

Numerische Simulation der Wärmeleitung mit finiten Elementen und künstlichen neuronalen Netzwerken

Bachelorarbeit

ZUR ERLANGUNG DES AKADEMISCHEN GRADES

Bachelor of Science

EINGEREICHT AN DER
MATHEMATISCHE-NATURWISSENSCHAFTLICH-TECHNISCHEM FAKULTÄT
DER UNIVERSITÄT AUGSBURG

VON

Daniel Larin

AUGSBURG, 24. APRIL 2022



**Universität
Augsburg
University**

Inhaltsverzeichnis

1 Einleitung	1
2 Die Wärmeleitungsgleichung	2
2.1 Problemstellung	2
2.2 Physikalische Motivation	4
2.3 Funktionalanalytische Vorbereitungen	5
2.4 Abstraktes Evolutionsproblem	9
2.5 Schwache Formulierung	10
2.6 Essentielle Randbedingungen	11
2.7 Natürliche Randbedingungen	13
2.8 Gemischte Randbedingungen	15
3 Finite-Elemente-Methode	17
3.1 Galerkin-Verfahren	18
3.2 Triangulierung	19
3.3 Simpliziale Lagrange-Elemente	23
3.4 Interpolation	26
3.5 Ortsdiskretisierung	28
3.6 Zeitdiskretisierung	30
3.6.1 Implizites Euler-Verfahren	32
3.6.2 Explizites Euler-Verfahren	34
4 Physics-informed neuronale Netzwerke	36
4.1 Künstliche neuronale Netze	36
4.2 Kostenfunktion	38
4.3 Optimierungsverfahren	42
4.4 Implementierung	43
4.5 Fehlerdiskussion	44
5 Numerische Experimente	46
5.1 1D-Testproblem	47
5.2 2D-Testproblem	60
6 Fazit	69
Literatur	70

1 Einleitung

Ein grundlegendes, physikalisches Phänomen aus der Thermodynamik, mit dem wir alle im Alltag in Kontakt kommen, ist die Wärmeleitung. Unter der Wärmeleitung verstehen wir dabei den physikalischen Wärmefluss in oder zwischen einem Feststoff, Gas oder Fluid aufgrund eines Temperaturunterschiedes. Mathematisch wird dieses Phänomen durch eine partielle Differentialgleichung, der Wärmeleitungsgleichung beschrieben. Sind wir nun an der Temperaturentwicklung eines Mediums, z. B. eines Heizrohrs oder den Kühlrippen eines Kühlers interessiert, so kann dies mit der Lösung jener Differentialgleichung unter gegebenen Anfangs- und Randbedingungen beschrieben werden.

In dieser Arbeit diskutieren wir numerische Verfahren zur Lösung jener Gleichung, d. h. der näherungsweisen Berechnung der Lösung durch Approximationsalgorithmen mit Hilfe von Computern. Zu Beginn führen wir in Kapitel 2 die Wärmeleitungsgleichung im Detail ein und treffen einige Vorbereitungen für das approximative Lösen in den nachfolgenden Abschnitten.

Für die näherungsweise Lösung der Wärmeleitungsgleichung ziehen wir zwei, grundlegend verschiedene Methoden heran. Auf der einen Seite diskutieren wir ein klassisches Verfahren aus der numerischen Mathematik zur Lösung von partiellen Differentialgleichungen, die Finite-Elemente-Methode in Kapitel 3. Auf der anderen Seite, motiviert durch die jüngsten Erfolge von künstlichen neuronalen Netzwerken in diversen Anwendungsgebieten, wie z. B. der Mustererkennung oder Simulation komplexer Systeme, untersuchen wir in Kapitel 4 die physics-informed neuronalen Netzwerke.

Abschließend vergleichen wir in Kapitel 5 die Performance, d. h. die Güte der Näherungen und die resultierte Rechenzeit beider Verfahren anhand zweier Testprobleme. Dabei sind die Algorithmen und die anschließende Auswertung in der Programmiersprache Python implementiert.

2 Die Wärmeleitungsgleichung

In diesem Kapitel betrachten wir die Wärmeleitungsgleichung kombiniert mit passenden Anfangs- und Randbedingungen. Hierfür orientieren wir uns in den Abschnitten 2.1 und 2.2 an [Eva10, Abschnitt 1.1 und 2.3] und [EG04, Abschnitt 3.1.1]. Um diese Gleichungen hinreichend diskutieren zu können, werden einige Konzepte aus der Funktionalanalysis herangezogen. Dabei sind die Abschnitte 2.3 und 2.4 an [Eva10, Abschnitt 5.2] und [EG04, Abschnitt 6.1.1 und 6.1.2] angelehnt.

2.1 Problemstellung

Sei u eine reellwertige Funktion mehrerer reeller Variablen. Eine partielle Differentialgleichung (engl.: partial differential equation, PDE) k -ter Ordnung ist eine Gleichung der Form

$$F(x, (\partial^\alpha u(x))_{|\alpha| \leq k}) = 0. \quad (2.1)$$

Diesbezüglich verwenden wir die Multiindex-Schreibweise, d. h. für einen Multiindex $\alpha = (\alpha_1, \dots, \alpha_n) \in \mathbb{N}_0^n$ mit $n \in \mathbb{N}$ ist $|\alpha| := \alpha_1 + \dots + \alpha_n$ und $\partial^\alpha := \partial_1^{\alpha_1} \dots \partial_n^{\alpha_n}$, wobei $\partial_i^{\alpha_i} := \frac{\partial^{\alpha_i}}{\partial x_i^{\alpha_i}}$ die partielle Ableitung nach x_i von der Ordnung α_i ist. Im Gegensatz zu gewöhnlichen Differentialgleichungen (engl.: ordinary differential equation, ODE) gibt es keine allgemeine Lösungstheorie oder -methoden für Gleichungen der Form (2.1). Die Theorie partieller Differentialgleichungen besteht vielmehr aus sehr vielen, methodisch durchaus verschiedenen Herangehensweisen an spezielle Klassen von PDEs. Dabei werden diese meist einem der Grundtypen *elliptisch*, *parabolisch* und *hyperbolisch* zugeordnet. Jedoch ist diese Klassifikation nicht erschöpfend, d. h. es kann nicht jede Gleichung einem dieser Typen zugeordnet werden. Im Rahmen dieser Arbeit behandeln wir die *d-dimensionale Wärmeleitungsgleichung*

$$\partial_t u - \Delta u = f \quad \text{bzw.} \quad \partial_t u - \Delta u = 0, \quad (2.2)$$

die ein „Paradebeispiel“ der parabolischen PDEs darstellt. Außerdem ist die PDE (2.2) linear und von zweiter Ordnung. Die gesuchte Funktion u ist hier eine Funktion vom Ort x auf einer Teilmenge $\Omega \subset \mathbb{R}^d$ mit $d \in \mathbb{N}$ und der Zeit t im Intervall $(0, T]$ mit $T > 0$. Beachte, dass der Laplaceoperator $\Delta := \sum_{i=1}^d \partial_i^2$ hier nur auf die „räumlichen Koordinaten“ (x_1, \dots, x_d) wirkt, so dass wir genauer eigentlich Δ_x schreiben sollten.

Der Zustand des untersuchten Systems wird außerdem durch äußere Einflüsse am *parabolischen Rand* $\Sigma_T := \overline{\Omega}_T \setminus \Omega_T$ beeinflusst, während $\Omega_T := \Omega \times (0, T]$ das *parabolische Innere des parabolischen Zylinders* $\overline{\Omega}_T = \overline{\Omega} \times [0, T]$ über Ω bezeichnet. Eine Darstellung dieser Mengen ist in Abbildung 1a zu finden. Die Randbedingungen auf dem parabolischen Rand setzen sich diesbezüglich aus einer Anfangsbedingung

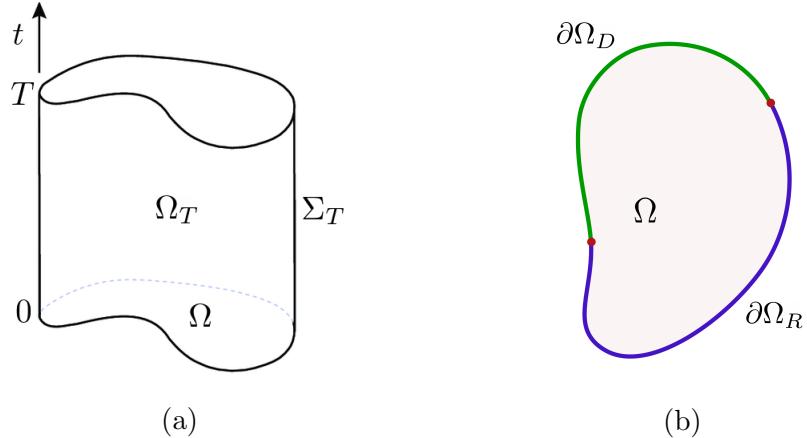


Abbildung 1: (a) Veranschaulichung des Raum-Zeit-Zylinders Ω_T und ihrem Rand Σ_T über das Ortsgebiet Ω , in Anlehnung an [Baz+10, Abbildung 1]. (b) Partitionierung des Randes $\partial\Omega$ in disjunkte Teilgebiete $\partial\Omega_D$ und $\partial\Omega_R$, in Anlehnung an [Wik22].

auf $\Omega \times \{0\}$ und einer Randbedingung auf $\Gamma_T := \partial\Omega \times (0, T]$ zusammen, die das Verhalten von u am Rand festlegen. Die Anfangs- und Randbedingungen sind hier von der Form

$$\begin{aligned} u(\cdot, 0) &= u_0 \quad \text{in } \Omega, \\ \mathcal{B}(u) &= g \quad \text{auf } \Gamma_T, \end{aligned}$$

für gegebene Funktionen $u_0 = u_0(x)$ und $g = g(x, t)$. Wir unterscheiden in folgende Typen von Randbedingungen:

- **Dirichlet-Randbedingung**, definiert durch den Randoperator

$$\mathcal{B}(u) := u \quad \text{auf } \Gamma_T.$$

Im Spezialfall $g = 0$ sprechen wir auch von einer *homogenen* Randbedingung, wohingegen wir im allgemeinen Fall $g \neq 0$ von einer *inhomogenen* Randbedingung sprechen.

- **Neumann-Randbedingung**, definiert durch den Randoperator

$$\mathcal{B}(u) := \partial_\nu u \quad \text{auf } \Gamma_T,$$

wobei $\partial_\nu u := \nu \cdot \nabla u$ die Ableitung von u in Richtung der äußeren Normalen ν an $\partial\Omega$ bezeichnet. In diesem Zusammenhang wirkt der Gradient $\nabla := (\partial_1, \dots, \partial_d)^\top$ nur auf die räumlichen Koordinaten, so dass wir genauer eigentlich ∇_x schreiben sollten.

- **Robin-Randbedingung**, definiert durch den Randoperator

$$\mathcal{B}(u) := \partial_\nu u + \rho u \quad \text{auf } \Gamma_T$$

mit einer gegebenen Funktion $\rho = \rho(x, t)$. Ferner erhalten wir im Grenzfall $\rho \rightarrow \infty$ bzw. $\rho \rightarrow 0$ wieder die Dirichlet- bzw. Neumann-Randbedingung.

Des Weiteren können wir den Fall der *gemischten* Randbedingungen anschauen. In diesem Fall partitionieren wir den Rand $\partial\Omega$ in $n \in \mathbb{N}$ disjunkte Teilgebiete $\partial\Omega_1, \dots, \partial\Omega_n$, so dass $\partial\Omega = \partial\Omega_1 \cup \dots \cup \partial\Omega_n$ gilt; vgl. Abbildung 1b für $n = 2$. Anschließend soll in jedem der Teilgebiete genau eine der oben genannten Randbedingungen gelten.

2.2 Physikalische Motivation

Im vorherigen Abschnitt haben wir die Wärmeleitungsgleichung (2.2) kennengelernt. Es verbleibt noch die Herleitung dieser Gleichung, sowie Ihre physikalische Bedeutung zu zeigen.

Betrachte ein homogenes Medium, das durch eine Punktmenge $A \subset \mathbb{R}^d$ beschrieben wird. Ziel ist es eine Gesetzmäßigkeit für die Temperatur $u = u(x, t)$ in jedem Raumpunkt des Mediums $x \in A$ zur Zeit $t > 0$ zu konstruieren. Dann ist die Änderung der Wärmemenge in A einerseits gegeben durch

$$\frac{d}{dt} \int_A u \, dx = \int_A \partial_t u \, dx. \quad (2.3)$$

Andererseits wird die zeitliche Änderung der Wärmemenge über die Grenzen des Mediums A durch die Wärmestromdichte $q := -\lambda \nabla u$ beschrieben. Der physikalische Parameter λ ist hierbei die Temperaturleitfähigkeit des Mediums. Diesen Zusammenhang nennt man das *Fouriersche Gesetz* der Wärmeleitung. Es folgt

$$-\int_{\partial A} q \cdot \nu \, dS = -\int_A \nabla \cdot q \, dx = \int_A \lambda \Delta u \, dx, \quad (2.4)$$

wobei ν die äußere Normale an ∂A bezeichnet. Nach Reskalierung können wir uns o. B. d. A. auf den Fall $\lambda = 1$ beschränken; sonst betrachte $\hat{u}(x, t) := u(x, \lambda t)$. Zumal die Ausdrücke (2.3) und (2.4) für jede Wahl von A gleich sein müssen, bekommen wir

$$\partial_t u - \Delta u = 0.$$

Die inhomogene Gleichung erhalten wir, wenn sich die Temperatur in A nicht nur durch Heraus- und Hineinfließen von Wärme ändern kann, ergo wenn Wärmequellen und -senken vorliegen. Wird diese „Quellen- bzw. Senkenstärke“ durch eine Dichte

$f = f(x, t)$ beschrieben, so liefert das den zusätzlichen Term $\int_A f dx$ in unseren Betrachtungen und wir erhalten

$$\partial_t u - \Delta u = f.$$

Es stellt sich noch die Frage der physikalischen Bedeutung der Anfangs- und Randbedingungen. Diesbezüglich ist in Einklang mit unserer physikalischen Intuition eine anfängliche Temperaturverteilung vorauszusetzen. Die Dirichlet- bzw. Neumann-Randbedingung kann unterdessen als das Festhalten der Temperatur u bzw. des Wärmestromes $\partial_\nu u$ in Richtung der äußeren Normalen am Rand interpretiert werden. Abschließend stellt die Robin-Randbedingung eine gewichtete Linearkombination von Dirichlet- und Neumann-Randbedingungen dar. Je nach gegebener Problemstellung muss daher zwischen den unterschiedlichen Randbedingungen gewählt werden.

2.3 Funktionalanalytische Vorbereitungen

In diesem Abschnitt gehen wir auf einige Konzepte der Funktionalanalysis ein, die insbesondere hilfreich sind für die Betrachtung von zeitabhängigen Funktionen.

Sei $\Omega \subset \mathbb{R}^d$ mit $d \in \mathbb{N}$ ein Gebiet, $T > 0$ der Endzeitpunkt und $u = u(x, t)$ eine Funktion definiert auf dem Raum-Zeit-Zylinder $\Omega \times (0, T)$. Alternativ können wir u als eine Funktion von der Zeit t auffassen mit Funktionswerten in einem Banachraum V . Dabei sind Elemente des Raumes V selbst wiederum Funktionen, die aber nur vom Ort x abhängen. Folglich diskutieren wir Funktionen u von der Form

$$u : (0, T) \rightarrow V, \quad t \mapsto u(t) \equiv u(\cdot, t). \quad (2.5)$$

Die erste Schwierigkeit ist die Wahl passender Räume von Funktionen der Form (2.5), die insbesondere Lösungen der PDE (2.2) sind. Eine erste Idee sind die Räume hinreichend stetig differenzierbarer Funktionen und ihre Unterräume, die vorgegebene Anfangs- und Randbedingungen berücksichtigen. Diese Räume sind allerdings aus funktionalanalytischer Sicht nicht besonders gutartig. Für Banachräume V sind hingegen die *Sobolev-Räume* genau die richtige Wahl, da diese den klassischen Differenziationsbegriff verallgemeinern.

Zuvor erinnern wir uns an die *Lebesgue-Räume*. Für $p \in [1, \infty]$ ist der Lebesgue-Raum $L^p(\Omega)$ der Raum aller p -fach integrierbarer Funktionen von Ω in die reellen Zahlen. Hinzu kommt, dass $L^p(\Omega)$ ein Banachraum mit der Norm

$$\|u\|_{L^p(\Omega)} := \begin{cases} \left(\int_{\Omega} |u(x)|^p dx \right)^{\frac{1}{p}}, & p < \infty, \\ \operatorname{ess\,sup}_{x \in \Omega} |u(x)|, & p = \infty, \end{cases}$$

ist. Zudem hat der Raum $L^2(\Omega)$ eine besondere Rolle unter den Lebesgue-Räumen. Dieser ist nämlich ein Hilbertraum mit dem Skalarprodukt

$$(u, v)_{L^2(\Omega)} := \int_{\Omega} u(x) v(x) dx.$$

Schwache Ableitung. Endlich können wir den Begriff der schwachen Differenzierbarkeit einführen. Seien $u, v \in L^1(\Omega)$ und $\alpha \in \mathbb{N}_0^d$ ein Multiindex. Dann heißt v die α -te schwache Ableitung von u , geschrieben $v = \partial^\alpha u$, wenn gilt

$$\int_{\Omega} u \partial^\alpha \chi dx = (-1)^{|\alpha|} \int_{\Omega} v \chi dx \quad \text{für alle } \chi \in \mathcal{C}_c^\infty(\Omega).$$

Dabei ist

$$\mathcal{C}_c^\infty(\Omega) := \{\chi \in \mathcal{C}^\infty(\Omega) \mid \text{supp } \chi \subset \Omega \text{ ist kompakt}\}$$

der Raum aller beliebig oft differenzierbaren Funktionen mit kompaktem Träger. Es lässt sich leicht überprüfen, dass die klassische Ableitung mit der schwachen Ableitung übereinstimmt, falls diese existiert. Für $p \in [1, \infty]$ ist der Sobolev-Raum $W^{k,p}(\Omega)$ der Raum aller $L^p(\Omega)$ -Funktionen, die k -mal schwach differenzierbar sind mit Ableitungen im Raum $L^p(\Omega)$. Obendrein ist $W^{k,p}(\Omega)$ ein Banachraum mit der Norm

$$\|u\|_{W^{k,p}(\Omega)} := \begin{cases} \left(\sum_{|\alpha| \leq k} \|\partial^\alpha u\|_{L^p(\Omega)}^p \right)^{\frac{1}{p}}, & p < \infty, \\ \sum_{|\alpha| \leq k} \|\partial^\alpha u\|_{L^\infty(\Omega)}, & p = \infty, \end{cases}$$

und der Halbnorm

$$[u]_{W^{k,p}(\Omega)} := \begin{cases} \left(\sum_{|\alpha|=k} \|\partial^\alpha u\|_{L^p(\Omega)}^p \right)^{\frac{1}{p}}, & p < \infty, \\ \sum_{|\alpha|=k} \|\partial^\alpha u\|_{L^\infty(\Omega)}, & p = \infty. \end{cases}$$

Wir merken an, dass $W^{0,p}(\Omega) = L^p(\Omega)$ gilt. Im Folgenden ist insbesondere der Hilbertraum $H^k(\Omega) := W^{k,2}(\Omega)$ mit dem Skalarprodukt

$$(u, v)_{H^k(\Omega)} := \sum_{|\alpha| \leq k} (\partial^\alpha u, \partial^\alpha v)_{L^2(\Omega)}$$

von Bedeutung.

Die schwache Ableitung bzw. die Sobolev-Räume wurden zum Lösen von PDEs entwickelt. Jedoch verbleibt noch die Schwierigkeit, inwiefern die Randbedingungen zu behandeln sind, da $W^{k,p}(\Omega)$ - und $L^p(\Omega)$ -Funktionen auf der Nullmenge $\partial\Omega$ nicht wohldefiniert sind. Hierfür betrachte folgendes Resultat im Falle $k = 1$ und $p = 2$ mit Sobolev-Räumen $H^s(\Omega)$ für fraktionale Exponenten $s > 0$ gegeben wie in [EG04, Satz B.30].

Satz 2.1 (Spursatz). Sei $\Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}$ mit Lipschitz-Rand $\partial\Omega$. Dann gibt es einen linearen und stetigen Spuroperator

$$\gamma_0 : H^1(\Omega) \rightarrow H^{\frac{1}{2}}(\partial\Omega)$$

so dass γ_0 surjektiv ist und $\gamma_0(u) = u|_{\partial\Omega}$ für $u \in \mathcal{C}^0(\overline{\Omega})$ gilt.

Beweis. Siehe [EG04, Satz B.52]. \square

Folglich ist der Spuroperator γ_0 eine stetige Fortsetzung der Restriktionsabbildung $u \mapsto u|_{\partial\Omega}$ und wir fassen die Aussage $u = g$ auf $\partial\Omega$ für Sobolev-Funktionen im Folgenden also präzise auf als $\gamma_0(u) = g$ fast überall auf $\partial\Omega$.

Null-Randbedingungen. Besonders wichtig für die Diskussion der homogenen Dirichlet-Randbedingung ist der Sobolev-Raum mit Null-Randbedingungen. Wir definieren $W_0^{k,p}(\Omega)$ als den Abschluss des Raumes $\mathcal{C}_c^\infty(\Omega)$ in $W^{k,p}(\Omega)$, d. h.

$$W_0^{k,p}(\Omega) := \overline{\mathcal{C}_c^\infty(\Omega)}^{W^{k,p}(\Omega)}.$$

Äquivalent gilt $u \in W_0^{k,p}(\Omega)$ genau dann, wenn es eine Folge $\{u_m\}_{m \in \mathbb{N}} \subset \mathcal{C}_c^\infty(\Omega)$ gibt mit $u_m \rightarrow u$ in $W^{k,p}(\Omega)$. Für $k = 1$ lässt sich zeigen, dass diese Menge genau die Sobolev-Funktionen mit Null-Randbedingungen sind. Hat also Ω einen Lipschitz-Rand, dann gilt $u \in W_0^{1,p}$ genau dann, wenn $u|_{\partial\Omega} = 0$ im Sinne von Spuren gilt. Des Weiteren ist nach der Poincaré-Ungleichung $\|\cdot\|_{W_0^{1,p}(\Omega)} := [\cdot]_{W^{1,p}(\Omega)}$ eine zu $\|\cdot\|_{W^{1,p}(\Omega)}$ äquivalente Norm für Funktionen im Unterraum $W_0^{1,p}(\Omega)$; vgl. [EG04, Lemma B.61 und Bemerkung B.62]. Für $p = 2$ ist $H_0^1(\Omega) := W_0^{1,2}(\Omega)$ außerdem ein Hilbertraum mit dem Skalarprodukt

$$(v, w)_{H_0^1(\Omega)} := \sum_{j=1}^d (\partial_j v, \partial_j w)_{L^2(\Omega)} = (\nabla v, \nabla w)_{L^2(\Omega)}.$$

Nach der Diskussion passender Banachräume für (2.5), betrachte folgende Räume von Banachraum-wertigen Funktionen.

- **Raum der stetigen Funktionen.** Für $k \in \mathbb{N}_0$ ist $\mathcal{C}^k([0, T]; V)$ der Raum der alle V -wertigen Funktionen zugehörig zur Differentialklasse \mathcal{C}^k umfasst, d. h. die k -mal stetig differenzierbaren Funktionen von $[0, T]$ nach V . Mit \dot{u} bezeichnen wir die erste Ableitung und mit $d_t^j u$ die j -te Ableitung von u nach der Zeit. Ein klassisches Resultat der Funktionalanalysis besagt, dass $\mathcal{C}^k([0, T]; V)$ ein Banachraum ist mit der Norm

$$\|u\|_{\mathcal{C}^k([0, T]; V)} := \sup_{t \in [0, T]} \sum_{j=0}^k \|d_t^j u(t)\|_V.$$

- **Lebesgue- bzw. Bochner-Raum.** Für $p \in [1, \infty]$ ist der Bochner-Raum $L^p(0, T; V)$ der Raum der alle V -wertigen Funktionen u umfasst, so dass Funktionen $\|u(\cdot)\|_V$ im Lebesgue-Raum $L^p((0, T))$ enthalten sind. Dann ist $L^p(0, T; V)$ ein Banachraum mit der Norm

$$\|u\|_{L^p(0, T; V)} := \begin{cases} \left(\int_0^T \|u(t)\|_V^p dt \right)^{\frac{1}{p}}, & p < \infty, \\ \operatorname{ess\,sup}_{t \in (0, T)} \|u(t)\|_V, & p = \infty. \end{cases}$$

- **Zeitabhängiger Sobolev-Raum.** Seien $p, q \in (1, \infty)$ und X_0, X_1 reflexive Banachräume mit stetiger Einbettung $X_0 \hookrightarrow X_1$. Dann ist der Raum

$$\mathcal{W}(X_0, X_1) := \{u \in L^p(0, T; X_0) \mid \dot{u} \in L^q(0, T; X_1)\}$$

ein Banachraum versehen mit der Norm

$$\|u\|_{\mathcal{W}(X_0, X_1)} := \|u\|_{L^p(0, T; X_0)} + \|\dot{u}\|_{L^q(0, T; X_1)}.$$

Hier bezeichnet \dot{u} die schwache Ableitung von u , d. h.

$$\int_0^T u(t) \dot{\chi}(t) dt = - \int_0^T \dot{u}(t) \chi(t) dt \quad \text{für alle } \chi \in \mathcal{C}_c^\infty([0, T]).$$

Aufgrund der Tatsache, dass zeitliche Evolutionsprobleme Anfangswertprobleme sind, müssen wir überprüfen, ob es für Funktionen auf $\Omega \times (0, T)$ legitim ist, Anfangswerte auf $\Omega \times \{0\}$ zu betrachten. Das folgendes Resultat liefert uns die gewünschte Eigenschaft.

