Proof of the Central Limit Theorem Using Measures as Operators

Jimmy Risk

May 26, 2015

Preface

Most the content except for the appendix comes from [1], a book by Adam Bobrowski entitled Functional Analysis for Probability and Stochastic Processes. This book definitely helps tie a gap for probabilists to better understand functional analysis, and for analysts to better understand probability theory. The reader may assume that, unless cited otherwise, the content being read comes from this book, though there were some personal comments added, and some proofs were omitted from the book so I wrote them myself. The order is as follows. We first go over some measure theoretic preliminary results, and then define an operator based on a measure. Convergence of measure (in some sense) is then equivalent to convergence of these operators in the correct topology, and for a specific case we show that a sequence of operators converges, proving the central limit theorem.

Remark 0.1. This is an extremely untraditional way of proving the Central Limit Theorem; in fact, I was unable to find any other source that proved it using this method, while every other method I found uses the Fourier transform. I chose this approach because it uses what we have covered all year and gives further results about these topics (including measure theory, Banach spaces, weak* topology, $C(X)$ spaces, and operators).

Remark 0.2. Traditionally in probability theory the symbol Ω is a set, $\mathcal{F}$ is a $\sigma-$algebra, and $\mathbb{P}$ is a probability measure, so $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space. Furthermore, the letter X is typically used to denote a random variable (a measurable function) on the probability space. To avoid notational confusion, I will stick with (X, Ω, μ) as a measure (or probability if $\mu(X) = 1$) space, and measurable functions will be denoted f, g, etc..

1 Measure Theory Preliminaries

Suppose that (X, Ω, μ) is a measure space and f is a measurable map from (X, Ω) to another measurable space (X', Ω'). Consider the set function μ_f on Ω' defined by

$$\mu_f(A) = \mu(f^{-1}(A)) = \mu(\{x : f(x) \in A\}).$$

It is easy to show that μ_f is a measure on (X', Ω'). The measure μ_f is called the **transport of the measure μ via f**, or a measure **induced** on (X', Ω') by μ and f. In particular, if μ is a probability measure (i.e. $\mu(X) = 1$) and $(X', \Omega') = (\mathbb{R}^n, \mathcal{M}_n(\mathbb{R}^n))$, where $\mathcal{M}_n(\mathbb{R}^n)$ are the Lebesgue measurable sets on $\mathbb{R}^n$, then μ_f is called the **distribution** or **law** of f.

Proposition 1.1. A measurable function ϕ defined on X' is integrable with respect to μ_f iff $\phi \circ f$ is integrable with respect to μ, and

$$\int_{X'} \phi d\mu_f = \int_X \phi \circ f d\mu. \tag{1}$$

Proof. First suppose $\phi = \sum a_i \mathbb{1}_{A_i}$ for $A_i \in \Omega'$ is a positive integrable simple function. Compute

$$\begin{aligned}
\int_{X'} \phi(x) d\mu_f(x) &= \int_{X'} \sum a_i \mathbb{1}_{A_i}(x) d\mu_f(x) = \sum a_i \mu_f(A_i) \\
&= \sum a_i \mu(y : f(y) \in A_i) = \int_X \sum a_i \mathbb{1}_{A_i}(f(y)) d\mu(y) \\
&= \int_X (\phi \circ f)(y) d\mu(y).
\end{aligned}$$

Therefore the proposition holds in the simple function case. We can extend the result to positive functions by the monotone convergence theorem, and finally all functions by linearity and triangle inequality. □

Equation 1 is called the **change of variables formula**. A particular case is that where a measure, say ν, is already defined on $(X', \mathcal{F}')$, and μ_f is absolutely continuous with respect to ν. If the Radon-Nikodym derivative is $\psi = \frac{d\mu_f}{d\nu}$, then the change of variables formula reads

$$\int_X \phi \circ f d\mu = \int_{X'} \phi d\mu_f = \int_{X'} \phi \psi d\nu.$$

Example 1.2. Suppose we have a probability space (X, Ω, μ) and a measure space $(\mathbb{R}, \mathcal{B}(\mathbb{R}), m)$, where m is the Lebesgue measure. Let $f : X \to \mathbb{R}$ be a function defined through its distribution which satisfies $\frac{d\mu_f}{dm}(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. Then for $A \in \mathcal{B}(\mathbb{R})$, we have

$$\mu(f \in A) = \int \mathbb{1}_A(f(y)) d\mu(y) = \int \mathbb{1}_A(x) \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx.$$

The function f is said to be a **standard normal random variable** and its distribution is defined by $\mu_f(dx) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$.

2 Measures as operators

Notation.

- $BM(\mathbb{R})$ is the space of bounded Borel measurable functions on $\mathbb{R}$;
- $C_b(\mathbb{R}) \subset C(\mathbb{R})$ is the subspace consisting of bounded continuous functions;

- $UC_b(\mathbb{R})$ is the space of bounded uniformly continuous Borel measurable functions on $\mathbb{R}$;

- $C_0(\mathbb{R})$ is the space of continuous functions that vanish at infinity.

- $M_b(X)$ is the space of bounded scalar-value Borel measures on a topological space X.

All of the function spaces above are equipped with the sup norm.

Definition 2.1. Given a finite measure μ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ we may define an operator T_μ acting in $BM(\mathbb{R})$ by the formula

$$(T_\mu f)(x) = \int_{\mathbb{R}} f(x+y)d\mu(y). \tag{2}$$

Proposition 2.2. T_μ is well defined.

