Uncertainty Quantification and Quasi-Monte Carlo

Sommersemester 2025

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

First lecture, April 15, 2025
Second lecture, April 22, 2025
Preliminaries

Practical matters

- Lectures on Mondays at 10:15-11:45 in A6/032 by Vesa Kaarnioja.
- Exercises on Tuesdays at 10:15-11:45 in A6/032 by Vesa Kaarnioja.
- *There will be no lectures on April 14, April 21, and June 9.*
- *The first and second lecture will be held April 15 and April 22 in room A3/120 in place of the exercise session*.
- Exercise sheets will be published regularly on the course Whiteboard page. Please submit your solutions to the exercises before the deadlines specified on each exercise sheet.
- The conditions for completing this course are
  (1) *successfully earning a cumulative 60% of points from the exercises* (active participation + regular attendance), and
  (2) *successfully passing the course exam.*

  The course evaluation is based on the oral exam at the end of the course.

Exercise guidelines

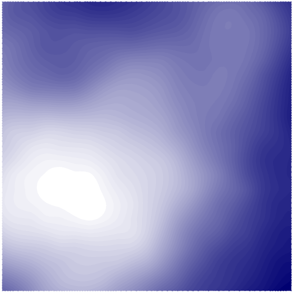
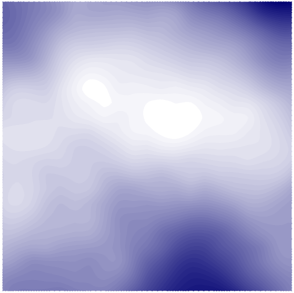
- Solutions to exercises can be submitted either via email or by handing in your solutions at the exercise session by the specified deadline. Late submissions will *not* be considered.
- Please present your calculations clearly and neatly, providing explanation for all steps.
- Ensure that your arguments are coherent and presented in an orderly fashion. Organize your solutions logically, starting from the problem statement and proceeding step-by-step to the solution.
- Typeset or write your solutions in clear handwriting for easy readability.
- Avoid ambiguity in your solutions: consider the perspective or the reader and ensure that your solutions are understandable from their point of view (i.e., the reader should not have to guess what you have written).
- Use appropriate mathematical notation and terminology.
- Double-check your solutions for errors and correctness before submission. Aim for precision and accuracy in your mathematical expressions and calculations.
- In programming tasks, ensure that your program executes successfully. Include the source code as well as the output of the program as part of your submission.

Uncertainty in groundwater flow

Risk analysis of radwaste disposal or CO_2 sequestration.

$$\begin{array}{ll} \text{Darcy's law} & \boldsymbol{q}(\boldsymbol{x}) + \boldsymbol{a}(\boldsymbol{x}) \nabla \boldsymbol{p}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x}) \\ \text{mass conservation law} & \nabla \cdot \boldsymbol{q}(\boldsymbol{x}) = 0 \end{array}$$

in $D \subset \mathbb{R}^d$, $d \in \{1, 2, 3\}$
together with boundary conditions

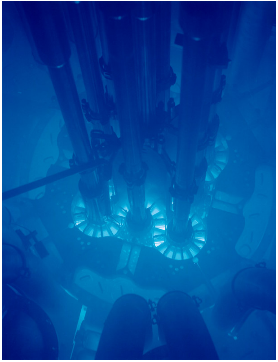
 

Uncertainty in $\boldsymbol{a}(\boldsymbol{x}, \boldsymbol{\omega})$ leads to uncertainty in $\boldsymbol{q}(\boldsymbol{x}, \boldsymbol{\omega})$ and $\boldsymbol{p}(\boldsymbol{x}, \boldsymbol{\omega})$

Criticality problem for nuclear reactors

$$-\nabla \cdot (\underbrace{a(\boldsymbol{x})}_{\text{diffusion}} \nabla u(\boldsymbol{x})) + \underbrace{b(\boldsymbol{x})}_{\text{absorption}} u(\boldsymbol{x}) = \lambda \underbrace{c(\boldsymbol{x})}_{\text{fission}} u(\boldsymbol{x})$$

- The smallest eigenvalue $\lambda_1 \in \mathbb{R}$ measures *criticality* of a reactor.
- Eigenfunction $u_1(\boldsymbol{x})$ is the *neutron flux* at the point $\boldsymbol{x}$.



Source: Argonne National Laboratory on Flickr

- $\lambda_1 \approx 1 \Rightarrow$ operating efficiently
- $\lambda_1 > 1 \Rightarrow$ not self-sustaining
- $\lambda_1 < 1 \Rightarrow$ supercritical

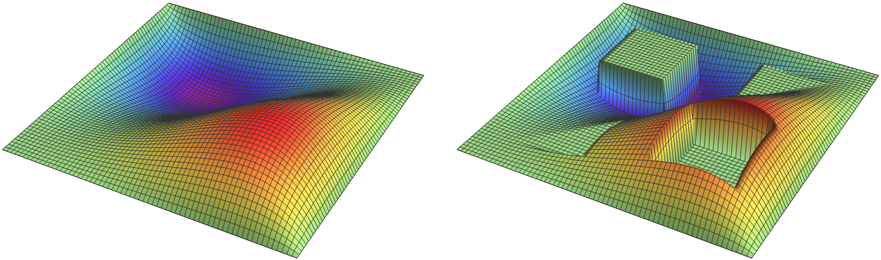
Optimization under uncertainty

Find $\min_{z \in L^2(D)} J(u, z)$,

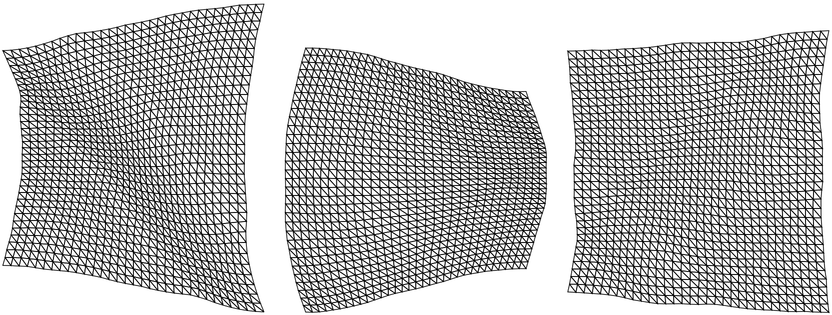
$$J(u,z) := \frac{1}{2}\int_{\Omega}\int_{D}(u(\boldsymbol{x},\omega) - g(\boldsymbol{x}))^2 \,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\mathbb{P}(\omega) + \frac{\alpha}{2}\int_{D} z(\boldsymbol{x})^2 \,\mathrm{d}\boldsymbol{x},$$

subject to

$$\begin{cases} -\nabla \cdot (a(\boldsymbol{x},\omega)\nabla u(\boldsymbol{x},\omega)) = z(\boldsymbol{x}), & \boldsymbol{x} \in D,\ \text{a.e. } \omega \in \Omega \\ u(\boldsymbol{x},\omega) = 0, & \boldsymbol{x} \in \partial D,\ \text{a.e. } \omega \in \Omega, \\ z_{\min}(\boldsymbol{x}) \leq z(\boldsymbol{x}) \leq z_{\max}(\boldsymbol{x}), & \text{a.e. } \boldsymbol{x} \in D. \end{cases}$$



Domain uncertainty quantification

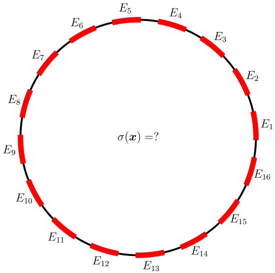


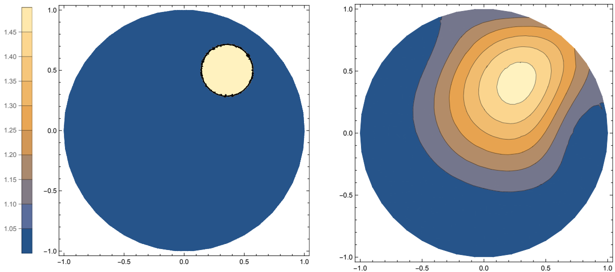
Three realizations of a random spatial domain

Electrical impedance tomography

Use measurements of current and voltage collected at electrodes covering part of the boundary to infer the interior conductivity of an object/body.

$$\begin{cases} \nabla \cdot (\sigma \nabla u) = 0 & \text{in } D, \\ \sigma \frac{\partial u}{\partial \boldsymbol{n}} = 0 & \text{on } \partial D \setminus \bigcup_{k=1}^{L} \overline{E_k}, \\ u + z_k \sigma \frac{\partial u}{\partial \boldsymbol{n}} = U_k & \text{on } E_k,\ k \in \{1, \ldots, L\}, \\ \int_{E_k} \sigma \frac{\partial u}{\partial \boldsymbol{n}} \, \mathrm{d}S = I_k, & k \in \{1, \ldots, L\}, \end{cases}$$





Consider the elliptic PDE problem:

$$\begin{cases} -\nabla \cdot \big(a(\boldsymbol{x})\nabla u(\boldsymbol{x})\big) = f(\boldsymbol{x}) & \text{for } \boldsymbol{x} \in D, \\ +\text{boundary conditions.} \end{cases}$$

In practice, one or several of the material/system parameters may be uncertain or incompletely known and modeled as random fields:

- PDE coefficient a may be uncertain;
- Source term f may be uncertain;
- Boundary conditions may be uncertain;
- The domain D itself may be uncertain.

In forward uncertainty quantification, one is interested in assessing how uncertainties in the inputs of a mathematical model affect the output.
$\Rightarrow$ If the uncertain inputs are modeled as random fields, then the output of the PDE is also a random field. One may be interested in assessing the statistical response of the system, for example, the expectation or variance of the PDE solution (or some other quantity of interest thereof).