Satz 2.2. *Seien $p, q \in (1, \infty)$ und $X_0 \hookrightarrow X \hookrightarrow X_1$ reflexive Banachräume mit stetigen Einbettungen. Dann gilt $\mathcal{W}(X_0, X_1) \subset \mathcal{C}^0([0, T]; X)$.*

Beweis. Siehe [EG04, Lemma 6.2]. □

In den folgenden Anwendungen ist besonders der Fall von Hilberträumen mit $p = q = 2$ von Bedeutung. Seien $V \hookrightarrow H$ Hilberträume mit stetiger und dichter Einbettung. Wenn wir H mit dem zugehörigen Dualraum H' gemäß dem Rieszschen Darstellungssatz [Eva10, Abschnitt D.3, Satz 2] identifizieren, folgt

$$V \hookrightarrow H \cong H' \hookrightarrow V',$$

d. h. die duale Paarung $\langle \cdot, \cdot \rangle_V$ kann als stetige Fortsetzung des Skalarproduktes $(\cdot, \cdot)_H$ interpretiert werden. Dabei nennt man das Tripel (V, H, V') einen *Gelfandschen Tripel* oder auch *Evolutionstripel*. Ein Evolutionstripel (V, H, V') induziert die dichten Einbettungen $L^p(0, T; V) \hookrightarrow L^p(0, T; H) \hookrightarrow L^p(0, T; V')$.

2.4 Abstraktes Evolutionsproblem

Wir sind jetzt in der Lage, ein grundlegendes Resultat für zeitabhängige (z. B. parabolische) Probleme zu formulieren. Dieses abstrakte Resultat spielt eine ähnlich bedeutende Rolle wie der Satz von Lax-Milgram [Eva10, Abschnitt 6.2.1] für elliptische Probleme.

Sei (V, H, V') ein Evolutionstripel. Sei $B : (0, T] \times V \times V \rightarrow \mathbb{R}$ eine Abbildung, so dass $B(t, \cdot, \cdot)$ für jedes $t \in (0, T]$ eine Bilinearform ist. Des Weiteren seien folgende Voraussetzungen an B erfüllt:

- (P1) Die Abbildung $t \rightarrow B(t, v, w)$ ist messbar für alle $v, w \in V$.
- (P2) Die Abbildung $(v, w) \rightarrow B(t, v, w)$ ist stetig, d. h. es existiert eine Konstante $C > 0$, so dass für alle $v, w \in V$ und $t \in [0, T]$ gilt:

$$|B(t, v, w)| \leq C\|v\|_V\|w\|_V.$$

- (P3) Die Abbildung $(v, w) \rightarrow B(t, v, w)$ ist schwach koerativ, d. h. es existieren Konstanten $\alpha, \gamma > 0$, so dass für alle $v \in V$ und $t \in [0, T]$ gilt:

$$B(t, v, v) \geq \alpha\|v\|_V^2 - \gamma\|v\|_H^2.$$

Für $u_0 \in H$ und $F \in L^2(0, T; V')$ betrachte folgende Problemstellung:

$$\begin{cases} \text{Gesucht ist } u \in \mathcal{W}(V, V'), \text{ so dass} \\ \langle \dot{u}(t), v \rangle_V + B(t, u(t), v) = \langle F(t), v \rangle_V \quad \text{für alle } v \in V, \quad \text{f.f.a. } t \in (0, T], \\ u(0) = u_0. \end{cases} \quad (2.6)$$

Nach Satz 2.2 ist die Anfangsbedingung $u(0) = u_0$ sinnvoll.

Bemerkung 2.3. Mit Hilfe eines Wechsels der Variablen ist es möglich die Voraussetzung (P3) so zu modifizieren, dass $\gamma = 0$ ist. Für eine detaillierte Erklärung siehe [EG04, S. 282]. Dann ist die Voraussetzungen (P3) äquivalent zu

- ($\widetilde{P3}$) Es gibt $\alpha > 0$, so dass für alle $v \in V$ und $t \in [0, T]$ gilt: $B(t, v, v) \geq \alpha\|v\|_V^2$.

Die folgenden Sätze zeigen, dass das Problem (2.6) *wohlgestellt* ist.

Satz 2.4 (Existenz und Eindeutigkeit). *Unter den Voraussetzungen (P1) bis (P3) hat das Anfangswertproblem (2.6) genau eine Lösung.*

Beweis. Siehe [EG04, Satz 6.6]. □

Satz 2.5 (A-priori Abschätzungen). Für $F \in L^2(0, T; V')$ erfüllt die Lösung u von (2.6) unter den Voraussetzungen (P1) bis (P3) folgende Abschätzungen:

$$\begin{aligned}\|u\|_{C^0([0,T],H)} &\leq \|u_0\|_H e^{-\frac{1}{2}\alpha c_P t} + \frac{1}{\sqrt{\alpha}} \|F\|_{L^2(0,T;V')}, \\ \|u\|_{L^2([0,T],V)} &\leq \frac{1}{\sqrt{\alpha}} \|u_0\|_H + \frac{1}{\alpha} \|F\|_{L^2(0,T;V')},\end{aligned}$$

wobei α, c_p so gegeben sind, dass $c_P^{-\frac{1}{2}}$ die Norm der Einbettung $V \hookrightarrow H$ und α die Koerzivitätskonstante aus ($\widetilde{P}3$) ist. Des Weiteren falls $F \in L^\infty(0, \infty; V')$ gilt:

$$\limsup_{t \rightarrow \infty} \|u(t)\|_H \leq \frac{1}{\alpha \sqrt{c_P}} \|F\|_{L^\infty(0, \infty; V')}.$$

Beweis. Siehe [EG04, Satz 6.7]. \square

Bemerkung 2.6. Satz 2.5 liefert eine stetige Abhängigkeit der Lösung u von den Daten (u_0, F) . Wir werden später sehen, dass numerische Lösungen auf Basis der Finite-Elemente-Methode vergleichbare (Stabilitäts-)Eigenschaften aufweisen.

2.5 Schwache Formulierung

Nachdem wir nun die notwendigen Begriffe der Funktionalanalysis für die Diskussion der Wärmeleitungsgleichung kennengelernt haben, können wir uns dem Anfangs- und Randwertproblem

$$\partial_t u - \Delta u = f \quad \text{in } \Omega_T, \tag{2.7a}$$

$$u(\cdot, 0) = u_0 \quad \text{in } \Omega, \tag{2.7b}$$

$$\mathcal{B}(u) = g \quad \text{auf } \Gamma_T, \tag{2.7c}$$

für gegebene Funktionen $f : \Omega_T \rightarrow \mathbb{R}$, $u_0 : \Omega \rightarrow \mathbb{R}$ und $g : \Gamma_T \rightarrow \mathbb{R}$ widmen. Derweil sei $u : \overline{\Omega}_T \rightarrow \mathbb{R}$ die gesuchte Funktion. Ziel ist es aus dem Anfangs- und Randwertproblem eine Problemstellung der Form (2.6) zu gewinnen. Diese bezeichnen wir auch als die *schwache Formulierung* von (2.7). Hierfür multiplizieren wir (2.7a) mit einer hinreichend glatten Testfunktion v und integrieren partiell nach x und erhalten

$$\int_{\Omega} \partial_t u v \, dx + \int_{\Omega} \nabla u \cdot \nabla v \, dx - \int_{\partial\Omega} \partial_{\nu} u v \, dS = \int_{\Omega} f v \, dx. \tag{2.8}$$

Im Folgenden fassen wir u nach Abschnitt 2.3 als eine Banachraum-wertige Funktion $u : [0, T] \rightarrow V$ auf. In diesem Fall stimmt $(u(t))(x)$ mit $u(x, t)$ aus (2.7) überein. Betrachte das Evolutionstripel (V, H, V') mit der Inklusionsabbildung $\text{Id} : V \hookrightarrow H$ für

$$H := L^2(\Omega) \quad \text{und} \quad H_0^1(\Omega) \subseteq V \subseteq H^1(\Omega),$$

wobei V abhängig von den Randbedingungen ist. In Analogie zur elliptischen Theorie [Eva10, Abschnitt 6.1.2] fassen wir $-\Delta u$ und f als Funktionale in $L^2(0, T; V')$

auf. Da $L^2(\Omega_T)$ und $L^2(0, T; L^2(\Omega))$ kanonisch isometrisch zueinander sind, erhalten wir die Einbettung

$$L^2(\Omega_T) \cong L^2(0, T; H) \hookrightarrow L^2(0, T; V').$$

Das legt nahe, Funktionen u mit $\dot{u} \in L^2(0, T; V')$ zu suchen. Unter Verwendung von (2.8) mit Testfunktionen in V erhalten wir

$$\langle \dot{u}(t), v \rangle_V + B(u(t), v) - (\partial_\nu u(t), v)_{L^2(\partial\Omega)} = \langle f(t), v \rangle_V \quad \text{f.f.a. } t \in (0, T]. \quad (2.9)$$

Dabei ist hier die Bilinearform $B : V \times V \rightarrow \mathbb{R}$, so dass

$$B(u, v) := (\nabla u, \nabla v)_{L^2(\Omega)} = (u, v)_{H_0^1(\Omega)} \quad \text{für alle } u, v \in V. \quad (2.10)$$

Letztendlich stellt Gleichung (2.9) einen Anhaltspunkt zur Konstruktion der schwachen Formulierung von (2.7) für die verschiedenen Randbedingungen dar. Hierbei unterscheiden wir im Folgenden zwischen *essentiellen* bzw. *wesentlichen* und *natürlichen* Randbedingungen. Wohingegen essentielle Randbedingungen entscheidend in der Wahl des Raumes V sind, beeinflussen natürliche Randbedingungen die Gleichung der schwachen Formulierung.

2.6 Essentielle Randbedingungen

In diesem Abschnitt konstruieren wir die schwache Formulierung von (2.7) für Dirichlet-Randbedingungen und überprüfen anschließend schwache Lösungen auf Existenz und Eindeutigkeit.

Homogene Dirichlet-Randbedingungen. Wir beginnen zunächst mit homogenen Randbedingungen, d. h. $u = 0$ auf $\partial\Omega$ im Sinne von Spuren. Die Null-Randbedingung können wir insofern erzwingen, indem wir $V = H_0^1(\Omega)$ wählen. Somit wird klar, inwiefern homogene Dirichlet-Randbedingungen der Klasse der essentiellen Randbedingungen zugehörig sind. Folglich verschwindet das Randintegral aus der Gleichung (2.9) unter Berücksichtigung von $v \in H_0^1(\Omega)$. Die schwache Formulierung ist schließlich wie folgt gegeben:

$$\begin{cases} \text{Gesucht ist } u \in \mathcal{W}(V, V') \text{ mit } V = H_0^1(\Omega), \text{ so dass} \\ \langle \dot{u}(t), v \rangle_V + B(u(t), v) = \langle f(t), v \rangle_V \quad \text{für alle } v \in V, \quad \text{f.f.a. } t \in (0, T], \\ u(0) = u_0. \end{cases}$$

Satz 2.7 (Existenz und Eindeutigkeit). Seien $u_0 \in L^2(\Omega)$ und $f \in L^2(\Omega_T)$. Dann hat das Anfangs- und Randwertproblem

$$\begin{aligned} \partial_t u - \Delta u &= f && \text{in } \Omega_T, \\ u(\cdot, 0) &= u_0 && \text{in } \Omega, \\ u &= 0 && \text{auf } \Gamma_T, \end{aligned}$$

eine eindeutige schwache Lösung $u \in \mathcal{W}(H_0^1(\Omega), (H_0^1(\Omega))')$.

Beweis. Die Behauptung ist nach Satz 2.4 gezeigt, wenn wir überprüft haben, dass die Bilinearform B aus (2.10) und die Daten f die Voraussetzungen von Satz 2.4 erfüllen. Die Stetigkeitsbedingung ergibt sich durch

$$|B(v, w)| = |(v, w)_{H_0^1(\Omega)}| \leq \|v\|_{H_0^1(\Omega)} \|w\|_{H_0^1(\Omega)} \quad \text{für alle } v, w \in H_0^1(\Omega)$$

mit Hilfe der Cauchy-Schwarzschen Ungleichung. Des Weiteren ist wegen der Nichtnegativität der Norm mit

$$B(v, v) = \|v\|_{H_0^1(\Omega)}^2 \geq \|v\|_{H_0^1(\Omega)}^2 - \|v\|_{L^2(\Omega)}^2 \quad \text{für alle } v \in H_0^1(\Omega)$$

die Koerzivitätsbedingung erfüllt. Dass f den Voraussetzungen von Satz 2.4 genügt, ergibt sich aus

$$f \in L^2(\Omega_T) \cong L^2(0, T; L^2(\Omega)) \hookrightarrow L^2(0, T; (H_0^1(\Omega))').$$

□

Inhomogene Dirichlet-Randbedingungen. Weiterhin betrachten wir die inhomogene Dirichlet-Randbedingung $u = g$ auf $\partial\Omega$ im Sinne von Spuren für eine gegebene Funktion $g : \Gamma_T \rightarrow \mathbb{R}$. Zur Einfachheit diskutieren wir nur zeitunabhängige rechte Seiten der Randbedingung. Hierbei nehmen wir an, dass $g(\cdot, t) \equiv g \in H^{\frac{1}{2}}(\partial\Omega)$ für alle $t \in (0, T]$ ist. Nach Satz 2.1 existiert ein $\omega \in H^1(\Omega)$, so dass $\omega|_{\partial\Omega} = g$ im Sinne von Spuren. Zur Rückführung auf den Fall mit homogenen Randbedingungen betrachte die schwache Lösung von

$$\begin{cases} \text{Gesucht ist } \psi \in \mathcal{W}(V, V') \text{ mit } V = H_0^1(\Omega), \text{ so dass} \\ \langle \dot{\psi}(t), v \rangle_V + B(\psi(t), v) = \langle F(t), v \rangle_V \quad \text{für alle } v \in V, \quad \text{f.f.a. } t \in (0, T], \\ \psi(0) = u_0 - \omega \end{cases}$$

mit $f \in L^2(\Omega_T)$ und $u_0 \in H$. Dabei sind B aus (2.10) und $F : [0, T] \rightarrow V'$, so dass

$$\langle F(t), v \rangle_V := \langle f(t), v \rangle_V - B(\omega, v) \quad \text{für alle } v \in V, \quad \text{f.f.a. } t \in (0, T].$$

Dass F den Voraussetzungen von Satz 2.4 genügt, ergibt sich aus $f \in L^2(0, T; V')$ und $B(\omega, \cdot) \in V'$. Dann lässt sich leicht überprüfen, dass u mit

$$u(t) = \psi(t) + \omega \quad \text{für } t \in [0, T]$$

die eindeutige schwache Lösung der Wärmeleitungsgleichung ist für inhomogene Dirichlet-Randbedingungen

$$u = g \quad \text{auf } \Gamma_T.$$

2.7 Natürliche Randbedingungen

In diesem Abschnitt konstruieren wir die schwache Formulierung von (2.7) für Robin- und Neumann-Randbedingungen und überprüfen anschließend schwache Lösungen auf Existenz und Eindeutigkeit.

Robin-Randbedingungen. Wir betrachten mit gegebenen Funktionen $g \in L^2(\Gamma_T)$ und $\rho \in L^\infty(\Gamma_T)$, wobei ρ nichtnegativ auf Mengen mit Maß > 0 ist, die Robin-Randbedingung

$$\partial_\nu u + \rho u = g \quad \text{auf } \Gamma_T. \quad (2.11)$$

Um die Randbedingung sinnvoll im schwachen Sinne interpretieren zu können, ziehen wir das Evolutionstripel $(H^{\frac{1}{2}}(\partial\Omega), L^2(\partial\Omega), H^{-\frac{1}{2}}(\partial\Omega))$ heran, wobei $H^{-\frac{1}{2}}(\partial\Omega)$ der zugehörige Dualraum von $H^{\frac{1}{2}}(\partial\Omega)$ ist. Ausgehend von Gleichung (2.9) für Testfunktionen in $V = H^1(\Omega)$ (d. h. nicht nur $H_0^1(\Omega)$, wie bei der Dirichlet-Randbedingung), ersetze die Richtungsableitung $\partial_\nu u$ durch $g - \rho u$ um die Robin-Randbedingung zu erzwingen. Da $\langle \cdot, \cdot \rangle_{H^{\frac{1}{2}}(\partial\Omega)}$ eine stetige Fortsetzung von $\langle \cdot, \cdot \rangle_{L^2(\partial\Omega)}$ ist, sind die Randintegrale äquivalent zu Auswertungen von Funktionalen in $H^{-\frac{1}{2}}(\partial\Omega)$. Nach passenden Umformungen ist die schwache Formulierung wie folgt gegeben:

$$\begin{cases} \text{Gesucht ist } u \in \mathcal{W}(V, V') \text{ mit } V = H^1(\Omega), \text{ so dass} \\ \langle \dot{u}(t), v \rangle_V + \tilde{B}(t, u(t), v) = \langle F(t), v \rangle_V \quad \text{für alle } v \in V, \quad \text{f.f.a. } t \in (0, T], \\ u(0) = u_0. \end{cases}$$

Einerseits ist hier die Abbildung $\tilde{B} : (0, T] \times V \times V \rightarrow \mathbb{R}$, so dass

$$\tilde{B}(t, u, v) := B(u, v) + (\rho(t)u, v)_{L^2(\partial\Omega)} \quad \text{für alle } u, v \in V, \quad \text{f.f.a. } t \in (0, T] \quad (2.12)$$

mit B aus (2.10). Andererseits ist $F : (0, T] \rightarrow V'$ so gegeben, dass

$$\langle F(t), v \rangle_V := \langle f(t), v \rangle_V + \langle g(t), v \rangle_{H^{\frac{1}{2}}(\partial\Omega)} \quad \text{für alle } v \in V, \quad \text{f.f.a. } t \in (0, T]. \quad (2.13)$$

Dass F tatsächlich ein lineares Funktional in $L^2(0, T; V')$ ist und somit die Schreibweise $\langle \cdot, \cdot \rangle_V$ gerechtfertigt ist wird im Beweis des folgenden Satzes begründet. Außerdem ist nach Satz 2.1 die Relation $v|_{\partial\Omega} \in H^{\frac{1}{2}}(\partial\Omega)$ für (2.13) sinnvoll.

Satz 2.8 (Existenz und Eindeutigkeit). Seien $u_0 \in L^2(\Omega)$, $f \in L^2(\Omega_T)$, $g \in L^2(\Gamma_T)$ und $\rho \in L^\infty(\Gamma_T)$, wobei ρ nichtnegativ auf Mengen mit Maß > 0 ist. Dann hat das Anfangs- und Randwertproblem

$$\begin{aligned} \partial_t u - \Delta u &= f && \text{in } \Omega_T, \\ u(\cdot, 0) &= u_0 && \text{in } \Omega, \\ \partial_\nu u + \rho u &= g && \text{auf } \Gamma_T, \end{aligned}$$

eine eindeutige schwache Lösung $u \in \mathcal{W}(H^1(\Omega), (H^1(\Omega))')$.

Beweis. Seien $v, w \in H^1(\Omega)$. Unter Verwendung der Cauchy-Schwarzschen Ungleichung für $(\cdot, \cdot)_{L^2(\partial\Omega)}$ und der Stetigkeit des Spuroperators $\gamma_0 : H^1(\Omega) \rightarrow L^2(\partial\Omega)$ ergibt sich

$$\begin{aligned} (\rho(t)v, w)_{L^2(\partial\Omega)} &\leq \|\rho(t)\|_{L^\infty(\partial\Omega)} \|v\|_{L^2(\partial\Omega)} \|w\|_{L^2(\partial\Omega)} \\ &\leq C_{\gamma_0}^2 \|\rho\|_{L^\infty(0,T;L^\infty(\partial\Omega))} \|v\|_{H^1(\Omega)} \|w\|_{H^1(\Omega)} \end{aligned} \quad (2.14)$$

für fast alle $t \in (0, T]$. Dabei nutzen wir die Identifikation

$$L^\infty(\Gamma_T) \cong L^\infty(0, T; L^\infty(\partial\Omega)) \quad \text{mit} \quad \|\rho\|_{L^\infty(\Gamma_T)} = \|\rho\|_{L^\infty(0, T; L^\infty(\partial\Omega))}.$$

Wir stellen fest dass, nach der Definition der $H^1(\Omega)$ -Norm, d. h.

$$\|v\|_{H^1(\Omega)}^2 = \|v\|_{L^2(\Omega)}^2 + \|\nabla v\|_{L^2(\Omega)}^2,$$

die Ungleichung $\|\nabla v\|_{L^2(\Omega)} \leq \|v\|_{H^1(\Omega)}$ gilt. Mit der Cauchy-Schwarzschen Ungleichung folgt die Stetigkeit von B in $H^1(\Omega) \times H^1(\Omega)$ wegen

$$|B(v, w)| = |(\nabla v, \nabla w)_{L^2(\Omega)}| \leq \|\nabla v\|_{L^2(\Omega)} \|\nabla w\|_{L^2(\Omega)} \leq \|v\|_{H^1(\Omega)} \|w\|_{H^1(\Omega)}. \quad (2.15)$$

Kombinieren wir die Resultate (2.14) und (2.15), so folgt

$$|\tilde{B}(t, v, w)| \leq |B(v, w)| + |(\rho(t)v, w)_{L^2(\partial\Omega)}| \leq C \|v\|_{H^1(\Omega)} \|w\|_{H^1(\Omega)}$$

mit $C := 1 + C_{\gamma_0}^2 \|\rho\|_{L^\infty(\Gamma_T)} < \infty$ und die Stetigkeitsbedingung aus Satz 2.4 ist erfüllt. Des Weiteren ergibt sich die Koerzivitätsbedingung durch

$$\tilde{B}(t, v, v) = \|\nabla v\|_{L^2(\Omega)}^2 + (\rho(t)v, v)_{L^2(\partial\Omega)} \geq \|\nabla v\|_{L^2(\Omega)}^2 = \|v\|_{H^1(\Omega)}^2 - \|v\|_{L^2(\Omega)}^2,$$

für fast alle $t \in (0, T]$ unter der Voraussetzung dass ρ fast überall nichtnegativ ist. Abschließend genügt F den Voraussetzungen von Satz 2.4, da f und g lineare, stetige Funktionale auf $L^2(0, T; H^1(\Omega))$ sind. \square

Neumann-Randbedingungen. Die Neumann-Randbedingung erlangen wir bei Betrachtung der Robin-Randbedingung, gesetzt dem Fall, dass ρ auf Γ_T verschwindet. Daher erhalten wir folgende schwache Formulierung für $\rho \equiv 0$ in (2.11):

$$\begin{cases} \text{Gesucht ist } u \in \mathcal{W}(V, V') \text{ mit } V = H^1(\Omega), \text{ so dass} \\ \langle \dot{u}(t), v \rangle_V + B(u(t), v) = \langle F(t), v \rangle_V \quad \text{für alle } v \in V, \quad \text{f.f.a. } t \in (0, T], \\ u(0) = u_0. \end{cases}$$

Hier sind B aus (2.10) und F aus (2.13).

Satz 2.9 (Existenz und Eindeutigkeit). Seien $u_0 \in L^2(\Omega)$, $f \in L^2(\Omega_T)$ und $g \in L^2(\Gamma_T)$. Dann hat das Anfangs- und Randwertproblem

$$\begin{aligned}\partial_t u - \Delta u &= f && \text{in } \Omega_T, \\ u(\cdot, 0) &= u_0 && \text{in } \Omega, \\ \partial_\nu u &= g && \text{auf } \Gamma_T,\end{aligned}$$

eine eindeutige schwache Lösung $u \in \mathcal{W}(H^1(\Omega), (H^1(\Omega))')$.

Beweis. Setze $\rho \equiv 0$ in Satz 2.8. \square

2.8 Gemischte Randbedingungen

In diesem Abschnitt gehen wir auf die schwache Formulierung von (2.7) für gemischten Randbedingungen ein. Aussagen über Existenz und Eindeutigkeit von schwachen Lösungen in diesem Fall wird im Rahmen dieser Arbeit nicht eingegangen. Wir diskutieren o. B. d. A. die Kombination von Dirichlet- und Robin-Randbedingungen.