Proof. Let $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ be our measurable space with a finite measure μ and T_μ defined above. First we check that T_μ maps $BM(\mathbb{R})$ into itself. If $f = \mathbb{1}_{(a,b]}$ for some real numbers $a < b$, then

$$T_\mu f(x) = \mu(-\infty, b-x] - \mu(-\infty, a-x]$$

is of bounded variation (see appendix) and hence measurable by Corollary A.5. Let $\mathcal{G}$ be the class of measurable sets such that for $A \in \mathcal{G}$, $T_\mu \mathbb{1}_A$ is measurable. We show that $\mathcal{G}$ is a $\lambda-$system (see Appendix B). First, $T_\mu \mathbb{1}_{\mathbb{R}}(x) = 0$, so $\mathbb{R} \in \mathcal{G}$. Next, let $A, B \in \mathcal{G}$ and $A \supset B$. Then

$$\begin{aligned} T_\mu \mathbb{1}_{A\setminus B}(x) &= \int \mathbb{1}_{A\setminus B}(x+y)d\mu(y) = \int \mathbb{1}_A(x+y)d\mu(y) - \int \mathbb{1}_B(x+y)d\mu(y) \\ &= T_\mu \mathbb{1}_A(x) - T_\mu \mathbb{1}_B(x), \end{aligned}$$

so it is a difference of measurable functions and hence measurable. Lastly let $(A_n) \in \mathcal{G}$ be a sequence of increasing sets such that $\bigcup_n A_n = A$. Then

$$T_\mu \mathbb{1}_A(x) = \int \mathbb{1}_{\bigcup_n A_n}(x+y)d\mu(y) = \lim_{m\to\infty} \int \mathbb{1}_{\bigcup_n^m A_n}(x+y)d\mu(y)$$

by monotone convergence theorem. Since the limit of measurable functions on $\mathbb{R}$ is measurable (Weaver Exercise 2.7), $T_\mu \mathbb{1}_A(x)$ is measurable. We have already shown that intervals are contained in $\mathcal{G}$, and intervals generate $\mathcal{B}(\mathbb{R})$, so by the monotone class theorem (Theorem B.4), $\mathcal{G} = \mathcal{B}(\mathbb{R})$. Hence $T_\mu f$ is measurable for any $f = \mathbb{1}_A$, where $A \in \mathcal{B}(\mathbb{R})$. By linearity of integral this is also true for simple functions. Since pointwise limits of measurable functions are measurable, this extends to all $f \in BM(\mathbb{R})$. As for its boundedness, we have

$$\|T_\mu f\| \le \sup_{x\in\mathbb{R}} |(T_\mu f)(x)| \le \sup_{x\in\mathbb{R}} \sup_{y\in\mathbb{R}} |f(x+y)|\mu(\mathbb{R}) = \|f\|\mu(\mathbb{R}),$$

with equality for $f = \mathbb{1}_{\mathbb{R}}$. Thus $\|T_\mu\| = \mu(\mathbb{R})$ and in particular $\|T_\mu\| = 1$ for a probability measure μ. □

Remark 2.3. T_μ is invariant for the subspaces $C_b(\mathbb{R})$ and $UC_b(\mathbb{R})$. The $C_b(\mathbb{R})$ case follows by Lebesgue Dominated Convergence Theorem and the $UC_b(\mathbb{R})$ case follows from

$$|(T_\mu f)(x) - (T_\mu f)(y)| \leq \sup_{z\in\mathbb{R}} |f(x+z) - f(y+z)|\mu(\mathbb{R}), \quad x, y \in \mathbb{R}.$$

It can similarly be shown that T_μ maps $C_0(\mathbb{R})$ into itself.

Proposition 2.4. If $T_\mu \in \mathcal{B}(B(\mathbb{R}))$, then it uniquely determines μ.

Proof. Note that if $T_\mu = T_\nu$, then

$$(T_\mu f)(0) = \int f d\mu,$$
$$(T_\nu f)(0) = \int f d\nu.$$

The result follows from Lemma A.1. □

Corollary 2.5. The map $\mu \mapsto T_\mu$ is a linear invertible map from $M_b(\mathbb{R})$ into $\mathcal{B}(BM(\mathbb{R}))$ (bounded linear operators on $BM(\mathbb{R})$)

3 Relating to Probability

3.1 Independence

Definition 3.1. Given a measure space (X, Ω, μ), a topological space (X', Ω'), and a function $f : X \to X'$, we say the $\sigma-$algebra generated by f, written $\sigma(f)$, is the smallest algebra containing the sets $f^{-1}(A)$ for all $A \in \Omega'$.

Definition 3.2. Let space (X, Ω, μ) be a **probability** space. Let $\mathcal{F}_j, j \in J$ be a family of classes of measurable subsets (J is an abstract index set). The classes are termed **mutually independent** if for all $n \in \mathbb{N}$, all $j_1, \dots, j_n \in J$, and all $A_i \in F_{j_i}, i = 1, \dots, n$ the following holds:

$$\mu\left(\bigcap_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} \mu(A_i).$$

The classes are termed **pairwisely independent** if for all $n \in \mathbb{N}$, all $t_1, t_2 \in J$, and all $A_i \in \mathcal{F}_{j_i}, i = 1, 2,$

$$\mu(A_1 \cap A_2) = \mu(A_1)\mu(A_2).$$

Functions $f_j, j \in J$ are said to be mutually (pairwisely) independent if the $\sigma-$algebras $\mathcal{F}_j = \sigma(f_j)$ generated by f_j are mutually (pairwisely) independent.

From now on, the phrase "classes (functions) are independent" should be understood as "classes (functions) are mutually independent".

3.2 Convolution

Definition 3.3. For any two finite Borel measures μ, ν, we may define the convolution $\mu * \nu$ as

$$(\mu * \nu)(E) = \int\int \mathbb{1}_E(y+z)d\mu(y)d\nu(z)$$

for all Borel measurable E.

Proposition 3.4. Two functions f and g defined on a probability space into a Borel measure space are independent iff the distribution of (f, g) is $\mu_f \otimes \mu_g$.

Proof. ($\Leftarrow$) Let $A \in \sigma(f), B \in \sigma(g)$. Then $A = \{x : f(x) \in \tilde{A}\}$ and $B = \{x : g(x) \in \tilde{B}\}$. If $\mu_{(f,g)} = \mu_f \otimes \mu_g$, then

$$\mu(A \cap B) = \mu_{(f,g)}(\tilde{A} \times \tilde{B}) = \mu_f(\tilde{A})\mu_g(\tilde{B}) = \mu(A)\mu(B).$$

($\Rightarrow$) If $D = A \times B$ is an open set, then

$$\mu_{(f,g)}(D) = \mu_{(f,g)}(A \times B) = \mu(\{x : f(x) \in A\} \cap \{y : g(y) \in B\}) = \mu_f(A)\mu_g(B).$$

Since the open sets generate the Borel $\sigma-$algebra, it holds for all Borel sets.

