Introduction to Functional Analysis

Foundations for Understanding Neural Networks

Table of Contents

Motivation: Why Functional Analysis?

From Finite to Infinite Dimensions

The Challenge

In machine learning, we work with:

- Finite-dimensional vectors: $\boldsymbol{x} \in \mathbb{R}^n$ Easy
- Infinite-dimensional objects: functions, images, time series Hard

Key Question

Can we extend concepts like length (norm), angle (inner product), and convergence from $\mathbb{R}^n$ to infinite-dimensional spaces?

Answer: Yes! Functional analysis gives us the tools. And it's essential to understand the Cybenko theorem, which proves that neural networks can approximate any continuous function.

Vector Spaces

Vector Spaces: The Foundation

Definition

A real vector space V is a set with operations $+ : V \times V \to V$ and $\cdot : \mathbb{R} \times V \to V$ satisfying:

- Commutativity: $\boldsymbol{u} + \boldsymbol{v} = \boldsymbol{v} + \boldsymbol{u}$
- Associativity: $(\boldsymbol{u} + \boldsymbol{v}) + \boldsymbol{w} = \boldsymbol{u} + (\boldsymbol{v} + \boldsymbol{w})$
- Distributivity: $\lambda \cdot (\boldsymbol{u} + \boldsymbol{v}) = \lambda \cdot \boldsymbol{u} + \lambda \cdot \boldsymbol{v}$
- Existence of zero: $\boldsymbol{u} + 0 = \boldsymbol{u}$ for all $\boldsymbol{u} \in V$

Importance

Vector spaces provide the abstract framework allowing us to treat functions and infinite-dimensional objects with the same algebraic rules as familiar vectors.

Vector Spaces: Examples

Common Examples

- $\mathbb{R}^n$: all n-dimensional real vectors
- $\mathcal{P}_k(I)$: polynomials of degree $\leq k$ on interval I
- $C^0(I)$: continuous functions on interval I
- $L^2(\Omega)$: square-integrable functions on domain Ω

Basis and Dimension

A **basis** of V is a minimal set of linearly independent vectors that span V.
The number of basis elements is the **dimension**.
Any $\boldsymbol{v} \in V$ can be written as:

$$\boldsymbol{v} = c_1\phi_1 + c_2\phi_2 + \cdots + c_n\phi_n$$

Inner Products and Norms

Inner Product: Measuring Angles

Definition

An **inner product** on vector space V is a function $(\cdot, \cdot) : V \times V \to \mathbb{R}$ such that:

1. $(\boldsymbol{u}, \boldsymbol{u}) \geq 0$ and $(\boldsymbol{u}, \boldsymbol{u}) = 0 \Leftrightarrow \boldsymbol{u} = 0$
2. $(\boldsymbol{u}, \boldsymbol{v}) = (\boldsymbol{v}, \boldsymbol{u})$ (symmetry)
3. $(\alpha \boldsymbol{u} + \beta \boldsymbol{v}, \boldsymbol{w}) = \alpha(\boldsymbol{u}, \boldsymbol{w}) + \beta(\boldsymbol{v}, \boldsymbol{w})$ (linearity)

Why It Matters

Inner products let us measure **angles** and **orthogonality** in infinite-dimensional spaces!

Inner Product: Examples

In $\mathbb{R}^n$ (Euclidean)

$$(\boldsymbol{v}, \boldsymbol{w}) = \sum_{i=1}^{n} v_i w_i = \boldsymbol{v} \cdot \boldsymbol{w}$$

In $C([0,1])$ (Continuous functions)

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

In $C^1([0,1])$ (Differentiable functions)

$$(f, g) = \int_0^1 [f(x)g(x) + f'(x)g'(x)]\,dx$$

Note: An inner product *always* induces a norm: $\|\boldsymbol{u}\| = \sqrt{(\boldsymbol{u}, \boldsymbol{u})}$

Norm: Measuring Length

Definition

A **norm** $\|\cdot\| : V \to \mathbb{R}$ satisfies:

1. Triangle inequality: $\|\boldsymbol{u} + \boldsymbol{v}\| \leq \|\boldsymbol{u}\| + \|\boldsymbol{v}\|$
2. Homogeneity: $\|\lambda \boldsymbol{u}\| = |\lambda| \|\boldsymbol{u}\|$
3. Positivity: $\|\boldsymbol{u}\| \geq 0$, with equality iff $\boldsymbol{u} = 0$

Importance

Norms generalize the concept of "distance" from $\mathbb{R}^n$ to any vector space, enabling analysis of convergence and approximation.