Sei $\partial\Omega_D$ eine nichtleere, geschlossene Teilmenge von $\partial\Omega$ und setze $\partial\Omega_R := \partial\Omega \setminus \partial\Omega_D$. Außerdem nehmen wir an, dass beide Teilmengen $\partial\Omega_D$ und $\partial\Omega_R$ entsprechend Abbildung 1b positives Maß besitzen. Definiere $\Gamma_{T,D} := \partial\Omega_D \times (0, T]$ und $\Gamma_{T,R} := \partial\Omega_R \times (0, T]$. Anschließend soll in $\Gamma_{T,D}$ bzw. $\Gamma_{T,R}$ die Dirichlet- bzw. Robin-Randbedingung gelten, d. h.

$$u = g_D \quad \text{auf } \Gamma_{T,D}, \tag{2.16a}$$

$$\partial_\nu u + \rho u = g_R \quad \text{auf } \Gamma_{T,R} \tag{2.16b}$$

mit gegebenen Funktionen g_D und g_R über $\Gamma_{T,D}$ und $\Gamma_{T,R}$ und $\rho \in L^\infty(\Gamma_{T,R})$ nichtnegativ fast überall. Hierbei diskutieren wir nur zeitunabhängige Funktionen $g_D(\cdot, t) \equiv g_D$ für alle $t \in (0, T]$. Wir nehmen an, dass eine Fortsetzung \tilde{g}_D auf $\partial\Omega$ von g_D existiert und in $H^{\frac{1}{2}}(\partial\Omega)$ liegt. Nach dem Spursatz 2.1 existiert ein $\omega_D \in H^1(\Omega)$, so dass $\omega_D|_{\partial\Omega} = \tilde{g}_D$, ergo $\omega_D|_{\partial\Omega_D} = g_D$ gilt. Des Weiteren setzen wir voraus, dass $g_R \in L^2(\Gamma_{T,R}) \hookrightarrow L^2(0, T; H^{-\frac{1}{2}}(\partial\Omega_R))$. Angelehnt an die Vorgehensweise bei inhomogenen Dirichlet- und Robin-Randbedingungen, betrachten wir folgende schwache Formulierung:

$$\left\{ \begin{array}{l} \text{Gesucht ist } \psi \in \mathcal{W}(V, V') \text{ mit } V = \{v \in H^1(\Omega) \mid v|_{\partial\Omega_D} = 0\}, \text{ so dass} \\ \langle \dot{\psi}(t), v \rangle_V + \tilde{B}(t, \psi(t), v) = \langle F(t), v \rangle_V \quad \text{für alle } v \in V, \quad \text{f.f.a. } t \in (0, T], \\ \psi(0) = u_0 - \omega_D. \end{array} \right.$$

Einerseits ist hier die Abbildung $\tilde{B} : (0, T] \times V \times V \rightarrow \mathbb{R}$, so dass

$$\tilde{B}(t, u, v) := B(u, v) + (\rho(t)u, v)_{L^2(\partial\Omega_R)} \quad \text{für alle } u, v \in V, \quad \text{f.f.a. } t \in (0, T] \tag{2.17}$$

mit B aus (2.10). Andererseits ist $F : (0, T] \rightarrow V'$ so gegeben, dass für alle $v \in V$ und für fast alle $t \in (0, T]$ gilt:

$$\langle F(t), v \rangle_V := \langle f(t), v \rangle_V + \langle g_R(t), v \rangle_{H^{\frac{1}{2}}(\partial\Omega_R)} - \tilde{B}(t, \omega_D, v).$$

Dann ist mit den obigen Voraussetzungen an g_D und g_R die Funktion u mit

$$u(t) = \psi(t) + \omega_D \quad \text{für } t \in [0, T]$$

die eindeutige schwache Lösung für Randbedingungen (2.16).

3 Finite-Elemente-Methode

In diesem Kapitel werden wir uns mit der Approximation von parabolischen PDEs zweiter Ordnung mit Hilfe der *Finite-Elemente-Methode* (FEM) beschäftigen. Zuvor erinnern wir uns an die schwache Formulierung der Wärme-leitungsgleichung:

$$\begin{cases} \text{Gesucht ist } u \in \mathcal{W}(V, V'), \text{ so dass} \\ \langle \dot{u}(t), v \rangle_V + B(t, u(t), v) = \langle F(t), v \rangle_V \quad \text{für alle } v \in V, \quad \text{f.f.a. } t \in (0, T], \\ u(0) = u_0 \in H := L^2(\Omega) \end{cases} \quad (3.1)$$

mit den Abbildungen $B : (0, T] \times V \times V \rightarrow \mathbb{R}$ und $F : (0, T] \rightarrow V'$ für den Funktionenraum V , so dass $H_0^1(\Omega) \subseteq V \subseteq H^1(\Omega)$, abhängig von den Randbedingungen; siehe Abschnitte 2.4 bis 2.7.

Um eine Näherungslösung von (3.1) zu erhalten, müssen wir sowohl im Ort als auch in der Zeit diskretisieren, wobei wir diese separat voneinander betrachten. Wir unterscheiden hierbei in folgenden beiden Vorgehensweisen:

- **Linienmethode.** Die Idee hinter der (vertikalen) Linienmethode besteht darin, zuerst eine Diskretisierung hinsichtlich der Ortsvariablen und danach das resultierende Problem hinsichtlich der Zeit zu diskretisieren. Dabei erhält man nach der Ortsdiskretisierung ein lineares System von ODEs. Diese können nun durch klassische Verfahren, wie z. B. das *implizite Euler-Verfahren* gelöst werden. Der Nachteil ist, dass die Ortsdiskretisierung fest ist und daher lokale, zeitabhängige Verfeinerungen nicht möglich sind.
- **Rothe-Methode.** Der vertikalen Linienmethode steht die horizontale Linienmethode gegenüber, welche auch unter dem Namen Rothe-Methode bekannt ist. Die Idee der Rothe-Methode besteht darin, zuerst eine Diskretisierung hinsichtlich der Zeitvariable vorzunehmen, um so eine ODE in einem Banachraum zu erhalten und diese anschließend nach den Ortsvariablen zu diskretisieren. Infolgedessen hat die Rothe-Methode den Vorteil, dass in jedem Zeitschritt ein anderes Verfahren zur Ortsdiskretisierung gewählt werden kann.

Im Folgenden schränken wir uns auf die Untersuchung der vertikalen Linienmethode ein. Einfachheitshalber sprechen wir hier von der FEM, obwohl wir damit eigentlich die Linienmethode, d. h. die Diskretisierung im Ort und in der Zeit meinen (im Allgemeinen bezieht sich der Begriff der FEM nur auf die Ortsdiskretisierung). Eine detaillierte Betrachtung der Rothe-Methode ist in [Tho04, Kapitel 7-9] zu finden.

3.1 Galerkin-Verfahren

Als Erstes diskutieren wir die Diskretisierung im Ort und betrachten hierfür das zu (2.7) zugehörige stationäre (d. h. zeitunabhängige) Problem, die *Poisson-Gleichung*

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega, \\ \mathcal{B}(u) &= g \quad \text{auf } \partial\Omega, \end{aligned}$$

mit zugehöriger schwachen Formulierung

$$\text{Gesucht ist } u \in V, \text{ so dass } \quad B(u, v) = \langle F, v \rangle_V \quad \text{für alle } v \in V. \quad (3.2)$$

Für diesen Fall nehmen wir an, dass $F \in V'$ und $B : V \times V \rightarrow \mathbb{R}$ eine stetige und koerzive Bilinearform mit Stetigkeitskonstante $C > 0$ und der Koerzivitätskonstante $\alpha > 0$ ist, d. h.

$$|B(v, w)| \leq C\|v\|_V\|w\|_V \quad \text{und} \quad B(v, v) \geq \alpha\|v\|_V^2 \quad (3.3)$$

für alle $v, w \in V$.

Die Grundidee der Finite-Elemente-Methode stellt das *Galerkin-Verfahren* dar, ein numerisches Verfahren zur näherungsweisen Lösung von PDEs. Hierbei konstruieren wir eine Näherung u_h der exakten Lösung u im endlichdimensionalen Teilraum V_h des unendlichdimensionalen Funktionenraumes V mit praktisch gut zu verarbeitenden Basen; vgl. [EG04, Abschnitt 2.2]. Hier stellt $h > 0$ ein Parameter zur Diskretisierung des Ortsraums Ω dar. Aufgrund der Tatsache, dass die Bilinearform B und das Funktional F auf $V_h \times V_h$ bzw. V_h definiert sind, betrachten wir folgende diskrete Problemstellung:

$$\text{Gesucht ist } u_h \in V_h, \text{ so dass } \quad B(u_h, v_h) = \langle F, v_h \rangle_V \quad \text{für alle } v_h \in V_h. \quad (3.4)$$

Bemerkung 3.1. Die Existenz und Eindeutigkeit der Lösung u von (3.2) und u_h von (3.4) folgen unmittelbar aus dem Satz von Lax-Milgram, da der Teilraum V_h wieder ein Hilbertraum mit dem aus V geerbten Skalarprodukt ist.

Der Fehler $u - u_h$ zwischen der exakten und diskreten Lösung erfüllt dabei folgende abstrakte Abschätzung.

Satz 3.2 (Céa). *Seien $V_h \subset V$ mit $\dim V_h < \infty$, die Bilinearform $B : V \times V \rightarrow \mathbb{R}$ stetig und koerziv wie in (3.3) und $F \in V'$. Dann gilt für die Lösung u von (3.2) und für die Lösung u_h von (3.4), dass*

$$\|u - u_h\|_V \leq \frac{C}{\alpha} \inf_{v_h \in V_h} \|u - v_h\|_V.$$

Beweis. Aufgrund der Billinearität von B folgt

$$B(u - u_h, \chi) = 0 \quad \text{für alle } \chi \in V_h$$

und wird *Galerkin-Orthogonalität* genannt. Infolge der Stetigkeit und Koerzivität der Bilinearform B gilt für $v_h \in V_h$, dass

$$\begin{aligned} \alpha \|u - u_h\|_V^2 &\leq B(u - u_h, u - u_h) \\ &= B(u - u_h, u - v_h) + B(u - u_h, v_h - u_h) \\ &= B(u - u_h, u - v_h) \\ &\leq C \|u - u_h\|_V \|u - v_h\|_V. \end{aligned}$$

Die Ungleichungskette kann nun durch $\alpha \|u - u_h\|_V$ geteilt werden. Schließlich ist $v_h \in V_h$ beliebig und es kann das Infimum gewählt werden, wodurch die Behauptung folgt. \square

Diese abstrakte Fehlerabschätzung zeigt, dass der Fehler zwischen der exakten und diskreten Lösung von der bestmöglichen Approximation V_h des Funktionenraumes V abhängt. Des Weiteren steht der Fehler orthogonal zum Raum V_h bzgl. des Skalarproduktes $B(\cdot, \cdot)$ und folglich ist u_h die Bestapproximierende von u in V_h bezüglich der Energienorm $\|\cdot\|_B := \sqrt{B(\cdot, \cdot)}$.

Ziel der nächsten Abschnitte 3.2 bis 3.4 ist die Konstruktion geeigneter endlichdimensionaler Teilräume V_h , die auf einer simplizialen Zerlegung bzw. Triangulierung des Gebietes Ω beruhen. Hierfür orientieren wir uns an [Dzi10, Abschnitte 3.2.1 und 3.2.2].

3.2 Triangulierung

In diesem Abschnitt gehen wir auf die simpliziale Zerlegung bzw. Triangulierung des Gebietes Ω ein. Zu Beginn definieren wir den Begriff eines Simplex.

Definition 3.3 (Simplex). Für $n = 1, \dots, d$ seien die $n+1$ Punkte $a_0, \dots, a_n \in \mathbb{R}^d$ affin unabhängig, d. h. $a_1 - a_0, \dots, a_n - a_0$ sind linear unabhängig. Dann bezeichne die konvexe Hülle von $\{a_j\}_{0 \leq j \leq n}$ gegeben durch

$$K := \left\{ x \in \mathbb{R}^d \mid x = \sum_{j=0}^n \lambda_j a_j, \sum_{j=0}^n \lambda_j = 1, \lambda_j \geq 0 \right\}$$

als einen n -Simplex in \mathbb{R}^d mit Eckpunkten $\{a_j\}_{0 \leq j \leq n}$.

Für $I \subset \{0, 1, \dots, n\}$ mit $|I| = m$ bezeichne die konvexe Hülle von $\{a_j\}_{j \in I}$ als den m -dimensionalen Seitensimplex von K . Die nulldimensionalen Seitensimplizes von K sind dabei genau die Eckpunkte von K .

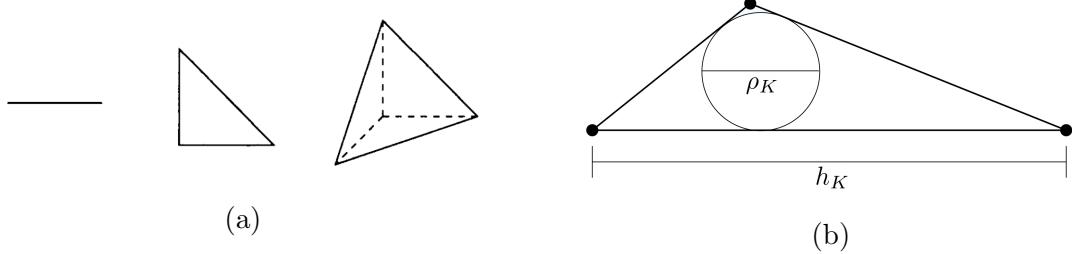


Abbildung 2: (a) n -Simplizes für $n = 1, 2, 3$, siehe [EG04, Abbildung 1.15]. (b) 2-Simplex mit zugehörigen Durchmessern h_K und ρ_K , in Anlehnung an [Zü13, Abbildung 2.1(c)].

Einige Beispiele eines Simplex sind in Abbildung 2a zu sehen. Des Weiteren definieren wir folgende geometrische Parameter, veranschaulicht in Abbildung 2b, für ein n -Simplex K :

- (1) der *Durchmesser* $h_K := \max_{0 \leq i, j \leq n} |a_j - a_i|$ ist der größte Abstand zweier Eckpunkte,
- (2) der *Innendurchmesser* $\rho_K := 2 \sup\{r > 0 \mid B_r(x) \subset K, x \in K\}$ ist der Durchmesser der größten, im Simplex vollständig liegenden Kugel $B_r(x) := \{y \in \mathbb{R}^d : |y - x| < r\}$,
- (3) das *Verhältnis* $\sigma_K := \frac{h_K}{\rho_K}$ ist ein Maß für die Nicht-Entartung von K .

Für eine einfachere Beschreibung von Punkten in Simplizes ist es praktischer ein Koordinatensystem zu betrachten, welches nicht dem kartesischen Koordinaten entspricht. Diesbezüglich führen wir den Begriff der baryzentrischen Koordinaten ein.

Definition 3.4 (Baryzentrische Koordinaten). Für alle $0 \leq j \leq d$ sind die *baryzentrischen Koordinaten* λ_j eines Punktes $x = (x_1, \dots, x_d)^\top \in \mathbb{R}^d$ im d -Simplex K mit Eckpunkten $\{a_j\}_{0 \leq j \leq d}$ die Polynome erster Ordnung von der Form

$$\lambda_j(x) := \alpha_{j,0} + \alpha_{j,1}x_1 + \cdots + \alpha_{j,d}x_d,$$

so dass $\lambda_j(a_i) = \delta_{ij}$ für alle $0 \leq i, j \leq d$ gilt.

Satz 3.5. Für alle $x \in K$ existiert genau ein eindeutiger Vektor $(\lambda_j(x))_{0 \leq j \leq d}$, so dass folgende Eigenschaften erfüllt sind:

$$x = \sum_{j=0}^d a_j \lambda_j(x) \quad \text{und} \quad \sum_{j=0}^d \lambda_j(x) = 1. \tag{3.5}$$

Beweis. Die Gleichungen (3.5) lassen sich äquivalent mit $\lambda_j = \lambda_j(x)$ wie folgt formulieren:

$$\begin{pmatrix} | & | & & | \\ a_0 & a_1 & \cdots & a_d \\ | & | & & | \\ 1 & 1 & & 1 \end{pmatrix} \begin{pmatrix} \lambda_0 \\ \lambda_1 \\ \vdots \\ \lambda_d \end{pmatrix} = \begin{pmatrix} x_1 \\ \vdots \\ x_d \\ 1 \end{pmatrix}. \quad (3.6)$$

Aufgrund der Tatsache, dass die Eckpunkte $\{a_j\}_{0 \leq j \leq d}$ affin unabhängig sind, hat die Systemmatrix wegen

$$\text{Rang} \begin{pmatrix} | & | & & | \\ a_0 & a_1 & \cdots & a_d \\ | & | & & | \\ 1 & 1 & & 1 \end{pmatrix} = \text{Rang} \begin{pmatrix} | & | & & | \\ a_0 & a_1 - a_0 & \cdots & a_d - a_0 \\ | & | & & | \\ 1 & 0 & & 0 \end{pmatrix}$$

vollen Rang und das lineare Gleichungssystem (3.6) ist für alle $x = (x_j)_{1 \leq j \leq d}$ eindeutig lösbar. \square

Bemerkung 3.6. Es lässt sich leicht durch Einsetzen überprüfen, dass der Vektor $(\lambda_j(x))_{0 \leq j \leq d}$ aus Satz 3.5 genau die baryzentrischen Koordinaten von x sind. Vor dem Hintergrund, dass die λ_j affine Funktionen bzgl. x sind, gilt

$$K = \{x \in \mathbb{R}^d \mid \lambda_j(x) \in [0, 1] \text{ für alle } 0 \leq j \leq d\},$$

so dass die Seitenflächen von K genau der Schnitt von K mit Hyperflächen $\{x \in \mathbb{R}^d \mid \lambda_j(x) = 0\}$ für $0 \leq j \leq d$ sind.

Wir setzen nun die Simplizes zu einer Triangulierung von Ω zusammen. Vorerst setzen wir voraus, dass Ω polygonal berandet ist, d. h. der Rand $\partial\Omega$ ist eine Vereinigung von $(d-1)$ -Simplizes.

Definition 3.7 (Triangulierung). Sei $K_j \subset \overline{\Omega}$ ein d -Simplex für $1 \leq j \leq N_T$ mit $N_T \in \mathbb{N}$. Dann ist die Menge der Simplizes $\mathcal{T}_h := \{K_j\}_{1 \leq j \leq N_T}$ eine *Triangulierung* von Ω , falls

- (1) $\overline{\Omega} = \bigcup_{j=1}^{N_T} K_j$,
- (2) für alle $1 \leq i, j \leq N_T$ ist der Schnitt $K_i \cap K_j$ entweder leer oder ein m -Simplex mit $0 \leq m \leq d-1$, so dass $K_i \cap K_j$ ein gemeinsamer m -dimensionaler Seitensimplex von K_i und K_j ist.

Außerdem heißt $h := \max_{K \in \mathcal{T}_h} h_K$ die *Gitterweite* von \mathcal{T}_h .

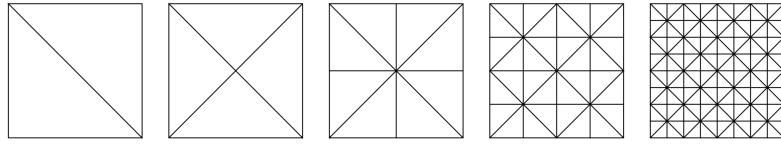


Abbildung 3: Triangulierungen \mathcal{T}_h eines Rechteckes Ω in \mathbb{R}^2 für kleiner werdende Gitterweiten $h > 0$, siehe [Dzi10, Abbildung 3.4].

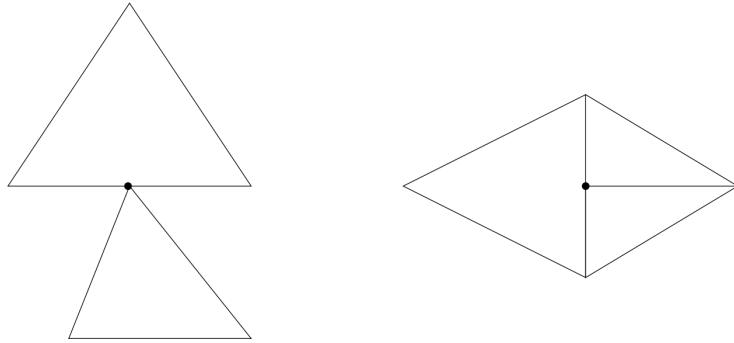


Abbildung 4: Triangulierungen, die in beiden Fällen Definition 3.7(2) verletzen, siehe [AU18, Abbildung 9.11].

In zwei Dimensionen ist der Schnitt zweier Dreiecke (2-Simplex), falls dieser nicht leer ist, entweder eine gemeinsame Ecke oder Kante; siehe Abbildungen 3 und 4. Gleichermaßen in drei Dimensionen ist der Schnitt zweier Tetraeder (3-Simplex) entweder leer oder eine gemeinsame Entität, d. h. eine Ecke, Kante oder Seitenfläche.

Als Nächstes führen wir den Begriff einer regulären Triangulierung ein um hinreichende Approximationseigenschaften von Funktionen über ein trianguliertes Gebiet zu gewinnen.

Definition 3.8. Eine Familie von Triangulierungen $\{\mathcal{T}_h\}_{h>0}$ heißt *regulär*, falls eine Konstante $C > 0$ existiert, so dass für alle $h > 0$ und $K \in \mathcal{T}_h$ gilt, dass $\sigma_K \leq C$ ist.

Bemerkung 3.9. In zwei Dimensionen, d. h. für ein Dreieck K mit kleinstem Innensinkel θ_K , gilt die Abschätzung $\sigma_K \leq \frac{2}{\sin(\theta_K)}$. Somit ist die Bedingung aus Definition 3.8 an \mathcal{T}_h , dass die einzelnen Dreiecke nicht zu spitze Winkel haben können und folglich im Grenzfall $h \rightarrow 0$ nicht entartet sind, siehe [EG04, Bemerkung 1.108(i)].

Gebiete mit krummlinigem Rand. Für nicht-polygonal berandete Gebiete Ω , wie z. B. $\Omega = B_1(0)$, betrachte eine Triangulierung \mathcal{T}_h einer Approximation Ω_h von Ω , so dass die äußeren Eckpunkte von Ω_h auf $\partial\Omega$ liegen. Weiter betrachte polynomiale Transformationen (vom Grad > 1) der Simplizes $K \in \mathcal{T}_h$, die einen nichtleeren Schnitt mit $\partial\Omega$ vorweisen, für eine bessere Approximation des Randes. Ersetze anschließend die Elemente K der Triangulierung durch die transformierten Elemente

\widehat{K} ; vgl. Abbildung 5. Für eine detaillierte Betrachtung, siehe [EG04, S. 35-36]. Zur Vereinfachung betrachten wir im Folgenden Triangulierungen \mathcal{T}_h von Ω , wobei im Falle der nicht-polygonal berandete Gebiete $\Omega \neq \Omega_h$ gilt.

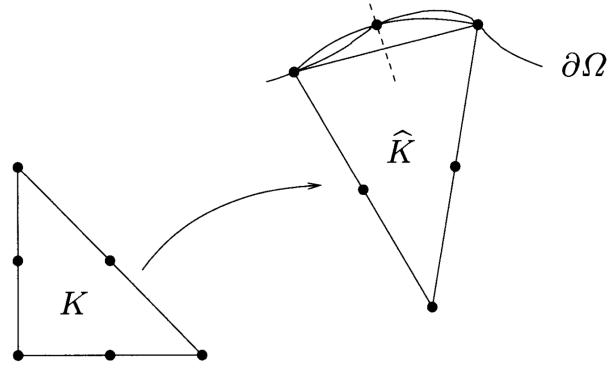


Abbildung 5: Geometrische Konstruktion eines krummlinigen Dreiecks \widehat{K} ausgehend von dem 2-Simplex K , in Anlehnung an [EG04, Abbildung 1.14].

3.3 Simpliziale Lagrange-Elemente

Die in Definition 3.4 eingeführten baryzentrischen Koordinaten eignen sich hervorragend zur Darstellung von Polynomen auf Simplizes. Wir bezeichnen den Raum der Polynome vom Grad höchstens $k \in \mathbb{N}$ mit

$$\mathbb{P}_k(X) := \left\{ p : X \rightarrow \mathbb{R} \mid p(x) = \sum_{|\alpha| \leq k} c_\alpha x^\alpha, c_\alpha \in \mathbb{R} \right\}$$

für eine Menge $X \subseteq \mathbb{R}^d$ in Multiindex-Schreibweise, d. h. für $\alpha = (\alpha_1, \dots, \alpha_d) \in \mathbb{N}_0^d$ ist $x^\alpha := x_1^{\alpha_1} \dots x_d^{\alpha_d}$. Dann ist die Dimension des Vektorraums $\mathbb{P}_k(X)$ gegeben durch

$$\dim \mathbb{P}_k(X) = \binom{d+k}{k} = \begin{cases} k+1 & d=1, \\ \frac{1}{2}(k+1)(k+2) & d=2, \\ \frac{1}{6}(k+1)(k+2)(k+3) & d=3. \end{cases}$$

Des Weiteren führen wir die folgende Menge von Punkten eines Simplizes K ein.

Definition 3.10. Es seien K ein d -Simplex und $k \in \mathbb{N}$. Dann heißt die Menge

$$\mathbb{G}_k(K) := \left\{ x \in K \mid \lambda_j(x) \in \left\{ \frac{i}{k} \mid i = 0, \dots, k \right\} \text{ für alle } 0 \leq j \leq d \right\}$$

das *Lagrange-Gitter* k -ter Ordnung.

Bemerkung 3.11. Für $k = 1$ ist das Gitter genau die Menge der Eckpunkte $\{a_j\}_{0 \leq j \leq d}$ in K und für $k = 2$ besteht die Menge aus Eckpunkten und den zugehörigen Mittelpunkten $\{a_{ij}\}_{0 \leq i \leq j \leq d}$ mit $a_{ij} := \frac{1}{2}(a_i + a_j)$ in K ; siehe Abbildung 6. Im Allgemeinen ist das Gitter $\mathbb{G}_k(K)$ eine endliche Menge von Punkten $\{\sigma_j\}_{1 \leq j \leq N_K}$ in K .

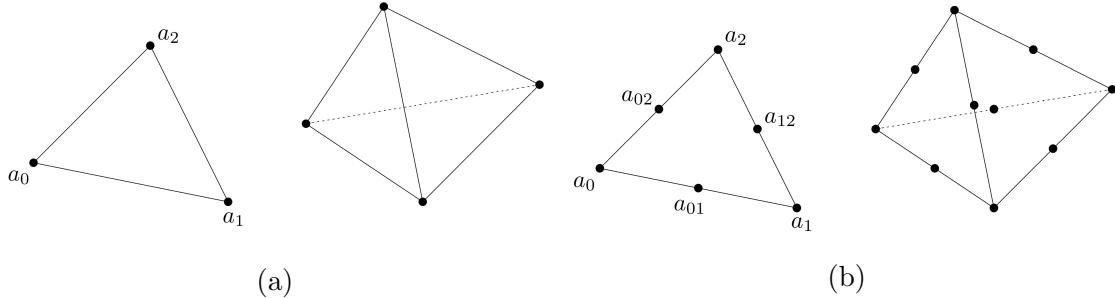


Abbildung 6: Dreiecke und Tetraeder (2- und 3-Simplex) mit zugehörigen Lagrange-Gitter für (a) $k = 1$ und (b) $k = 2$, in Anlehnung an [Joh, Beispiel 10.8].