□

Corollary 3.5. If f and g are two independent functions on a probability space (X, Ω, μ), then the distribution of their sum is the convolution of their distributions:

$$\mu_{f+g} = \mu_f * \mu_g.$$

Proof. This is a simple computation using Proposition 3.4:

$$\begin{aligned}
\mu_{f+g}(A) &= \mu\left(\{x : f(x) + g(x) \in A\}\right) = \int \mathbb{1}_{\{f(x)+g(x)\in A\}}d\mu(x) \\
&= \int\int \mathbb{1}_{\{s+t\in A\}}d\left(\mu_f \otimes \mu_g\right)(s,t) = \int\int \mathbb{1}_{\{s+t\in A\}}d\mu_f(s)d\mu_g(t) \\
&= (\mu_f * \mu_g)(A).
\end{aligned}$$

□

4 Convergence of Measure

On a measure space (X, Ω, μ), if $f : X \to X'$ for a target space (X', Ω'), then we may assign it to the operator T_f defined by $T_f := T_{\mu_f}$ where μ_f is the transport measure of f.

Recall the general Riesz-Markov theorem from Weaver.

Theorem 4.1. Let X be a second countable locally compact Hausdorff space. Then every bounded linear functional on $C_0(X)$ is given by integrating against a scalar-valued Borel measure on X, and this pairing implements an isometric isomorphism between $C_0(X)'$ and $M(X)$.

From this, we can consider weak* convergence of Borel measures on X. Actually, if X is only locally compact, it is often more convenient to consider X^*,the one-point compactification of X, and to use $C(X^*)$ as the set of test functions for weak* convergence. The reason is that it may happen that some mass of the measure may escape to infinity, as in the following example.

Example 4.2. Let $X = \mathbb{R}^+$ and $\mu_n = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_n$, where δ_k is the Dirac measure equaling 1 if $k \in A$ and 0 otherwise. In this case, for any $f \in C_0(\mathbb{R}^+)$, $\int f d\mu_n$ converges to $\frac{1}{2}f(0)$, so that μ_n as functionals on $C_0(\mathbb{R}^+)$ converge in the weak* topology to an (improper[1]) distribution $\frac{1}{2}\delta_0$. In this approach it is unclear what happened with the missing mass. Taking an $f \in C((R^+)^*) = C([0,\infty])$ clarifies the situation, because for $f \in C([0,\infty])$ we see that $\int f d\mu_n$ converges to $\frac{1}{2}f(0) + \frac{1}{2}f(\infty)$, and so μ_n converges to $\frac{1}{2}\delta_0 + \frac{1}{2}\delta_\infty$.

The following lemma serves very well as a useful tool in proving weak* convergence of measures. It is the main tool in proving the Central Limit Theorem. It involves use of the Arzela-Ascoli Theorem – see Appendix C. First we define the strong operator topology.

Definition 4.3. If X and Y are Banach spaces, the **strong operator topology** is the topology defined on $\mathcal{B}(X,Y)$ generated by the family of seminorms $\{p_x : x \in X\}$, where $p_x(T) = \|Tx\|_Y$ for $T \in \mathcal{B}(X,Y)$. Hence if $\|T_n x - Tx\|_Y \to 0$ as $n \to \infty$ for all $x \in X$, we say T_n **converges strongly** to T.

Lemma 4.4. A sequence μ_n of probability measures on $[-\infty,\infty]$ converges to a probability measure μ in the weak* topology iff the corresponding operators T_{μ_n} in $C([-\infty,\infty])$ converge strongly to T_μ.

Proof. Suppose $T_{\mu_n} \to T$ strongly, that is, by Definition 2.1, $\|T_{\mu_n} f - T_\mu f\| \to 0$ for all $f \in C([-\infty,\infty])$. That is,

$$\sup_x \left| \int f(x+y)d\mu_n(y) - \int f(x+y)d\mu(y) \right| \to 0.$$

However, the supremum is over all x, so the case where $x = 0$ yields $|\int f(y)d\mu_n(y) - \int f(y)d\mu(y)| \to 0$. But in this case, the left hand side is actually equal to $|\mu_n(f) - \mu(f)|$. This holds for all $f \in C([-\infty,\infty])$, so $\mu_n \to \mu$ in the weak* topology.

Now suppose $\mu_n \to \mu$ in the weak* topology. By assumption, for any $f \in C([-\infty,\infty])$ and any $x \in (-\infty,\infty)$ there exists the limit of $g_n(x) = \int f(x+y)d\mu_n(y)$, as $n \to \infty$, and it equals $g(x) = \int f(x+y)d\mu(y)$. Moreover, We will prove that g_n converges to g uniformly, i.e. strongly in $C[-\infty,\infty]$.

We claim that the family (g_n) is bounded by $\|f\|$ and is equicontinuous on $[-\infty,\infty]$, so that the assumptions of the Arzela-Ascoli Theorem (Theorem C.2) are satisfied. In this case, the family (g_n) will be uniformly bounded, and the result follows: toward contradiction suppose $g_n \nrightarrow g$ strongly, and choose a subsequence that stays at some distance $\epsilon > 0$ from g.

[1]An *improper* distribution is a distribution that is not a probability distribution.

By the Arzela-Ascoli Theorem, there exists a subsequence of our subsequence that converges uniformly to some $\tilde{g} \in C([-\infty, \infty])$. By uniqueness of limit of g_n, this subsequence must also converge (pointwise) to g, implying that $\tilde{g} = g$, a contradiction.