Common Norms

Norms in $\mathbb{R}^n$

- $\|\boldsymbol{v}\|_2 = \left(\sum_{i=1}^n |v_i|^2\right)^{1/2}$ (Euclidean norm)
- $\|\boldsymbol{v}\|_\infty = \max_{1\leq i\leq n} |v_i|$ (Max norm)
- $\|\boldsymbol{v}\|_1 = \sum_{i=1}^n |v_i|$ (Manhattan norm)

Norms in Function Spaces

- $\|f\|_{L^2(I)} = \left(\int_I |f(x)|^2\, dx\right)^{1/2}$ (Energy norm)
- $\|f\|_{L^\infty(I)} = \sup_{x\in I} |f(x)|$ (Supremum norm)

Different norms measure different notions of "smallness"!

Cauchy-Schwarz Inequality

Theorem

For an inner product space with induced norm $\|\boldsymbol{u}\| = \sqrt{(\boldsymbol{u}, \boldsymbol{u})}$:

$$|(\boldsymbol{u}, \boldsymbol{v})| \leq \|\boldsymbol{u}\| \cdot \|\boldsymbol{v}\|$$

Intuition

This is the abstract version of "$\boldsymbol{u} \cdot \boldsymbol{v} = \|\boldsymbol{u}\|\|\boldsymbol{v}\| \cos\theta$". The dot product can't exceed the product of lengths!

Example

In $C([0,1])$: $\left|\int_0^1 f(x)g(x)\,dx\right| \leq \sqrt{\int_0^1 f^2(x)\,dx} \cdot \sqrt{\int_0^1 g^2(x)\,dx}$

Convergence and Completeness

Sequences in Normed Spaces

Cauchy Sequence

A sequence $\{\boldsymbol{v}_i\}_{i=1}^{\infty}$ is **Cauchy** if for all $\epsilon > 0$, there exists n such that:

$$\|\boldsymbol{v}_i - \boldsymbol{v}_j\| \leq \epsilon \quad \text{for all } i, j \geq n$$

Convergent Sequence

$\{\boldsymbol{v}_i\}_{i=1}^{\infty}$ **converges** to $\boldsymbol{v}$ if for all $\epsilon > 0$, there exists n such that:

$$\|\boldsymbol{v} - \boldsymbol{v}_i\| \leq \epsilon \quad \text{for all } i \geq n$$

Question: Is every Cauchy sequence convergent?

Not Always! A Classic Example

Counterexample: Rationals

Consider $\mathbb{Q}$ (rational numbers) with absolute value as norm.
The sequence $\{u_n\}_{n\in\mathbb{N}}$ defined by:

$$u_n = \frac{1}{0!} + \frac{1}{1!} + \frac{1}{2!} + \cdots + \frac{1}{n!}$$

is Cauchy in $\mathbb{Q}$ but converges to $e \notin \mathbb{Q}$!

Completeness

A normed vector space is **complete** if every Cauchy sequence converges to an element in the space.

Key insight: Completeness ensures our analysis is "closed" — we don't escape the space!

Banach and Hilbert Spaces

Banach and Hilbert Spaces: The Main Characters

Banach Space

A **Banach space** is a complete normed vector space.

Hilbert Space

A **Hilbert space** is a complete inner product vector space.

The Relationship

Every Hilbert space is a Banach space (completeness + inner product $\Rightarrow$ completeness + norm), but not every Banach space is a Hilbert space (a Banach space might have only a norm, not an inner product).

Why they matter: These spaces are *closed* under limits, so analysis works reliably!

Examples of Banach and Hilbert Spaces

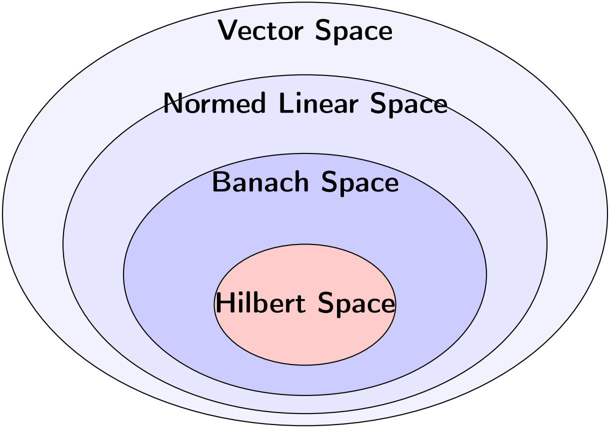
Banach Spaces

- $(\mathbb{R}^n, \|\cdot\|_p)$ for any $p \geq 1$
- $(L^p(\Omega), \|\cdot\|_{L^p})$ for any $p \geq 1$
- $(C(\Omega), \|\cdot\|_\infty)$ (continuous functions)