Unser Ziel ist es jetzt jedem Gitterpunkt σ_j eindeutig eine Basisfunktion φ_j von $\mathbb{P}_k(K)$ zuzuordnen, so dass sich jede Funktion über K als eine Linearkombination dieser Basisfunktionen auf den Gitterpunkten schreiben lassen kann. Vergleiche hierfür Abbildung 7.

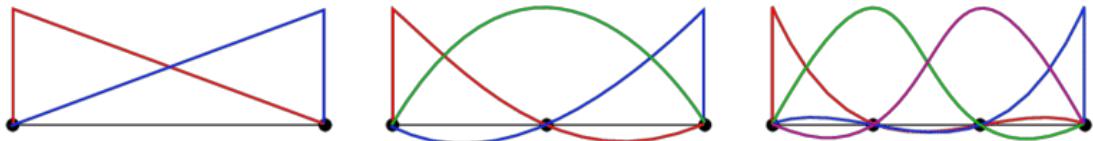


Abbildung 7: Basisfunktionen von $\mathbb{P}_k(K)$ für ein Intervall K zugehörig zu den Gitterpunkten aus $\mathbb{G}_k(K)$ für $k = 1, 2, 3$, siehe [Cor16, Abbildung 4.1].

Satz 3.12. Sei K ein d -Simplex mit Lagrange-Gitter $\mathbb{G}_k(K)$ der Ordnung $k \in \mathbb{N}$ bestehend aus Punkten $\{\sigma_j\}_{1 \leq j \leq N_K}$. Dann ist jedes Polynom $p \in \mathbb{P}_k(K)$ eindeutig durch seine Funktionswerte auf $\{\sigma_j\}_{1 \leq j \leq N_K}$ festgelegt. Ebenfalls gilt $\dim \mathbb{P}_k(K) = |\mathbb{G}_k(K)|$ und es existiert eine Basis $\{\varphi_j\}_{1 \leq j \leq N_K}$ von $\mathbb{P}_k(K)$, so dass

$$\varphi_j(\sigma_i) = \delta_{ij} \quad \text{für alle } 1 \leq i, j \leq N_K.$$

Beweis. Siehe [Dzi10, Hilfssatz 3.18]. □

Bemerkung 3.13.

- (1) Für $k = 1$ hat nach [Dzi10, Hilfssatz 3.16(1)] ein Polynom $p \in \mathbb{P}_1(K)$ die eindeutige Darstellung

$$p(x) = \sum_{j=0}^d p(a_j) \lambda_j(x)$$

für alle $x \in K$. Außerdem gilt $\dim \mathbb{P}_1(K) = |\mathbb{G}_1(K)| = d + 1$.

- (2) Für $k = 2$ hat nach [Dzi10, Hilfssatz 3.17(1)] ein Polynom $p \in \mathbb{P}_2(K)$ die eindeutige Darstellung

$$p(x) = \sum_{j=0}^d p(a_{jj}) \lambda_j(x) (2\lambda_j(x) - 1) + 4 \sum_{j=0}^d \sum_{i=0}^{j-1} p(a_{ij}) \lambda_i(x) \lambda_j(x)$$

für alle $x \in K$. Außerdem gilt $\dim \mathbb{P}_2(K) = |\mathbb{G}_2(K)| = \frac{1}{2}(d+1)(d+2)$.

Um schließlich zu einem endlichdimensionalen Unterraum von V zu gelangen, müssen wir zunächst festlegen, auf welche Weise die lokalen Funktionen in K global in Ω zusammengesetzt werden.

Definition 3.14 (Lagrange-Finite-Elemente-Raum). Sei \mathcal{T}_h eine Triangulierung von Ω . Definiere den *Lagrange-Finite-Elemente-Raum* \mathcal{V}_h^k als den Raum der stetigen, stückweise polynomiellen Funktionen

$$\mathcal{V}_h^k = \{v \in \mathcal{C}^0(\bar{\Omega}) \mid v|_K \in \mathbb{P}_k(K) \text{ für alle } K \in \mathcal{T}_h\}. \quad (3.7)$$

Hinzu definiere das Gitter \mathcal{G}_h^k der Triangulierung, dass eine Vereinigung der Gitter der einzelnen Simplizes ist, d. h.

$$\mathcal{G}_h^k = \bigcup_{K \in \mathcal{T}_h} \mathbb{G}_k(K). \quad (3.8)$$

Im Folgenden sei das Gitter \mathcal{G}_h^k die endliche Menge von Punkten $\{a_j\}_{1 \leq j \leq N}$, wobei N als die Anzahl der *Freiheitsgrade* bezeichnen.

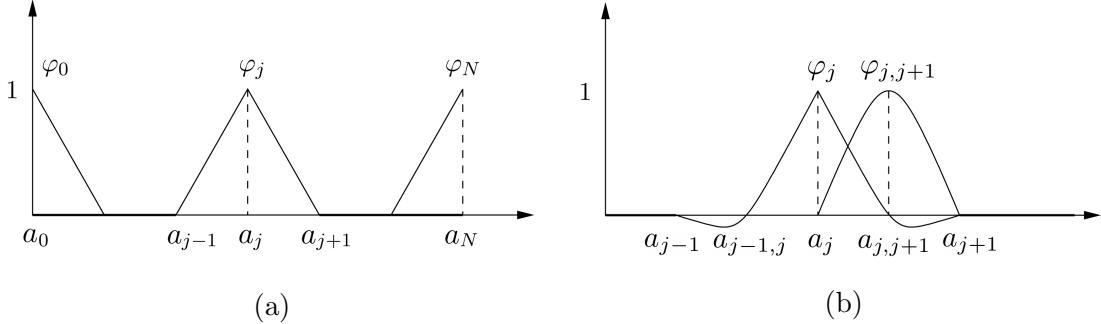


Abbildung 8: Basisfunktionen von \mathcal{V}_h^k für ein Intervall zugehörig zu den Gitterpunkten aus \mathcal{G}_h^k für (a) $k = 1$ und (b) $k = 2$, in Anlehnung an [Fre08, Abbildung 7.1 und 7.3].

In gleicher Weise wie in Satz 3.12 können wir für eine Triangulierung eine eindeutige Korrespondenz zwischen den Gitterpunkten aus \mathcal{G}_h^k und passenden Basisfunktionen von \mathcal{V}_h^k beweisen. Eine Veranschaulichung jener Basisfunktionen ist in Abbildung 8 zu finden.

Satz 3.15. Der Lagrange-Finite-Elemente-Raum \mathcal{V}_h^k ist ein Teilraum von $H^1(\Omega)$. Ebenfalls gilt $\dim \mathcal{V}_h^k = |\mathcal{G}_h^k| = N$ und es existiert eine Basis $\{\varphi_j\}_{1 \leq j \leq N}$ von \mathcal{V}_h^k , so dass

$$\varphi_j(a_i) = \delta_{ij} \quad \text{für alle } 1 \leq i, j \leq N.$$

Folglich hat jede Funktion $v_h \in \mathcal{V}_h^k$ die eindeutige Darstellung

$$v_h = \sum_{j=1}^N v_h(a_j) \varphi_j.$$

Beweis. Siehe [Dzi10, Element 3.19]. □

Finite-Elemente-Raum mit homogenen Dirichlet-Randbedingungen. Um einen endlichdimensionalen Teilraum von V mit $H_0^1(\Omega) \subseteq V \subseteq H^1(\Omega)$ zu erhalten, wähle $\mathcal{V}_h^k \cap V$. Im Fall der homogenen Dirichlet-Randbedingung, d. h. $V = H_0^1(\Omega)$, betrachten wir

$$\mathcal{V}_{0,h}^k := \mathcal{V}_h^k \cap H_0^1(\Omega) = \{v_h \in \mathcal{V}_h^k \mid v_h|_{\partial\Omega} = 0\}.$$

Mit Satz 3.15 hat eine Funktion $v_h \in \mathcal{V}_{0,h}^k$ die eindeutige Darstellung

$$v_h = \sum_{j=1}^N v_h(a_j) \varphi_j = \sum_{a_j \in \partial\Omega} v_h(a_j) \varphi_j + \sum_{a_j \notin \partial\Omega} v_h(a_j) \varphi_j = \sum_{a_j \notin \partial\Omega} v_h(a_j) \varphi_j$$

und folglich hat der Raum $\mathcal{V}_{0,h}^k$ die Basis $\{\varphi_j \mid a_j \notin \partial\Omega, 1 \leq j \leq N\}$ mit $\dim \mathcal{V}_{0,h}^k = |\mathcal{G}_h^k \setminus \partial\Omega|$. Dabei ist der erste Summand gleich 0, da die Funktion v_h am Rand verschwindet. Der Fall der gemischten Randbedingung, d. h. es gilt die homogene Dirichlet-Randbedingung auf einer Teilmenge $\partial\Omega_D \subset \partial\Omega$, funktioniert analog, wohingegen wir hier den Raum $\mathcal{V}_{D,h}^k := \{v_h \in \mathcal{V}_h^k \mid v_h|_{\partial\Omega_D} = 0\}$ betrachten. Im weiteren Verlauf schränken wir uns daher auf den Fall $V_h = \mathcal{V}_h^k$ ein, da die Vorgehensweise identisch für $\mathcal{V}_{0,h}^k$ und $\mathcal{V}_{D,h}^k$ ist.

3.4 Interpolation

In Abschnitt 3.1 haben wir gesehen, dass für das Galerkin-Verfahren und demnach auch für die Finite-Elemente-Methode die abstrakte Fehlerabschätzung

$$\|u - u_h\|_V \leq \frac{C}{\alpha} \inf_{v_h \in V_h} \|u - v_h\|_V$$

gilt. Ist $V_h = \mathcal{V}_h^k$, der Lagrange-Finite-Elemente-Raum, so können wir einen linearen und stetigen Operator $\mathcal{I}_h^k : V \rightarrow \mathcal{V}_h^k$ definieren. Wir bezeichnen \mathcal{I}_h^k als einen *Interpolationsoperator*, falls dieser die Eigenschaft besitzt, dass Funktionen aus \mathcal{V}_h^k invariant bleiben, d. h. für alle $v_h \in \mathcal{V}_h^k$ gilt die Eigenschaft $\mathcal{I}_h^k v_h = v_h$.

Mit der eindeutigen Darstellung von Funktionen in \mathcal{V}_h^k nach Satz 3.15 konstruieren wir \mathcal{I}_h^k , so dass

$$\mathcal{I}_h^k : \mathcal{C}^0(\overline{\Omega}) \rightarrow \mathcal{V}_h^k, \quad v \mapsto \sum_{i=1}^N v(a_i) \varphi_i.$$

Dieser Operator ist linear und stetig und erfüllt offensichtlich die Invarianz von \mathcal{V}_h^k . Nach dem Sobolevschen Einbettungssatz [EG04, Satz B.46(ii)] kann außerdem der Raum $H^{k+1}(\Omega)$ für $k \geq 1$ und $d \leq 3$ als Definitionsbereich des Operators \mathcal{I}_h^k gewählt werden. Der Interpolationsoperator \mathcal{I}_h^k hat die Charakteristik, dass Funktionen punktweise über \mathcal{G}_h^k auf sich selbst abgebildet werden, d. h.

$$(\mathcal{I}_h^k v)(x) = v(x), \quad \text{für alle } x \in \mathcal{G}_h^k$$

und Funktionen $v \in \text{dom } \mathcal{I}_h^k$, d. h. aus dem Definitionsbereich von \mathcal{I}_h^k . Jene Eigenschaft ist in Abbildung 9 illustriert.

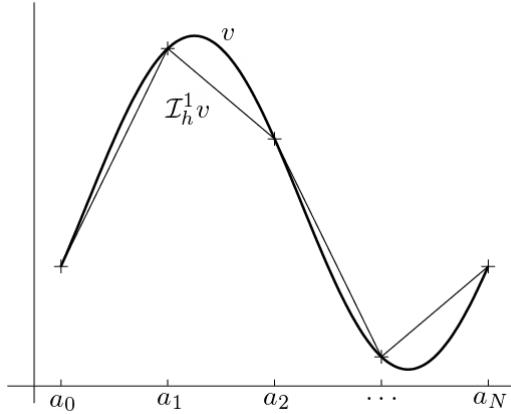


Abbildung 9: Anwendung des Interpolationsoperators \mathcal{I}_h^k für $k = 1$ mit dem Gitter $\mathcal{G}_h^k = \{a_j\}_{0 \leq j \leq N}$ auf eine glatte Funktion $v : [a_0, a_N] \rightarrow \mathbb{R}$, in Anlehnung an [DLM10, Abbildung 1].

Aus der abstrakten Fehlerabschätzung erhalten wir

$$\|u - u_h\|_V \leq \frac{C}{\alpha} \|u - \mathcal{I}_h^k u\|_V.$$

Um also zu einer Fehlerabschätzung für die Finite-Elemente-Methode für simpliziale Lagrange-Elemente zu kommen, benötigen wir eine Abschätzung für den Interpolationsfehler $\|u - \mathcal{I}_h^k u\|_V$. Der nächste Satz liefert genau die gewünschte Schätzung.

Satz 3.16. *Sei $\{\mathcal{T}_h\}_{h>0}$ eine Familie von regulären Triangulierungen von $\Omega \subset \mathbb{R}^d$ mit $d \leq 3$. Sei \mathcal{V}_h^k der Finite-Elemente-Raum vom Grad $k \geq 1$ und \mathcal{I}_h^k der zugehörige*

Interpolationsoperator. Dann existiert eine Konstante $C > 0$, so dass für alle $h > 0$ und $v \in H^{l+1}(\Omega)$ mit $1 \leq l \leq k$ gilt:

$$\|v - \mathcal{I}_h^k v\|_{L^2(\Omega)} + h [v - \mathcal{I}_h^k v]_{H^1(\Omega)} \leq Ch^{l+1} [v]_{H^{l+1}(\Omega)}. \quad (3.9)$$

Des Weiteren gilt für alle $v \in X$, wobei X entweder $H^1(\Omega)$ oder $L^2(\Omega)$ ist, folgendes Dichtheitsresultat:

$$\lim_{h \rightarrow 0} \left(\inf_{v_h \in V_h^k} \|v - v_h\|_X \right) = 0.$$

Beweis. Siehe [EG04, Korollar 1.109, 1.110 und Beispiel 1.111]. \square

Bemerkung 3.17. Die Fehlerabschätzung (3.9) ist optimal, falls v hinreichend glatt ist, d. h. falls $v \in H^{k+1}(\Omega)$. Jedoch sollte v in $H^s(\Omega)$ aber nicht in $H^{s+1}(\Omega)$ für ein $s \geq 2$ liegen, führt eine Erhöhung des Grades k über $s - 1$ hinaus zu keiner Verbesserung des Interpolationsfehlers. Für mehr Details siehe [EG04, Beispiel 1.111 und Abschnitt 3.2.5].

Unter Verwendung des Satzes 3.16 erhalten wir schlussendlich die asymptotische Fehlerabschätzung $\mathcal{O}(h^{k+1})$ bzw. $\mathcal{O}(h^k)$ für den Fehler $u - u_h$ in der H -Norm bzw. V -Norm im Grenzfall $h \rightarrow 0$.

3.5 Ortsdiskretisierung

Nach den Vorbereitungen für die Finite-Elemente-Methode in den Abschnitten 3.1 bis 3.4 gehen wir nun wieder auf das Anfangs-/Randwertproblem (2.7) zurück mit zugehöriger schwacher Formulierung (3.1). Die folgende Diskussion stützt sich dabei auf [EG04, Abschnitte 6.1.4 bis 6.1.6].

Zur Vereinfachung nehmen wir im weiteren Verlauf an, dass die Abbildung $B(\cdot, u, v)$ aus (3.1) stetig in $[0, T]$ ist für alle $u, v \in V$. Des Weiteren setzen wir voraus, dass $F \in \mathcal{C}^0([0, T]; V')$. Sei $\{V_h\}_{h>0}$ eine Familie von endlichdimensionalen Teilräumen von V . Beispielsweise im Falle $V = H^1(\Omega)$ wähle $V_h = \mathcal{V}_h^k$ aus (3.7) für $k \in \mathbb{N}$. Ganz analog zum Vorgehen beim stationären Problem untersuchen wir folgende (semi-)diskrete Problemstellung:

$$\begin{cases} \text{Gesucht ist } u_h \in \mathcal{C}^1([0, T]; V_h), \text{ so dass} \\ \langle \dot{u}_h(t), v_h \rangle_V + B(t, u_h(t), v_h) = \langle F(t), v_h \rangle_V \quad \text{für alle } v_h \in V_h, \quad \text{für alle } t \in (0, T], \\ u_h(0) = u_{0,h}, \end{cases} \quad (3.10)$$

wobei $u_{0,h} \in V_h$ eine Approximation von u_0 ist, auf dessen Konstruktion wir in Bemerkung 3.19(1) genauer eingehen.

Anfangswertproblem in Matrixdarstellung. Sei $\{\varphi_j\}_{1 \leq j \leq N}$ eine Basis von V_h . Dann hat die approximative Lösung $u_h(t) \in V_h$ für alle $t \in [0, T]$ die Darstellung

$$u_h(t) = \sum_{i=1}^N U_i(t) \varphi_i.$$

Für den Lagrange-Finite-Elemente-Raum $V_h = \mathcal{V}_h^k$ mit zugehörigem Gitter $\mathcal{G}_h^k = \{a_j\}_{1 \leq j \leq N}$ für $k \in \mathbb{N}$ gilt außerdem nach Satz 3.15, dass

$$U_i(t) = u_h(t)(a_i) \quad \text{für alle } 1 \leq i \leq N.$$

Eingesetzt in die diskrete Problemstellung (3.10) unter Verwendung der Bilinearität von $B(t, \cdot, \cdot)$ erhalten wir

$$\sum_{i=1}^N \dot{U}_i(t)(\varphi_i, \varphi_j)_H + \sum_{i=1}^N U_i(t)B(t, \varphi_i, \varphi_j) = \langle F(t), \varphi_j \rangle_V \quad \text{für alle } j = 1, \dots, N,$$

mit $U_i(0) = \gamma_i$, wohingegen die approximierte Anfangsbedingung $u_{0,h}$ die Darstellung $u_{0,h} = \sum_{i=1}^N \gamma_i \varphi_i$ besitzt. Es reicht unterdessen aus, sich auf die Basisfunktionen $\{\varphi_j\}_{1 \leq j \leq N}$ von V_h für die Wahl der Testfunktionen v_h zu beschränken. Für $t \in [0, T]$ definieren wir den Koeffizientenvektor $U(t) := (U_i(t))_{1 \leq i \leq N} \in \mathbb{R}^N$ und den Lastvektor $\mathcal{F}(t) := (\langle F(t), \varphi_j \rangle_V)_{1 \leq j \leq N} \in \mathbb{R}^N$. Des Weiteren führen wir die *Massematrix* $\mathcal{M} \in \mathbb{R}^{N \times N}$ und die (zeitabhängige) *Steifigkeitsmatrix* $\mathcal{S}(t) \in \mathbb{R}^{N \times N}$ für alle $t \in [0, T]$ ein, so dass

$$\mathcal{M}_{ij} := (\varphi_i, \varphi_j)_H \quad \text{und} \quad \mathcal{S}_{ij}(t) := B(t, \varphi_i, \varphi_j) \quad \text{für alle } 1 \leq i, j \leq N.$$

Die Massematrix \mathcal{M} ist symmetrisch und positiv definit. Die Symmetrie der Steifigkeitsmatrix \mathcal{S} ist hingegen nur für die in der Arbeit genannten Varianten der Abbildung B garantiert; siehe (2.10), (2.12) und (2.17). Folglich erhalten wir das Anfangswertproblem der Form

$$\begin{cases} \mathcal{M}\dot{U}(t) + \mathcal{S}(t)U(t) = \mathcal{F}(t) & \text{für alle } t \in (0, T], \\ U(0) = U_0, \end{cases} \quad (3.11)$$

während die Anfangsbedingung $U_0 \in \mathbb{R}^N$ von der Form $U_0 := (\gamma_i)_{1 \leq i \leq N}$ ist.

Aufgrund der Symmetrie und positiven Definitheit von \mathcal{M} ist die Existenz der Inversen \mathcal{M}^{-1} sichergestellt und das obige System von gewöhnlichen Differentialgleichungen hat die Form

$$\dot{U}(t) = -\mathcal{M}^{-1}\mathcal{S}(t)U(t) + \mathcal{M}^{-1}\mathcal{F}(t).$$

Demnach ist nach dem Satz von Picard-Lindelöf [SB05, Satz 7.1.1] die Existenz und Eindeutigkeit einer Lösung der diskreten Problemstellung (3.10) bzw. (3.11) gewährleistet.

Unser Ziel ist es jetzt herauszufinden, inwiefern $u_h(t)$ eine gute Approximation von $u(t)$ darstellt. Dafür schränken wir uns jetzt auf den Finite-Elemente-Raum \mathcal{V}_h^k für $k \in \mathbb{N}$ als endlichdimensionalen Teilraum von V ein. Dann ist der Fehler $u - u_h$ zwischen der exakten und semidiskreten Lösung nach folgendem Satz gegeben.

Satz 3.18. *Sei $\{\mathcal{T}_h\}_{h>0}$ eine Familie von regulären Triangulierungen von $\Omega \subset \mathbb{R}^d$ mit $d \leq 3$. Sei \mathcal{V}_h^k der Finite-Elemente-Raum vom Grad $k \in \mathbb{N}$ und \mathcal{I}_h^k der zugehörige Interpolationsoperator. Dann gilt für die Lösung $u \in \mathcal{C}^1([0, T]; W)$ von (3.1) mit $W := H^{k+1}(\Omega) \cap V$ und für die Lösung $u_h \in \mathcal{C}^1([0, T]; \mathcal{V}_h^k)$ von (3.10) bzw. (3.11), dass*

$$\|u - u_h\|_{\mathcal{C}^0([0, T]; H)} \leq \|u_0 - u_{0,h}\|_H e^{-c_P \alpha \frac{T}{2}} + c \left(1 + \frac{1}{\alpha c_P}\right) h^{k+1} \|u\|_{\mathcal{C}^1([0, T]; W)},$$

$$\frac{1}{\sqrt{T}} \|u - u_h\|_{L^2(0, T; V)} \leq \frac{1}{\sqrt{\alpha T}} \|u_0 - u_{0,h}\|_H + c \left(1 + \frac{1}{\alpha \sqrt{c_P}} + \frac{1}{\sqrt{T}}\right) h^k \|u\|_{\mathcal{C}^1([0, T]; W)},$$

mit Konstanten $\alpha, c_P, c > 0$ für alle $h > 0$.

Beweis. Siehe [EG04, Satz 6.14]. □

Bemerkung 3.19.

- (1) Falls $u_0 \in W$, wähle $u_{0,h} = \mathcal{I}_h^k u_0$ als Approximation der Anfangsdaten. Nach Satz 3.16 gilt $\|u_0 - u_{0,h}\|_H \leq Ch^{k+1} \|u_0\|_W$ und die Fehlerabschätzungen aus Satz 3.18 sind optimal, d. h. $\mathcal{O}(h^{k+1})$ in der $\mathcal{C}^0([0, T]; H)$ -Norm bzw. $\mathcal{O}(h^k)$ in der $L^2(0, T; V)$ -Norm für $h \rightarrow 0$.
- (2) Der Fehler, der durch die Approximation der Anfangsdaten entsteht, nimmt exponentiell mit T in der H -Norm ab. In der V -Norm nimmt das Mittel des Fehlers proportional zu $T^{-\frac{1}{2}}$ ab. Demzufolge sind die Anfangsdaten unempfindlich gegenüber Störungen für fortlaufenden Zeiten; eine charakteristische Eigenschaft von parabolischen Gleichungen. Im Speziellen gilt nach [EG04, Bemerkung 6.15(ii)] die Abschätzung

$$\limsup_{t \rightarrow \infty} \left(\|u(t) - u_h(t)\|_H + \frac{h}{\sqrt{t}} \|u - u_h\|_{L^2(0, t; V)} \right) \leq Ch^{k+1} \|u\|_{\mathcal{C}^1([0, \infty), W)}.$$

3.6 Zeitdiskretisierung

Es gibt unzählige Herangehensweise zur Approximation einer Lösung von (3.10) bzw. (3.11). Dessen ungeachtet sind explizite Ein- und Mehrschrittverfahren für diese Systeme von linearen ODEs kaum geeignet. Solche ODEs besitzen eine spezielle Eigenschaft, die als *Steifigkeit* bezeichnet wird. In Bezug darauf heißt ein System von linearen ODEs steif, falls der Realteil des minimalen Eigenwertes der Systemmatrix deutlich kleiner ist als der Realteil des maximalen Eigenwertes. Diese

Eigenschaft führt dazu, dass in expliziten Verfahren wir gezwungen sind deutlich kleinere Schrittweiten verwenden zu müssen gegenüber impliziten Verfahren, siehe [SB05, Abschnitt 7.2.16]. Für Systeme (3.10) bzw. (3.11) ist die Steifigkeit stark von der Qualität der Triangulierung von Ω abhängig und insbesondere für kleine Gitterweiten $h \rightarrow 0$ von Bedeutung, vgl. [EG04, Abschnitt 9.1.4]. Im Rahmen dieser Arbeit untersuchen wir das *implizite* und *explizite Euler-Verfahren*, welche zur Klasse der Einschrittverfahren gehören.