High-dimensional numerical integration

$$\int_{[0,1]^s} f(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} \approx \sum_{i=1}^{n} w_i f(\boldsymbol{t}_i)$$

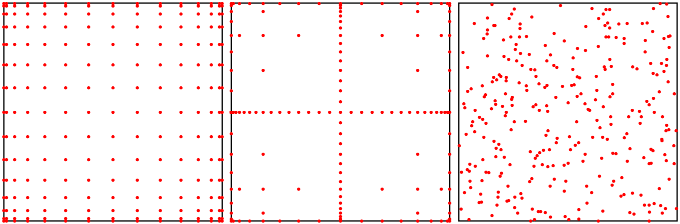


Figure: Tensor product grid, sparse grid, Monte Carlo nodes (not QMC rules)

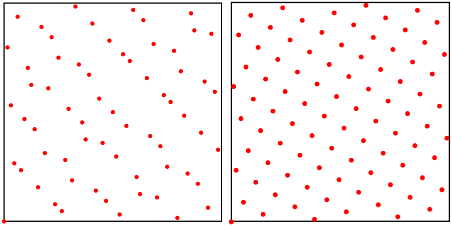


Figure: Sobol′ points, lattice rule (examples of QMC rules)

Quasi-Monte Carlo (QMC) methods are a class of *equal weight* cubature rules

$$\int_{[0,1]^s} f(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} \approx \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{t}_i),$$

where $(\boldsymbol{t}_i)_{i=1}^n$ is an ensemble of *deterministic* nodes in $[0,1]^s$.

The nodes $(\boldsymbol{t}_i)_{i=1}^n$ are NOT random! Instead, they are *deterministically chosen*.

QMC methods exploit the smoothness and anisotropy of an integrand in order to achieve better-than-Monte Carlo rates.

Course contents

- Preliminaries: Hilbert spaces, Sobolev spaces, elliptic partial differential equations (PDEs)
- Finite element (FE) method
- Modeling random field inputs
- Elliptic PDEs with random coefficients
- Quasi-Monte Carlo (QMC) methods
- QMC-FE methods for uncertainty quantification of elliptic PDEs with random coefficients

Preliminary functional analysis

Inner product space

A real vector space X is an *inner product space* if there exists a mapping $\langle \cdot, \cdot \rangle \colon X \times X \to \mathbb{R}$ satisfying

- $\langle ax_1 + bx_2, y \rangle = a\langle x_1, y \rangle + b\langle x_2, y \rangle$ for all $x_1, x_2, y \in X$ and $a, b \in \mathbb{R}$;
- $\langle x, y \rangle = \langle y, x \rangle$ for all $x, y \in X$;
- $\langle x, x \rangle \geq 0$ for all $x \in X$, where equality holds iff $x = 0$.

A mapping $\langle \cdot, \cdot \rangle$ satisfying these conditions is called an *inner product*.

Example

i) $\mathbb{R}^n = \{(x_1, \ldots, x_n) \mid x_k \in \mathbb{R}\}$. Then the inner product is the Euclidean dot product

$$\langle x, y \rangle = \sum_{k=1}^{n} x_k y_k, \quad x = (x_1, \ldots, x_n),\ y = (y_1, \ldots, y_n).$$

ii) Let $X = C([a, b]) = \{f \mid f \colon [a, b] \to \mathbb{R} \text{ is continuous}\}$ and define

$$\langle f, g \rangle = \int_a^b f(x)g(x)\,\mathrm{d}x.$$

Then this is an inner product on $C([a, b])$.

iii) Let $X = \ell^2(\mathbb{R}) = \left\{(z_k)_{k=1}^{\infty} \mid \sum_{k=1}^{\infty} |z_k|^2 < \infty\right\}$. Then $\ell^2(\mathbb{R})$ is an inner product space when

$$\langle x, y \rangle = \sum_{k=1}^{\infty} x_k y_k, \quad x = (x_1, x_2, \ldots),\ y = (y_1, y_2, \ldots).$$

Definition

A real vector space X is a *normed space* if there exists a mapping $\|\cdot\|: X \to \mathbb{R}$ satisfying

- $\|ax\| = |a|\|x\|$ for all $a \in \mathbb{R}$ and $x \in X$;
- $\|x\| \geq 0$ for all $x \in X$, where equality holds iff $x = 0$.
- $\|x + y\| \leq \|x\| + \|y\|$ for all $x, y \in X$ (triangle inequality).

If X is an inner product space, then it is a normed space in a canonical way with the induced norm $\|\cdot\|: X \to \mathbb{R}$ defined by

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

The first two postulates follow immediately from the properties of inner product spaces, the triangle inequality follows from the Cauchy–Schwarz inequality.

Proposition (Cauchy–Schwarz inequality)

If $(X, \langle\cdot,\cdot\rangle)$ is an inner product space, then

$$|\langle x, y\rangle| \leq \|x\|\|y\| \quad \textit{for all } x, y \in X.$$

Proof. Let $x, y \in X$ and $t \in \mathbb{R}$. If $x = 0$ or $y = 0$, then the claim is trivial. Suppose that $x \neq 0 \neq y$. Then

$$0 \leq \langle x + ty, x + ty \rangle = \|x\|^2 + 2t\langle x, y \rangle + t^2\|y\|^2.$$

This is a second degree polynomial w.r.t. t with at most 1 real root. Hence,

$$\begin{aligned} \text{discriminant} \leq 0 \quad &\Leftrightarrow \quad 4|\langle x, y \rangle|^2 - 4\|x\|^2\|y\|^2 \leq 0 \\ &\Leftrightarrow \quad |\langle x, y \rangle|^2 \leq \|x\|^2\|y\|^2. \end{aligned}$$

Note that if $y = ax$, $a \in \mathbb{R}$, then discriminant $= 0$ and Cauchy–Schwarz holds with equality. $\square$

The triangle inequality is an immediate consequence of Cauchy–Schwarz:

$$\begin{aligned} \|x + y\|^2 &= \langle x + y, x + y \rangle = \|x\|^2 + \|y\|^2 + 2\langle x, y \rangle \\ &\leq \|x\|^2 + \|y\|^2 + 2|\langle x, y \rangle| \leq \|x\|^2 + \|y\|^2 + 2\|x\|\|y\| \\ &= (\|x\| + \|y\|)^2 \quad \text{for all } x, y \in X. \end{aligned}$$

For our purposes, having an inner product is not enough. We need to know that these spaces are also *complete* normed spaces.

Definition (Cauchy sequence)

A sequence $(x_k)_{k=1}^{\infty}$ of elements of $(X, \|\cdot\|)$ is called a *Cauchy sequence* if for all $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that

$$m, n > N \quad \Rightarrow \quad \|x_m - x_n\| < \varepsilon.$$

Definition (Complete space)

A normed space $(X, \|\cdot\|)$ is *complete* if all Cauchy sequences in X converge to an element of X.

Definition (Banach space)

A normed space $(X, \|\cdot\|)$ which is complete with respect to $\|\cdot\|$ is a *Banach space*.

Definition (Hilbert space)

An inner product space $(H, \langle\cdot,\cdot\rangle)$ which is complete with respect to $\|\cdot\| = \sqrt{\langle\cdot,\cdot\rangle}$ defined by the inner product is a *Hilbert space*.

Example

i) $\mathbb{R}^n$ and $\ell^2(\mathbb{R})$ are complete.

ii) $C([a,b])$ is *not* complete w.r.t. the norm

$$\|f\|^2 = \int_a^b |f(x)|^2 \, \mathrm{d}x.$$

Let $a = -1$, $b = 1$, and define

$$f_n(x) := \begin{cases} 0, & -1 \le x < 0, \\ nx, & 0 \le x \le \frac{1}{n}, \\ 1, & \frac{1}{n} < x \le 1. \end{cases}$$

Then f_n is continuous, and if $H(x) = \chi_{[0,1]}(x) = \begin{cases} 0, & -1 \le x \le 0, \\ 1, & 0 < x \le 1, \end{cases}$ we have

$$\int_{-1}^1 |f_n(x) - H(x)|^2 \, \mathrm{d}x = \int_0^{1/n} |nx - 1|^2 \, \mathrm{d}x = \int_0^{1/n} (n^2x^2 - 2nx + 1) \, \mathrm{d}x$$

$$= \left[\frac{n^2x^3}{3} - nx^2 + x \right]_{x=0}^{x=1/n} = \frac{1}{3n} - \frac{1}{n} + \frac{1}{n} = \frac{1}{3n} \xrightarrow{n\to\infty} 0.$$

We have $\|f_n - H\| \to 0$, but $H \notin C([-1,1])$.

However, note that $C([a,b])$ *is* complete w.r.t. the sup-norm $\|f\|_\infty = \sup_{a\le x\le b} |f(x)|$, but $\|\cdot\|_\infty \neq \|\cdot\|$ and *there is no inner product inducing* $\|\cdot\|_\infty$*-norm* (exercise).

If one wishes to consider function spaces equipped with inner product norms, one is led to L^2 spaces.

Definition

Let $D \subset \mathbb{R}^n$ be a Lebesgue measurable set. Then

$$L^2(D) := \{f \mid f : D \to \mathbb{R} \text{ measurable}, \int_D |f(\boldsymbol{x})|^2 \,\mathrm{d}\boldsymbol{x} < \infty\}.$$

We define the inner product

$$\langle f, g \rangle_{L^2(D)} = \int_D f(\boldsymbol{x}) g(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}, \tag{1}$$

which induces the norm

$$\|f\|_{L^2(D)} = \left(\int_D |f(\boldsymbol{x})|^2 \,\mathrm{d}\boldsymbol{x} \right)^{1/2}.$$

Theorem

$L^2(D)$ is a Hilbert space with the inner product (1).

Remark. In practice, we treat the elements of $L^2(D)$ (resp. $L^p(D)$) as functions. Strictly speaking, elements of $L^2(D)$ (resp. $L^p(D)$) are equivalence classes of measurable functions that are equal almost everywhere on D. That is, if $f, g \in L^2(D)$ and $f(\boldsymbol{x}) = g(\boldsymbol{x})$ for almost every $\boldsymbol{x} \in D$, then f and g represent the same element in $L^2(D)$. This identification ensures that $L^2(D)$ is a true normed space (and in fact a Hilbert space), since the norm is zero if and only if the function is zero almost everywhere.

