions was not new, but it required experience to determine the best method of implementing them in Isabelle. Another consideration in the case of integration was that when working with concrete functions, one often wishes to prove integrability by computing the integral (e.g. the integral of e^{-x} over $[0, \infty)$ is 1), and this is generally very inconvenient to do when integrability must be verified separately. A better plan in such cases is to have an instantiation `has_integral` of the relation R from the preceding paragraph, and obtain integrability by proving that f `has_integral` c for appropriate c when working with specific integrals. Thus it is convenient to have not only an integral operator and an integrability predicate, but also a `has_integral` relation. In some sense it does not matter which is taken as fundamental, but as noted in the preceding paragraph it seems more natural to take `has_integral` as fundamental.

The change from a fundamental integral operator and integrability predicate to a fundamental `has_integral` relation was accomplished by Hölzl when reimplementing the integration library using the Bochner integral. This change was motivated largely by a desire to provide a unified framework for vector-valued integrals, and was of direct importance to the central limit theorem formalization because the complex-valued integrals arising from characteristic functions can be handled as Bochner integrals.

There is also the question of how to handle improper integrals; for example, the sinc function ($x^{-1} \sin x$ away from zero, and 1 at zero) is not integrable over $[0, \infty)$, and yet

$$\lim_{t \to \infty} \int_0^t \operatorname{sinc} x \, dx = \frac{\pi}{2}.$$

Often this is written simply as $\int_0^\infty \operatorname{sinc} x \, dx = \pi/2$, perhaps with a warning that notation is being abused (see note regarding this limit in [6], p. 223). Improper integrals also occur over finite intervals, for example

$$\lim_{t \to 0} \int_t^1 \frac{(-1)^{\lfloor x \rfloor + 1}}{\lfloor x \rfloor} \mathbb{1}_{(0,1]} \, dx = \ln 2,$$

but the integrand is not integrable over $(0, 1]$ because the harmonic series diverges. It would be possible to implement the integral over an interval to include improper integrals, as is standard practice in calculus texts, by including a limit in the definition of such an integral. However, this seems to carry with it too many disadvantages, simply because of the complication that a hidden limit introduces; it is inconvenient for users to constantly need to eliminate the limit whenever they use integrals, and difficult to set up automated tools to deal with it effectively.

5.4.2 The Sine Integral Function

First a note on theorems fundamental to working with integrals: The fundamental theorem of calculus was present in some form when we started, but we found it useful to extend it significantly. We also proved lemmata concerning

integration by substitution (change of variables), and defined integrals over sets and intervals. During the course of the formalization Hölzl improved our fundamental lemmata and our definitions of integrals over sets and intervals and incorporated them into the integration library he was developing. We do not include pieces of that library in this paper, but the reader may readily find them on the Isabelle website (under HOL-Probability).

On p. 223 of [6], Billingsley notes that it is "an important analytic fact" that

$$\lim_{t\to\infty} \int_0^t \frac{\sin x}{x}\, dx = \frac{\pi}{2}.$$

Partly because of this, but mostly because this important analytic fact is used in the proof of the Lévy inversion theorem, we decided to formalize the proof of this limit.

The function $\frac{\sin x}{x}$ has a removable discontinuity (due to not being defined) at zero; the result of filling in that discontinuity (with the value 1) is called the function sinc: $\operatorname{sinc} x = \frac{\sin x}{x}$ if $x \neq 0$, $\operatorname{sinc} 0 = 1$. The indefinite integral of the sinc function, starting at 0, is called the *sine integral function*:

$$\operatorname{Si}(x) = \int_0^t \operatorname{sinc} x\, dx.$$

These definitions, and the basic properties of the sinc and Si functions, are formalized as expected. Note that `LINT` is ASCII notation for the Lebesgue integral, and `LBINT` is ASCII notation for the Lebesgue integral with respect to Lebesgue measure (otherwise known as Lebesgue-Borel measure).

```
abbreviation sinc :: "real ⇒ real" where
  "sinc ≡ (λx. if x = 0 then 1 else sin x / x)"

lemma sinc_at_0: "((λx. sin x / x::real) ---> 1) (at 0)"
using DERIV_sin [of 0] by (auto simp add: has_field_derivative_def field_has_derivative_at)

lemma isCont_sinc: "isCont sinc x"
apply (case_tac "x = 0", auto)
  apply (simp add: isCont_def)
  apply (subst LIM_equal [where g = "λx. sin x / x"], auto intro: sinc_at_0)
  apply (rule continuous_transform_at [where d = "abs x" and f = "λx. sin
x / x"])
  by (auto simp add: dist_real_def)

lemma continuous_on_sinc[continuous_intros]:
  "continuous_on S f ⟹ continuous_on S (λx. sinc (f x))"
using continuous_on_compose[of S f sinc, OF _ continuous_at_imp_continuous_on]
  by (auto simp: isCont_sinc)

lemma borel_measurable_sinc[measurable]: "sinc ∈ borel_measurable borel"
by (intro borel_measurable_continuous_on1 continuous_at_imp_continuous_on
ballI isCont_sinc)
```

```
lemma sinc_AE: "AE x in lborel. sin x / x = sinc x"
by (rule AE_I [where N = "{0}"], auto)

definition Si :: "real ⇒ real" where "Si t ≡ LBINT x=0..t. sin x / x"

lemma Si_alt_def : "Si t = LBINT x=0..t. sinc x"
apply (case_tac "0 ≤ t")
  unfolding Si_def apply (rule interval_lebesgue_integral_cong_AE, auto)
  apply (rule AE_I' [where N = "{0}"], auto)
  apply (subst (1 2) interval_integral_endpoints_reverse, simp)
  apply (rule interval_lebesgue_integral_cong_AE, auto)
by (rule AE_I' [where N = "{0}"], auto)
lemma sinc_neg [simp]: "sinc (- x) = sinc x"by auto

lemma Si_neg:
  fixes T :: real
  assumes "T ≥ 0"
  shows "Si (- T) = - Si T"
proof -
  have "LBINT x=ereal 0..T. -1 *R sinc (- x) = LBINT y= ereal (- 0)..ereal
(- T). sinc y"
    apply (rule interval_integral_substitution_finite [OF assms])
    by (auto intro: derivative_intros continuous_at_imp_continuous_on isCont_sinc)
  also have "(LBINT x=ereal 0..T. -1 *R sinc (- x)) = -(LBINT x=ereal 0..T.
sinc x)"
    by (subst sinc_neg) (simp_all add: interval_lebesgue_integral_uminus)
  finally have *: "-(LBINT x=ereal 0..T. sinc x) = LBINT y= ereal 0..ereal
(- T). sinc y"
    by simp
  show ?thesis
    using assms unfolding Si_alt_def
     apply (subst zero_ereal_def)+
     by (auto simp add: * [symmetric])
qed

lemma iSi_isCont: "isCont Si x"
proof -
  have "Si = (λt. LBINT x=ereal 0..ereal t. sinc x)"
    apply (rule ext, simp add: Si_def zero_ereal_def)
    apply (rule interval_integral_cong_AE)
    by (auto intro!: AE_I' [where N = "{0}"])
  thus ?thesis
    apply (elim ssubst)
    apply (rule DERIV_isCont)
    apply (subst has_field_derivative_within_open [symmetric,
      where s = "{(min (x - 1) (- 1))<..<(max 1 (x+1))}"], auto)
    apply (rule DERIV_subset [where s = "{(min (x - 2) (- 2))..(max 2 (x+2))}"])
    unfolding has_field_derivative_iff_has_vector_derivative
    apply (rule interval_integral_FTC2)
```

```
    by (auto intro!: continuous_on_sinc continuous_on_id)
qed
lemma borel_measurable_iSi[measurable]: "Si ∈ borel_measurable borel"
by (auto intro: iSi_isCont continuous_at_imp_continuous_on borel_measurable_continuous_on1)
```

