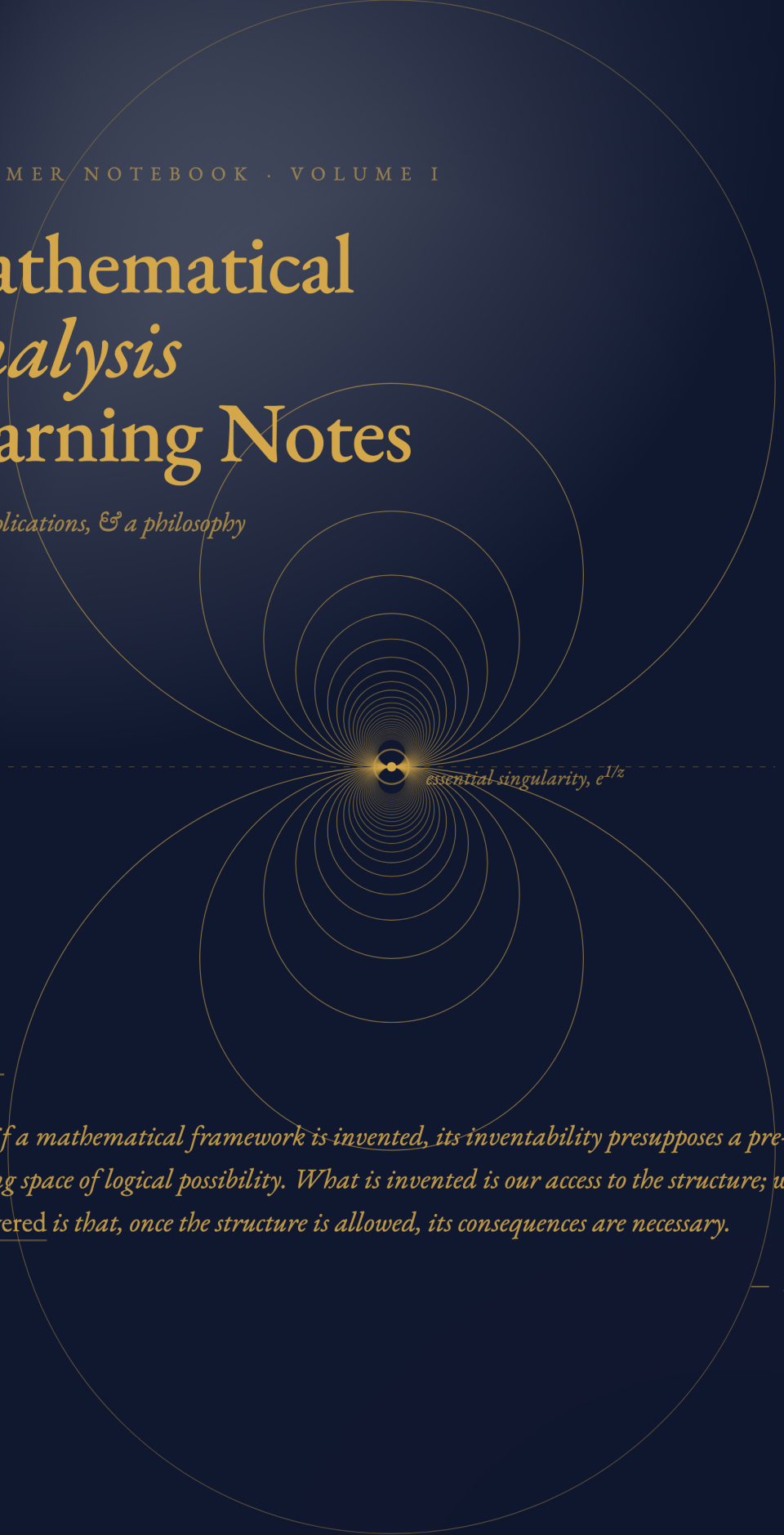
A SUMMER NOTEBOOK · VOLUME I

Mathematical *Analysis* Learning Notes

on its applications, & a philosophy



Even if a mathematical framework is invented, its inventability presupposes a pre-existing space of logical possibility. What is invented is our access to the structure; what is discovered *is that, once the structure is allowed, its consequences are necessary.*

— X. L.

Xuran Lyu

SUMMER 2026

Author's Preface

Xuran Lyu

These notes are a record of learning as it happens: definitions I make myself state precisely, examples I rework until they stop feeling strange, questions I keep around long enough that they eventually turn into structure. I am writing them as a companion for a summer of analysis, topology, measure theory, functional analysis, elliptic PDE, complex analysis, and the philosophy around mathematical practice.

The aim is not only to collect results but to keep the path toward them: what each idea is trying to solve, why its formal language ends up feeling natural, and how distant parts of mathematics turn out to talk to one another. So the pages that follow are part reference and part notebook — tidied up where I had the time, still rough where I was only feeling my way.

My Story of Learning Mathematics. I majored in mathematics, philosophy, and economics as an undergraduate. Philosophy of mathematics, metaphysics, epistemology, and philosophy of science pushed me toward analysis because they taught me to ask what a statement really means, what counts as a rigorous argument, and how hidden assumptions shape a theory. Philosophy of science especially trained me to think more rigorously about models, evidence, explanation, and the structure of a discipline.

A philosophical instinct. *My own instinct is somewhere between mathematical Platonism and structuralism. Even if a mathematical framework is invented, its inventability presupposes a pre-existing space of logical possibility. What is invented is our access to the structure; what is discovered is that, once the structure is allowed, its consequences are necessary. Here "allowed" means logically possible, or possibly true. Once the structure is fixed, its consequences are true no matter what within that structure.*

I took point-set topology in Fall 2023 with Professor Nikolai Saveliev at the University of Miami, and real analysis I in Fall 2024 with Professor Ilie Grigorescu at the University of Miami. In Fall

2025 at Duke, I took complex analysis with Professor Calvin McPhail-Snyder. I also took ODE theory and applied PDE in Fall 2025, but I was overly ambitious. That experience is part of why I am writing these notes: I want a record of my reflections, mistakes, and improved approach to mathematics.

In Spring 2026, I attempted to take functional analysis with Professor Mark Stern at Duke, but I was afraid that I could end up with a horrible grade. So I decided to self-study the subject more carefully. These teachers still had an impact on me: not only through the material, but through the way they approached mathematics and the way they seemed to enjoy it.

I am currently learning measure theory, functional analysis, and PDE analysis and plan to take them in the Fall 2026 semester. I plan to apply to mathematics PhD programs and operations research PhD programs in Fall 2026. Wish me luck.

Why Analysis Feels Worth It. I believe mathematics is one of the keys to the universe, in the spirit of Roger Penrose's *The Road to Reality*. It is also a practical language for operations research, economics, quantitative trading, and quantitative research. Analysis is difficult because it takes time to adapt to its language, but I am enjoying that adaptation. The precision is part of the beauty.

Complex analysis made this especially vivid for me. Ideas such as the Riemann surface, the Riemann sphere, essential singularities, and the major theorems of the subject were simply appealing. I was also astonished to learn that contour integration is not just a beautiful formal technique: different contour methods help solve ODEs, PDEs, and even problems in financial mathematics, such as the Carr–Madan formula I met while taking a course on financial derivatives. After complex analysis, and while learning functional analysis and operator theory, quantum mechanics also began to feel more approachable. Functional analysis gave me a similar feeling with time series analysis: once functions, spaces, projections, operators, and convergence become more natural, the subject starts looking much less mysterious. If someone has no feeling for such ideas, then they have not yet found the gut-level part of mathematics.

Books and Channels That Shaped These Notes. The Rudin books have become landmarks for the path I am trying to walk:

- Baby Rudin: *Principles of Mathematical Analysis*.
- Papa Rudin: *Real and Complex Analysis*.
- Grandpa Rudin: *Functional Analysis*.

I was also influenced by math exposition online, especially ThatMathThing, whose videos helped keep alive the sense that mathematics has personalities, history, and taste, not only formal statements.

Xuran Lyu

Contents

1 Real Analysis

Foundations to Strengthen.

- Drill ε–δ and sequence arguments until they stop requiring thought.
- Be fluent with compactness in $\mathbb{R}^n$ in all its guises: Heine–Borel, Bolzano–Weierstrass, sequential compactness.
- Never conflate pointwise with uniform convergence; the gap between them is where the counterexamples live.
- Keep careful track of when a limit may be moved past continuity, differentiation, or integration—and when it may not.

2 Complex Analysis

Foundations to Strengthen.

- Get fluent with Cauchy's theorem, Cauchy's integral formula, and the residue theorem; nearly everything else runs through them.
- Practice contour estimates. The ML inequality is the workhorse here.
- Read off the type of a singularity from its Laurent series.
- Keep the geometric side in view: conformal maps, harmonic functions, and analytic continuation.
- Use residues to evaluate real integrals, and learn to spot when a contour argument is the right tool.

3 Topology

Foundations to Strengthen.

- Bases and subbases: the economical way to specify a topology without listing every open set.
- Read continuity as "preimages of open sets are open."
- Treat subspace, product, and quotient topologies as the three standard construction rules.
- Keep compactness, connectedness, and Hausdorffness apart; they are independent properties.
- Collect counterexamples along the way — most of topology is learned by watching a definition fail.

Definition 3.1.

Let X be a set. A *topology* on X is a collection τ of subsets of X, called open sets, such that $\varnothing, X \in \tau$, arbitrary unions of sets in τ lie in τ, and finite intersections of sets in τ lie in τ.

Checkpoint. Given two topologies on the same set, what does it mean for one to be "stronger" or "weaker" than the other?

3.1 How to Prove Something Is a Topology

Intuition. There is no need to understand every open set one at a time. To show a family of subsets is a topology, we only check that it survives the three operations the axioms demand.

The axioms themselves are not arbitrary. Open sets record stable local information about points. A union of allowed regions should stay allowed, since a point belongs to a union as soon as it belongs to one of the pieces. A finite intersection should stay allowed too: with finitely many constraints we can always shrink a neighborhood until it meets all of them at once. The empty set and the whole space are the two degenerate cases at the ends.

Proposition 3.2: Topology Proof Strategy.

Let X be a set and let $\tau \subseteq \mathcal{P}(X)$ be a collection of subsets of X. To prove that τ is a topology on X, prove:

(i) $\varnothing \in \tau$ and $X \in \tau$;

(ii) if $\{U_\alpha\}_{\alpha \in A} \subseteq \tau$, then $\bigcup_{\alpha \in A} U_\alpha \in \tau$;

(iii) if $U_1, \dots, U_n \in \tau$, then $\bigcap_{j=1}^{n} U_j \in \tau$.

Remark 3.3. A common slip is to verify closure under finite unions instead of arbitrary ones. The asymmetry is deliberate: topology allows arbitrary unions but only finite intersections, because infinitely many local restrictions, intersected together, can shrink away every bit of room around a point.

Example 3.4. On $\mathbb{R}$ with its usual topology the union step is immediate: a point in a union already lies in one member of the family, and it carries that member's small interval along with it. Intersections cost a little more, but only the minimum radius. If x has room ε_1 inside U_1 and room ε_2 inside U_2, then it has room $\min\{\varepsilon_1, \varepsilon_2\}$ inside $U_1 \cap U_2$.

■

Takeaway. Topology gives the general vocabulary for continuity and convergence.

May 25, 2026

3.2 Topological Spaces and Hausdorff Spaces

Learning Goals. Understand what a topological space is and why the open sets are the primitive data, and see why the Hausdorff condition is the separation property that makes limits behave the way they do in metric spaces.

Intuition. The open sets are the data: they record which subsets count as observable regions. Once they are fixed, neighborhoods, continuity, and convergence all follow. A Hausdorff space is one where any two distinct points can be walled off inside disjoint open neighborhoods.

Begin with a bare set X. It has points, but nothing in it tells us when one point is "near" another or what a continuous map should be. Doing analysis means adding just enough structure to declare which regions around a point are open, and that structure is the topology. Wanting limits to behave as they do in $\mathbb{R}^n$ asks for slightly more: a way to tell two distinct points apart by purely local observations. That extra demand is the Hausdorff condition.

Definition 3.5.

A *topological space* is a pair (X, τ), where X is a set and τ is a collection of subsets of X such that

(i) $\varnothing \in \tau$ and $X \in \tau$;

(ii) arbitrary unions of sets in τ are in τ;

(iii) finite intersections of sets in τ are in τ.

The elements of τ are called *open sets*.

Definition 3.6.

Let (X, τ) be a topological space. A set $U \subseteq X$ is a *neighborhood* of $x \in X$ if there exists an open set $O \in \tau$ such that

$$x \in O \subseteq U.$$

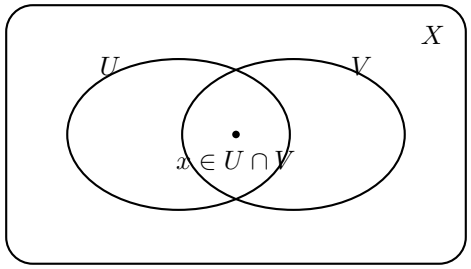
Topology does not start from distance; it starts from open sets. When a metric d happens to be available, the open balls

$$B(x, r) = \{y \in X : d(x, y) < r\}$$

generate a topology, and we recover the familiar picture. Many of the topologies we care about, though, arrive without any distance to write down first.

Example 3.7. Every metric space (X, d) is a topological space: call $U \subseteq X$ open if for every $x \in U$ there exists $r > 0$ such that $B(x, r) \subseteq U$.

■



Finite intersections of open sets are open: if $U, V \in \tau$, then $U \cap V \in \tau$.

Figure 1. Open sets are the primitive objects in a topology.

3.2.1 The Usual Topology

Intuition. The usual topology is the one calculus already uses, long before anyone names it. On $\mathbb{R}$ the basic local regions are open intervals; on $\mathbb{R}^n$ they are open balls.

The point of analysis on $\mathbb{R}$ is to say when a point has room around it. The right primitive is neither the whole line nor a lone point but a small interval centered at the point. A set counts as open precisely when every one of its points has some positive amount of such room left inside the set.

Definition 3.8.

The *usual topology* on $\mathbb{R}$ is the topology generated by all open intervals

$$(a, b) = \{x \in \mathbb{R} : a < x < b\}.$$

Equivalently, $U \subseteq \mathbb{R}$ is open in the usual topology if for every $x \in U$ there exists $\varepsilon > 0$ such that

$$(x - \varepsilon, x + \varepsilon) \subseteq U.$$

Proposition 3.9.

The usual topology on $\mathbb{R}$ is a topology.

Proof. Let

$$\tau_{\text{usual}} = \{U \subseteq \mathbb{R} : \text{for every } x \in U, \text{ there exists } \varepsilon > 0 \text{ with } (x - \varepsilon, x + \varepsilon) \subseteq U\}.$$

We check the three topology axioms.

First, $\varnothing \in \tau_{\text{usual}}$ because there is no $x \in \varnothing$ to check. Also $\mathbb{R} \in \tau_{\text{usual}}$, since for every $x \in \mathbb{R}$ we can take any $\varepsilon > 0$, and then $(x - \varepsilon, x + \varepsilon) \subseteq \mathbb{R}$.

Second, let $\{U_\alpha\}_{\alpha \in A}$ be any family of sets in τ_{usual}. If

$$x \in \bigcup_{\alpha \in A} U_\alpha,$$

then $x \in U_{\alpha_0}$ for some $\alpha_0 \in A$. Since U_{α_0} is open, there exists $\varepsilon > 0$ such that

$$(x - \varepsilon, x + \varepsilon) \subseteq U_{\alpha_0} \subseteq \bigcup_{\alpha \in A} U_\alpha.$$

Thus arbitrary unions are open.

Third, let $U, V \in \tau_{\text{usual}}$. If $x \in U \cap V$, then there exist $\varepsilon_U, \varepsilon_V > 0$ such that

$$(x - \varepsilon_U, x + \varepsilon_U) \subseteq U, \qquad (x - \varepsilon_V, x + \varepsilon_V) \subseteq V.$$

Put $\varepsilon = \min\{\varepsilon_U, \varepsilon_V\}$. Then

$$(x - \varepsilon, x + \varepsilon) \subseteq U \cap V.$$

So $U \cap V$ is open. The same argument, using the minimum of finitely many positive radii, proves that every finite intersection of usual open sets is open. Therefore τ_{usual} is a topology on $\mathbb{R}$. □

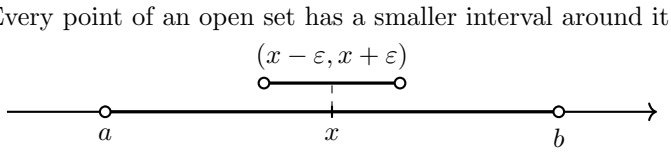


Figure 2. In the usual topology on $\mathbb{R}$, open intervals are the basic neighborhoods.

Proposition 3.10.

A subset $U \subseteq \mathbb{R}$ is open in the usual topology if and only if it is a union of open intervals.

Proof. If U is a union of open intervals, then every point $x \in U$ lies in some interval $(a, b) \subseteq U$. Choose

$$\varepsilon < \min\{x - a, b - x\}.$$

Then $(x - \varepsilon, x + \varepsilon) \subseteq (a, b) \subseteq U$, so U is open.

Conversely, if U is open, then for every $x \in U$ choose $\varepsilon_x > 0$ such that $(x - \varepsilon_x, x + \varepsilon_x) \subseteq U$. Then

$$U = \bigcup_{x \in U} (x - \varepsilon_x, x + \varepsilon_x),$$

so U is a union of open intervals. □

Example 3.11. The sets $(0, 1)$, $(0, 1) \cup (2, 5)$, $\mathbb{R}$, and $\varnothing$ are open in the usual topology on $\mathbb{R}$. ■

Example 3.12. The sets $[0, 1]$, $(0, 1]$, and $\{0\}$ are not open in the usual topology on $\mathbb{R}$. For example, $0 \in [0, 1]$, but every interval $(-\varepsilon, \varepsilon)$ around 0 contains negative numbers, so it is not contained in $[0, 1]$. ■

Definition 3.13.

The *usual topology* on $\mathbb{R}^n$ is the topology generated by Euclidean open balls

$$B(x, r) = \{y \in \mathbb{R}^n : \|x - y\|_2 < r\}, \qquad r > 0.$$

Equivalently, $U \subseteq \mathbb{R}^n$ is open if every $x \in U$ has some Euclidean ball $B(x, r)$ contained in U.

Remark 3.14. With the usual topology, the open-set definition of continuity reproduces the ε–δ definition from calculus. A map $f : \mathbb{R}^n \to \mathbb{R}^m$ is continuous exactly when the preimage of every usual open set in $\mathbb{R}^m$ is again usual open in $\mathbb{R}^n$.

Common Pitfall. Open is not the same as "having no boundary" in some loose visual sense. It means every point of the set keeps a full small neighborhood inside the set. And openness and

closedness are not opposites: a set can be neither, as $(0, 1]$ shows in the usual topology on $\mathbb{R}$.

Takeaway. The usual topology is the working topology of real analysis: intervals on $\mathbb{R}$, balls on $\mathbb{R}^n$, and limits and continuity in their familiar sense.

3.2.2 Common Types of Topologies

Intuition. Putting different topologies on the same set is a way of choosing how much information is observable. A finer topology has more open sets, so points and local behavior can be told apart more sharply; a coarser one has fewer, and the space sees less.

A useful habit when a new topology appears is to ask where its open sets came from. Sometimes a distance, sometimes an order on the set, sometimes a larger space the set sits inside, sometimes the demand that certain maps be continuous. Tracing that source usually reveals what the topology was built to measure.

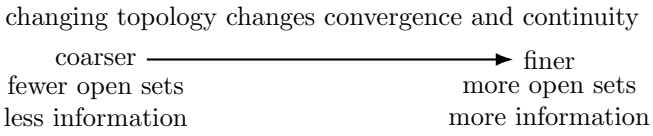


Figure 3. A topology is a choice of observational resolution.

Definition 3.15: Discrete and Indiscrete Topologies.

Let X be a set.

- The *discrete topology* is $\tau = \mathcal{P}(X)$. Every subset of X is open.
- The *indiscrete topology* is $\tau = \{\varnothing, X\}$. Only $\varnothing$ and X are open.

Remark 3.16. These two are the extremes. The discrete topology is the finest possible on X, with every subset open; the indiscrete topology is the coarsest, with nothing open but $\varnothing$ and X.

Definition 3.17: Metric Topology.

If (X, d) is a metric space, the *metric topology* is the topology whose open sets are those $U \subseteq X$ such that for every $x \in U$ there exists $r > 0$ with

$$B(x, r) = \{y \in X : d(x, y) < r\} \subseteq U.$$

The usual topology on $\mathbb{R}^n$ is the metric topology generated by the Euclidean metric.

Definition 3.18: Cofinite Topology.

Let X be a set. The *cofinite topology* on X is

$$\tau_{\text{cofinite}} = \{\varnothing\} \cup \{U \subseteq X : X \setminus U \text{ is finite}\}.$$

So the nonempty open sets are exactly the sets whose complements are finite.

Example 3.19. On an infinite set X the cofinite topology is far coarser than the discrete one: a nonempty open set is forced to be large, missing only finitely many points. Two such sets can never be disjoint, since their union would already omit only finitely many points while their complements are finite. So an infinite cofinite space fails to be Hausdorff.

■

Definition 3.20: Subspace Topology.

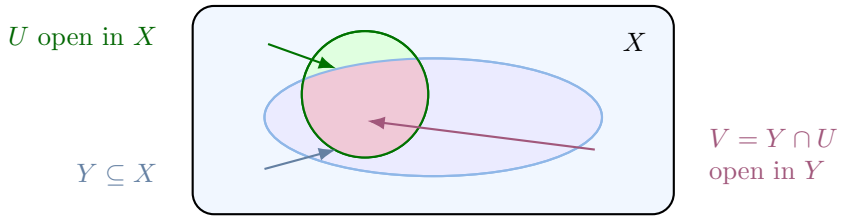
If (X, τ) is a topological space and $Y \subseteq X$, the *subspace topology* on Y is

$$\tau_Y = \{Y \cap U : U \in \tau\}.$$

The open sets of Y are just the open sets of X cut down to Y. This is what it means for Y to *inherit* a topology: nothing is invented on Y from scratch. A set $V \subseteq Y$ is open in Y exactly when

$$V = Y \cap U$$

for some open U in the ambient space X. Openness, in other words, is always relative to the space we are working in.



Y inherits open sets by intersecting ambient open sets with Y.

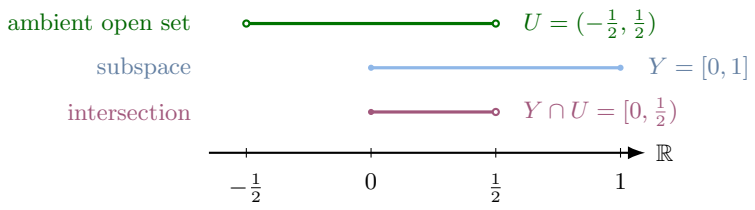
Example 3.21. Take $[0, 1]$ as a subspace of $\mathbb{R}$ in the sense of Definition 3.20. The set $[0, \frac{1}{2})$ is open in $[0, 1]$ because

$$[0, \tfrac{1}{2}) = [0, 1] \cap (-\tfrac{1}{2}, \tfrac{1}{2}),$$

and $(-\frac{1}{2}, \frac{1}{2})$ is open in the ambient space $\mathbb{R}$. That is exactly the criterion: by Definition 3.20, a set $V \subseteq [0, 1]$ is open in $[0, 1]$ if there exists an open set $U \subseteq \mathbb{R}$ such that

$$V = [0, 1] \cap U.$$

Here we take $U = (-\frac{1}{2}, \frac{1}{2})$ and get $V = [0, \frac{1}{2})$. Notice what "open in $[0, 1]$" does and does not mean: it is openness after we restrict attention to the smaller world $[0, 1]$, not openness on the whole real line. At the endpoint 0 no room to the left is required, because the points to the left of 0 never belonged to $[0, 1]$ in the first place.



even though $[0, \frac{1}{2})$ is not open in $\mathbb{R}$ with the usual topology.

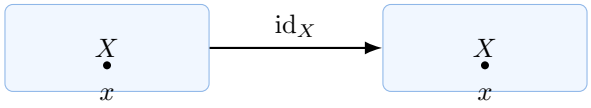
■

Definition 3.22: Identity Map.

For a set X, the *identity map* on X is the function

$$\mathrm{id}_X : X \to X, \qquad \mathrm{id}_X(x) = x.$$

Neither the point nor the surrounding space changes.



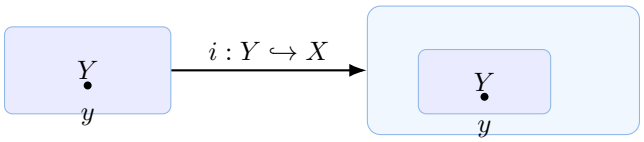
identity map: same point, same space

Definition 3.23: Inclusion Map.

If $Y \subseteq X$, the *inclusion map* is

$$i : Y \hookrightarrow X, \qquad i(y) = y.$$

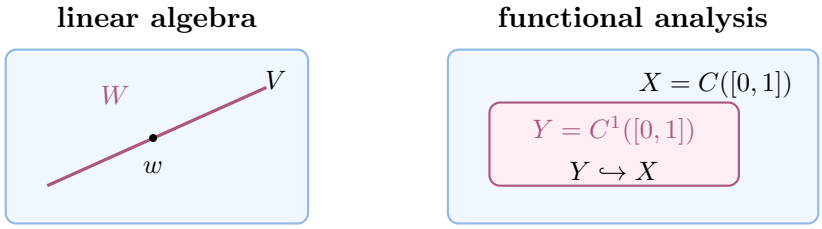
The point does not move. What changes is the space we regard it in: first y is a point of Y, then the very same y is a point of X.



inclusion map: same point, larger ambient space

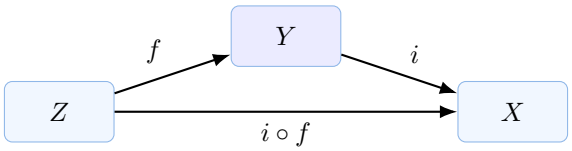
Definition 3.24: Subspaces in Linear Algebra and Functional Analysis.

A *linear subspace* $W \subseteq V$ is a subset that is itself a vector space under the operations of V. Its natural map into V is the inclusion $W \hookrightarrow V$.
The same pattern recurs in functional analysis, now with spaces whose points are functions. A smaller function space Y sits inside a larger one X, and the subspace topology lets Y inherit convergence and continuity from X without ever leaving Y.



A subspace is a smaller world inside a larger one. The elements do not change; the available ambient structure changes.

Remark 3.25: Why the Subspace Topology Matters. Most analysis takes place on smaller worlds: intervals, domains with boundary, constraint sets, function subspaces. That is why the subspace topology is used so constantly. It is the coarsest topology on Y making the inclusion $i : Y \hookrightarrow X$ continuous, and this has a clean consequence: a map $f : Z \to Y$ is continuous exactly when the composite $i \circ f : Z \to X$ is.



continuity into the subspace can be checked after viewing the same values in the ambient space

Definition 3.26: Product Topology.

If X and Y are topological spaces, the *product topology* on $X \times Y$ is generated by products

$$U \times V,$$

where $U \subseteq X$ and $V \subseteq Y$ are open. These rectangles are the basic open sets.

Definition 3.27: Quotient Topology.

Let $q : X \to Y$ be a surjective map from a topological space X onto a set Y. The *quotient topology* on Y is defined by declaring $U \subseteq Y$ open exactly when

$$q^{-1}(U)$$

is open in X.

Remark 3.28. Three constructions, three slogans. Subspace: inherit open sets from a larger space. Product: take just enough open sets to make the coordinate projections continuous. Quotient: glue points together, then keep exactly those open sets whose preimages were open before the gluing.

Definition 3.29: Order Topology.

On a linearly ordered set, the *order topology* is generated by open intervals

$$(a, b) = \{x : a < x < b\},$$

together with the appropriate one-sided intervals near endpoints. The usual topology on $\mathbb{R}$ is also its order topology.

Definition 3.30: Weak Topology.

Let X be a set and let $\mathcal{F}$ be a collection of maps $f : X \to Y_f$, where each Y_f is a topological space. The *weak topology generated by* $\mathcal{F}$ is the coarsest topology on X that makes every $f \in \mathcal{F}$ continuous.

Example 3.31. For a normed vector space X, the *weak topology* $\sigma(X, X^*)$ is the coarsest topology making every continuous linear functional $f \in X^*$ continuous. Unwinding the definition, $x_n \to x$ weakly means

$$f(x_n) \to f(x) \qquad \text{for every } f \in X^*.$$

As a rule this topology is strictly weaker than the norm topology.

■

Definition 3.32: Zariski Topology.

On $\mathbb{R}^n$ or $\mathbb{C}^n$, the *Zariski topology* declares algebraic sets, meaning common zero sets of polynomials, to be closed. This topology is central in algebraic geometry and is much coarser than the usual topology.

Takeaway. The topologies worth recognizing on sight are the discrete and indiscrete, the metric/usual, cofinite, subspace, product, quotient, order, weak, and Zariski topologies. In each case the same question applies: what information are its open sets meant to preserve?

Definition 3.33.

A topological space X is *Hausdorff* if for every pair of distinct points $x, y \in X$, there exist open sets $U, V \subseteq X$ such that

$$x \in U, \qquad y \in V, \qquad U \cap V = \varnothing.$$

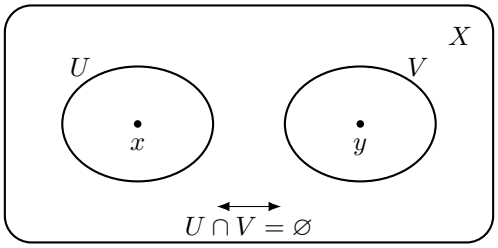


Figure 4. The Hausdorff condition separates distinct points by disjoint open neighborhoods.

Proposition 3.34.

Every metric space is Hausdorff.

Proof. Let (X, d) be a metric space and let $x \neq y$. Then $d(x, y) > 0$. Put

$$r = \frac{d(x, y)}{3}.$$

The balls $B(x, r)$ and $B(y, r)$ are disjoint. Indeed, if some z belonged to both balls, then

$$d(x, y) \leq d(x, z) + d(z, y) < r + r = \frac{2d(x, y)}{3},$$

which is impossible. Thus distinct points can be separated by disjoint open neighborhoods. □

Proposition 3.35.

In a Hausdorff space, a convergent sequence has at most one limit.

Proof. Suppose $x_n \to x$ and $x_n \to y$ with $x \neq y$. Since the space is Hausdorff, choose disjoint open sets U, V with $x \in U$ and $y \in V$. Because $x_n \to x$, eventually $x_n \in U$. Because $x_n \to y$, eventually $x_n \in V$. Therefore, for all sufficiently large n, we would have $x_n \in U \cap V$, contradicting $U \cap V = \varnothing$. □

3.2.3 How to Think About Hausdorff Spaces

Intuition. Hausdorffness is the statement that points are locally distinguishable. When $x \neq y$, the topology carries enough open sets to give each point a private open region of its own, and the two regions never touch.

It helps to read a topology as a system of tests: the open neighborhoods are the local observations available to us. If no pair of disjoint observations ever separates two distinct points, the topology is simply too coarse to tell those points apart. Hausdorffness is the least separation we can demand and still recover the metric-space behavior we rely on, namely that a limit, when it exists, lands at a single place.

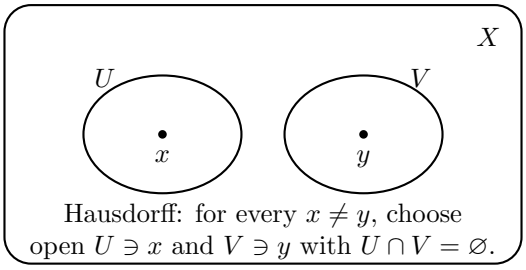


Figure 5. A Hausdorff space can separate any two distinct points.

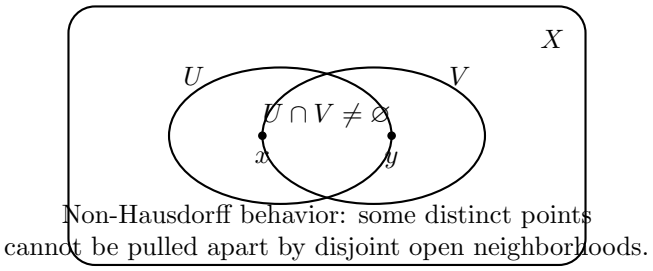


Figure 6. A non-Hausdorff space has points the topology cannot fully distinguish.

Example 3.36. The real line $\mathbb{R}$ with its usual topology is Hausdorff. If $x < y$, set

$$r = \frac{y - x}{3}.$$

Then $(x - r, x + r)$ and $(y - r, y + r)$ are disjoint open neighborhoods of x and y.

■

Example 3.37. Every discrete space is Hausdorff. If $x \neq y$, then $\{x\}$ and $\{y\}$ are disjoint open neighborhoods.

■

Example 3.38. If X has at least two points, the indiscrete topology $\{\varnothing, X\}$ is not Hausdorff. The only open neighborhood of any point is X itself, so two distinct points cannot have disjoint open neighborhoods.

■

Proposition 3.39.

A topological space X is Hausdorff if and only if the diagonal

$$\Delta = \{(x, x) : x \in X\}$$

is closed in the product space $X \times X$.

Proof. Assume X is Hausdorff. If $(x, y) \notin \Delta$, then $x \neq y$. Choose disjoint open sets $U, V \subseteq X$ with $x \in U$ and $y \in V$. Then $U \times V$ is an open neighborhood of (x, y) in $X \times X$, and

$$(U \times V) \cap \Delta = \varnothing.$$

Thus every point outside Δ has an open neighborhood contained in $(X \times X) \setminus \Delta$, so $(X \times X) \setminus \Delta$ is open. Therefore Δ is closed.

Conversely, assume Δ is closed. If $x \neq y$, then $(x, y) \notin \Delta$. Since $(X \times X) \setminus \Delta$ is open, there exist open sets $U, V \subseteq X$ such that

$$(x, y) \in U \times V \subseteq (X \times X) \setminus \Delta.$$

This means $x \in U$, $y \in V$, and $U \cap V = \varnothing$; otherwise a point $z \in U \cap V$ would give $(z, z) \in U \times V \cap \Delta$. Therefore X is Hausdorff. □

Remark 3.40. In full generality, sequences do not see everything a topology does. The sharp statement is that in a Hausdorff space limits of nets are unique. For metric spaces, and more generally first-countable spaces, sequences suffice, which is why the sequence version already carries the analytic intuition.

Remark 3.41. Hausdorffness is one of the places where topology pays off in functional analysis. A topological vector space is normally assumed Hausdorff, so that limits of vectors are unique. On a locally convex space the condition has a dual reading: the topology is Hausdorff exactly when the continuous linear functionals separate points, that is, for every $x \neq 0$ some continuous f has $f(x) \neq 0$. In particular the weak topology $\sigma(X, X^*)$ is Hausdorff precisely when X^* separates the points of X.

3.2.4 Why Hausdorff Spaces Matter

Intuition. So much of analysis turns on pinning down where a process ends. Once a sequence, net, or approximating family is allowed to converge to two different points at once, a great many arguments lose their grip.

Think of a limit as the location that all sufficiently small observations finally force on us. When two points cannot be fenced off by disjoint neighborhoods, no local observation can ever decide which of them a process is heading toward. Hausdorffness is precisely the promise that distinct points eventually demand distinct local evidence.

1. **It gives uniqueness of limits.** In metric spaces this is reflexive; in general topology it is a theorem. Hausdorff separation is what stops a single convergent process from having two limits.
2. **It makes compact sets behave like closed bounded sets.** Compact subsets of a Hausdorff space are closed. That fact is part of why compactness can stand in for finiteness throughout analysis.
3. **It makes graphs of continuous maps closed.** When Y is Hausdorff and $f : X \to Y$ is continuous, the graph
$$\text{graph}(f) = \{(x, f(x)) : x \in X\}$$
is closed in $X \times Y$ — and closed graphs sit at the center of functional analysis.
4. **It lets functions separate information cleanly.** Hausdorffness makes points topologically distinguishable, and in functional analysis that reappears as the question of whether continuous linear functionals separate vectors.
5. **It rules out pathological identifications.** A non-Hausdorff space can glue points so tightly that the topology can never pull them apart again. Some geometry exploits this, but for ordinary analysis it is usually too weak.

Proposition 3.42.

If X is compact and Y is Hausdorff, then every continuous bijection $f : X \to Y$ is a homeomorphism.

Proof. It is enough to prove that f sends closed sets to closed sets. Let $C \subseteq X$ be closed. Since X is compact, C is compact. Since f is continuous, $f(C)$ is compact in Y. Because Y is Hausdorff, compact subsets of Y are closed. Thus $f(C)$ is closed. Therefore f is a closed map, so the inverse map $f^{-1} : Y \to X$ is continuous. □

Takeaway. Hausdorffness is the condition that keeps distinct points distinct as far as limits are concerned. Analysis, functional analysis, and PDE all lean on it: drop it, and convergence can stop behaving like convergence.