Bounded linear operators in Hilbert spaces

Definition

Let X and Y be normed spaces with norms $\|\cdot\|_X$ and $\|\cdot\|_Y$. A linear operator $A\colon X \to Y$ is said to be *bounded* if there exists $C > 0$ such that

$$\|Ax\|_Y \le C\|x\|_X \quad \text{for all } x \in X.$$

Lemma

Let $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ be normed spaces. Then a linear operator $A\colon X \to Y$ is bounded iff

$$\|A\| := \|A\|_{X\to Y} := \sup_{\|x\|_X \le 1} \|Ax\|_Y < \infty. \qquad \text{(operator norm)}$$

Proof. "$\Rightarrow$" If there is $C > 0$ s.t. $\|Ax\|_Y \le C\|x\|_X$ for all $x \in X$, then clearly $\|A\| = \sup_{\|x\|_X\le 1} \|Ax\|_Y \le C$.
"$\Leftarrow$" Let $\|A\| < \infty$. Since $\|\frac{x}{\|x\|_X}\|_X = 1$ for all $x \neq 0$, from the linearity of A we infer

$$\frac{\|Ax\|_Y}{\|x\|_X} = \|A(\tfrac{x}{\|x\|_X})\|_Y \le \|A\| \quad \text{for all } x \in X.$$

This implies the important estimate

$$\|Ax\|_Y \le \|A\|\|x\|_X \quad \text{for all } x \in X.$$

□

A linear operator is continuous precisely when it is bounded.

Proposition

Let $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ be normed spaces and $A\colon X \to Y$ a linear operator. Then the following are equivalent:

(i) *A is a bounded operator;*

(ii) *A is continuous (in X);*

(iii) *A is continuous at one point $x_0 \in X$.*

Proof. (i) $\Rightarrow$ (ii): if $x, y \in X$ and $\varepsilon > 0$, then

$$\|x-y\|_X \leq \frac{\varepsilon}{\|A\|} =: \delta \quad \Rightarrow \quad \|Ax - Ay\|_Y \overset{A \text{ linear}}{=} \|A(x-y)\|_Y \leq \|A\|\|x-y\|_X \leq \varepsilon.$$

(ii) $\Rightarrow$ (iii): trivial.
(iii) $\Rightarrow$ (i): let A be continuous at $x_0 \in X$. By definition, there exists $\delta > 0$ such that

$$\|y - x_0\|_X \leq \delta \quad \Rightarrow \quad \|Ay - Ax_0\|_Y \leq 1.$$

If $x \in X$ is such that $\|x\|_X \leq \delta$, then by taking $y = x + x_0$:

$$\|Ax\|_Y = \|A(x + x_0) - Ax_0\|_Y \leq 1.$$

On the other hand, for any $\|x\|_X \leq 1$, there holds $\|\delta x\|_X = \delta\|x\|_X \leq \delta$ and thus

$$\delta\|Ax\|_Y = \|A(\delta x)\|_Y \leq 1, \quad \text{i.e.,} \quad \|Ax\|_Y \leq \frac{1}{\delta} \quad \text{for all } \|x\|_X \leq 1.$$

Therefore $\|A\| \leq \frac{1}{\delta}$, meaning that A is bounded. $\square$

Let H be a real Hilbert space.

Definition

Two elements $x, y \in H$ are said to be *orthogonal* if $\langle x, y \rangle = 0$.

Let $M \subset H$ be a subset. The orthogonal complement of M in H is defined as

$$M^\perp := \{y \in H \mid \langle x, y \rangle = 0 \quad \text{for all } x \in M\}.$$

We state the following easy consequences.

Lemma

For any subset $M \subset H$, $M^\perp$ is a closed subspace of H and $M \subset (M^\perp)^\perp$.

Lemma

If M is a subspace of H, then $(M^\perp)^\perp = \overline{M}$.

Proof. Exercise. □

Proposition (Hilbert projection theorem)

Let M be a nonempty, closed, and convex[†] *subset of a real Hilbert space H. Then there exists a unique element $x_0 \in M$ satisfying*

$$\|x_0\| \leq \|x\| \quad \textit{for all } x \in M.$$

Proof. Let $\delta = \inf\{\|x\| \mid x \in M\}$. We use the parallelogram identity $\|u+v\|^2 + \|u-v\|^2 = 2\|u\|^2 + 2\|v\|^2$ (exercise) applied to vectors $u = \frac{1}{2}x$ and $v = \frac{1}{2}y$, $x, y \in M$, to obtain

$$\frac{1}{4}\|x-y\|^2 = \frac{1}{2}\|x\|^2 + \frac{1}{2}\|y\|^2 - \left\|\frac{x+y}{2}\right\|^2.$$

Due to convexity $\frac{1}{2}(x+y) \in M$, so

$$\|x-y\|^2 \leq 2\|x\|^2 + 2\|y\|^2 - 4\delta^2 \quad \text{for all } x, y \in M. \tag{2}$$

Existence: let $(x_k)_{k=1}^{\infty} \subset M$ s.t. $\|x_k\| \xrightarrow{k\to\infty} \delta$. Substituting $x \leftarrow x_n$ and $y \leftarrow x_m$ in (2) yields $\|x_n - x_m\|^2 \leq 2\|x_n\|^2 + 2\|x_m\|^2 - 4\delta^2$, since $\frac{1}{2}(x_n + x_m) \in M$ for all n, m. Thus $\|x_n - x_m\| \to 0$ as $n, m \to \infty$. $(x_k)_{k=1}^{\infty}$ is Cauchy in the Hilbert space H, so there exists $x_0 := \lim_{k\to\infty} x_k \in H$. Since $\|\cdot\|$ is continuous, $\|x_0\| = \lim_{k\to\infty}\|x_k\| = \delta$. Since M is closed and $(x_k)_{k=1}^{\infty} \subset M$, the limit $x_0 \in M$.

Uniqueness: If $\|x\| = \|y\| = \delta \Rightarrow \|x-y\|^2 \leq 0$ by (2) and so $x = y$. □

[†] $tx + (1-t)y \in M$ for all $x, y \in M$, $t \in (0,1)$.

Corollary

Let H be a real Hilbert space, M a nonempty, closed, and convex subset of H, and $x \in H$. Then there exists a unique element $y_0 \in M$ such that

$$\|x - y_0\| = \inf\{\|x - y\| \mid y \in M\}.$$

Proof. The set $x - M := \{x - y \mid y \in M\}$ is closed and convex, and $\min\{\|x - y\| \mid x - y \in x - M\} = \min\{\|x - y\| \mid y \in M\}$. The claim follows from the previous result. $\square$

Proposition (Orthogonal decomposition)

If M is a closed subspace of a real Hilbert space H, then

$$H = M \oplus M^{\perp},$$

which means that every element $y \in H$ can be uniquely represented as

$$y = x + x^{\perp}, \quad x \in M, \; x^{\perp} \in M^{\perp}.$$

Proof. It suffices to prove that $M \cap M^{\perp} = \{0\}$ and $M + M^{\perp} = H$.
• If $x \in M \cap M^{\perp}$, then $0 = \langle x, x \rangle = \|x\|^2$ (i.e., $x \perp x$) so $x = 0$.
$\therefore M \cap M^{\perp} = \{0\}$.
• Let $x \in H$. The Hilbert projection theorem guarantees that there exists a unique $y_0 \in M$ such that

$$\|x - y_0\| \leq \|x - y\| \quad \text{for all } y \in M. \tag{3}$$

Let $x_0 = x - y_0$ so that $x = y_0 + x_0 \in M + x_0$. It remains to show that $x_0 \in M^{\perp}$.
The inequality (3) can be written as

$$\|x_0\| \leq \|z\| \quad \text{for all } z \in x - M.$$

Since $y_0 \in M$ and M is a vector space, $y_0 + M = M$ and $M = -M$ which implies $x - M = x + M = y_0 + x_0 + M = x_0 + M$. The previous inequality can be recast as

$$\|x_0\| \leq \|z\| \quad \text{for all } z \in x_0 + M \quad \Leftrightarrow \quad \|x_0\| \leq \|x_0 + z\| \quad \text{for all } z \in M.$$

This statement is true if and only if $\langle x_0, z \rangle = 0$ for all $z \in M$. Therefore $x_0 \in M^{\perp}$. $\square$

Let M be a closed subspace. The orthogonal decomposition implies that every element $y \in H$ can be uniquely represented as

$$y = x + x^{\perp},\ x \in M,\ x^{\perp} \in M^{\perp}.$$

Lemma

Let $M \subset H$ be a closed subspace. The mapping $P_M\colon H \to M$, $y \mapsto x$, is an orthogonal projection, i.e., $P_M^2 = P_M$ and $\mathrm{Ran}(P_M) \perp \mathrm{Ran}(I - P_M)$. It satisfies the following properties:

- P_M *is linear;*
- $\|P_M\| = 1$ *if* $M \neq \{0\}$;
- $I - P_M = P_{M^{\perp}}$;
- $\|y - P_M y\| \leq \|y - z\|$ *for all* $z \in M$;
- $y \in M \ \Rightarrow\ P_M y = y,\ (I - P_M)y = 0;$
  $y \in M^{\perp} \Rightarrow\ P_M y = 0,\ (I - P_M)y = y.$
- $\|y\|^2 = \|P_M y\|^2 + \|(I - P_M)y\|^2$ *(Pythagoras)*

Proof. See for example [Rudin, Real and Complex Analysis, pp. 34–35]. □

Example

Let H_1 and H_2 be real Hilbert spaces and let $A\colon H_1 \to H_2$ be a continuous linear operator.

The kernel (or null space) of operator A is defined as

$$\mathrm{Ker}(A) := \{x \in H_1 \mid Ax = 0\}.$$

The range (or image) of operator A is defined as

$$\mathrm{Ran}(A) := \{y \in H_2 \mid y = Ax,\ x \in H_1\}.$$

Then we have the following:

- $\mathrm{Ker}(A)$ is a *closed* subspace of H_1, and $\mathrm{Ran}(A)$ is a subspace of H_2.
- $H_1 = \mathrm{Ker}(A) \oplus (\mathrm{Ker}(A))^{\perp}$.
- $H_2 = \overline{\mathrm{Ran}(A)} \oplus (\mathrm{Ran}(A))^{\perp}$.

We denote

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

Proposition

Let X and Y be normed spaces. If Y is complete, then $\mathcal{L}(X, Y)$ is complete w.r.t. the operator norm (i.e., it is a Banach space).