It remains to prove the claim, that is, that (g_n) is bounded by $\|f\|$ and is equicontinuous on $[-\infty, \infty]$. Since $f \in C([-\infty, \infty])$, for a given $\epsilon > 0$ there exists $\delta > 0$ such that for $y, y' \in \mathbb{R}$, $|f(y) - f(y')| < \epsilon$ provided $|y - y'| < \delta$. Hence, for any $x \in \mathbb{R}$ and $|h| \leq \delta$, we also have that

$$|g_n(x+h) - g_n(x)| \leq \int |f(x+h+y) - f(x+y)| d\mu_n(y) < \epsilon,$$

since μ_n is a probability measure. Since this holds for all n, this proves that g_n is equicontinuous at $x \in \mathbb{R}$. To prove it is equicontinuous at ∞, we first take $t > 0$ and define $f_k \in C([-\infty, \infty]), k \geq 1$ as $f_k(x) = \frac{1}{1+k\max\{t-x,0\}}$. Then, $\lim_{k\to\infty} f_k(x) = \mathbb{1}_{[t,\infty)}(x), x \in \mathbb{R}$. Hence

$$\limsup_{n\to\infty} \mu_n[t, \infty) \leq \lim_{n\to\infty} \int x_k d\mu_n = \int x_k d\mu_n \underset{k\to\infty}{\to} \mu[t, \infty).$$

This implies that given an $\epsilon > 0$ we may choose a $t > 0$ so that $\mu_n[t, \infty) < \epsilon$, for sufficiently large n. Since such a t may be chosen for each $n \geq 1$ individually, as well, and x belongs to $C([-\infty, \infty])$, we may choose a t so that $\mu_n[t, \infty) < \epsilon$ for all $n \geq 1$ and $|f(x) - f(\infty)| < \epsilon$, for $x > t$. Now, for $x > 2t$,

$$|g_n(\infty) - g_n(x)| \leq \int_{y\leq t} + \int_{y>t} |f(\infty) - f(x+y)| d\mu_n(y) \leq \epsilon + 2\|f\|\epsilon$$

proving that y_n are equicontinuous at ∞. The case of $-\infty$ is treated in the same way. □

5 Central Limit Theorem

5.1 Sum of Normal Random Variables

Recall that f has a standard normal distribution if $\frac{d\mu_f}{dm}(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. It turns out that if g has a standard normal distribution and is independent of f, then $f + g$ has a normal distribution (no longer standard, meaning that $\frac{d\mu_{f+g}}{dm}(x)$ will differ by constants in some fashion).

Example 5.1. We compute μ_{f+g}. By Corollary 3.5, $\mu_{f+g} = \mu_f * \mu_g$.

$$\begin{aligned}
\mu_{f+g}(A) = (\mu_f * \mu_g)(A) &= \int\int \mathbb{1}_A(y+z) d\mu_f(y) d\mu_g(z) \\
&= \int\int \mathbb{1}_A(y+z) \frac{1}{2\pi} e^{-y^2/2} e^{-z^2/2} dy dz \\
&= \int\int \mathbb{1}_A(u) \frac{1}{2\pi} e^{-(u-v)^2/2} e^{-v^2/2} dv du \qquad (*)
\end{aligned}$$

by the change of variable $y + z = u, z = v$. Expand the exponent and complete the square:

$$\frac{-(u-v)^2}{2} - \frac{v^2}{2} = -v^2 - \frac{u^2}{2} + uv = -\left(v - \frac{u}{2}\right)^2 - \frac{u^2}{4}$$

In general, change of variable implies that the following equality holds

$$\int \frac{1}{\sqrt{2\pi}} e^{-(x-\mu)^2} dx = \int \frac{1}{\sqrt{2}\sqrt{2\pi}} e^{-y^2/2} dy = \frac{1}{\sqrt{2}}.$$

Hence (*) becomes

$$\int\int \mathbb{1}_A(u) \frac{1}{2\pi} e^{-\left(v-\frac{u}{2}\right)^2 - \frac{u^2}{4}} dv du = \int \mathbb{1}_A(u) \frac{1}{\sqrt{2}\sqrt{2\pi}} e^{-u^2/4} du.$$

Hence $\frac{d\mu_{f+g}}{dm}(x) = \frac{1}{\sqrt{2}\sqrt{2\pi}} e^{-x^2/4}$. In fact, we say a function h who has the density

$$\frac{d\mu_h}{dm}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}$$

is a normal random variable with mean μ and variance σ^2. We have therefore shown that $f+g$ is a normal random variable with mean 0 and variance 2. One can show by induction that if $f_1, \ldots, f_n$ is a sequence of independent standard normal random variables, then $f_1 + f_2 + \cdots + f_n$ has the normal distribution with mean 0 and variance n. In an even easier fashion, one may show af_1 has the normal distribution with mean 0 and variance a^2. Consequently,

$$\frac{f_1}{\sqrt{n}} + \cdots + \frac{f_n}{\sqrt{n}}$$

has the standard normal distribution.

5.2 Operator Composition

Proposition 5.2. $(T_\mu T_\nu \phi)(x) = T_{\mu*\nu}(\phi)$. The map $\mu \mapsto T_\mu$ is a homomorphism of two Banach algebras. $\mu \mapsto T_\mu$ changes convolution into operator composition.

Proof.

$$\begin{aligned}(T_\mu T_\nu \phi)(x) &= \int (T_\nu\phi)(x+y)\mu(dy) = \int\int \phi(x+y+z)\nu(dz)\mu(dy) \\ &= \int \phi(x+t)(\mu*\nu)(dt) = T_{\mu*\nu}(x)\end{aligned}$$

□

Since for each function we have the measure μ_f, we can assign f to the operator $T_f = T_{\mu_f}$. By Corollary 3.5 $\mu_{f+g} = \mu_f * \mu_g$, so Proposition 5.2 implies $T_{f+g} = T_f T_g$.

Corollary 5.3. If f and $f_1, \ldots, f_n$ are independent standard normal random variables, then $T_f = T_{\frac{1}{\sqrt{n}}\sum_{i=1}^n f_i} = \left(T_{\frac{1}{\sqrt{n}} f_1}\right)^{\circ n}$, where $(\cdot)^{\circ n}$ means the $n-$fold composition.

Proof. The first equality comes from Example 5.1 and Proposition 2.4, and the second comes from Proposition 5.2. □

The following lemma gives us a bound for the norm of the composition of operators.