Als Erstes betrachte eine äquidistante Zerlegung des Zeitintervalls $[0, T]$ in endliche Punkte

$$\mathbb{I}_\tau := \{t_m = m\tau \mid m = 0, 1, \dots, M\} \quad \text{mit} \quad \tau := \frac{T}{M}. \quad (3.12)$$

Damit folgt die Darstellung $[0, T] = \bigcup_{m=0}^{M-1} [t_m, t_{m+1}]$. Letztendlich ist das Ziel der Zeitdiskretisierung eine endliche Folge

$$u_{h,\tau} := (u_h^0, u_h^1, \dots, u_h^M) \in V_h^{M+1}$$

zu konstruieren, so dass u_h^m den Funktionswert $u_h(t_m)$ mit u_h aus Satz 3.18 und demzufolge $u(t_m)$ für $m = 0, 1, \dots, M$ approximiert. Infolgedessen bezeichnen wir τ als die *Schrittweite* der Zeitdiskretisierung.

Konvergenztheorie. Für eine detaillierte Untersuchung der Diskretisierung führen wir zuerst folgende Operatoren ein. Sei S der lineare Operator, der einem Paar von Anfangsdaten u_0 und rechte Seiten F die eindeutige Lösung u von (3.1) zuordnet. Wir bezeichnen S als den *exakten Lösungsoperator*. Des Weiteren betrachte die Abbildungen

$$u_0 \longmapsto \rho_{1,h}(u_0) \in V_h \quad \text{und} \quad F \longmapsto \rho_{2,h,\tau}(F) \in V_h^M$$

die endlichdimensionale Approximationen von u_0 und F liefern. Dann definieren wir den *diskreten Lösungsoperator* $S_{h,\tau}$, der einem Paar von approximativen Daten $\rho_{1,h}(u_0)$ und $\rho_{2,h,\tau}(F)$ die Näherungslösung $u_{h,\tau}$ zuordnet. Der Operator $S_{h,\tau}$ hängt hier vom gewählten Schema der Zeitdiskretisierung ab und wir setzen

$$u_{h,\tau} = S_{h,\tau}(\rho_{1,h}(u_0), \rho_{2,h,\tau}(F)).$$

Falls $u_0 \in W$ mit $W := H^{k+1}(\Omega) \cap V$ für $k \in \mathbb{N}$, wähle $\rho_{1,h} = \mathcal{I}_h^k$, den Interpolationsoperator zum Finite-Elemente-Raum V_h^k . Für Daten $F \in \mathcal{C}^0([0, T]; V')$, wähle $\rho_{2,h,\tau}(F) = (\tilde{\mathcal{P}}_h F(t_1), \dots, \tilde{\mathcal{P}}_h F(t_M))$. Dabei ist $\tilde{\mathcal{P}}_h$ die stetige Fortsetzung auf V' der orthogonalen Projektion \mathcal{P}_h von H nach V_h , d. h. für alle $v \in H$ ist $\mathcal{P}_h v$ die Lösung von

$$(\mathcal{P}_h v, \chi)_H = (v, \chi)_H \quad \text{für alle } \chi \in V_h.$$

Wir nehmen im Folgenden jetzt an, dass die Lösung $u = S(u_0, F)$ in $\mathcal{C}^0([0, T]; X)$ liegt, wobei der Banachraum X entweder V oder H ist. Zur Zweckmäßigkeit definieren wir die endliche Folge

$$u_\tau := (u(t_m))_{0 \leq m \leq M} \in X^{M+1},$$

wobei Funktionsauswertungen von u aufgrund der Stetigkeit (in der Zeit) sinnvoll sind. Aufgrund der Inklusionen

$$V_h^{M+1} \subset V^{M+1} \subset H^{M+1}$$

können wir die exakte Lösung u_τ und die diskrete Lösung $u_{h,\tau}$ für passende Normen im Produktraum X^{M+1} miteinander vergleichen und dadurch sinnvoll Fehlerabschätzungen angeben. Für eine endliche Folge $v = (v_0, v_1, \dots, v_M)$ von Funktionen in X^{M+1} definieren wir daher die folgenden diskreten Normen

$$\begin{aligned}\|v\|_{\ell^2([t_m, t_n]; X)}^2 &:= \tau \sum_{l=m}^n \|v_l\|_X^2, \\ \|v\|_{\ell^\infty([t_m, t_n]; X)} &:= \max_{m \leq l \leq n} \|v_l\|_X\end{aligned}$$

für $0 \leq m \leq n \leq M$. In Folgendem sind wir insbesondere daran interessiert die Fehler $\|u_\tau - u_{h,\tau}\|_{\ell^2([t_1, T]; V)}$ und $\|u_\tau - u_{h,\tau}\|_{\ell^\infty([0, T]; H)}$ abzuschätzen.

3.6.1 Implizites Euler-Verfahren

Das implizite Euler-Verfahren liefert eine Konstruktion der diskreten Lösung $u_{h,\tau} = (u_h^0, \dots, u_h^M) \in V_h^{M+1}$ mit gegebenem Startwert $(u_h^0, v_h)_H = (u_0, v_h)_H$ für alle $v_h \in V_h$. Dabei ist die Rekursionsvorschrift für $m = 0, 1, \dots, M-1$ gegeben durch

$$\frac{1}{\tau}(u_h^{m+1} - u_h^m, v_h)_H + B(t_{m+1}, u_h^{m+1}, v_h) = \langle F(t_{m+1}), v_h \rangle_V \quad \text{für alle } v_h \in V_h. \quad (3.13)$$

Schema in Matrixdarstellung. Sei wieder $\{\varphi_j\}_{1 \leq j \leq N}$ eine Basis von V_h . Dann besitzen die Komponenten von $u_{h,\tau}$ die Darstellung

$$u_h^m = \sum_{i=1}^N U_i^m \varphi_i \quad \text{für alle } m = 0, 1, \dots, M$$

mit Koeffizienten $U^m := (U_i^m)_{1 \leq i \leq N} \in \mathbb{R}^N$. Eingesetzt in die Rekursionsvorschrift (3.13) erhalten wir

$$\frac{1}{\tau} \mathcal{M} (U^{m+1} - U^m) + \mathcal{S}(t_{m+1}) U^{m+1} = \mathcal{F}(t_{m+1}).$$

Dies ist äquivalent zum Lösen eines linearen Gleichungssystems in jeden Zeitschritt von der Form

$$(\mathcal{M} + \tau \mathcal{S}(t_{m+1})) U^{m+1} = \mathcal{M} U^m + \tau \mathcal{F}(t_{m+1}).$$

Die Matrix $\mathcal{M} + \tau \mathcal{S}(t_{m+1})$ ist hier für die in der Arbeit genannten Varianten der Abbildung B und hinreichend kleinen Schrittweiten $\tau > 0$ symmetrisch und positiv definit und folglich invertierbar.

Konvergenz und Stabilität. Im Folgenden sind wir daran interessiert das Schema (3.13) genauer zu untersuchen. Dafür betrachten wir den diskreten Lösungsoperator $S_{h,\tau} : V_h \times V_h^M \rightarrow V_h^{M+1}$, so dass für alle $x_h \in V_h$ und $y_{h,\tau} := (y_h^1, \dots, y_h^M) \in V_h^M$ die endliche Folge $z_{h,\tau} := (z_h^0, \dots, z_h^M) = S_{h,\tau}(x_h, y_{h,\tau})$ rekursiv gegeben ist durch $z_h^0 = x_h$ und

$$\frac{1}{\tau}(z_h^{m+1} - z_h^m, v_h)_H + B(t_{m+1}, z_h^{m+1}, v_h) = (y_h^{m+1}, v_h)_H \quad \text{für alle } v_h \in V_h \quad (3.14)$$

und $m = 0, 1, \dots, M-1$. Für $\rho_{1,h}(u_0) = \mathcal{P}_h u_0$ und $\rho_{2,h,\tau}(F) = (\tilde{\mathcal{P}}_h F(t_1), \dots, \tilde{\mathcal{P}}_h F(t_M))$ folgt die Darstellung

$$u_{h,\tau} = S_{h,\tau}(\rho_{1,h}(u_0), \rho_{2,h,\tau}(F)).$$

Der Operator $S_{h,\tau}$ besitzt Stabilitätseigenschaften, d. h. eine stetige Abhängigkeit von den approximativen Daten $(\rho_{1,h}(u_0), \rho_{2,h,\tau}(F))$. Diese sind vergleichbar mit dem Resultat aus Satz 2.5 des exakten Problems (3.1).

Satz 3.20. *Sei $z_{h,\tau} = S_{h,\tau}(x_h, y_{h,\tau})$ wie in (3.14). Falls $\tau \leq \frac{1}{\alpha c_P}$, dann gilt:*

$$\begin{aligned} \|z_{h,\tau}\|_{\ell^\infty([0,T];H)} &\leq \begin{cases} e^{-\frac{1}{4}\alpha c_P T} \|x_h\|_H + \frac{1}{\sqrt{\alpha}} \|y_{h,\tau}\|_{\ell^2([t_1,T];V')}, \\ e^{-\frac{1}{4}\alpha c_P T} \|x_h\|_H + \frac{1}{\alpha \sqrt{c_P}} \|y_{h,\tau}\|_{\ell^\infty([0,T];V')}, \end{cases} \\ \|z_{h,\tau}\|_{\ell^2([t_1,T];V)} &\leq \begin{cases} \frac{1}{\sqrt{\alpha}} \|x_h\|_H + \frac{1}{\alpha} \|y_{h,\tau}\|_{\ell^2([t_1,T];V')}, \\ \frac{1}{\sqrt{\alpha}} \|x_h\|_H + \frac{\sqrt{T}}{\alpha} \|y_{h,\tau}\|_{\ell^\infty([0,T];V')}. \end{cases} \end{aligned}$$

Beweis. Siehe [EG04, Lemma 6.25]. □

Für die Konvergenz des impliziten Euler-Verfahrens gilt folgende Aussage.

Satz 3.21. *Falls die Lösung von (3.1) in $Z := \mathcal{C}^1([0,T];W) \cap \mathcal{C}^2([0;T];V')$ liegt, existiert eine Konstante $C > 0$ unabhängig von h, τ und T , so dass*

$$\begin{aligned} \|u_\tau - u_{h,\tau}\|_{\ell^\infty([0,T];H)} &\leq C(h^{k+1} + \tau) \|u\|_Z, \\ \frac{1}{\sqrt{T}} \|u_\tau - u_{h,\tau}\|_{\ell^2([t_1,T];V)} &\leq C \left(1 + \frac{1}{\sqrt{T}}\right) (h^k + \tau) \|u\|_Z. \end{aligned}$$

Beweis. Siehe [EG04, Satz 6.29]. □

Bemerkung 3.22. Falls $\|u\|_Z$ gleichmäßig beschränkt nach T ist, gilt

$$\begin{aligned} \limsup_{T \rightarrow \infty} \|u(T) - u_h^M\|_H &\leq C(h^{k+1} + \tau) \|u\|_Z, \\ \limsup_{T \rightarrow \infty} \frac{1}{\sqrt{T}} \|u_\tau - u_{h,\tau}\|_{\ell^2([t_1,T];V)} &\leq C(h^k + \tau) \|u\|_Z, \end{aligned}$$

d. h. der Fehler ist gleichmäßig beschränkt für beliebig große Zeiten; vgl. [EG04, Bemerkung 6.30(i)]. Diese Eigenschaft ist charakteristisch für parabolische Gleichungen; die Informationen der Anfangsbedingung sowie die akkumulierten Fehler der vorherigen Iterationsschritte gehen in den Gleichungen unter. Insgesamt wächst der Fehler durch Näherungen für fortlaufende Zeiten nicht.

3.6.2 Explizites Euler-Verfahren

Das explizite Euler-Verfahren liefert eine Konstruktion der diskreten Lösung $u_{h,\tau}$ mit gegebenem Startwert $(u_h^0, v_h)_H = (u_0, v_h)_H$ für alle $v_h \in V_h$. Dabei ist die Rekursionsvorschrift für $m = 0, 1, \dots, M-1$ gegeben durch

$$\frac{1}{\tau}(u_h^{m+1} - u_h^m, v_h)_H + B(t_m, u_h^m, v_h) = \langle F(t_m), v_h \rangle_V \quad \text{für alle } v_h \in V_h.$$

So ist der Unterschied zwischen dem impliziten und expliziten Euler-Verfahren im Schema auf die Wahl des Rückwärts- bzw. Vorwärtsdifferenzenquotienten zur Approximation der Zeitableitung zurückzuführen.

Schema in Matrixdarstellung. Analog zur Vorgehensweise im impliziten Verfahren erhalten wir die folgende Matrixdarstellung der Rekursionsvorschrift für das explizite Verfahren

$$\frac{1}{\tau} \mathcal{M} (U^{m+1} - U^m) + \mathcal{S}(t_m) U^m = \mathcal{F}(t_m)$$

mit äquivalenter Darstellung

$$U^{m+1} = (\mathcal{I} - \tau \mathcal{M}^{-1} \mathcal{S}(t_m)) U^m + \tau \mathcal{M}^{-1} \mathcal{F}(t_m).$$

Konvergenz und Stabilität. Erneut betrachten wir den diskreten Lösungsoperator $S_{h,\tau}$, so dass die endliche Folge $z_{h,\tau} = S_{h,\tau}(x_h, y_{h,\tau})$ rekursiv gegeben ist durch $z_h^0 = x_h$ und

$$\frac{1}{\tau}(z_h^{m+1} - z_h^m, v_h)_H + B(t_m, z_h^m, v_h) = (y_h^m, v_h)_H \quad \text{für alle } v_h \in V_h \quad (3.15)$$

für $m = 0, 1, \dots, M-1$. Für $\rho_{1,h}(u_0) = \mathcal{P}_h u_0$ und $\rho_{2,h,\tau}(F) = (\tilde{\mathcal{P}}_h F(t_1), \dots, \tilde{\mathcal{P}}_h F(t_M))$ folgt die Darstellung $u_{h,\tau} = S_{h,\tau}(\rho_{1,h}(u_0), \rho_{2,h,\tau}(F))$. Für die Untersuchung der Stabilitätseigenschaften des Operators $S_{h,\tau}$ definiere unter anderem

$$c_i(h) := \max_{v_h \in V_h} \frac{\|v_h\|_V}{\|v_h\|_H}.$$

Diese Größe ist endlich aufgrund $\dim V_h < \infty$. Falls $V_h = \mathcal{V}_h^k$ für quasi-uniforme Triangulierungen ist, gilt $c_i(h) \leq ch^{-1}$; siehe [EG04, Abschnitt 1.7].

Satz 3.23. Sei $\kappa \in (0, 1)$. Sei $z_{h,\tau} = S_{h,\tau}(x_h, y_{h,\tau})$ wie in (3.15). Falls $\tau \leq \frac{\kappa\alpha}{\|B\|^2} \frac{1}{c_i(h)^2}$, dann gilt:

$$\begin{aligned}\|z_{h,\tau}\|_{\ell^\infty([0,T];H)} &\leq e^{-\alpha c_P(1-\kappa)T} \|x_h\|_H + \frac{1}{\alpha\sqrt{c_P(1-\kappa)}} \|y_{h,\tau}\|_{\ell^\infty([0,T];V')}, \\ \frac{1}{\sqrt{T}} \|z_{h,\tau}\|_{\ell^2([t_1,T];V)} &\leq \frac{1}{\sqrt{\alpha T(1-\kappa)}} \|x_h\|_H + \frac{1}{\alpha\sqrt{1-\kappa}} \|y_{h,\tau}\|_{\ell^\infty([0,T];V')}.\end{aligned}$$

Beweis. Siehe [EG04, Lemma 6.31]. \square

Für die Konvergenz des expliziten Euler-Verfahrens gilt folgende Aussage.

Satz 3.24. Wir nehmen an, dass $\tau < \frac{\alpha}{\|B\|^2} \frac{1}{c_i(h)^2}$. Falls die Lösung von (3.1) in $Z := \mathcal{C}^1([0, T]; W) \cap \mathcal{C}^2([0; T]; V')$ liegt, existiert eine Konstante $C > 0$ unabhängig von h , τ und T , so dass

$$\begin{aligned}\|u_\tau - u_{h,\tau}\|_{\ell^\infty([0,T];H)} &\leq C (h^{k+1} + \tau) \|u\|_Z, \\ \frac{1}{\sqrt{T}} \|u_\tau - u_{h,\tau}\|_{\ell^2([t_1,T];V)} &\leq C \left(1 + \frac{1}{\sqrt{T}}\right) (h^k + \tau) \|u\|_Z.\end{aligned}$$

Beweis. Siehe [EG04, Satz 6.32]. \square

Zusammenfassend gilt nach Satz 3.21 und 3.24 im Falle des impliziten und expliziten Euler-Verfahrens die asymptotische Fehlerabschätzung $\mathcal{O}(h^{k+1} + \tau)$ in der $\ell^\infty([0, T]; H)$ -Norm bzw. $\mathcal{O}(h^k + \tau)$ in der $\ell^2([t_1, T]; V)$ -Norm für verschwindende Gitter- und Schrittweiten $h \rightarrow 0$ und $\tau \rightarrow 0$. Folglich ist die Konvergenz von Näherungslösungen nach der FEM gegen die exakte Lösung (2.7) gewährleistet. Das explizite Euler-Verfahren weist zwar einen geringeren Rechenaufwand in der Anwendung (Systemmatrix des linearen Gleichungssystems ist von einfacherer Form; im Speziellen symmetrisch, positiv definit und zeitunabhängig) gegenüber dem Impliziten auf. Letztendlich ist jedoch das explizite Verfahren dem Impliziten wegen des beschränkten Stabilitätsgebietes unterlegen. Hierbei müssen in beiden Verfahren hinreichend kleine Schrittweiten τ gewählt werden, wobei im expliziten Verfahren diese Wahl noch von der Gitterweite h abhängt. Diese Feststellung steht im Einklang mit der anfänglichen Aussage, dass explizite Verfahren gegenüber impliziten Verfahren kleinere Schrittweiten τ fordern.

4 Physics-informed neuronale Netzwerke

In diesem Kapitel führen wir *physics-informed neuronale Netzwerke* (engl.: physics-informed neural networks, PINN) ein zum Lösen von PDEs mit zugehörigen Anfangs- und Randbedingungen. Die Idee des Verfahrens besteht darin, die exakte Lösung u durch *künstliche neuronale Netze* (KNN) zu approximieren, d. h. wir konstruieren eine Näherungslösung u_θ der exakten Lösung u für Parameter $\theta \in \Theta$. Passende θ im endlichdimensionalen Parameterraum Θ werden hiernach durch das Minimieren eines speziellen Funktionals, welches eng mit der PDE zusammenhängt, gefunden. Solch ein Ansatz ist vergleichbar mit der Finite-Elemente-Methode, wohingegen wir den Raum der stetigen, stückweise polynomiellen Funktionen durch den Raum der neuronalen Netze ersetzen. Die folgende Diskussion stützt sich dabei auf die Werke [Lu+20] und [Mee19].

4.1 Künstliche neuronale Netze

Als Erstes gehen wir auf den Begriff der künstlichen neuronalen Netze ein. So ist ein KNN eine Vernetzung von *künstlichen Neuronen*, die an den biologischen Neuronen angelehnt sind. Das biologische Neuron empfängt chemische Signale an seinen Eingangskanälen, den sogenannten Dendriten, die es dazu anregen können, einen elektrischen Impuls an seinen Ausgangskanal, das sogenannte Axon, entlang zu schicken. Dieses Signal kann über den Ausgangskanal wiederum die Eingangskanäle anderer Neuronen beeinflussen.

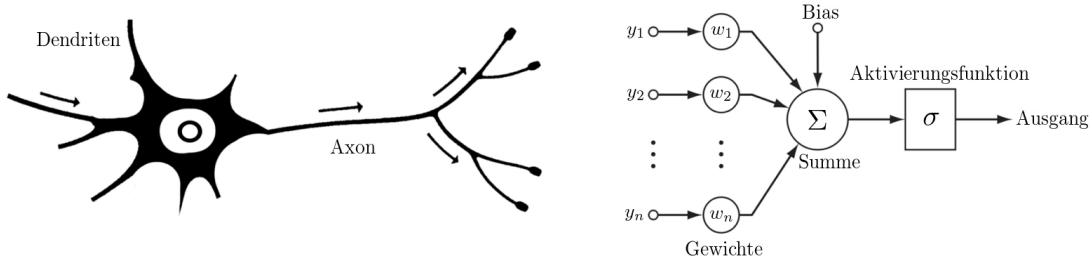


Abbildung 10: Vergleich zwischen dem (links) biologischen und (rechts) künstlichen Neuron, in Anlehnung an [AD18, Abbildung 2].

Das einzelne Neuron. Ein künstliches Neuron stellt eine abstrahierte Version des biologischen Neurons dar und ist eine nichtlineare Abbildung der Form

$$\mathbb{R}^n \ni y \mapsto \sigma(w \cdot y + b) \in \mathbb{R}.$$

Wir bezeichnen $w \in \mathbb{R}^n$ für $n \in \mathbb{N}$ als die *Gewichte* und $b \in \mathbb{R}$ als den *Bias* des Neurons. So können die Gewichte und der Bias als Verbindungsstärken der verschie-

denen Eingangssignale interpretiert werden. Nach anschließender Linearkombination der Gewichte mit der Eingabe $w \cdot y$ und einer Verschiebung durch den Bias b wird das Signal in die *Aktivierungsfunktion* $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ eingesetzt. Die Aktivierungsfunktion ist in diesem Fall ein Modell der Anregung und Hemmung eines Signals im biologischen Neuron. Zur Veranschaulichung der Gemeinsamkeiten und Unterschiede eines biologischen und künstlichen Neurons siehe Abbildung 10.

Typische Aktivierungsfunktionen sind hierbei die Sigmoidfunktionen. Diese sind durch einen S-förmigen Graphen charakterisiert, vgl. Abbildung 11a.

Definition 4.1. Eine *Sigmoidfunktion* $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ ist eine beschränkte und stetig differenzierbare, reelle Funktion mit durchweg positiver oder durchweg negativer ersten Ableitung und genau einem Wendepunkt.

Oft wird der Begriff der Sigmoidfunktion auf den Spezialfall der logistischen Funktion $(1 + e^{-y})^{-1}$ bezogen. Außer der logistischen Funktion enthält die Menge der Sigmoidfunktionen beispielsweise die Tangens-hyperbolicus-Funktion (\tanh) mit $\tanh(y) = (e^y - e^{-y})(e^y + e^{-y})^{-1}$.

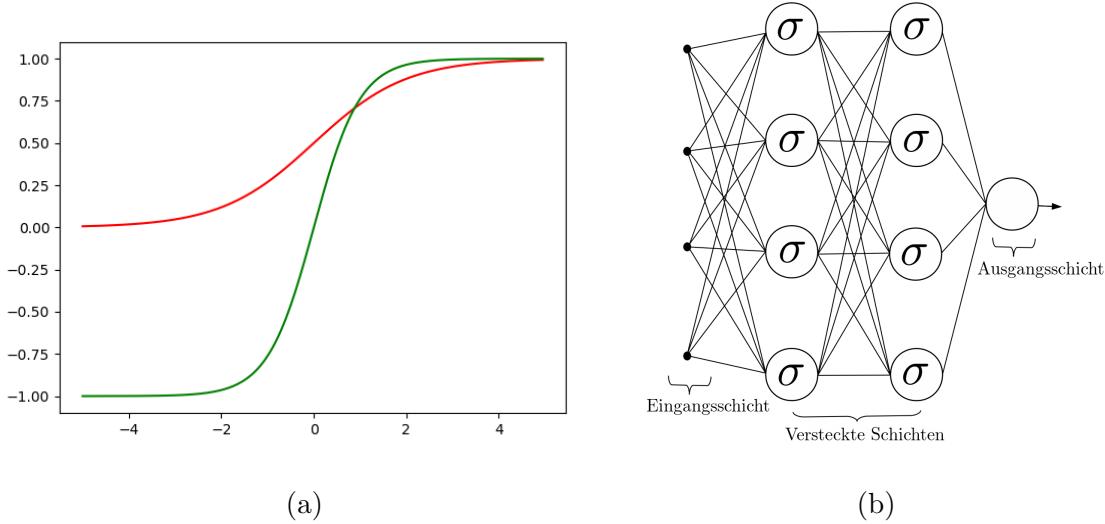


Abbildung 11: (a) Plot der Tangens-hyperbolicus-Funktion (grün) und logistischen Funktion (rot). (b) Veranschaulichung eines MLP als Graph, d. h. die Neuronen sind als Knoten und ihre Verbindungen als Kanten dargestellt.

Das Netzwerk. Die künstlichen Neuronen werden nun in mehreren Schichten konfiguriert und miteinander verbunden. Auf welche Art und Weise diese konfiguriert und verbunden werden bezeichnet man auch als die Topologie oder Struktur des KNes. Im Rahmen dieser Arbeit betrachten wir das *mehrlagige Perzeptron* (MLP), illustriert in Abbildung 11b. Grundsätzlich besitzt ein MLP neben einer Eingangsschicht mit N_0 Neuronen und einer Ausgangsschicht mit N_L Neuronen auch $L - 1$

versteckte Schichten mit je N_l Neuronen für $l = 1, \dots, L - 1$ und $L \geq 2$. Ferner ist zu beachten, dass die Ausgänge der Neuronen mit den Eingängen der nachfolgenden Schicht verknüpft sind, so dass der Informationsfluss nur in einer Richtung verläuft. Hier spricht man auch von *Feedforward-Netzen*. In vektorieller Schreibweise sind die Ausgangssignale der l -ten Schicht $\mathcal{N}^l(y) \in \mathbb{R}^{N_l}$ für Eingangssignale $y \in \mathbb{R}^{N_0}$ gegeben durch

$$\mathcal{N}^l(y) := \begin{cases} y & l = 0, \\ \sigma(W^l \mathcal{N}^{l-1}(y) + b^l) & l = 1, \dots, L - 1, \\ W^L \mathcal{N}^{L-1}(y) + b^L & l = L. \end{cases}$$

Dabei betrachten wir die Gewichte $W^l \in \mathbb{R}^{N_l \times N_{l-1}}$ und Bias $b^l \in \mathbb{R}^{N_l}$ des KNNe für $l = 1, \dots, L$. Des Weiteren merken wir an, dass die Aktivierungsfunktion σ komponentenweise auf das vektorwertige Argument angewendet wird. Abschließend fassen wir die Gewichte und Bias in den Parameter $\theta = \{W^l, b^l\}_{l=1, \dots, L}$ mit dem zugrunde liegenden Parameterraum $\Theta \cong \mathbb{R}^{N_\Theta}$, $N_\Theta := \sum_{l=1}^L N_l(N_{l-1} + 1)$ zusammen.