Hilbert Spaces (Special Banach Spaces)

- $(\mathbb{R}^n, \|\cdot\|_2)$ with dot product
- $(L^2(\Omega), \|\cdot\|_{L^2})$ with $\langle f, g \rangle = \int_\Omega fg \, d\mu$
- $(H^1(\Omega), \|\cdot\|_{H^1})$ (Sobolev spaces, more on this later!)

Hierarchy of Spaces



The Riemann Integral and Its Limitations

Riemann Integral: A Quick Reminder

The Idea

For a function $f : [a, b] \to \mathbb{R}$, partition the interval into boxes and approximate the area:

- Lower sum: $L(f, P) = \sum_{k=1}^{N} m_k(x_k - x_{k-1})$ where $m_k = \inf f$ on $[x_{k-1}, x_k]$
- Upper sum: $U(f, P) = \sum_{k=1}^{N} M_k(x_k - x_{k-1})$ where $M_k = \sup f$ on $[x_{k-1}, x_k]$

Riemann Integrability

f is Riemann integrable if:

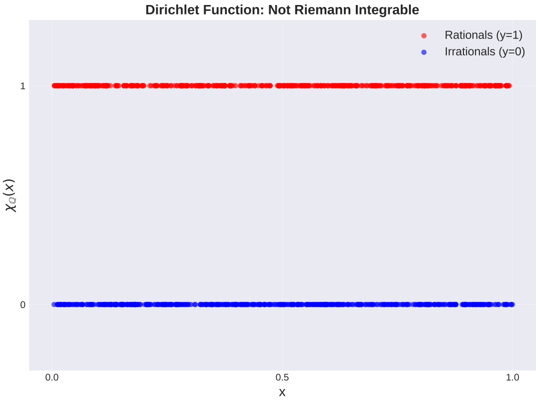
$$\inf_P U(f, P) = \sup_P L(f, P)$$

Simple, intuitive, and works for most "nice" functions!

A Problematic Function - I

The Dirichlet Function

$$\chi_{\mathbb{Q}}(x) = \begin{cases} 1 & \text{if } x \in \mathbb{Q} \\ 0 & \text{if } x \notin \mathbb{Q} \end{cases}$$



A Problematic Function - II

The Problem

For *any* partition:

- $L(f, P) = 0$ (infimum on each interval is 0)
- $U(f, P) = 1$ (supremum on each interval is 1)

So $L(f, P) \neq U(f, P)$, and this function is **not Riemann integrable**!

Why Should We Care About This?

Engineering Perspective

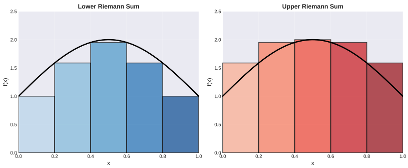
- We need to work with *pathological* functions (e.g., discontinuous signals, noise)
- The Dirichlet function, while strange, appears in theory (sets of measure zero, etc.)
- We want to integrate in multiple dimensions and swap limits with integrals

Three Motivations for Lebesgue Integral

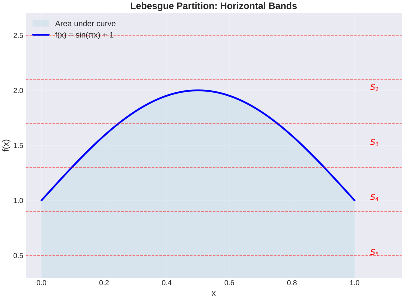
1. **Handle pathological functions:** Integrate discontinuous, "weird" functions
2. **Multidimensional integration:** Easier to prove Fubini's theorem
3. **Limit exchange:** Can swap $\int$ and lim under mild conditions

The Lebesgue Integral

Riemann: Partition the domain



Lebesgue: Partition the range



The Key Difference

Instead of dividing the x-axis, we divide the y-axis and sum areas of *level sets*:

$$\int f \, d\mu \approx \sum_i \alpha_i \mu(S_i)$$

where $S_i = \{x : f(x) \in [\alpha_i, \alpha_{i+1}]\}$

Measure: The Foundation of Lebesgue Integration

Measure μ

A **measure** is a function $\mu : \mathcal{A} \to [0, \infty]$ that assigns "size" to sets:

- $\mu(\emptyset) = 0$
- Countable additivity: $\mu(\bigcup_i A_i) = \sum_i \mu(A_i)$ for disjoint sets

A Crucial Fact

Every countable set has measure zero.
This is why the Dirichlet function finally becomes integrable!