Checkpoint. The test for Hausdorffness: take two arbitrary distinct points and ask whether the topology can supply disjoint open neighborhoods, one around each.

Common Pitfall. Hausdorffness is not built into the definition of a topological space — plenty of spaces lack it. Analysis simply tends to insist on it, since uniqueness of limits is usually too important to give up.

Takeaway. A topology fixes the open sets. A Hausdorff topology adds just enough separation between distinct points to force limits to be unique.

4 Measure Theory

Foundations to Strengthen.

- Read a σ-algebra as the collection of sets (or events) a given theory is permitted to measure.
- See measurable functions as maps that respect the measurable sets.
- Get the three convergence theorems—monotone convergence, Fatou, dominated convergence—to the point of reflex.
- Learn L^p spaces, Young's inequality, Holder's inequality, Minkowski's inequality, and completeness.
- Study product measures, Fubini–Tonelli, and Radon–Nikodym.

4.1 Measurable Spaces and Measure Spaces

Big Idea. Two questions usually get run together, and measure theory keeps them apart:

which sets are allowed to be measured? what size do those sets have?

Answering the first gives a measurable space; answering the second turns it into a measure space.

Definition 4.1: Measurable Space.

A measurable space is a pair

$$(X, \mathcal{A}),$$

where X is a set and $\mathcal{A}$ is a σ-algebra of subsets of X. This means:

1. $\varnothing \in \mathcal{A}$ and $X \in \mathcal{A}$;
2. if $A \in \mathcal{A}$, then $X \setminus A \in \mathcal{A}$;
3. if $A_1, A_2, \cdots \in \mathcal{A}$, then
$$\bigcup_{n=1}^{\infty} A_n \in \mathcal{A}.$$

The sets in $\mathcal{A}$ are called measurable sets.

Remark 4.2: What This Means. No sizes have been assigned yet. A measurable space records only which subsets of X count as legitimate questions. When $A \in \mathcal{A}$, asking "is the point in A?" is allowed; when $A \notin \mathcal{A}$, the theory simply declines to answer.

Definition 4.3: Measure Space.

A measure space is a triple

$$(X, \mathcal{A}, \mu),$$

where $(X, \mathcal{A})$ is a measurable space and

$$\mu : \mathcal{A} \to [0, \infty]$$

is a measure, meaning

$$\mu(\varnothing) = 0$$

and, whenever $A_1, A_2, \cdots \in \mathcal{A}$ are pairwise disjoint,

$$\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n).$$

This property is called countable additivity.

Remark 4.4: Measurable Space Versus Measure Space. The measurable space fixes the domain of discourse:

$$(X, \mathcal{A}).$$

A measure space is that, together with a rule for size:

$$(X, \mathcal{A}, \mu).$$

In short, $\mathcal{A}$ records what can be measured and μ records how large each measurable set turns out to be.

Example 4.5: Real Line. On $\mathbb{R}$ the standard measurable space is

$$(\mathbb{R}, \mathcal{B}(\mathbb{R})),$$

where $\mathcal{B}(\mathbb{R})$ is the Borel σ-algebra generated by the open sets. Attaching Lebesgue measure produces the measure space

$$(\mathbb{R}, \mathcal{B}(\mathbb{R}), m),$$

at least before completion, with $m([a, b]) = b - a$.

■

Remark 4.6: A Small Catalogue of Measures and d-Symbols. Every integral fits the same template:

$$\int_X f \, d\mu.$$

What sits after the d names the notion of size that weights the integral. Usually it is not f that is being measured but the input space, measured in some particular way.

Symbol	Measure or object	Meaning
dx, $dm(x)$	Lebesgue measure on $\mathbb{R}^n$	Length, area, or volume. For example, $\int_{\mathbb{R}} f(x)\,dx$ weights f by ordinary length.
dt	Lebesgue measure on time	Time length. For example, $\int_0^T h(t)\,dt$ is an ordinary time integral.
$d\mathbb{P}(\omega)$	Probability measure on Ω	Probability weight. Expectation is $\mathbb{E}[X] = \int_\Omega X(\omega)\,d\mathbb{P}(\omega)$.
$d\mathcal{L}_X(x)$	Law of a random variable X	The pushforward measure $X_\#\mathbb{P}$ on the output space. It measures sets of possible values of X.
$p(x)\,dx$	Measure with density p	Probability or mass distributed continuously relative to Lebesgue measure. Then $\int g(x)p(x)\,dx = \int g\,d\mu$.
δ_a	Dirac measure	Point mass at a. It gives $\int f\,d\delta_a = f(a)$.
counting measure	Measure on a countable set	Counts points. Integration becomes summation: $\int f\,d\# = \sum_x f(x)$.
$d\mathbb{P}\,dt$	Product measure	Measures random-time pairs (ω, t). This is the natural measure behind many $L^2(\Omega \times [0, T])$ norms.

Fixing a measure already brings function spaces with it:

$$L^p(X,\mu) = \left\{ f : \int_X |f|^p \, d\mu < \infty \right\}.$$

So $L^2([0,1], dx)$ is square-integrability against length, whereas $L^2(\Omega, d\mathbb{P})$ is square-integrability against probability:

$$\|X\|^2_{L^2(\Omega)} = \mathbb{E}[X^2] = \int_\Omega X(\omega)^2 \, d\mathbb{P}(\omega).$$

The L^2 inner product depends on the measure in the same way:

$$\langle f, g \rangle_{L^2(X,\mu)} = \int_X f g \, d\mu.$$

One caution: dB_t in stochastic calculus is not the same kind of symbol as dx, dt, or $d\mathbb{P}$. It denotes a Brownian increment and is taken up later in the probability chapter.

4.2 Measurable Functions

Big Idea. A measurable function is one that respects the measurable sets on both the domain and the target. Integrability is a separate matter. All that is asked here is that questions about the output pull back to measurable questions about the input.

Definition 4.7: Measurable Function.

Let $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ be measurable spaces. A function

$$f : X \to Y$$

is measurable if, for every $B \in \mathcal{B}$,

$$f^{-1}(B) = \{x \in X : f(x) \in B\} \in \mathcal{A}.$$

Remark 4.8: Measurable Does Not Mean Length. The word "measurable" invites the wrong reading—as if the function itself were being measured, say by the length of its graph or its norm. Nothing of the sort is meant.
A measure μ on X measures subsets of X, never functions. It can size an input set

$$A \subseteq X, \qquad A \in \mathcal{A}.$$

If instead $f : X \to Y$ is some observed quantity, the questions we care about are about its output,

$$\text{"is } f(x) \text{ in } B\text{?"} \qquad B \subseteq Y,$$

and μ can only get at them after they are translated back to an input set:

$$f^{-1}(B) = \{x \in X : f(x) \in B\} \subseteq X.$$

We call f measurable precisely when this translated set always lands among the sets μ is allowed to measure:

$$B \in \mathcal{B} \quad \Longrightarrow \quad f^{-1}(B) \in \mathcal{A}.$$

A measurable function, then, is not one whose "size" has been computed in advance. It is one whose output events can be legally measured on the input space. The size of the function arrives only later, through quantities like

$$\int_X |f| \, d\mu, \qquad \|f\|_{L^p}.$$

Proposition 4.9: Real-Valued Measurability Tests.

Let $(X, \mathcal{A})$ be a measurable space and let

$$f : X \to \mathbb{R}.$$

Then f is Borel measurable if and only if any one of the following equivalent tests

holds:

$$\{x : f(x) > a\} \in \mathcal{A} \quad \text{for every } a \in \mathbb{R},$$

or

$$\{x : f(x) < a\} \in \mathcal{A} \quad \text{for every } a \in \mathbb{R}.$$

It is enough to check these tests for rational numbers $a \in \mathbb{Q}$.

Proof. Define

$$\mathcal{C} = \{B \subseteq \mathbb{R} : f^{-1}(B) \in \mathcal{A}\}.$$

Since preimages preserve complements and countable unions, $\mathcal{C}$ is a σ-algebra of subsets of $\mathbb{R}$. So once $\mathcal{C}$ contains a generating family for $\mathcal{B}(\mathbb{R})$, it contains every Borel set.

The open rays (a, ∞) generate $\mathcal{B}(\mathbb{R})$, so checking

$$f^{-1}((a, \infty)) = \{x : f(x) > a\} \in \mathcal{A}$$

for every a already proves Borel measurability. The rays $(-\infty, a)$ work the same way. Rational thresholds suffice because, for example,

$$\{f > a\} = \bigcup_{\substack{q \in \mathbb{Q} \\ q > a}} \{f > q\}.$$

□

Remark 4.10: How to Prove Measurability in Practice. For a real-valued function there are a few standard routes:

1. **Direct preimage test.** Show $\{f > a\}$ or $\{f < a\}$ is measurable for every threshold a. This is the most hands-on route.
2. **Use known measurable building blocks.** Continuous functions are Borel measurable; an indicator 1_A is measurable exactly when A is; and sums, products, absolute values, maxima, minima, and pointwise limits of measurable real-valued functions are again measurable.
3. **Use composition.** If $f : X \to Y$ and $g : Y \to Z$ are measurable, so is $g \circ f : X \to Z$.
4. **Use pieces.** Cut X into countably many measurable pieces; if the formula for f is measurable on each piece, the whole piecewise-defined function is measurable.

Underneath all of them is the same demand: output questions have to pull back to measurable input questions.

Remark 4.11: Point Map Versus Set Test. The notation

$$f : X \to Y$$

records a point-level function:

$$x \in X \quad \longmapsto \quad f(x) \in Y.$$

Each element of X goes to an element of Y. It is not, in the first instance, a map from subsets

of X to subsets of Y. And at this stage X and Y carry only σ-algebras; they need not be vector spaces, and f need not be linear.
Keeping this straight heads off a common confusion with functional analysis:

general measurable function	$f : X \to Y$
linear operator	$T : X \to Y$ between vector spaces
linear functional	$\varphi : X \to \mathbb{F}$

A measurable function turns into an operator only once X and Y carry vector-space structure and the map is linear, and into a functional only when the target is the scalar field $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$ and the map is linear.
Measuring anything about the output, though, requires set-level questions. For a set $B \subseteq Y$ of possible output values,

$$f^{-1}(B) = \{x \in X : f(x) \in B\}$$

collects the input points whose images land in B. Note that $f^{-1}(B)$ is the *preimage of the set* B and carries no assumption that f is invertible.

Remark 4.12: Why Preimages Appear. Points move forward under $f : X \to Y$, but measurability is checked backward, on sets. An observable output event $B \subseteq Y$ corresponds to the input event $f^{-1}(B) \subseteq X$ that drives the output into B, and measurability is the demand

$$\text{measurable output question} \quad \Longrightarrow \quad \text{measurable input question.}$$

Remark 4.13: Where the Definition Comes From. Measurability exists to make output measurements legal. Suppose $f : X \to Y$ records a quantity observed on X and we want to answer

$$\text{“does } f(x) \text{ lie in } B\text{?”}$$

for measurable target sets $B \subseteq Y$. The measure μ sits on X, not on Y, so the only available way to assign a size to the output event B is to pull it back:

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

When $f^{-1}(B)$ fails to be measurable this number has no meaning at all. A measurable function is, then, exactly one for which every observable output question turns into a measurable input question.

Remark 4.14: Analogy with Continuous Functions. The definition is patterned on the topological characterization of continuity:

$$f \text{ continuous} \quad \Longleftrightarrow \quad f^{-1}(U) \text{ is open whenever } U \text{ is open.}$$

Measurability just swaps “open” for “measurable”:

$$f \text{ measurable} \quad \Longleftrightarrow \quad f^{-1}(B) \text{ is measurable whenever } B \text{ is measurable.}$$

The shape of the condition is identical—structure on the target has to pull back to structure on the domain—only the kind of structure changes. Topology tracks nearness; measure theory tracks which sets have a size and can be integrated over.

Example 4.15: Continuous Functions Are Measurable. Let

$$f : \mathbb{R} \to \mathbb{R}, \qquad f(x) = x^2,$$

with the Borel σ-algebra on both copies of $\mathbb{R}$. This f is measurable. Take the output question

$$f(x) < 4;$$

it pulls back to

$$f^{-1}((-\infty, 4)) = (-2, 2),$$

a Borel set. The same goes for any continuous

$$f : \mathbb{R} \to \mathbb{R},$$

since preimages of open sets are open and hence Borel.

■

Example 4.16: A Finite Random Variable. Let

$$\Omega = \{1, 2, 3, 4, 5, 6\}, \qquad \mathcal{F} = \mathcal{P}(\Omega),$$

and define the die-roll random variable

$$X : \Omega \to \mathbb{R}, \qquad X(\omega) = \omega.$$

Since every subset of Ω lies in $\mathcal{F}$, this is measurable. For the output event

$$B = \{2, 4, 6\},$$

we get

$$X^{-1}(B) = \{2, 4, 6\}.$$

The variable maps outcomes to numbers; the events are recovered by pulling output sets back to subsets of Ω.

■

Example 4.17: Indicator Function. Let $(X, \mathcal{A})$ be a measurable space and let $A \subseteq X$. The indicator function

$$1_A(x) = \begin{cases} 1, & x \in A, \\ 0, & x \notin A \end{cases}$$

is measurable if and only if $A \in \mathcal{A}$. Measurable functions therefore carry the information of the measurable sets inside them.

■

Example 4.18: Measurable Does Not Mean Integrable. On the measure space $((0, 1], \mathcal{B}((0, 1]), m)$, the function

$$f(x) = \frac{1}{x}$$

is measurable, being continuous on $(0, 1]$. Yet

$$\int_0^1 \frac{1}{x}\, dx = \infty.$$

Measurability only guarantees that the measure can see the function; it says nothing about whether the integral is finite.

■

Example 4.19: A Measurable but Discontinuous Function. The function

$$1_{\mathbb{Q}} : \mathbb{R} \to \mathbb{R}$$

is measurable, since $\mathbb{Q}$ is countable and so Borel, yet it is discontinuous at every point of $\mathbb{R}$. Measurability is the weaker requirement: it asks for measurable preimages, not open ones.

■

4.3 Pushforward Measure

Big Idea. A measurable function can carry a measure from its domain over to its target; the result is the pushforward measure. In probability this pushforward is precisely the law, or distribution, of a random variable.

Definition 4.20: Pushforward Measure.

Let $(X, \mathcal{A}, \mu)$ be a measure space, let $(Y, \mathcal{B})$ be a measurable space, and let

$$f : X \to Y$$

be measurable. The pushforward measure of μ by f is the measure $f_{\#}\mu$ on $(Y, \mathcal{B})$ defined by

$$(f_{\#}\mu)(B) = \mu(f^{-1}(B)), \qquad B \in \mathcal{B}.$$

Remark 4.21: What Pushforward Means. The measure lives on X, but through f we can size subsets of Y by pulling them back. A target set $B \subseteq Y$ inherits the size of all the points in X whose image lands in B:

$$B \quad \mapsto \quad f^{-1}(B) \quad \mapsto \quad \mu(f^{-1}(B)).$$

Example 4.22: A Die Roll Pushes Probability to Numbers. Let

$$\Omega = \{\text{rolls of one fair die}\}, \qquad \mathbb{P}(\{\omega\}) = \frac{1}{6},$$

and let

$$X : \Omega \to \mathbb{R}$$

be the number showing on the die. Its pushforward

$$X_{\#}\mathbb{P}$$

is the distribution of that numerical output. For instance,

$$(X_{\#}\mathbb{P})(\{2,4,6\}) = \mathbb{P}(X^{-1}(\{2,4,6\})) = \mathbb{P}(\{\text{even rolls}\}) = \frac{1}{2}.$$

A measure on the outcomes Ω has become a measure on numbers.

■

Example 4.23: A Constant Map Creates a Point Mass. Let $(X, \mathcal{A}, \mu)$ be a probability space and let

$$f : X \to \mathbb{R}, \qquad f(x) = c$$

for a fixed number c. Then $f_{\#}\mu$ is the point mass at c:

$$f_{\#}\mu = \delta_c.$$

To see it, take a Borel set $B \subseteq \mathbb{R}$ and compute

$$(f_{\#}\mu)(B) = \mu(f^{-1}(B)) = \begin{cases} 1, & c \in B, \\ 0, & c \notin B. \end{cases}$$

A pushforward can collapse an entire space onto a single output value.

■

Example 4.24: Squaring Transports Uniform Measure. Let μ be Lebesgue measure on $[0,1]$ and let

$$f : [0,1] \to [0,1], \qquad f(x) = x^2.$$

For $0 \le a \le 1$,

$$(f_{\#}\mu)([0,a]) = \mu(f^{-1}([0,a])) = \mu([0,\sqrt{a}]) = \sqrt{a}.$$

The output is no longer uniform. Squaring crowds the mass toward 0, since a wide band of small inputs all map to small outputs.

■

Theorem 4.25: Integral Against a Pushforward.

Let $f : X \to Y$ be measurable and let $g : Y \to \mathbb{R}$ be measurable. Whenever one side is well-defined,

$$\int_Y g(y)\, d(f_{\#}\mu)(y) = \int_X g(f(x))\, d\mu(x).$$

Proof. Start with indicator functions. For $g = 1_B$,

$$\int_Y 1_B(y)\, d(f_{\#}\mu)(y) = (f_{\#}\mu)(B) = \mu(f^{-1}(B)),$$

while

$$\int_X 1_B(f(x))\, d\mu(x) = \int_X 1_{f^{-1}(B)}(x)\, d\mu(x) = \mu(f^{-1}(B)).$$

Linearity carries this to nonnegative simple functions, monotone convergence then to all nonnegative measurable functions, and splitting into positive and negative parts handles the integrable case. □

Remark 4.26: Probability Translation. With $\mu = \mathbb{P}$ and $f = X$ a random variable,

$$X_{\#}\mathbb{P}$$

is the distribution of X, and the theorem reads as the familiar

$$\mathbb{E}[g(X)] = \int_{\mathbb{R}} g(y)\, d\mathcal{L}_X(y).$$

4.4 Integrable Functions

Big Idea. Measurable means the measure can see the function. Integrable means its signed area, or average, comes out finite. Integrability forces measurability, but the converse fails: plenty of measurable functions are not integrable.

Definition 4.27: Integrable Function.

Let $(X, \mathcal{A}, \mu)$ be a measure space. A measurable function

$$f : X \to \mathbb{R}$$

is integrable if

$$\int_X |f| \, d\mu < \infty.$$

In that case, the integral

$$\int_X f \, d\mu$$

is defined as a finite number.

Proposition 4.28: Common Integrability Tests.

Let $(X, \mathcal{A}, \mu)$ be a measure space and let $f : X \to \mathbb{R}$ be measurable.

1. If
$$\int_X |f| \, d\mu < \infty,$$
then f is integrable. This is the definition.
2. If g is integrable and
$$|f(x)| \leq g(x) \qquad \text{for every } x \in X,$$
then f is integrable.
3. If $\mu(X) < \infty$ and $|f(x)| \leq M$ for every $x \in X$ and some constant $M < \infty$, then f is integrable.

Proof. The first claim restates the definition. For the comparison test, monotonicity of the integral gives

$$\int_X |f| \, d\mu \leq \int_X g \, d\mu < \infty.$$

For the bounded-on-finite-measure test, set $g = M 1_X$:

$$\int_X |f| \, d\mu \leq \int_X M \, d\mu = M\mu(X) < \infty.$$

□

Remark 4.29: How to Prove Integrability in Practice. The order is not negotiable:

1. Establish measurability of f first. Until then the Lebesgue integral is not even legally defined.
2. Then show the absolute integral is finite:
$$\int_X |f| \, d\mu < \infty.$$
3. One does this by computing the integral outright, by bounding $|f|$ by a known integrable function, or by invoking boundedness on a finite-measure domain.

The template is

$$\boxed{\text{measurable first}} \quad \Longrightarrow \quad \boxed{\text{finite absolute size second}}.$$

A word of warning: signed cancellation does not count. Lebesgue integrability demands $\int_X |f| \, d\mu < \infty$, not just some formal finite value for $\int_X f \, d\mu$.

Remark 4.30: Measurable Versus Integrable. Measurability is a compatibility condition:

$$f^{-1}(B) \in \mathcal{A}.$$

Integrability is a condition on size:

$$\int_X |f| \, d\mu < \infty.$$

One says "the function is admissible," the other says "the function has finite total size." The second only makes sense once the first holds.

Example 4.31: A Complete Measurable-and-Integrable Proof. Let

$$f : \mathbb{R} \to \mathbb{R}, \qquad f(x) = x^2 1_{[0,1]}(x).$$

We check that f is integrable on $(\mathbb{R}, \mathcal{B}(\mathbb{R}), m)$.
For measurability: $x \mapsto x^2$ is continuous, hence Borel measurable, and $[0,1]$ is Borel, so $1_{[0,1]}$ is measurable. A product of measurable real-valued functions is measurable, which settles f.
For finiteness, note

$$0 \le |f(x)| = x^2 1_{[0,1]}(x) \le 1_{[0,1]}(x),$$

so comparison gives

$$\int_{\mathbb{R}} |f| \, dm \le \int_{\mathbb{R}} 1_{[0,1]} \, dm = m([0,1]) = 1 < \infty.$$

Hence f is integrable.

■

Example 4.32: Measurable But Not Integrable. On the measure space

$$((0,1), \mathcal{B}((0,1)), m),$$

the function

$$f(x) = \frac{1}{x}$$

is measurable but not integrable, since

$$\int_0^1 \frac{1}{x}\, dx = \infty.$$

Measurability by itself buys nothing about finiteness of the integral.

■

5 Functional Analysis

Foundations to Strengthen.

- Keep a concrete example list: ℓ^p, L^p, $C([0,1])$, Hilbert spaces, and spaces of operators.
- Know why bounded linear maps are exactly continuous linear maps.
- Practice operator norm estimates.
- Learn dual spaces through examples before the abstract theorems.
- Treat Hahn–Banach, uniform boundedness, open mapping, and closed graph as the core Banach-space theorems.
- Separate norm convergence, weak convergence, and weak-* convergence.

5.1 Hierarchy of Spaces

Big Idea. The picture shows the subject being built one layer of structure at a time. Moving inward adds information; the inner layers know everything the outer ones do, and more.

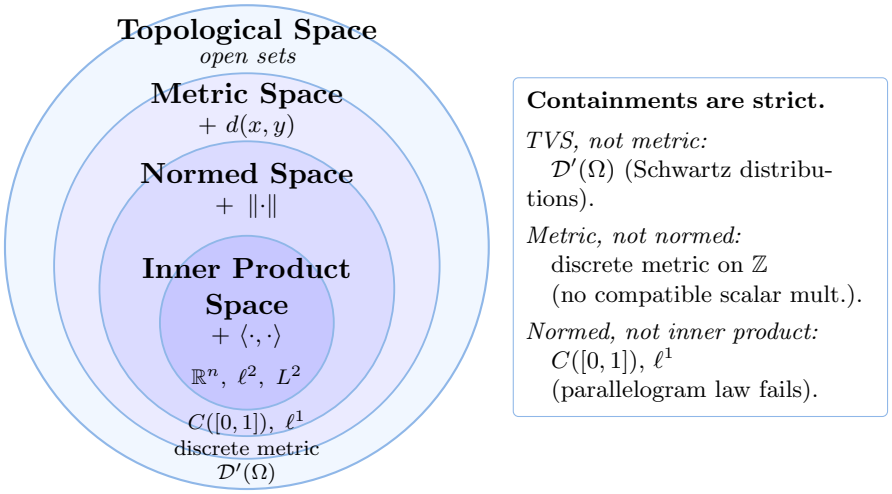


Figure 7. A hierarchy of spaces. Each ring adds the italicized structure to the one outside it. The label inside each annulus is a canonical space whose richest structure is exactly that layer, so it also witnesses that the containment is strict; the right-hand callout collects the same spaces as named counterexamples.

Intuition. Reading inward, each ring knows more. A topological space knows only its open sets. A metric space knows distances. A normed space adds vector addition, scaling, and a notion of length. An inner product space knows angles too.

One way to see why the layers fall in this order is to ask what the weakest version of each question requires. To talk about continuity at all, open sets suffice, which is the topological space. Wanting actual numerical distances forces a metric. Wanting to add and scale vectors while measuring their size forces a vector space and a norm. And wanting genuine geometry inside that space, angles and orthogonality and projection, forces an inner product. Each demand is strictly stronger than the one before it.

Proposition 5.1.

Every inner product space is a normed vector space, every normed vector space is a metric space, and every metric space is a topological space.

Proof. If V is an inner product space, define

$$\|x\| = \sqrt{\langle x, x\rangle}.$$

This is a norm, so V becomes a normed vector space. If V is a normed vector space, define

$$d(x, y) = \|x - y\|.$$

This is a metric. If (X, d) is a metric space, define a set $U \subseteq X$ to be open when for every $x \in U$ there exists $r > 0$ such that $B(x, r) \subseteq U$. These open sets form a topology. □

Remark 5.2. None of these implications reverse. A topological space need not come from a metric; a metric space need not carry any vector space structure; a normed space need not come from an inner product. The diagram records a hierarchy of added structure, not a chain of equivalences.

Example 5.3: A Normed Space That Is Not Hilbert. The space $C([0,1])$ with the supremum norm

$$\|f\|_\infty = \sup_{x\in[0,1]} |f(x)|$$

is a normed vector space, hence a metric space and a topological space. Its norm does not come from an inner product, though. Any norm that does satisfies the parallelogram law,

$$\|f+g\|^2 + \|f-g\|^2 = 2\|f\|^2 + 2\|g\|^2,$$

and the supremum norm fails this identity in general.

■

Remark 5.4. Function spaces, sequence spaces, and operator spaces are not new rungs sitting between "normed" and "inner product." They are the objects the whole subject is about. Which rung any one of them occupies depends on the structure we put on it: it might end up a topological vector space, a Banach space, or, with luck, a Hilbert space.

Remark 5.5. There is another object lurking near this picture: the *topological vector space*. It does not fit as a tidy circle inside metric spaces, since many topological vector spaces carry no norm and need not even be metrizable. What holds is the one-way inclusion: every normed space is a topological vector space, but the converse fails.

Takeaway. Keep this hierarchy in view as a map of the subject: topology buys continuity, a metric buys distance, a norm buys linear size, and an inner product buys geometry.

5.2 Function and Sequence Spaces

Big Idea. Before anything is said about functionals or operators, pin down the space whose points are under study. Here those points are usually functions, sequences, or linear maps.

Example 5.6: Function Space. The space $C([0,1])$ is a function space. A point of this space is a continuous function

$$f : [0,1] \to \mathbb{R}.$$

For instance $f(x) = x^2$ is one point of $C([0,1])$.
Distance only makes sense after we put structure on this set. Addition and scalar multiplication are defined pointwise:

$$(f+g)(x) = f(x) + g(x), \qquad (af)(x) = af(x).$$

With these operations, $C([0,1])$ is a vector space. Now define the supremum norm by

$$\|f\|_\infty = \sup_{x\in[0,1]} |f(x)|\,.$$

This is a norm: nonnegative, zero only when $f = 0$, homogeneous, and it obeys the triangle inequality because

$$|(f+g)(x)| \le |f(x)| + |g(x)|$$

holds at every $x \in [0,1]$. So $C([0,1])$ is a normed vector space, and the norm hands us the metric

$$d(f,g) = \|f-g\|_\infty = \sup_{x\in[0,1]} |f(x) - g(x)|\,.$$

Only with this norm in hand can we say that two functions are close in the uniform distance. ■

Example 5.7: Sequence Spaces. Typical *sequence spaces* include

$$\ell^p = \left\{ x = (x_n)_{n=1}^\infty : \sum_{n=1}^\infty |x_n|^p < \infty \right\}, \qquad 1 \le p < \infty,$$

with norm

$$\|x\|_p = \left(\sum_{n=1}^\infty |x_n|^p \right)^{1/p}.$$

These are normed vector spaces, and in fact complete, so Banach spaces. The case ℓ^2 stands out: it is a Hilbert space, carrying the inner product

$$\langle x, y \rangle = \sum_{n=1}^\infty x_n \overline{y_n}.$$

■

Example 5.8: More Function Spaces. Typical *function spaces* include $C([0,1])$, $L^p(\Omega)$, and the Sobolev spaces $H^k(\Omega)$. Each treats a function as a single point of a vector space, and a norm then decides when two such points are close. On $C([0,1])$, for instance,

$$\|f\|_\infty = \sup_{x\in[0,1]} |f(x)|$$

measures uniform closeness.

■

5.2.1 *Operator Spaces:* $L(X, Y)$

Definition 5.9: Operator Space $L(X, Y)$.

Let X and Y be normed vector spaces over the same field $\mathbb{F}$. The operator space from X to Y is

$$L(X, Y) = \mathcal{L}(X, Y) = \{T : X \to Y : T \text{ is linear and bounded}\}.$$

Here *bounded* means that there exists a constant $C \geq 0$ such that

$$\|Tx\|_Y \leq C \|x\|_X \qquad \text{for every } x \in X.$$

The norm on this operator space is the *operator norm*

$$\|T\|_{L(X,Y)} = \sup_{\|x\|_X \leq 1} \|Tx\|_Y = \sup_{x \neq 0} \frac{\|Tx\|_Y}{\|x\|_X}.$$

If $Y = \mathbb{F}$, then

$$X^* = L(X, \mathbb{F})$$

is the dual space of continuous linear functionals on X.
When no norms are in play, $\mathrm{Lin}(X, Y)$ is the safer symbol for the vector space of all linear maps $X \to Y$. Throughout these notes $L(X, Y)$ and $\mathcal{L}(X, Y)$ both mean the bounded linear maps between normed spaces.

Remark 5.10: Two Notations for Operator Spaces. In these notes $L(X, Y)$ and $\mathcal{L}(X, Y)$ carry exactly the same meaning, the bounded linear maps from X to Y; the choice between them is pure convention.
Books differ on this. Some reserve $L(X, Y)$ for all linear maps and use $\mathcal{L}(X, Y)$ only for the bounded ones. To sidestep the ambiguity, the convention adopted here is $\mathrm{Lin}(X, Y)$ when boundedness is not assumed, and $L(X, Y) = \mathcal{L}(X, Y)$ for bounded operators between normed spaces.

Example 5.11: Matrix Operators. Every matrix $A \in \mathbb{F}^{m \times n}$ defines a linear map

$$T_A : \mathbb{F}^n \to \mathbb{F}^m, \qquad T_A x = Ax.$$

With the usual Euclidean norms, $T_A \in L(\mathbb{F}^n, \mathbb{F}^m)$. In finite dimensions every linear map is automatically bounded, which makes matrix maps the basic finite-dimensional members of $L(X, Y)$.

■

Example 5.12: Evaluation Functional. Fix $a \in [0, 1]$. On $C([0, 1])$ with the supremum norm, define

$$\delta_a : C([0, 1]) \to \mathbb{F}, \qquad \delta_a(f) = f(a).$$

Then δ_a is linear, and

$$|\delta_a(f)| = |f(a)| \le \|f\|_\infty .$$

so

$$\delta_a \in L(C([0,1]), \mathbb{F}) = C([0,1])^*.$$

Its input is a function and its output is a scalar.

■

Example 5.13: Integral Operator. Define $K : C([0,1]) \to C([0,1])$ by

$$(Kf)(x) = \int_0^x f(t)\, dt.$$

The map K is linear. Also,

$$|(Kf)(x)| \le \int_0^x |f(t)|\, dt \le \|f\|_\infty \qquad (0 \le x \le 1),$$

so

$$\|Kf\|_\infty \le \|f\|_\infty .$$

Hence $K \in L(C([0,1]), C([0,1]))$. This is a typical infinite-dimensional operator: a function goes in, a function comes out.

■

Example 5.14: A Linear Map Not in $L(X,Y)$. The derivative map

$$D : C^1([0,1]) \to C([0,1]), \qquad Df = f',$$

is linear. But if the domain $C^1([0,1])$ is given only the supremum norm $\|f\|_\infty$, then D is not bounded. Indeed, for

$$f_n(x) = \sin(nx),$$

we have $\|f_n\|_\infty \le 1$, but

$$\|Df_n\|_\infty = \|n\cos(nx)\|_\infty = n.$$

No single constant C can make $\|Df\|_\infty \le C\,\|f\|_\infty$ hold for every f. Whether a map lands in $L(X,Y)$ is not intrinsic to the map; it depends on the norms placed on the domain and target.

■

Remark 5.15: Function Space versus Functional. A *function space* has functions for its points. A *space of functionals*, or *dual space*, has scalar-valued maps on some other vector space for its points. Functionals show up later precisely because they act *on* spaces rather than being the first spaces in sight.

Takeaway. Locate the point before anything else. In $C([0,1])$ it is a function, in ℓ^p a sequence, and in an operator space a bounded linear map, that is, a point of $L(X,Y)$.

5.3 Norms and Completeness

Intuition. A norm is a decision about what "size" will mean. Make that decision and a distance comes for free,

$$d(x, y) = \|x - y\| .$$

Different norms emphasize different features, maximum size or square-integrable energy among them.

The underlying point is short. Subtraction in a vector space produces an error vector $x - y$; a norm is the rule that collapses that vector into a single nonnegative number. Choosing a norm is choosing which kind of error we have agreed to care about.

Definition 5.16: Norm.

Let V be a vector space over $\mathbb{R}$ or $\mathbb{C}$. A *norm* on V is a map $\|\cdot\| : V \to [0, \infty)$ such that for all $x, y \in V$ and all scalars a,

(i) $\|x\| = 0$ if and only if $x = 0$;

(ii) $\|ax\| = |a| \, \|x\|$;

(iii) $\|x + y\| \leq \|x\| + \|y\|$.

Definition 5.17: Seminorm.

Let V be a vector space over $\mathbb{R}$ or $\mathbb{C}$. A *seminorm* on V is a map $p : V \to [0, \infty)$ such that for all $x, y \in V$ and all scalars a,

(i) $p(ax) = |a| \, p(x)$;

(ii) $p(x + y) \leq p(x) + p(y)$.

The one difference from a norm is that a seminorm may vanish on nonzero vectors:

$$p(x) = 0 \qquad \text{does not necessarily imply} \qquad x = 0.$$

Remark 5.18: Why Seminorms Matter. A seminorm reports on one feature of a vector, not on the whole of it. That is exactly what is wanted in spaces controlled by several measurements at once, say size together with derivative size. To recover a genuine norm one usually combines several seminorms, or passes to a quotient in which the invisible directions have been collapsed to 0.

Example 5.19: Derivative Seminorm. On $C^1([0, 1])$, define

$$p(f) = \left\| f' \right\|_\infty = \sup_{x \in [0,1]} \left| f'(x) \right| .$$

This is a seminorm because derivatives are linear:

$$(af)' = af', \qquad (f + g)' = f' + g'.$$

Therefore

$$p(af) = |a|\,p(f), \qquad p(f+g) \le p(f) + p(g).$$

It is not a norm, though. Take $f(x) = 3$: then $f \neq 0$, yet $f'(x) = 0$ everywhere, so

$$p(f) = 0.$$

What p records is variation, not height, so every constant function is invisible to it.

■

Definition 5.20: Banach Space.

A *Banach space* is a normed vector space that is complete in the metric induced by its norm. Complete here means that every Cauchy sequence has a limit, and that limit lies in the same space.

Example 5.21: Euclidean Norm. On $\mathbb{R}^n$, the Euclidean norm is

$$\|x\|_2 = \left(\sum_{j=1}^{n} |x_j|^2 \right)^{1/2}.$$

It comes from the inner product

$$\langle x, y \rangle = \sum_{j=1}^{n} x_j y_j,$$

and what it measures is ordinary geometric length.

■

Example 5.22: ℓ^p and ℓ^∞ Norms. For a sequence $x = (x_n)$ and $1 \le p < \infty$, define

$$\|x\|_p = \left(\sum_{n=1}^{\infty} |x_n|^p \right)^{1/p}.$$

The limiting maximum-size norm is

$$\|x\|_\infty = \sup_{n \ge 1} |x_n|\,.$$

So ℓ^p tracks summable size, whereas ℓ^∞ tracks the uniform boundedness of a sequence.

■

Example 5.23: L^p Norms. For a measurable function f on a measure space Ω and $1 \le p < \infty$, the L^p norm is

$$\|f\|_{L^p} = \left(\int_\Omega |f(x)|^p \, dx \right)^{1/p}.$$

The case $p = 2$ stands apart because it comes from the inner product

$$\langle f, g \rangle = \int_\Omega f(x)\overline{g(x)}\, dx.$$

■

Remark 5.24: Why the L^p Formula Looks This Way. The pattern

$$\left(\int_\Omega |f(x)|^p \, dx \right)^{1/p}$$

does three things. The factor $|f(x)|^p$ measures local size with no cancellation. The integral then sums those local sizes over the whole domain, exactly as the ℓ^p sum adds up coordinates. Finally the power $1/p$ pulls the answer back to the same scale as f.
At $p = 2$ this reads

$$\|f\|_{L^2} = \left(\int_\Omega |f(x)|^2 \, dx \right)^{1/2}.$$

The square sends positive and negative values to positive contributions and weights large values more heavily. The integral does real work here, collecting the small contributions scattered across the domain; drop it and there is no total size left to speak of. The square root is just as necessary, since it gives

$$\|cf\|_{L^2} = |c| \, \|f\|_{L^2},$$

while $\int_\Omega |f|^2$ on its own scales like $|c|^2$, making it an energy rather than a norm.