Proof. Let $x \in X$ and assume that $A_k \in \mathcal{L}(X, Y)$, $k \in \mathbb{N}$, is a Cauchy sequence. If $x = 0$, then $A_k 0 = 0$ and the limit $A(0) := \lim_{k\to\infty} A_k 0 = 0$ trivially exists. On the other hand, if $x \neq 0$, then for all $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that

$$m, n > N \quad \Rightarrow \quad \|A_m - A_n\| < \frac{\varepsilon}{\|x\|_X}.$$

Especially,

$$\|A_m x - A_n x\|_Y \leq \|A_m - A_n\| \|x\|_X < \varepsilon \quad \text{when } m, n > N,$$

so $(A_k x)$ is a Cauchy sequence in Y and therefore the limit

$$A(x) := \lim_{k\to\infty} A_k x$$

exists.

It is easy to see that $A(x) := \lim_{k\to\infty} A_k x$ is linear. It is also bounded: there exists $N \in \mathbb{N}$ such that

$$m, n > N \quad \Rightarrow \quad \|A_m - A_n\| < 1.$$

Fix $m > N$. Then for all $n > m$,

$$\|A_n\| < 1 + \|A_m\|$$

and thus

$$\|A_n x\|_Y \leq (1 + \|A_m\|)\|x\|_X.$$

But $\|Ax\|_Y = \lim_{n\to\infty} \|A_n x\|_Y \leq (1 + \|A_m\|)\|x\|_X$. Therefore A is bounded.

Finally, we need to show that $\|A_n - A\| \to 0$ as $n \to \infty$. Since we assumed $(A_k)_{k=1}^{\infty}$ to be Cauchy, let $\varepsilon > 0$ be s.t. for $m, n > N$, there holds $\|A_m - A_n\| < \varepsilon$. Then

$$\|(A - A_n)x\|_Y = \lim_{m\to\infty} \|A_m x - A_n x\|_Y \leq \varepsilon \|x\|_X \quad \text{for all } x \in X$$

$$\Rightarrow \quad \|A - A_n\| < \varepsilon.$$

Hence $\|A - A_n\| \to 0$ as $n \to \infty$. $\square$

If $X = H_1$ and $Y = H_2$ are Hilbert spaces, then $\mathcal{L}(H_1, H_2)$ is a complete normed space.

Definition

Let H be a Hilbert space. The space $H' := \mathcal{L}(H, \mathbb{R})$ is called the *topological dual space* of H.

Corollary

If H is a Hilbert space, then H' is complete w.r.t. the operator norm.

Proof. This is an immediate consequence of the previous proposition since $\mathbb{R}$ is a complete Hilbert space. □

Remark. In general, $\mathcal{L}(H_1, H_2)$ is *not* a Hilbert space even when both H_1 and H_2 are. However, in the special case $H' = \mathcal{L}(H, \mathbb{R})$ it turns out that indeed one can associate an inner product that induces the operator norm $\| \cdot \|$ – meaning that H' is a Hilbert space! This is made possible by the *Riesz representation theorem*.

Existence results

Proposition (Riesz representation theorem)

Let H be a real Hilbert space. If $A\colon H \to \mathbb{R}$ is a bounded linear functional, i.e., A is linear and there exists $C > 0$ such that

$$|A(x)| \leq C\|x\| \quad \textit{for all } x \in H,$$

then there exists a unique *$y \in H$ such that*

$$A(x) = \langle x, y \rangle \quad \textit{for all } x \in H.$$

Proof. If $A \equiv 0$, then $y = 0$ and this is unique. Suppose $A \neq 0$ and let

$$M := \mathrm{Ker}(A) = \{x \in H \mid A(x) = 0\}.$$

Since A is continuous, M is a *closed* subspace of H. Furthermore, by the orthogonal decomposition $H = M \oplus M^{\perp}$, our assumption $A \neq 0$ implies that $M \neq H \Rightarrow M^{\perp} \neq \{0\}$.

Let $x \in H$ and $z \in M^{\perp}$ with $\|z\| = 1$. Define

$$u := A(x)z - A(z)x.$$

Then

$$A(u) = A(x)A(z) - A(z)A(x) = 0$$

meaning that $u \in M$. In particular $\langle u, z \rangle = \langle A(x)z - A(z)x, z \rangle = 0$ and

$$\begin{aligned} A(x) &= A(x) \underbrace{\langle z, z \rangle}_{=\|z\|^2=1} = \langle A(x)z, z \rangle \\ &= \langle A(z)x, z \rangle = A(z)\langle x, z \rangle = \langle x, zA(z) \rangle. \end{aligned}$$

$\therefore$ The element $y = zA(z)$ satisfies $A(x) = \langle x, y \rangle$.
To prove uniqueness, suppose that there exist $y_1, y_2 \in H$ such that

$$A(x) = \langle x, y_1 \rangle = \langle x, y_2 \rangle.$$

Then $\langle x, y_1 - y_2 \rangle = 0$ for all $x \in H$. Choose $x = y_1 - y_2$. Then

$$0 = \langle y_1 - y_2, y_1 - y_2 \rangle = \|y_1 - y_2\|^2 \quad \Leftrightarrow \quad y_1 = y_2.$$

$\square$

The Riesz operator

Let $x \in H$ and consider the linear mapping $f_x : H \to \mathbb{R}$, $z \mapsto \langle z, x\rangle_H$. Note that $f_x \in H'$ since it follows from the Cauchy–Schwarz inequality that

$$|f_x(z)| = |\langle z, x\rangle_H| \leq \|z\|_H \|x\|_H \quad \text{for all } z \in H. \tag{4}$$

Now define the *Riesz operator* $R_H : H \to H'$ as $x \mapsto f_x$.

- R_H is linear: $R_H(ax_1 + bx_2) = f_{ax_1+bx_2} = \langle \cdot, ax_1 + bx_2\rangle_H = a\langle \cdot, x_1\rangle_H + b\langle \cdot, x_2\rangle_H = af_{x_1} + bf_{x_2} = aR_H x_1 + bR_H x_2$ for $x_1, x_2 \in H$, $a, b \in \mathbb{R}$.
- R_H is an isometry ($\|R_H x\|_{H'} = \|x\|_H$): it follows from (4) that $\|R_H x\|_{H'} = \|f_x\|_{H'} = \sup_{\|z\|_H \leq 1} |\langle z, x\rangle_H| \leq \|x\|_H$. The other direction follows from $\|x\|_H^2 = \langle x, x\rangle_H = f_x(x) = |f_x(x)| \leq \|f_x\|_{H'} \|x\|_H = \|R_H x\|_{H'} \|x\|_H$.
- R_H is injective (one-to-one): let $R_H x = R_H y$ for some $x, y \in H$. From linearity, $R_H(x - y) = 0 \Rightarrow f_{x-y} = 0 \Rightarrow \langle x - y, z\rangle_H = 0$ for all $z \in H \Rightarrow x = y$.
- R_H is surjective (onto): by Riesz representation theorem, given $A \in H'$, there exists a unique $x \in H$ satisfying $A(z) = \langle z, x\rangle_H = f_x(z)$ for all $z \in H$. In other words, $A = \langle \cdot, x\rangle_H = f_x = R_H x$.

$\therefore$ The Riesz operator $R_H : H \to H'$ is a bijective linear operator isometry.

Lemma

Let H be a Hilbert space. The dual space $H' := \mathcal{L}(H, \mathbb{R})$ is a Hilbert space induced by

$$\|A\|_{H'} := \sup_{\|x\|_H \leq 1} |Ax| = \sqrt{\langle A, A\rangle_{H'}}, \quad \langle A, B\rangle_{H'} := \langle R_H^{-1} A, R_H^{-1} B\rangle_H.$$

Adjoint operator

Proposition

Let H_1 and H_2 be real Hilbert spaces and suppose that $A \in \mathcal{L}(H_1, H_2)$. Then there exists a unique bounded linear operator $A^*: H_2 \to H_1$, called the adjoint *of A, satisfying $\langle Ax, y\rangle_{H_2} = \langle x, A^* y\rangle_{H_1}$. Moreover, $\|A\|_{H_1 \to H_2} = \|A^*\|_{H_2 \to H_1}$.*

Proof. Let $y \in H_2$ and consider $T_y: H_1 \to \mathbb{R}$, $x \mapsto \langle Ax, y\rangle_{H_2}$. Clearly, T_y is linear and bounded so by the Riesz representation theorem there exists a *unique* $z \in H_1$ s.t.

$$\langle Ax, y\rangle_{H_2} = T_y(x) = \langle x, z\rangle_{H_1} \quad \text{for all } x \in H_1.$$

Define $A^* y := z$.

- Let $a, b \in \mathbb{R}$ and $y_1, y_2 \in H_2$. Linearity follows from $\langle x, A^*(ay_1 + by_2)\rangle = \langle Ax, ay_1 + by_2\rangle = a\langle Ax, y_1\rangle + b\langle Ax, y_2\rangle = a\langle x, A^* y_1\rangle + b\langle x, A^* y_2\rangle = \langle x, aA^* y_1 + bA^* y_2\rangle$. Since $x \in H_1$ was arbitrary, $A^*(ay_1 + by_2) = aA^* y_1 + bA^* y_2$.
- $\|A^*\|_{H_2 \to H_1} = \sup_{\|y\|_{H_2} \le 1} \|A^* y\|_{H_1} \overset{(*)}{=} \sup_{\|y\|_{H_2} \le 1} \sup_{\|x\|_{H_1} \le 1} |\langle A^* y, x\rangle|$

  $= \sup_{\|y\|_{H_2} \le 1} \sup_{\|x\|_{H_1} \le 1} |\langle y, Ax\rangle| \overset{(*)}{=} \sup_{\|x\|_{H_1} \le 1} \|Ax\|_{H_2} = \|A\|_{H_1 \to H_2} < \infty.$ □

$^{(*)}$Let $\Lambda \in \mathcal{L}(H, K)$, H, K Hilbert spaces. Cauchy–Schwarz: $\sup_{\|y\|_K \le 1} |\langle \Lambda x, y\rangle_K| \le \|\Lambda x\|_K$.
Other direction: $\sup_{\|y\|_K \le 1} |\langle \Lambda x, y\rangle_K| \ge |\langle \Lambda x, \frac{1}{\|\Lambda x\|_K} \Lambda x\rangle|_K = \|\Lambda x\|_K$.
$\therefore \|\Lambda x\|_K = \sup_{\|y\|_K \le 1} |\langle \Lambda x, y\rangle_K|$.