Unser Ziel ist es, die Näherungslösung des Anfangs- und Randwertproblems (2.7) durch ein MLP zu parametrisieren, so dass für Eingangsdaten $(x, t) \in \bar{\Omega}_T$ und passenden Parameter $\theta \in \Theta$ die Ausgangsgröße

$$u_\theta(x, t) := \mathcal{N}^L(x, t)$$

eine Näherung der exakten Lösung $u(x, t)$ von (2.7) ist. Folglich sind die Anzahl der Neuronen der Eingangs- und Ausgangsschicht des MLP festgelegt durch $N_0 = d + 1$ und $N_L = 1$. Ergo stellen die Funktionen im *Raum der neuronalen Netze*

$$\mathcal{V}_\Theta := \{u_\theta : \bar{\Omega}_T \rightarrow \mathbb{R} \mid \theta \in \Theta\} \quad \text{mit} \quad \dim \mathcal{V}_\Theta = N_\Theta$$

die möglichen Kandidaten der Näherungslösung dar. Man vergleiche den Raum der neuronalen Netze \mathcal{V}_Θ mit dem Produkt des Lagrange-Finite-Elemente-Raumes $\prod_{j=0}^M \mathcal{V}_h^k$ aus (3.7) in Abschnitt 3.6. In PINN ist außerdem eine hinreichende Differenzierbarkeit der Funktionen in \mathcal{V}_Θ nach den Eingangsgrößen (x, t) und den Parametern θ notwendig. Dies wird durch die Wahl einer hinreichend glatten Aktivierungsfunktion σ , wie z. B. $\sigma = \tanh \in \mathcal{C}^\infty(\mathbb{R})$ erreicht.

4.2 Kostenfunktion

Im letzten Abschnitt haben wir den Raum der neuronalen Netze \mathcal{V}_Θ kennengelernt. Jedoch ist noch ungeklärt, inwiefern die Parameter $\theta \in \Theta$ gewählt werden müssen, so

dass $u_\theta \in \mathcal{V}_\Theta$ eine hinreichend gute Approximation der Lösung u von (2.7) darstellt. Hierfür konstruieren wir ein *Kostenfunktional* \mathcal{J} , so dass

$$u = \arg \min_{v \in \text{dom } \mathcal{J}} \mathcal{J}(v).$$

Dementsprechend handelt es sich hier um ein Optimierungsproblem; Verfahren zur Lösung solcher Probleme werden in Abschnitt 4.3 diskutiert. Im Allgemeinen ist es jedoch schwer Aussagen über ein Funktional über den Parameterraum Θ bzw. den Raum der neuronalen Netze \mathcal{V}_Θ zu treffen. Daher betrachten wir zunächst ein Funktional $\mathcal{J} : \mathcal{X} \rightarrow \mathbb{R}$ für passende Banachräume \mathcal{X} . In diesem Zusammenhang sei \mathcal{X} so gewählt, dass das Anfangs- und Randwertproblem unter hinreichend glatten Daten (u_0, f, g) wohlgestellt ist und die Lösung in \mathcal{X} liegt. Man vergleiche $\mathcal{X} \cong \mathcal{W}(V, V')$ mit $H_0^1(\Omega) \subseteq V \subseteq H^1(\Omega)$ je nach Randbedingungen wie in Abschnitten 2.6 und 2.7.

Damit Optimierungsverfahren zum Lösen jener Problemstellungen hinreichend gute Ergebnisse erzielen können, setzen wir folgendes an das Funktional \mathcal{J} voraus:

(L1) Der Minimierer von \mathcal{J} stimmt mit der Lösung u von (2.7) überein, d. h.

$$u = \arg \min_{v \in \mathcal{X}} \mathcal{J}(v).$$

(L2) Das Funktional \mathcal{J} hat genau ein globales Minimum.

Im Folgenden nehme o. B. d. A. an, dass $\mathcal{J}(u) = 0$ gilt. Wenn das nicht der Fall ist, betrachte $\widehat{\mathcal{J}}(\cdot) := \mathcal{J}(\cdot) - \mathcal{J}(u)$.

(L3) Für alle $\epsilon > 0$ existiert ein $\delta > 0$, so dass für alle $v \in \mathcal{X}$ gilt:

$$\mathcal{J}(v) < \delta \implies \|u - v\|_{\mathcal{X}} < \epsilon.$$

(L4) Für alle $\epsilon > 0$ existiert ein $\delta > 0$, so dass für alle $v \in \mathcal{X}$ gilt:

$$\|u - v\|_{\mathcal{X}} > \delta \implies \mathcal{J}(v) > \epsilon.$$

Ohne Bedingung (L2) kann der Fall eintreten, dass Optimierungsverfahren gegen lokale Minimas, die nicht das globale Minimum sind, konvergieren und dort verharren. Folglich führt mit Bedingung (L1) das Minimieren des Funktionales \mathcal{J} stets zur Lösung von (2.7). Die Bedingung (L3) besagt, dass Funktionen v mit hinreichend kleinem Funktionswert $\mathcal{J}(v)$ hinreichend nah an der Lösung u sind. Auf der anderen Seite liefert (L4) dass der Funktionswert $\mathcal{J}(v)$ größer wird, falls Funktionen v sich von der Lösung u entfernen. Des Weiteren ermöglichen uns (L2) und (L4) die Konstruktion eines beschränkten Gebietes in \mathcal{X} , sodass der zurückgelegene Pfad eines

Optimierungsverfahrens vollständig in jenem Gebiet enthalten ist. Solche Pfade terminieren genau im Minimum. Unter den Voraussetzungen (L1) bis (L4) ist also die Konvergenz jener Verfahren mit infinitesimalen Schrittweiten gegen u gewährleistet; vgl. [Mee19, Eigenschaften 3.1 bis 3.4].

Funktional im Banachraum. Da wir nun die notwendigen Bedingungen an das Kostenfunktional diskutiert haben, kommen wir zur Konstruktion eines Kandidaten $\mathcal{J} : \mathcal{X} \rightarrow \mathbb{R}$. Für Parameter $\lambda_{\mathcal{H}}, \lambda_{\mathcal{I}}, \lambda_{\mathcal{B}} > 0$ definiere

$$\mathcal{J}(v) := \lambda_{\mathcal{H}} \mathcal{J}_{\mathcal{H}}(v) + \lambda_{\mathcal{I}} \mathcal{J}_{\mathcal{I}}(v) + \lambda_{\mathcal{B}} \mathcal{J}_{\mathcal{B}}(v), \quad (4.1)$$

für $v \in \mathcal{X}$ mit

$$\begin{aligned} \mathcal{J}_{\mathcal{H}}(v) &:= \|\mathcal{H}(v) - f\|_{L^2(\Omega_T)}^2, \\ \mathcal{J}_{\mathcal{I}}(v) &:= \|v(\cdot, 0) - u_0\|_{L^2(\Omega)}^2, \\ \mathcal{J}_{\mathcal{B}}(v) &:= \|\mathcal{B}(v) - g\|_{L^2(\Gamma_T)}^2. \end{aligned}$$

Hier ist $\mathcal{H} := \partial_t - \Delta$ der lineare, parabolische Differentialoperator zweiter Ordnung der Wärmeleitungsgleichung. Dann gilt:

Satz 4.2. *Das Funktional \mathcal{J} aus (4.1) erfüllt die Voraussetzungen (L1) bis (L4).*

Beweis. Siehe [Mee19, Satz 1 bis 4]. □

Gemischte Randbedingungen. Im Falle der gemischten Randbedingungen, d. h. für eine disjunkte Partitionierung des Randes $\partial\Omega$ in $n \in \mathbb{N}$ Teilgebiete $\partial\Omega_1, \dots, \partial\Omega_n$ und mit $\Gamma_{T,i} := \partial\Omega_i \times (0, T]$ für $i = 1, \dots, n$, gilt für passende Randoperatoren \mathcal{B}_i und rechte Seiten g_i

$$\mathcal{B}(v) - g = \begin{cases} \mathcal{B}_1(v) - g_1 & \text{auf } \Gamma_{T,1}, \\ \dots \\ \mathcal{B}_n(v) - g_n & \text{auf } \Gamma_{T,n}. \end{cases}$$

Aus der Linearität des Lebesgue-Integrals über die Teilgebiete $\Gamma_{T,i}$ folgt

$$\mathcal{J}_{\mathcal{B}}(v) = \|\mathcal{B}(v) - g\|_{L^2(\Gamma_T)}^2 = \sum_{i=1}^n \|\mathcal{B}_i(v) - g_i\|_{L^2(\Gamma_{T,i})}^2.$$

In der Anwendung ist diese Wahl des Kostenfunktional nicht ideal, da die einzelnen Randbedingungen nicht notwendigerweise von gleicher Größenordnung sind. Daher führen wir ähnlich wie bei der Gewichtung der einzelnen Funktionale $\mathcal{J}_{\mathcal{H}}$, $\mathcal{J}_{\mathcal{I}}$ und $\mathcal{J}_{\mathcal{B}}$ noch die Parameter $\lambda_{\mathcal{B},1}, \dots, \lambda_{\mathcal{B},n} > 0$ ein und setzen $\lambda_{\mathcal{B}} = 1$. Demnach wähle im Fall der gemischten Randbedingungen für $\mathcal{J}_{\mathcal{B}}$ das Funktional

$$\mathcal{J}_{\mathcal{B}}(v) = \sum_{i=1}^n \lambda_{\mathcal{B},i} \|\mathcal{B}_i(v) - g_i\|_{L^2(\Gamma_{T,i})}^2.$$

Zum Schluss merken wir noch an, dass die Anfangsbedingung als Dirichlet-Randbedingung auf der Teilmenge $\Omega \times \{0\}$ des parabolischen Randes Σ_T interpretiert und äquivalent als Bestandteil der gemischten Randbedingungen behandelt werden kann.

Funktional im Parameterraum. Im Allgemeinen ist der Banachraum \mathcal{X} unendlichdimensional und somit ungeeignet in der praktischen Anwendung. Infolgedessen schränken wir uns auf den Fall der neuronalen Netze ein, d. h. setze $\mathcal{X} = \mathcal{V}_\Theta$. Dann ist das Funktional \mathcal{L} über dem Parameterraum Θ von der Form

$$\mathcal{L} : \Theta \rightarrow \mathbb{R}, \quad \theta \mapsto \mathcal{J}(u_\theta). \quad (4.2)$$

Auf gleiche Art und Weise setze die Funktionale $\mathcal{L}_H, \mathcal{L}_I, \mathcal{L}_B : \Theta \rightarrow \mathbb{R}$.

Diese Einschränkung auf Funktionen im Raum der neuronalen Netze führt jedoch dazu, dass die Bedingungen (L1) bis (L4) nicht notwendigerweise erfüllt sind. Beispielsweise ist es nicht legitim anzunehmen, dass die Lösung u von (2.7) durch endliche neuronale Netze dargestellt werden kann. In diesem Fall gilt

$$\min_{\theta \in \Theta} \mathcal{L}(\theta) > \mathcal{J}(u) = 0.$$

Monte-Carlo-Integration. Wir bemerken, dass die Summanden im Kostenfunktional \mathcal{L} aus (4.2) von der Form

$$I = \int_A v \, dy$$

sind für integrierbare Funktionen v über beschränkte Gebiete $A \subset \mathbb{R}^n$, $n \in \mathbb{N}$. Für die Berechnung solcher hochdimensionaler Integrale werden im Allgemeinen numerische Integrationsverfahren hinzugezogen. Aus diesem Grund betrachten wir die *Monte-Carlo-Integration*. Hierfür werden $N \in \mathbb{N}$ zufällige Punkte y_1, \dots, y_N gleichverteilt im Integrationsgebiet A erzeugt. Dann ergibt sich eine Näherung des Integrals I als Durchschnitt der Funktionswerte dieser Stellen

$$I \approx I_N := \frac{|A|}{N} \sum_{i=1}^N v(y_i),$$

wobei $|A|$ das Maß des Gebietes A bezeichnet. Der Vorteil ist die vergleichsweise einfache Implementierung sowie die relativ einfache Erweiterbarkeit auf mehrdimensionale Integrale. Hier sind klassische Integrationsalgorithmen stark vom „Fluch der Dimensionalität“ betroffen und für hochdimensionale Probleme nicht mehr praktikabel. Im Vergleich dazu erfüllt die Monte-Carlo-Integration die asymptotische Fehlerabschätzung

$$|I - I_N| = \mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$$

für $N \rightarrow \infty$ unabhängig von der Dimension n ; vgl. [ES00, Abschnitt 6.2].

Wir wenden jetzt die Monte-Carlo-Integration auf \mathcal{L} von (4.2) an. Zu diesem Zweck betrachten wir die Mengen \mathcal{T}_H , \mathcal{T}_I und \mathcal{T}_B von zufälligen Punkten gleichverteilt in Ω_T , Ω und Γ_T . Dann gilt für die *Kostenfunktion* $\mathcal{L}_{\mathcal{T}}$, dass $\mathcal{L}(\theta) \approx \mathcal{L}_{\mathcal{T}}(\theta)$ mit

$$\mathcal{L}_{\mathcal{T}}(\theta) := \lambda_H \mathcal{L}_{H,\mathcal{T}}(\theta) + \lambda_I \mathcal{L}_{I,\mathcal{T}}(\theta) + \lambda_B \mathcal{L}_{B,\mathcal{T}}(\theta), \quad (4.3)$$

für $\theta \in \Theta$ und

$$\begin{aligned}\mathcal{L}_{H,\mathcal{T}}(\theta) &:= \frac{1}{|\mathcal{T}_H|} \sum_{(x_i, t_i) \in \mathcal{T}_H} |\mathcal{H}(u_\theta)(x_i, t_i) - f(x_i, t_i)|^2, \\ \mathcal{L}_{I,\mathcal{T}}(\theta) &:= \frac{1}{|\mathcal{T}_I|} \sum_{x_i \in \mathcal{T}_I} |u_\theta(x_i, 0) - u_0(x_i)|^2, \\ \mathcal{L}_{B,\mathcal{T}}(\theta) &:= \frac{1}{|\mathcal{T}_B|} \sum_{(x_i, t_i) \in \mathcal{T}_B} |\mathcal{B}(u_\theta)(x_i, t_i) - g(x_i, t_i)|^2.\end{aligned}$$

Dabei haben wir die Maße der Integrationsgebiete $|\Omega_T|$, $|\Omega|$ und $|\Gamma_T|$ in den nicht-genauer spezifizierten Parametern λ_H , λ_I und λ_B berücksichtigt. Die Elemente der Mengen \mathcal{T}_H , \mathcal{T}_I und \mathcal{T}_B bezeichnen wir als die *Eingangsdaten* des KNNes.

4.3 Optimierungsverfahren

Im diesem Abschnitt sind wir daran interessiert das Optimierungsproblem

$$\theta_{\mathcal{T}} = \arg \min_{\theta \in \Theta} \mathcal{L}_{\mathcal{T}}(\theta) \quad (4.4)$$

für $\mathcal{L}_{\mathcal{T}}$ von (4.3) zu lösen. Im Rahmen dieser Arbeit schränken wir uns auf die Diskussion der *Adam-Optimierung* ein. Dieses Optimierungsverfahren wird häufig für das Minimieren einer Kostenfunktion zugehörig zu einem KNN verwendet. Hierbei ist Adam ein iteratives Verfahren, so dass $\mathcal{L}_{\mathcal{T}}$ schrittweise minimiert wird, d. h. für die Folge $\{\theta_i\}_{i \in \mathbb{N}_0}$ gegeben durch den Startwert θ_0 und der Iterationsvorschrift

$$\begin{aligned}g_{i+1} &= \nabla \mathcal{L}_{\mathcal{T}}(\theta_i), \\ m_{i+1} &= \beta_1 m_i + (1 - \beta_1) g_{i+1}, \\ v_{i+1} &= \beta_2 v_i + (1 - \beta_2) g_{i+1} \otimes g_{i+1}, \\ \widehat{m}_{i+1} &= \frac{m_{i+1}}{1 - \beta_1^{i+1}}, \\ \widehat{v}_{i+1} &= \frac{v_{i+1}}{1 - \beta_2^{i+1}}, \\ \theta_{i+1} &= \theta_i - \alpha \widehat{m}_{i+1} \oslash \left(\sqrt{\widehat{v}_{i+1}} + \epsilon \right),\end{aligned}$$

für $i = 0, 1, 2, \dots$ mit $m_0 = v_0 = 0$ gilt, dass $\mathcal{L}_{\mathcal{T}}(\theta_k) \approx \mathcal{L}_{\mathcal{T}}(\theta_{\mathcal{T}})$ für hinreichend große $k \in \mathbb{N}_0$ ist. Hier bezeichnen \otimes und \oslash die komponentenweise Multiplikation und Division zweier Vektoren. Des Weiteren sind die Momente $\beta_1, \beta_2 \in [0, 1)$, die

Schrittweite $\alpha > 0$ und $\epsilon > 0$ die Parameter des Optimierungsverfahrens. Im Kontext von KNNen wird die Schrittweite auch als *Lernrate* und der Optimierungsprozess als *Lernprozess* bezeichnet. Für eine detaillierte Diskussion von Adam, siehe [Cal20, Abschnitt 4.7].

Gewichtsinitialisierung. Es bleibt noch die Frage offen inwiefern der Startwert $\theta_0 \in \Theta$ gewählt werden muss. So kann im Falle von betragsmäßig sehr kleinen oder großen Gewichten, der Gradient der Kostenfunktion $\nabla \mathcal{L}_T(\theta)$ für tiefe KNN, d. h. $L \gg 1$, verschwinden (≈ 0) bzw. explodieren ($\gg 0$). Beide Fälle stellen ein großes Problem im Lernprozess dar. Aus diesem Grund betrachte die *Glorot-Initialisierung*. Diese Wahl der Gewichtsinitialisierung führt zu einer Erhaltung der Varianz der verschiedenen Schichten des MLP und folglich dem Gradienten. Hierfür wird der Startwert $\theta_0 = \{W^{0l}, b^{0l}\}_{l=1,\dots,L}$ wie folgt gewählt: Für $l = 1, \dots, L$ werden die Bias-Vektoren b^{0l} konstant (für gewöhnlich gleich 0) gesetzt und die Einträge der Gewichtsmatrizen W_{ij}^{0l} werden gleichverteilt im Intervall

$$\left[-\frac{\sqrt{6}}{\sqrt{N_{l-1} + N_l}}, \frac{\sqrt{6}}{\sqrt{N_{l-1} + N_l}} \right] \quad (4.5)$$

erzeugt. Dabei wurde angenommen, dass die Aktivierungsfunktion σ näherungsweise linear im Bereich der auftretenden Eingangsgrößen und ihre Ableitung σ' in diesem Bereich näherungsweise 1 ist.

Für flache KNN, d. h. kleinen $L \geq 2$, ist die Problematik von verschwindenden und explodierenden Gradienten weniger bedeutend. Demungeachtet soll die Initialisierung die Wahl der Aktivierungsfunktion berücksichtigen, da diese von sehr großer Bedeutung für die Güte der Approximation von (2.7) ist. Im Fall $\sigma = \tanh$, betrachte die Gleichverteilung im Intervall

$$\left[-\frac{\sqrt{6}}{\sqrt{N_{l-1} \tanh 2}}, \frac{\sqrt{6}}{\sqrt{N_{l-1} \tanh 2}} \right] \quad (4.6)$$

anstelle von (4.5). Für mehr Details zur Gewichtsinitialisierung, siehe [Cal20, Abschnitt 6.3] und [Mee19, Abschnitt 3.2.4].

4.4 Implementierung

Wir fassen nun die einzelnen Schritte des Algorithmus zusammen; vgl. Abbildung 12. Für eine gegebene Netzwerkstruktur, d. h. Anzahl an Schichten $L \geq 2$ mit Anzahl an Neuronen $N_1, \dots, N_{L-1} \in \mathbb{N}$ und der Aktivierungsfunktion $\sigma : \mathbb{R} \rightarrow \mathbb{R}$, initialisiere die Startgewichte und -bias $\theta_0 = \{W^{0l}, b^{0l}\}_{l=1,\dots,L}$ wie in Abschnitt 4.3 nach (4.5) bzw. (4.6). Zudem wähle Eingangsdaten $\mathcal{T}_H, \mathcal{T}_I$ und \mathcal{T}_B gleichverteilt in den zugehörigen Integrationsgebieten und wähle passende Gewichte $\lambda_H, \lambda_I, \lambda_B > 0$ der

einzelnen Terme der Kostenfunktion $\mathcal{L}_{\mathcal{T}}$ von (4.3). Ausgehend von θ_i berechne θ_{i+1} für $i = 0, 1, 2, \dots$ nach der Iterationsvorschrift der Adam-Optimierung in Abschnitt 4.3. Hierbei lassen wir das Optimierungsverfahren terminieren, falls entweder keine nennenswerte Verbesserung der Kostenfunktion in nachfolgenden Iterationsschritten auftreten oder falls die Kostenfunktion eine gegebene Toleranz unterschreitet. Außerdem soll bei der Implementierung eine maximale Anzahl an Iterationsschritten i_{\max} berücksichtigt werden. Schlussendlich terminiert der Algorithmus in θ^* in einer hinreichend kleinen Umgebung von $\theta_{\mathcal{T}}$ aus (4.4). Folglich stellt u_{θ^*} die Approximation der exakten Lösung u von (2.7) durch neuronale Netze dar.

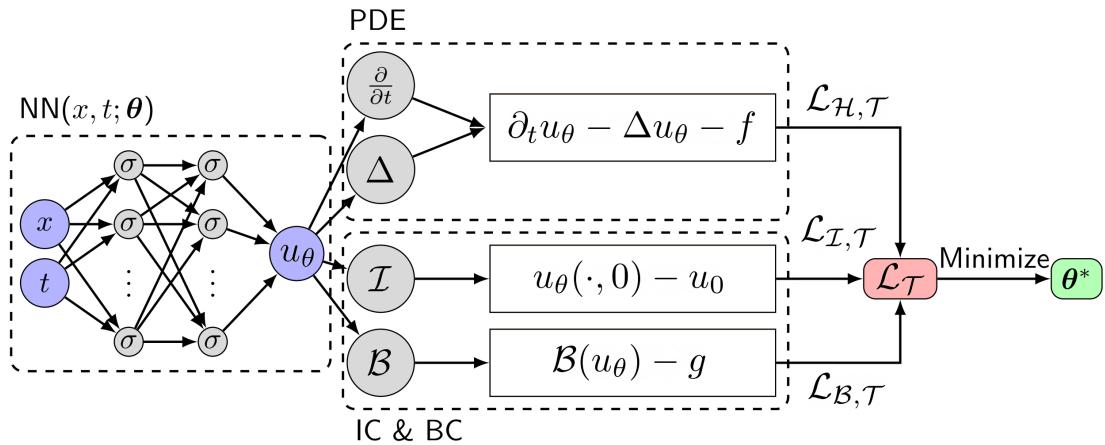


Abbildung 12: Schema von PINN für das Lösen von (2.7), in Anlehnung an [Lu+20, Abbildung 1].

4.5 Fehlerdiskussion

In den letzten Abschnitten ist es uns gelungen ein Verfahren zur Berechnung möglicher Näherungslösungen von (2.7) durch neuronale Netze zu konstruieren. Es bleibt aber noch die Diskussion des Fehlers zwischen der exakten Lösung u und der Näherungslösung offen. Als Erstes wollen wir dabei ein wichtiges Resultat für KNNe mit einer versteckten Schicht ($L = 2$) nennen. Hierfür führen wir zuerst folgende Notation ein. Für einen Multiindex $k = (k_1, \dots, k_n) \in \mathbb{N}_0^n$ mit $n \in \mathbb{N}$, gilt $v \in \mathcal{C}^k(\mathbb{R}^n)$ falls $\partial^\alpha v \in \mathcal{C}^0(\mathbb{R}^n)$ für alle $\alpha \leq k$ mit $\alpha = (\alpha_1, \dots, \alpha_n) \in \mathbb{N}_0^n$, d. h. $\alpha_i \leq k_i$ für alle $i = 1, \dots, n$.

Satz 4.3 (Universelles Approximationstheorem). Seien $k_i \in \mathbb{N}_0^n$ für $i = 1, \dots, s$ mit $n, s \in \mathbb{N}$ und setze $\kappa = \max_{i=1, \dots, s} |k_i|$. Des Weiteren sei $\sigma \in \mathcal{C}^\kappa(\mathbb{R})$ und σ kein Polynom. Dann ist der Raum der neuronalen Netze mit einer versteckten Schicht

$$\mathcal{M}(\sigma) := \text{span}\{\sigma(w \cdot y + b) \mid w \in \mathbb{R}^n, b \in \mathbb{R}\}$$

dicht in

$$\mathcal{C}^{k_1, \dots, k_s}(\mathbb{R}^n) := \bigcap_{i=1}^s \mathcal{C}^{k_i}(\mathbb{R}^n),$$

d. h. für alle $\epsilon > 0$, $v \in \mathcal{C}^{k_1, \dots, k_s}(\mathbb{R}^n)$ und kompakte Teilmengen $K \subset \mathbb{R}^d$ existiert ein $w \in \mathcal{M}(\sigma)$, so dass

$$\max_{y \in K} |\partial^\alpha v(y) - \partial^\alpha w(y)| < \epsilon,$$

für alle $\alpha \in \mathbb{N}_0^n$ mit $\alpha \leq k_i$ für ein $i = 1, \dots, s$.