Lemma 5.4. Let S_i, T_i be linear operators in a Banach space and let $M = \max_{i=1,\ldots,n}\{\|S_i\|, \|T_i\|\}$. Then

$$\|S_n S_{n-1} \cdots S_1 - T_n T_{n-1} \cdots T_1\| \leq M^{n-1} \sum_{i=1}^{n} \|T_i - S_i\|.$$

In particular, for any S and T and $M = \max\{\|S\|, \|T\|\}$,

$$\|S^n - T^n\| \leq M^{n-1} n \|S - T\|.$$

Proof. We show this by induction. It clearly holds for $n = 1$. Assume it holds for $n = k$, that is,

$$\|S_k S_{k-1} \cdots S_1 - T_k T_{k-1} \cdots T_1\| \leq M^{k-1} \sum_{i=1}^{k} \|T_i - S_i\|.$$

Define $A_k = S_k S_{k-1} \cdots S_1$ and $B_k = T_k T_{k-1} \cdots T_1$. Then

$$\begin{aligned}
\|A_{k+1} - B_{k+1}\| &= \|S_{k+1}A_k - T_{k+1}B_k\| = \|(S_{k+1}A_k - T_{k+1}A_k) + (T_{k+1}A_k - T_{k+1}B_k)\| \\
&\leq \|S_{k+1}A_k - T_{k+1}A_k\| + \|T_{k+1}A_k - T_{k+1}B_k\| \\
&\leq \|A_k\|\|S_{k+1} - T_{k+1}\| + \|T_{k+1}\|\|A_k - B_k\| \\
&\leq \|A_k\|\|S_{k+1} - T_{k+1}\| + \|T_{k+1}\| M^{k-1} \sum_{i=1}^{k} \|T_i - S_i\| \\
&\leq M^k \|S_{k+1} - T_{k+1}\| + M M^{k-1} \sum_{i=1}^{k} \|T_i - S_i\| \\
&= M^k \sum_{i=1}^{k+1} \|T_i - S_i\|.
\end{aligned}$$

We could have used two different $M's$, one being the max up to k and the other up to $k+1$, but the proof is clear. Note that we used the operator composition inequality $\|S \circ T\| \leq \|S\|\|T\|$. □

5.3 Proof of Central Limit Theorem

Lemma 5.5. Let (X, Ω, μ) be a probability space and $f \in L^2(X)$. Let $\int f d\mu = 0$ and $\int f^2 d\mu = 1$. Also, let a_n be a sequence of positive numbers such that $\lim_{n\to\infty} a_n = 0$. Then, for any $\phi \in C^2[-\infty, \infty]$, the space of continuous twice differentiable functions on $[-\infty, \infty]$ with continuous derivatives, the limit of $\frac{1}{a_n^2}(T_{a_n f}\phi - \phi)$ exists and does not depend on f. In fact, it equals $\frac{1}{2}\phi''$.

Proof. By the Taylor formula, for a twice differentiable ϕ and real numbers x and y,

$$\phi(x+y) = \phi(x) + y\phi'(x) + \frac{y^2}{2}\phi''(x + \theta y),$$

for some $0 \le \theta \le 1$ depending on x, y and ϕ. Thus,

$$\begin{aligned}
\frac{1}{a_n^2}[T_{a_n f}\phi(x) - \phi(x)] &= \frac{1}{a_n^2}\left[\int \phi(x+y)d\mu_{a_n f}(y) - \phi(x)\right] \\
&= \frac{1}{a_n^2}\left[\int \phi(x+a_n f(y))d\mu(y) - \phi(x)\right] \\
&= \frac{1}{a_n^2}\int \left[\phi(x+a_n f(y)) - \phi(x)\right] d\mu(y) \\
&= \frac{1}{a_n^2}\int \left[a_n f(y)\phi'(x) + \frac{a_n^2 f^2(y)}{2}\phi''(x+\theta a_n f(y))\right] d\mu(y) \\
&= \int \frac{f^2(y)}{2}\phi''(x+\theta a_n f(y))d\mu(y)
\end{aligned}$$

Fix $\epsilon > 0$. By the above equality, we would like to show

$$\left|\frac{1}{a_n^2}[T_{a_n f}\phi(x) - \phi(x)] - \frac{1}{2}\phi''(x)\right| = \left|\int \frac{f^2(y)}{2}\phi''(x+\theta a_n f(y))d\mu(y) - \frac{1}{2}\phi''(x)\right| \to 0.(*)$$

Since $\phi'' \in C[-\infty, \infty]$, for fixed $\epsilon > 0$ we can choose δ such that $|y| < \delta$ implies $|\phi''(x+y) - \phi''(y)| < \epsilon$. Break the integral into the set where $|f| \ge \frac{\delta}{a_n}$ and where $|f| < \frac{\delta}{a_n}$. Note that $\|f\|_2^2 = 1$ implies $\frac{1}{2}\phi''(x) = \int \frac{f^2(y)}{2}\phi''(x)d\mu$. Hence

$$\begin{aligned}
\left|\int_{|f|<\frac{\delta}{a_n}} \frac{f^2(y)}{2}\phi''(x+\theta a_n f(y))d\mu(y) - \frac{1}{2}\phi''(x)\right| &\le \int_{|f|<\frac{\delta}{a_n}} \frac{f^2(y)}{2}\underbrace{\left|\phi''(x+\theta a_n f(y)) - \frac{1}{2}\phi''(x)\right|}_{<\epsilon} d\mu(y) \\
&< \epsilon \int \frac{f^2(y)}{2} d\mu(y) \\
&\le \frac{\epsilon}{2},
\end{aligned}$$

where we subtly used $0 \le \theta \le 1$ so that $|\theta a_n f(y)| \le |f(y)| < \delta$ on the set $|f| < \frac{\delta}{a_n}$.

For the opposite, simply note

$$\int_{|f|\ge\frac{\delta}{a_n}} \frac{f^2(y)}{2}\phi''(x+\theta a_n f(y))d\mu(y) \le \|\phi''\| \int_{|f|\ge\frac{\delta}{a_n}} \frac{f^2(y)}{2} d\mu(y) \to 0$$

by the monotone convergence theorem since $\mu(f = \infty) = 0$ from $\|f\|_2^2 = 1$.

Putting it all together, we have shown the left hand side of (*) is less than a sequence converging to 0 plus $\epsilon/2$, with no x appearing. Hence taking the supremum over all x and $n \to 0$ shows the result. □

Theorem 5.6. Central Limit Theorem. If $f_n, n \in \mathbb{N}$ is a sequence of independent random variables with the same distribution with $m := \int f_n d\mu$ and $\sigma^2 := \int f_n^2 d\mu - m^2$, then the distribution of

$$\frac{1}{\sqrt{n\sigma^2}}\sum_{k=1}^{n}(f_k - m)$$

converges weak* to the standard normal distribution.