Remark 5.25: Sup Norm and L^∞ Norm. On $C([0,1])$, the supremum norm is often written like an L^∞ norm:

$$\|f\|_\infty = \sup_{x \in [0,1]} |f(x)|.$$

The reason is that the p-norm

$$\|f\|_p = \left(\int_0^1 |f(x)|^p \, dx \right)^{1/p}$$

becomes dominated by the largest value of $|f|$ as $p \to \infty$. So L^∞ is the limiting case in which only that largest size survives.

Remark 5.26. There is a small but important distinction. For general measurable functions, the L^∞ norm is defined by the *essential supremum*

$$\|f\|_{L^\infty} = \operatorname*{ess\,sup}_{x \in [0,1]} |f(x)|.$$

It overlooks anything happening on sets of measure zero, which is forced on us because L^p spaces identify functions agreeing almost everywhere. For a continuous function on $[0, 1]$ the

essential supremum and the ordinary supremum coincide, so on $C([0,1])$ we may safely write

$$\|f\|_{L^\infty} = \|f\|_\infty .$$

Example 5.27: Why L^∞ Uses Essential Supremum. Let

$$h(x) = \begin{cases} 1, & x = 0, \\ 0, & x \neq 0. \end{cases}$$

As an ordinary bounded function it has $\sup_{x\in[0,1]} |h(x)| = 1$. But $h = 0$ almost everywhere, so as an element of $L^\infty([0,1])$,

$$\|h\|_{L^\infty} = 0.$$

The single-point spike is what the essential supremum is built to ignore, and it is exactly why L^∞ uses it. A continuous function cannot hide such a spike, so on $C([0,1])$ the two notions agree.

■

Example 5.28: Sobolev Norm. In PDE work a norm often has to see a function and its derivatives at once. The standard Sobolev norm does this:

$$\|u\|_{H^1(\Omega)} = \left(\int_\Omega |u(x)|^2 \, dx + \int_\Omega |\nabla u(x)|^2 \, dx \right)^{1/2} .$$

By it, u counts as small only when u and its first derivatives are both small in L^2.

■

Takeaway. The norms worth recognizing on sight are the Euclidean, ℓ^p, L^p, supremum, L^∞, and Sobolev norms. The operator norm waits until bounded linear operators are in hand, since it sizes a map rather than a vector or a function.

5.4 Inequalities

Big Idea. Inequalities are the grammar of estimates. They trade a quantity we cannot handle for simpler ones that norms, integrals, or energy methods can.

Theorem 5.29: Young's Inequality.

Let $p, q > 1$ satisfy

$$\frac{1}{p} + \frac{1}{q} = 1.$$

If $a, b \geq 0$, then

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}.$$

Proof. If $a = 0$ or $b = 0$, the statement is immediate. Assume $a, b > 0$. Since $\log x$ is concave on $(0, \infty)$, and since the weights $1/p$ and $1/q$ add to 1,

$$\begin{aligned} \log\left(\frac{a^p}{p} + \frac{b^q}{q}\right) &\geq \frac{1}{p}\log(a^p) + \frac{1}{q}\log(b^q) && \text{([concavity of log] 1)} \\ &= \log a + \log b && \text{([power rule for log] 2)} \\ &= \log(ab). && \text{([log of product] 3)} \end{aligned}$$

Exponentiating both sides gives

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}. \qquad \text{([exponentiate (1)–(3)] 4)}$$

□

Remark 5.30: Why We Use Young's Inequality. The inequality earns its place whenever a proof throws up a product term while the estimates on hand only control powers of the two factors separately. It splits the mixed quantity

$$ab$$

into the separate pieces

$$a^p \qquad \text{and} \qquad b^q.$$

Energy estimates lean on this: one piece is often shrunk and absorbed into the left side, leaving the other on the right as controlled data.

Remark 5.31: The ε Version. In estimates, the most useful form is

$$ab \leq \varepsilon a^p + C_\varepsilon b^q \qquad (\varepsilon > 0),$$

with C_ε a positive constant depending on ε, p, q. When $p = q = 2$ it specializes to the workhorse estimate

$$ab \leq \varepsilon a^2 + \frac{1}{4\varepsilon} b^2.$$

This is the form that shows up in PDE energy estimates: the small εa^2 term is absorbed into the left side, and the $C_\varepsilon b^2$ term stays on the right.

Example 5.32: Splitting a Product Term. Suppose an estimate carries a product like $\|u\|\,\|f\|$. Young's inequality with $p = q = 2$ gives

$$\|u\|\,\|f\| \le \varepsilon\,\|u\|^2 + \frac{1}{4\varepsilon}\,\|f\|^2,$$

splitting the mixed term into one piece that sees only u and one that sees only f.

■

Theorem 5.33: Cauchy–Schwarz Inequality.

If H is an inner product space, then for all $x, y \in H$,

$$|\langle x, y\rangle| \le \|x\|\,\|y\|.$$

Proof. If $y = 0$, the statement is immediate. Assume $y \neq 0$ and define

$$\lambda = \frac{\langle x, y\rangle}{\|y\|^2}. \qquad \text{([projection coefficient] 1)}$$

Then

$$\begin{aligned} 0 &\le \|x - \lambda y\|^2 && \text{([nonnegativity of norm] 2)} \\ &= \|x\|^2 - \frac{|\langle x, y\rangle|^2}{\|y\|^2}. && \text{([expand using (1)] 3)} \end{aligned}$$

Thus

$$\begin{aligned} \frac{|\langle x, y\rangle|^2}{\|y\|^2} &\le \|x\|^2 && \text{([rearrange (3)] 4)} \\ |\langle x, y\rangle|^2 &\le \|x\|^2\,\|y\|^2 && \text{([multiply by } \|y\|^2\text{] 5)} \\ |\langle x, y\rangle| &\le \|x\|\,\|y\|. && \text{([take square roots] 6)} \end{aligned}$$

□

Remark 5.34: Why Cauchy–Schwarz Matters. The inequality says an inner product can never exceed the product of the two lengths. It converts the geometric interaction $\langle x, y\rangle$ into a pure size bound $\|x\|\,\|y\|$, and that single move underlies projection, orthogonality, and a long list of energy arguments.

Example 5.35: Cauchy–Schwarz in $\mathbb{R}^n$. For $x, y \in \mathbb{R}^n$,

$$\left|\sum_{j=1}^{n} x_j y_j\right| = |\langle x, y\rangle| \le \left(\sum_{j=1}^{n} x_j^2\right)^{1/2} \left(\sum_{j=1}^{n} y_j^2\right)^{1/2}.$$

The left side measures how the two vectors interact; the right side only needs their lengths. ■

Theorem 5.36: Holder's Inequality.

Let $1 < p, q < \infty$ satisfy

$$\frac{1}{p} + \frac{1}{q} = 1.$$

If $f \in L^p(\Omega)$ and $g \in L^q(\Omega)$, then

$$\int_\Omega |f(x)g(x)|\, dx \leq \|f\|_{L^p} \|g\|_{L^q}.$$

Remark 5.37: First-Principles Motivation for Holder. The quantity we want to control is the total interaction

$$\int_\Omega |fg|\, dx.$$

At a single point, $|f(x)g(x)|$ can blow up from either factor, a large f or a large g. Holder's inequality says that controlling f in the p-power sense and g in the dual q-power sense already pins down the total size of the product.
The exponents have to satisfy

$$\frac{1}{p} + \frac{1}{q} = 1,$$

and the reason is mechanical: the proof normalizes both functions and applies Young's inequality at each point,

$$FG \leq \frac{F^p}{p} + \frac{G^q}{q}.$$

Strip away the packaging and Holder is just Young applied pointwise and then integrated over the domain.

Proof. If $\|f\|_{L^p} = 0$ or $\|g\|_{L^q} = 0$, the result is immediate. Otherwise define

$$F = \frac{|f|}{\|f\|_{L^p}}, \qquad G = \frac{|g|}{\|g\|_{L^q}}. \qquad \text{([normalize the two functions] 1)}$$

Then $\|F\|_{L^p} = 1$ and $\|G\|_{L^q} = 1$. For every $x \in \Omega$,

$$F(x)G(x) \leq \frac{F(x)^p}{p} + \frac{G(x)^q}{q}. \qquad \text{([Young's inequality pointwise] 2)}$$

Integrating this pointwise estimate gives

$$\begin{aligned} \int_\Omega FG\, dx &\leq \frac{1}{p}\int_\Omega F^p\, dx + \frac{1}{q}\int_\Omega G^q\, dx && \text{([integrate (2)] 3)} \\ &= \frac{1}{p} + \frac{1}{q} && \text{([normalization in (1)] 4)} \\ &= 1. && \text{([conjugate exponents] 5)} \end{aligned}$$

Finally,

$$\int_\Omega |fg|\,dx = \|f\|_{L^p}\,\|g\|_{L^q}\int_\Omega FG\,dx \qquad \text{([definition of } F,G\text{] 6)}$$
$$\leq \|f\|_{L^p}\,\|g\|_{L^q}\,. \qquad \text{([use (5)] 7)}$$

□

Remark 5.38: Holder as the L^p Cauchy–Schwarz. Holder is the product estimate for L^p spaces, and at $p = q = 2$ it collapses to Cauchy–Schwarz in L^2:

$$\int_\Omega |fg|\,dx \leq \|f\|_{L^2}\,\|g\|_{L^2}\,.$$

Cauchy–Schwarz is the Hilbert-space special case; Holder is the general L^p statement.

Remark 5.39: Why a Measure Space Appears Here. The L^p inequalities need a measure space because L^p size is built by adding up local contributions:

$$\|f\|_{L^p} = \left(\int_\Omega |f|^p\,d\mu\right)^{1/p}.$$

For that integral to mean anything we need $(\Omega, \mathcal{A}, \mu)$, which records which sets are measurable and how much weight each part of the domain carries.
For finite-dimensional Cauchy–Schwarz in $\mathbb{R}^n$, no measure language is necessary; the coordinate sum already does the job. Yet that case is the same statement with counting measure:

$$\sum_{j=1}^n |x_j y_j| \leq \left(\sum_{j=1}^n |x_j|^p\right)^{1/p}\left(\sum_{j=1}^n |y_j|^q\right)^{1/q}.$$

Measure spaces are simply the common language in which sums, integrals, and probability expectations all become one kind of object.

Remark 5.40: Holder Points Toward Dual Spaces. A preview of item 5.75. For a normed vector space X over $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$, the *dual space* is

$$X^* = L(X, \mathbb{F}),$$

the bounded linear functionals $\varphi : X \to \mathbb{F}$, each a rule that eats a vector and returns a scalar. Its norm is

$$\|\varphi\| = \sup_{\|x\|_X \leq 1} |\varphi(x)|\,.$$

Holder manufactures such functionals. Fix $g \in L^q(\Omega)$ and set

$$\Phi_g(f) = \int_\Omega f(x)g(x)\,d\mu(x) \qquad (f \in L^p(\Omega)).$$

This Φ_g is linear in f, and Holder bounds it:

$$|\Phi_g(f)| \leq \|f\|_{L^p}\,\|g\|_{L^q}\,.$$

So Φ_g is bounded, that is,

$$\Phi_g \in (L^p(\Omega))^*.$$

This is the door from Holder into duality: any element of L^q can measure every element of L^p through integration.

Example 5.41: Holder with $p = 3$. If $f \in L^3(\Omega)$ and $g \in L^{3/2}(\Omega)$, then

$$\int_\Omega |fg| \, dx \le \|f\|_{L^3} \, \|g\|_{L^{3/2}} .$$

The product-control idea is the same as in Cauchy–Schwarz, except the two functions now live in different L^p spaces.

■

Theorem 5.42: Minkowski Inequality.

For $1 \le p < \infty$ and $f, g \in L^p(\Omega)$,

$$\|f + g\|_{L^p} \le \|f\|_{L^p} + \|g\|_{L^p} .$$

Proof. For $p = 1$, integrate the pointwise triangle inequality:

$$\int_\Omega |f + g| \, dx \le \int_\Omega |f| \, dx + \int_\Omega |g| \, dx. \qquad \text{([pointwise triangle inequality] 1)}$$

Now assume $1 < p < \infty$, and let q be the conjugate exponent. If $f + g = 0$ in L^p, there is nothing to prove. Otherwise,

$$\|f + g\|_{L^p}^p = \int_\Omega |f + g|^p \, dx. \qquad \text{([definition of } L^p \text{ norm] 2)}$$

Using $|f + g| \le |f| + |g|$,

$$\|f + g\|_{L^p}^p \le \int_\Omega (|f| + |g|) \, |f + g|^{p-1} \, dx. \qquad \text{([pointwise triangle inequality] 3)}$$

Apply Holder's inequality to each term:

$$\begin{aligned} \|f + g\|_{L^p}^p &\le \|f\|_{L^p} \left\| |f + g|^{p-1} \right\|_{L^q} + \|g\|_{L^p} \left\| |f + g|^{p-1} \right\|_{L^q} && \text{([Holder on both terms] 4)} \\ &= (\|f\|_{L^p} + \|g\|_{L^p}) \left\| |f + g|^{p-1} \right\|_{L^q} . && \text{([factor common term] 5)} \end{aligned}$$

Since $q = p/(p - 1)$,

$$\begin{aligned} \left\| |f + g|^{p-1} \right\|_{L^q} &= \left(\int_\Omega |f + g|^p \, dx \right)^{1/q} && \text{([compute } (p-1)q = p\text{] 6)} \\ &= \|f + g\|_{L^p}^{p-1} . && \text{([because } p/q = p - 1\text{] 7)} \end{aligned}$$

Therefore

$$\|f+g\|_{L^p}^p \le (\|f\|_{L^p} + \|g\|_{L^p})\,\|f+g\|_{L^p}^{p-1}. \qquad \text{([combine (5) and (7)] 8)}$$

Dividing by $\|f+g\|_{L^p}^{p-1}$ gives

$$\|f+g\|_{L^p} \le \|f\|_{L^p} + \|g\|_{L^p}. \qquad \text{([divide by nonzero factor] 9)}$$

□

Remark 5.43: Why Minkowski Matters. Minkowski is the triangle inequality for L^p. Without it,

$$\|f\|_{L^p} = \left(\int_\Omega |f|^p \, dx\right)^{1/p}$$

would be just a way of measuring size; with it, the measurement is an honest norm.

Example 5.44: Minkowski for Two Functions. For $f, g \in L^2(\Omega)$,

$$\left(\int_\Omega |f+g|^2 \, dx\right)^{1/2} \le \left(\int_\Omega |f|^2 \, dx\right)^{1/2} + \left(\int_\Omega |g|^2 \, dx\right)^{1/2}.$$

The L^2 distance behaves like ordinary length, then: the size of a sum never exceeds the sum of the sizes.

■

Takeaway. Four inequalities, four jobs: Young splits a product, Cauchy–Schwarz bounds an inner product, Holder bounds a product in L^p, and Minkowski supplies the triangle inequality for L^p norms.

5.5 Estimate Toolkit for Continuity

May 26, 2026

Big Idea. Continuity in normed spaces is almost always an estimate in disguise. A handful of inequalities settles continuity of the norm, of addition, of scalar multiplication, and of every bounded linear map.

Remark 5.45: Common Triangle Inequality Patterns. The same idea appears in several forms.

$$|a+b| \le |a| + |b| \qquad \text{for real or complex numbers,}$$

$$\|u+v\| \le \|u\| + \|v\| \qquad \text{for vectors in a normed space,}$$

and

$$d(x,z) \le d(x,y) + d(y,z) \qquad \text{for points in a metric space.}$$

Each reads the same way: the direct route is never longer than a detour through an intermediate point.

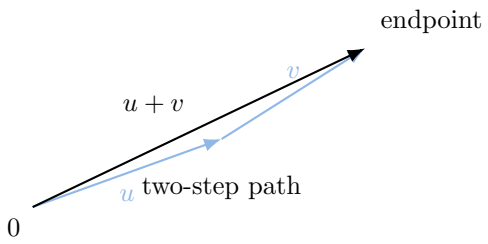


Figure 8. The vector triangle inequality says the direct displacement $u+v$ has length at most the two-step path with lengths $\|u\|$ and $\|v\|$.

Definition 5.46: Lipschitz Map.

Let (X, d_X) and (Y, d_Y) be metric spaces. A map $f : X \to Y$ is *Lipschitz* if there exists a constant $L \ge 0$ such that

$$d_Y(f(x), f(y)) \le L\, d_X(x,y) \qquad \text{for all } x, y \in X.$$

The number L is a Lipschitz constant. Read loosely, f may stretch distances, but never by more than the fixed factor L.

Proposition 5.47: Lipschitz Implies Continuous.

Every Lipschitz map is continuous.

Proof. Let $f : X \to Y$ be Lipschitz with constant L. If $L = 0$, then f is constant and therefore

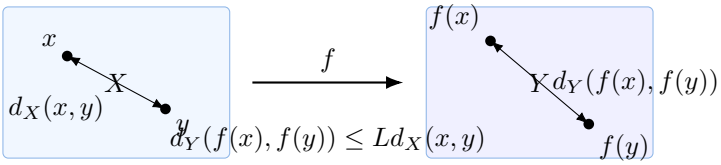


Figure 9. A Lipschitz map has controlled stretching: close inputs remain close outputs, with one fixed constant controlling all pairs.

continuous. If $L > 0$, given $\varepsilon > 0$, choose

$$\delta = \frac{\varepsilon}{L}.$$

Whenever $d_X(x, y) < \delta$,

$$d_Y(f(x), f(y)) \le L d_X(x, y) < L\delta = \varepsilon.$$

That is the ε–δ statement of continuity. □

Example 5.48: Examples of Lipschitz Maps. The map $f : \mathbb{R} \to \mathbb{R}$ given by

$$f(x) = 3x - 2$$

is Lipschitz with constant $L = 3$, because

$$|f(x) - f(y)| = |3x - 3y| = 3\,|x - y|\,.$$

The sine function is Lipschitz with constant $L = 1$:

$$|\sin x - \sin y| \le |x - y|\,.$$

This follows from the mean value theorem, since

$$\big|(\sin)'(t)\big| = |\cos t| \le 1.$$

In any normed space, the norm map

$$x \mapsto \|x\|$$

is Lipschitz with constant $L = 1$, because the reverse triangle inequality gives

$$|\|x\| - \|y\|| \le \|x - y\|\,.$$

The map $f(x) = x^2$ is not Lipschitz on all of $\mathbb{R}$; its slope runs off to infinity. It is Lipschitz on any bounded interval $[-R, R]$, though, since

$$\big|x^2 - y^2\big| = |x - y|\,|x + y| \le 2R\,|x - y|$$

whenever $x, y \in [-R, R]$.

■

Remark 5.49: How to Choose δ in an ε–δ Proof. The goal is always to force

$$d_Y(f(x), f(a)) < \varepsilon$$

from the assumption

$$d_X(x, a) < \delta.$$

In practice we bound the output error by the input error, then pick δ small enough that the bound drops below ε.
For a Lipschitz map,

$$d_Y(f(x), f(a)) \leq L d_X(x, a),$$

so it suffices to arrange

$$L d_X(x, a) < \varepsilon,$$

which means taking

$$\delta = \frac{\varepsilon}{L} \qquad (L > 0).$$

This is the easiest case there is: a Lipschitz estimate hands over δ with no further work.

Example 5.50: A Non-Lipschitz Delta Choice: x^2. Prove that $f(x) = x^2$ is continuous at a fixed point $a \in \mathbb{R}$.
Solution. We want: for every $\varepsilon > 0$ some $\delta > 0$ with

$$|x - a| < \delta \quad \Longrightarrow \quad \left|x^2 - a^2\right| < \varepsilon.$$

Factor the output error:

$$\left|x^2 - a^2\right| = |(x - a)(x + a)| = |x - a|\,|x + a|\,. \qquad \text{([difference of squares] 1)}$$

The input error $|x - a|$ is right there; the loose factor $|x + a|$ is what still needs controlling. Localize x near a with a preliminary restriction

$$|x - a| < 1. \qquad \text{([local restriction] 2)}$$

Then

$$\begin{aligned} |x| &= |(x - a) + a| && \text{([rewrite } x \text{ near } a\text{] 3)} \\ &\leq |x - a| + |a| && \text{([triangle inequality] 4)} \\ &< 1 + |a|\,. && \text{([use (2)] 5)} \end{aligned}$$

Bound the extra factor:

$$\begin{aligned} |x + a| &\leq |x| + |a| && \text{([triangle inequality] 6)} \\ &< 1 + 2\,|a|\,. && \text{([use (5)] 7)} \end{aligned}$$

Return to the output error. Combining (1) and (7),

$$\begin{aligned} \left|x^2 - a^2\right| &= |x - a|\,|x + a| && \text{([from (1)] 8)} \\ &< (1 + 2\,|a|)\,|x - a|\,. && \text{([use (7)] 9)} \end{aligned}$$

Therefore, to make $\left|x^2 - a^2\right| < \varepsilon$, it is enough to require

$$(1 + 2\,|a|)\,|x - a| < \varepsilon. \qquad \text{([target inequality] 10)}$$

This is guaranteed by

$$|x - a| < \frac{\varepsilon}{1 + 2\,|a|}. \qquad \text{([solve for input error] 11)}$$

Restrictions (2) and (11) must both hold, so take

$$\delta = \min\left\{1, \frac{\varepsilon}{1 + 2\,|a|}\right\}. \qquad \text{([satisfy both restrictions] 12)}$$

Then $|x - a| < \delta$ forces both $|x - a| < 1$ and $|x - a| < \varepsilon/(1 + 2\,|a|)$, and (9) gives

$$\left|x^2 - a^2\right| < \varepsilon.$$

So x^2 is continuous at a.

■

Remark 5.51: Why the Triangle Inequality Gives Continuity. On its own the triangle inequality says nothing about continuity. What it provides is the estimate that carries a small input error to a small output error.
Take addition and compare two nearby input pairs (x_1, y_1) and (x_2, y_2). Their output error is

$$\|(x_1 + y_1) - (x_2 + y_2)\| = \|(x_1 - x_2) + (y_1 - y_2)\|\,.$$

By the triangle inequality,

$$\|(x_1 + y_1) - (x_2 + y_2)\| \le \|x_1 - x_2\| + \|y_1 - y_2\|\,.$$

It is the same pattern as the familiar real-number inequality

$$|a + b| \le |a| + |b|\,,$$

only with vectors in place of numbers. Set

$$u = x_1 - x_2, \qquad v = y_1 - y_2.$$

Then

$$(x_1 + y_1) - (x_2 + y_2) = u + v,$$

so the normed-space triangle inequality gives

$$\|u + v\| \le \|u\| + \|v\|\,.$$

So when x_1 is close to x_2 and y_1 to y_2, the sum $x_1 + y_1$ is close to $x_2 + y_2$. In fact, with the product metric

$$D((x_1, y_1), (x_2, y_2)) = \|x_1 - x_2\| + \|y_1 - y_2\|\,,$$

addition is 1-Lipschitz:

$$\|(x_1 + y_1) - (x_2 + y_2)\| \le D((x_1, y_1), (x_2, y_2)).$$

By item 5.47, addition is continuous.

Remark 5.52: Scalar Multiplication. Scalar multiplication adds a wrinkle: the scalar and the vector can both move at once. The decomposition that handles this is

$$bx - ay = b(x - y) + (b - a)y.$$

Therefore

$$\|bx - ay\| \leq |b| \, \|x - y\| + |b - a| \, \|y\| \, .$$

If (b, x) is close to (a, y), then b is close to a, so b stays bounded, say $|b| \leq |a| + 1$ near a. Hence

$$\|bx - ay\| \leq (|a| + 1) \, \|x - y\| + |b - a| \, \|y\| \, ,$$

and the right-hand side tends to 0 as $b \to a$ and $x \to y$. That is continuity of the map

$$\mathbb{F} \times V \to V, \qquad (a, x) \mapsto ax.$$

Takeaway. Continuity proofs and the choice of δ almost always run backward. Bound the output error, see what has to be made small, and impose the input restriction that forces it.

5.6 Topological Vector Spaces

Learning Goals. See topological vector spaces as vector spaces whose topology gets along with addition and scalar multiplication. This is the shared setting that sits underneath normed spaces, Banach and Hilbert spaces, weak topologies, and many distribution spaces.

Intuition. A topological vector space is one where "these vectors are close" has a meaning, and where the algebraic operations honor it. If $x_n \to x$ and $y_n \to y$, the sums $x_n + y_n$ should approach $x + y$; if $a_n \to a$ and $x_n \to x$, the products $a_n x_n$ should approach ax.

The definition is more or less forced. Start from a vector space V, which gives addition and scalar multiplication. Analysis wants limits, so we hand V a topology. That topology is only worth having if it respects the algebra, meaning small changes in the inputs produce small changes in sums and scalar multiples. Demanding exactly that of the two structure maps

$$(x, y) \mapsto x + y \qquad \text{and} \qquad (a, x) \mapsto ax$$

is the same as demanding they be continuous.

Definition 5.53: Topological Vector Space.

Let V be a vector space over $\mathbb{F}$, where $\mathbb{F}$ is either $\mathbb{R}$ or $\mathbb{C}$. A *topological vector space* is a vector space V together with a topology such that the maps

$$V \times V \to V, \qquad (x, y) \mapsto x + y,$$

and

$$\mathbb{F} \times V \to V, \qquad (a, x) \mapsto ax,$$

are continuous.

The first condition is continuity of addition, the second continuity of scalar multiplication. Between them they say the topology and the linear structure are not working against each other.

Figure 10. A topological vector space requires both structure maps to be continuous.

Example 5.54: Normed Spaces are Topological Vector Spaces. Every normed vector space is a topological vector space. The norm hands V a metric,

$$d(x, y) = \|x - y\|,$$

and addition is continuous because

$$\|(x_1 + y_1) - (x_2 + y_2)\| \leq \|x_1 - x_2\| + \|y_1 - y_2\|.$$

Scalar multiplication is continuous because

$$\|a_n x_n - ax\| \le |a_n| \, \|x_n - x\| + |a_n - a| \, \|x\| ,$$

so $a_n \to a$ and $x_n \to x$ force $a_n x_n \to ax$. For why these inequalities amount to continuity, see items 5.51 and 5.52.

■

Example 5.55: Weak Topology. The weak topology on a normed space X is the topology in which $x_n \rightharpoonup x$ means

$$\varphi(x_n) \to \varphi(x) \qquad \text{for every } \varphi \in X^*.$$

It is usually coarser than the norm topology, yet it still respects vector addition and scalar multiplication, so it too is a topological vector space topology.

■

Remark 5.56. Conventions split on whether Hausdorffness belongs in the definition or is stated separately. In the Hausdorff case limits are unique, which is almost always the setting one wants in functional analysis.

Big Idea. In a topological vector space the local picture at every point is just the local picture at 0, shifted over. Translation $x \mapsto x + a$ is continuous with continuous inverse $x \mapsto x - a$, so neighborhoods of 0 are the only local data we really need.

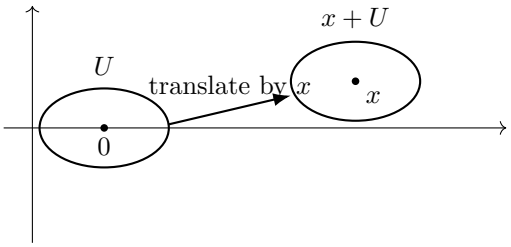


Figure 11. Neighborhoods of any point are translates of neighborhoods of 0.

Checkpoint. Why is a normed vector space automatically a topological vector space?

Solution. The norm gives a metric and so a topology. The triangle inequality makes addition continuous, and the norm estimate for $a_n x_n - ax$ makes scalar multiplication continuous. □

Takeaway. Topological vector spaces are where topology and linear algebra meet. Normed spaces are the familiar examples; weak topologies are the reason the subject needs the more general language at all.

5.7 Operators and Functionals

Big Idea. With the spaces settled, the maps between them come next. An operator sends vectors or functions to vectors or functions; a functional is the special case whose output is a single scalar.

Definition 5.57: Bounded Linear Operator.

Let X and Y be normed spaces. A linear map $T : X \to Y$ is *bounded* if there exists $C \geq 0$ such that

$$\|Tx\|_Y \leq C \|x\|_X \qquad \text{for every } x \in X.$$

The least possible such constant is the *operator norm*

$$\|T\| = \sup_{\|x\|_X \leq 1} \|Tx\|_Y .$$

Remark 5.58: Equivalent Ways to Read the Operator Norm. For a bounded linear operator $T : X \to Y$, the operator norm can be read in three equivalent ways:

$$\|T\| = \sup_{\|x\|_X \leq 1} \|Tx\|_Y = \sup_{\|x\|_X = 1} \|Tx\|_Y = \sup_{x \neq 0} \frac{\|Tx\|_Y}{\|x\|_X}.$$

The last form is usually the clearest: it asks for the largest amplification ratio

$$\frac{\text{output size}}{\text{input size}} = \frac{\|Tx\|_Y}{\|x\|_X}.$$

Why does this match the version that only inspects unit vectors? For $x \neq 0$, write

$$u = \frac{x}{\|x\|_X}.$$

Then $\|u\|_X = 1$ and

$$x = \|x\|_X \, u.$$

By linearity,

$$Tx = T(\|x\|_X \, u) = \|x\|_X \, Tu,$$

so

$$\frac{\|Tx\|_Y}{\|x\|_X} = \|Tu\|_Y .$$

So the ratio for any nonzero x is already attained at the normalized direction $u = x/\|x\|_X$.

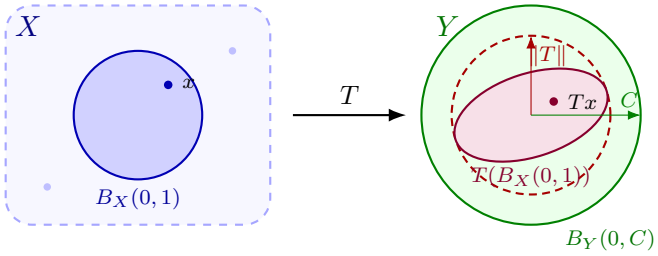


Figure 12. Schematic picture of boundedness. Because T is linear, the image $T(B_X(0,1))$ is an ellipse (purple), not an arbitrary blob. The smallest closed ball that contains it has radius $\|T\|$ (dashed red), so the ellipse is tangent to that circle. Any larger $C \geq \|T\|$ also works as a bound, so the green ball $B_Y(0,C)$ trivially contains the image; the operator norm $\|T\|$ is the infimum over all admissible C.

Definition 5.59: Bounded Linear Functional.

Let X be a normed vector space over $\mathbb{F}$. A *linear functional* on X is a linear map

$$\varphi : X \to \mathbb{F}.$$

It is a *bounded linear functional* if there exists $C \geq 0$ such that

$$|\varphi(x)| \leq C \|x\|_X \qquad \text{for every } x \in X.$$

A bounded linear functional is therefore nothing more than a bounded linear operator whose target happens to be the scalar field $\mathbb{F}$.

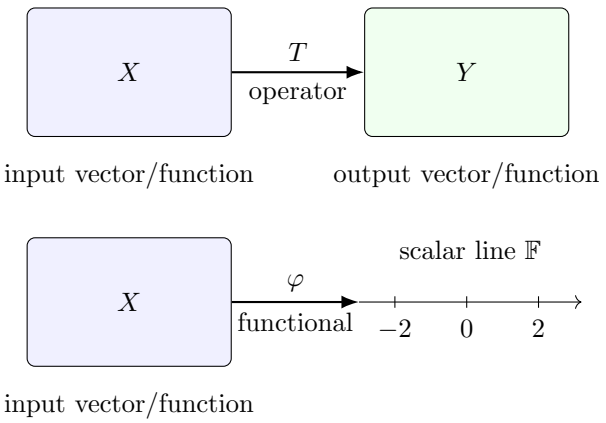
Remark 5.60: Operator Versus Functional. The difference is the output.

$$T : X \to Y \qquad \text{operator: output lives in another normed space } Y,$$

while

$$\varphi : X \to \mathbb{F} \qquad \text{functional: output is one scalar.}$$

Every bounded linear functional is a bounded linear operator, but the reverse fails. Take $T(f) = f'$, an operator since its output is another function, against $\varphi(f) = f(a)$, a functional since its output is the single number $f(a)$.



Visual. An operator preserves vector-valued information; a functional crushes the input down to one scalar reading.

Theorem 5.61: Bounded if and only if Continuous.

Let X and Y be normed spaces and let $T : X \to Y$ be linear. The following three statements are equivalent:

(A) T is bounded: there exists $M \geq 0$ such that

$$\|Tx\|_Y \leq M \|x\|_X \qquad \text{for every } x \in X.$$

(B) T is continuous on all of X.

(C) T is continuous at 0.

Remark 5.62: How to Read This Theorem. With an arbitrary function, continuity at one point tells us next to nothing about continuity anywhere else. Linearity changes that: continuity at 0 alone is enough. Control T near 0, scale any large vector x down into that neighborhood, and then use linearity to scale the output back up. The argument yields a single global constant C with

$$\|Tx\|_Y \leq C \|x\|_X \qquad \text{for every } x \in X,$$

which is boundedness.

Remark 5.63: Why Choose $\varepsilon = 1$?. For the direction "continuous at 0 implies bounded" there is no new limit to establish; one finite output bound near 0 is all we are after. So we run continuity at 0 with the convenient tolerance $\varepsilon = 1$, which produces a radius $\delta > 0$ with

$$\|u\|_X < \delta \qquad \Longrightarrow \qquad \|Tu\|_Y < 1.$$

Nothing depends on the value 1; any fixed positive tolerance does the same job. Nor is δ a free choice: continuity guarantees some such radius, and any smaller one works too.

Remark 5.64: Why Define $u = \frac{\delta}{2\|x\|}x$?. Fix a nonzero x. The vector

$$\frac{x}{\|x\|_X}$$

is x normalized: same direction, norm 1. Scaling it by $\delta/2$ gives

$$u = \frac{\delta}{2}\frac{x}{\|x\|_X} = \frac{\delta}{2\|x\|_X}x,$$

a short vector pointing along x, with

$$\|u\|_X = \frac{\delta}{2} < \delta.$$

The $1/2$ carries no significance beyond keeping u safely inside the open ball $\|u\|_X < \delta$; $\delta/3$ or 0.99δ would serve just as well. What we cannot use is δ itself, since that gives $\|u\|_X = \delta$ and the continuity statement demands the strict inequality $\|u\|_X < \delta$.

Remark 5.65: What Does "Prove Continuity" Mean Here?. Proving T continuous at a point a comes down to one task:

$$\text{given any allowed output error } \varepsilon > 0, \text{ find an input radius } \delta > 0$$

such that

$$\|x - a\|_X < \delta \qquad \Longrightarrow \qquad \|Tx - Ta\|_Y < \varepsilon.$$

At heart it is error control:

$$\text{small input error} \quad \Longrightarrow \quad \text{small output error.}$$

When T is linear, the output error takes a special form:

$$Tx - Ta = T(x - a).$$

If T is bounded, then

$$\|Tx - Ta\|_Y = \|T(x - a)\|_Y \leq M \|x - a\|_X .$$

So boundedness is a machine for producing δ: to force $M\|x - a\|_X < \varepsilon$ it is enough to require

$$\|x - a\|_X < \frac{\varepsilon}{M}.$$

That is all there is to $(A) \Rightarrow (B)$.

Proof. **Step 1: (A) implies (B).** Assume T is bounded. Then there is $M \geq 0$ such that

$$\|Tx\|_Y \leq M \|x\|_X \qquad \text{for every } x \in X.$$

We prove continuity at an arbitrary point $a \in X$.

Continuity target. Given any $\varepsilon > 0$, we must find $\delta > 0$ such that

$$\|x - a\|_X < \delta \qquad \Longrightarrow \qquad \|Tx - Ta\|_Y < \varepsilon.$$

If $M = 0$, then $T = 0$, so every output error is 0 and continuity is immediate. Now assume $M > 0$ and choose

$$\delta = \frac{\varepsilon}{M}.$$

Whenever $\|x - a\|_X < \delta$,

$$\begin{aligned}
\|Tx - Ta\|_Y &= \|T(x - a)\|_Y && \text{([turn output error into input error] 1)} \\
&\le M \|x - a\|_X && \text{([boundedness controls the error] 2)} \\
&< M\delta && \text{([because } \|x - a\|_X < \delta\text{] 3)} \\
&= \varepsilon. && \text{([choice of } \delta\text{] 4)}
\end{aligned}$$

Thus T is continuous at a. Since a was arbitrary, T is continuous on all of X.

Step 2: (B) implies (C). If T is continuous on all of X, then it is continuous at the particular point 0.