Example: The rational numbers $\mathbb{Q} \subset [0, 1]$ have measure zero, so they contribute nothing to the integral.

The Dirichlet Function Revisited

Lebesgue Integration of $\chi_{\mathbb{Q}}$

The Dirichlet function on $[0,1]$:

$$\chi_{\mathbb{Q}}(x) = \begin{cases} 1 & \text{if } x \in \mathbb{Q} \\ 0 & \text{if } x \notin \mathbb{Q} \end{cases}$$

Lebesgue Integral

$$\int_0^1 \chi_{\mathbb{Q}}(x)\,dx = 1 \cdot \mu(\mathbb{Q} \cap [0,1]) + 0 \cdot \mu([0,1] \setminus \mathbb{Q})$$
$$= 1 \cdot 0 + 0 \cdot 1 = 0$$

because $\mathbb{Q} \cap [0,1]$ is countable and thus has measure zero!

Remark: With Lebesgue, pathological functions become manageable!

Three Reasons to Use Lebesgue

1. Handle Pathological Functions

Riemann: Cannot integrate discontinuous-everywhere functions.
Lebesgue: Integrates discontinuities on sets of measure zero without trouble.

2. Multidimensional Integration (Fubini)

For Riemann, Fubini's theorem is complicated.
For Lebesgue, Fubini is elegant: $\int_{\mathbb{R}^2} f\, d\mu = \int \left(\int f(x, y)\, dy \right) dx$

3. Limit Exchange

Monotone Convergence Theorem: If $f_n \uparrow f$, then $\lim_n \int f_n = \int f$
Dominated Convergence Theorem: If $|f_n| \leq g$ and $f_n \to f$ a.e., then $\lim_n \int f_n = \int f$
(These are essential for neural networks and approximation theory!)

Spaces of Integrable Functions

L^p Spaces: Functions with Finite Power

Definition

For $1 \leq p < \infty$: $L^p(\Omega) = \left\{ f : \Omega \to \mathbb{R} \,\middle|\, \|f\|_{L^p} < \infty \right\}$

where the L^p norm is:

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

$$\|f\|_{L^\infty(\Omega)} = \sup_{x \in \Omega} |f(x)|, \quad p = \infty$$

Intuition

- L^1: Absolutely integrable functions
- L^2: Square-integrable, "finite energy" functions
- L^∞: Bounded functions

Properties of L^p Spaces I

Theorem

For all $1 \leq p \leq \infty$, $L^p(\Omega)$ is a **Banach space**.

Special Case: L^2

For $p = 2$, $L^2(\Omega)$ is actually a **Hilbert space** with inner product:

$$\langle f, g \rangle_{L^2} = \int_\Omega f(x)g(x)\,dx$$

and induced norm:

$$\|f\|_{L^2} = \sqrt{\langle f, f \rangle_{L^2}} = \sqrt{\int_\Omega |f(x)|^2\,dx}$$

Properties of L^p Spaces II

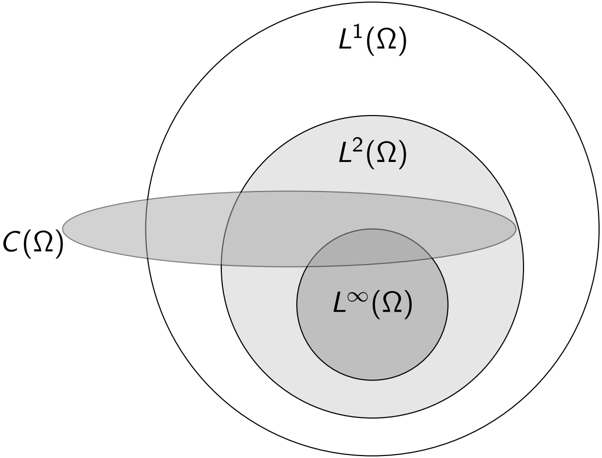
Why This Matters

Being a Hilbert space, L^2 has all the geometric structure we love: orthogonality, projections, Fourier series, etc. This is *essential* for approximation theory and the Cybenko theorem!