The `abbreviation` keyword indicates a definition which is always to be expanded; normally Isabelle does not expand definitions unless explicitly told to do so in order to avoid the associated explosion in proof search possibilities. Whether a definition should be made using `abbreviation` (so it is conveniently always expanded) or `definition` is an important design decision affecting the usability of a library, and there can be many conflicting factors. For example, case splits are difficult for automated tools to work with, but in the case of the sinc function we chose an abbreviation because the function was not used in a sufficiently fundamental way to warrant developing a reasonably complete list of its basic properties, as is needed for a definition made with the `definition` keyword to function well (without requiring that it constantly be expanded).

Billingsely's proof that $\lim_{x\to\infty} \mathrm{Si}(x) = \frac{\pi}{2}$ uses Fubini's theorem, a result concerning product measures which allows integrals with respect to product measures to be computed as iterated integrals, in either order, under very general circumstances. The product of two measure spaces (X_1, Σ_1, μ_1) and (X_2, Σ_2, μ_2) has as σ-algebra the algebra $\sigma(\Sigma_1 \times \Sigma_2)$ generated by rectangles $A \times B$ where $A \in \Sigma_1$ and $B \in \Sigma_2$, and measure given by $(\mu_1 \times \mu_2)(A \times B) = \mu_1(A)\mu_2(B)$ for $A \in \Sigma_1$ and $B \in \Sigma_2$, extended to the entire σ-algebra $\sigma(\Sigma_1 \times \Sigma_2)$ by Carathéodory's theorem. For a formal statement of Fubini's theorem we refer the reader to any standard treatment of measure theory, such as that in [6]. The basics of product measures had already been formalized by Hölzl [18], including Fubini's theorem. We formalized the result displayed in figure 2 to facilitate the use of Fubini's theorem in proofs involving concrete integrals.

Let us now examine the informal computation of the limit of $\mathrm{Si}(x)$ at infinity, following Billingsley [6] as usual. The fundamental theorem of calculus immediately yields

$$\int_0^t e^{-ux} \sin x \, dx \frac{1}{1+u^2}[1 - e^{-ut}(u \sin t + \cos t)].$$

Taking $t \to \infty$, we see that

$$\int_0^t \left(\int_0^\infty |e^{-ux} \sin x| \, du \right) dx = \int_0^t x^{-1} |\sin x| \, dx \leq t,$$

so Fubini's theorem may be used in the integration of $e^{-ux}\sin x$ over $(0,t)\times(0,\infty)$:

$$\begin{aligned}\int_0^t \frac{\sin x}{x}\,dx &= \int_0^t \sin x\left(\int_0^\infty e^{-ux}\,du\right)dx\\ &= \int_0^\infty \left(\int_0^t e^{-ux}\sin x\,dx\right)du\\ &= \int_0^\infty \frac{du}{1+u^2} - \int_0^\infty \frac{e^{-ut}}{1+u^2}(u\sin t+\cos t)\,du.\end{aligned}$$

It is an elementary fact that $\int_0^\infty \frac{du}{1+u^2} = \frac{\pi}{2}$, and the change of variable $v = ut$ can be used to see that the second integral in the final result of the above calculation converges to 0 as $t\to\infty$. Hence

$$\lim_{t\to\infty}\mathrm{Si}(t) = \lim_{t\to\infty}\int_0^t \frac{\sin x}{x}\,dx = \frac{\pi}{2}.$$

Formalizing this argument was quite possibly the most painful part of our formalization of the CLT, but it paid off with improvements to the integration library made because of lessons learned; see the preceding section for an outline of these. I won't include all the technical lemmata we employed during this painful formalization, but give formal statements of the main result and the fact that the subtracted "error term" in the last line of the informal calculation goes to zero (in inverted order, of course).

```
lemma Si_at_top_lemma:
  shows "⋀t. t ≥ 0 ⟹ interval_lebesgue_integrable lborel 0 ∞
     (λx. exp (- (x * t)) * (x * sin t + cos t) / (1 + x²))"
  and
    "((λt. (LBINT x=0..∞. exp (-(x * t)) * (x * sin t + cos t) / (1 + x^2)))
---> 0) at_top"

lemma Si_at_top:
  shows "(Si ---> pi / 2) at_top"
```

We shall see more use of the Isabelle integration library when working with characteristic functions, and in particular the proofs of the Lévy inversion and continuity theorems. Much more than is outlined in this paper is present in the full formalization.