Step 3: (C) implies (A). Assume T is continuous at 0. Apply continuity at 0 with output tolerance $\varepsilon = 1$. Then there exists $\delta > 0$ such that

$$\|u\|_X < \delta \qquad \Longrightarrow \qquad \|Tu\|_Y < 1.$$

Fix $x \neq 0$ and define

$$u = \frac{\delta}{2} \frac{x}{\|x\|_X} = \frac{\delta}{2 \|x\|_X} x.$$

This is the normalized direction of x, scaled down to length $\delta/2$. Then

$$\begin{aligned}
\|u\|_X &= \frac{\delta}{2 \|x\|_X} \|x\|_X && \text{([definition of } u\text{] 5)} \\
&= \frac{\delta}{2} && \text{([cancel the input size] 6)} \\
&< \delta. && \text{([strictly inside the continuity ball] 7)}
\end{aligned}$$

Therefore $\|Tu\|_Y < 1$. Using linearity again,

$$\begin{aligned}
\frac{\delta}{2 \|x\|_X} \|Tx\|_Y &= \left\| \frac{\delta}{2 \|x\|_X} Tx \right\|_Y && \text{([homogeneity of the norm] 8)} \\
&= \left\| T\left(\frac{\delta}{2 \|x\|_X} x \right) \right\|_Y && \text{([linearity] 9)} \\
&= \|Tu\|_Y && \text{([definition of } u\text{] 10)} \\
&< 1. && \text{([continuity at 0] 11)}
\end{aligned}$$

Hence

$$\|Tx\|_Y < \frac{2}{\delta} \|x\|_X .$$

The same estimate is trivial when $x = 0$. Thus T is bounded with constant $M = 2/\delta$. This proves (A). □

Remark 5.66: How to Rediscover This Proof. The proof should not read like a rabbit pulled from a hat. To rediscover it, set the *target* and the *available tool* side by side.
Target. Boundedness asks for one constant M with

$$\|Tx\|_Y \le M \|x\|_X \qquad \text{for every } x \in X.$$

The job is to control $\|Tx\|_Y$ for an arbitrarily large input x.
Available tool. Continuity at 0 is purely local: with $\varepsilon = 1$ there is a $\delta > 0$ such that

$$\|u\|_X < \delta \qquad \Longrightarrow \qquad \|Tu\|_Y < 1.$$

This says nothing directly about a large x, only about inputs u near 0.
Bridge. Linearity lets us rescale. Given a large nonzero x, shrink it into the local ball:

$$u = \frac{\delta}{2} \frac{x}{\|x\|_X}.$$

Then $\|u\|_X = \delta/2 < \delta$, so the available tool says

$$\|Tu\|_Y < 1.$$

But u is a scalar multiple of x, so linearity gives

$$Tu = T\left(\frac{\delta}{2\|x\|_X} x\right) = \frac{\delta}{2\|x\|_X} Tx.$$

Therefore

$$\frac{\delta}{2\|x\|_X} \|Tx\|_Y < 1,$$

which rearranges to

$$\|Tx\|_Y < \frac{2}{\delta} \|x\|_X .$$

The constant has appeared on its own: $M = 2/\delta$.
The recipe behind the discovery is:

write the target
$\Downarrow$
write what the hypothesis gives
$\Downarrow$
build a bridge between them.

The bridge here is normalization together with linearity.

Remark 5.67: Which "Bounded" Is This?. The word *bounded* here has nothing to do with an input function being pointwise bounded, nor with its belonging to L^2 or L^p; those are separate notions.

For a function f, pointwise bounded means

$$\sup_x |f(x)| < \infty.$$

For an L^p function, integrability means

$$\int |f|^p < \infty.$$

For an operator $T : X \to Y$, bounded means

$$\|Tx\|_Y \leq C \|x\|_X \qquad \text{for every input } x.$$

A bounded linear operator may well act on L^2 or L^p spaces, but boundedness is a property of the rule T itself, not a claim that T is an L^2 or L^p function.

Remark 5.68: Why Bounded and Linear Belong Together. Linearity gives algebra:

$$T(x+y) = Tx + Ty, \qquad T(\alpha x) = \alpha Tx.$$

Boundedness gives analysis:

$$\|Tx\|_Y \leq C \|x\|_X .$$

Together they make T respect both the vector-space structure and limits. That is why bounded linear maps are the right morphisms for normed spaces: they are precisely the continuous linear maps. Drop boundedness and a linear map can wreck convergence; keep it and limits, Cauchy sequences, and dual spaces all stay under control.

Example 5.69: Operator Norm. If $T : X \to Y$ is a bounded linear operator between normed spaces, its operator norm

$$\|T\| = \sup_{\|x\|_X \leq 1} \|Tx\|_Y$$

records the most T can stretch a vector of size at most 1.

■

Example 5.70: Operator Space Revisited. By item 5.9, if X and Y are normed spaces, the space

$$L(X,Y) = \mathcal{L}(X,Y)$$

of bounded linear operators $T : X \to Y$ is an operator space. It has the operator norm

$$\|T\| = \sup_{\|x\|_X \leq 1} \|Tx\|_Y .$$

The dual space $X^* = L(X, \mathbb{F})$ is the operator space of continuous linear functionals on X.

■

5.7.1 Functions as Points and Functionals as Points

Intuition. The distinction is clean: a point of a function space is a function, while a point of a space of functionals is a rule that consumes a function or vector and returns a scalar.

The same object can play either role depending on the space it sits in. A function $g \in L^2(\Omega)$ is a point of the Hilbert space $L^2(\Omega)$, yet the same g also generates a functional

$$F_g(f) = \int_\Omega f(x)\overline{g(x)}\,dx.$$

For all that, g and F_g are different kinds of object: g is an input vector, F_g a scalar-valued map on vectors.

Example 5.71: Space of Functionals. The dual space $C([0,1])^*$ is a space whose points are continuous linear functionals on $C([0,1])$. For example, for a fixed $a \in [0,1]$, the evaluation map

$$\delta_a : C([0,1]) \to \mathbb{R}, \qquad \delta_a(f) = f(a),$$

is a point of $C([0,1])^*$. It is not a function on $[0,1]$ at all; its input is itself a function.

■

Example 5.72: Integral Functional. If $g \in C([0,1])$, then

$$I_g(f) = \int_0^1 f(x)g(x)\,dx$$

defines a continuous linear functional $I_g \in C([0,1])^*$. The same g wears two hats: a point of the function space $C([0,1])$, and, through I_g, a point of the functional space $C([0,1])^*$.

■

Example 5.73: Examples of Integral Functionals. Here are common scalar-valued functionals on a function space:

$$A(f) = \int_0^1 f(t)\,dt \qquad \text{average or total mass,}$$

$$M_1(f) = \int_0^1 tf(t)\,dt \qquad \text{first moment,}$$

and, more generally,

$$I_g(f) = \int_0^1 f(t)g(t)\,dt.$$

In every case the input is a function f and the output is one number, which is exactly what makes them functionals.

■

Remark 5.74: Delta as an Evaluation Functional. The evaluation functional

$$\delta_a(f) = f(a)$$

is not an integral against any continuous g. It can still be written formally as

$$\delta_a(f) = \int_0^1 f(t)\,\delta_a(t)\,dt,$$

with δ_a the Dirac delta concentrated at a, and more honestly as integration against the Dirac measure:

$$\delta_a(f) = \int_{[0,1]} f(t)\,d\delta_a(t).$$

The delta symbol is no ordinary function; it is shorthand for the functional "evaluate f at a."

Takeaway. The output settles the name: scalar output makes a functional, vector or function output makes an operator.

5.8 Dual Space

Big Idea. The dual space gathers up every continuous linear measurement of a normed space. Rather than jump straight to scalar-valued maps, it pays to prove the general statement first: a complete target makes the whole space of bounded linear operators complete.

Definition 5.75: Dual Space.

Let X be a normed vector space over $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$. The *dual space* of X is

$$X^* = L(X, \mathbb{F}),$$

the normed space of all bounded linear functionals $\varphi : X \to \mathbb{F}$ with norm

$$\|\varphi\| = \sup_{\|x\|_X \leq 1} |\varphi(x)| .$$

Remark 5.76: Existence and Uniqueness of the Dual Space. Set-theoretically the dual space X^* always exists: it is the collection of all bounded linear maps from X into $\mathbb{F}$. At least one such functional is always present, the zero functional

$$0(x) = 0.$$

Existence is not the real question. The question is whether X^* holds enough nonzero functionals to detect the geometry of X, and this is part of why the Hahn–Banach theorem matters: it guarantees a generous supply of continuous linear functionals on a normed space.
Once X and the scalar field are fixed, X^* is also unique:

$$X^* = L(X, \mathbb{F})$$

is a definition, not a construction involving choices. Two authors working with the same bounded functionals and the same operator norm end up with the same normed space, and any difference in notation is settled by a canonical identification.

5.8.1 The Banach Space of Bounded Linear Operators

Intuition. A Cauchy sequence (A_n) in operator norm is one whose operators eventually act almost identically on every unit vector. For each fixed $x \in X$, then, the outputs $A_n x$ ought to converge somewhere in Y. The argument goes through because Y is complete, so those pointwise limits genuinely exist.

Watch where completeness enters: in the target Y, not the domain X. The limits being taken are limits of the output sequence $(A_n x)$ inside Y, not of vectors inside X.

> **Remark 5.77: Source.** The next theorem is Theorem 1.3.1 in Theo Bühler and Dietmar A. Salamon, *Functional Analysis*. What follows fills in the local definitions and the skipped estimates, so their proof can be read without pausing.

Definition 5.78: Bounded Operators and the Operator Norm, Local Reminder.

Let X and Y be normed vector spaces over the same field. A linear map $T : X \to Y$ is *bounded* if there exists a constant $C \geq 0$ such that

$$\|Tx\|_Y \leq C \|x\|_X \qquad \text{for every } x \in X.$$

The space of bounded linear operators from X to Y is

$$L(X,Y) = \{T : X \to Y : T \text{ is linear and bounded}\}.$$

For $T \in L(X,Y)$, the *operator norm* is

$$\|T\|_{L(X,Y)} = \sup_{\|x\|_X \leq 1} \|Tx\|_Y .$$

It is the largest output size T can produce from inputs of size at most 1.

Lemma 5.79: Basic Operator-Norm Estimate.

If $T \in L(X,Y)$, then

$$\|Tx\|_Y \leq \|T\|_{L(X,Y)} \|x\|_X \qquad \text{for every } x \in X.$$

Proof. If $x = 0$, then $Tx = T0 = 0$, so the estimate is immediate. Now suppose $x \neq 0$ and define

$$u = \frac{x}{\|x\|_X}.$$

Then $\|u\|_X = 1$. Since u is in the unit ball, the definition of operator norm gives

$$\|Tu\|_Y \leq \|T\|_{L(X,Y)} . \qquad \text{([unit ball] 1)}$$

Because $x = \|x\|_X\, u$, linearity gives

$$\begin{aligned} \|Tx\|_Y &= \|T(\|x\|_X\, u)\|_Y && \text{([rewrite } x\text{] 2)} \\ &= \|\|x\|_X\, Tu\|_Y && \text{([linearity] 3)} \\ &= \|x\|_X\, \|Tu\|_Y && \text{([homogeneity] 4)} \\ &\le \|x\|_X\, \|T\|_{L(X,Y)}\,. && \text{([use (1)] 5)} \end{aligned}$$

This proves the estimate. □

Definition 5.80: Cauchy in Operator Norm.

A sequence (A_n) in $L(X,Y)$ is Cauchy in operator norm if

$$\forall \rho > 0,\ \exists N \in \mathbb{N},\ \forall m, n \ge N, \qquad \|A_m - A_n\|_{L(X,Y)} < \rho.$$

The tolerance ρ lives in the operator space $L(X,Y)$. Fixing a single input x later turns this operator error into an output error inside Y.

Theorem 5.81: Completeness of Operator Spaces.

Let X be a normed vector space and let Y be a Banach space over the same field. Then $L(X,Y)$ is a Banach space with respect to the operator norm.

Proof. Let $(A_n)_{n\in\mathbb{N}}$ be a Cauchy sequence in $L(X,Y)$. We must show that it converges, in operator norm, to some bounded linear operator $A : X \to Y$.

Step 1: operator Cauchy implies pointwise Cauchy. Fix $x \in X$. We prove that $(A_n x)$ is Cauchy in Y. Let $\eta > 0$ be arbitrary. Choose

$$\rho = \frac{\eta}{\|x\|_X + 1}. \qquad \text{([choose operator tolerance] 1)}$$

This number is positive because $\eta > 0$ and $\|x\|_X + 1 > 0$. Since (A_n) is Cauchy in operator norm, there exists $N \in \mathbb{N}$ such that

$$m, n \ge N \quad \Longrightarrow \quad \|A_m - A_n\|_{L(X,Y)} < \frac{\eta}{\|x\|_X + 1}. \qquad \text{([Cauchy choice of } N\text{] 2)}$$

Then, for $m, n \ge N$,

$$\begin{aligned} \|A_m x - A_n x\|_Y &= \|(A_m - A_n)x\|_Y && \text{([linearity of difference] 3)} \\ &\le \|A_m - A_n\|_{L(X,Y)}\, \|x\|_X && \text{([Lemma 5.79] 4)} \\ &< \frac{\eta}{\|x\|_X + 1}\, \|x\|_X && \text{([use (2)] 5)} \\ &= \eta \frac{\|x\|_X}{\|x\|_X + 1} && \text{([algebra] 6)} \\ &< \eta. && \text{([}t/(t+1) < 1\text{] 7)} \end{aligned}$$

Therefore $(A_n x)$ is Cauchy in Y for every fixed $x \in X$.

Step 2: use completeness of Y to define the candidate limit. Because Y is Banach, every Cauchy sequence in Y converges in Y. Hence, for every fixed $x \in X$, the sequence $(A_n x)$ has a limit in Y. Define

$$Ax := \lim_{n\to\infty} A_n x. \tag{BS 1.3.1}$$

This defines a map $A : X \to Y$, and it is the one spot where completeness of Y is essential: the limits live in the output space Y, not in the domain X.

Step 3: prove that the pointwise limit is linear. Let $x, z \in X$ and let α, β be scalars. Then

$$\begin{aligned}
A(\alpha x + \beta z) &= \lim_{n\to\infty} A_n(\alpha x + \beta z) && ([\text{definition of } A]\ 3)\\
&= \lim_{n\to\infty} (\alpha A_n x + \beta A_n z) && ([\text{linearity of } A_n]\ 4)\\
&= \alpha \lim_{n\to\infty} A_n x + \beta \lim_{n\to\infty} A_n z && ([\text{limit algebra}]\ 5)\\
&= \alpha Ax + \beta Az. && ([\text{definition of } A]\ 6)
\end{aligned}$$

Thus A is linear.

Step 4: prove that A is bounded. To prove $A \in L(X, Y)$, we need one constant C such that

$$\|Ax\|_Y \le C \|x\|_X \qquad \text{for every } x \in X.$$

Apply the operator-norm Cauchy condition with tolerance 1. There exists $N_0 \in \mathbb{N}$ such that

$$m \ge N_0 \quad \Longrightarrow \quad \|A_m - A_{N_0}\|_{L(X,Y)} < 1. \tag{([Cauchy with tolerance 1]\ 8)}$$

Fix $x \in X$. For every $m \ge N_0$,

$$\begin{aligned}
\|A_m x - A_{N_0} x\|_Y &= \|(A_m - A_{N_0})x\|_Y && ([\text{difference}]\ 9)\\
&\le \|A_m - A_{N_0}\|_{L(X,Y)} \|x\|_X && ([\text{Lemma } 5.79]\ 10)\\
&< \|x\|_X . && ([\text{use (8)}]\ 11)
\end{aligned}$$

Let $m \to \infty$. Since $A_m x \to Ax$ in Y, we get

$$A_m x - A_{N_0} x \to Ax - A_{N_0} x. \tag{([subtract fixed vector]\ 12)}$$

The norm is continuous, so

$$\|Ax - A_{N_0} x\|_Y \le \|x\|_X . \tag{([limit of (11)]\ 13)}$$

Now use the triangle inequality:

$$\begin{aligned}
\|Ax\|_Y &= \|(Ax - A_{N_0}x) + A_{N_0}x\|_Y && ([\text{add and subtract}]\ 14)\\
&\le \|Ax - A_{N_0}x\|_Y + \|A_{N_0}x\|_Y && ([\text{triangle}]\ 15)\\
&\le \|x\|_X + \|A_{N_0}\|_{L(X,Y)} \|x\|_X && ([\text{use (13) and boundedness}]\ 16)\\
&= \left(1 + \|A_{N_0}\|_{L(X,Y)}\right) \|x\|_X . && ([\text{factor}]\ 17)
\end{aligned}$$

So A is bounded. Since Step 3 already proved that A is linear, we have $A \in L(X, Y)$.

Step 5: prove convergence in operator norm. Let $\varepsilon > 0$. Since (A_n) is Cauchy in operator norm, there exists $N \in \mathbb{N}$ such that

$$m, n \geq N \implies \|A_m - A_n\|_{L(X,Y)} < \varepsilon. \qquad \text{([Cauchy choice of } N\text{] 18)}$$

Fix $n \geq N$ and $x \in X$. For every $m \geq N$,

$$\begin{aligned} \|A_m x - A_n x\|_Y &= \|(A_m - A_n)x\|_Y && \text{([difference] 19)} \\ &\leq \|A_m - A_n\|_{L(X,Y)} \|x\|_X && \text{([Lemma 5.79] 20)} \\ &< \varepsilon \|x\|_X . && \text{([use (18)] 21)} \end{aligned}$$

Let $m \to \infty$. Since $A_m x \to Ax$ in Y and the norm is continuous,

$$\|(A - A_n)x\|_Y = \|Ax - A_n x\|_Y \leq \varepsilon \|x\|_X \qquad (x \in X). \qquad \text{(BS 1.3.2)}$$

Taking the supremum over all $\|x\|_X \leq 1$,

$$\begin{aligned} \|A - A_n\|_{L(X,Y)} &= \sup_{\|x\|_X \leq 1} \|(A - A_n)x\|_Y && \text{([op norm] 22)} \\ &\leq \sup_{\|x\|_X \leq 1} \varepsilon \|x\|_X && \text{([use (BS 1.3.2)] 23)} \\ &\leq \varepsilon. && \text{([unit ball] 24)} \end{aligned}$$

Therefore, for each $\varepsilon > 0$, there is $N \in \mathbb{N}$ such that $\|A - A_n\|_{L(X,Y)} \leq \varepsilon$ whenever $n \geq N$. Hence $A_n \to A$ in operator norm. Every Cauchy sequence in $L(X, Y)$ therefore converges to an element of $L(X, Y)$, so $L(X, Y)$ is Banach. □

Corollary 5.82: Dual Spaces Are Banach.

If X is a normed vector space over $\mathbb{F}$, where $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$, then its dual space $X^* = L(X, \mathbb{F})$ is a Banach space.

Proof. Apply Theorem 5.81 with $Y = \mathbb{F}$. The scalar field $\mathbb{R}$ or $\mathbb{C}$ is complete, so the target is Banach, and hence $L(X, \mathbb{F}) = X^*$ is Banach. □

Remark 5.83: Why This Matters. X itself need not be complete for X^* to be. Completeness of X^* is inherited from the scalar target $\mathbb{F}$, which is what makes dual spaces such stable objects: even if X is missing limits, its bounded linear measurements still assemble into a Banach space.

Takeaway. The dual space is more than a catalogue of functionals. It is a Banach space in its own right, its distance given by the operator norm:

$$\|\varphi - \psi\| = \sup_{\|x\|_X \leq 1} |\varphi(x) - \psi(x)| .$$

5.9 Integral Operators and Green's Functions

Big Idea. An integral functional collapses a function down to one number. An integral operator leaves a free variable standing, so its output is a whole new function.

Remark 5.84: Fredholm and Volterra Kernels. Fredholm and Volterra integral equations belong first to the world of *integral operators*, not merely integral functionals.
A Fredholm-type kernel $K(x,t)$ produces an operator

$$(Tf)(x) = \int_a^b K(x,t)f(t)\,dt.$$

A function f goes in and a function of x comes out, so

$$T : X \to Y$$

is a map between function spaces, a point of an operator space like $L(X,Y)$ once it is bounded and linear.
A Volterra-type kernel carries a variable upper limit:

$$(Vf)(x) = \int_a^x K(x,t)f(t)\,dt.$$

Its output still depends on x, so $V : f \mapsto Vf$ is again an operator.

Intuition. The quick test is to look at the output. Function to a number means functional; function to another function means operator.

A functional squeezes all the information in f into one scalar,

$$f \longmapsto \int_0^1 f(t)g(t)\,dt,$$

while an integral operator keeps the variable x in play,

$$f \longmapsto \left(x \mapsto \int_a^b K(x,t)f(t)\,dt\right).$$

A kernel $K(x,t)$ is thus a whole family of functionals, one per fixed x.

Example 5.85: Fixed x Gives a Functional. For each fixed x, the rule

$$f \longmapsto \int_a^b K(x,t)f(t)\,dt$$

is a scalar-valued functional of f. Let x run, collect all those scalars into

$$x \longmapsto (Tf)(x),$$

and the result is the integral operator T. This is why Fredholm and Volterra kernels live in operator theory, whereas the single integral $I_g(f)$ above is only a functional.

■

Example 5.86: Green's Function for an ODE BVP. Consider the boundary value problem

$$-u''(x) = f(x), \qquad u(0) = u(1) = 0.$$

The Green's function $G(x,t)$ is defined by

$$-\frac{\partial^2}{\partial x^2} G(x,t) = \delta_t(x), \qquad G(0,t) = G(1,t) = 0.$$

For this problem,

$$G(x,t) = \begin{cases} x(1-t), & 0 \le x \le t, \\ t(1-x), & t \le x \le 1. \end{cases}$$

Then the solution is

$$u(x) = \int_0^1 G(x,t) f(t)\, dt.$$

Hold x fixed and the rule

$$f \longmapsto u(x) = \int_0^1 G(x,t) f(t)\, dt$$

is a functional of f, returning the single number $u(x)$, the value of the solution at that x. Letting x stay free recovers the full Green's formula:

$$T : f \longmapsto u = Tf, \qquad (Tf)(x) = u(x) = \int_0^1 G(x,t) f(t)\, dt, \qquad 0 \le x \le 1.$$

For the explicit Green's function above, this is

$$u(x) = (1-x) \int_0^x t f(t)\, dt + x \int_x^1 (1-t) f(t)\, dt.$$

This is the integral operator at work, sending the input f to the whole output function u. ■

Intuition. Green's functions bridge point sources and operators. The equation

$$L_x G(x,t) = \delta_t(x)$$

reads: fix the source point t, strike the system with a unit point source there, and record the response as a function of x.

A general forcing $f(t)$ is built up from many tiny point sources. Each source at t provokes the response $G(x,t)$, so the total response is the superposition

$$u(x) = \int G(x,t) f(t)\, dt.$$

The delta sits in the definition of G; the final solution formula is an integral operator.

Takeaway. What lands a Green's formula in operator theory is its output: the whole solution function u, not one scalar value of it.

5.10 Hilbert Spaces and Riesz Representation

Big Idea. A Hilbert space is a complete normed space with inner-product geometry on top. That geometry is exactly what lets every continuous linear functional be written as the inner product with one fixed vector.

Theorem 5.87: Riesz Representation for Hilbert Spaces.

Let H be a Hilbert space. For every continuous linear functional $F \in H^*$, there exists a unique vector $g \in H$ such that

$$F(f) = \langle f, g \rangle \qquad \text{for every } f \in H.$$

Remark 5.88. Riesz is not the claim that every function carries a unique functional in any function space whatsoever. Its content is narrower and sharper: in a Hilbert space, each continuous linear functional is represented uniquely as an inner product with some vector. For $L^2(\Omega)$ this says every $F \in (L^2(\Omega))^*$ has the form

$$F(f) = \int_\Omega f(x)\overline{g(x)}\,dx$$

for one $g \in L^2(\Omega)$. Here g is a function-as-point and F a functional-as-point, and Riesz matches the two roles one-to-one.

Remark 5.89: Inner Products and Functionals in Quantum Mechanics. The physical intuition is right, but it needs stating carefully. An inner product is not a functional on H, since it takes two inputs,

$$\langle \cdot, \cdot \rangle : H \times H \to \mathbb{C}.$$

It is a scalar-valued form. Fix one of its two slots, though, and what remains is a functional of the other vector.
In Dirac notation, a ket $|\psi\rangle$ is a vector in the Hilbert space H. The corresponding bra $\langle\psi|$ is the functional

$$\langle\psi| : H \to \mathbb{C}, \qquad \langle\psi|(|\phi\rangle) = \langle\psi, \phi\rangle.$$

A bra is literally a linear functional on the state space, and the number

$$\langle\psi, \phi\rangle$$

is the value it takes on $|\phi\rangle$.

Intuition. Quantum mechanics turns the Riesz theorem into notation. A state is a vector $|\psi\rangle$, but the same state can probe another state by inner product,

$$|\phi\rangle \longmapsto \langle\psi, \phi\rangle.$$

A vector holds geometric information; a functional asks one scalar question of an input vector. The inner product is the device that turns a fixed vector into such a question. In finite dimensions

this is the move from a column vector to a row vector, and Riesz says the move still goes through in any Hilbert space.

Takeaway. Reserve *function space* for spaces like $C([0,1])$ and $L^2(\Omega)$, whose points are functions, and *dual space* or *space of functionals* for spaces like X^*, whose points are scalar-valued maps on X. In a Hilbert space, Riesz says the two worlds can be identified, but it is worth keeping them apart in one's head.

6 Probability Theory

Foundations to Strengthen.

- Start from probability spaces as measure spaces with total mass one.
- Read random variables as measurable maps and laws as pushforward measures.
- Learn conditional expectation before martingales.
- Separate almost sure convergence, convergence in probability, and L^p convergence.
- Before stochastic calculus, know filtrations, adapted processes, stopping times, and martingales.

Functional-Analysis Lens for This Chapter. A lot of probability becomes easier to organize once random variables are treated as points in function spaces. The translation we lean on is:

$$\text{random variable} \quad \longleftrightarrow \quad \text{measurable function on } (\Omega, \mathcal{F}, \mathbb{P}).$$

With that in hand, $L^p(\Omega)$ collects the random variables whose pth moment is finite:

$$\|X\|_{L^p} = \big(\mathbb{E}|X|^p\big)^{1/p},$$

and L^p convergence is just norm convergence of random variables. The case $L^2(\Omega)$ is a Hilbert space, with inner product

$$\langle X, Y \rangle_{L^2} = \mathbb{E}[XY].$$

This is the reason functional analysis keeps surfacing here. Conditional expectation turns out to be a projection, martingales are projections that fit together over time, the Ito integral is built from L^2 limits, and Markov processes drag in semigroups and generators.

Probability object	Functional-analysis reading
Random variable X	point of a function space such as $L^p(\Omega)$
Expectation $\mathbb{E}[X]$	integral, hence a linear functional when defined
Variance and covariance	norm and inner product after centering
Conditional expectation	projection onto a smaller information subspace
Martingale	process consistent with those projections over time
Ito integral	limit in a norm, usually built first in L^2
Markov process	transition operators, semigroups, and generators

6.1 Probability Space as a Measure Space

Big Idea. Probability theory starts out as measure theory with one normalization: the whole sample space has measure one. We rename μ to $\mathbb{P}$, but nothing structural changes underneath.

Definition 6.1: Probability Space.

A probability space is a triple

$$(\Omega, \mathcal{F}, \mathbb{P}),$$

where

- Ω is the sample space, the set of all possible outcomes;
- $\mathcal{F}$ is a σ-algebra of events, the subsets of Ω whose probabilities are allowed to be measured;
- $\mathbb{P} : \mathcal{F} \to [0, 1]$ is a measure with

$$\mathbb{P}(\Omega) = 1.$$

A probability space is then a measure space whose total mass is one; compare item 4.3.

Remark 6.2: Measure-Theory Lens. Nothing here goes beyond measure theory. We have the same object

$$(\Omega, \mathcal{F}, \mu),$$

carrying the extra normalization $\mu(\Omega) = 1$. What "probability" supplies is interpretation: measurable sets become events, and the measure of an event is read as its probability.

Remark 6.3: Why the Sigma-Algebra Is There. Elementary probability is happy to write $\mathbb{P}(A)$ for any subset $A \subseteq \Omega$. Measure theory forces more care: $\mathbb{P}(A)$ is defined only when $A \in \mathcal{F}$. Think of the σ-algebra as the list of questions about the outcome that the model is allowed to answer.

Example 6.4: Coin Toss. For one coin toss,

$$\Omega = \{H, T\}, \qquad \mathcal{F} = \{\varnothing, \{H\}, \{T\}, \Omega\},$$

and, for a fair coin,

$$\mathbb{P}(\{H\}) = \mathbb{P}(\{T\}) = \frac{1}{2}.$$

A finite measure space, total mass one.

■

6.2 Random Variables as Measurable Maps

Big Idea. There is nothing mysterious about a random variable. It is a function on the outcome space; the randomness is only our ignorance of which $\omega \in \Omega$ actually occurred.

Measure Versus Random Variable. Three objects need to stay separate. A random variable need not be real-valued; in general it lands in some measurable space $(S, \mathcal{S})$.

$$\begin{aligned} \mathbb{P}(A) &: \text{measure an event } A \in \mathcal{F}, \\ X(\omega) &: \text{read a value from an outcome } \omega \in \Omega, \\ (X_{\#}\mathbb{P})(B) &: \text{measure output values } B \in \mathcal{S}. \end{aligned}$$

The random variable X is not the probability measure. It is a map that converts output questions into events:

$$\{X \in B\} = X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\}.$$

The measure $\mathbb{P}$ then assigns probability to that event:

$$\mathbb{P}(X \in B) = \mathbb{P}(X^{-1}(B)).$$

Definition 6.5: Random Variable.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $(S, \mathcal{S})$ be a measurable space. An S-valued random variable, also called a random element, is a measurable map

$$X : (\Omega, \mathcal{F}) \to (S, \mathcal{S}).$$

This is the notion of measurable function from item 4.7. Measurable means that for every measurable output set $B \in \mathcal{S}$,

$$X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}.$$

Every question about the value of X has to correspond to an event in Ω.
When $S = \mathbb{R}$ and $\mathcal{S} = \mathcal{B}(\mathbb{R})$, X is a real-valued random variable. We meet this case first because expectation, variance, and stochastic integrals all need numerical output.

Remark 6.6: Measure-Theory Lens. A random variable is a measurable function, no more. The condition $X^{-1}(B) \in \mathcal{F}$ asks that every observable question about the value of X be answerable inside the original probability space. Probability enters only once we measure those preimages:

$$\mathbb{P}(X \in B) = \mathbb{P}(X^{-1}(B)).$$

Remark 6.7: Inputs of the Measure Versus the Random Variable. The measure and the random variable take different kinds of input, and conflating them causes endless confusion.

For an S-valued random variable,

object	input	output
$\mathbb{P}$	$A \in \mathcal{F}$	$\mathbb{P}(A) \in [0,1]$
X	$\omega \in \Omega$	$X(\omega) \in S$

The measure $\mathbb{P}$ acts on events and answers

$$\text{how likely is the event } A?$$

The random variable X labels each outcome by a value in the target space and answers

$$\text{what value do we observe if the outcome is } \omega?$$

The two meet through preimages. To ask for the probability that X lands in a measurable set $B \in \mathcal{S}$, first convert that output question into an event in Ω:

$$\{X \in B\} = X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\}.$$

Then apply the measure:

$$\mathbb{P}(X \in B) = \mathbb{P}(X^{-1}(B)).$$

The random variable manufactures events out of output questions; the measure assigns probabilities to them.

Remark 6.8: Why Real-Valued Random Variables Appear First. The output need not be a real number. Still, the real-valued case comes first, simply because numbers can be averaged and integrated:

$$X : \Omega \to \mathbb{R}.$$

Plenty of important random objects are not real-valued:

random object	target space
coin result	$\{H, T\}$
random vector	$\mathbb{R}^d$
Brownian path	$C([0,T])$
random matrix	$\mathbb{R}^{m\times n}$

Each is a random variable in the general measurable-map sense. It turns real-valued only after a numerical observation is applied to it: a payoff, a coordinate, a norm, an evaluation map.

Example 6.9: Measure, Random Variable, and Law. For one fair coin toss, let

$$\Omega = \{H, T\}, \qquad \mathbb{P}(\{H\}) = \mathbb{P}(\{T\}) = \frac{1}{2}.$$

Here $\mathbb{P}$ is the measure: it assigns sizes to events like $\{H\}$ and $\{T\}$.
Define a payoff random variable

$$X(H) = 1, \qquad X(T) = -1.$$

This X is not a measure but a real-valued function from outcomes to numbers. The event that the payoff is positive is

$$\{X > 0\} = X^{-1}((0, \infty)) = \{H\}.$$

Therefore

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

The law of X is the pushforward measure $X_{\#}\mathbb{P}$ on $\mathbb{R}$:

$$X_{\#}\mathbb{P} = \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_{1}.$$

Three layers, then: the original measure sits on outcomes, X carries outcomes to numbers, and the law is the induced measure back on the numbers.

■

Remark 6.10: Measurable Versus Integrable. These are two different levels of structure, and it pays to keep them apart.
Measurability is what makes X a legitimate random variable: for every reasonable set of output values $B \subseteq \mathbb{R}$, the question "did X land in B?" must correspond to an event

$$\{\omega : X(\omega) \in B\} \in \mathcal{F}.$$

It governs whether probabilities of output events are even defined.
Integrability is the stronger demand that the average size be finite:

$$\mathbb{E}|X| = \int_{\Omega} |X(\omega)| \, d\mathbb{P}(\omega) < \infty.$$

Only then is $\mathbb{E}[X]$ a finite number. A random variable can be perfectly measurable and still fail to be integrable.

Remark 6.11: Why $X : \Omega \to S$, Not $\mathcal{F} \to S$?. The sample point $\omega \in \Omega$ is the actual outcome, and a random variable assigns a value to each possible one:

$$\omega \longmapsto X(\omega).$$

The σ-algebra $\mathcal{F}$ is not the set of outcomes. It is the collection of events, the subsets of Ω whose probabilities may be measured. So $\mathcal{F}$ is the bookkeeping for questions, while Ω is where the function actually lives.
In one coin toss, for instance, the outcome is H or T, and a payoff random variable might read

$$X(H) = 1, \qquad X(T) = -1.$$

A function on outcomes. The event $\{H\} \in \mathcal{F}$ is what we use to phrase and measure a question like "is $X = 1$?"

Remark 6.12: Distribution. The distribution, or law, of X is the pushforward measure

$$\mathcal{L}_X(B) = \mathbb{P}(X \in B) = \mathbb{P}(X^{-1}(B)).$$

It transports probability from the abstract sample space Ω over to the target S. In the real-valued case $S = \mathbb{R}$, so the law lives on the familiar real line. In the notation of item 4.20,

$$\mathcal{L}_X = X_{\#}\mathbb{P}.$$

Definition 6.13: Expectation.

If X is integrable, its expectation is the integral of X over the probability space:

$$\mathbb{E}[X] = \int_\Omega X(\omega)\, d\mathbb{P}(\omega).$$

Expectation is not some new operation; it is Lebesgue integration against a probability measure. And "integrable" here is the same condition as in item 4.27:

$$\int_\Omega |X|\ d\mathbb{P} < \infty.$$

Remark 6.14: Analysis Lens. Expectation is linear on integrable random variables:

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y].$$

On a space such as $L^1(\Omega)$ it is therefore a linear functional. The probabilistic word is "average"; the analytic object underneath is an integral.

Remark 6.15: Expectation Versus Inner Product. On its own,

$$X \longmapsto \mathbb{E}[X],$$

expectation is a linear functional, not an inner product. The inner product shows up only when two square-integrable random variables are paired:

$$\langle X, Y\rangle_{L^2} = \mathbb{E}[XY] \qquad \text{for real-valued random variables.}$$

For complex-valued random variables, use

$$\langle X, Y\rangle_{L^2} = \mathbb{E}[X\overline{Y}].$$

Strictly, $L^2(\Omega)$ identifies random variables that agree almost surely, and with that identification the pairing is a genuine inner product. After centering, covariance is precisely this inner product:

$$\mathrm{Cov}(X, Y) = \langle X - \mathbb{E}X,\ Y - \mathbb{E}Y\rangle_{L^2}.$$

Takeaway. The dictionary is:

measure theory	probability theory
measure space $(\Omega, \mathcal{F}, \mu)$	probability space $(\Omega, \mathcal{F}, \mathbb{P})$
measurable set	event
measurable function	random variable
integral	expectation

6.3 Distribution Functions and the Limit Theorems

Big Idea. The law of a random variable (the pushforward measure above) says *where* its probability mass sits. The *distribution function* repackages that mass into one increasing curve, and a *density*, when it exists, is its derivative. Two theorems then govern averages of independent samples: the law of large numbers pins the average to the mean, and the central limit theorem says the leftover fluctuation is Gaussian.

6.3.1 The distribution function (CDF)

Definition 6.16: Cumulative Distribution Function.