Proof. By Lemma 4.4, it is enough to show convergence of the corresponding operators in the strong topology. Without loss of generality, assume $m = 0$ and $\sigma^2 = 1$ (otherwise we consider the function $\frac{f-m}{\sigma}$ which will satisfy these properties). Let $T_n = T_{\frac{1}{\sqrt{n}f}}$ where f is any of the f_i (the choice is irrelevant since they have the same distribution). By the independence assumption, Proposition 5.2 implies $T_{\frac{1}{\sqrt{n}}\sum_{k=1}^n f_k} = (T_n)^{\circ n}$, so we need to show that $(T_n)^{\circ n}$ converges strongly to T_g, where g has the standard normal distribution. Note that the set $C^2[-\infty,\infty]$ is dense in $C[-\infty,\infty]$, hence it sufficies to show convergence for $\phi \in C^2[-\infty,\infty]$. Now by Corollary 5.3 and Lemma 5.4,

$$\begin{aligned}\|(T_n)^{\circ n}\phi - T_g\phi\| &= \|(T_n)^{\circ n}\phi - (T_{\frac{1}{\sqrt{n}}g})^{\circ n}\phi\| \leq n\|T_n\phi - T_{\frac{1}{\sqrt{n}}g}\phi\| \\ &\leq \|n\left(T_{\frac{1}{\sqrt{n}}f}\phi - \phi\right) - n\left(T_{\frac{1}{\sqrt{n}}g}\phi - \phi\right)\|,\end{aligned}$$

which, by adding and subtracting $\frac{1}{2}\phi''$, using triangle inequality, and applying Lemma 5.5 (with $a_n = \frac{1}{\sqrt{n}}$), goes to 0 as $n \to \infty$. □

A Measure Theory Results

Lemma A.1. Let μ and ν be two finite Borel measures on $\mathbb{R}$ and assume that $\int f d\mu = \int f d\nu$ for every bounded continuous function f. Then $\mu = \nu$.

Proof. It suffices to show $\mu(a,b] = \nu(a,b], a < b \in \mathbb{R}$, since half open sets generate the $\sigma-$algebra. Consider $f_t = \frac{1}{t}\mathbb{1}_{[0,t)} * \mathbb{1}_{(a,b]}, t \geq 0$. Since

$$f_t(x) = \frac{1}{t}\int_{-\infty}^{x} \mathbb{1}_{[0,t)}(x-y)\mathbb{1}_{(a,b]}(y)dy, \tag{*}$$

then $|f_t(x)| \leq 1$ and $|f_t(x+y) - f_t(x)| \leq \frac{y}{t}$, so that f_t is bounded and continuous. Hence by assumption

$$\int x_t d\mu = \int x_t d\nu. \tag{**}$$

If $x \leq a$, then (*) implies $f_t(x) = 0$. If $x > b$, we write

$$f_t(x) = \frac{1}{t}\int_{a}^{b} \mathbb{1}_{[0,t)}(x-y)dy = \frac{1}{t}\int_{x-b}^{x-a} \mathbb{1}_{[0,t)}(y)dy$$

to see that $f_t(x) = 0$, if $x - b > t$. Finally, if $a < x \leq b, f_t(x) = \frac{1}{t}\int_0^t dy = 1$, for $t < x - a$. Consequently, $\lim_{t\to 0} f_t(x) = \mathbb{1}_{(a,b]}(x), x \in \mathbb{R}$. By the Lebesgue Dominated Convergence Theorem, we may let $t \to 0$ in (**) to obtain $\mu(a,b] = \int \mathbb{1}_{(a,b]}d\mu = \int \mathbb{1}_{(a,b]}d\nu = \nu(a,b]$. □

Definition A.2. A function $f : [a,b] \to \mathbb{R}$ is said to be of bounded variation on $[a,b]$ if and only if there is a constant $M > 0$ such that

$$\sum_{i=1}^{n} |f(x_i) - f(x_{i-1})| \leq M$$

for all partitions $\pi = \{x_0, x_1, \ldots, x_n\}$ of $[a,b]$.

Theorem A.3. (Jordan Decomposotion) If f is of bounded variation, then f can be written as a difference of two monotone increasing functions, that is, $f = g - h$ for g, h monotone increasing.

Proof. See [2] (Provides a more detailed result than the one above). □

Theorem A.4. (Froda's theorem) Let f be a real-valued monotone function on an interval I. Then the set of discontinuities of the first kind is at most countable.

Proof. See [3]. □

Corollary A.5. If $f : [a,b] \to \mathbb{R}$ is of bounded variation, then f is Borel measurable.

Proof. By the Jordan Decomposotion, $f = g - h$. Then g and h are monotonically increasing and by Froda's theorem have countably many discontinuities. Hence they are Borel measurable. □

B $\pi - \lambda$ Theorem

This section is taken mostly from [4] Probability and Stochastics by Erhan Çinlar, my favorite reference on probability theory. Simply googling "π λ theorem" should yield similar results.

Definition B.1. A collection $\mathcal{C}$ of subsets of a set E is called a $\pi-$system if it closed under intersections. A collection $\mathcal{D}$ of subsets of E is called a $\lambda-$system (sometimes called a Dynkin system) if

a) $E \in \mathcal{D}$

b) $A, B \in \mathcal{D}$ and $A \supset B$ implies $A \backslash B \in \mathcal{D}$,

c) $(A_n) \subset \mathcal{D}$ and $A_n \nearrow A$ implies $A \subset \mathcal{D}$,

where $(A_n) \subset \mathcal{D}$ means (A_n) is a sequence of elements of $\mathcal{D}$ and $A_n \nearrow A$ means the sequence is increasing to A in the following sense:

$$A_1 \subset A_2 \subset \dots, \quad \bigcup_n A_n = A.$$

It is obvious that a $\sigma-$algebra is both a $\pi-$system and a $\lambda-$system. The converse is also true, and it is not too difficult to show. The proof is left as an exercise.