5.5 Characteristic Functions

As noted in the summary section 3, in a probabilistic context the Fourier transform of a probability measure (equivalently, of a random variable distributed as the given measure) is called its characteristic function. Here we describe our formalization of the definition and basic properties of characteristic functions.

First we need some properties of e^{ix}:

```
abbreviation iexp :: "real ⇒ complex" where
  "iexp ≡ (λx. Exp (i * complex_of_real x))"

lemma isCont_iexp [simp]: "isCont iexp x"
by (intro continuous_intros)
lemma cmod_iexp [simp]: "cmod (Exp (i * (x::real))) = 1"
by simp
lemma iexp_alt: "iexp x = cos x + i * sin x"
by (simp add: complex_eq_iff cis_conv_exp[symmetric] cos_of_real sin_of_real)
lemma has_vector_derivative_iexp[derivative_intros]:
  "(iexp has_vector_derivative i * iexp x) (at x within s)"
by (auto intro!: derivative_eq_intros simp: Re_exp Im_exp has_vector_derivative_complex_iff)
```

When we initially began formalizing properties of characteristic functions, the Isabelle library had no support for integrals of functions of type $\mathbb{R} \to \mathbb{C}$; fortunately the formalization of Bochner integration corrects that problem.

```
lemma interval_integral_iexp:
  fixes a b :: real
  shows "(CLBINT x=a..b. iexp x) = ii * iexp a - ii * iexp b"
by (subst interval_integral_FTC_finite [where F = "λx. -ii * iexp x"])
    (auto intro!: derivative_eq_intros continuous_intros)
```

Here CLBINT is ASCII notation for the Lebesgue integral of a function of type $\mathbb{R} \to \mathbb{C}$ with the canonical Lebesgue measure. As discussed in section 5.4.1, computing the value of an integral is useless without also showing the function is integrable, so we do that as well.

```
lemma (in prob_space) integrable_iexp:
  assumes f: "f ∈ borel_measurable M" "⋀x. Im (f x) = 0"
  shows "integrable M (λx. Exp (ii * (f x)))"
proof (intro integrable_const_bound [of _ 1])
  from f have "⋀x. of_real (Re (f x)) = f x"
    by (simp add: complex_eq_iff)
  then show "AE x in M. cmod (Exp (i * f x)) ≤ 1"
    using cmod_iexp[of "Re (f x)" for x] by simp
qed (insert f, simp)
```

We can now define the characteristic function of a measure and prove some basic properties. Informally, the characteristic function of a probability measure μ is simply the function $\varphi(t) = \int e^{itx}\,\mu(dx)$, which is of course continuous (in fact uniformly continuous, see [6] p. 343).

```
definition
  char :: "real measure ⇒ real ⇒ complex"
where
  "char M t ≡ complex_lebesgue_integral M (λx. iexp (t * x))"

lemma (in real_distribution) char_zero: "char M 0 = 1"
unfolding char_def by (simp del: space_eq_univ add: prob_space) lemma (in
real_distribution) cmod_char_le_1: "norm (char M t) ≤ 1"
proof -
```

```
  have "norm (char M t) ≤ (∫ x. norm (iexp (t * x)) ∂M)"
    unfolding char_def by (intro integral_norm_bound integrable_iexp) auto
  also have "... ≤ 1"
    by (simp del: of_real_mult)
  finally show ?thesis .
qed lemma (in real_distribution) isCont_char: "isCont (char M) t"
unfolding continuous_at_sequentially
proof safe
  fix X assume X: "X ----> t"
  show "(char M ∘ X) ----> char M t"
    unfolding comp_def char_def
    by (rule integral_dominated_convergence[where w="λ_. 1"])
       (auto simp del: of_real_mult intro!: AE_I2 tendsto_intros X)
qed lemma (in real_distribution) char_measurable [measurable]: "char M ∈
borel_measurable borel"
by (auto intro!: borel_measurable_continuous_on1 continuous_at_imp_continuous_on
isCont_char)
```

It is instrumental to the proof of the central limit theorem that if X and Y are independent random variables (say on the space $(\Omega, \mathcal{F}, \mathbb{P})$), then the characteristic function of their sum is the (pointwise) product of their characteristic functions. To see this, note that the random vectors $(\cos X, \sin X)$ and $(\cos Y, \sin Y)$ are independent, and so, letting φ_1 be the characteristic function of X, φ_2 be the characteristic function of Y, and $t \in \mathbb{R}$, we have (following Billingsley [6] as usual)

$$\begin{aligned}\varphi_1(t)\varphi_2(t) &= (\mathbb{E}(\cos X) + i\mathbb{E}(\sin X))(\mathbb{E}(\cos Y) + i\mathbb{E}(\sin Y)) \\ &= \mathbb{E}(\cos X)\mathbb{E}(\cos Y) - \mathbb{E}(\sin X)\mathbb{E}(\sin Y) \\ &\quad + i(\mathbb{E}(\cos X)\mathbb{E}(\sin Y) + \mathbb{E}(\sin X)\mathbb{E}(\cos Y)) \\ &= \mathbb{E}(\cos X \cos Y - \sin X \sin Y + i(\cos X \sin Y + \sin X \cos Y)) \\ &= \mathbb{E}(e^{it(X+Y)}).\end{aligned}$$

Since if $X \perp\!\!\!\perp Y$ and $X \sim \mu$, $Y \sim \nu$, then $X + Y \sim \mu * \nu$, the convulution of μ and ν, we see that in general the characteristic function of a convolution is the (pointwise) product of the characteristic functions.