The *cumulative distribution function* (CDF) of a real random variable X is

$$F_X(x) = \mathbb{P}(X \leq x), \qquad x \in \mathbb{R}.$$

Intuition. On a continuous scale the probability of hitting any single exact value is 0, so asking for $\mathbb{P}(X = x)$ is usually hopeless. The cure is to stop asking about points and sweep instead: start at $-\infty$ and accumulate all the probability up to x. That turns scattered mass into one curve that only ever climbs, from 0 on the far left to 1 on the far right — a single object on which any two distributions can be compared.

Proposition 6.17: Characterizing Properties of a CDF.

A function $F : \mathbb{R} \to [0, 1]$ is the CDF of some random variable if and only if it is

(i) nondecreasing,

(ii) right-continuous, and

(iii) $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to +\infty} F(x) = 1$.

Probabilities then read off directly: $\mathbb{P}(a < X \leq b) = F(b) - F(a)$, and a jump $F(x) - F(x^-) = \mathbb{P}(X = x)$ is the mass sitting at a single point. The CDF determines the law completely — equal F means equal distribution — which is why it is the common currency for comparing laws.

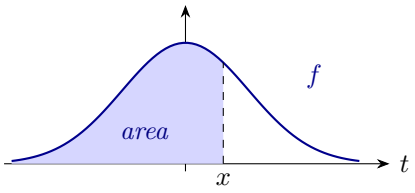
6.3.2 Densities and masses (PDF and PMF)

Definition 6.18: Density and Mass Functions.

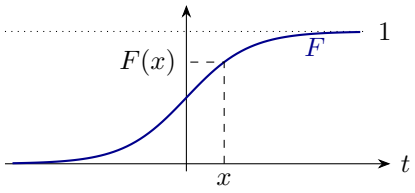
X is *(absolutely) continuous* with *probability density function* (PDF) $f_X \geq 0$ if

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt, \qquad \text{equivalently} \qquad \mathbb{P}(X \in B) = \int_B f_X(t)\,dt,$$

in which case $f_X = F_X'$ almost everywhere and $\int_{\mathbb{R}} f_X = 1$. A *discrete* X instead has a *probability mass function* (PMF) $p_X(x) = \mathbb{P}(X = x)$ with $\sum_x p_X(x) = 1$.



density f: the area up to x...



... is the height $F(x)$, climbing $0 \to 1$

Remark 6.19: Three Kinds of Law. By Lebesgue's decomposition every distribution on $\mathbb{R}$ is a mixture of three pure types: *discrete* (mass at isolated points, a PMF), *absolutely continuous* (smeared out with a density, a PDF), and *singular* (mass on a set of length zero with no atoms — the Cantor distribution is the standard specimen). Models almost always use the first two; densities are convenient precisely because integration replaces summation.

Example 6.20: Reading a Density: the Normal Law. The standard normal $N(0,1)$ has density

$$f(x) = \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2},$$

with mean 0 and variance 1, and *no* elementary CDF: the function $\Phi(x) = \int_{-\infty}^{x} f$ is tabulated, not closed-form. Shifting and scaling gives $N(\mu, \sigma^2)$, with density $\frac{1}{\sigma\sqrt{2\pi}}\, e^{-(x-\mu)^2/2\sigma^2}$. This is the law the central limit theorem below makes universal.

■

6.3.3 The Law of Large Numbers

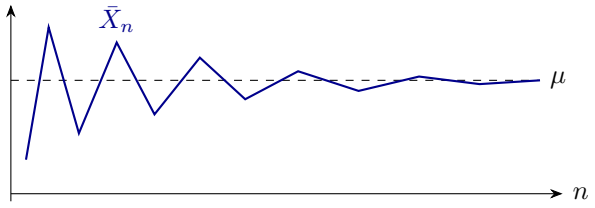
Theorem 6.21: Law of Large Numbers.

Let $X_1, X_2, \dots$ be independent and identically distributed with finite mean $\mathbb{E}[X_i] = \mu$, and write $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ for the sample average. Then

$$\bar{X}_n \longrightarrow \mu$$

as $n \to \infty$ — in probability (the *weak* law), and in fact almost surely (the *strong* law).

Intuition. One draw is noise; an average of many draws is signal. Independent errors partly cancel, and as n grows the cancellation wins, so the random average collapses onto the deterministic mean. This is the precise content of "probability is long-run frequency": a relative frequency is just the average of indicator variables, so it converges to the underlying probability.



early averages swing wildly; as n grows they settle onto the mean μ

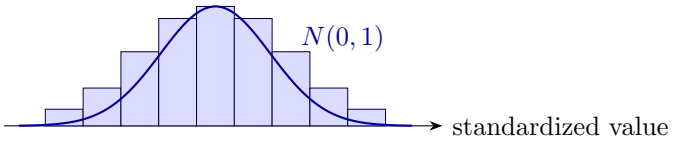
6.3.4 The Central Limit Theorem

Theorem 6.22: Central Limit Theorem.

Let $X_1, X_2, \ldots$ be i.i.d. with mean μ and finite variance $\sigma^2 > 0$. Then the centered, rescaled average converges in distribution to a standard normal:

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} \xrightarrow[n\to\infty]{d} N(0,1).$$

Intuition. The law of large numbers kills the average's fluctuation; the central limit theorem measures what is left. Magnify the leftover wobble by exactly $\sqrt{n}$ and it neither blows up nor vanishes — it settles into a bell curve, *whatever* the shape of the individual X_i. That universality is why the normal law is everywhere, and why the $\sqrt{n}$ scaling — not n — is the heartbeat of diffusion: it is the same $\sqrt{\Delta t}$ that builds Brownian motion and turns the binomial tree into geometric Brownian motion (section 6.6).



whatever the shape of each draw, the rescaled sum piles up into the bell

Takeaway. The CDF is the universal description of a law and always exists; a PDF or PMF is the convenient special case (a density, or point masses). Across many independent samples the average locks onto the mean (LLN) and its $\sqrt{n}$-magnified error is Gaussian (CLT) — the two facts that make averages predictable and the normal law inevitable.

6.4 Stochastic Processes as Families of Random Variables

Big Idea. A stochastic process is a random object that evolves in time. Formally it is a family of random variables indexed by time,

$$(X_t)_{t\in T},$$

or equivalently a map of two variables,

$$X : T \times \Omega \to \mathbb{R}, \qquad (t,\omega) \mapsto X(t,\omega).$$

The variables play different roles: t is time, ω is the outcome of the experiment.

Definition 6.23: Stochastic Process.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let T be an index set, usually $T = \mathbb{N}$, $T = \mathbb{Z}$, or $T = [0,\infty)$. A stochastic process is a family

$$(X_t)_{t\in T}$$

such that each

$$X_t : \Omega \to \mathbb{R}$$

is a random variable. Equivalently, a process is a map

$$X : T \times \Omega \to \mathbb{R}, \qquad X_t(\omega) = X(t,\omega).$$

Remark 6.24: Function-Space Lens. Fix t and X_t is a point of a space like $L^p(\Omega)$. Fix ω and the map $t \mapsto X(t,\omega)$ is a path. A stochastic process can therefore be read as a curve in a function space,

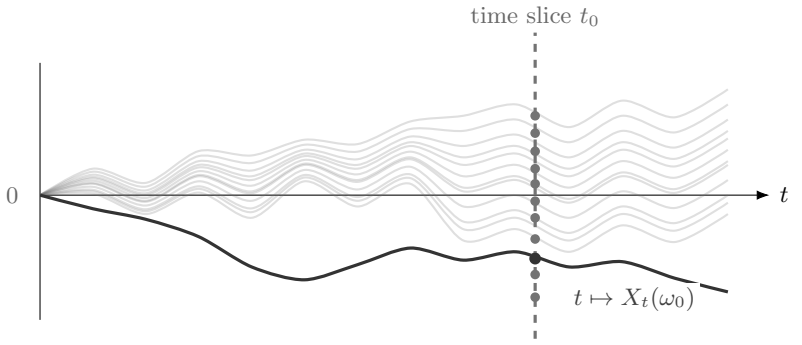
$$t \longmapsto X_t \in L^p(\Omega),$$

or as a random element of a path space such as $C([0,T])$. A good part of stochastic analysis is keeping track of which reading is in play at any moment.

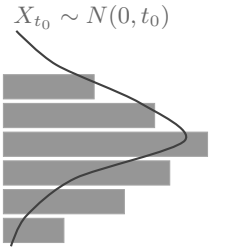
Reading a stochastic process $X(t, \omega)$

The same process reads two ways: along time for a single outcome, or vertically at one time across many outcomes.

Paths: vary ω, vary t
the whole process



Fix $t = t_0$
X_{t_0} is one random variable



ω-slice: horizontal read

Pick one outcome ω_0. Then $t \mapsto X_t(\omega_0)$ is a single realized sample path. For Brownian motion the path sits in $C([0, T])$, yet it is almost surely nowhere differentiable.

t-slice: vertical read

Freeze time at t_0. The map $\omega \mapsto X(t_0, \omega) = X_{t_0}(\omega)$ is a random variable, and where it cuts across all the paths is its distribution.

Takeaway. The notation X_t conceals two readings. Fixing t leaves a random variable on Ω; fixing ω leaves one path through time. Stochastic calculus has to handle both at once.

Remark 6.25: The Two Slices. Two slices through the same picture.

$$\begin{aligned} &\text{Fix time } t = t_0: && \omega \mapsto X(t_0, \omega) \text{ is a random variable.} \\ &\text{Fix outcome } \omega = \omega_0: && t \mapsto X(t, \omega_0) \text{ is one realized path.} \end{aligned}$$

This is the mental model to hold on to: a process is not a single function of time but a random family of them.

Example 6.26: Random Walk. Let $\xi_1, \xi_2, \dots$ be independent random variables with

$$\mathbb{P}(\xi_k = 1) = \mathbb{P}(\xi_k = -1) = \frac{1}{2}.$$

Define

$$S_n = \sum_{k=1}^{n} \xi_k.$$

Then $(S_n)_{n \geq 0}$ is a discrete-time stochastic process. Fix n and S_n is a random variable; fix an outcome ω and the sequence

$$n \mapsto S_n(\omega)$$

is one path of the walk.

■

A Small Catalogue of Common Processes. Once a process is seen as a family $(X_t)_{t \in T}$, the next question is what structure the family carries. Named processes earn their names by imposing a particular rule, usually on dependence, increments, or path regularity.

Stochastic process

a family $(X_t)_{t\in T}$; each fixed t gives a random variable, and each fixed ω gives one path

Discrete time $T = \mathbb{N}$ or $\mathbb{Z}$

White noise. Shocks ε_t have no linear memory: $\mathbb{E}\varepsilon_t = 0$, and $\mathrm{Cov}(\varepsilon_t, \varepsilon_s) = 0$ for $t \neq s$.

Random walk. The level accumulates shocks: $X_t = X_0 + \sum_{k=1}^{t} \varepsilon_k$. It remembers the past through its current level.

AR/MA/ARMA. AR remembers levels; MA remembers shocks. A simple mixed form is $X_t = \phi X_{t-1} + \varepsilon_t + \theta\varepsilon_{t-1}$.

ARCH/GARCH. The variance is dynamic: $X_t = \sigma_t\varepsilon_t$. The size of tomorrow's shock depends on the past.

Continuous time $T = [0, \infty)$

Poisson process. A jump-count process: $N_t - N_s \sim \mathrm{Poisson}(\lambda(t-s))$. Paths are step functions with jumps of size 1.

Brownian motion. Continuous Gaussian noise: $B_t - B_s \sim N(0, t-s)$. Paths are continuous but almost surely nowhere differentiable.

Diffusion / stochastic ODE. A random IVP driven by Brownian noise: $dX_t = b(X_t, t)\,dt + \sigma(X_t, t)\,dB_t$.

OU, GBM, CIR. Same diffusion template, different b and σ. The solution is the whole process $(X_t)_{t\geq 0}$, not a single number.

Structural labels properties can overlap

Markov. The present state summarizes the past: $\mathbb{P}(\text{future} \mid \mathcal{F}_t) = \mathbb{P}(\text{future} \mid X_t)$.

Martingale. Zero predictable drift: $\mathbb{E}[X_t \mid \mathcal{F}_s] = X_s$ for $s \leq t$.

Gaussian process. Every finite vector has a multivariate normal law: $(X_{t_1}, \ldots, X_{t_n})$ is Gaussian.

Stationary. The finite-dimensional distributions do not change under time shifts.

These categories overlap freely. Brownian motion is continuous-time, Gaussian, Markov, and a martingale all at once; a random walk is discrete-time, Markov, and often a martingale.

Definition 6.27: White Noise.

A white-noise process is a sequence $(\varepsilon_t)_{t\in\mathbb{Z}}$ with

$$\mathbb{E}[\varepsilon_t] = 0, \qquad \operatorname{Cov}(\varepsilon_t, \varepsilon_s) = 0 \quad (t \neq s).$$

It is the prototype of a process with no linear memory. Many time-series models begin with white noise and then run it through a filter.

Definition 6.28: Random Walk.

A random walk accumulates independent shocks:

$$X_t = X_{t-1} + \varepsilon_t, \qquad \text{so} \qquad X_t = X_0 + \sum_{k=1}^{t} \varepsilon_k.$$

Because the level remembers every past shock, it is usually not stationary. The increments, on the other hand, are simple:

$$X_t - X_{t-1} = \varepsilon_t.$$

Brownian motion is the continuous-time scaling limit of this idea.

Definition 6.29: Markov Process.

A Markov process is one whose future depends on the past only through the present. In discrete time,

$$\mathbb{P}(X_{n+1} \in A \mid X_n, X_{n-1}, \ldots) = \mathbb{P}(X_{n+1} \in A \mid X_n).$$

The present state summarizes everything the past has to offer. On a finite state space the central object is a transition matrix; in continuous time it becomes a transition semigroup with its generator.

Definition 6.30: Martingale.

A martingale is a process whose conditional future value equals its current value:

$$\mathbb{E}[X_{t+1} \mid \mathcal{F}_t] = X_t.$$

Informally it is a fair game relative to the information $\mathcal{F}_t$, which we make precise in the next subsection. A martingale's increments need not be independent; the statement is only that the predictable drift, given current information, vanishes.

Example 6.31: A Random Walk Is a Martingale. Return to the random walk

$$X_t = X_0 + \sum_{k=1}^{t} \varepsilon_k,$$

and write $\mathcal{F}_t = \sigma(\varepsilon_1, \ldots, \varepsilon_t)$ for the information carried by the first t shocks. Suppose the increments are integrable and centered given the past:

$$\mathbb{E}\,|\varepsilon_t| < \infty \qquad \text{and} \qquad \mathbb{E}[\varepsilon_t \mid \mathcal{F}_{t-1}] = 0.$$

The simple symmetric walk, where each ε_k equals $+1$ or -1 with probability $\frac{1}{2}$, is the standard case. Since X_t is built from $\varepsilon_1, \ldots, \varepsilon_t$ it is adapted and integrable, and

$$\mathbb{E}[X_t \mid \mathcal{F}_{t-1}] = \mathbb{E}[X_{t-1} + \varepsilon_t \mid \mathcal{F}_{t-1}] = X_{t-1} + \mathbb{E}[\varepsilon_t \mid \mathcal{F}_{t-1}] = X_{t-1}.$$

Two properties of conditional expectation carry the computation. The level X_{t-1} is already known at time $t-1$, hence $\mathcal{F}_{t-1}$-measurable, so it passes through the conditional expectation unchanged. The new shock ε_t has conditional mean zero, so it adds nothing. What is left is X_{t-1}, which is the martingale property.

■

Remark 6.32: Where the Fair Game Lives. What matters is the centered increment, not independence. If instead $\mathbb{E}[\varepsilon_t \mid \mathcal{F}_{t-1}] = \mu$, the same line gives

$$\mathbb{E}[X_t \mid \mathcal{F}_{t-1}] = X_{t-1} + \mu,$$

a submartingale when $\mu > 0$ and a supermartingale when $\mu < 0$. The martingale case is the one with no predictable drift: given everything known today, the best forecast of tomorrow's level is today's level. That is the precise content of "fair game."

Definition 6.33: Poisson Process.

A Poisson process $(N_t)_{t \geq 0}$ counts random arrivals. It starts at $N_0 = 0$, has independent increments, and for rate $\lambda > 0$ satisfies

$$N_t - N_s \sim \text{Poisson}(\lambda(t-s)), \qquad 0 \leq s < t.$$

Its paths are step functions that jump by one. This is the model to reach for when the randomness lives in event counts or arrival times.

Definition 6.34: Gaussian Process.

A Gaussian process is a process $(X_t)_{t \in T}$ such that every finite vector

$$(X_{t_1}, \ldots, X_{t_n})$$

has a multivariate Gaussian distribution. Its law is determined by its mean function

$$m(t) = \mathbb{E}[X_t]$$

and covariance kernel

$$K(s,t) = \mathrm{Cov}(X_s, X_t).$$

Brownian motion is one Gaussian process among many; the converse does not hold.

Definition 6.35: Brownian Motion.

Brownian motion is the canonical continuous random path. It starts at 0, has independent Gaussian increments, and satisfies

$$B_t - B_s \sim N(0, t-s), \qquad 0 \le s < t.$$

Its paths are continuous but almost surely nowhere differentiable. This roughness is exactly what forces Ito integration in place of ordinary calculus.

Remark 6.36: Time-Series Processes. The standard discrete-time processes are best read as filters of white noise. An MA(q) process has finite shock memory:

$$X_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}.$$

An AR(p) process has memory in the level:

$$X_t = c + \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + \varepsilon_t.$$

An ARMA process combines the two. ARCH and GARCH push the memory into the volatility instead, so the size of the next shock depends on the past even while its sign stays unpredictable.

Remark 6.37: Continuous-Time Diffusions. Many continuous-time models are written as stochastic differential equations. For now, read the formulas as names for a few recurring shapes of random motion:

Process	Typical form	Main idea
Ornstein–Uhlenbeck	$dX_t = \kappa(\theta - X_t)\,dt + \sigma\,dB_t$	Gaussian mean reversion
Geometric Brownian motion	$dS_t = \mu S_t\,dt + \sigma S_t\,dB_t$	positive multiplicative noise
CIR process	$dV_t = \kappa(\theta - V_t)\,dt + \xi\sqrt{V_t}\,dB_t$	nonnegative mean reversion

These will land better once filtrations, Brownian motion, Ito integration, and SDEs are on the table.

Takeaway. "Process" does not name a model. It names a format: a family of random variables indexed by time. Which process appears depends on the rule imposed on dependence –

independent noise, accumulated shocks, Markov dependence, fair-game behavior, and so on.

6.5 Filtrations: Information Growing in Time

Big Idea. Stochastic calculus is about random paths, but just as much about what is known at each instant. A filtration is the device that records that growing information.

Definition 6.38: Filtration.

A filtration is an increasing family of σ-algebras

$$(\mathcal{F}_t)_{t\geq 0}$$

such that

$$s \leq t \qquad \Longrightarrow \qquad \mathcal{F}_s \subseteq \mathcal{F}_t.$$

Read $\mathcal{F}_t$ as the collection of events whose truth can be settled from information available up to time t.

Remark 6.39: Measure-Theory Lens. A filtration is a sequence of σ-algebras, not of numbers. A larger σ-algebra contains more measurable sets, so the inclusion

$$\mathcal{F}_s \subseteq \mathcal{F}_t$$

says later times can observe more events. Information, in this language, is just measurability.

Definition 6.40: Adapted Process.

A process $(X_t)_{t\geq 0}$ is adapted to a filtration $(\mathcal{F}_t)_{t\geq 0}$ if

$$X_t \text{ is } \mathcal{F}_t\text{-measurable for every } t.$$

In words, the value X_t is already known at time t: an adapted process never peeks ahead.

Intuition. Adapted is best read as *non-anticipating*, and the opposite is easy to picture: $Y_t := B_1$, the Brownian value at time 1 held for all t, is *not* adapted — at $t = \frac{1}{2}$ it already "knows" B_1, which lies in the future, outside $\mathcal{F}_{1/2}$.

So how does one ever *know* that a given process is adapted? Anything assembled from the path only up to the present is adapted automatically. A running integral of Brownian increments, $\int_0^t H_s\, dB_s$, uses increments with $s \leq t$ and nothing later, so its value at time t depends on $\mathcal{F}_t$ alone. This is exactly why the solution of a stochastic differential equation,

$$X_t = X_0 + \int_0^t b\, ds + \int_0^t \sigma\, dB_s,$$

is adapted *by construction*: by time t it has accumulated only the increments up to t, so it cannot depend on anything later. Indeed the Ito integral is *defined* only for adapted integrands, so adaptedness is already baked into what these objects mean.

Remark 6.41: Measure-Theory Lens. Adaptedness is, once more, a measurability condition. That X_t is $\mathcal{F}_t$-measurable means each event

$$\{X_t \in B\}$$

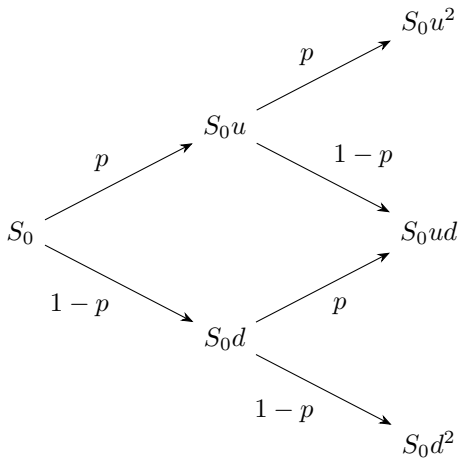
lies inside the information available by time t, which is how "no future information" turns into something precise.

Remark 6.42: Why This Matters. An ordinary integral may use the whole function at once. A stochastic integral cannot: its integrand has to be adapted, so the decision at time t about how much to multiply by dB_t draws on the past and present but never on future Brownian motion. That is the precise content of "no future information."

6.6 The Binomial Tree: A Discrete Stock Model

Big Idea. The binomial tree is the smallest complete probabilistic model of a stock: finite, fully explicit, and yet already carrying a sample space, a σ-algebra, a filtration, a martingale, and a pricing rule. It is the discrete shadow of the Brownian-motion models in the rest of the chapter — refine the time step and it converges to geometric Brownian motion.

The Model. Start from a price S_0. In each period the price is multiplied by one of two factors: *up* by $u > 1$ with probability p, or *down* by $0 < d < 1$ with probability $1 - p$, independently across periods. Over n periods the prices fan out into a tree. Because an up–down move and a down–up move land on the same price ($ud = du$), the *prices* recombine, even though the two *paths* stay distinct.



Example 6.43: The Probability Space of a Binomial Tree. Take the two-period tree (this is the exercise solved in section 12). Each outcome is the record of two moves, so the sample space is the set of length-two words

$$\Omega = \{U, D\}^2 = \{UU,\ UD,\ DU,\ DD\}, \qquad |\Omega| = 4.$$

Since Ω is finite, take the σ-algebra to be the full power set

$$\mathcal{F} = 2^{\Omega}, \qquad |\mathcal{F}| = 2^4 = 16,$$

so every subset counts as an event. The moves are independent, up with probability p and down with $1 - p$, so each path's probability is the product of its step probabilities:

$$\mathbb{P}(UU) = p^2, \qquad \mathbb{P}(UD) = \mathbb{P}(DU) = p(1-p), \qquad \mathbb{P}(DD) = (1-p)^2,$$

which sum to $\big(p + (1-p)\big)^2 = 1$. The probability of any event is the sum of the path probabilities it contains. For n periods the same recipe gives $\Omega = \{U, D\}^n$, $\mathcal{F} = 2^{\Omega}$, and $\mathbb{P}(\omega) = p^{\#\text{up}}(1-p)^{\#\text{down}}$.

■

Remark 6.44: Paths Versus Prices. The space Ω has *four* outcomes, but the terminal *price* S_2 takes only *three* values, because UD and DU both give S_0ud. So

$$\mathbb{P}(S_2 = S_0u^2) = p^2, \quad \mathbb{P}(S_2 = S_0ud) = 2p(1-p), \quad \mathbb{P}(S_2 = S_0d^2) = (1-p)^2.$$

The recombining picture is about prices; the probability space still tracks every path.

The Filtration and the Price Process. Information arrives one move at a time — which is exactly a filtration. Before any move nothing is known, $\mathcal{F}_0 = \{\emptyset, \Omega\}$. After the first move only its direction is known,

$$\mathcal{F}_1 = \sigma(\text{first move}) = \big\{\emptyset,\ \{UU, UD\},\ \{DU, DD\},\ \Omega\big\},$$

and after the second everything is known, $\mathcal{F}_2 = 2^\Omega = \mathcal{F}$. The price after k moves,

$$S_k = S_0\, u^{\#\text{ups in first } k}\, d^{\#\text{downs in first } k},$$

is $\mathcal{F}_k$-measurable: an *adapted* process, fixed by what is known at time k and no more.

Risk-Neutral Pricing in One Step. Here is the payoff of all this structure. Say one dollar held over a period grows to the risk-free factor $e^{r\,\Delta t}$, and assume no arbitrage, $d < e^{r\,\Delta t} < u$. Then there is a unique probability $q \in (0,1)$ making the *discounted* stock a martingale, $e^{-r\,\Delta t}\,\mathbb{E}^{\mathbb{Q}}[S_1] = S_0$, namely

$$q = \frac{e^{r\,\Delta t} - d}{u - d}.$$

Under this *risk-neutral* measure $\mathbb{Q}$, any derivative paying V_1 next period is worth its discounted expected payoff,

$$V_0 = e^{-r\,\Delta t}\,\mathbb{E}^{\mathbb{Q}}[V_1] = e^{-r\,\Delta t}\big(q\,V_u + (1-q)\,V_d\big).$$

The striking part: the real-world probability p has *vanished* from the price. Multi-period values follow by applying this one-step rule backward through the tree — backward induction.

Remark 6.45: The Bridge to Brownian Motion. Fix a horizon T, cut it into n steps of length $\Delta t = T/n$, and scale the factors as $u = e^{\sigma\sqrt{\Delta t}}$ and $d = e^{-\sigma\sqrt{\Delta t}}$. Then $\log S_k$ is a rescaled random walk, and by the central limit theorem it converges, as $n \to \infty$, to a Brownian motion with drift — so the stock price converges to geometric Brownian motion and the binomial price converges to the Black–Scholes price. This is the Cox–Ross–Rubinstein model, and it is why the discrete tree and the continuous SDE tell the same story.

Takeaway. Everything in this chapter is already visible in the binomial tree, with no limits taken: $(\Omega, \mathcal{F}, \mathbb{P})$ is four points and their power set, the filtration is the move-by-move reveal, the discounted price is a martingale under $\mathbb{Q}$, and a derivative's price is a discounted risk-neutral expectation. The continuous theory is this same skeleton with the time step sent to zero.

6.7 Brownian Motion

Big Idea. Brownian motion is the canonical continuous-time random path: the limit object sitting behind random walks, and the noise source that drives most stochastic differential equations.

Definition 6.46: Brownian Motion.

A process $(B_t)_{t\geq 0}$ is a standard Brownian motion if

- $B_0 = 0$;
- $B_t - B_s \sim N(0, t-s)$ whenever $0 \leq s < t$;
- increments over disjoint time intervals are independent;
- the paths $t \mapsto B_t(\omega)$ are continuous with probability one.

Remark 6.47: Analysis Lens. Brownian motion wears two hats at once. Analytically, its sample paths live in $C([0,T])$ with probability one; probabilistically, each fixed B_t is a Gaussian random variable. The tension between them is that the paths stay continuous while being far rougher than any differentiable curve.

Remark 6.48: The Strange Part. Brownian paths are continuous, yet with probability one nowhere differentiable. The expression

$$\frac{dB_t}{dt}$$

simply has no meaning as an ordinary derivative. This single obstruction is what splits stochastic calculus off from the ordinary kind.

Definition 6.49: Quadratic Variation.

For a partition

$$0 = t_0 < t_1 < \cdots < t_n = T,$$

the quadratic variation sum of a path is

$$\sum_{k=0}^{n-1} \left(B_{t_{k+1}} - B_{t_k}\right)^2.$$

For Brownian motion, as the mesh of the partition goes to zero,

$$\sum_{k=0}^{n-1} \left(B_{t_{k+1}} - B_{t_k}\right)^2 \longrightarrow T$$

in probability, and in stronger senses under appropriate refinements.

Remark 6.50: Analysis Lens. Quadratic variation gauges roughness at second order. A smooth path has quadratic variation zero; a Brownian path has it nonzero. Here is the analytic reason the two calculi part ways: Brownian increments are small, but not so small that their squares wash out once summed.

Remark 6.51: Why Ordinary Calculus Fails. For a smooth path $x(t)$, the quadratic variation is zero:

$$\sum (x(t_{k+1}) - x(t_k))^2 \to 0.$$

For Brownian motion, the quadratic variation is nonzero:

$$[B]_T = T.$$

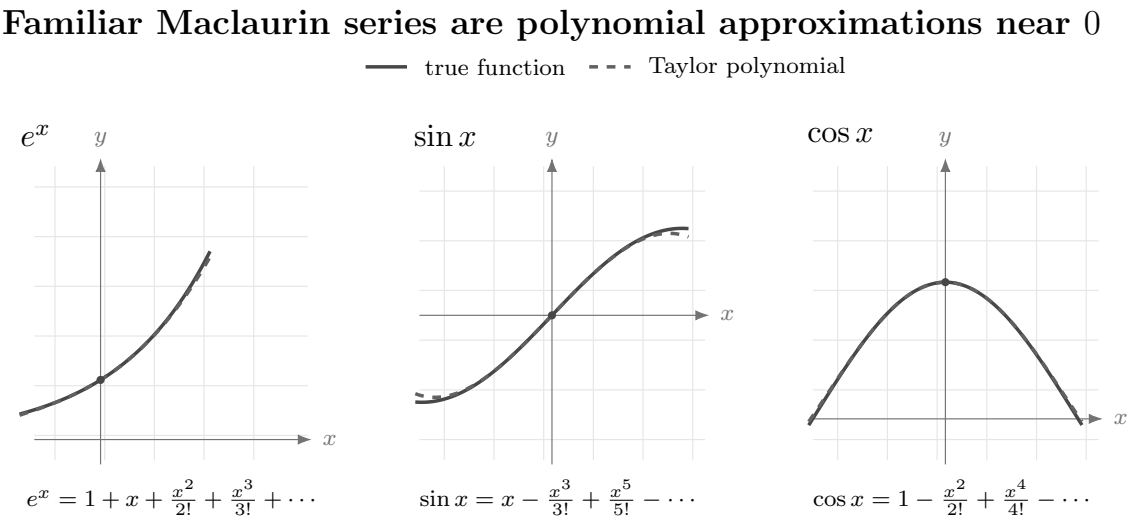
That one fact rewrites the chain rule: in stochastic calculus the second-derivative term refuses to vanish.

Remark 6.52: The Taylor Expansion Term That Does Not Vanish. The Ito correction comes straight out of Taylor expansion. The series one first remembers are the famous ones for transcendental functions, expanded around 0:

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots,$$

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots, \qquad \cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \cdots.$$

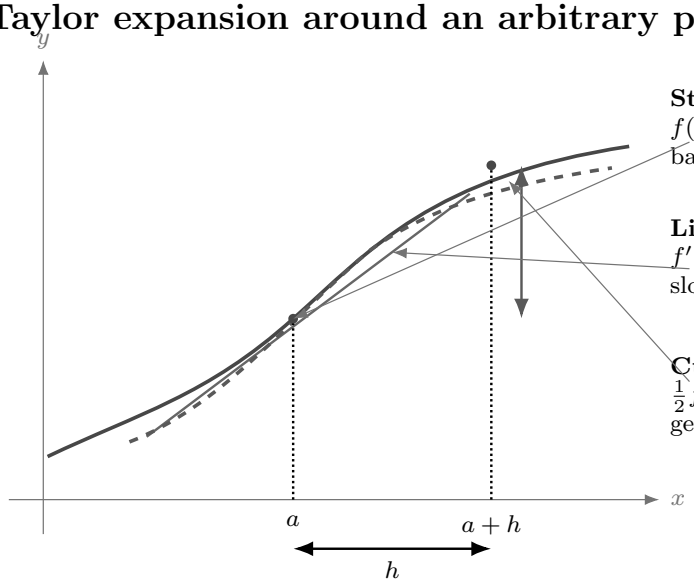
All of these are special cases of one idea: near a point, a smooth function is well approximated by a polynomial whose coefficients are its derivatives. Expanding around 0 goes by the name Maclaurin; expanding around an arbitrary point a is the general Taylor expansion.



For stochastic calculus the memorized series for e^x, $\sin x$, and $\cos x$ are beside the point. What we want is the local expansion of an arbitrary smooth f. For small h, around a point a,

$$f(a+h) = f(a) + f'(a)h + \frac{1}{2}f''(a)h^2 + \frac{1}{6}f'''(a)h^3 + \cdots.$$

Taylor expansion around an arbitrary point a



After subtracting $f(a)$, the increment form is

$$f(a+h) - f(a) = f'(a)h + \frac{1}{2}f''(a)h^2 + \frac{1}{6}f'''(a)h^3 + \cdots .$$

The zeroth-order term $f(a)$ drops out only because we are tracking the change in f rather than f itself.
Apply this to a path. On a small interval $[t_k, t_{k+1}]$, write

$$\Delta t_k = t_{k+1} - t_k, \qquad \Delta B_k = B_{t_{k+1}} - B_{t_k}.$$

Taylor expansion with

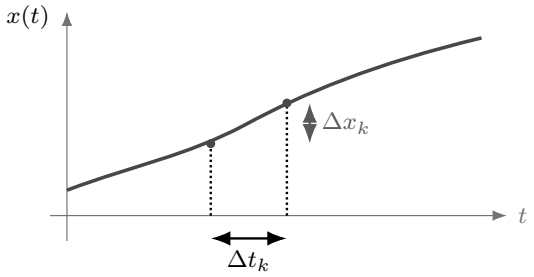
$$a = B_{t_k}, \qquad h = \Delta B_k$$

gives, informally,

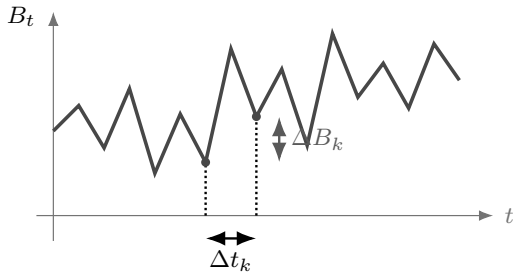
$$f(B_{t_{k+1}}) - f(B_{t_k}) \approx f'(B_{t_k})\Delta B_k + \frac{1}{2}f''(B_{t_k})(\Delta B_k)^2 + \frac{1}{6}f'''(B_{t_k})(\Delta B_k)^3 + \cdots .$$

Putting Taylor on a path: $a = B_{t_k}$ and $h = \Delta B_k$

Ordinary smooth path
increments are proportional to time



Brownian path
increments are typically much larger



What survives after summing over a fine partition

$f'(B_{t_k})\Delta B_k$	$\frac{1}{2}f''(B_{t_k})(\Delta B_k)^2$	$\frac{1}{6}f'''(B_{t_k})(\Delta B_k)^3$
size $\sqrt{\Delta t_k}$	size Δt_k	size $(\Delta t_k)^{3/2}$
↓	↓	↓
stochastic integral	Ito correction	vanishes in the limit
$\int f'(B_t)\,dB_t$	$\frac{1}{2}\int f''(B_t)\,dt$	

For an ordinary smooth path $x(t)$, the increment satisfies

$$\Delta x_k = x(t_{k+1}) - x(t_k) \approx x'(t_k)\Delta t_k.$$

The first-order term has size Δt_k, so the square has size $(\Delta t_k)^2$. Summing over a fine partition,

$$\sum_k (\Delta x_k)^2 \approx \sum_k (x'(t_k))^2(\Delta t_k)^2 \le \max_k \Delta t_k \sum_k (x'(t_k))^2 \Delta t_k \longrightarrow 0.$$

The second-order term vanishes in the limit, which is exactly why ordinary calculus can keep only the first-order one.
Brownian motion scales differently. Its increments satisfy

$$\Delta B_k \sim N(0, \Delta t_k),$$

so their typical size is

$$\Delta B_k \text{ has size } \sqrt{\Delta t_k}.$$

Therefore

$$(\Delta B_k)^2 \text{ has size } \Delta t_k, \qquad (\Delta B_k)^3 \text{ has size } (\Delta t_k)^{3/2}.$$

When these terms are summed over roughly $1/\Delta t$ intervals, the power counting is:

Taylor term	size of one increment	size after summing
ΔB_k	$(\Delta t_k)^{1/2}$	stochastic integral term
$(\Delta B_k)^2$	Δt_k	order 1 term
$(\Delta B_k)^3$	$(\Delta t_k)^{3/2}$	vanishes

More precisely,

$$\sum_k (\Delta B_k)^2 \to T,$$

while the higher-order Taylor terms die in the Ito limit. The second-order term thus survives as a genuine first-order-in-time contribution:

$$\frac{1}{2}\sum_k f''(B_{t_k})(\Delta B_k)^2 \rightsquigarrow \frac{1}{2}\int_0^T f''(B_t)\,dt.$$

That surviving second derivative is the Ito correction:

$$df(B_t) = f'(B_t)\,dB_t + \frac{1}{2}f''(B_t)\,dt.$$

The formula is nothing more than ordinary Taylor expansion together with the special quadratic variation of Brownian motion.