Proposition B.2. A collection of subsets of a set E is a $\sigma-$algebra if and only if it is both a $\pi-$system and a $\lambda-$system on E.

The next lemma is in preparation for the main theorem of this section. Its proof is left as an exercise in checking the $\lambda-$system conditions one by one.

Lemma B.3. Let $\mathcal{D}$ be a $\lambda-$system on E. Fix $D \in \mathcal{D}$ and let

$$\hat{\mathcal{D}} = \{A \in \mathcal{D} : A \cap D \in \mathcal{D}\}.$$

Then $\hat{\mathcal{D}}$ is again a $\lambda-$system.

Notation. For a collection of subsets of a set E, we denote $\sigma\mathcal{C}$ as the smallest $\sigma-$algebra containing $\mathcal{C}$, that is, if $\mathcal{C}_1, \mathcal{C}_2, \dots,$ are all $\sigma-$algebras containing $\mathcal{C}$, then

$$\sigma\mathcal{C} = \bigcap_n \mathcal{C}_n.$$

The following theorem is the main result of this section, sometimes called the $\pi - \lambda$ theorem.

Theorem B.4. *Monotone Class Theorem.* If a $\lambda-$system contains a $\pi-$system, then it also contains the $\sigma-$algebra generated by that $\pi-$system.

Proof. Let $\mathcal{C}$ be a $\pi-$system. Let $\mathcal{D}$ be the smallest $\lambda-$system on E that contains $\mathcal{C}$, that is, $\mathcal{D}$ is the intersection of all $\lambda-$systems containing $\mathcal{C}$. The claim is that $\mathcal{D} \supset \sigma\mathcal{C}$. To show it, since $\sigma\mathcal{C}$ is the smallest $\sigma-$algebra containing $\mathcal{C}$, it is enough to show that $\mathcal{D}$ is a $\sigma-$algebra. In view of Proposition B.2, it is thus enough to show that the $\lambda-$system $\mathcal{D}$ is also a $\pi-$system.

To that end, fix $B \in \mathcal{C}$ and let

$$\mathcal{D}_1 = \{A \in \mathcal{D} : A \cap B \in \mathcal{D}\}.$$

Since $\mathcal{C}$ is contained in $\mathcal{D}$, the set B is in $\mathcal{D}$; and Lemma B.3 implies that $\mathcal{D}_1$ is a $\lambda-$system. It also contains $\mathcal{C}$: if $A \in \mathcal{C}$ then $A \cap B \in \mathcal{C}$ since B is in $\mathcal{C}$ and $\mathcal{C}$ is a $\pi-$system. Hence, $\mathcal{D}_1$ must contain the smallest $\lambda-$system containing $\mathcal{C}$, that is, $\mathcal{D}_1 \supset \mathcal{D}$. In other words, $A \cap B \in \mathcal{D}$ for every $A \in \mathcal{D}$ and $B \in \mathcal{C}$.

Consequently, for fixed A in $\mathcal{D}$, the collection

$$\mathcal{D}_2 = \{B \in \mathcal{D} : A \cap B \in \mathcal{D}\}$$

contains $\mathcal{C}$. By Lemma B.3, $\mathcal{D}_2$ is a $\lambda-$system. Thus, $\mathcal{D}_2$ must contain $\mathcal{D}$. In other words, $A \cap B \in \mathcal{D}$ whenever A and B are in $\mathcal{D}$, that is, $\mathcal{D}$ is a $\pi-$system.

□

This result is surprisingly powerful; the route to proving existence and uniqueness of the Lebesgue measure provided in most probability theory texts involve use of the Monotone Class Theorem. There are also several other applications such as the following. *(The following proposition is irrelevant to the rest of the text but is an interesting application of the Monotone Class Theorem.)*

Proposition B.5. Every set $A \in \mathcal{B}(\mathbb{R}^n)$ has the following property: for every probability measure $\mathbb{P}$ on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ and for every $\epsilon > 0$, there is a closed set F and an open set G such that $F \subset A \subset G$ and $\mathbb{P}(G \backslash F) < \epsilon$.

Proof. Let $\mathcal{F}$ be the collection of all sets satisfying the described property. We will perform the following: (i) Show $\mathcal{F}$ is a $\lambda-$system. (ii) Show $\mathcal{F}$ contains all of the closed sets. (iii) If $\mathcal{C}$ is the collection of all closed subsets of $\mathbb{R}^n$, then $\mathcal{C}$ is a $\pi-$system and $\sigma\mathcal{C} = \mathcal{B}(\mathbb{R}^n)$. (iv) $\mathcal{F}$ is a $\lambda-$system and contains $\mathcal{C}$ which is a $\pi-$system, so the Monotone Class Theorem implies that $\mathcal{F} \supset \mathcal{B}(\mathbb{R}^n)$. Therefore all $A \in \mathcal{B}(\mathbb{R}^n)$ have the desired property.

(i) $\emptyset \in \mathcal{F}$ is obvious. To show $\mathcal{F}$ is closed under complementation, suppose $A \in \mathcal{F}$, so for each $\epsilon > 0$, we have a cloased set F and an open set G such that $F \subset A \subset G$ and $\mathbb{P}(G \backslash F) < \epsilon$. But then F^c is open, G^c is closed, $G^c \subset A^c \subset F^c$, and $\mathbb{P}(F^c \backslash G^c) = \mathbb{P}(G \backslash F) < \epsilon$; therefore $A^c \subset \mathcal{F}$. To show $\mathcal{F}$ is closed under countable unions, let $A = \bigcup_{k=1}^{\infty}$ where $A_k \in \mathcal{F}$ for each k. For each $\epsilon > 0$, there is a corresponding sequence of closed sets (F_k) and sequence of open sets (G_k) such that $F_k \subset A_k \subset G_k$ and $\mathbb{P}(G_k \backslash F_k) < \epsilon/2^{k+1}, k = 1, 2, \ldots$. Let $G = \bigcup_{k=1}^{\infty} G_k$ and $F = \bigcup_{k=1}^{m} F_k$, where m is chosen such that $\mathbb{P}(\bigcup_{k=1}^{\infty} F_k \backslash \bigcup_{k=1}^{m} F_k) < \epsilon/2$. Then G is open, F is closed, $F \subset A \subset G$, and

$$\mathbb{P}(G \backslash F) \leq \mathbb{P}\left(G \backslash \bigcup_{k=1}^{\infty} F_k\right) + \mathbb{P}\left((\bigcup_{k=1}^{\infty} F_k \backslash F\right) < \sum_{k=1}^{\infty} \frac{\epsilon}{2^{k+1}} + \frac{\epsilon}{2} = \epsilon.$$

Therefore $\mathcal{F}$ is a $\sigma-$algebra and by Proposition B.2 it is a $\lambda-$system.