All this is formalized as follows. The setsum version handles finite sums, as will occur in the proof of the CLT.

```
lemma (in prob_space) char_distr_sum:
  fixes X1 X2 :: "'a ⇒ real" and t :: real
  assumes "indep_var borel X1 borel X2"
  shows "char (distr M borel (λω. X1 ω + X2 ω)) t =
    char (distr M borel X1) t * char (distr M borel X2) t"
proof -
  from assms have [measurable]: "random_variable borel X1" by (elim indep_var_rv1)
  from assms have [measurable]: "random_variable borel X2" by (elim indep_var_rv2)

  have "char (distr M borel (λω. X1 ω + X2 ω)) t = (CLINT x|M. iexp (t *
(X1 x + X2 x)))"
```

```
    by (simp add: char_def integral_distr)
  also have "... = (CLINT x|M. iexp (t * (X1 x)) * iexp (t * (X2 x))) "
    by (simp add: field_simps exp_add)
  also have "... = (CLINT x|M. iexp (t * (X1 x))) * (CLINT x|M. iexp (t * (X2
x)))"
    by (intro indep_var_lebesgue_integral indep_var_compose[unfolded comp_def,
OF assms])
       (auto intro!: integrable_iexp)
  also have "... = char (distr M borel X1) t * char (distr M borel X2) t"
    by (simp add: char_def integral_distr)
  finally show ?thesis .
qed

lemma (in prob_space) char_distr_setsum:
  "indep_vars (λi. borel) X A ⟹
    char (distr M borel (λω. ∑i∈A. X i ω)) t = (∏i∈A. char (distr M borel
(X i)) t)"
proof (induct A rule: infinite_finite_induct)
  case (insert x F) with indep_vars_subset[of "λ_. borel" X "insert x F" F]
show ?case
    by (auto simp add: char_distr_sum indep_vars_setsum)
qed (simp_all add: char_def integral_distr prob_space del: distr_const)
```

We shall also need the characteristic function of the standard normal distribution, in order to show that the product of characteristic functions of independent identically distributed random variables of finite variance converges to it. The characteristic function of the standard normal distribution is $e^{-t^2/2}$; the detailed calculation and its associated technical lemmata are ommited.

```
abbreviation
  "std_normal_distribution ≡ density lborel std_normal_density"

lemma real_dist_normal_dist: "real_distribution std_normal_distribution"
unfolding real_distribution_def
  apply (rule conjI)
  apply (rule prob_space_normal_density, auto)
unfolding real_distribution_axioms_def by auto

theorem char_std_normal_distribution:
  "char std_normal_distribution = (λt. complex_of_real (exp (- (t^2) / 2)))"
```

5.6 The Lévy Theorems

Here we formalize some significant results about characteristic functions which are essential to their usefullness for studying distribution functions. The Lévy inversion theorem shows that the characteristic function of a distribution uniquely determines that distribution, while the Lévy continuity theorem shows that weak convergence of distributions is equivalent to pointwise convergence of the associated characteristic functions.

5.6.1 The Lévy Inversion Theorem

In preparation for the informal proof of the Lévy inversion theorem, note that

$$(*) \qquad \int_0^t \frac{\sin x\theta}{x}\,dx = \operatorname{sgn}\theta\,\mathrm{Si}(t|\theta|),$$

where $\operatorname{sgn} x$ is the sign of x ($\operatorname{sgn} x = 1$ if $x > 0$, $\operatorname{sgn} 0 = 0$, and $\operatorname{sgn} x = -1$ if $x < 0$).

Theorem 5.16. Let μ be a probability measure, and φ be the characteristic function of μ. If a and b are continuity points of μ and $a < b$, then

$$\mu(a,b] = \lim_{T\to\infty} \frac{1}{2\pi}\int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it}\varphi(t)\,dt.$$

Hence distinct probability measures have distinct characteristic functions.

Proof. Define

$$I(T) = \frac{1}{2\pi}\int_{-T}^{T} \frac{e^{ita} - e^{itb}}{it}\varphi(t)\,dt.$$

Since $[-T,T]\times\mathbb{R}$ has finite measure with respect to $\lambda\times\mu$ (where λ is Lebesgue measure), and $|\varphi(t)| \le 1$ and

$$\left|\frac{e^{ita} - e^{itb}}{it}\right| \le |b-a|$$

for all t, we have by Fubini's theorem that

$$I(T) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-T}^{T} \frac{e^{it(x-a)} - e^{it(x-b)}}{it}\,dt\,\mu(dx).$$

Using DeMoivre's formula to rewrite the integrand and noting $(*)$ (from before the statement of the theorem) and the fact that sin is an odd, and cos an even, function reveals that

$$I(T) = \int_{-\infty}^{\infty}\left(\frac{\operatorname{sgn}(x-a)}{\pi}\mathrm{Si}(T|x-a|) - \frac{\operatorname{sgn}(x-b)}{\pi}\mathrm{Si}(T|x-b|)\right)\mu(dx).$$

Since $\lim_{T\to\infty}\mathrm{Si}(T) = \frac{\pi}{2}$, we have that Si is bounded, and hence the integrand is bounded. Moreover, the integrand converges to the function given by

$$\psi_{a,b}(x) = \begin{cases} 0 & \text{if } x < a, \\ \frac{1}{2} & \text{if } x = a, \\ 1 & \text{if } a < x < b, \\ \frac{1}{2} & \text{if } x = b, \\ 0 & \text{if } b < x. \end{cases}$$

Thus by the bounded convergence theorem we have that $I(T) \to \int \psi_{a,b}\,d\mu$ as $T\to\infty$, which implies the desired conclusion when $\mu\{a\} = \mu\{b\} = 0$. □

Uniqueness follows because if two Borel probability measures μ, ν have the same characteristic function, they agree on the π-system of half-open intervals $(a, b]$ where $\mu\{a\} = \nu\{a\} = \mu\{b\} = \nu\{b\} = 0$ (this is a π-system generating the Borel sets because μ and ν each have countably many atoms), and hence everywhere by the Carathéodory extension theorem.