Key Fact

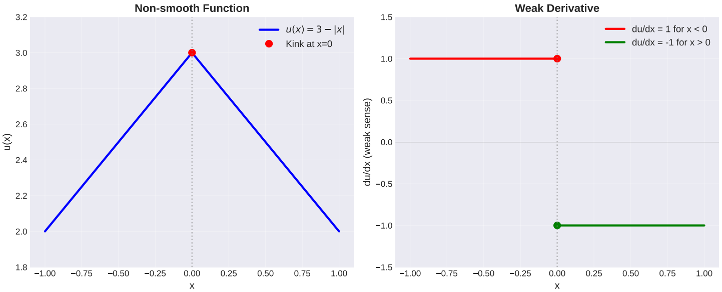
$L^1(\Omega) \supset L^2(\Omega) \supset L^\infty(\Omega)$ for finite measure domains.
(If $\mu(\Omega) < \infty$ and $\|f\|_{L^\infty}$ is small, then f is also in L^p for any p.)

Space Inclusions (Bounded Domain)



Weak Derivatives and Sobolev Spaces

The Motivation: Non-Smooth Functions



The Problem

Consider a function like $u(x) = 3 - |x|$. It's continuous but has a *kink* at $x = 0$!

The classical derivative doesn't exist at that point. But can we still make sense of "the derivative" in a weaker sense?

Weak Derivative: Integration by Parts

Classical Integration by Parts

If $u, \phi \in C^1$ and ϕ has compact support:

$$\int_\Omega u'(x)\phi(x)\,dx = -\int_\Omega u(x)\phi'(x)\,dx$$

Weak Derivative Definition

We say $g \in L^1_{\text{loc}}(\Omega)$ is the **weak derivative** of $\boldsymbol{u}$ if:

$$\int_\Omega g(x)\phi(x)\,dx = -\int_\Omega u(x)\phi'(x)\,dx$$

for *all* test functions $\phi \in \mathcal{D}(\Omega)$ (smooth, compactly supported).
We write $g = \frac{du}{dx}$ in the weak sense.

Weak Derivative: The Kink Example

Example: $u(x) = 3 - |x|$ on $(-1, 1)$

The weak derivative is:

$$g(x) = \begin{cases} 1 & -1 < x \leq 0 \\ -1 & 0 < x < 1 \end{cases}$$

Even though u is not classically differentiable at $x = 0$!

Why This Works

The key is **integration by parts**. The kink doesn't matter for the integral — only the jump in the derivative does.

Remark

Weak and classical derivatives share properties: linearity, product rule, chain rule. When a function is classically differentiable, its weak derivative equals its classical derivative.

Definition

The **Sobolev space** $W^{k,p}(\Omega)$ is:

$$W^{k,p}(\Omega) = \left\{ u \in L^p(\Omega) \,\middle|\, D^\alpha u \in L^p(\Omega) \text{ for all } |\alpha| \leq k \right\}$$

with norm:

$$\|u\|_{W^{k,p}} = \left(\sum_{|\alpha| \leq k} \|D^\alpha u\|_{L^p}^p \right)^{1/p}$$

Sobolev Spaces: Functions with Weak Derivatives II

Important Case: $p = 2$

The notation $H^k(\Omega) = W^{k,2}(\Omega)$ is standard. For $k = 1$:

$$H^1(\Omega) = \left\{ u \in L^2(\Omega) \,\middle|\, \nabla u \in [L^2(\Omega)]^d \right\}$$

with inner product and norm:

$$\langle u, v \rangle_{H^1} = \int_\Omega (uv + \nabla u \cdot \nabla v)\, dx, \quad \|u\|_{H^1} = \sqrt{\|u\|_{L^2}^2 + \|\nabla u\|_{L^2}^2}$$

Sobolev Spaces: Why They Matter

Key Properties

- $W^{k,p}(\Omega)$ is a **Banach space** for all $1 \leq p < \infty$
- $H^k(\Omega) = W^{k,2}(\Omega)$ is a **Hilbert space**
- Weak derivatives allow us to work with non-smooth functions

For PDEs and Neural Networks

Sobolev spaces are the **natural setting** for solving differential equations and analyzing neural networks because:

- Functions can be non-smooth but still belong to H^1
- We can take weak derivatives and do calculus
- The Hilbert space structure (H^2) provides geometric tools

Remark: Functions in $H^1(\Omega)$ are continuous for $d = 1$, might have isolated discontinuities for $d = 2$, and can be discontinuous on curves for $d = 3$.