Some properties of the adjoint operator

Proposition

Let H_1 and H_2 be real Hilbert spaces and suppose that $A, B \in \mathcal{L}(H_1, H_2)$. Then

(i) $\|A^*A\|_{H_1\to H_1} = \|A\|^2_{H_1\to H_2}$;

(ii) $A^{**} = A$, *where* $A^{**} = (A^*)^*$;

(iii) $(c_1 A + c_2 B)^* = c_1 A^* + c_2 B^*$, $c_1, c_2 \in \mathbb{R}$.

Proof. (i) Let $x \in H_1$, $\|x\|_{H_1} = 1$. By the Cauchy–Schwarz inequality,

$$\|Ax\|^2_{H_2} = \langle Ax, Ax\rangle_{H_2} = \langle x, A^*Ax\rangle_{H_1} \le \|A^*Ax\|_{H_1} \;\Rightarrow\; \|A\|^2_{H_1\to H_2} \le \|A^*A\|_{H_1\to H_1}.$$

Other direction: $\|A^*A\| \le \|A^*\| \cdot \|A\| = \|A\|^2$.
(ii) If $x \in H_1$ and $y \in H_2$, then

$$\langle A^{**}x, y\rangle_{H_2} = \langle x, A^*y\rangle_{H_1} = \langle A^*y, x\rangle_{H_1} = \langle y, Ax\rangle_{H_2} = \langle Ax, y\rangle_{H_2}.$$

Hence $\langle A^{**}x - Ax, y\rangle_{H_2} = 0$ for all $y \in H_2 \Rightarrow A^{**}x = Ax$ for all $x \in H_1 \Rightarrow A^{**} = A$.
(iii) Let $x \in H_1$ and $y \in H_2$. Then

$$\begin{aligned}\langle (c_1A + c_2B)^*y, x\rangle_{H_1} &= \langle y, (c_1A + c_2B)x\rangle_{H_2} = c_1\langle y, Ax\rangle_{H_2} + c_2\langle y, Bx\rangle_{H_2} \\ &= c_1\langle A^*y, x\rangle_{H_1} + c_2\langle B^*y, x\rangle_{H_1} = \langle (c_1A^* + c_2B^*)y, x\rangle_{H_1}.\end{aligned}$$

Similarly to the previous part, we conclude that $(c_1A + c_2B)^* = c_1A^* + c_2B^*$. $\square$

Self-adjoint operators

Definition

Let H be a Hilbert space. The operator $A \in \mathcal{L}(H) := \mathcal{L}(H, H)$ is called *self-adjoint* if $A^* = A$, i.e.,

$$\langle Ax, y\rangle = \langle x, Ay\rangle \quad \text{for all } x, y \in H.$$

Example

Let H be a Hilbert space and let $A, B \in \mathcal{L}(H)$ be self-adjoint operators. Then

(i) $A + B$ is self-adjoint.

(ii) if $c \in \mathbb{R}$, then cA is self-adjoint.

(iii) if $AB = BA$, then AB is self-adjoint.

Parts (i) and (ii) follow immediately from part (iii) on the previous slide. If $x, y \in H$, then

$$\langle ABx, y\rangle = \langle BAx, y\rangle = \langle Ax, By\rangle = \langle x, ABy\rangle \quad \Rightarrow \quad (AB)^* = AB.$$

Example

Let H be a Hilbert space and $M \subset H$ a closed subspace. Then the orthogonal projections $P_M \colon H \to M$ and $I - P_M =: P_{M^\perp} : H \to M^\perp$ are self-adjoint.

Lax–Milgram lemma

Proposition (Lax–Milgram lemma)

Let H be a real Hilbert space and let $B\colon H \times H \to \mathbb{R}$ be a bilinear mapping[†] *with $C, c > 0$ such that*

$$|B(u,v)| \le C\|u\| \cdot \|v\| \quad \text{for all } u, v \in H, \qquad \text{(boundedness)}$$
$$B(u,u) \ge c\|u\|^2 \quad \text{for all } u \in H. \qquad \text{(coercivity)}$$

Let $F\colon H \to \mathbb{R}$ be a bounded linear mapping. Then there exists a unique element $u \in H$ satisfying

$$B(u,v) = F(v) \quad \text{for all } v \in H$$

and

$$\|u\| \le \frac{1}{c}\|F\|.$$

[†] $B(u+v,w) = B(u,w) + B(v,w)$, $B(au,v) = aB(u,v)$,
$B(u,v+w) = B(u,v) + B(u,w)$, $B(u,av) = aB(u,v)$
for all $u, v, w \in H$ and $a \in \mathbb{R}$.

Proof. We split the proof into several steps.
Step 1. Let $v \in H$ be fixed. Then the mapping

$$T\colon w \mapsto B(v, w),\ H \to \mathbb{R},$$

is bounded and linear. It follows from the Riesz representation theorem that there exists a unique element $a \in H$ with

$$Tw = \langle a, w \rangle \quad \text{for all } w \in H.$$

Let us define the mapping $A\colon H \to H$ by setting

$$Av = a.$$

Then

$$B(v, w) = \langle Av, w \rangle \quad \text{for all } v, w \in H.$$

Step 2. We show that the mapping $A\colon H \to H$ is linear and bounded. Clearly,

$$\begin{aligned} \langle A(c_1 v_1 + c_2 v_2), w\rangle &= B(c_1 v_1 + c_2 v_2, w) \\ &= c_1 B(v_1, w) + c_2 B(v_2, w) \\ &= \langle c_1 A v_1 + c_2 A v_2, w\rangle \end{aligned}$$

for all $w \in H$, so $A(c_1 v_1 + c_2 v_2) = c_1 A v_1 + c_2 A v_2$. Moreover,

$$\begin{aligned} \|Av\|^2 &= \langle Av, Av\rangle \\ &= B(v, Av) \\ &\le C\|v\|\|Av\| \end{aligned}$$

which implies that

$$\|Av\| \le C\|v\|.$$

Step 3. We show that

$$\begin{cases} A \text{ is one-to-one,} \\ \mathrm{Ran}(A) = AH \text{ is closed in } H. \end{cases}$$

We begin by noting that

$$c\|v\|^2 \leq B(v, v) = \langle Av, v\rangle \leq \|Av\|\|v\|$$

and thus

$$\|Av\| \geq c\|v\| \quad \text{for all } v \in H. \tag{5}$$

Especially

$$Av = Aw \Rightarrow A(v - w) = 0 \Rightarrow 0 = \|A(v - w)\| \geq c\|v - w\| \geq 0 \Rightarrow v = w$$

so A is one-to-one.

To see that $\mathrm{Ran}(A)$ is closed, let $y_j = Ax_j \in \mathrm{Ran}(A)$. The goal is to show that $y := \lim_{j\to\infty} y_j \in \mathrm{Ran}(A)$. We observe that

$$\lim_{j,k\to\infty} \|x_j - x_k\| \overset{(5)}{\leq} \lim_{j,k\to\infty} \frac{1}{c}\|y_j - y_k\| = 0,$$

i.e., $(x_j)_{j=1}^{\infty}$ is Cauchy and $x := \lim_{j\to\infty} x_j \in H$ exists by completeness. Moreover,

$$\lim_{j\to\infty} \|Ax_j - Ax\| \leq \lim_{j\to\infty} \|A\|\|x_j - x\| \leq C \lim_{j\to\infty} \|x_j - x\| = 0$$

and therefore

$$y = \lim_{j\to\infty} Ax_j = Ax \in \mathrm{Ran}(A).$$

Step 4. We show that $\overline{\mathrm{Ran}(A)} = H$. We prove this by contradiction: suppose that $\mathrm{Ran}(A) = \overline{\mathrm{Ran}(A)} \neq H$. Then there exists $w \in \mathrm{Ran}(A)^\perp$, $w \neq 0$.[†] This implies that

$$\|w\|^2 \leq \frac{1}{c} B(w, w) = \frac{1}{c} \langle Aw, w \rangle = 0,$$

i.e., $w = 0$. This contradiction shows that $\mathrm{Ran}(A) = H$. Therefore $A \colon H \to H$ is a continuous bijection.

Step 5. Existence of a solution. We use the Riesz representation theorem: since $F \colon H \to \mathbb{R}$ is linear and continuous, there exists $b \in H$ such that

$$F(v) = \langle b, v \rangle \quad \text{for all } v \in H.$$

Define $u := A^{-1} b$. Hence

$$\begin{aligned} Au = b \quad &\Leftrightarrow \quad \langle Au, v \rangle = \langle b, v \rangle \quad \text{for all } v \in H \\ &\Leftrightarrow \quad B(u, v) = F(v) \quad \text{for all } v \in H. \end{aligned}$$

[†] Since $(\mathrm{Ran}(A)^\perp)^\perp = \overline{\mathrm{Ran}(A)} \neq H \Rightarrow (\mathrm{Ran}(A))^\perp \neq \{0\}$.

Step 6. Uniqueness. Suppose that

$$B(u_1, w) = F(w) \quad \text{for all } w \in H,$$
$$B(u_2, w) = F(w) \quad \text{for all } w \in H.$$

Let $u := u_1 - u_2$. By linearity,

$$B(u, w) = 0 \quad \text{for all } w \in H.$$

The coercivity of B implies that

$$\|u\|^2 \leq \frac{1}{c} B(u, u) = 0$$

so that $u = 0$, i.e., $u_1 = u_2$.