The bound

$$\left|\frac{e^{ita} - e^{itb}}{it}\right| \leq |b - a|$$

is formalized as

```
lemma Levy_Inversion_aux2:
  fixes a b t :: real
  assumes "a ≤ b" and "t ≠ 0"
  shows "cmod ((iexp (t * b) - iexp (t * a)) / (ii * t)) ≤ b - a"
    (is "?F ≤ _")
proof -
  have "?F = cmod (iexp (t * a) * (iexp (t * (b - a)) - 1) / (ii * t))"
    using ‘t ≠ 0‘ by (intro arg_cong[where f=norm]) (simp add: field_simps
exp_diff exp_minus)
  also have "... = cmod (iexp (t * (b - a)) - 1) / abs t"
    apply (subst norm_divide)
    apply (subst norm_mult)
    apply (subst cmod_iexp)
    using ‘t ≠ 0‘ by (simp add: complex_eq_iff norm_mult)
  also have "... ≤ abs (t * (b - a)) / abs t"
    apply (rule divide_right_mono)
    using equation_26p4a [of "t * (b - a)" 0] apply (simp add: field_simps
eval_nat_numeral)
    by force
  also have "... = b - a"
    using assms by (auto simp add: abs_mult)
  finally show ?thesis .
qed
```

The formal statements of the inversion and uniqueness theorems follow; both proofs are long and omitted, though it should be noted that obtaining uniqueness from inversion was not as straightforward as it appears it ought to be from the informal proof.

```
theorem Levy_Inversion:
  fixes M :: "real measure"
  and a b :: real
  assumes "a ≤ b"
  defines "μ ≡ measure M" and "φ ≡ char M"
  assumes "real_distribution M"
  and "μ {a} = 0" and "μ {b} = 0"
  shows
  "((λT :: nat. 1 / (2 * pi) * (CLBINT t=-T..T. (iexp (-(t * a)) -
  iexp (-(t * b))) / (ii * t) * φ t)) ---> μ {a<..b}) at_top"
  (is "((λT :: nat. 1 / (2 * pi) * (CLBINT t=-T..T. ?F t * φ t)) --->
```

```
      of_real (μ {a<..b})) at_top")

theorem Levy_uniqueness:
  fixes M1 M2 :: "real measure"
  assumes "real_distribution M1" "real_distribution M2" and
    "char M1 = char M2"
  shows "M1 = M2"
```

5.6.2 The Lévy Continuity Theorem

We are finally ready to connect characteristic functions to weak convergence.

Theorem 5.17. Let $\langle \mu_n \mid n \in \mathbb{N} \rangle$ be a sequence of probability measures with characteristic functions $\langle \varphi_n \mid n \in \mathbb{N} \rangle$, and μ be a probability measure with characteristic function φ. Then $\mu_n \Rightarrow \mu$ iff $\varphi_n \to \varphi$ pointwise.

Proof. ($\Longrightarrow$): Since e^{itx} has bounded modulus and is continuous in x for each $t \in \mathbb{R}$, this follows immediately from the portmanteau theorem applied to the real and imaginary parts of e^{itx}.

($\Longleftarrow$): We have another opportunity to use Fubini's theorem.

$$\begin{aligned}
\frac{1}{u}\int_{-u}^{u}(1-\varphi_n(t))\,dt &= \int_{-\infty}^{\infty}\left(\frac{1}{u}\int_{-u}^{u}(1-e^{itx})\,dt\right)\mu_n(dx)\\
&= 2\int_{-\infty}^{\infty}(1-\operatorname{sinc} ux)\,\mu_n(dx)\\
&\geq 2\int_{|x|\geq 2/u}\left(1-\frac{1}{|ux|}\right)\mu_n(dx)\\
&\geq \mu_n\{x\in\mathbb{R}\mid |x|\geq\frac{2}{u}\}.
\end{aligned}$$

Since φ is continuous and $\varphi(0) = 1$, for every $\varepsilon > 0$ there exists $u \in \mathbb{R}$ such that

$$\frac{1}{u}\int_{-u}^{u}(1-\varphi(t))\,dt < \varepsilon.$$

Because $\varphi_n \to \varphi$ pointwise, by the bounded convergence theorem there is $n_0 \in \mathbb{N}$ such that

$$\frac{1}{u}\int_{-u}^{u}(1-\varphi_n(t))\,dt < 2\varepsilon.$$

for $n \geq n_0$. Thus $\mu_n\{x \in \mathbb{R} \mid |x| \geq 2u^{-1}\} < 2\varepsilon$, and so for some $a \geq 2u^{-1}$ we have $\mu_n\{x \in \mathbb{R} \mid |x| \geq a\} < 2\varepsilon$. Consequently the sequence $\langle \mu_n \mid n \in \mathbb{N} \rangle$ is tight.

By the corollary concerning tightness (see the end of section 5.3) it is sufficient to show that every subsequece $\langle \mu_{n_k} \mid k \in \mathbb{N} \rangle$ which has a weak limit in fact converges weakly to μ. If $\mu_{n_k} \Rightarrow \nu$ for some probability measure ν, then by the first half of the theorem we have that $\lim_{k\to\infty} \varphi_{n_k}(t) = \varphi(t)$ is the characteristic function of ν, and hence $\nu = \mu$ by the uniqueness theorem. □

As the reader may anticipate, the formal proof is long and difficult. The formal statement of the Lévy continuity theorem suffices to be written here.

```
theorem levy_continuity1:
  fixes
    M :: "nat ⇒ real measure" and
    M' :: "real measure"
  assumes
    real_distr_M : "⋀n. real_distribution (M n)" and
    real_distr_M': "real_distribution M'" and
    measure_conv: "weak_conv_m M M'"
  shows
    "(λn. char (M n) t) ----> char M' t"
```