6.8 Ito Integration

Big Idea. The stochastic integral

$$\int_0^T H_t \, dB_t$$

cannot be defined by ordinary Riemann integration – Brownian motion is too rough for that. Instead it is built first for simple adapted processes and then extended by an L^2 limit.

What the Integrand H_t Is. H_t is the *integrand*: the process being integrated against Brownian motion, the stochastic counterpart of the $h(t)$ in $\int h(t)\,dt$. Like X_t it is a process, $H_t = H_t(\omega)$, one random variable for each time. The one real requirement is that it be *adapted*: H_t has to be decided from information available up to time t, with no glimpse of the future.

The trading picture is the cleanest. Read H_t as the position held at time t — shares of an asset whose price is B_t — so that $\int_0^T H_t \, dB_t$ is the total gain from running that strategy through the random price moves.

The symbol H_{t_k} is just H_t sampled at the *left* endpoint t_k of the partition piece $[t_k, t_{k+1}]$. The endpoint matters here in a way it never does for ordinary integrals: the Ito integral fixes the position H_{t_k} at the *start* of the interval, before the increment $B_{t_{k+1}} - B_{t_k}$ is revealed. That non-anticipating choice — bet first, then watch the coin — is what makes $\int_0^T H_t \, dB_t$ a martingale, and it is exactly what separates the Ito integral from the Stratonovich one, which samples the midpoint instead.

Whenever a dB term appears, H is simply whatever multiplies it. In Ito's formula the stochastic part is $\int_0^t \sigma(s, X_s) f_x(s, X_s)\, dB_s$, so there $H_s = \sigma(s, X_s) f_x(s, X_s)$.

Remark 6.53: Do Not Read dB_t Like dt or $d\mathbb{P}$. The notation

$$\int_0^T H_t \, dB_t$$

looks like $\int h(t)\,dt$ or $\int X \, d\mathbb{P}$, but dB_t plays a different role.

symbol	kind of object	what it does
dt	time measure	weights ordinary time integrals
$d\mathbb{P}$	probability measure	weights expectations
dB_t	Brownian increment	drives stochastic sums

For a small time interval,

$$dB_t \quad \text{means heuristically} \quad B_{t+dt} - B_t.$$

So the Ito integral starts life as a sum of random increments:

$$\sum_k H_{t_k}\big(B_{t_{k+1}} - B_{t_k}\big).$$

The probability measure $d\mathbb{P}$ comes in when we measure the size of the resulting random variable, usually in $L^2(\Omega, d\mathbb{P})$. The time measure dt comes in through the norm of the integrand:

$$H \in L^2(\Omega \times [0,T], d\mathbb{P}\, dt) \quad \Longrightarrow \quad \int_0^T H_t \, dB_t \in L^2(\Omega, d\mathbb{P}).$$

That split is why the Ito integral is constructed by quadratic-mean limits rather than by pathwise Riemann integration.

Remark 6.54: Visual: Area Versus Ito Accumulation. Fix an outcome ω. The realized Brownian path

$$t \mapsto B_t(\omega)$$

is continuous, so the ordinary time integral

$$\int_0^T B_t(\omega)\, dt$$

reads as signed area under that path – a plain dt-integral.
The Ito integral

$$\int_0^T H_t\, dB_t$$

means something else. Built from the products

$$H_{t_k}\big(B_{t_{k+1}} - B_{t_k}\big),$$

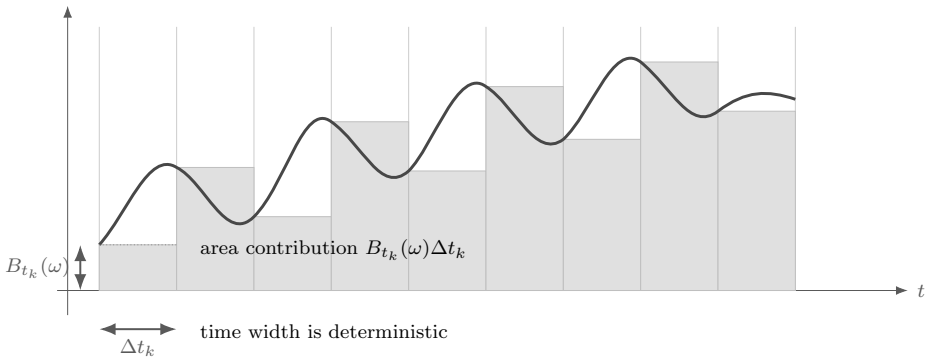
it accumulates Brownian increments weighted by the current exposure H_{t_k}. The word "exposure" carries the point: H_{t_k} is locked in before the next Brownian increment shows up.

Two different limits: area against dt, versus accumulation against dB_t

A. Ordinary time integral: rectangles approximate signed area

Fix one realized path $t \mapsto B_t(\omega)$. A Riemann sum takes width Δt_k and height $B_{t_k}(\omega)$.

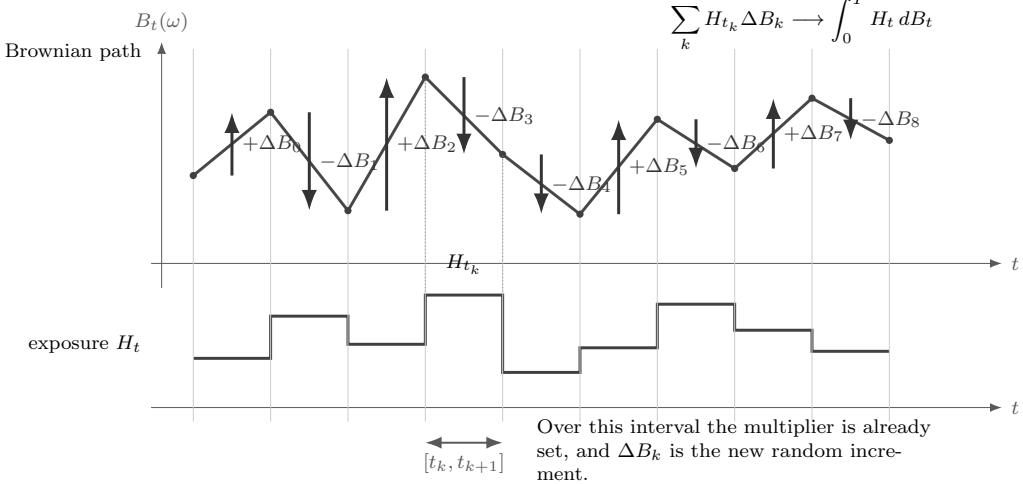
$$\sum_k B_{t_k}(\omega)\Delta t_k \longrightarrow \int_0^T B_t(\omega)\,dt$$



B. Ito integral: Brownian increments weighted by left-endpoint exposure

The summand on $[t_k, t_{k+1}]$ is $H_{t_k}(B_{t_{k+1}} - B_{t_k})$. The exposure H_{t_k} is fixed from information known at t_k, before the next shock arrives.

$$\sum_k H_{t_k}\Delta B_k \longrightarrow \int_0^T H_t\,dB_t$$



Thus

$$\int_0^T H_t\,dt \quad \text{is area over time,}$$

$$\int_0^T H_t\,dB_t \quad \text{accumulates } H_t\text{-weighted Brownian increments.}$$

In the second, H_t is the exposure chosen from current information and dB_t is the fresh Brownian increment it weights.

Definition 6.55: Ito Integral for Step Processes.

Let H_t be an adapted step process of the form

$$H_t = \sum_{k=0}^{n-1} H_k \, 1_{(t_k, t_{k+1}]}(t),$$

where each H_k is $\mathcal{F}_{t_k}$-measurable. Define

$$\int_0^T H_t \, dB_t = \sum_{k=0}^{n-1} H_k \big(B_{t_{k+1}} - B_{t_k}\big).$$

Each coefficient H_k uses only information available at the left endpoint t_k. That left-endpoint rule is precisely what keeps the integral non-anticipating.

Remark 6.56: Measure-Theory and Functional-Analysis Lens. Two ideas combine here. The condition $H_k \in \mathcal{F}_{t_k}$ is the measure-theory side: the integrand sees only information already in hand. The passage from step processes to general integrands is the functional-analysis side: define the integral on a dense class, prove a norm estimate, complete the space.

Theorem 6.57: Ito Isometry.

For square-integrable adapted processes,

$$\mathbb{E}\left[\left(\int_0^T H_t \, dB_t\right)^2\right] = \mathbb{E}\left[\int_0^T H_t^2 \, dt\right].$$

This identity is the estimate on which the whole L^2 completion of the Ito integral rests.

Remark 6.58: Functional-Analysis Lens. The isometry says the Ito integral is controlled by an L^2 norm:

$$\left\| \int_0^T H_t \, dB_t \right\|_{L^2(\Omega)}^2 = \mathbb{E} \int_0^T H_t^2 \, dt.$$

The stochastic integral is, at first, an isometry from a space of integrands into $L^2(\Omega)$. That is what makes the construction a Hilbert-space construction rather than a pathwise Riemann integral.

Remark 6.59: The Construction in Words. Spelled out, the construction runs:

adapted step processes
$\downarrow$
define sums against Brownian increments
$\downarrow$
prove the Ito isometry
$\downarrow$
complete the space in L^2
$\downarrow$
obtain $\int_0^T H_t \, dB_t$ for general H.

Hilbert-space thinking sits quietly inside stochastic calculus: the integral is assembled from an L^2 norm and an isometry.

6.9 Ito Formula and Stochastic Differential Equations

Big Idea. Ito's formula is the chain rule for stochastic calculus. It carries an extra second-derivative term, and the reason is the nonzero quadratic variation of Brownian motion.

Theorem 6.60: Ito Formula, One-Dimensional Form.

Suppose

$$dX_t = b(t, X_t)\,dt + \sigma(t, X_t)\,dB_t.$$

If $f(t, x)$ is sufficiently smooth, then

$$\begin{aligned} df(t, X_t) &= f_t(t, X_t)\,dt + f_x(t, X_t)\,dX_t + \frac{1}{2} f_{xx}(t, X_t)\,\sigma(t, X_t)^2\,dt \\ &= \left(f_t + b f_x + \frac{1}{2}\sigma^2 f_{xx} \right)(t, X_t)\,dt + \sigma(t, X_t) f_x(t, X_t)\,dB_t. \end{aligned}$$

What b and σ Are. The coefficients $b(t, x)$ and $\sigma(t, x)$ are deterministic functions of the same two variables as f: a time t and a state x. Writing $b(t, X_t)$ and $\sigma(t, X_t)$ just drops the current value of the process into the space slot, exactly as for $f(t, X_t)$ below, so each becomes a process — the coefficient read off along the path.

Their roles, though, are different. Reading the equation one increment at a time,

$$\underbrace{X_{t+dt} - X_t}_{\text{the change}} \approx \underbrace{b(t, X_t)\,dt}_{\text{predictable drift}} + \underbrace{\sigma(t, X_t)\,(B_{t+dt} - B_t)}_{\text{random kick}}.$$

So b is the *drift*: the average velocity of the process, the part that would remain if the noise were switched off, leaving the ODE $\dot{x} = b(t, x)$. And σ is the *diffusion*, or volatility: it scales the mean-zero Brownian increment dB_t, fixing how violently the path jitters when it sits at (t, X_t). Large σ means a wild path; $\sigma \equiv 0$ collapses the SDE back to a deterministic ODE.

Letting them depend on the state is the whole point — it is what makes the dynamics more than a straight line plus noise:

Ornstein–Uhlenbeck:	$b(x) = \kappa(\theta - x)$,	$\sigma(x) = \sigma$ (constant)
GBM:	$b(x) = \mu x$,	$\sigma(x) = \sigma x$
CIR:	$b(v) = \kappa(\theta - v)$,	$\sigma(v) = \xi\sqrt{v}$

The OU drift pulls the state back toward θ; the CIR diffusion shrinks as $v \to 0$, which is what keeps that process nonnegative. These are the same two coefficients the generator collects, $\mathcal{L}f = bf' + \frac{1}{2}\sigma^2 f''$ (item 6.87): b rides the first derivative, σ^2 the second.

What $f(t, X_t)$ Means. Here $f(t, x)$ is an ordinary deterministic function of two *independent* real variables — a time t and a space variable x — with no randomness in it at all. Its partial derivatives f_t, f_x, f_{xx} are computed by plain calculus, holding the other variable fixed: if $f(t, x) = e^{-t}x^2$, then $f_t = -e^{-t}x^2$, $f_x = 2e^{-t}x$, and $f_{xx} = 2e^{-t}$, whatever x later turns out to be.

The randomness appears only when the process is dropped into the space slot. Writing $f(t, X_t)$ means $f(t, x)$ evaluated at $x = X_t(\omega)$ — the height of the fixed surface f read off along

the random curve $t \mapsto (t, X_t)$ — and the outcome is a new process. So yes, $f(t, X_t)$ is $f(t, x(t))$ with $x(t) = X_t$, and $f_t(t, X_t)$, $f_x(t, X_t)$, $f_{xx}(t, X_t)$ are those same deterministic derivatives, simply sampled at the running point (t, X_t).

One caution: f_t is the partial in the first slot, with x held fixed. It is *not* the rate of change of $f(t, X_t)$ along the path. Assembling the partials into the true differential $df(t, X_t)$ — with the quadratic-variation correction — is exactly what the theorem does.

The first slot exists for explicit time dependence. An option price $C(t, s)$ depends on calendar time as well as the stock level, which is why a C_t term shows up. When f has no explicit t, written simply $f(x)$, the f_t term drops and Ito's formula shortens to

$$df(X_t) = f'(X_t)\, dX_t + \tfrac{1}{2} f''(X_t)\, \sigma(t, X_t)^2\, dt.$$

With $f(x) = x^2$ and $X_t = B_t$ this is $d(B_t^2) = 2B_t\, dB_t + dt$, the identity in the pitfall remark below.

Remark 6.61: Analysis Lens. Read Ito's formula as Taylor expansion plus quadratic variation. The first-derivative terms behave just as in ordinary calculus; the second-derivative term survives because

$$(dB_t)^2 = dt.$$

The theorem is a chain rule fitted to paths whose first-order variation is wild but whose quadratic variation is under control.

Remark 6.62: Where the Extra Term Comes From. Ordinary calculus treats $(dt)^2$ as negligible. Ito calculus runs on the mnemonic rules

$$(dt)^2 = 0, \qquad dt\, dB_t = 0, \qquad (dB_t)^2 = dt.$$

The last one is just Brownian quadratic variation written informally.

Remark 6.63: Notation Pitfall: $(dB_t)^2$ Versus $d(B_t^2)$. The Ito mnemonic

$$(dB_t)^2 = dt$$

does *not* say that $d(B_t^2) = dt$. Feeding

$$f(x) = x^2$$

into Ito's formula gives

$$d(B_t^2) = 2B_t\, dB_t + dt.$$

The square of the increment does contribute a dt term, but the differential of the squared Brownian motion keeps both a random part and a drift correction.

6.9.1 Using Ito's Formula: Worked Examples

The Recipe. Applying Ito's formula (Theorem 6.60) is mechanical once the pieces are named:

(1) **Read off** b **and** σ from the SDE $dX_t = b\,dt + \sigma\,dB_t$.

(2) **Differentiate** f by ordinary calculus — compute f_t, f_x, and f_{xx}.

(3) **Substitute** into

$$df(t, X_t) = \big(f_t + b f_x + \tfrac{1}{2}\sigma^2 f_{xx}\big)dt + \sigma f_x\, dB_t.$$

The only thing beyond ordinary calculus is the $\frac{1}{2}\sigma^2 f_{xx}$ term. When the process is Brownian motion itself, $X_t = B_t$ (so $b = 0$, $\sigma = 1$), the formula shortens to

$$df(t, B_t) = \big(f_t + \tfrac{1}{2} f_{xx}\big)dt + f_x\, dB_t,$$

which already covers the next two examples.

Example 6.64: Ito Applied to $f(x) = x^2$. The plainest case: $X_t = B_t$ and $f(x) = x^2$, so $b = 0$, $\sigma = 1$, and (no time dependence) $f_t = 0$, $f'(x) = 2x$, $f''(x) = 2$.

$$\begin{aligned} d(B_t^2) &= \big(f_t + \tfrac{1}{2} f''\big)dt + f'\, dB_t && \text{([the } X_t = B_t \text{ recipe] 1)} \\ &= \big(0 + \tfrac{1}{2}\cdot 2\big)dt + 2B_t\, dB_t && \text{([} f_t = 0,\ f'' = 2,\ f' = 2B_t\text{] 2)} \\ &= 2B_t\, dB_t + dt && \text{([simplify (2)] 3)} \end{aligned}$$

The lone dt is the Ito correction; ordinary calculus would have stopped at $d(x^2) = 2x\,dx$. This is the identity flagged in the pitfall remark above. Integrating from 0 to t and using $B_0 = 0$ rearranges it into a famous formula,

$$\int_0^t B_s\, dB_s = \tfrac{1}{2}B_t^2 - \tfrac{1}{2}t,$$

where the $-\frac{1}{2}t$ is exactly what the ordinary $\int x\,dx = \frac{1}{2}x^2$ is missing. It is also why $\int_0^t B_s\, dB_s$ has mean zero — it is a martingale — even though B_t^2 drifts upward.

■

Example 6.65: A Time-Dependent f: the Exponential Martingale. Now let f carry the time variable. Keep $X_t = B_t$ and set

$$f(t, x) = \exp\Big(\theta x - \tfrac{1}{2}\theta^2 t\Big), \qquad Z_t := f(t, B_t),$$

for a constant θ. Every partial is a multiple of f itself: $f_t = -\frac{1}{2}\theta^2 f$, $f_x = \theta f$, and $f_{xx} = \theta^2 f$.

$$\begin{aligned} dZ_t &= \big(f_t + \tfrac{1}{2} f_{xx}\big)dt + f_x\, dB_t && \text{([the } X_t = B_t \text{ recipe] 1)} \\ &= \Big(-\tfrac{1}{2}\theta^2 f + \tfrac{1}{2}\theta^2 f\Big)dt + \theta f\, dB_t && \text{([} f_t = -\tfrac{1}{2}\theta^2 f,\ f_{xx} = \theta^2 f,\ f_x = \theta f\text{] 2)} \\ &= \theta Z_t\, dB_t && \text{([the two } dt \text{ terms cancel (2)] 3)} \end{aligned}$$

The drift has vanished: $dZ_t = \theta Z_t\, dB_t$ has no dt part, so Z_t is a martingale. The $-\frac{1}{2}\theta^2 t$ planted in the exponent was engineered for exactly this — to produce an f_t that cancels the

Ito correction $\frac{1}{2} f_{xx}$. This Z_t is the exponential martingale behind the change of measure in Girsanov's theorem (Theorem 6.95).

■

6.9.2 SDEs and Their Solutions

Definition 6.66: Stochastic Differential Equation.

A stochastic differential equation, or SDE, has the form

$$dX_t = b(t, X_t)\,dt + \sigma(t, X_t)\,dB_t.$$

Equivalently, it means the integral equation

$$X_t = X_0 + \int_0^t b(s, X_s)\,ds + \int_0^t \sigma(s, X_s)\,dB_s.$$

The first integral is an ordinary time integral; the second is an Ito integral.

What the Solution of an SDE Is. The solution of an SDE is the entire process

$$(X_t)_{t\geq 0},$$

not a single random variable. Recall that a process is a map

$$X : [0, T] \times \Omega \to \mathbb{R}, \qquad (t, \omega) \mapsto X(t, \omega),$$

usually abbreviated

$$X_t(\omega) = X(t, \omega).$$

Each fixed time t gives a random variable

$$X_t : \Omega \to \mathbb{R},$$

and the whole family, as t ranges, is the process. Each fixed outcome ω_0 gives

$$t \mapsto X_t(\omega_0),$$

one sample path of the solution. Probabilists routinely suppress the ω, writing X_t for $X_t(\omega)$.

The condition

$$X_0 = x_0$$

is an initial condition. Basic SDEs are initial-value problems, not boundary-value problems. Boundary behavior does turn up in special models – absorbing barriers, reflecting Brownian motion, or a state constraint like $V_t \geq 0$ in the CIR process.

Remark 6.67: Ordinary or Partial, Initial or Boundary. Two pairs of labels are worth fixing here.

Ordinary versus partial. The equation above is a *stochastic ordinary* differential equation — a SODE, though nearly everyone just writes SDE — because the unknown X_t is a process in a single variable, time, driven by a finite-dimensional Brownian motion. Strip out the noise and it is an ODE. When the unknown is instead a field $u(t, x)$ in time *and* space, and noise is added to a PDE, the object is a *stochastic partial* differential equation, an SPDE. The

stochastic heat equation

$$\partial_t u = \Delta u + \dot{W},$$

driven by space–time white noise $\dot{W}$, is the standard example; its solution is a random field, not a single path. This chapter stays with SODEs.

Initial-value versus boundary-value. An SDE carries an initial condition $X_0 = x_0$ and is run forward in time, so it is an *initial-value problem* (IVP), just like an ODE. An elliptic equation such as $-\Delta u = f$ on a region Ω instead fixes data on the spatial boundary $\partial\Omega$ — a *boundary-value problem* (BVP) — and is taken up in section 7. The slogan: time evolves from a start, space is pinned at its edges.

Remark 6.68: Stock Path Versus Option Price. In the Black–Scholes model S_t is the underlying stock price process, not the option price. A fixed sample path

$$t \mapsto S_t(\omega_0)$$

is one continuous-time stock-price trajectory, and observed market data are a discrete sampling of one such realized path.

The option price is another object entirely, usually written

$$C_t = C(t, S_t) \qquad \text{or} \qquad C(t, s),$$

where $C(t, s)$ is the pricing function evaluated at time t and stock level s.

Remark 6.69: Analysis Lens. An SDE is best read through its integral form. The differential notation just compresses two maps:

$$X \mapsto \int_0^t b(s, X_s)\, ds, \qquad X \mapsto \int_0^t \sigma(s, X_s)\, dB_s.$$

Existence and uniqueness then become fixed-point questions in spaces of adapted processes.

Remark 6.70: Named SDEs as Specific Processes. OU, GBM, and CIR are themselves stochastic processes – each one defined by a particular SDE, once an initial value and a driving Brownian motion have been chosen:

Ornstein–Uhlenbeck:	$dX_t = \kappa(\theta - X_t)\, dt + \sigma\, dB_t,$	mean reversion
GBM under $\mathbb{P}$:	$dS_t = \mu S_t\, dt + \sigma S_t\, dB_t,$	physical stock dynamics
Black–Scholes under $\mathbb{Q}$:	$dS_t = r S_t\, dt + \sigma S_t\, dW_t^{\mathbb{Q}},$	risk-neutral stock dynamics
CIR:	$dV_t = \kappa(\theta - V_t)\, dt + \xi\sqrt{V_t}\, dB_t,$	nonnegative mean reversion

Every one of these defines a whole time-indexed family of random variables, not a lone random variable.

Example 6.71: Solving Geometric Brownian Motion. Geometric Brownian motion is

the stock model under the physical measure $\mathbb{P}$,

$$dS_t = \mu S_t\,dt + \sigma S_t\,dB_t, \qquad S_0 > 0,$$

the multiplicative analogue of exponential growth: both the drift and the noise scale with the current level S_t.

Deterministic warm-up. Switch the noise off ($\sigma = 0$) and the equation collapses to the ODE $\dot{S} = \mu S$. Separating variables gives $d(\log S) = \mu\,dt$, hence $S_t = S_0 e^{\mu t}$ — pure exponential growth. The multiplicative structure is begging for the substitution $\log S$; the only question is what the noise adds to it.

Step 1 — choose a change of variable. This is not an assumption about S_t but a free choice. Ito's formula (Theorem 6.60) returns the differential of $f(S_t)$ for *any* smooth f we care to name, exactly as a u-substitution lets us view an integral through any convenient change of variable; we are simply free to watch the process $f(S_t)$ for an f of our choosing. The art is picking the f that makes the equation solvable. Which one? The deterministic case was cracked by taking logs, and log is precisely what turns multiplicative growth into an additive quantity we can integrate. So follow the process $\log S_t$: take $f(x) = \log x$, with $f'(x) = 1/x$ and $f''(x) = -1/x^2$. Matching the SDE against $dX_t = b\,dt + \sigma\,dB_t$ reads off the coefficients $b(t,x) = \mu x$ and $\sigma(t,x) = \sigma x$, so $\sigma(t,S_t)^2 = \sigma^2 S_t^2$.

Step 2 — apply Ito's formula. Since f has no explicit time dependence, the $f(x)$ form of Ito's formula (Theorem 6.60) applies:

$$\begin{aligned}
d(\log S_t) &= f'(S_t)\,dS_t + \tfrac{1}{2} f''(S_t)\,\sigma^2 S_t^2\,dt && \text{([Ito's formula, no } f_t \text{ term] 1)}\\
&= \frac{1}{S_t}\,dS_t - \frac{1}{2}\,\frac{1}{S_t^2}\,\sigma^2 S_t^2\,dt && \text{([}f' = 1/x,\ f'' = -1/x^2\text{] 2)}\\
&= \frac{1}{S_t}\big(\mu S_t\,dt + \sigma S_t\,dB_t\big) - \tfrac{1}{2}\sigma^2\,dt && \text{([substitute } dS_t\text{; reduce the last term (2)] 3)}\\
&= \Big(\mu - \tfrac{1}{2}\sigma^2\Big)dt + \sigma\,dB_t && \text{([cancel } S_t\text{] 4)}
\end{aligned}$$

The extra $-\frac{1}{2}\sigma^2\,dt$ is the *Ito correction*; the ordinary chain rule would have stopped at $d(\log S_t) = \mu\,dt + \sigma\,dB_t$.

Step 3 — integrate and exponentiate. The right-hand side has constant coefficients, so it integrates termwise:

$$\begin{aligned}
\log S_t - \log S_0 &= \Big(\mu - \tfrac{1}{2}\sigma^2\Big)t + \sigma B_t && \text{([integrate } 0 \to t\text{; } B_0 = 0\text{] 1)}\\
\log \frac{S_t}{S_0} &= \Big(\mu - \tfrac{1}{2}\sigma^2\Big)t + \sigma B_t && \text{([combine the logarithms] 2)}\\
S_t &= S_0 \exp\Big(\big(\mu - \tfrac{1}{2}\sigma^2\big)t + \sigma B_t\Big) && \text{([exponentiate both sides] 3)}
\end{aligned}$$

This is the genuine solution, not a lucky guess: feeding it back through Ito's formula returns the original SDE $dS_t = \mu S_t\,dt + \sigma S_t\,dB_t$, so the choice of log is vindicated after the fact.

Step 4 — read off the distribution. Because $B_t \sim N(0,t)$, the exponent is normal, so S_t is *lognormal*:

$$\log S_t \sim N\Big(\log S_0 + \big(\mu - \tfrac{1}{2}\sigma^2\big)t,\ \sigma^2 t\Big).$$

Its mean recovers the deterministic answer, using $\mathbb{E}\big[e^{\sigma B_t}\big] = e^{\sigma^2 t/2}$ (the normal moment generating function):

$$\begin{aligned}
\mathbb{E}[S_t] &= S_0\, e^{(\mu-\sigma^2/2)t}\, \mathbb{E}\big[e^{\sigma B_t}\big] && \text{([pull out the deterministic factor] 1)}\\
&= S_0\, e^{(\mu-\sigma^2/2)t}\, e^{\sigma^2 t/2} && \text{([}B_t \sim N(0,t)\text{, so } \mathbb{E}[e^{\sigma B_t}] = e^{\sigma^2 t/2}\text{] 2)}\\
&= S_0\, e^{\mu t} && \text{([the two corrections cancel (2)] 3)}
\end{aligned}$$

So the *mean* grows at rate μ — exactly the boxed deterministic solution $S_0 e^{\mu t}$ — while the *median* grows only at the reduced rate $\mu - \frac{1}{2}\sigma^2$. That gap is the volatility drag: the deterministic curve tracks the average, not the typical path.

■

What People Mean by "Black–Scholes". "Black–Scholes" names a whole modeling package, not a single equation. The physical stock model

$$dS_t = \mu S_t\, dt + \sigma S_t\, dB_t$$

is geometric Brownian motion under the real-world measure $\mathbb{P}$. Option pricing usually moves to a risk-neutral measure $\mathbb{Q}$, under which the stock obeys

$$dS_t = rS_t\, dt + \sigma S_t\, dW_t^{\mathbb{Q}}.$$

This risk-neutral SDE is what people call the Black–Scholes stock dynamics. The famous price formula then drops out of combining this process with a payoff, discounting, and risk-neutral expectation.

Example 6.72: Solving the Black–Scholes Risk-Neutral SDE. Under the risk-neutral measure $\mathbb{Q}$, the Black–Scholes stock price satisfies

$$dS_t = rS_t\, dt + \sigma S_t\, dW_t^{\mathbb{Q}}, \qquad S_0 > 0.$$

This is geometric Brownian motion with the drift μ replaced by the risk-free rate r, so the computation of Example 6.71 carries over verbatim (with $B_t \to W_t^{\mathbb{Q}}$):

$$S_t = S_0 \exp\left(\left(r - \tfrac{1}{2}\sigma^2\right) t + \sigma W_t^{\mathbb{Q}}\right).$$

The only change is the drift in the exponent; the Ito correction $-\frac{1}{2}\sigma^2 t$ is the same. The solution is the process $(S_t)_{t\geq 0}$, and a single value like S_1 is just one random variable sitting inside it.

■

Remark 6.73: From Stock Process to Option Price. Solving the stock SDE is not yet pricing the option; it only hands us the distribution of the future price S_T. For a European call with strike K and maturity T, the payoff is

$$(S_T - K)^+.$$

The time-zero Black–Scholes price is

$$C_0 = e^{-rT}\mathbb{E}^{\mathbb{Q}}\big[(S_T - K)^+\big].$$

In the simplest no-dividend version, the five inputs of the Black–Scholes call price are

$$S_0, \qquad K, \qquad T, \qquad r, \qquad \sigma,$$

and they enter at different points:

S_0	initial stock price
r, σ	coefficients in the risk-neutral stock SDE
T	time horizon and option maturity
K	strike in the payoff $(S_T - K)^+$

So K has nothing to do with the stock dynamics; it belongs to the option contract. The SDE says how the underlying moves; the payoff says what cash flow the option draws from that movement.

Remark 6.74: Implied Volatility. The Black–Scholes formula can be used in the forward direction:

$$\sigma \quad \longmapsto \quad C_{\mathrm{BS}}(S_0, K, T, r, \sigma).$$

Implied volatility runs the relationship backward. Given a market price C_{mkt}, it is the σ_{impl} solving

$$C_{\mathrm{BS}}(S_0, K, T, r, \sigma_{\mathrm{impl}}) = C_{\mathrm{mkt}}.$$

Solving the SDE still matters: it produces the lognormal law of S_T and hence the pricing formula. Implied volatility merely inverts that formula to read the market price off as a volatility number.

Remark 6.75: Measure-Theory Lens. The physical stock model begins on a filtered probability space

$$(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P}).$$

Both the Brownian motion and the solution process must be adapted: at time t, the value S_t relies only on information available up to t.
For pricing we view the same measurable world under a risk-neutral measure $\mathbb{Q}$. Here $W^{\mathbb{Q}}$ is Brownian motion under $\mathbb{Q}$, and the equality

$$S_t = S_0 \exp\left(\left(r - \frac{1}{2}\sigma^2\right)t + \sigma W_t^{\mathbb{Q}}\right)$$

holds as an equality of random variables for each fixed t, with the usual convention of identifying versions that agree as processes almost surely.
Under $\mathbb{Q}$, the discounted price $e^{-rt}S_t$ is a martingale – a measure-theoretic statement about conditional expectation.

Remark 6.76: Functional-Analysis Lens. Solving an SDE is, at bottom, a problem set in spaces of processes. A typical choice is a Banach space of adapted continuous processes, say

$$\mathcal{S}_T^2 = \left\{ X \text{ adapted and continuous} : \mathbb{E} \sup_{0\le t\le T} |X_t|^2 < \infty \right\}.$$

The SDE integral form

$$X_t = X_0 + \int_0^t b(s, X_s)\, ds + \int_0^t \sigma(s, X_s)\, dB_s$$

defines an operator on that space. Existence and uniqueness follow by estimating the operator, typically through the Ito isometry and related inequalities, and then closing the argument with a fixed point.
Functional analysis returns in option pricing. A payoff $h(S_T)$ is a random variable, often placed in an L^2 space, and conditional expectation is a projection onto the information available at time t. The Black–Scholes pricing PDE itself comes from the generator

$$\mathcal{L}f(s) = rsf'(s) + \frac{1}{2}\sigma^2 s^2 f''(s),$$

an operator acting on functions of the state s.

6.10 The Probability Theory of SDEs

Big Idea. The previous subsection manipulated SDEs with Ito's formula and solved a few by hand. Now the probabilistic questions: does a solution exist, is it unique, and what is its law? The answers turn an SDE into a genuine probabilistic object – a diffusion – whose distribution is governed by a single differential operator, the generator.

6.10.1 Existence and Uniqueness

Theorem 6.77: Existence and Uniqueness of Strong Solutions.

Consider

$$dX_t = b(t, X_t)\,dt + \sigma(t, X_t)\,dB_t, \qquad X_0 = \xi,$$

on $[0, T]$, with ξ independent of the Brownian motion and $\mathbb{E}\,|\xi|^2 < \infty$. Suppose b and σ are Lipschitz in the space variable and grow at most linearly: there is a constant K with

$$|b(t,x) - b(t,y)| + |\sigma(t,x) - \sigma(t,y)| \le K\,|x - y|\,,$$

$$|b(t,x)| + |\sigma(t,x)| \le K(1 + |x|).$$

Then the SDE has a solution $(X_t)_{0\le t\le T}$ that is continuous, adapted, and square integrable, and any two such solutions are indistinguishable:

$$\mathbb{P}\big(X_t = Y_t \text{ for all } t \in [0,T]\big) = 1.$$

Remark 6.78: Functional-Analysis Lens. This is the Picard–Lindelof theorem for ODEs, transplanted to processes; the deterministic original is spelled out in item 6.80 below. Solving the integral form

$$X_t = \xi + \int_0^t b(s, X_s)\,ds + \int_0^t \sigma(s, X_s)\,dB_s$$

means finding a fixed point of the map that sends a process to the right-hand side. The Lipschitz hypothesis makes that map a contraction on the Banach space $\mathcal{S}_T^2$ of square-integrable adapted continuous processes, with the Ito isometry controlling the stochastic integral; linear growth keeps the iterates from blowing up in finite time. Banach's fixed-point theorem then delivers existence and uniqueness together.

Remark 6.79: When Lipschitz Fails. Lipschitz continuity is sufficient, not necessary. The CIR diffusion coefficient $\sigma(v) = \xi\sqrt{v}$ is only Holder-$\frac{1}{2}$ near the origin, so the theorem does not apply as stated; existence and the constraint $V_t \ge 0$ still hold, but they need square-root specific arguments. This is the same caution the next subsection raises about Heston.

Theorem 6.80: Picard–Lindelof, the Deterministic Original.

Consider the ordinary differential equation as an initial-value problem

$$\dot{x}(t) = F(t, x(t)), \qquad x(0) = x_0,$$

on $[0, T]$. Suppose F is continuous in t and Lipschitz in the space variable, uniformly in t: there is a constant L with

$$|F(t,x) - F(t,y)| \le L\,|x - y| \qquad \text{for all } t \in [0,T] \text{ and all } x, y.$$

Then the problem has exactly one C^1 solution $x : [0, T] \to \mathbb{R}$.

Intuition. This proof is the blueprint that item 6.77 copies. First turn the differential equation into an integral one: a C^1 function solves the initial-value problem exactly when it solves

$$x(t) = x_0 + \int_0^t F(s, x(s))\, ds.$$

The right-hand side defines a map Φ on the space $C([0,T])$ of continuous functions,

$$(\Phi x)(t) = x_0 + \int_0^t F(s, x(s))\, ds,$$

and a solution is precisely a *fixed point*, $x = \Phi x$. With the supremum norm $\|x\|_\infty = \sup_{0\le t\le T} |x(t)|$, the Lipschitz bound gives

$$|(\Phi x)(t) - (\Phi y)(t)| \le \int_0^t L\, |x(s) - y(s)|\, ds \le Lt\, \|x - y\|_\infty .$$

On a short interval — length h with $Lh < 1$ — the constant Lh is below one, so Φ contracts and Banach's fixed-point theorem returns a unique fixed point there. Patching finitely many such intervals reaches all of $[0,T]$; cleaner still, the weighted (Bielecki) norm $\|x\|_\lambda = \sup_t e^{-\lambda t} |x(t)|$ with $\lambda > L$ turns Φ into a single global contraction,

$$\|\Phi x - \Phi y\|_\lambda \le \frac{L}{\lambda}\, \|x - y\|_\lambda , \qquad \frac{L}{\lambda} < 1.$$

The reach is genuinely global because the Lipschitz constant is *global*: it also forces linear growth, $|F(t,x)| \le M_0 + L\,|x|$ with $M_0 = \sup_t |F(t,0)|$, so by Gronwall's inequality the solution cannot blow up in finite time. (Merely *local* Lipschitz would give only a short-time solution — think of $\dot x = x^2$, which escapes to infinity before $t = 1/x_0$.) The iterates $x_{n+1} = \Phi x_n$, started from the constant function x_0, converge uniformly to it.