Step 7. *A priori bound*. If $B(u, w) = F(w)$ for all $w \in H$, then by setting $w = u$ we obtain

$$\|u\|^2 \leq \frac{1}{c} B(u, u) = \frac{1}{c} F(u) \leq \frac{1}{c} \|F\| \|u\|$$

which immediately yields

$$\|u\| \leq \frac{1}{c} \|F\|.$$

$\square$

Density argument

Lemma

Let X, Y be Banach spaces and let $Z \subset X$ be a dense subspace. If $T : Z \to Y$ is a linear mapping such that

$$\|Tx\|_Y \leq C\|x\|_X, \quad x \in Z, \tag{6}$$

then there exists a unique extension $\widetilde{T} : X \to Y$ with $\widetilde{T}|_Z = T$ and

$$\|\widetilde{T}x\|_Y \leq C\|x\|_X, \quad x \in X. \tag{7}$$

Moreover, if (6) holds with equality, then so does (7).

Proof. Let $x \in X$. Because $Z \subset X$ is dense, there exists a sequence $(z_k)_{k=1}^{\infty} \subset Z$ s.t. $\|z_k - x\|_X \xrightarrow{k\to\infty} 0$. Let $\varepsilon > 0$. Since $(z_k)_{k=1}^{\infty}$ is a Cauchy sequence, there exists $N \in \mathbb{N}$ s.t.

$$m, n \geq N \quad \Rightarrow \quad \|z_m - z_n\|_X < \frac{\varepsilon}{C}.$$

Then there holds

$$\|Tz_m - Tz_n\|_Y = \|T(z_m - z_n)\|_Y \leq C\|z_m - z_n\|_X < \varepsilon,$$

which means that $(Tz_k)_{k=1}^{\infty}$ is a Cauchy sequence in Y. Since Y is complete, there exists $y := \lim_{k\to\infty} Tz_k$. Hence we may define $\widetilde{T} : X \to Y$ by setting $\widetilde{T}(x) = y$.

We begin by showing that $\widetilde{T}$ is well-defined. Let $(z_k)_{k=1}^{\infty}$, $(\widetilde{z}_k)_{k=1}^{\infty}$ be two sequences in Z s.t. $z_k, \widetilde{z}_k \xrightarrow{k\to\infty} x$ in X. Then

$$\|Tz_k - T\widetilde{z}_k\|_Y = \|T(z_k - \widetilde{z}_k)\|_Y \le C\|z_k - \widetilde{z}_k\| \le C\|z_k - x\| + C\|\widetilde{z}_k - x\| \overset{k\to\infty}{\to} 0.$$

Recalling that $\widetilde{T}(x) := \lim_{k\to\infty} Tz_k$, we obtain

$$\|T\widetilde{z}_k - \widetilde{T}(x)\| \le \|T\widetilde{z}_k - Tz_k\| + \|Tz_k - \widetilde{T}(x)\| \overset{k\to\infty}{\to} 0,$$

showing that $\widetilde{T}$ is well-defined.
Next we show that $\widetilde{T}$ is linear. Let $x, \widetilde{x} \in X$ and $a, b \in \mathbb{R}$. Let $Z \ni z_k \xrightarrow{k\to\infty} x$ and $Z \ni \widetilde{z}_k \xrightarrow{k\to\infty} \widetilde{x}$. Now $ax + b\widetilde{x} \in X$ and $Z \ni az_k + b\widetilde{z}_k \to ax + b\widetilde{x}$. Thus

$$\widetilde{T}(ax + b\widetilde{x}) = \lim_{k\to\infty} T(az_k + b\widetilde{z}_k) = a\lim_{k\to\infty} Tz_k + b\lim_{k\to\infty} T\widetilde{z}_k = a\widetilde{T}x + b\widetilde{T}x,$$

since the limit is linear.†
Since the norm is continuous,

$$\|\widetilde{T}x\| = \|\lim_{k\to\infty} Tx_k\| = \lim_{k\to\infty}\|Tx_k\| \le C\lim_{k\to\infty}\|x_k\| = C\|x\|.$$

Finally, $\widetilde{T}|_Z = T$ holds by construction and the uniqueness of the limit $Tz_k \to y$ ensures that there cannot exist another mapping $L\colon X \to Y$ s.t. $L|_Z = T$ and $\|Lx\| \le C\|x\|$. $\square$

†Let $y := \lim_{k\to\infty} Tz_k$ and $\widetilde{y} := \lim_{k\to\infty} T\widetilde{z}_k$.
Then $\|T(az_k + b\widetilde{z}_k) - ay - b\widetilde{y}\| \le a\|Tz_k - y\| + b\|T\widetilde{z}_k - \widetilde{y}\| \to 0$.
Hence $\lim_{k\to\infty} T(az_k + b\widetilde{z}_k) = a\lim_{k\to\infty} Tz_k + b\lim_{k\to\infty} T\widetilde{z}_k$.

Multi-index notation

A vector of the form $\boldsymbol{\alpha} := (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d$ is called a *multi-index*. We denote the j^{th} component of multi-index $\boldsymbol{\alpha}$ by α_j.

The *order* (or *modulus*) of a multi-index is defined as

$$|\boldsymbol{\alpha}| := \alpha_1 + \cdots + \alpha_d.$$

Let $\boldsymbol{x} := (x_j)_{j=1}^d \in \mathbb{R}^d$. We define the monomial notation

$$\boldsymbol{x}^{\boldsymbol{\alpha}} := \prod_{j=1}^d x_j^{\alpha_j},$$

where $0^0 := 1$, and the corresponding formula for partial derivatives

$$\partial^{\boldsymbol{\alpha}} := \partial_{\boldsymbol{x}}^{\boldsymbol{\alpha}} := \prod_{j=1}^d \frac{\partial^{\alpha_j}}{\partial x_j^{\alpha_j}}.$$

Other often used multi-index notations include $\binom{\boldsymbol{\alpha}}{\boldsymbol{\beta}} := \prod_{j=1}^d \binom{\alpha_j}{\beta_j}$, $\boldsymbol{\alpha}! := \alpha_1! \cdots \alpha_d!$ (but $|\boldsymbol{\alpha}|! := (\alpha_1 + \cdots + \alpha_d)!$), etc.

Some function spaces

Let $D \subset \mathbb{R}^d$ be a nonempty open set. Let us recall the following function spaces.

Definition

$C(D) := \{u \colon D \to \mathbb{R} \mid u \text{ is continuous}\}$,

$C^k(D) := \{u \colon D \to \mathbb{R} \mid \exists \partial^{\boldsymbol{\alpha}} u \text{ is continuous for all } |\boldsymbol{\alpha}| \le k,\ \boldsymbol{\alpha} \in \mathbb{N}_0^d\}$,

$$C^\infty(D) := \{u \colon D \to \mathbb{R} \mid \exists \partial^{\boldsymbol{\alpha}} u \text{ is continuous for all } \boldsymbol{\alpha} \in \mathbb{N}_0^d\} = \bigcap_{k=0}^{\infty} C^k(D),$$

$C_0^k(D) := \{u \in C^k(D) \mid \mathrm{supp}(u) \subset D \text{ is a compact set}\}$,

$C_0^\infty(D) := \{u \in C^\infty(D) \mid \mathrm{supp}(u) \subset D \text{ is a compact set}\}$,

where $\mathrm{supp}(u) := \overline{\{\boldsymbol{x} \in D \mid u(\boldsymbol{x}) \neq 0\}}$,

$$L^1(D) := \{u \colon D \to \mathbb{R} \mid u \text{ is measurable, } \|u\|_{L^1(D)} := \int_D |u(\boldsymbol{x})| \,\mathrm{d}\boldsymbol{x} < \infty\}.$$

Remark. Recall that in the Euclidean space $\mathbb{R}^d$, a set is compact iff it is closed and bounded. This is the *Heine–Borel theorem*.

Let $D \subset \mathbb{R}^d$ be open.

Definition

$$L^1_{\mathrm{loc}}(D) := \{u \colon D \to \mathbb{R} \mid u \in L^1(K) \text{ for all compact } K \subset D\}$$

Example

Let $u \in C^1(D)$. Then integration by parts yields

$$\int_D u(\boldsymbol{x}) \partial_{x_i} \varphi(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} = -\int_D \partial_{x_i} u(\boldsymbol{x}) \varphi(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} \quad \text{for all } \varphi \in C_0^\infty(D). \tag{8}$$

If $u \in C^k(D)$ and $\boldsymbol{\alpha} \in \mathbb{N}_0^d$ is a multi-index with order $|\boldsymbol{\alpha}| := \nu_1 + \cdots + \nu_d = k$, then we obtain from repeated application of (8) that

$$\int_D u(\boldsymbol{x}) \partial^{\boldsymbol{\alpha}} \varphi(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} = (-1)^{|\boldsymbol{\alpha}|} \int_D \partial^{\boldsymbol{\alpha}} u(\boldsymbol{x}) \varphi(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} \quad \text{for all } \varphi \in C_0^\infty(D).$$

The so-called *weak derivative* is a generalization of the classical derivative.

Definition (Weak derivative)

Let $u, w \in L^1_{\mathrm{loc}}(D)$. We call w the *weak ∂_{x_i} derivative of u* and denote $w = \partial_{x_i} u$ if

$$\int_D w(\boldsymbol{x})\varphi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = -\int_D u(\boldsymbol{x})\partial_{x_i}\varphi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \quad \text{for all } \varphi \in C_0^\infty(D).$$

Moreover, we call w the *weak $\partial^{\boldsymbol{\alpha}}$ derivative of u* and denote $w = \partial^{\boldsymbol{\alpha}} u$ if

$$\int_D w(\boldsymbol{x})\varphi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = (-1)^{|\boldsymbol{\alpha}|}\int_D u(\boldsymbol{x})\partial^{\boldsymbol{\alpha}}\varphi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \quad \text{for all } \varphi \in C_0^\infty(D).$$

This definition ensures that the integration by parts formula is valid if the weak derivative exists.

Remark. This definition generalizes the classical derivative: if $u \in C^1(D)$, then the weak derivative coincides with the classical one.