(ii) Now choose a closed set F. Let $G_k = \{x \in \mathbb{R}^n : \|x - y\| < 1/l$, for some $y \in F\}$. Then each G_k is open, $G_1 \supset G_2 \supset \dots$, and $\bigcap_{k=1}^{\infty} G_k = F$. Thus, for each $\epsilon > 0$, there exists an m such that $\mathbb{P}(G_m \backslash F) < \epsilon$. It follows that $F \in \mathcal{F}$.

By statements (iii) and (iv) above, the proof is complete. □

C Arzela-Ascoli Theorem

Most of this section comes from [5] A Course in Functional Analysis by John Conway . Googling "Arzela-Ascoli Theorem" should yield similar results.

Definition C.1. If X is completely regular and $\mathcal{F} \subset C(X)$, then $\mathcal{F}$ is **equicontinuous** if for every $\epsilon > 0$ and every $x_0 \in X$ there is a neighborhood U of x_0 such that $|f(x) - f(x_0)| < \epsilon$ for all $x \in U$ and for all $f \in \mathcal{F}$.

Note that for a single function $f \in C(X), \mathcal{F} = \{f\}$ is equicontinuous. The concept of equicontinuity states that one neighborhood works for all $f \in \mathcal{F}$.

Recall that a metric space X is bounded if there exists some number r such that $d(x, y) \leq r$ for all $x, y \in X$. The space X is totally bounded if for all $r > 0$ there exists finitely many open balls of radius r whose union covers X. A consequence is that every totally bounded space is bounded.

Theorem C.2. The Arzela–Ascoli Theorem. If X is compact and $\mathcal{F} \subset C(X)$, then $\mathcal{F}$ is totally bounded iff $\mathcal{F}$ is bounded and equicontinuous.

Proof. Suppose $\mathcal{F}$ is totally bounded. Then $\mathcal{F}$ is bounded. To show equicontinuity, if $\epsilon > 0$, then there are $f_1, \dots, f_n \in \mathcal{F}$ such that $\mathcal{F} \subset \bigcup_{k=1}^{n} \{f \in C(X) : \|f - f_k\| < \epsilon/3\}$. If $x_0 \in X$, let U be an open neighborhood of x_0 such that for $1 \leq k \leq n$ and $x \in U, |f_k(x) - f_k(x_0)| < \epsilon/3$. If $f \in \mathcal{F}$, let f_k such that $\|f - f_k\| < \epsilon/3$. Then for $x \in U$,

$$\begin{aligned} |f(x) - f(x_0)| &\leq |f(x) - f_k(x)| + |f_k(x) - f_k(x_0)| + |f_k(x_0) - f(x_0)| \\ &< \epsilon. \end{aligned}$$

Hence $\mathcal{F}$ is equicontinuous.

Now assume that $\mathcal{F}$ is equicontinuous and $\mathcal{F} \subset \text{ball}C(X)$. Let $\epsilon > 0$. For each $x \in X$, let U_x be an open neighborhood of x such that $|f(x) - f(y)| < \epsilon/3$ for $f \in \mathcal{F}$ and $y \in U_x$. Now $\{U_x : x \in X\}$ is an open covering of X. Since X is compact, there are points $x_1, \dots, x_n \in X$ such that $X = \bigcup_{j=1}^{n} U_x$.

Let $\{\alpha_1, \dots, \alpha_n\} \subset \{z \in \mathbb{F} : |z| < 1\}$ such that $\overline{\{z \in \mathbb{F} : |z| < 1\}} \subset \bigcup_{k=1}^{m} \{\alpha : |\alpha - \alpha_k| < \epsilon/6\}$. Consider the collection B of those ordered $n-$tuples $b = (\beta_1, \dots, \beta_n)$ for which there is a function $f_b \in \mathcal{F}$ such that $|f_b(x_j) - \beta_j) < \epsilon/6$ for $1 \leq j \leq n$. Note that B is not empty since $f(x) \subset \overline{\{z \in \mathbb{F} : |z| < 1\}}$ for every $f \in \mathcal{F}$. In fact, each $f \in \mathcal{F}$ gives rise to such a $b \in B$. Moreover B is finite. Fix one function $f_b \in \mathcal{F}$ associated as above with $b \in B$. It is enough to show that $\mathcal{F} \subset \bigcup_{b \in B} \{f : \|f - f_b\| < \epsilon\}$, since this would imply that $\mathcal{F}$ is totally bounded.

If $f \in \mathcal{F}$, there is a $b \in B$ such that $|f(x_j) - f_b(x_j)| < \epsilon/3$ for $1 \le j \le n$. Therefore if $x \in X$, let x_j be chosen such that $x \in U_{x_j}$. Thus $|f(x) - f_b(x)| < |f(x) - f(x_j)| + |f(x_j) - f_b(x_j)| + |f_b(x_j) - f_b(x)| < \epsilon$. Since x was arbitrary, $\|f - f_b\| < \epsilon$. □

References

[1] A. Bobrowski, *Functional analysis for probability and stochastic processes: an introduction.* Cambridge University Press, 2005.

[2] S. Bianchini and D. Tonon, “A decomposition theorem for bv functions,” *Commun. Pure Appl. Anal.*, vol. 10, no. 6, pp. 1549–1566, 2011.

[3] Wikipedia, “Froda’s theorem,” 11 April 2014. [Online; accessed 5-15-2015].

[4] Ç. Erhan *et al.*, *Probability and stochastics*, vol. 261. Springer Science & Business Media, 2011.

[5] J. B. Conway, *A course in functional analysis*, vol. 96. Springer Science & Business Media, 1990.