5.7 The Central Limit Theorem

All the pieces are now in place for the proof of the central limit theorem; we just need to put them together. To remind the reader: we shall show that if μ is the distribution of a random variable with finite nonzero variance and φ is the characteristic function of μ, then $\varphi(\sigma^{-1}n^{-1/2}x)^n \to e^{-x^2/2}$ for each $x \in \mathbb{R}$, where $e^{-x^2/2}$ is the characteristic function of a standard normal distribution. From this and facts proven earlier about characteristic functions of independent random variables it follows that the normalized sums $\frac{1}{\sigma\sqrt{n}}\sum_{k=0}^{n}(X_k - \mathbb{E}(X_k))$ converge weakly to a standard normal distribution. For the remainder of this section, we assume without loss of generality that the random variables we work with have mean zero and variance one (otherwise a random variable with finite nonzero variance can be translated and scaled).

The central limit theorem is proved simply by showing that if X is a random variable with zero mean and unit variance, then the characteristic function φ of x satisfies $\varphi(t) = 1 - t^2/2 + o(t^2)$, as follows from Taylor expansion about zero. Thus

$$\varphi\left(\frac{t}{\sqrt{n}}\right)^n = \left(1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right) \to e^{-t^2/2},$$

by basic facts about limits. Since $e^{-t^2/2}$ is the characteristic function of a standard normal distribution, and the characteristic function of $\frac{1}{\sqrt{n}}\sum_{k=0}^{n} X_k$ is $\varphi(t/\sqrt{n})^n$, we have by the Lévy continuity theorem that

$$\frac{1}{\sqrt{n}}\sum_{k=0}^{n} X_k \Rightarrow N,$$

where N is a random variable with standard normal distribution, as desired.