Example 6.81: Picard Iteration Builds the Exponential. Take $\dot x = ax$ with $x(0) = 1$, so $F(t,x) = ax$. Starting from $x_0(t) \equiv 1$,

$$x_1(t) = 1 + \int_0^t a\, ds = 1 + at,$$

$$x_2(t) = 1 + \int_0^t a\,(1 + as)\, ds = 1 + at + \frac{(at)^2}{2},$$

and in general

$$x_n(t) = \sum_{k=0}^{n} \frac{(at)^k}{k!} \;\longrightarrow\; e^{at} \qquad (n \to \infty).$$

The fixed point is $x(t) = e^{at}$, exactly the solution of $\dot x = ax$. The contraction does not merely promise a solution; the iteration constructs it.

■

Remark 6.82: Why Lipschitz, and Not Just Continuity. Continuity of F alone already

buys a solution — this is Peano's theorem — but not a unique one. The standard failure is

$$\dot{x} = \sqrt{|x|}, \qquad x(0) = 0,$$

whose right-hand side is continuous yet not Lipschitz at the origin, where its slope is infinite. Both

$$x(t) \equiv 0 \qquad \text{and} \qquad x(t) = \frac{t^2}{4}$$

solve it — for the second, $\dot{x} = t/2 = \sqrt{t^2/4}$ — and the path may in fact rest at 0 for any stretch of time before taking off, so there are infinitely many solutions. Lipschitz continuity is precisely what forbids this. It is also the deterministic shadow of the caution above: the square root in $\sqrt{|x|}$ is the same culprit as $\sigma(v) = \xi\sqrt{v}$ in the CIR model, which is why that diffusion needs an argument of its own.

Remark 6.83: The Dictionary: ODE to SDE. Set item 6.80 and item 6.77 side by side and the stochastic result is the same proof carrying one extra term.

ODE (Picard–Lindelof)	SDE (item 6.77)	what changes
unknown function $x(t)$	unknown process X_t	randomness enters
$\dot{x} = F(t,x)$	$dX_t = b\,dt + \sigma\,dB_t$	a noise term is added
$x_0 + \int_0^t F\,ds$	$\xi + \int_0^t b\,ds + \int_0^t \sigma\,dB_s$	an Ito integral appears
$C([0,T])$, sup norm	$\mathcal{S}_T^2$, $\mathbb{E}\sup\lvert X\rvert^2$	L^2 over paths
Φ contracts via L	Φ contracts via the Ito isometry	isometry replaces the $\int ds$ bound
Banach fixed point	Banach fixed point	unchanged
needs F Lipschitz	needs b, σ Lipschitz	same hypothesis
non-unique for $\sqrt{\lvert x\rvert}$	delicate for $\sqrt{v}$ in CIR	same square-root gap

The single genuinely new ingredient is the Ito isometry, which sizes the stochastic integral $\int \sigma\,dB$ the way the elementary bound $\left|\int F\,ds\right| \le \int |F|\,ds$ sizes the deterministic one.

6.10.2 Strong Versus Weak Solutions

Definition 6.84: Strong and Weak Solutions.

Fix a filtered probability space carrying a Brownian motion B. A *strong* solution is a process X, adapted to the filtration generated by B and the initial condition, that satisfies the integral equation: the driving Brownian motion is handed over in advance and X is built from it. A *weak* solution loosens this. It is a triple — a probability space, a Brownian motion $\widetilde{B}$, and a process $\widetilde{X}$ on it — that jointly satisfies the equation, with the driving noise produced alongside the solution rather than fixed beforehand.

Remark 6.85: Two Notions of Uniqueness. Each notion of solution comes with its own uniqueness. *Pathwise uniqueness* says two solutions driven by the same Brownian motion coincide almost surely. *Uniqueness in law* says two weak solutions share the same distribution as processes, even on different spaces. Pathwise uniqueness is the stronger of the two, and the Yamada–Watanabe theorem joins the threads: weak existence together with pathwise uniqueness forces a unique strong solution.

Remark 6.86: Why Weak Solutions Earn Their Keep. For pricing and most distributional questions only the law of the solution matters, which is exactly what a weak solution pins down. Weak solutions also survive under far milder hypotheses — continuity of the coefficients can already be enough — so they reach models where no solution adapted to a prescribed Brownian motion exists.

6.10.3 The Generator and the Markov Property

Autonomous Coefficients. For the rest of this subsection take the coefficients time-homogeneous,

$$dX_t = b(X_t)\,dt + \sigma(X_t)\,dB_t,$$

which is where the clean diffusion theory lives. Write $\mathbb{E}^x$ for expectation under the law of the solution started at $X_0 = x$.

Definition 6.87: Infinitesimal Generator.

The *generator* of the diffusion is the second-order operator

$$\mathcal{L}f(x) = b(x)f'(x) + \tfrac{1}{2}\sigma(x)^2 f''(x),$$

defined on smooth functions f. It is exactly the drift part of Ito's formula applied to $f(X_t)$.

Remark 6.88: Probability Lens. The solution of an autonomous SDE is a *strong Markov process*: given the present state, the future is independent of the past, and this holds even at suitable random times. Its law is encoded in the transition semigroup

$$P_t f(x) = \mathbb{E}^x[f(X_t)],$$

and the generator is the time derivative of that semigroup at the origin,

$$\mathcal{L}f = \lim_{t \downarrow 0} \frac{P_t f - f}{t}.$$

One operator on functions of the state thus summarizes the entire diffusion.

Theorem 6.89: Dynkin's Formula.

For smooth f with suitable growth,

$$\mathbb{E}^x[f(X_t)] = f(x) + \mathbb{E}^x \int_0^t \mathcal{L}f(X_s)\,ds.$$

Equivalently, the process

$$M_t^f = f(X_t) - \int_0^t \mathcal{L}f(X_s)\,ds$$

is a martingale.

Remark 6.90: Where Dynkin Comes From. Apply Ito's formula to $f(X_t)$:

$$f(X_t) = f(x) + \int_0^t \mathcal{L}f(X_s)\,ds + \int_0^t \sigma(X_s)f'(X_s)\,dB_s.$$

The dt part is precisely $\mathcal{L}f$, and the dB part is an Ito integral, hence a mean-zero martingale. Taking expectations erases the stochastic integral and leaves Dynkin's formula. The martingale M^f is just $f(X_t)$ with its predictable drift removed — the continuous-time echo of item 6.31.

6.10.4 From SDE to PDE: Kolmogorov and Feynman–Kac

Theorem 6.91: Kolmogorov Backward Equation.

Let $u(t,x) = \mathbb{E}^x[g(X_t)]$ for a bounded smooth g. Then u solves

$$\partial_t u = \mathcal{L}u, \qquad u(0,x) = g(x),$$

where $\mathcal{L}$ acts on the state variable x.

Theorem 6.92: Feynman–Kac Formula.

Let $r(x) \geq 0$ be a discount rate and g a payoff. The function

$$u(t,x) = \mathbb{E}^x\left[e^{-\int_0^t r(X_s)\,ds}\, g(X_t)\right]$$

solves

$$\partial_t u = \mathcal{L}u - r\,u, \qquad u(0,x) = g(x).$$

Remark 6.93: What Feynman–Kac Buys. The formula is a two-way bridge. Left to right, it writes the solution of a linear parabolic PDE as an average over diffusion paths, which is what Monte Carlo PDE solvers exploit. Right to left, it converts an expectation over random paths into a deterministic PDE for finite differences or finite elements. Probability and analysis are describing the same object from opposite sides.

Example 6.94: Black–Scholes as Feynman–Kac. Under the risk-neutral measure the stock follows $dS_t = rS_t\,dt + \sigma S_t\,dW_t^{\mathbb{Q}}$, with generator

$$\mathcal{L}f(s) = rsf'(s) + \tfrac{1}{2}\sigma^2 s^2 f''(s),$$

the same operator that closed the previous subsection. The price of a claim paying $h(S_T)$ at maturity is the discounted risk-neutral expectation

$$C(t,s) = \mathbb{E}^{\mathbb{Q}}\left[e^{-r(T-t)}h(S_T) \mid S_t = s\right],$$

and Feynman–Kac makes C solve the Black–Scholes PDE

$$\partial_t C + \mathcal{L}C - rC = 0, \qquad C(T,s) = h(s).$$

This is the initial-value equation of item 6.92 run backward through the substitution $\tau = T - t$. The model that earlier produced a closed-form price now reappears as a PDE, and the two routes agree.

■

6.10.5 Changing the Drift: Girsanov

Theorem 6.95: Girsanov, Drift Form.

Let θ_t be adapted with $\mathbb{E}\exp\!\left(\frac{1}{2}\int_0^T \theta_s^2\,ds\right) < \infty$ (Novikov's condition), and set the exponential martingale

$$Z_t = \exp\!\left(-\int_0^t \theta_s\,dB_s - \tfrac{1}{2}\int_0^t \theta_s^2\,ds\right).$$

Under the measure $\mathbb{Q}$ defined by $d\mathbb{Q} = Z_T\,d\mathbb{P}$, the process

$$\widetilde{B}_t = B_t + \int_0^t \theta_s\,ds$$

is a Brownian motion. Changing measure therefore shifts the drift of an SDE while leaving its diffusion coefficient untouched.

Remark 6.96: Measure-Theory Lens. Girsanov is a Radon–Nikodym statement: Z_T is the density that reweights outcomes so that a drifting Brownian motion becomes driftless. This is the machinery behind the move from the physical measure $\mathbb{P}$ to the risk-neutral measure $\mathbb{Q}$ in the previous subsection. Choosing θ to absorb the excess return $\mu - r$ is exactly what turns the discounted stock price into a martingale under $\mathbb{Q}$.

Takeaway. An SDE is more than a formula to differentiate. Lipschitz and growth conditions give it a unique strong solution; that solution is a Markov diffusion captured by its generator $\mathcal{L}$. From there Dynkin and Feynman–Kac trade path expectations for PDEs, and Girsanov shows the drift is really a question of which measure we compute under. These are the probabilistic facts the mechanical Ito calculus quietly rests on.

6.11 Stochastic Volatility: Heston and Rough Heston

Big Idea. Geometric Brownian motion holds volatility constant. Real markets do not oblige: volatility clusters, drifts over time, and tends to move against the asset price. Stochastic volatility models respond by making the volatility itself random.

Prerequisites for Heston-Type Models. The Heston model only makes sense once these objects are kept straight:

- A Brownian motion models continuous random noise.
- An SDE describes a quantity whose infinitesimal change has a drift part and a random noise part.
- A filtration records the information available up to time t.
- An adapted process cannot use future information.
- Quadratic variation explains why Ito's formula has a second-order correction.
- In option pricing, the risk-neutral measure is the probability measure under which the discounted asset price is a martingale.
- Volatility is the standard deviation scale, while variance is its square.

Definition 6.97: Correlated Brownian Motions.

Two Brownian motions W^S and W^V have correlation $\rho \in [-1, 1]$ if their quadratic covariation satisfies

$$d\langle W^S, W^V \rangle_t = \rho \, dt.$$

Informally, their small random shocks satisfy

$$dW_t^S \, dW_t^V = \rho \, dt.$$

Remark 6.98: Analysis Lens. Quadratic covariation is how correlation of Brownian motions is recorded. In finite-dimensional linear algebra correlation is an inner-product notion; here the same notion reappears infinitesimally through $d\langle W^S, W^V \rangle_t$.

Remark 6.99: Why Correlation Matters. In equity markets price and volatility tend to move in opposite directions: a falling stock price usually comes with rising volatility. The parameter ρ captures this leverage effect, and a typical equity model takes

$$\rho < 0.$$

Definition 6.100: The Heston Model.

Under a risk-neutral probability measure, the Heston model is

$$\begin{aligned} dS_t &= rS_t\,dt + \sqrt{V_t}\,S_t\,dW_t^S, \\ dV_t &= \kappa(\theta - V_t)\,dt + \xi\sqrt{V_t}\,dW_t^V, \\ d\langle W^S, W^V\rangle_t &= \rho\,dt. \end{aligned}$$

Here S_t is the asset price and V_t is the instantaneous variance. The parameters have the following meanings:

r	risk-free interest rate
κ	speed of mean reversion of variance
θ	long-run variance level
ξ	volatility of volatility
ρ	correlation between price noise and variance noise

Remark 6.101: Structural Lens. Heston is a coupled system of SDEs. Its state variable is

$$(S_t, V_t),$$

so the model lives in a two-dimensional state space. Behind the pricing sits an operator or semigroup acting on payoff functions, which is what ties Heston together with PDE and Fourier methods.

Remark 6.102: How to Read the Heston System. In the first equation the price still behaves like geometric Brownian motion, only with the constant volatility σ swapped for the random $\sqrt{V_t}$:

$$\sigma \quad \rightsquigarrow \quad \sqrt{V_t}.$$

The second equation pulls variance back toward θ. When $V_t > \theta$ we have $\kappa(\theta - V_t) < 0$ and the drift pushes variance down; when $V_t < \theta$ it pushes variance up.

Definition 6.103: CIR Variance Process.

The variance process in the Heston model,

$$dV_t = \kappa(\theta - V_t)\,dt + \xi\sqrt{V_t}\,dW_t^V,$$

is a Cox–Ingersoll–Ross, or CIR, process. The square-root noise term

$$\xi\sqrt{V_t}\,dW_t^V$$

is designed to keep the variance nonnegative.

Remark 6.104: Analysis Lens. The CIR process is a diffusion with the state constraint $V_t \geq 0$. Its coefficient $\sqrt{V_t}$ degenerates at the boundary, so ordinary Lipschitz intuition does not tell the whole story. A useful reminder that SDE theory is really analysis of coefficients and boundary behavior.

Remark 6.105: Feller Condition. The condition

$$2\kappa\theta \geq \xi^2$$

is the Feller condition. When it holds and $V_0 > 0$, the CIR variance stays strictly positive. When it fails, the variance can touch 0, though the square-root structure still stops it from going negative in the usual Heston setup.

Example 6.106: What Heston Adds Beyond Black–Scholes. In Black–Scholes,

$$dS_t = rS_t\,dt + \sigma S_t\,dW_t,$$

the volatility σ is a single constant, so one number drives every option price.
Heston makes that number random:

$$\sigma_t = \sqrt{V_t}.$$

The extra freedom is what lets the model reproduce market features like volatility smiles, clustering, and the leverage effect.

■

Remark 6.107: Why Heston Is Tractable. Heston owes its popularity to keeping strong analytic structure despite the random volatility. It is an *affine* model: the logarithm of the characteristic function of $\log S_t$ can be recovered from Riccati-type ordinary differential equations. That structure is exactly what makes Fourier pricing efficient.

Definition 6.108: Rough Volatility Idea.

A volatility process is called *rough* when, at small time scales, its paths are less regular than Brownian-looking ones. Roughness is usually indexed by a Hurst parameter

$$H \in (0, \tfrac{1}{2}),$$

with smaller H meaning rougher local behavior.

Remark 6.109: Why Rough Volatility Appears. At very small time scales, empirical volatility does not resemble a smooth Markov diffusion: it shows memory and irregular local fluctuation. Rough volatility models try to match this by trading the Markovian variance equation for a Volterra equation, with the past entering through a singular kernel.

Definition 6.110: Volterra Kernel.

For $\alpha \in (\frac{1}{2}, 1)$, the kernel

$$K(t) = \frac{t^{\alpha-1}}{\Gamma(\alpha)}$$

has a weak singularity at $t = 0$. A Volterra integral

$$\int_0^t K(t-s) f(s)\, ds$$

weights each past value $f(s)$ by how close s lies to the present time t.

Remark 6.111: Functional-Analysis Lens. A Volterra kernel is just an integral operator:

$$(K * f)(t) = \int_0^t K(t-s) f(s)\, ds.$$

Rough volatility is therefore more than "extra random volatility." It turns a Markovian SDE into an equation with memory, where the past acts through an operator.

Definition 6.112: Rough Heston Model.

A common rough Heston formulation is

$$\begin{aligned} dS_t &= r S_t\, dt + \sqrt{V_t}\, S_t\, dW_t^S, \\ V_t &= V_0 + \int_0^t K(t-s)\kappa(\theta - V_s)\, ds + \int_0^t K(t-s)\xi\sqrt{V_s}\, dW_s^V, \end{aligned}$$

with

$$K(t) = \frac{t^{\alpha-1}}{\Gamma(\alpha)}, \qquad \alpha = H + \frac{1}{2}, \qquad H \in (0, \tfrac{1}{2}).$$

As before,

$$d\langle W^S, W^V \rangle_t = \rho\, dt.$$

Remark 6.113: Structural Lens. Classical Heston is local in time – the next move depends only on the current state. Rough Heston is history-dependent: a Volterra operator applied to the whole past produces the variance. This is the jump from finite-dimensional Markov state variables to path-dependent, function-space dynamics.

Remark 6.114: Heston Versus Rough Heston. The Heston variance is Markovian:

$$dV_t = \kappa(\theta - V_t)\, dt + \xi\sqrt{V_t}\, dW_t^V,$$

its future fixed by the present value V_t.

Rough Heston is non-Markovian in V_t alone:

$$V_t = V_0 + \int_0^t K(t-s)(\cdots)\,ds + \int_0^t K(t-s)(\cdots)\,dW_s^V.$$

Through the kernel $K(t-s)$, the current variance carries a memory of the entire past.

Takeaway. The conceptual path is:

$$\begin{gathered}\text{Brownian motion} \to \text{Ito integral} \to \text{SDEs} \\ \to \text{geometric Brownian motion} \to \text{random variance } V_t \\ \to \text{Heston} \\ \to \text{Volterra memory kernel} \to \text{rough Heston.}\end{gathered}$$

Heston makes volatility random; rough Heston makes it random, locally rough, and endowed with memory.

Takeaway. The route into stochastic calculus runs:

$$\begin{gathered}\text{measure space} \to \text{probability space} \to \text{random variables} \\ \to \text{stochastic processes } X(t,\omega) \to \text{adapted processes} \to \text{Brownian motion} \\ \to \text{Ito integral} \to \text{Ito formula and SDEs.}\end{gathered}$$

Probability becomes analysis on random paths.

7 PDE Analysis: Elliptic PDE

Foundations to Strengthen.

- Know weak derivatives and Sobolev spaces, especially H^1 and H^1_0.
- Practice integration by parts and Green's identities.
- Translate classical boundary value problems into weak formulations.
- Learn Lax–Milgram as the basic existence theorem.
- Get comfortable with Young's inequality for absorbing product terms in energy estimates; compactness and the maximum principle round out the toolkit.

Example 7.1. The Poisson problem

$$-\Delta u = f \quad \text{in } \Omega, \qquad u = 0 \quad \text{on } \partial\Omega$$

rarely has a classical solution under mild assumptions on f. The usual move is to relax the derivatives to weak ones and look for u inside a Sobolev space instead.

■

Checkpoint. What makes the weak formulation more tractable than the classical one, even though the two are supposed to describe the same solution?

Takeaway. Elliptic PDE analysis trades the differential equation for a question about function spaces, where existence reduces to an estimate and compactness does much of the remaining work.

8 Applications

Foundations to Strengthen.

- In optimization, know convexity, gradients, Hessians, constraints, and KKT conditions.
- In machine learning, separate approximation error, estimation error, regularization, and generalization.
- In RKHS theory, remember: kernels encode inner products and make point evaluation continuous.
- In time series, know stationarity, autocovariance, ARMA models, and spectral ideas.
- In quantum mechanics, focus on Hilbert spaces, self-adjoint operators, spectra, and unitary evolution.

How to Read This Chapter Through Functional Analysis. The fastest way to organize an application is to ask four questions:

1. What is the space? Are the objects vectors, functions, random variables, signals, or wave functions?
2. What is the norm or inner product? What does "small error" mean in this model?
3. What are the maps? Are they operators, functionals, projections, kernels, or semigroups?
4. What theorem is being used? Boundedness, continuity, compactness, projection, spectral theory, or duality?

The examples that follow look unrelated on the surface, but they all run through the same three steps:

$$\text{choose a space} \longrightarrow \text{put structure on it} \longrightarrow \text{study maps on that space.}$$

Application	Functional-analysis object	What it explains
Machine learning	function class with a norm	generalization and regularization
RKHS	Hilbert space of functions	point evaluation as inner product
Time series	L^2 random variables and shifts	prediction, covariance, filters
Quantum mechanics	Hilbert space and operators	states, observables, evolution
PDE models	function spaces and weak derivatives	existence, stability, energy estimates

8.1 Theory of Machine Learning

Big Idea. The mathematics starts once we look past the algorithm and ask what is actually being learned: which function, fit to which data, against which loss, with what brake on complexity. The reason analysis shows up at all is that this is an approximation problem, usually in infinite dimensions. We pick a function out of a large class, fit it to a finite sample, and then ask whether it still does anything sensible off that sample.

Definition 8.1: Learning Problem.

Let (X, Y) be random variables with values in an input space $\mathcal{X}$ and an output space $\mathcal{Y}$. A hypothesis is a function $f : \mathcal{X} \to \mathcal{Y}$. Given a loss function $\ell(f(x), y)$, the population risk is

$$R(f) = \mathbb{E}[\ell(f(X), Y)].$$

From data $(x_i, y_i)_{i=1}^n$, the empirical risk is

$$R_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i).$$

Learning means driving $R(f)$ down while only ever seeing $R_n(f)$.

Remark 8.2: Why Analysis Appears. Fitting the data is the easy part. The trouble is the size of the function class: too large and we fit noise, too small and we cannot represent the pattern at all. Learning theory tries to control the gap between the two risks with estimates of the form

$$R(f) \leq R_n(f) + \text{complexity penalty} + \text{probability error}.$$

Making such a bound precise is what pulls in norms, compactness arguments, concentration inequalities, and Hilbert-space geometry.

Remark 8.3: Functional-Analysis Lens: the Hypothesis Space. Here the object to keep an eye on is the hypothesis space $\mathcal{F}$. Running a learning algorithm amounts to selecting a single point $f \in \mathcal{F}$, and the penalty $\Omega(f)$ records which of those points count as simple. Regularization, read this way, is just a choice of geometry on the space of candidate functions.

Example 8.4: Regularized Empirical Risk. A common learning rule chooses

$$f_n \in \arg\min_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i) + \lambda \Omega(f) \right\},$$

where $\Omega(f)$ measures complexity and $\lambda > 0$ sets the price we pay for a complicated function. For linear regression $\Omega(f)$ is typically a Euclidean norm; in kernel methods it is usually the Hilbert-space norm $\|f\|_{\mathcal{H}}^2$.

■

Takeaway. Learning theory sits at the seam between approximation and probability. We fit

the sample we have, but we have to rein in the function class enough that the fit still means something on the data we never saw.

8.2 Reproducing Kernel Hilbert Spaces

Big Idea. A reproducing kernel Hilbert space is one where evaluating a function is the same as taking an inner product. That small fact is what lets us carry out nonlinear function learning with nothing more exotic than Hilbert-space geometry.

Definition 8.5: Positive Definite Kernel.

A function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive definite kernel if it is symmetric and, for every finite set $x_1, \dots, x_n \in \mathcal{X}$ and scalars $c_1, \dots, c_n \in \mathbb{R}$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j K(x_i, x_j) \geq 0.$$

Equivalently, the Gram matrix

$$(K(x_i, x_j))_{i,j=1}^{n}$$

is positive semidefinite for every finite sample.

Definition 8.6: Reproducing Kernel Hilbert Space.

A Hilbert space $\mathcal{H}$ of functions $f : \mathcal{X} \to \mathbb{R}$ is a reproducing kernel Hilbert space, or RKHS, if for every $x \in \mathcal{X}$ the evaluation map

$$E_x : \mathcal{H} \to \mathbb{R}, \qquad E_x(f) = f(x),$$

is a bounded linear functional. By the Riesz representation theorem, there exists a unique element $K_x \in \mathcal{H}$ such that

$$f(x) = \langle f, K_x \rangle_{\mathcal{H}} \qquad \text{for every } f \in \mathcal{H}.$$

The reproducing kernel is

$$K(x, t) = K_t(x) = \langle K_t, K_x \rangle_{\mathcal{H}}.$$

Remark 8.7: Why "Reproducing"?. The formula

$$f(x) = \langle f, K_x \rangle_{\mathcal{H}}$$

hands back the value of f at x from a single inner product against the kernel section K_x. Evaluating a function has turned into a geometric operation.

Example 8.8: Kernel Ridge Regression. Given data $(x_i, y_i)_{i=1}^n$ and an RKHS $\mathcal{H}$, kernel ridge regression solves

$$\min_{f\in\mathcal{H}} \left\{ \frac{1}{n}\sum_{i=1}^{n}(f(x_i)-y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2 \right\}.$$

By the representer theorem the minimizer is finite-dimensional after all:

$$f_n(x) = \sum_{i=1}^{n} \alpha_i K(x, x_i).$$

Even with $\mathcal{H}$ infinite dimensional, the solution never leaves the span of the kernel sections sitting at the data points.

■

Remark 8.9: The Functional-Analytic Meaning. What makes the kernel method work is that the data only ever query point values $f(x_i)$. In an RKHS each such query is a bounded linear functional, and Riesz rewrites it as an inner product with K_{x_i}. A problem posed over a space of functions collapses into linear algebra with the Gram matrix

$$K_{ij} = K(x_i, x_j),$$

which is exactly where functional analysis touches the practice of machine learning.

Remark 8.10: Structural Translation. Ordinary regression treats the unknown as a formula to be tuned. In an RKHS the unknown is instead a vector $f \in \mathcal{H}$, and the data act on it through linear measurements

$$E_{x_i}(f) = f(x_i).$$

Riesz representation does the heavy lifting: every bounded functional E_{x_i} corresponds to some vector K_{x_i}, and that correspondence is what reduces function fitting to Hilbert-space linear algebra.

Takeaway. An RKHS is just a function space in which evaluation happens to be continuous, and that one analytic property is the whole reason kernels are useful. It turns nonlinear function fitting into Hilbert-space geometry and, in the end, a finite Gram-matrix computation.

8.3 Time Series Analysis

Big Idea. A time series is more than a list of numbers; it is a signal indexed by time, and most of the interesting questions turn out to be about operators acting on it, shifts and projections chief among them. The functional analysis kicks in the moment we agree to treat the random variables themselves as points in a function space.

Definition 8.11: Time Series.

A discrete-time time series is a sequence of random variables

$$(X_t)_{t\in\mathbb{Z}}.$$

It is weakly stationary if

$$\mathbb{E}[X_t] = \mu \qquad \text{and} \qquad \operatorname{Cov}(X_t, X_{t+h}) = \gamma(h)$$

depend only on the lag h, not on the absolute time t.

Remark 8.12: The Hilbert-Space View. Once $X_t \in L^2(\Omega)$, each observation is a vector in the Hilbert space of square-integrable random variables, with inner product

$$\langle X, Y\rangle_{L^2} = \mathbb{E}[XY],$$

or, after centering,

$$\langle X - \mathbb{E}X, Y - \mathbb{E}Y\rangle_{L^2} = \operatorname{Cov}(X, Y).$$

Covariance, in other words, is the inner product once we have subtracted the means. That is the whole reason orthogonality and least-squares prediction feel so at home in time series.

Remark 8.13: Functional-Analysis Lens: Projection and Prediction. The space is usually $L^2(\Omega)$, the random variables with finite second moment, and the maps that matter are the shift operator, the covariance operator, and the projection onto the span of the past. Cast this way, forecasting loses its mystique: more often than not it is an orthogonal projection.

Definition 8.14: Backshift Operator.

The backshift operator B acts on a time series by

$$(BX)_t = X_{t-1}.$$

For example, an autoregressive model of order one can be written

$$X_t = \phi X_{t-1} + \varepsilon_t, \qquad \text{or} \qquad (I - \phi B)X_t = \varepsilon_t.$$

An AR model is then an operator equation, and solving it means inverting a polynomial in the shift operator.

Example 8.15: AR(1) as an Operator Equation. For

$$X_t = \phi X_{t-1} + \varepsilon_t, \qquad |\phi| < 1,$$

we formally invert $I - \phi B$:

$$X_t = (I - \phi B)^{-1}\varepsilon_t = \sum_{k=0}^{\infty} \phi^k \varepsilon_{t-k}.$$

The hypothesis $|\phi| < 1$ is precisely what keeps this infinite series under control. Stability here is an operator-norm condition wearing different clothes.

■

Remark 8.16: Spectral Intuition. A stationary series has a second description on the frequency side. The autocovariance $\gamma(h)$ and the spectral density $f(\omega)$ are a Fourier-transform pair,

$$\gamma(h) = \int_{-\pi}^{\pi} e^{ih\omega} f(\omega)\, d\omega,$$

so the same dependence structure can be read in two coordinate systems at once. That dual reading is what brings Fourier analysis and operator theory into the subject together.

Takeaway. Read as Hilbert-space geometry with a few operators attached, time series analysis loses much of its mystery. Prediction becomes projection, centered covariance becomes an inner product, and ARMA models become equations in the shift operator; the spectral picture is then just Fourier analysis applied to the dependence.

8.4 Quantum Mechanics

Big Idea. Quantum mechanics is written in the language of Hilbert space. States are the vectors, observables the operators, and measurement is a spectral decomposition of one of those operators, with evolution generated by another. Once that dictionary is in place, functional analysis and operator theory carry most of the conceptual weight.

Definition 8.17: State Space.

In the Hilbert-space formulation of quantum mechanics, a pure state is a unit vector

$$\psi \in \mathcal{H}, \qquad \|\psi\|_{\mathcal{H}} = 1.$$

Vectors that differ by a nonzero complex scalar describe the same physical ray, so what we are really working with is Hilbert-space geometry taken up to phase.

Definition 8.18: Observable.

An observable is represented by a self-adjoint operator A on $\mathcal{H}$. If the system is in the state ψ, the expected value of the observable is

$$\langle A\psi, \psi\rangle_{\mathcal{H}}.$$

Self-adjointness is what guarantees these expectation values are real and what licenses the spectral reading of a measurement.

Example 8.19: Position and Momentum. On $\mathcal{H} = L^2(\mathbb{R})$, the position operator is formally

$$(Q\psi)(x) = x\psi(x),$$

and the momentum operator is formally

$$(P\psi)(x) = -i\hbar\,\psi'(x).$$

They satisfy the commutation relation

$$QP - PQ = i\hbar I$$

on suitable domains. Here is one place domains genuinely matter: the central operators of the theory are unbounded, and treating them like ordinary matrices invites trouble.

■

Definition 8.20: Schrodinger Evolution.

The Schrodinger equation is

$$i\hbar \frac{d}{dt}\psi(t) = H\psi(t),$$

where H is the Hamiltonian operator. Formally, the solution is

$$\psi(t) = e^{-itH/\hbar}\psi(0).$$

Time evolution is therefore an operator exponential. As long as H is self-adjoint the evolution stays unitary, which is what keeps the norm of the state fixed.

Remark 8.21: Why Functional Analysis Matters. In finite dimensions quantum mechanics is essentially linear algebra: vectors, matrices, eigenvalues. The infinite-dimensional version needs genuine functional analysis, because the wave functions live in spaces like $L^2(\mathbb{R}^n)$ and the operators in play may be unbounded. At that point domains and adjoints stop being formalities and the spectrum has to be handled with care.

Remark 8.22: Functional-Analysis Lens: States and Spectra. Now the space is the Hilbert space $\mathcal{H}$ of states: a state is a vector, an observable a self-adjoint operator, and time evolution a one-parameter unitary group. The analysis comes down to the same triple of questions as everywhere else,

What is the space? What is the operator? What is its spectrum?

and answering them is most of what lies under the physics.

Takeaway. Quantum mechanics is about as close as analysis comes to having a flagship application. Hilbert spaces, self-adjoint operators, spectral theory, and PDE all meet here and speak one language.

9 Computational Mathematics

Foundations to Strengthen.

- Keep the different sources of error distinct: truncation, discretization, roundoff, and the error already baked into the model.
- Get clear on conditioning versus stability. A problem can be inherently sensitive, or a perfectly good problem can be wrecked by an unstable algorithm.
- Make LU, QR, SVD, eigenvalue methods, and least squares routine.
- Know basic ODE solvers, stiffness, and finite difference / finite element ideas for PDE.
- Always ask whether the algorithm converges to the intended mathematical object.

9.1 Numerical Analysis

Big Idea. Numerical analysis is about approximating mathematical objects while keeping error under control. Getting an answer out of an algorithm is one thing. The real work is knowing that the answer is accurate, that it survives small perturbations, and that it converges to the object we meant to compute in the first place.

Learning Goals.

- Track approximation error, roundoff error, and truncation error.
- Understand convergence rates and asymptotic error estimates.
- Learn why stability is often the difference between useful computation and meaningless output.

Takeaway. Numerical analysis turns proofs into procedures we can run, and then asks the uncomfortable follow-up question: do those procedures deserve to be trusted?

9.2 Numerical Linear Algebra

Big Idea. Numerical linear algebra concerns algorithms for linear systems, eigenvalue problems, matrix factorizations, and least-squares fitting. Most nonlinear and differential-equation algorithms eventually unwind into repeated linear algebra steps, which is why this subject ends up carrying so much of applied computation on its back.

Learning Goals.

- Understand conditioning, matrix norms, and backward error.
- Study Gaussian elimination, QR factorization, SVD, and eigenvalue algorithms.
- Learn why large sparse matrices require iterative methods.

Takeaway. Linear algebra lives comfortably in finite dimensions. Numerical linear algebra is where that finite-dimensional structure runs up against finite precision and finite time.

9.3 Numerical Differential Equations

Big Idea. Here the goal is to approximate the evolution in time, or the spatial structure, that an ODE or PDE describes. We trade a continuous problem for a discrete one, and the whole difficulty is doing that trade without throwing away the qualitative behavior that made the original equation worth solving.

Learning Goals.

- Learn time-stepping methods for ODEs, including Euler and Runge–Kutta methods.
- Understand consistency, stability, and convergence.
- Study finite difference, finite element, and spectral approaches to PDEs.

Takeaway. A discretization is more than a sampling of the equation. If the discrete problem fails to carry over enough of the original structure, its solutions stop meaning anything.

9.4 Numerical Solvers

Big Idea. A numerical solver is a machine for finding unknowns: a root, a fixed point, a minimizer, the solution of a linear system, the solution of a differential equation. Designing one is where abstract estimates turn into practical choices about stopping criteria and how much computation we can afford.

Learning Goals.

- Compare direct and iterative solvers.
- Understand Newton's method, fixed-point iteration, gradient methods, and Krylov subspace methods.
- Learn how tolerances, residuals, and stopping criteria affect the reliability of a computed answer.

Takeaway. A solver is only as good as its error control. A small residual is encouraging, but on its own it means little until we read it against the conditioning of the problem and the model behind it.

9.5 Computational Modeling

Big Idea. Computational modeling uses mathematics to build a simplified world we can simulate. The model is never the world itself; it is a structured approximation whose assumptions need to stay visible rather than buried in the code.

Learning Goals.

- Identify state variables, parameters, equations, and constraints.
- Separate modeling error from numerical error.
- Understand calibration, validation, sensitivity, and uncertainty.
- Connect mathematical structure with scientific, economic, or engineering interpretation.

Takeaway. Computation does not replace theory; it gives theory a laboratory. How much that laboratory is worth comes down to how good the model is.

10 Philosophy of Math

Foundations to Strengthen.

- Hold Platonism, formalism, intuitionism, and structuralism side by side without collapsing any of them into a slogan.
- Take seriously what a definition does. It fixes a structure, and the consequences then become necessary inside it.
- Keep proof and computation apart, and keep both apart from the intuition that explains why a result is true.
- Sit with existence theorems. What have we shown when we prove something exists but cannot exhibit it?
- Tie mathematical explanation back to science, not merely to calculation.

11 Philosophy of Science and Quant Research

Foundations to Strengthen.

- Keep model, mechanism, prediction, explanation, and causation apart.
- Treat uncertainty carefully: sampling error and regime change are not the same failure, and neither is a hidden assumption.
- In quant research, know why backtests overfit and why out-of-sample testing matters.
- Learn stationarity, dependence, transaction costs, and risk controls.
- Ask what evidence would weaken a model, not only what would support it.

Learning Goals. This section connects philosophy of science with quantitative research. A quant model is a structured claim about the world. Pinning down that claim means naming which variables matter, how uncertainty enters, what data could test it, and where it should be expected to break.

Big Idea. Quant research is applied epistemology with mathematics attached. "Can this formula fit the past?" is the easy question. The harder ones are "What would make me believe this model, what would make me distrust it, and what kind of world would have to exist for the strategy to keep working?"

Intuition. A good model simplifies on purpose. It throws away detail so that structure becomes visible, and the detail it throws away can come back as risk. Philosophy of science has a vocabulary for this tension—idealization, falsification, robustness—and quant research runs into the same thing in backtests, noisy data, and markets that change underneath the assumptions.