Linear Operators and Functionals

Linear Operators: Generalized Matrices

Definition

A **linear operator** $T : X \to Y$ (where X, Y are vector spaces) satisfies:

$$T(\alpha \boldsymbol{u} + \beta \boldsymbol{v}) = \alpha T(\boldsymbol{u}) + \beta T(\boldsymbol{v})$$

for all $\boldsymbol{u}, \boldsymbol{v} \in X$ and scalars $\alpha, \beta \in \mathbb{R}$.

Bounded (Continuous) Operator

T is **bounded** if there exists $K > 0$ such that:

$$\|T\boldsymbol{u}\|_Y \leq K\|\boldsymbol{u}\|_X \quad \text{for all } \boldsymbol{u} \in X$$

The **operator norm** is: $\|T\| = \sup_{\boldsymbol{u} \neq 0} \frac{\|T\boldsymbol{u}\|_Y}{\|\boldsymbol{u}\|_X}$

Intuition: Bounded operators don't "blow up" inputs. Like matrices with finite norm!

Linear Functionals and Dual Spaces

Linear Functional

A **linear functional** $\ell : X \to \mathbb{R}$ is a linear operator mapping to scalars.

Dual Space

The **dual space** X' is the set of all bounded linear functionals on X:

$$X' = \{\ell : X \to \mathbb{R} \mid \ell \text{ is linear and bounded}\}$$

The dual norm is:

$$\|\ell\|_{X'} = \sup_{\boldsymbol{u} \neq 0} \frac{|\ell(\boldsymbol{u})|}{\|\boldsymbol{u}\|_X}$$

Importance

The dual space is itself a Banach space! It represents all possible "measurements" we can make on X.

The Riesz Representation Theorem

Theorem (Riesz Representation)

Let H be a Hilbert space and ℓ a bounded linear functional on H. Then there exists a **unique** element $\boldsymbol{u} \in H$ such that:

$$\ell(\boldsymbol{v}) = \langle \boldsymbol{v}, \boldsymbol{u} \rangle \quad \text{for all } \boldsymbol{v} \in H$$

Moreover, $\|\ell\|_{H'} = \|\boldsymbol{u}\|_H$.

Why This Is Profound

In a Hilbert space, **every linear functional looks like an inner product!** This is the bridge between abstract functionals and concrete inner products.

Riesz Representation: Examples I

Example 1: In $\mathbb{R}^n$

Every linear functional $f : \mathbb{R}^n \to \mathbb{R}$ can be written as:

$$f(\boldsymbol{x}) = \boldsymbol{x} \cdot \boldsymbol{y}$$

for some $\boldsymbol{y} \in \mathbb{R}^n$.
Example: $f(\boldsymbol{x}) = x_1 + x_2 + \cdots + x_n = \boldsymbol{x} \cdot (1, 1, \ldots, 1)$

Riesz Representation: Examples II

Example 2: In $L^2(0,1)$

Consider the functional:

$$\ell(f) = \int_0^{1/2} f(x)\, dx$$

By Riesz, there exists unique $u \in L^2(0,1)$ such that:

$$\int_0^{1/2} f(x)\, dx = \int_0^1 f(x)u(x)\, dx$$

Clearly: $u(x) = \begin{cases} 1 & 0 < x \leq 1/2 \\ 0 & 1/2 < x < 1 \end{cases}$

Approximation Theory

The Approximation Problem

Central Question

Can we approximate arbitrary continuous functions by simpler functions?

Why This Matters

- **Neural networks:** We approximate with compositions of simple nonlinearities
- **Numerical methods:** We approximate solutions with polynomials
- **Compression:** We approximate high-dimensional data with low-rank structures

Key Theorem (Weierstrass)

Any continuous function on a closed interval can be uniformly approximated by polynomials.
More formally: For $f \in C([a, b])$ and $\epsilon > 0$, there exists polynomial $p(x)$ such that $\|f - p\|_\infty < \epsilon$.

Density and Closure

Density

A subset $S \subseteq X$ is **dense** in normed space X if for every $\boldsymbol{u} \in X$ and $\epsilon > 0$, there exists $\boldsymbol{s} \in S$ such that:

$$\|\boldsymbol{u} - \boldsymbol{s}\| < \epsilon$$

Meaning

A family of functions is *dense* in a space if, no matter which target function you pick in that space, and no matter how small an error you allow, you can find a function from the family that is as close as you like to the target.

Examples

- Polynomials are dense in $C([a, b])$ (Weierstrass)
- Trigonometric polynomials are dense in $L^2(0, 2\pi)$ (Fourier series)
- Continuous functions are dense in $L^p(\Omega)$