Because it is the primary goal of our formalization, we present the formal proof of the central limit theorem in full.

```
theorem (in prob_space) central_limit_theorem:
  fixes
    X :: "nat ⇒ 'a ⇒ real" and
```

```
    μ :: "real measure" and
    σ :: real and
    S :: "nat ⇒ 'a ⇒ real"
  assumes
    X_indep: "indep_vars (λi. borel) X UNIV" and
    X_integrable: "⋀n. integrable M (X n)" and
    X_mean_0: "⋀n. expectation (X n) = 0" and
    σ_pos: "σ > 0" and
    X_square_integrable: "⋀n. integrable M (λx. (X n x)^2)" and
    X_variance: "⋀n. variance (X n) = σ^2" and
    X_distrib: "⋀n. distr M borel (X n) = μ"
  defines
    "S n ≡ λx. ∑i<n. X i x"
  shows
    "weak_conv_m (λn. distr M borel (λx. S n x / sqrt (n * σ^2)))
        (density lborel std_normal_density)"
proof -
  def S' ≡ "λn x. S n x / sqrt (n * σ^2)"
  def φ ≡ "λn. char (distr M borel (S' n))"
  def ψ ≡ "λn t. char μ (t / sqrt (σ^2 * n))"

  have X_rv [simp, measurable]: "⋀n. random_variable borel (X n)"
    using X_indep unfolding indep_vars_def2 by simp
  interpret μ: real_distribution μ
    by (subst X_distrib [symmetric, of 0], rule real_distribution_distr, simp)

  have μ_integrable [simp]: "integrable μ (λx. x)"
    apply (simp add: X_distrib [symmetric, of 0])
    using assms by (subst integrable_distr_eq, auto)
  have μ_mean_integrable [simp]: "μ.expectation (λx. x) = 0"
    apply (simp add: X_distrib [symmetric, of 0])
    using assms by (subst integral_distr, auto)
  have μ_square_integrable [simp]: "integrable μ (λx. x^2)"
    apply (simp add: X_distrib [symmetric, of 0])
    using assms by (subst integrable_distr_eq, auto)
  have μ_variance [simp]: "μ.expectation (λx. x^2) = σ^2"
    apply (simp add: X_distrib [symmetric, of 0])
    using assms by (subst integral_distr, auto)

  have main: "⋀t. eventually (λn. cmod (φ n t - (1 + (-(t^2) / 2) / n)^n)
≤
        (t^2 / (6 * σ^2) * (LINT x|μ. min (6 * x^2) (|t / sqrt (σ^2 * n)| * |x|
^ 3))))
      sequentially"
  proof (rule eventually_sequentiallyI)
    fix n :: nat and t :: real
    assume "n ≥ nat (ceiling (t^2 / 4))"
    hence n: "n ≥ t^2 / 4" by (subst nat_ceiling_le_eq [symmetric])
    let ?t = "t / sqrt (σ^2 * n)"
```

```
    def ψ' ≡ "λn i. char (distr M borel (λx. X i x / sqrt (σ^2 * n)))"
    have *: "⋀n i t. ψ' n i t = ψ n t"
      unfolding ψ_def ψ'_def char_def apply auto
      apply (subst X_distrib [symmetric])
      apply (subst integral_distr, auto)
      by (subst integral_distr, auto)

    have 1: "S' n = (λx. (∑ i < n. X i x / sqrt (σ^2 * n)))"
      by (rule ext, simp add: S'_def S_def setsum_divide_distrib ac_simps)

    have "φ n t = (∏ i < n. ψ' n i t)"
      unfolding φ_def ψ'_def apply (subst 1)
      apply (rule char_distr_setsum)
      apply (rule indep_vars_compose2[where X=X])
      apply (rule indep_vars_subset)
      apply (rule X_indep)
      apply auto
      done
    also have "... = (ψ n t)^n"
      by (auto simp add: * setprod_constant)
    finally have 2: "φ n t =(ψ n t)^n" .

    have "cmod (ψ n t - (1 - ?t^2 * σ^2 / 2)) ≤
        ?t^2 / 6 * (LINT x|μ. min (6 * x^2) (|?t| * |x| ^ 3))"
      unfolding ψ_def by (rule μ.aux, auto)
    also have "?t^2 * σ^2 = t^2 / n"
      using σ_pos by (simp add: power_divide)
    also have "t^2 / n / 2 = (t^2 / 2) / n" by simp
    finally have **: "cmod (ψ n t - (1 + (-(t^2) / 2) / n)) ≤
      ?t^2 / 6 * (LINT x|μ. min (6 * x^2) (|?t| * |x| ^ 3))" by simp

    have "cmod (φ n t - (complex_of_real (1 + (-(t^2) / 2) / n))^n) ≤
         n * cmod (ψ n t - (complex_of_real (1 + (-(t^2) / 2) / n)))"
      apply (subst 2, rule norm_power_diff)
      unfolding ψ_def apply (rule μ.cmod_char_le_1)
      apply (simp only: norm_of_real)
      apply (auto intro!: abs_leI)
      using n by (subst divide_le_eq, auto)
    also have "... ≤ n * (?t^2 / 6 * (LINT x|μ. min (6 * x^2) (|?t| * |x| ^
3)))"
      by (rule mult_left_mono [OF **], simp)
    also have "... = (t^2 / (6 * σ^2) * (LINT x|μ. min (6 * x^2) (|?t| * |x| ^
3)))"
      using σ_pos by (simp add: field_simps min_absorb2)
    finally show "cmod (φ n t - (1 + (-(t^2) / 2) / n)^n) ≤
        (t^2 / (6 * σ^2) * (LINT x|μ. min (6 * x^2) (|?t| * |x| ^ 3)))"
      by simp
  qed

  have S_rv [simp, measurable]: "⋀n. random_variable borel (λx. S n x / sqrt
```

```
(n * σ²))"
    unfolding S_def by measurable
  have "⋀t. (λn. φ n t) ----> char std_normal_distribution t"
  proof -
    fix t
    let ?t = "λn. t / sqrt (σ² * n)"
    have *: "⋀n. integrable μ (λx. 6 * x^2)" by auto
    have **: "⋀n. integrable μ (λx. min (6 * x²) (|t / sqrt (σ² * real n)|
* |x| ^ 3))"
      by (rule integrable_bound [OF *]) auto
    have ***: "⋀x. (λn. |t| * |x| ^ 3 / |sqrt (σ² * real n)|) ----> 0"
      apply (subst divide_inverse)
      apply (rule tendsto_mult_right_zero)
      using σ_pos apply (subst abs_of_nonneg, simp)
      apply (simp add: real_sqrt_mult)
      apply (rule tendsto_mult_right_zero)
      apply (rule tendsto_inverse_0_at_top)
      by (rule filterlim_compose [OF sqrt_at_top filterlim_real_sequentially])
    have "(λn. LINT x|μ. min (6 * x²) (|?t n| * |x| ^ 3)) ----> (LINT x|μ.
0)"
      apply (rule integral_dominated_convergence [where w = "λx. 6 * x^2"])
      using σ_pos apply (auto intro!: AE_I2)
      apply (rule tendsto_sandwich [OF _ _ tendsto_const ***])
      apply (auto intro!: always_eventually min.cobounded2)
      done
    hence "(λn. LINT x|μ. min (6 * x²) (|?t n| * |x| ^ 3)) ----> 0" by simp
    hence main2: "(λn. t² / (6 * σ²) * (LINT x|μ. min (6 * x²) (|?t n| *
|x| ^ 3))) ----> 0"
      by (rule tendsto_mult_right_zero)
    have **: "(λn. (1 + (-(t^2) / 2) / n)^n) ----> exp (-(t^2) / 2)"
      by (rule tendsto_exp_limit_sequentially)
    have "(λn. complex_of_real ((1 + (-(t^2) / 2) / n)^n)) ---->
        complex_of_real (exp (-(t^2) / 2))"
      by (rule isCont_tendsto_compose [OF _ **], auto)
    hence "(λn. φ n t) ----> complex_of_real (exp (-(t^2) / 2))"
      apply (rule Lim_transform)
      by (rule Lim_null_comparison [OF main main2])
    thus "(λn. φ n t) ----> char std_normal_distribution t"
      by (subst char_std_normal_distribution)
  qed
  thus ?thesis
    apply (intro levy_continuity)
    apply (rule real_distribution_distr [OF S_rv])
    unfolding real_distribution_def real_distribution_axioms_def
    apply (simp add: prob_space_normal_density)
    unfolding φ_def S'_def by -
qed
```

6 Conclusion: Opportunities for Improvement and Extension

The version of the central limit theorem we proved is not the most general presented in Billingsley [6]. With some more calculational effort we could formalize the Lindeberg central limit theorem, which relaxes the condition that the random variables being summed be identically distributed (they merely need to not deviate too much in distribution, as made precise by the *Lindeberg condition* given on p. 359 of [6]). Even the condition that the variables being summed be independent can be weakened to a condition of weak dependence; Billingsley begins to outline this on p. 363. Both of these are good candidates for further formalization now that the background theory of characteristic functions is in place, with the dependent variables generalization being the more ambitious. Other generalizations include the CLT for random vectors ([6], p. 385) and various versions of the CLT for martingales ([6], pp. 475–478). Many other refinements and generalizations for the central limit theorem exist in the mathematical literature.

During the formalization process, it was often surprising how far the analysis libraries of Isabelle extend, but at the same time frustrating that the automated tools would get stuck on seemingly trivial matters like determining whether an instance of zero should be interpreted as a real or an extended real. Clearly much more remains to be done to encode basic human analytical intuition into proof procedures.

Our formalization sometimes approximated an informal presentation quite well, but as the reader can perceive looking through the proof scripts we have presented formal proofs still tend to be much longer. This should be remedied as the library is developed further and automated tools are improved.