Beweis. Siehe [Lu+20, Satz 2.1]. \square

Diese Eigenschaft liefert uns die Aussage, dass beliebige, glatte Funktionen hinreichend nah durch flache KNNe approximiert werden können. Dass dieses Resultat auch für die Ableitungen der Funktionen gilt, ist insbesondere für Lösungen von Differentialgleichungen bedeutsam. Nichtsdestotrotz haben wir in der Diskussion des Kostenfunktionalen im Parameterraum bereits festgestellt, dass die Lösung u von (2.7) im Allgemeinen nicht durch ein neuronales Netz beschrieben werden kann. Im Folgenden betrachte daher die bestmögliche Approximation u_{θ_N} von u durch ein KNN mit $\theta_N := \arg \min_{\theta \in \Theta} \|u - u_\theta\|_{\mathcal{X}}$. Des Weiteren betrachten wir in der Anwendung die Näherung $\mathcal{L}_{\mathcal{T}}$ des Kostenfunktionalen \mathcal{L} , d. h. wir trainieren das KNN nur mit endlich vielen Trainingsdaten. Ergo ist u_{θ_T} aus (4.4) die bestmögliche Näherung von u , die durch das Minimieren von $\mathcal{L}_{\mathcal{T}}$ erreicht werden kann. Wir stellen aber fest, dass Optimierungsverfahren im Allgemeinen lediglich näherungsweise das Minimum liefern, d. h. das Verfahren terminiert bei θ^* in einer hinreichend kleinen Umgebung von θ_T wie in Abschnitt 4.4. Zur Vereinfachung nehmen wir an, dass u_{θ_N} und u_{θ_T} wohldefiniert und eindeutig sind. Dann können wir den Fehler \mathcal{E} zwischen u und u_{θ^*} wie folgt zerlegen:

$$\mathcal{E} := \|u - u_{\theta^*}\|_{\mathcal{X}} \leq \underbrace{\|u - u_{\theta_N}\|_{\mathcal{X}}}_{\mathcal{E}_{app}} + \underbrace{\|u_{\theta_N} - u_{\theta_T}\|_{\mathcal{X}}}_{\mathcal{E}_{gen}} + \underbrace{\|u_{\theta_T} - u_{\theta^*}\|_{\mathcal{X}}}_{\mathcal{E}_{opt}}.$$

Dann misst der Approximationsfehler \mathcal{E}_{app} die Güte der Approximation der Lösung durch neuronale Netze. Der Generalisierungsfehler \mathcal{E}_{gen} ist festgelegt durch die Anzahl und Lage der Eingangsdaten \mathcal{T}_H , \mathcal{T}_B und \mathcal{T}_I und gibt die Generalisierungsfähigkeit des Algorithmus auf unbekannte Eingangsdaten aus $\Omega_T \setminus \mathcal{T}_H$, $\Gamma_T \setminus \mathcal{T}_B$ und $\Omega \setminus \mathcal{T}_I$ an. Große und komplexe KNN führen hier zu kleinem Fehler \mathcal{E}_{app} , wobei es aber zu einem großen Fehler \mathcal{E}_{gen} führen kann. Diese Problematik bezeichnet man auch als das *Bias-Varianz-Dilemma*. Falls \mathcal{E}_{gen} gegenüber \mathcal{E}_{app} dominiert, spricht man auch von *Overfitting*. Hinzu kommt der Optimierungsfehler \mathcal{E}_{opt} abhängig von der Komplexität von $\mathcal{L}_{\mathcal{T}}$, den gewählten Parametern des Optimierungsverfahrens und der Anzahl an Iterationsschritten. Letztendlich ist jedoch die genaue Quantifizierung der einzelnen Fehlerterme noch ein offenes Problem im Bereich der KNNe; vgl. [Lu+20, Abschnitt 2.4].

5 Numerische Experimente

In diesem Kapitel wenden wir die beiden in der Arbeit diskutierten Verfahren zur Lösung von (2.7), die FEM und PINN, auf zwei Testprobleme an. Hierfür definieren wir passende Fehlermaße, um Aussagen über die Güte der Näherungen beider Verfahren treffen zu können. Zuvor aber erinnern wir uns an die Unterschiede beider Verfahren in ihrer Funktionsweise.

Vergleich zwischen FEM und PINN.

- Bei der FEM wird die exakte Lösung durch eine Linearkombination von stetigen, stückweise polynomiellen Funktionen mit unbekannten Koeffizienten genähert. Währenddessen ist für PINN die Näherung ein neuronales Netz parametrisiert durch die Gewichte und Bias.
- Die FEM ist gitterbasiert, d. h. die Problemstellung wird auf einer diskreten Untermenge des parabolischen Zylinders basierend auf einer Triangulierung gelöst. Auf der anderen Seite können PINN sowohl auf einem Gitter als auch auf zufälligen Punkten im Zylinder ausgewertet werden.
- Die FEM wandelt die Problemstellung mithilfe der schwachen Formulierung in ein lineares Gleichungssystem bestehend aus der Masse- und Steifigkeitsmatrix und dem Lastvektor um. PINN hingegen betten die Informationen aus der Problemstellung in die Kostenfunktion ein.
- Im letzten Schritt werden die Gleichungssysteme in der FEM durch lineare Solver gelöst, während die Gewichte und Bias in PINN durch gradienten-basierte Optimierer erlernt werden.

Im Gesamten liefert die FEM eine lineare bzw. polynomische Approximation, wo hingegen PINN nichtlineare Approximationen der Lösung und ihrer Ableitungen bereitstellen, vgl. [Lu+20, Abschnitt 2.5].

Fehlermaß. Um im Folgenden die Näherungslösungen der FEM und PINN miteinander vergleichen zu können, benötigen wir kompatible Fehlermaße. Jedoch stellen wir fest, dass die Näherungslösungen beider Verfahren nicht im selben „Raum“ liegen. Wohingegen PINN eine Funktion über den parabolischen Zylinder $\bar{\Omega}_T$ liefert, erhalten wir durch die FEM eine endliche Folge von Funktionen in \mathcal{V}_h^k aus (3.7) für die einzelnen Zeitschritte in \mathbb{I}_τ aus (3.12). Doch nach Satz 3.15 sind Funktionen in \mathcal{V}_h^k eindeutig durch Funktionsauswertungen auf dem Gitter \mathcal{G}_h^k aus (3.8) festgelegt. Aus diesem Grund ist die Näherungslösung der FEM eindeutig durch die Funktionswerte auf der Punktmenge $\mathcal{G}_h^k \times \mathbb{I}_\tau \subset \bar{\Omega}_T$ definiert. Folglich ist es sinnvoll die

Näherungslösungen der FEM und PINN mit der exakten Lösung punktweise über $\mathcal{G}_h^k \times \mathbb{I}_\tau$ zu vergleichen.

Seien $u_{h,\tau} = (u_h^0, u_h^1, \dots, u_h^M) \in \prod_{j=0}^M \mathcal{V}_h^k$ aus Abschnitt 3.6 die Näherungslösung von (2.7) nach der FEM und u_θ aus Abschnitt 4.1 die Näherungslösung von (2.7) nach PINN für Gewichte und Bias $\theta = \{W^l, b^l\}_{l=1,\dots,L}$. Wir betrachten den punktweisen Fehler im Ort zu jedem Zeitpunkt aus $\mathbb{I}_\tau = \{t_0, t_1, \dots, t_M\}$, d. h. für alle $j = 0, 1, \dots, M$ und $p \in [1, \infty]$ definiere die Fehler

$$E_{\text{FEM}}(t_j) := \begin{cases} \left(\sum_{x_i \in \mathcal{G}_h^k} |u(x_i, t_j) - u_h^j(x_i)|^p \right)^{\frac{1}{p}} & p < \infty, \\ \max_{x_i \in \mathcal{G}_h^k} |u(x_i, t_j) - u_h^j(x_i)| & p = \infty, \end{cases}$$

$$E_{\text{PINN}}(t_j) := \begin{cases} \left(\sum_{x_i \in \mathcal{G}_h^k} |u(x_i, t_j) - u_\theta(x_i, t_j)|^p \right)^{\frac{1}{p}} & p < \infty, \\ \max_{x_i \in \mathcal{G}_h^k} |u(x_i, t_j) - u_\theta(x_i, t_j)| & p = \infty. \end{cases}$$

Es gilt, falls $E_{\text{FEM}} = 0$ auf \mathbb{I}_τ , dass $u(\cdot, t_j) = u_h^j$ auf \mathcal{G}_h^k für alle $j = 0, 1, \dots, M$ ist. Analog folgt aus $E_{\text{PINN}} = 0$ auf \mathbb{I}_τ , dass $u = u_\theta$ auf $\mathcal{G}_h^k \times \mathbb{I}_\tau$ gilt. Im Folgenden sind wir insbesondere an den Fällen $p = 2$ und $p = \infty$ interessiert.

Implementierung. Für die Berechnung der Näherungslösungen mithilfe der FEM nutzen wir die Python-Bibliothek **FEniCS**¹ und für PINN die Bibliothek **DeepXDE**². Außerdem werden noch weitere Bibliotheken für das wissenschaftliche Rechnen wie **NumPy**³ verwendet. Der zugehörige Code ist dabei dokumentiert auf

https://github.com/daniellarin22/heat_equation_solver

und dem beliegenden Datenträger zu finden. Die Berechnungen wurden hierbei auf einem Computer mit einer AMD Ryzen 7 5800X CPU, 32GB RAM und einer NVIDIA RTX 3060 Ti GPU (zur Beschleunigung für KNN, siehe NVIDIA cuDNN) ermittelt.

5.1 1D-Testproblem

Wir betrachten in diesem Abschnitt die homogene, eindimensionale Wärmeleitungsgleichung auf dem Intervall $\Omega = (0, 1)$ mit Endzeitpunkt $T = 0.1$

$$\partial_t u - \partial_x^2 u = 0 \quad \text{in } \Omega \times (0, T], \tag{5.1a}$$

$$u(\cdot, 0) = u_0 \quad \text{in } \Omega, \tag{5.1b}$$

$$u(0, \cdot) = u(1, \cdot) = 0 \quad \text{in } (0, T], \tag{5.1c}$$

¹Doku in <https://fenicsproject.org/pub/tutorial/pdf/fenics-tutorial-vol1.pdf>

²Doku in <https://deepxde.readthedocs.io/en/latest/index.html>

³Doku in <https://numpy.org/doc/stable/>

mit gegebener Anfangsbedingung $u_0(x) = \sin(2\pi x)$. Dann besitzt (5.1) nach Satz 2.7 eine eindeutige Lösung gegeben durch

$$u(x, t) = e^{-4\pi^2 t} \sin(2\pi x).$$

Sei $u_{h,\tau} = (u_h^0, u_h^1, \dots, u_h^M)$ die Näherungslösung von (5.1) nach der FEM, wobei wir uns auf die Verwendung des impliziten Euler-Verfahrens einschränken. Außerdem sind die Parameter der beiden Verfahren in Tabelle 1 gegeben.

Tabelle 1: Parameter der FEM für 1D-Testproblem

Grad der Lagrange-Elemente	$k = 2$
Gitterweite	$h = 0.02$
Zeitschrittweite	$\tau = 0.001$

Im Vergleich dazu sei u_θ die Näherungslösung von (5.1) nach PINN für Gewichte und Bias $\theta = \{W^l, b^l\}_{l=1, \dots, L}$ hinreichend nahe am Minimum der Kostenfunktion \mathcal{L}_τ aus (4.3). Hier sind die Parameter des Verfahrens (nicht zu verwechseln mit den Gewichten und Bias) in Tabelle 2 gegeben. Passende Verfahrensparameter werden hingegen mithilfe einer Rastersuche ermittelt, d. h. verschiedene Kombinationen von Werten werden ausprobiert und auf ihre Performance (Güte der Approximation) gemessen.

Tabelle 2: Parameter von PINN für 1D-Testproblem

Netzwerkstruktur	
Anzahl der Schichten	$L = 5$
Neuronen in jeder Schicht	$N = N_1 = \dots = N_{L-1} = 12$
Aktivierungsfunktion	$\sigma = \tanh$
Kostenfunktion	
Gewichte der Kostenfunktion	$\lambda_H = \lambda_I = \lambda_B = 1$
Anzahl der Trainingsdaten in \mathcal{T}_H	1000
Anzahl der Trainingsdaten in \mathcal{T}_I	1000
Anzahl der Trainingsdaten in \mathcal{T}_B	1000
Optimierung	
Verfahren	Adam
Gewichtsinitialisierung	Glorot
Lernrate	$\alpha = 0.001$
Anzahl der Iterationen	75000

Abbildung 13 zeigt den Lernprozess für PINN. So ist erkennbar, dass die Anfangs- und Randdaten nur näherungsweise erfüllt werden, da die zugehörigen Terme der

Kostenfunktion ungleich Null sind. Vergleichen wir dies mit der Näherungslösung nach der FEM, stellen wir fest, dass diese Daten bis auf Maschinengenauigkeit exakt erfüllt werden. Dieser Sachverhalt wird bestätigt im Langzeitverhalten des Fehlers der beiden Näherungslösungen in Abbildung 14. In der Tat ist der Fehler der FEM für jeden Zeitpunkt, insbesondere für den Anfangszeitpunkt, kleiner als der Fehler von PINN. Außerdem ist die sichtbare Abweichung der Anfangs- und Randbedingungen der Näherungslösung nach PINN in den Abbildungen 15 bis 17 erkennbar. Abschließend darf nicht unerwähnt bleiben, dass die Rechenzeit von PINN mit 287.41 s die von der FEM mit 0.15 s bei weitem übersteigt und folglich in der praktischen Anwendung der FEM deutlich unterlegen ist.

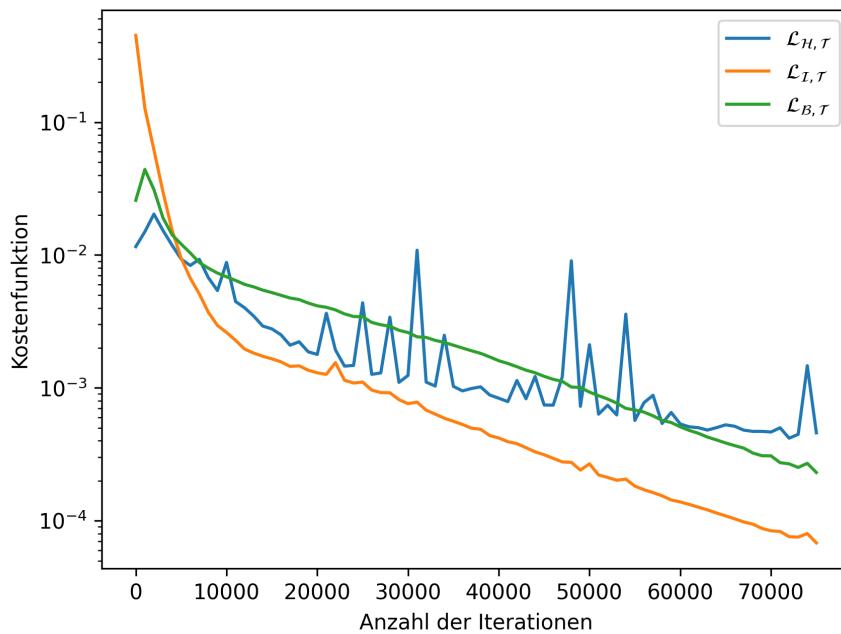
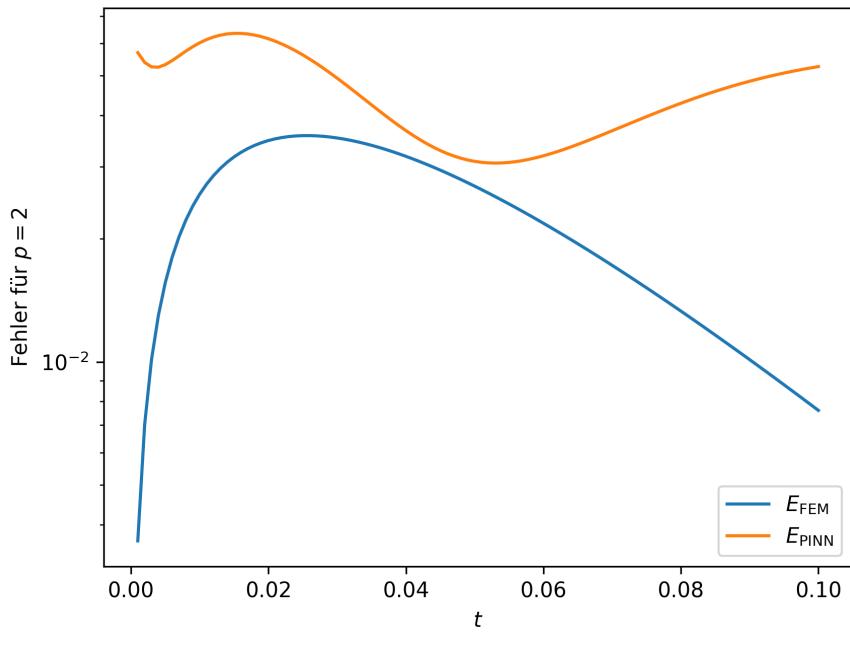
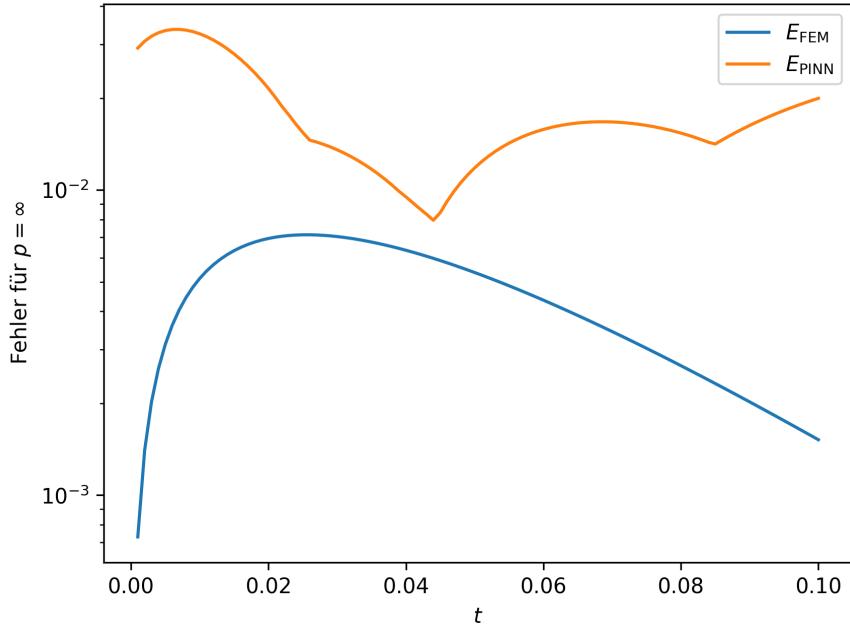


Abbildung 13: Plot der einzelnen Komponenten der Kostenfunktion (4.3) von PINN für jeden Iterationsschritt des Optimierungsverfahrens.



(a)



(b)

Abbildung 14: Fehlerplots von FEM und PINN in Abhängigkeit von der Zeit für das 1D-Testproblem (5.1) für (a) $p = 2$ und (b) $p = \infty$.

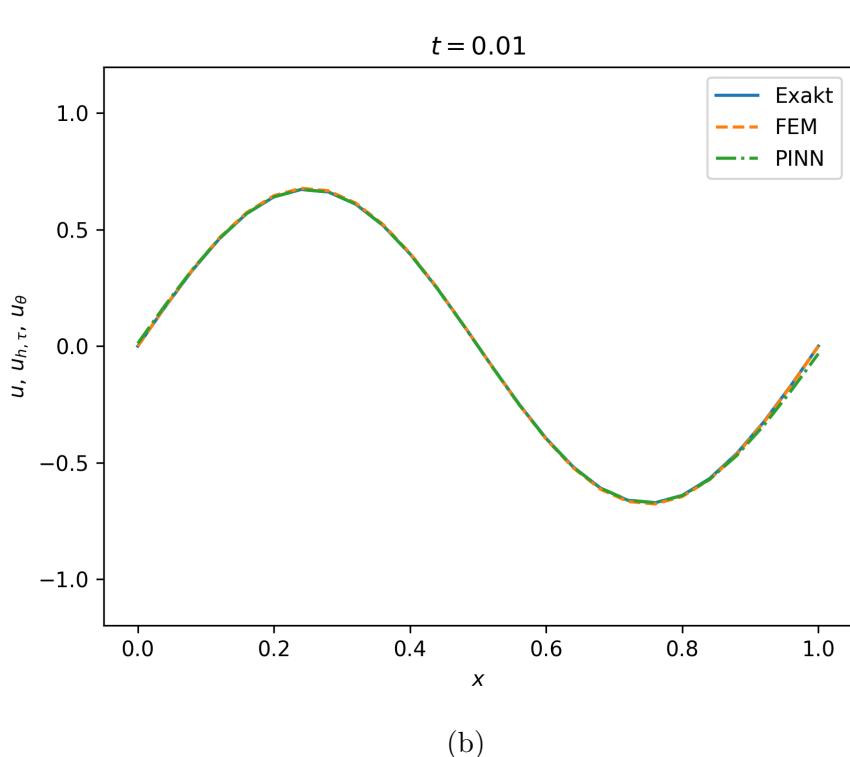
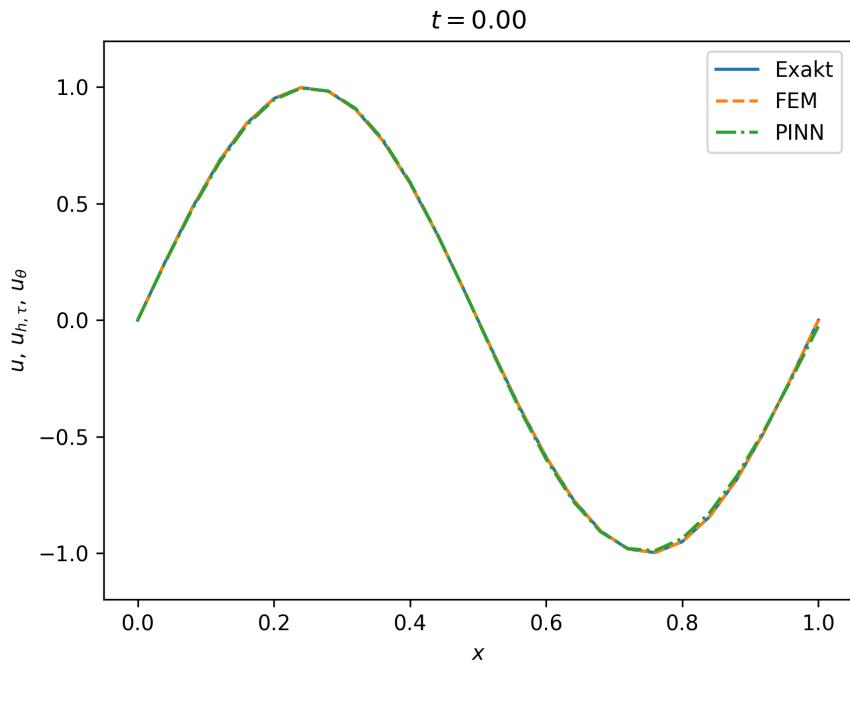


Abbildung 15: Plot der exakten Lösung und der Näherungslösungen der FEM und PINN für feste Zeiten (a) $t = 0$ und (b) $t = 0.01$ des 1D-Testproblems (5.1).