Weak derivative

Example

Let $d = 1$, $D = (0,2)$, and

$$u(x) = \begin{cases} x & \text{if } 0 < x \leq 1, \\ 1 & \text{if } 1 \leq x < 2. \end{cases}$$

Define

$$v(x) = \begin{cases} 1 & \text{if } 0 < x \leq 1, \\ 0 & \text{if } 1 \leq x < 2. \end{cases}$$

We claim $u' = v$ in the weak sense, i.e., $\int_0^2 u(x)\varphi'(x)\,\mathrm{d}x = -\int_0^2 v(x)\varphi(x)\,\mathrm{d}x$ for all $\varphi \in C_0^\infty(D)$. Let $\varphi \in C_0^\infty(D)$ be arbitrary. Then

$$\int_0^2 u(x)\varphi'(x)\,\mathrm{d}x = \int_0^1 x\varphi'(x)\,\mathrm{d}x + \int_1^2 \varphi'(x)\,\mathrm{d}x$$

$$= \underbrace{\left[x\varphi(x)\right]\Big|_{x=0}^{x=1}}_{=\cancel{\varphi(1)}-0} - \int_0^1 \varphi(x)\,\mathrm{d}x + \underbrace{\varphi(2)}_{=0} - \cancel{\varphi(1)} = -\int_0^1 \varphi(x)\,\mathrm{d}x = -\int_0^2 v(x)\varphi(x)\,\mathrm{d}x$$

as desired.

Sobolev spaces

Sobolev spaces

Definition

The *Sobolev space* of order k based on $L^2(D)$ is defined by

$$H^k(D) := \{u \in L^2(D) \mid \partial^{\boldsymbol{\alpha}} u \in L^2(D) \text{ for all } |\boldsymbol{\alpha}| \leq k\},$$

and we equip this space with the norm

$$\|u\|_{H^k(D)} = \left(\sum_{|\boldsymbol{\alpha}| \leq k} \int_D |\partial^{\boldsymbol{\alpha}} u(\boldsymbol{x})|^2 \, \mathrm{d}\boldsymbol{x} \right)^{1/2},$$

induced by the inner product

$$\langle u, v \rangle_{H^k(D)} = \sum_{|\boldsymbol{\alpha}| \leq k} \int_D \partial^{\boldsymbol{\alpha}} u(\boldsymbol{x}) \partial^{\boldsymbol{\alpha}} v(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}.$$

Moreover, we define

$$H_0^k(D) := \mathrm{cl}_{H^k(D)}(C_0^\infty(D)),$$

i.e., $H_0^k(D)$ is the closure of $C_0^\infty(D)$ in the topology of $H^k(D)$.

Proposition

- $\partial^{\boldsymbol{\alpha}}\colon H^k(D) \to H^{k-|\boldsymbol{\alpha}|}$, $k \geq |\boldsymbol{\alpha}|$, *is bounded.*
- $\partial^{\boldsymbol{\alpha}}(\partial^{\boldsymbol{\beta}} u) = \partial^{\boldsymbol{\beta}}(\partial^{\boldsymbol{\alpha}} u) = \partial^{\boldsymbol{\alpha}+\boldsymbol{\beta}} u$, $u \in H^{|\boldsymbol{\alpha}|+|\boldsymbol{\beta}|}(D)$, *where* $\boldsymbol{\alpha} + \boldsymbol{\beta} := (\alpha_1 + \beta_1, \ldots, \alpha_d + \beta_d)$.

Proposition

$H^k(D)$ is a Hilbert space for all $k \in \mathbb{N}$.

Proof. Let $(u_j)_{j=1}^{\infty}$ be a Cauchy sequence in $H^k(D)$. Then for all $|\boldsymbol{\alpha}| \le k$

$$\|D^{\boldsymbol{\alpha}} u_m - D^{\boldsymbol{\alpha}} u_n\|_{L^2(D)} \le \|u_m - u_n\|_{H^k} \xrightarrow{m,n\to\infty} 0,$$

so $(D^{\boldsymbol{\alpha}} u_j)_{j=1}^{\infty}$ is a Cauchy sequence in $L^2(D)$. Since $L^2(D)$ is complete, there exists $f^{\boldsymbol{\alpha}} \in L^2(D)$ such that $\|f_{\boldsymbol{\alpha}} - D^{\boldsymbol{\alpha}} u_j\|_{L^2(D)} \xrightarrow{j\to\infty} 0$.

Esp. $u_j \xrightarrow{j\to\infty} f^{\mathbf{0}} := u$ in $L^2(D)$.

We show that $D^{\boldsymbol{\alpha}} u \in L^2(D)$ for all $|\boldsymbol{\alpha}| \le k$, i.e., $u \in H^k(D)$. For $\varphi \in C_0^\infty(D)$,

$$\begin{aligned}
\int_D u(\boldsymbol{x}) \partial^{\boldsymbol{\alpha}} \varphi(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} &= \lim_{j\to\infty} \int_D u_j(\boldsymbol{x}) \partial^{\boldsymbol{\alpha}} \varphi(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \\
&= \lim_{j\to\infty} \int_D (-1)^{|\boldsymbol{\alpha}|} \partial^{\boldsymbol{\alpha}} u_j(\boldsymbol{x}) \cdot \varphi(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \\
&= \int_D (-1)^{|\boldsymbol{\alpha}|} f^{\boldsymbol{\alpha}}(\boldsymbol{x}) \cdot \varphi(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}
\end{aligned}$$

so $\partial^{\boldsymbol{\alpha}} u = f^{\boldsymbol{\alpha}} \in L^2(D)$, $|\boldsymbol{\alpha}| \le k$. Thus $u \in H^k(D)$.

Finally,

$$\|u_j - u\|^2_{H^k(D)} = \sum_{|\boldsymbol{\alpha}|\leq k} \|\partial^{\boldsymbol{\alpha}} u_j - f^{\boldsymbol{\alpha}}\|^2_{L^2(D)} \xrightarrow{j\to\infty} 0$$

which means that

$$\lim_{j\to\infty} u_j = u \quad \text{in } H^k(D). \quad \square$$

The case when D is a polygon (2D) or a polyhedron (3D) will be of special interest to us. In these cases, the boundary ∂D is not smooth, which needs to be accounted for by our theory. However, it turns out that working with Lipschitz domains is sufficient for our purposes. To this end, we recall the following.

Definition

Let $D \subset \mathbb{R}^d$ be a bounded, open set. A function $u\colon D \to \mathbb{R}$ is *Lipschitz continuous* if there exists $L > 0$ such that

$$|u(\boldsymbol{x}) - u(\boldsymbol{y})| \leq L|\boldsymbol{x} - \boldsymbol{y}|, \quad \boldsymbol{x}, \boldsymbol{y} \in D.$$

Theorem (Rademacher's theorem)

If $D \subset \mathbb{R}^d$ is an open subset and $f\colon D \to \mathbb{R}$ is Lipschitz continuous, then f is differentiable almost everywhere in D.

A Lipschitz hypograph $D \subset \mathbb{R}^d$ is a domain of the form

$$D = \{\boldsymbol{x} \in \mathbb{R}^d \mid x_d > \zeta(\boldsymbol{x}'),\ \boldsymbol{x}' := (x_1, \dots, x_{d-1}) \in \mathbb{R}^d\}$$

where $\zeta\colon \mathbb{R}^{d-1} \to \mathbb{R}$ is a Lipschitz function.

Definition (bounded Lipschitz domain)

An open, bounded set is a *Lipschitz* domain if its boundary ∂D is compact and if there exist $\{W_j\}_{j=1}^N$ and $\{D_j\}_{j=1}^N$ with the following properties:

(i) $\{W_j\}_{j=1}^N$ is a finite open cover of ∂D, i.e., each $W_j \subset \mathbb{R}^d$ is an open set and $\partial D \subset \bigcup_{j=1}^N W_j$.

(ii) Each D_j can be transformed into a Lipschitz hypograph by a rotation plus a translation.

(iii) $W_j \cap D = W_j \cap D_j$ for all $j \in \{1, \ldots, N\}$.

The class of Lipschitz domains is broad enough to cover most cases that arise in applications of partial differential equations. For example, any polygon in $\mathbb{R}^2$ or convex polyhedron in $\mathbb{R}^3$ is a Lipschitz domain. If $\kappa \colon \mathbb{R}^d \to \mathbb{R}^d$ is a C^1 diffeomorphism and D is a Lipschitz domain, then the set $\kappa(D)$ is again a Lipschitz domain.

Note that the outer normal vector is defined a.e. at the boundary and it is a.e. continuous.

Trace theorem on Lipschitz domains

Theorem

Let D be a bounded Lipschitz domain and let $\gamma\colon C^\infty(\overline{D}) \to C^\infty(\partial D)$ be the trace operator $\gamma u = u|_{\partial D}$. Then the trace operator has a unique extension to a bounded linear operator $\gamma\colon H^1(D) \to L^2(\partial D)$.

The significance of the trace theorem is that the boundary values of Sobolev functions belonging to $H^1(D)$ are well-defined in an unambiguous way.

Trace zero functions in $H_0^1(D)$

Theorem

Let $D \subset \mathbb{R}^d$ be a bounded Lipschitz domain, $u \in H^1(D)$, and $\gamma\colon H^1(D) \to L^2(\partial D)$ the trace operator. Then

$$u \in H_0^1(D) \quad \Leftrightarrow \quad \gamma u = 0\colon \partial D \to \mathbb{R}.$$

Proof. "$\Rightarrow$" follows from previous discussion. "$\Leftarrow$" is more difficult (see, e.g., L.C. Evans "Partial Differential Equations" Section 5.5 for details). □

This implies in particular the characterization $H_0^1(D) = \mathrm{Ker}(\gamma)$, meaning that elements in $H_0^1(D)$ are precisely those elements in $H^1(D)$ with zero trace.

Definition

Let $\|\cdot\|$ and $\|\cdot\|_*$ be two norms in a normed space X. The norms are called *equivalent* if there exist constants $c_1, c_2 > 0$ such that

$$c_1\|x\|_* \leq \|x\| \leq c_2\|x\|_* \quad \text{for all } x \in X.$$

The significance behind this notion lies in the fact that equivalent norms induce the same topology on X. That is, $\|\cdot\|$ and $\|\cdot\|_*$ induce *exactly the same* convergent sequences in X.