> **Remark 11.1: Functional Analysis and Time Series.** Functional analysis pays off in quant work because time series sit naturally as objects in a space. A signal can be projected, smoothed, or acted on by an operator. Time series analysis then feels less foreign after a course in functional analysis: data become functions or vectors, filters become operators, and estimation turns into a question about approximation and stability.

Takeaway. Philosophy keeps asking what a model means. Quant work cares about whether it survives contact with data, costs, and constraints. Done well, the subject holds both questions at once.

12 Solutions

How These Solutions Are Organized. Solutions are grouped by book, one subsection each, and labeled with the exercise number the source uses. Each entry states the problem briefly and then gives the argument in a `solution` block. Multi-step computations use the justified-step template: inside an `align*`, end each line with `\why{reason}` for a right-flushed, auto-numbered tag ([reason] N), and cite an earlier step with `\stepref`. Short logical arguments can stay as plain prose. To add a book, start a new `\subsection`.

12.1 Baby Rudin: Principles of Mathematical Analysis

Exercise 1.1. If $r \in \mathbb{Q}$ with $r \neq 0$ and x is irrational, prove that $r + x$ and rx are irrational.

Solution. Suppose $r + x$ were rational. Then $x = (r + x) - r$ would be a difference of two rationals, hence rational, contradicting that x is irrational. So $r + x$ is irrational. Likewise, if rx were rational, then $x = (rx)/r$ would be rational (using $r \neq 0$), again a contradiction. So rx is irrational. □

Exercise (template). For real numbers a, b, prove $2ab \leq a^2 + b^2$, with equality exactly when $a = b$.

Solution. Start from a square, which is never negative, and expand:

$$(a - b)^2 \geq 0 \qquad \text{([a square is nonnegative] 1)}$$
$$a^2 - 2ab + b^2 \geq 0 \qquad \text{([expand the square] 2)}$$
$$a^2 + b^2 \geq 2ab. \qquad \text{([rearrange (2)] 3)}$$

The single inequality is an equality precisely when $(a - b)^2 = 0$, that is, when $a = b$. □

12.2 Petters and Dong: An Introduction to Mathematical Finance with Applications

About the Book. A. O. Petters and X. Dong, *An Introduction to Mathematical Finance with Applications* (Springer, SUMAT). An undergraduate text organized around financial intuition: the time value of money, probability and statistics for markets, portfolio theory, and option pricing. Solutions below keep the book's own exercise numbering.

Exercise 6.1. Consider an oversimplified stock price described by a two-period binomial tree: in each period the price goes up by a factor u with probability p or down by a factor d with probability $1-p$. Identify the corresponding probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Solution. An outcome is the record of the two moves, each *up* (U) or *down* (D), so the sample space is the set of length-two words

$$\Omega = \{U, D\}^2 = \{UU,\ UD,\ DU,\ DD\}.$$

With only four outcomes, take the σ-algebra to be the power set $\mathcal{F} = 2^\Omega$ (all 16 subsets), so every subset is an event. The two periods are independent — up with probability p, down with $1-p$ — so each path's probability is the product of its step probabilities:

$$\mathbb{P}(UU) = p^2, \qquad \mathbb{P}(UD) = \mathbb{P}(DU) = p(1-p), \qquad \mathbb{P}(DD) = (1-p)^2.$$

These are nonnegative and sum to $\big(p + (1-p)\big)^2 = 1$, and the probability of any event is the sum of the path probabilities it contains, so $(\Omega, \mathcal{F}, \mathbb{P})$ is a genuine probability space. (Note the price S_2 takes only the three values $S_0u^2,\ S_0ud,\ S_0d^2$, since UD and DU recombine; the four-point Ω still records both paths.) The full discrete model — filtration, adapted price process, and risk-neutral pricing — is developed in section 6.6.

All σ-algebras on this Ω. On a finite set a σ-algebra is fixed entirely by its *atoms* — its smallest nonempty members — which partition Ω. So the σ-algebras on Ω correspond exactly to the partitions of its four points, and there are $B_4 = 15$ of them (the fourth Bell number). Listed by their atoms (a k-atom σ-algebra holds 2^k sets):

- *One atom* — the trivial σ-algebra $\{\emptyset, \Omega\}$, which is $\mathcal{F}_0$:

$$\{UU, UD, DU, DD\}.$$

- *Two atoms* — seven of them, each giving $\{\emptyset, A, A^c, \Omega\}$. Four split off a single path,

$$\begin{aligned} &\{UU\} \mid \{UD, DU, DD\}, \quad \{UD\} \mid \{UU, DU, DD\}, \\ &\{DU\} \mid \{UU, UD, DD\}, \quad \{DD\} \mid \{UU, UD, DU\}, \end{aligned}$$

  and three split into two pairs — the first being $\mathcal{F}_1$, the σ-algebra of the first move:

$$\{UU, UD\} \mid \{DU, DD\}, \qquad \{UU, DU\} \mid \{UD, DD\}, \qquad \{UU, DD\} \mid \{UD, DU\}.$$

- *Three atoms* — one pair and two singletons, six in all, each holding $2^3 = 8$ sets:

$$\begin{aligned} &\{UU, UD\} \mid \{DU\} \mid \{DD\}, \quad \{UU, DU\} \mid \{UD\} \mid \{DD\}, \\ &\{UU, DD\} \mid \{UD\} \mid \{DU\}, \quad \{UD, DU\} \mid \{UU\} \mid \{DD\}, \\ &\{UD, DD\} \mid \{UU\} \mid \{DU\}, \quad \{DU, DD\} \mid \{UU\} \mid \{UD\}. \end{aligned}$$

- *Four atoms* — the full power set 2^{Ω} (16 sets), which is $\mathcal{F}_2 = \mathcal{F}$:

$$\{UU\} \mid \{UD\} \mid \{DU\} \mid \{DD\}.$$

In total $1 + 7 + 6 + 1 = 15$. The model's filtration uses just three of these, $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2$, each a coarsening of the next — less information carried by fewer, larger atoms. □

Exercise 6.7. Let $\mathcal{B} = \{\mathcal{B}(t)\}$ be standard Brownian motion. What is the probability that $\mathcal{B}(1)$ lies between -1 and 1?

Solution. Standard Brownian motion has $\mathcal{B}(t) \sim N(0, t)$, so at $t = 1$ the value $\mathcal{B}(1) \sim N(0, 1)$ is standard normal. Writing Φ for its CDF,

$$\begin{aligned}
\mathbb{P}\big(-1 < \mathcal{B}(1) < 1\big) &= \Phi(1) - \Phi(-1) && ([\mathcal{B}(1) \sim N(0,1);\ \text{probability of an interval}]\ 1)\\
&= \Phi(1) - \big(1 - \Phi(1)\big) && ([\text{symmetry } \Phi(-x) = 1 - \Phi(x)]\ 2)\\
&= 2\Phi(1) - 1 && ([\text{combine (2)}]\ 3)\\
&\approx 2(0.8413) - 1 = 0.6827 && ([\Phi(1) \approx 0.8413]\ 4)
\end{aligned}$$

So $\mathbb{P}(-1 < \mathcal{B}(1) < 1) \approx 68.3\%$ — the one-standard-deviation rule, since $\mathcal{B}(1)$ has standard deviation 1. □

Exercise 6.8. Let $X = \{X_t\}$ and $Y = \{Y_t\}$ be two processes. Verify the identity

$$\Delta(X_t Y_t) = X_t\,\Delta Y_t + Y_t\,\Delta X_t + \Delta X_t\,\Delta Y_t.$$

Solution. Write $\Delta Z_t = Z_{t+1} - Z_t$ for the one-step increment. Split the product increment by adding and subtracting $X_t Y_{t+1}$:

$$\begin{aligned}
\Delta(X_t Y_t) &= X_{t+1}Y_{t+1} - X_t Y_t && ([\text{definition of the increment } \Delta]\ 1)\\
&= \big(X_{t+1}Y_{t+1} - X_t Y_{t+1}\big) + \big(X_t Y_{t+1} - X_t Y_t\big) && ([\text{add and subtract } X_t Y_{t+1}]\ 2)\\
&= (X_{t+1} - X_t)\,Y_{t+1} + X_t\,(Y_{t+1} - Y_t) && ([\text{factor each group (2)}]\ 3)\\
&= Y_{t+1}\,\Delta X_t + X_t\,\Delta Y_t && ([\Delta X_t = X_{t+1} - X_t,\ \Delta Y_t = Y_{t+1} - Y_t]\ 4)\\
&= (Y_t + \Delta Y_t)\,\Delta X_t + X_t\,\Delta Y_t && ([Y_{t+1} = Y_t + \Delta Y_t\ (4)]\ 5)\\
&= X_t\,\Delta Y_t + Y_t\,\Delta X_t + \Delta X_t\,\Delta Y_t && ([\text{expand and reorder}]\ 6)
\end{aligned}$$

which is the claimed identity. The cross term $\Delta X_t\,\Delta Y_t$ is the discrete shadow of the Ito correction: in the continuous limit it becomes the quadratic-covariation term $d\langle X, Y\rangle_t$, and for $X = Y = \mathcal{B}$ it is exactly what makes $(d\mathcal{B}_t)^2 = dt$. □

> **Remark 12.1: A Problem That Needs This.** The desk holds h_t shares of a stock S_t and re-hedges each day, and at the close must explain exactly where the day's P&L came from — the reported number refuses to match "delta times price change," and the gap has to be accounted for. The position is the product $V = h\,S$, so the only way to attribute ΔV is to expand it by this identity:
>
> $$\Delta V = \underbrace{h\,\Delta S}_{\text{holding gain}} + \underbrace{S\,\Delta h}_{\text{trades placed}} + \underbrace{\Delta h\,\Delta S}_{\text{rebalancing P\&L}}.$$
>
> The missing piece is the cross term $\Delta h\,\Delta S$: the P&L earned by re-hedging *while* the price

moves — the discrete cousin of gamma P&L. Drop it and the books do not reconcile. Faced instead with estimating a covariance or hedge ratio from intraday ticks, the quant computes the same sum $\sum \Delta X\, \Delta Y$. Whenever two moving quantities are multiplied, this is the rule that must be applied.

Exercise 6.13. For the SDE

$$dY = \Big(f_t + \mu f_x + \tfrac{1}{2}\sigma^2 f_{xx}\Big)dt + \sigma f_x\, d\mathcal{B},$$

compute $\mathbb{E}(dY \mid \mathcal{F}_t)$ and $\mathrm{Var}(dY \mid \mathcal{F}_t)$ for the Brownian filtration $\{\mathcal{F}_t\}$, and indicate the properties of conditional expectation used.

Solution. Write the equation as $dY = a\,dt + b\,d\mathcal{B}$ with drift $a = f_t + \mu f_x + \frac{1}{2}\sigma^2 f_{xx}$ and diffusion $b = \sigma f_x$. Both a and b are $\mathcal{F}_t$-measurable — known from the path up to time t. Here is why: each is a deterministic function of t and the *current* value X_t (the equation is the Ito differential of $Y = f(t, X_t)$ for $dX = \mu\,dt + \sigma\,d\mathcal{B}$), and X_t is *adapted* —

$$X_t = X_0 + \int_0^t \mu\, ds + \int_0^t \sigma\, d\mathcal{B}_s$$

is built from the Brownian path only up to t — so X_t is $\mathcal{F}_t$-measurable, and any deterministic function of (t, X_t) is too. The forward increment $d\mathcal{B} = \mathcal{B}_{t+dt} - \mathcal{B}_t$, by contrast, reaches *past* t: it is independent of $\mathcal{F}_t$, with $\mathbb{E}[d\mathcal{B}] = 0$ and $\mathbb{E}[(d\mathcal{B})^2] = dt$. In short, a and b live in the present, $d\mathcal{B}$ in the future.

Conditional mean.

$$\begin{aligned}
\mathbb{E}(dY \mid \mathcal{F}_t) &= \mathbb{E}(a\,dt \mid \mathcal{F}_t) + \mathbb{E}(b\,d\mathcal{B} \mid \mathcal{F}_t) && \text{([linearity of conditional expectation] 1)}\\
&= a\,dt + b\,\mathbb{E}(d\mathcal{B} \mid \mathcal{F}_t) && \text{([take out what is known } (a, b, dt \text{ are } \mathcal{F}_t\text{-measurable)] 2)}\\
&= a\,dt + b\,\mathbb{E}(d\mathcal{B}) && \text{([}d\mathcal{B} \text{ is independent of } \mathcal{F}_t\text{] 3)}\\
&= a\,dt && \text{([}\mathbb{E}(d\mathcal{B}) = 0\text{, so the term from (3) vanishes] 4)}
\end{aligned}$$

that is, $\mathbb{E}(dY \mid \mathcal{F}_t) = \big(f_t + \mu f_x + \frac{1}{2}\sigma^2 f_{xx}\big)dt$.

Conditional variance.

$$\begin{aligned}
\mathrm{Var}(dY \mid \mathcal{F}_t) &= \mathrm{Var}(a\,dt + b\,d\mathcal{B} \mid \mathcal{F}_t) && \text{([}dY = a\,dt + b\,d\mathcal{B}\text{] 1)}\\
&= \mathrm{Var}(b\,d\mathcal{B} \mid \mathcal{F}_t) && \text{([the } \mathcal{F}_t\text{-measurable } a\,dt \text{ adds no spread] 2)}\\
&= b^2\,\mathrm{Var}(d\mathcal{B} \mid \mathcal{F}_t) && \text{([pull out the known factor } b\text{, squared] 3)}\\
&= b^2\,\mathrm{Var}(d\mathcal{B}) && \text{([}d\mathcal{B} \text{ is independent of } \mathcal{F}_t\text{] 4)}\\
&= b^2\,dt && \text{([Var}(d\mathcal{B}) = \mathbb{E}[(d\mathcal{B})^2] = dt\text{] 5)}
\end{aligned}$$

that is, $\mathrm{Var}(dY \mid \mathcal{F}_t) = \sigma^2 f_x^2\, dt$.

Properties used. Linearity of conditional expectation; "taking out what is known," $\mathbb{E}(ZW \mid \mathcal{F}_t) = Z\,\mathbb{E}(W \mid \mathcal{F}_t)$ for $\mathcal{F}_t$-measurable Z; and the role of independence, $\mathbb{E}(d\mathcal{B} \mid \mathcal{F}_t) = \mathbb{E}(d\mathcal{B})$, since the forward increment is independent of the past $\mathcal{F}_t$. For the variance, an $\mathcal{F}_t$-measurable term contributes no conditional spread, a known factor comes out squared, and $\mathbb{E}[(d\mathcal{B})^2] = dt$ is the Brownian quadratic-variation rule $(d\mathcal{B})^2 = dt$. (This dY is just the Ito differential of $Y = f(t, X_t)$, so the answer says the drift is the conditional mean rate and $\sigma^2 f_x^2$ the conditional variance rate — the very meaning of "drift" and "diffusion.") □

Remark 12.2: A Problem That Needs This. Risk asks for a one-day number on an option position $Y = f(t, S)$ before tomorrow's open: the expected P&L and the variance that feeds Value-at-Risk. Producing it *is* this exercise. The conditional mean $\mathbb{E}(dY \mid \mathcal{F}_t)$ is the expected move, the conditional variance $\mathrm{Var}(dY \mid \mathcal{F}_t) = \sigma^2 f_x^2\, dt$ is the spread, and the one-day VaR comes out as a multiple of $\sigma\,|f_x|\sqrt{dt}$. Since f_x is the *delta*, the risk is delta-squared times the underlying's own variance — which is precisely why cancelling the $f_x\, dX$ term by delta-hedging is the first move to shrink it. Run the same computation in reverse and it sets the price: force $\mathbb{E}(dY \mid \mathcal{F}_t)$ to the risk-free rate, leaving only the mean-zero noise, and that is the risk-neutral pricing condition. Without this conditional mean and variance there is no risk number and no price.

Exercise (template) — continuous compounding. A principal earning annual rate r, compounded n times a year, grows by the factor $(1 + r/n)^n$ over one year. Show that as the compounding is made continuous, $n \to \infty$, this factor tends to e^r.

Solution. Set $h = r/n$, so $h \to 0$ as $n \to \infty$ and $n = r/h$. Pass to logarithms and the limit becomes a derivative:

$$\begin{aligned} \log\Big(1 + \tfrac{r}{n}\Big)^n &= n \log\Big(1 + \tfrac{r}{n}\Big) && ([\log a^n = n \log a]\ 1) \\ &= r\,\frac{\log(1+h)}{h} && ([h = r/n, \text{ so } n = r/h]\ 2) \\ &\longrightarrow\ r \cdot 1 = r && ([\tfrac{\log(1+h)}{h} \to (\log)'(1) = 1 \text{ as } h \to 0\ (2)]\ 3) \end{aligned}$$

Exponentiating, $(1 + r/n)^n \to e^r$. So continuous compounding multiplies the principal by e^r each year — the origin of the e^{rT} growth and discount factors used throughout the book. □

Takeaway. The entries above are format templates: a prose solution for a short logical argument, and a justified-step `align*` for a computation. Replace them with your own work, and add one `\subsection` per book as the collection grows.

13 Course Timestamps

13.1 Measure Theory — Claudio Landim, Lecture 1: Constructing a Non-Measurable Set (the Vitali Set)

Source. Full lecture: Measure Theory — the Vitali set and non-measurable sets. This is the classic opening of measure theory: by constructing the Vitali set, the lecture proves *by contradiction* a deep fact — no single function can assign a consistent "length" to every subset of $\mathbb{R}$. The timestamps below trace the argument from first principles, stripped of heavy notation.

13.1.1 The goal, and the ideal we ask for

- **00:14 — From interval length to a measure on all sets.** The length of a closed interval $[a, b]$ is plainly $b - a$. The ambition of the lecture is to promote this single number into a function λ — a *measure* — that hands a length to *every* subset of $\mathbb{R}$, not merely to intervals. What length, for example, should the set of all rationals carry?

- **02:40 — Four properties such a measure must have.** For λ to deserve the name "length," four demands are forced by ordinary intuition:

  (0) *Universality* — λ is defined on *every* subset of $\mathbb{R}$.

  (1) *Compatibility* — on an interval it returns the ordinary length, $\lambda([a, b]) = b - a$.

  (2) *Translation invariance* — sliding a set along the line cannot change its size, $\lambda(A + x) = \lambda(A)$.

  (3) *Countable (σ-) additivity* — the measure of a countable union of pairwise disjoint sets is the sum of their measures.

  The whole lecture is the discovery that these four cannot coexist.

13.1.2 The tools: equivalence classes and the axiom of choice

- **06:50 — Sorting the reals by rational differences.** Declare two reals equivalent, $x \sim y$, exactly when their difference $x - y$ is *rational*. This is an equivalence relation, so it shatters the whole line into classes: within one class any two points differ by a rational, while points from different classes differ by an irrational. There are uncountably many such classes.

- **09:01 — One representative per class: the Vitali set Ω.** Here is the single piece of magic. From *each* class pick exactly one representative, arranged so that every chosen point lands in $[0, 1]$. Gather these representatives into one set $\Omega \subset [0, 1]$. That such a simultaneous choice across uncountably many classes is even permitted is precisely the content of the *axiom of choice* — no explicit rule names the representatives.

13.1.3 The deduction: two nets that cannot both hold

- **11:53 — Distinct rational shifts are disjoint.** Slide Ω by two different rationals $p \neq q$. The copies $\Omega + p$ and $\Omega + q$ then share no point whatsoever. A common point would force two representatives in Ω to differ by a rational — that is, to sit in the same class — contradicting "one representative per class."

- **15:09 — The shifts stay inside a finite interval.** Restrict the shift to rationals $q \in [-1, 1]$. Since $\Omega \subset [0, 1]$, every copy $\Omega + q$ lands inside $[-1, 2]$, and therefore so does the union of all of them.

- **17:57 — Monotonicity: a part is no bigger than the whole.** Additivity forces $A \subset B \Rightarrow \lambda(A) \leq \lambda(B)$. Because the union of the shifts sits inside $[-1, 2]$, its total length can be at most $\lambda([-1, 2]) = 3$.

- **22:08 — First half: $\lambda(\Omega)$ is forced to be 0.** The shifts $\{\Omega + q : q \in \mathbb{Q} \cap [-1, 1]\}$ are countably many and pairwise disjoint, and by translation invariance each carries the *same* measure $\lambda(\Omega)$. By σ-additivity,

$$\sum_{q \in \mathbb{Q} \cap [-1,1]} \lambda(\Omega) \;=\; \lambda\Big(\bigcup\nolimits_q (\Omega + q)\Big) \;\leq\; 3.$$

  A countable sum of one fixed nonnegative number is either 0 (when that number is 0) or $+\infty$ (when it is positive). Capped at 3, the only survivor is $\lambda(\Omega) = 0$; hence the whole union has measure 0 as well.

- **24:36 — Second half: those same shifts cover all of $[0, 1]$.** Take any $x \in [0, 1]$. It belongs to some class, whose representative α was placed inside $\Omega \subset [0, 1]$. Then $q := x - \alpha$ is rational and lies in $[-1, 1]$, so $x = \alpha + q \in \Omega + q$. Every point of $[0, 1]$ is caught this way, giving $[0, 1] \subset \bigcup_q (\Omega + q)$.

- **27:38 — The collision: $1 \leq 0$.** Monotonicity applied to that covering of $[0, 1]$ gives

$$1 \;=\; \lambda([0, 1]) \;\leq\; \lambda\Big(\bigcup\nolimits_q (\Omega + q)\Big) \;=\; 0,$$

  where the final equality is the airtight count from 22:08. The conclusion $1 \leq 0$ is a flat absurdity.

13.1.4 The verdict

- **29:23 — Surrender omniscience, not the calculus.** Every step used only bedrock logic, so the fault must lie in the opening wish. Properties (1)–(3) — intervals measure their length, size is translation invariant, disjoint pieces add up — are the foundations that calculus is built on and cannot be given up. What must yield is Property (0): a measure can *never* be defined consistently on every subset of $\mathbb{R}$. Objects as pathological as Ω are genuinely *non-measurable*. The repair, taken up next, is to fence off a safe collection of "measurable" sets — a σ-algebra — and let Lebesgue measure live only there.

13.2 Measure Theory — Claudio Landim, Lecture 2: From Semi-Algebras to σ-Algebras

Source. Full lecture: Measure Theory — semi-algebras, algebras, and σ-algebras (Claudio Landim). The sequel to the Vitali construction: since no measure can live on *every* subset of $\mathbb{R}$, this lecture builds the staircase of set families — semi-algebra → algebra → σ-algebra — on which a measure can safely be defined, and then asks what kind of *additivity* such a function ought to obey.

13.2.1 Building the foundation: starting from intervals

- **01:43 — What is a semi-algebra?** To measure "length" at all, what is the most basic brick? On the line it is the *interval*. A semi-algebra $\mathcal{S}$ is the family that abstracts exactly the properties of intervals, asking for three things:
  (1) the whole space belongs, $\Omega \in \mathcal{S}$;
  (2) closed under finite intersection, $A, B \in \mathcal{S} \Rightarrow A \cap B \in \mathcal{S}$ — the overlap of two intervals is again an interval, e.g. $(a, b] \cap (c, d]$;
  (3) complements break into finitely many pieces — for $A \in \mathcal{S}$, the complement A^c need not be a single interval, but it is always a *finite disjoint union* of members of $\mathcal{S}$.

- **04:08 — The prototype.** The half-open intervals $(a, b]$, together with the rays $(-\infty, b]$ and (a, ∞) and the empty set, form a semi-algebra. This is the raw material from which Lebesgue measure is later assembled.

13.2.2 Upgrading the structure: the algebra

- **09:24 — What is an algebra?** The semi-algebra's complement rule is still weak: a complement is only a finite *union* of members, not necessarily a member itself. To make set operations as smooth as arithmetic, demand more. An algebra $\mathcal{A}$ requires
  (1) $\Omega \in \mathcal{A}$;
  (2) closure under intersection, $A, B \in \mathcal{A} \Rightarrow A \cap B \in \mathcal{A}$;
  (3) closure under complement *outright*, $A \in \mathcal{A} \Rightarrow A^c \in \mathcal{A}$ — far stronger than the semi-algebra.

  By De Morgan, closure under complement and intersection forces closure under finite *union* as well; so inside an algebra one may freely intersect, union, and complement finitely many sets and never escape it. Every algebra is in particular a semi-algebra.

13.2.3 The ultimate form: the σ-algebra

- **12:08 — What is a σ-algebra?** Calculus and limits teach that the real magic happens at infinity, and finitely many operations cannot reach it. So promote "closed under finite unions" to "closed under *countable* unions." A σ-algebra $\mathcal{F}$ keeps the complement rule but now asks: for any countable sequence $A_1, A_2, \cdots \in \mathcal{F}$, the countable intersection $\bigcap_{j=1}^{\infty} A_j$ (equivalently the countable union) lies again in $\mathcal{F}$. This is the stage on which modern measure theory and probability stand — only here can one safely speak of "almost everywhere" or "the probability of a limit."

13.2.4 Set-theoretic magic: generated algebras and σ-algebras

- 14:31 — **The power of intersection.** However many algebras one has — even uncountably many — their intersection is again an algebra. The same holds one level up: an arbitrary intersection of σ-algebras is again a σ-algebra.

- 21:02 — **Generated by a class.** Given only a scattered handful of sets $\mathcal{C}$, what is the smallest legal house — algebra, or σ-algebra — that contains them?

- 23:23 — **The construction by intersection.** Collect *every* algebra that contains $\mathcal{C}$ (the power set $\mathcal{P}(\Omega)$ is always one, so the family is never empty) and intersect them all. By the previous remark the result is itself an algebra, and by construction the *smallest* one containing $\mathcal{C}$; write it $\mathcal{A}(\mathcal{C})$. The generated σ-algebra $\sigma(\mathcal{C})$ is defined in exactly the same way.

- 27:32 — **The crucial asymmetry: algebra clear, σ-algebra opaque.** Here is a deep difficulty. If $\mathcal{S}$ is a semi-algebra, the members of the generated algebra $\mathcal{A}(\mathcal{S})$ are completely explicit — the proof at 30:01 shows they are exactly the finite disjoint unions of members of $\mathcal{S}$. But the members of the generated σ-algebra admit *no* explicit formula at all: there is no closed form for a general element. This is precisely why later arguments reach the σ-algebra only indirectly, through limit tools such as the monotone class theorem.

13.2.5 Assigning size: set functions and additivity

- 49:46 — **Finite additivity.** With the sets pinned down, a measure μ can be defined. A set function μ sending sets into $[0,\infty]$ is *finitely additive* when $\mu(\emptyset) = 0$ and, whenever E splits into finitely many disjoint pieces E_j,

$$\mu(E) = \sum_j \mu(E_j).$$

  This is the bare statement that the whole equals the sum of its parts.

- 1:00:46 — **Toward calculus: σ-additivity.** Echoing the σ-algebra, if E splits into *countably* many disjoint pieces A_j with $\mu(E) = \sum_{j=1}^{\infty} \mu(A_j)$, then μ is σ-additive. This is the property that lets the Lebesgue integral exchange limits with integration.

- 1:03:30 — **A warning example: finitely additive but not σ-additive.** On the half-open subintervals of $(0,1]$, define

$$\mu\big((a,b]\big) = b - a \quad (a > 0), \qquad \mu\big((0,b]\big) = \infty :$$

  any interval whose left endpoint is 0 is handed the value ∞. This μ is finitely additive.

- 1:06:51 — **The contradiction.** Yet split $(0,\frac{1}{2}]$ into the countable disjoint stack $(\frac{1}{4},\frac{1}{2}] \cup (\frac{1}{8},\frac{1}{4}] \cup \cdots$, that is the intervals $(2^{-j-1}, 2^{-j}]$. None of these touches 0, so

$$\sum_{j\geq 1} \mu\big((2^{-j-1}, 2^{-j}]\big) = \tfrac{1}{4} + \tfrac{1}{8} + \cdots = \tfrac{1}{2}, \qquad \text{yet} \qquad \mu\big((0,\tfrac{1}{2}]\big) = \infty.$$

  Since $\infty \neq \frac{1}{2}$, finite additivity does *not* force countable additivity — which is exactly what makes σ-additivity the demanding, and precious, property at the heart of measure theory.

13.3 Measure Theory — Claudio Landim, Lecture 3: Continuity and the Extension to an Algebra

Source. Full lecture: Measure Theory — set functions: continuity and the extension theorem (Claudio Landim). The structural heart of the subject. First it identifies measure-theoretic *continuity* with σ-additivity; then it uses that equivalence to lift a measure from a semi-algebra up to the algebra it generates — the first rung of the ladder that ends, a lecture later, at the Carathéodory extension onto a full σ-algebra.

13.3.1 Continuity of a measure

- **01:28 — Continuity from below.** In calculus, if $x_n \uparrow x$ then a continuous f obeys $f(x_n) \to f(x)$. Transplant the idea to sets: if sets swell and nest, $E_1 \subset E_2 \subset \cdots$ with union $\bigcup_n E_n = E$, a sensible measure ought to respect the limit,

$$\lim_{n\to\infty} \mu(E_n) = \mu(E).$$

  This is *continuity from below.*

- **03:32 — Continuity from above.** Symmetrically, if sets shrink and nest, $E_1 \supset E_2 \supset \cdots$, and peel down like an onion to a core $\bigcap_n E_n = E$, the measure should again follow the limit, $\lim_{n\to\infty} \mu(E_n) = \mu(E)$.

- **05:17 — A fatal trap in the shrinking case.** Continuity from above needs one extra hypothesis: *some* E_{n_0} must have finite measure. Why? Take $E_n = [n, \infty)$. These retreat rightward to the empty intersection $\bigcap_n E_n = \emptyset$, of measure 0, yet every E_n has infinite length. Without the finiteness clause one would conclude $+\infty \to 0$, an absurdity.

13.3.2 The key lemma: continuity equals σ-additivity

- **07:25 — The statement: continuity $\iff$ σ-additivity.** Calculus rests on limits (an infinite operation); an algebra supports only finite ones. The previous lecture showed finite additivity is strictly weaker than σ-additivity. This lecture supplies the missing bridge: *finite additivity together with continuity is exactly σ-additivity.*

- **10:38 — Proof: σ-additive $\Rightarrow$ continuous from below.** Let $E_n \uparrow E$. The trick (12:01) is to slice the nested tower into disjoint shells: $F_1 = E_1$ and $F_k = E_k \setminus E_{k-1}$. The F_k are pairwise disjoint with $\bigsqcup_k F_k = E$ and $E_n = \bigsqcup_{k\le n} F_k$. Summing the shells with σ-additivity turns the partial measures $\mu(E_n)$ into the full series adding up to $\mu(E)$.

- **25:00 — Proof: continuous from below $\Rightarrow$ σ-additive.** Conversely, assume finite additivity and continuity from below. Given countably many disjoint E_k, set $F_n = \bigsqcup_{k\le n} E_k$; then $F_n \uparrow E = \bigsqcup_k E_k$. Continuity gives $\mu(F_n) \to \mu(E)$, while finite additivity gives $\mu(F_n) = \sum_{k\le n} \mu(E_k)$. Letting $n \to \infty$ turns the finite sum into the infinite series — which is σ-additivity.

- **30:19 — Proof: continuity at $\emptyset$ from above $\Rightarrow$ σ-additive.** The same argument runs with a sequence shrinking to the empty set. This third form is the workhorse in practice: showing "the measure of sets collapsing to $\emptyset$ tends to 0" is usually far easier than summing an infinite series head-on.

13.3.3 Extending a measure from a semi-algebra to an algebra

- **40:34 — The goal of the extension theorem.** We hold a measure μ defined only on a semi-algebra $\mathcal{S}$ (say the intervals) and want to carry it up to the algebra $\mathcal{A} = \mathcal{A}(\mathcal{S})$ it generates: a function ν on $\mathcal{A}$ that (1) agrees with μ on the old ground $\mathcal{S}$, (2) is finitely additive, and (3) is the *unique* such extension.

- **44:52 — Defining the extension ν.** Last lecture's structure theorem says every $A \in \mathcal{A}$ is a finite disjoint union of semi-algebra pieces, $A = \bigsqcup_j E_j$ with $E_j \in \mathcal{S}$. Since the whole should equal the sum of its parts, the definition writes itself (46:11):

$$\nu(A) := \sum_j \mu(E_j).$$

- **48:40 — Checking it is well defined.** The crucial subtlety: the same A might be cut as $\bigsqcup_j E_j$ or as $\bigsqcup_k F_k$. If the two cuts gave different totals the definition would collapse. Refining both partitions by their mutual intersections $E_j \cap F_k$ into one common finer grid, and applying the finite additivity already known on $\mathcal{S}$, shows both totals equal the grid's total — so they agree (52:07).

- **53:07 — Finite additivity and uniqueness of ν.** With well-definedness secured, finite additivity reduces to merging sums (54:13). Uniqueness (57:31) is just as direct: every block is built from the basic bricks of $\mathcal{S}$, so once the bricks' measures are fixed the value on any assembled set is forced.

- **1:01:00 — The payoff: σ-additivity is inherited.** Finite additivity is not the end of the story. The finale shows that if the original μ on $\mathcal{S}$ was already σ-additive, the extended ν on $\mathcal{A}$ *inherits* σ-additivity automatically. The mechanism is once more a limit exchange over the intersection grid (1:09:09), converting an infinite sum on $\mathcal{A}$ back into one on $\mathcal{S}$ — the first clean expansion of the measure-theoretic skeleton.

This lecture is the springboard from elementary geometry to modern measure: by proving that *continuity is additivity* and exploiting set partitions, it widens a measure's reach from bare intervals (a semi-algebra) to a family closed under finite operations (an algebra), loading the ammunition for the next step — onto a full σ-algebra via the Carathéodory extension.

13.4 Probability Theory — Claudio Landim, Lecture 1: Introduction

Source. Full lecture: Master Program: Probability Theory — Lecture 1, Introduction (Claudio Landim). The lecture builds probability from purely measure-theoretic foundations; the timestamps below pair each idea with a first-principles reading of why it has to be set up the way it is, stripped of the heavy notation.

13.4.1 Foundations and maps

- **00:25 — Probability measure space.** To talk about probability rigorously we set up a three-part stage: Ω, the sample space (the set of all possible outcomes); $\mathcal{F}$, the σ-algebra (the events we are allowed to ask "what is its probability?" about); and P, the probability measure (the ruler that assigns each event a size, or likelihood). Why must the total probability be exactly 1? Because 1 stands for absolute certainty: "something happens" is the conservation law of the whole construction.

- **02:18 — Random variables.** Abstract events — "it rains tomorrow," "the coin shows tails" — cannot be added or multiplied as they stand. The random variable is the bridge. Despite the name it is not really a "variable" but a translation machine: a map that forces each abstract outcome into a real number. Only once outcomes have become numbers can calculus get to work on them.

- **03:33 — Probability distribution measures.** Once the random variable has translated events into real numbers, the probability mass they carried moves onto the real line. A distribution measure is the bookkeeping of how that mass — picture it as clay — is laid out, and how much weight sits where.

13.4.2 Tools for characterizing a distribution

- **05:14 — Distribution functions.** Why invent the cumulative distribution function $F(x) = P(X \leq x)$? On a continuous line the probability of any single exact point is 0 — no one is exactly some height to infinite precision. So we switch to cumulative thinking: sweep from $-\infty$ and collect all the probability up to x. This turns scattered mass into one curve that climbs monotonically from 0 to 1, giving every distribution a single common standard to be compared on.

- **13:03 — Three kinds of distribution, and the decomposition theorem.** How many ways can the probability "clay" lie on the line? The theory allows exactly three:

  (1) *Discrete* — the clay clumps at isolated points (a die's faces).

  (2) *Absolutely continuous* — the clay is spread smoothly, like butter, with no lumps and a density everywhere (height, temperature).

  (3) *Singular* (17:45) — the deeply counterintuitive case, the Cantor distribution: the clay sits on a fractal set of length zero yet genuinely carries mass.

  Every distribution on the line is a convex combination — a mixture — of these three pure types.

13.4.3 Core quantities and limit laws

- **22:11** — **Expectation.** Drop the integral sign and the expectation is the system's center of mass. Picture the real line as a seesaw and the probability as weights laid along it; the expectation is the single fulcrum at which the seesaw balances. It is the long-run average that repeated trials pile up around.

- **40:36** — **Jensen's inequality.** This is the essence of nonlinearity. A convex function curves upward on both sides (the bottom of a smile), so it amplifies spread and extremes. Hence averaging first and then applying the function never exceeds applying the function first and then averaging,

  $$\phi(\mathbb{E}[X]) \leq \mathbb{E}[\phi(X)].$$

  It is also why, with a convex payoff, embracing volatility can raise the long-run expectation.

- **47:29** — **Markov's and Chebyshev's inequalities.** An inference from conservation. Knowing only the mean of a nonnegative quantity, can we bound how often it takes an extreme value? Yes — the total is fixed, so if large values occurred too often they would overwhelm the average. These inequalities turn that single fact, the fixed mean, into an absolute ceiling on the probability of extreme events.