Our main goal for the central limit formalization project was to improve the Isabelle integration libraries, and this succeeded very well, first with the author and Avigad extending them as needed for proofs, and then with Hölzl unifying everything as he rewrote the library to accomodate vector-valued integrals, such as integrals of functions of type $\mathbb{R} \to \mathbb{C}$, natively.

As remarked in section 5.4.2, calculating with integrals was perhaps the most painful part of the formalization, and it is interesting to speculate how computer algebra systems could help remedy this situation. Perhaps a computer algebra system could be modified to keep track of enough information for Isabelle to reconstruct a proof (including a proof of integrability), or alternatively an interactive proof assistant could be used to verify the correctness of a computer algebra system and then the results obtained by that system could be used freely.

The depth of formalized analysis has increased dramatically in recent years, and we hope that our addition of the central limit theorem is valuable both as a milestone in the formalization of probability and statistics, and for the library infrastructure which was developed to support it (integration, Fourier transforms, cumulative distribution functions, etc.).

References

[1] Peter B Andrews. *An introduction to mathematical logic and type theory*, volume 27. Springer Science & Business Media, 2002.

[2] Jeremy Avigad, Kevin Donnelly, David Gray, and Paul Raff. A formally verified proof of the prime number theorem. *ACM Trans. Comput. Logic*, 9(1), December 2007.

[3] Jeremy Avigad, Johannes Hölzl, and Luke Serafin. A formally verified proof of the central limit theorem. *CoRR*, abs/1405.7012, 2014.

[4] Clemens Ballarin. Tutorial to locales and locale interpretation. In *Contribuciones científicas en honor de Mirian Andrés Gómez*, pages 123–140. Universidad de La Rioja, 2010.

[5] Bruno Barras, Samuel Boutin, Cristina Cornes, Judicaël Courant, Jean-Christophe Filliatre, Eduardo Gimenez, Hugo Herbelin, Gerard Huet, Cesar Munoz, Chetan Murthy, et al. The coq proof assistant reference manual: Version 6.1. 1997.

[6] Patrick Billingsley. *Probability and measure*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, third edition, 1995. A Wiley-Interscience Publication.

[7] Leonardo De Moura and Nikolaj Bjørner. Satisfiability modulo theories: introduction and applications. *Communications of the ACM*, 54(9):69–77, 2011.

[8] Herbert B Enderton. *Elements of set theory*. Academic Press, 1977.

[9] Hans Fischer. *A history of the central limit theorem: From classical to modern probability theory*. Springer Science & Business Media, 2010.

[10] Francis Galton. *Natural inheritance*. Macmillan, London, 1889.

[11] Michael JC Gordon. Edinburgh lcf: a mechanised logic of computation. 1979.

[12] Mike Gordon. Set theory, higher order logic or both? In *Theorem Proving in Higher Order Logics*, pages 191–201. Springer, 1996.

[13] Florian Haftmann. Haskell-style type classes with isabelle/isar, 2013.

[14] T. Hales, M. Adams, G. Bauer, D. Tat Dang, J. Harrison, T. Le Hoang, C. Kaliszyk, V. Magron, S. McLaughlin, T. Tat Nguyen, T. Quang Nguyen, T. Nipkow, S. Obua, J. Pleso, J. Rute, A. Solovyev, A. Hoai Thi Ta, T. N. Tran, D. Thi Trieu, J. Urban, K. Khac Vu, and R. Zumkeller. A formal proof of the Kepler conjecture. *ArXiv e-prints*, January 2015.

[15] Thomas C Hales. The jordan curve theorem, formally and informally. *The American Mathematical Monthly*, pages 882–894, 2007.

[16] John Harrison. Formalizing an analytic proof of the Prime Number Theorem (dedicated to Mike Gordon on the occasion of his 60th birthday). *Journal of Automated Reasoning*, 43:243–261, 2009.

[17] John Harrison. Hol light: An overview. In *Theorem Proving in Higher Order Logics*, pages 60–66. Springer, 2009.

[18] Johannes Hölzl and Armin Heller. Three chapters of measure theory in Isabelle/HOL. In Marko C. J. D. van Eekelen, Herman Geuvers, Julien Schmaltz, and Freek Wiedijk, editors, *Interactive Theorem Proving (ITP 2011)*, volume 6898 of *LNCS*, pages 135–151, 2011.

[19] Johannes Hölzl, Fabian Immler, and Brian Huffman. Type classes and filters for mathematical analysis in Isabelle/HOL. In Sandrine Blazy, Christine Paulin-Mohring, and David Pichardie, editors, *Interactive Theorem Proving (ITP 2013)*, volume 7998 of *LNCS*, pages 279–294. Springer, 2013.

[20] Yitzhak Katznelson. *An introduction to harmonic analysis*. Cambridge University Press, 2004.

[21] Tobias Nipkow. Programming and proving in isabelle. *HOL.[Online] Available: http://www. cl. cam. ac. uk/research/hvg/Isabelle/dist/Isabelle2013-2/do c/prog-prove. pdf*, 2013.

[22] Lawrence C Paulson and Jasmin Christian Blanchette. Three years of experience with sledgehammer, a practical link between automatic and interactive theorem provers. *IWIL-2010*, 2010.

[23] Bertrand Russell. Knowledge by acquaintance and knowledge by description. In *Proceedings of the Aristotelian Society*, pages 108–128. JSTOR, 1910.

[24] Atle Selberg. An elementary proof of the prime-number theorem. *Annals of Mathematics*, 50(2):305–313, 1949.

[25] Patrick Suppes. *Axiomatic set theory*. Courier Corporation, 1960.

[26] Giuseppe Vitali. Sul problema della misura dei gruppi di punti di una retta. *Bologna, Italy*, 1905.

[27] Markus Wenzel et al. *Isabelle/Isar—a versatile environment for human-readable formal proof documents*. PhD thesis, Institut für Informatik, Technische Universität München, 2002. http://tumb1. biblio. tu-muenchen. de/publ/diss/in/2002/wenzel. html, 2002.