For our purposes, let $A\colon X \to Y$ be a mapping between two normed spaces. Suppose that $c_X\|\cdot\|_{X,*} \leq \|\cdot\|_X \leq C_X\|\cdot\|_{X,*}$ and $c_Y\|\cdot\|_{Y,*} \leq \|\cdot\|_Y \leq C_Y\|\cdot\|_{Y,*}$ for $c_X, C_X, c_Y, C_Y > 0$. If

$$\|A(x)\|_Y \leq K\|x\|_X \quad \text{for some } x \in X,$$

then

$$\|A(x)\|_{Y,*} \leq \frac{C_X K}{c_Y}\|x\|_{X,*} \quad \text{for some } x \in X.$$

We can change between equivalent norms rather liberally since any results about boundedness, stability, etc., proved using one norm remain true for equivalent norms up to a trivial scaling of the (typically generic) coefficient.

Proposition (Poincaré inequality)

Let $D \subset \mathbb{R}^d$ be a bounded domain. Then there exists $C > 0$ (independently of u) such that

$$\|u\|_{L^2(D)} \leq C\|\nabla u\|_{L^2(D)} \quad \textit{for all } u \in H_0^1(D).$$

Proof. Let $\varphi \in C_0^\infty(D)$. Since we assumed D is bounded, we may assume $D \subset [-a, a]^d$ for suitably large $a > 0$. Extending φ by zero outside of D, we obtain

$$\begin{aligned}\varphi(x_1, x_2, \dots, x_d) &= \varphi(x_1, x_2, \dots, x_d) - \varphi(-a, x_2, \dots, x_d) \\ &= \int_{-a}^{x_1} \frac{\partial \varphi}{\partial x_1}(t_1, x_2, \dots, x_d)\,\mathrm{d}t_1.\end{aligned}$$

By the Cauchy–Schwarz inequality,

$$\begin{aligned}|\varphi(x_1, x_2, \dots, x_d)|^2 &\leq 2a \int_{-a}^{a} \left|\frac{\partial \varphi}{\partial x_1}(t_1, x_2, \dots, x_d)\right|^2 \mathrm{d}t_1 \\ \Rightarrow \quad \int_{-a}^{a} |\varphi(x_1, x_2, \dots, x_d)|^2\,\mathrm{d}x_1 &\leq 4a^2 \int_{-a}^{a} \left|\frac{\partial \varphi}{\partial x_1}(t_1, x_2, \dots, x_d)\right|^2 \mathrm{d}t_1.\end{aligned}$$

Repeated integrations w.r.t. $x_2, x_3, \dots, x_d$ over $[-a, a]$ together with the density of $C_0^\infty(D)$ in $H_0^1(D)$ prove the assertion. □

An equivalent norm in $H_0^1(D)$

For all $u \in H_0^1(D)$, the norm

$$\|u\|_{H_0^1(D)} := \|\nabla u\|_{L^2(D)} := \left(\int_D \|\nabla u(\boldsymbol{x})\|^2 \,\mathrm{d}\boldsymbol{x} \right)^{1/2}$$

is equivalent to $\|u\|_{H^1(D)} := (\|u\|^2_{L^2(D)} + \|\nabla u\|^2_{L^2(D)})^{1/2}$.

This can be seen as an immediate consequence of the Poincaré inequality:

$$\|u\|^2_{H_0^1(D)} \leq \|u\|^2_{L^2(D)} + \|\nabla u\|^2_{L^2(D)} \leq (1 + C^2)\|u\|^2_{H_0^1(D)}.$$

Sobolev inequality

We mention the following result.

Theorem

Let $D \subset \mathbb{R}^d$ be a bounded Lipschitz domain and $k > d/2$. Then

$$H^k(D) \subset C_B(D) := \{v \in C(D) \mid v \text{ is bounded}\}$$

and there is a constant $C > 0$ s.t.

$$\|u\|_{C_B(D)} := \sup_{\boldsymbol{x} \in D} |u(\boldsymbol{x})| \le C\|u\|_{H^1(D)} \quad \text{for all } u \in H^1(D).$$

Proof. Cf., e.g., Adams (1975) or Adams and Fournier (2003). □

- If $d = 1$, then $u \in H^1(D)$ has a continuous representative.
- If $d \in \{2, 3\}$, then $u \in H^2(D)$ has a continuous representative.

Elliptic PDEs

Let $D \subset \mathbb{R}^d$ be an open and bounded Lipschitz domain. We consider the problem

$$\begin{cases} -\nabla \cdot (a(\boldsymbol{x})\nabla u(\boldsymbol{x})) = f(\boldsymbol{x}), & \boldsymbol{x} \in D, \\ u|_{\partial D} = 0, \end{cases} \tag{9}$$

where $f: D \to \mathbb{R}$ is the *source* and $a: D \to \mathbb{R}$ is the *diffusion coefficient*.

Uniform ellipticity assumption: There exist constants $a_{\max}, a_{\min} > 0$ such that

$$0 < a_{\min} \le a(\boldsymbol{x}) \le a_{\max} < \infty \quad \text{for all } \boldsymbol{x} \in D.$$

Definition

Let $a \in C^1(D)$ and $f \in C(D)$. Then $u \in C^2(D)$ is the *classical solution* to (9) if (9) holds for all $\boldsymbol{x} \in D$ and $u(\boldsymbol{y}) = 0$, $\boldsymbol{y} \in \partial D$.

The requirement that f is continuous is usually much too restrictive for practical applications.

Definition (Strong solution)

Let $a\colon \overline{D} \to \mathbb{R}$ be Lipschitz and $f \in L^2(D)$. We call $u \in H^2(D) \cap H_0^1(D)$ a *strong solution* to (9) if

$$-\nabla \cdot (a(\boldsymbol{x})\nabla u(\boldsymbol{x})) = f(\boldsymbol{x}) \quad \text{for a.e. } \boldsymbol{x} \in D,$$

where the derivatives are the weak derivatives.

Note that we also have the following.

Lemma

Let $D \subset \mathbb{R}^d$ be a bounded Lipschitz domain. Then for $u, v \in H^1(D)$,

$$\int_D \partial_{x_j} u(\boldsymbol{x}) v(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} = -\int_D u(\boldsymbol{x}) \partial_{x_j} w(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} + \int_{\partial D} n_j u|_{\partial D} v|_{\partial D} \,\mathrm{d}S,$$

where $\cdot|_{\partial D}\colon H^1(D) \to L^2(\partial D)$ is the trace operator.

Proof. The formula holds for $u, v \in C^\infty(\overline{D})$. The assertion follows by exploiting the density of $C^\infty(\overline{D})$ in $H^1(D)$. □

If u is a strong solution to the PDE (9), then for all $v \in C_0^\infty(D)$

$$
\begin{aligned}
\langle -\nabla \cdot (a \nabla u), v \rangle_{L^2(D)} &= \int_D -\nabla \cdot (a(\boldsymbol{x}) \nabla u(\boldsymbol{x})) v(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \\
&\overset{\dagger}{=} \int_D a(\boldsymbol{x}) \nabla u(\boldsymbol{x}) \cdot \nabla v(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} + \int_{\partial D} \underbrace{v(\boldsymbol{x})}_{=0} (a \nabla u(\boldsymbol{x}) \cdot \boldsymbol{n}(\boldsymbol{x})) \, \mathrm{d}S \\
&= \int_D a(\boldsymbol{x}) \nabla u(\boldsymbol{x}) \cdot \nabla v(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} =: B(u, v).
\end{aligned}
$$

Define also

$$
F(v) := \int_D f(\boldsymbol{x}) v(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}.
$$

This leads us to consider the variational formulation

$$
B(u, v) = F(v) \quad \text{for all } v \in C_0^\infty(D).
$$

$^\dagger \nabla \cdot (v(a \nabla u)) = a \nabla v \cdot \nabla u + v \nabla \cdot (a \nabla u)$ + divergence theorem

The previous discussion motivates us to introduce the so-called *weak solution* to (9).

Definition

Let $a \in L^\infty(D)$ and $f \in L^2(D)$. Then $u \in H_0^1(D)$ is called a weak solution to (9) if

$$B(u,v) = F(v) \quad \text{for all } v \in H_0^1(D), \tag{10}$$

where

$$B(u,v) = \int_D a(\boldsymbol{x}) \nabla u(\boldsymbol{x}) \cdot \nabla v(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}$$

and

$$F(v) = \int_D f(\boldsymbol{x}) v(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}.$$

Remark. It is sufficient to enforce (10) for all $v \in C_0^\infty(D)$. Moreover, the definition can be extended for arbitrary $F \in (H_0^1(D))' =: H^{-1}(D)$.

Our variational problem is

$$B(u,v) = F(v) \quad \text{for all } v \in H_0^1(D), \tag{11}$$

where $B(u,v) = \int_D a(\boldsymbol{x})\nabla u(\boldsymbol{x}) \cdot \nabla v(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ and $F(v) = \int_D f(\boldsymbol{x})v(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$.

Let us use the norm $\|v\|_{H_0^1(D)} := \|\nabla v\|_{L^2(D)}$, which is equivalent to the usual Sobolev norm by Poincaré's inequality.

Provided that we have uniform ellipticity, i.e.,
$0 < a_{\min} \le a(\boldsymbol{x}) \le a_{\max} < \infty$ for all $\boldsymbol{x} \in D$, then

$$B(u,v) = \int_D a(\boldsymbol{x})\nabla u(\boldsymbol{x}) \cdot \nabla v(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \le a_{\max}\|u\|_{H_0^1(D)}\|v\|_{H_0^1(D)}$$

for all $u, v \in H_0^1(D)$ and

$$B(u,u) = \int_D a(\boldsymbol{x})\nabla u(\boldsymbol{x}) \cdot \nabla u(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \ge a_{\min}\|u\|^2_{H_0^1(D)} \quad \text{for all } u \in H_0^1(D).$$

$\therefore$ By the Lax–Milgram lemma, there exists a unique solution $u \in H_0^1(D)$ to (11) s.t. $\|u\|_{H_0^1(D)} \le \dfrac{\|F\|_{H^{-1}(D)}}{a_{\min}}$.

When does the weak solution coincide with the strong solution?

If $f \in L^2(D)$, the diffusion coefficient a is smooth enough (e.g., Lipschitz), and the boundary ∂D is "nice enough" (e.g., a convex polyhedron), then $u \in H^2(D) \cap H_0^1(D)$ and the weak solution coincides with the strong solution. These considerations belong to the purview of *elliptic regularity theory*.