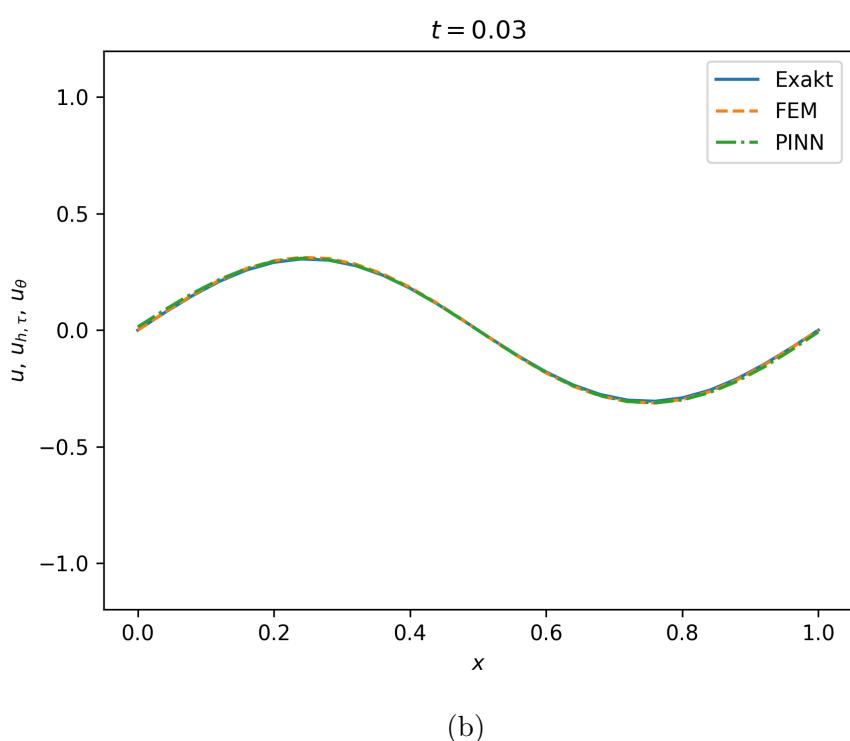
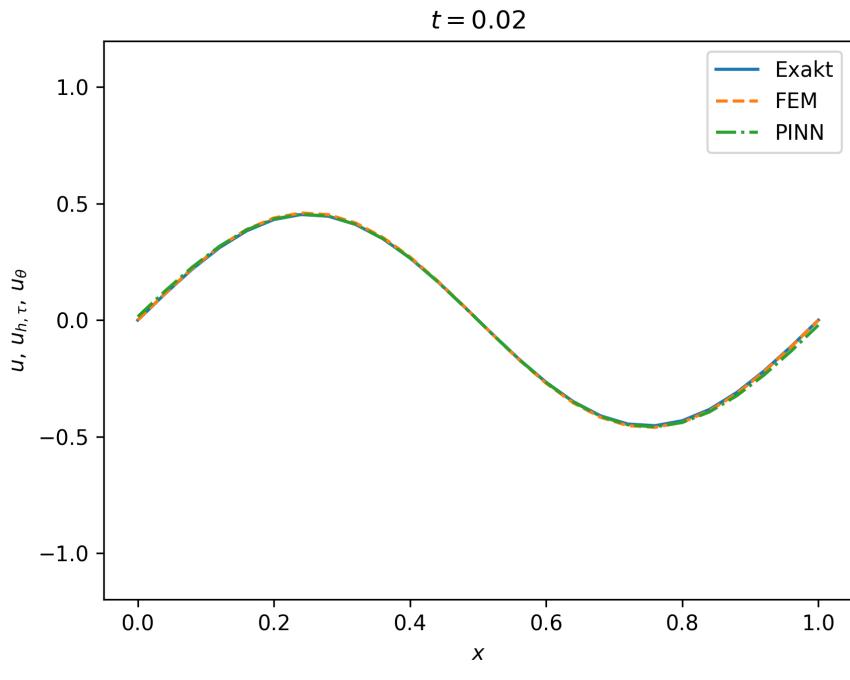
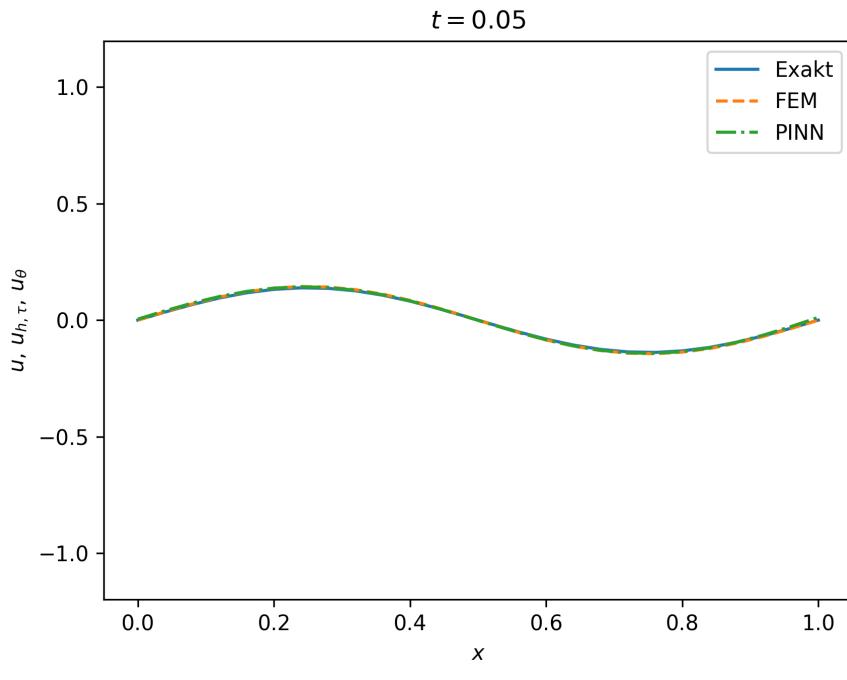
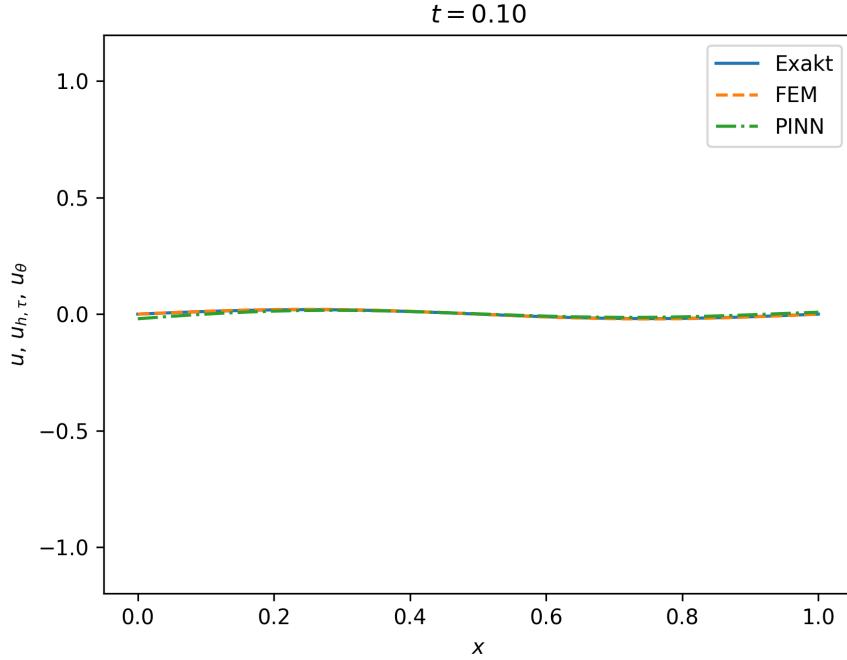


Abbildung 16: Plot der exakten Lösung und der Näherungslösungen der FEM und PINN für feste Zeiten (a) $t = 0.02$ und (b) $t = 0.03$ des 1D-Testproblems (5.1).



(a)



(b)

Abbildung 17: Plot der exakten Lösung und der Näherungslösungen der FEM und PINN für feste Zeiten (a) $t = 0.5$ und (b) $t = 0.1$ des 1D-Testproblems (5.1).

Exakte Dirichlet-Randbedingungen für PINN. Um die näherungsweise Einhaltung der Anfangs- und Randbedingungen in PINN zu überwinden, entwickeln wir eine neue Methodik für PINN mit exakten Bedingungen. Diesbezüglich können Dirichlet-Randbedingungen auf dem parabolischen Zylinder, d. h. Anfangs- und Randbedingungen von der Form

$$\begin{aligned} u(\cdot, 0) &= u_0 \quad \text{in } \Omega, \\ u &= g_D \quad \text{auf } \partial\Omega_D \times (0, T], \end{aligned}$$

für Teilmengen $\partial\Omega_D \subseteq \partial\Omega$ exakt eingebettet werden. Angelehnt an das Vorgehen in [Lu+21, Abschnitt 2.3], konstruieren wir hierfür hinreichend glatte Funktionen $\zeta, \eta : \bar{\Omega}_T \rightarrow \mathbb{R}$, so dass

$$\begin{aligned} \zeta(\cdot, 0) &= u_0 \quad \text{in } \Omega, \\ \zeta &= g_D \quad \text{auf } \partial\Omega_D \times (0, T], \end{aligned}$$

und

$$\begin{aligned} \eta &> 0 \quad \text{in } \Omega_T, \\ \eta(\cdot, 0) &= 0 \quad \text{in } \Omega, \\ \eta &= 0 \quad \text{auf } \partial\Omega_D \times (0, T], \end{aligned}$$

gilt. Im Allgemeinen ist jedoch die Konstruktion passender Funktionen ζ und η (ohne Kenntnis der Lösung) nur für einfache Gebiete Ω und $\partial\Omega_D$, sowie Anfangs- und Randbedingungen u_0 und g_D möglich. Im Falle des 1D-Testproblems (5.1), wähle die Funktionen

$$\begin{aligned} \zeta(x, t) &= \sin(2\pi x), \\ \eta(x, t) &= tx(1 - x). \end{aligned}$$

Damit die Näherungslösung u_θ nun die Dirichlet-Randbedingungen auf dem parabolischen Zylinder exakt erfüllt, setze

$$u_\theta(x, t) = \zeta(x, t) + \eta(x, t) \mathcal{N}^L(x, t),$$

wobei $\mathcal{N}^L(x, t)$ die Ausgangsgröße des KNNes für Eingangsdaten $(x, t) \in \bar{\Omega}_T$ ist, vgl. Abbildung 18. Folglich verschwinden die Terme der Kostenfunktion zugehörig zu den Anfangs- und Randbedingungen, d. h. $\mathcal{L}_T = \mathcal{L}_{\mathcal{H}, T}$ und es kann $\lambda_{\mathcal{I}} = \lambda_{\mathcal{B}} = 0$, sowie $\mathcal{T}_{\mathcal{I}} = \mathcal{T}_{\mathcal{B}} = \emptyset$ gewählt werden.

Unter Verwendung von nur 25000 Iterationen im Lernprozess, zu sehen in Abbildung 19, können wir feststellen, dass der Fehler in Abbildung 20 von PINN für

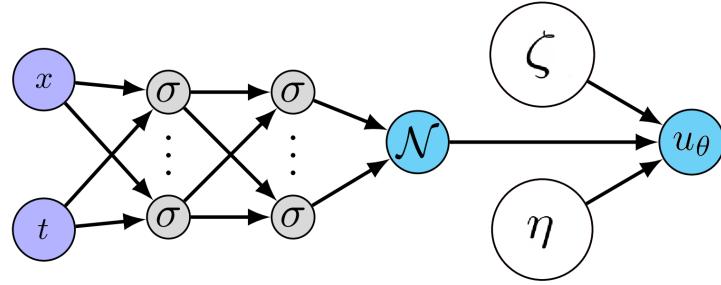


Abbildung 18: Schema des neuronalen Netzes in PINN für das Lösen von (2.7) mit exakten Anfangs- und Randdaten, in Anlehnung an [Lu+21, Abbildung 1B].

fast alle Zeitpunkte kleiner ist als der Fehler der FEM. Des Weiteren ist in den Abbildungen 21 bis 23 die exakte Erfüllung der Anfangs- und Randbedingungen erkennbar. Zusätzlich ist durch den reduzierten Rechenaufwand die Rechenzeit von PINN auf 164.10 s gesunken, wodurch es aber dennoch keine echte Alternative zur FEM darstellt.

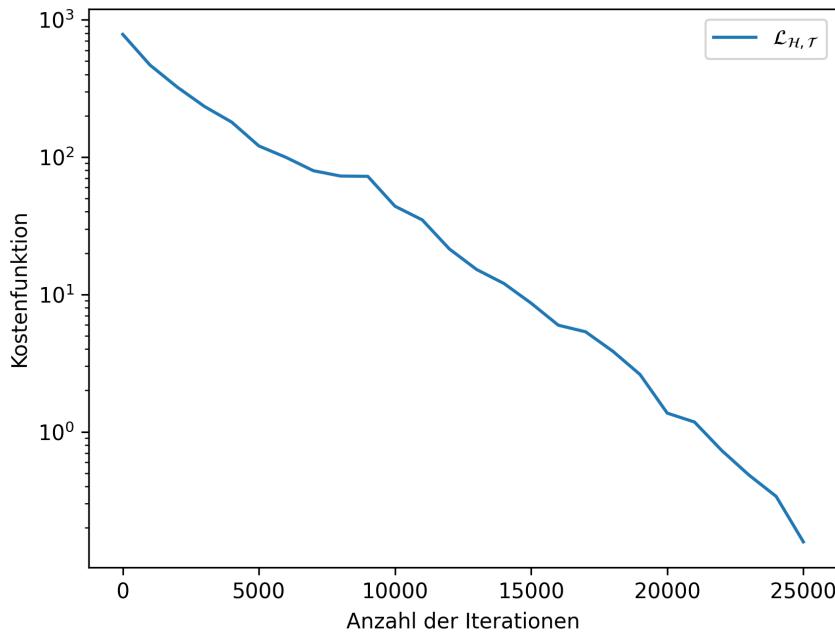
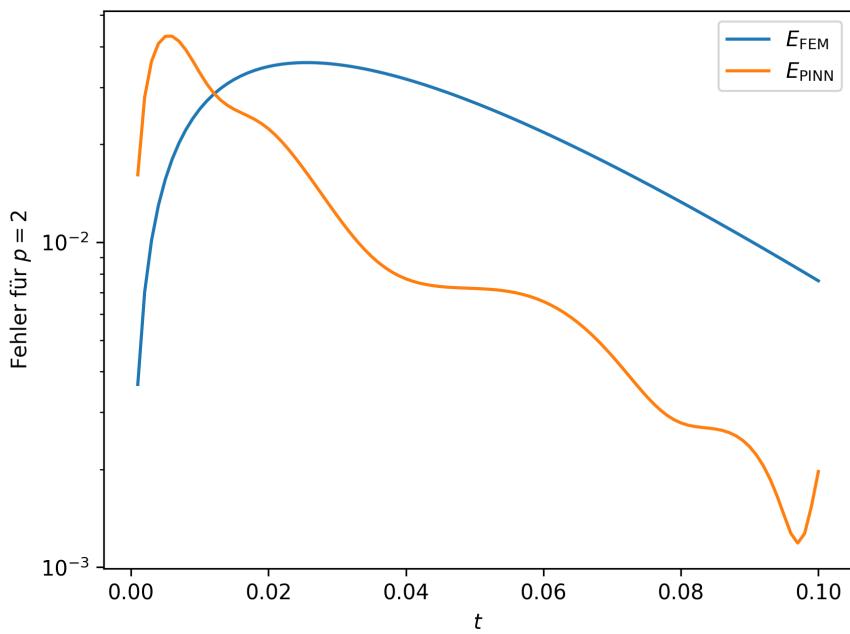
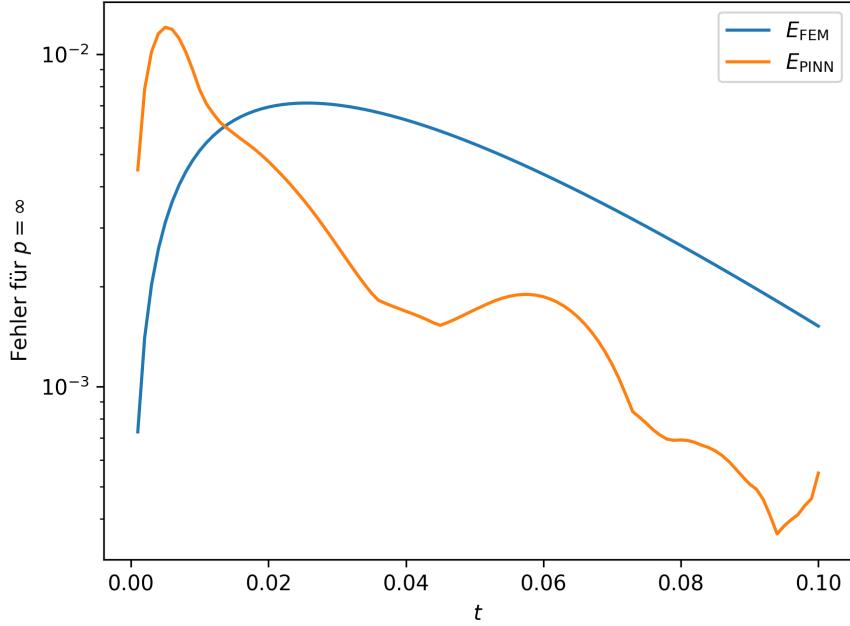


Abbildung 19: Plot der einzelnen Komponenten der Kostenfunktion (4.3) von PINN mit exakten Anfangs- und Randdaten für jeden Iterationsschritt des Optimierungsverfahrens.

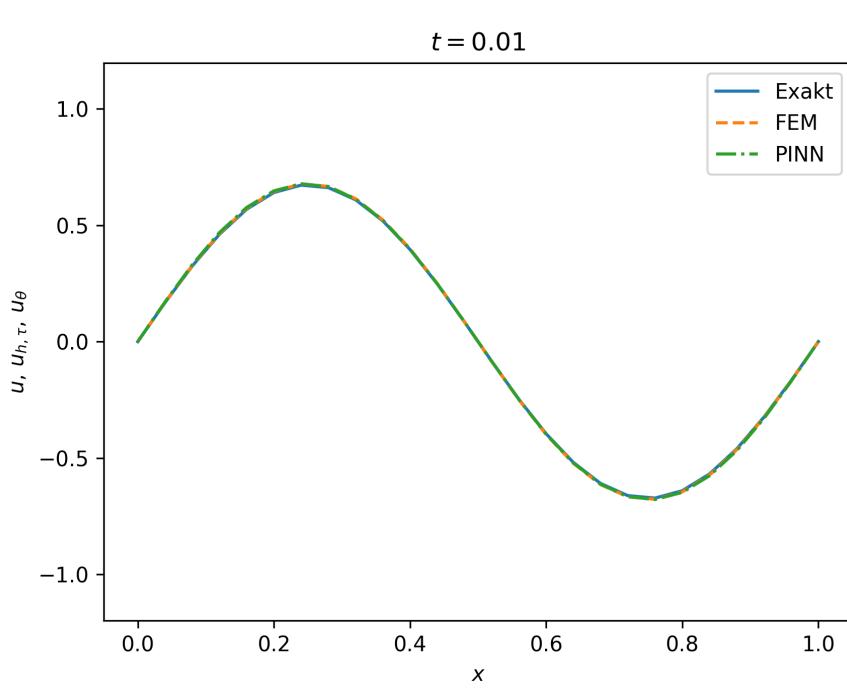
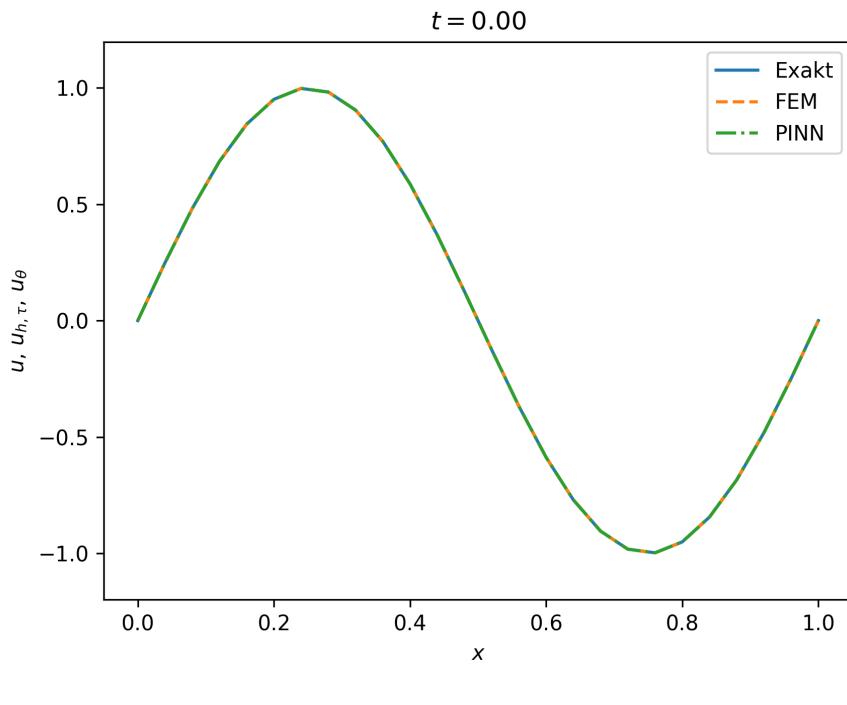


(a)



(b)

Abbildung 20: Fehlerplots von FEM und PINN mit exakten Anfangs- und Randdaten in Abhängigkeit von der Zeit für das 1D-Testproblem (5.1) für (a) $p = 2$ und (b) $p = \infty$.



(b)

Abbildung 21: Plot der exakten Lösung und der Näherungslösungen der FEM und PINN mit exakten Anfangs- und Randdaten für feste Zeiten (a) $t = 0$ und (b) $t = 0.01$ des 1D-Testproblems (5.1).

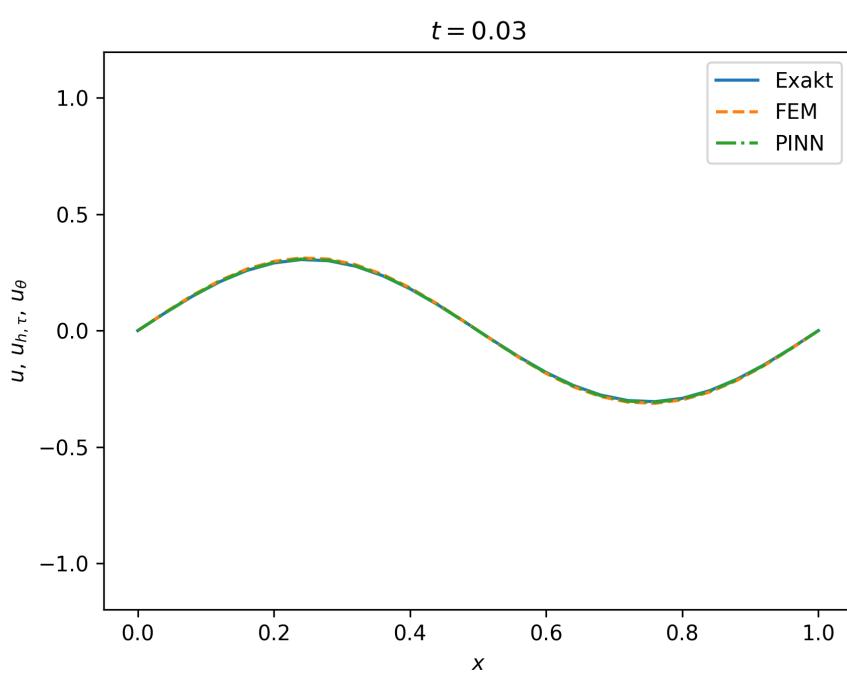
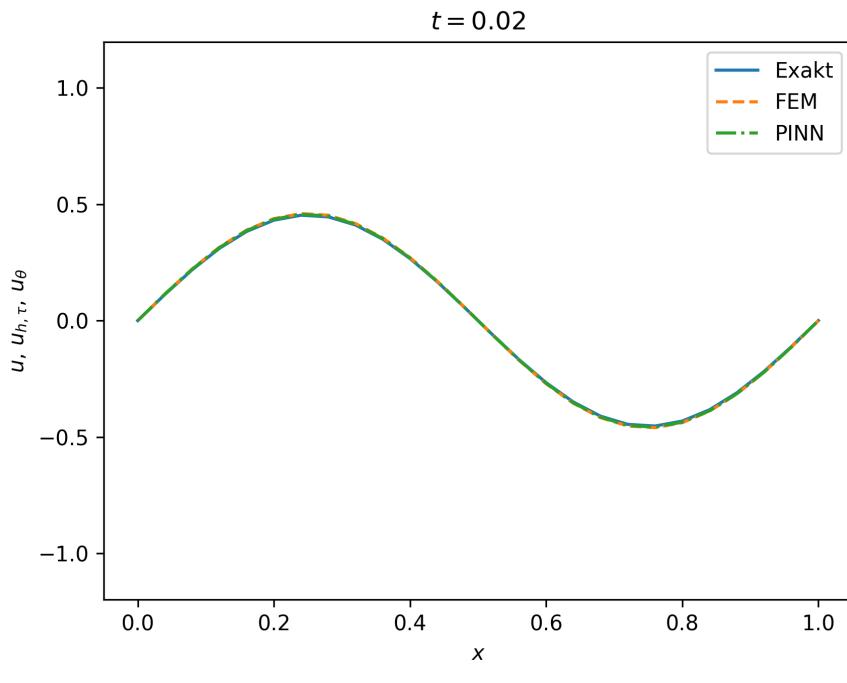
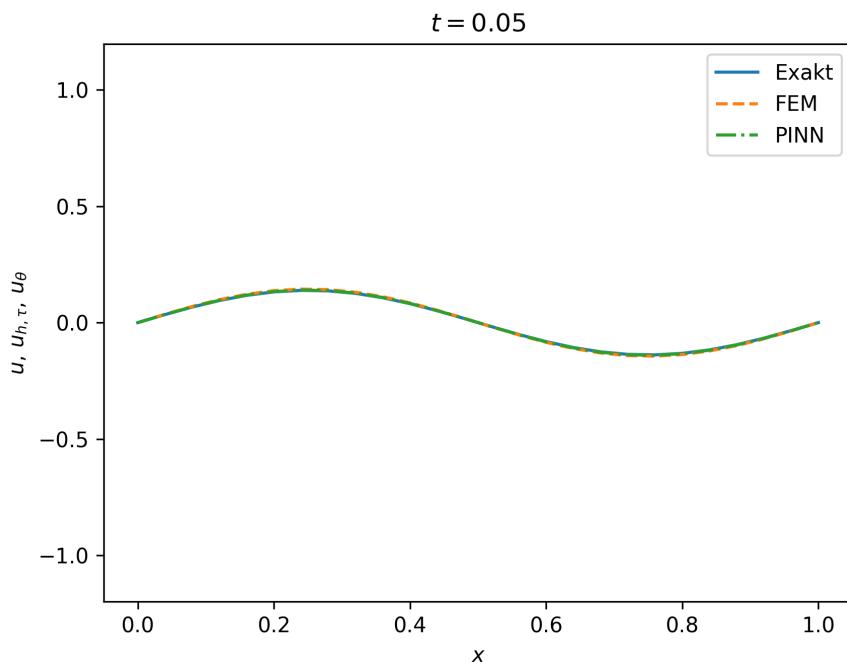
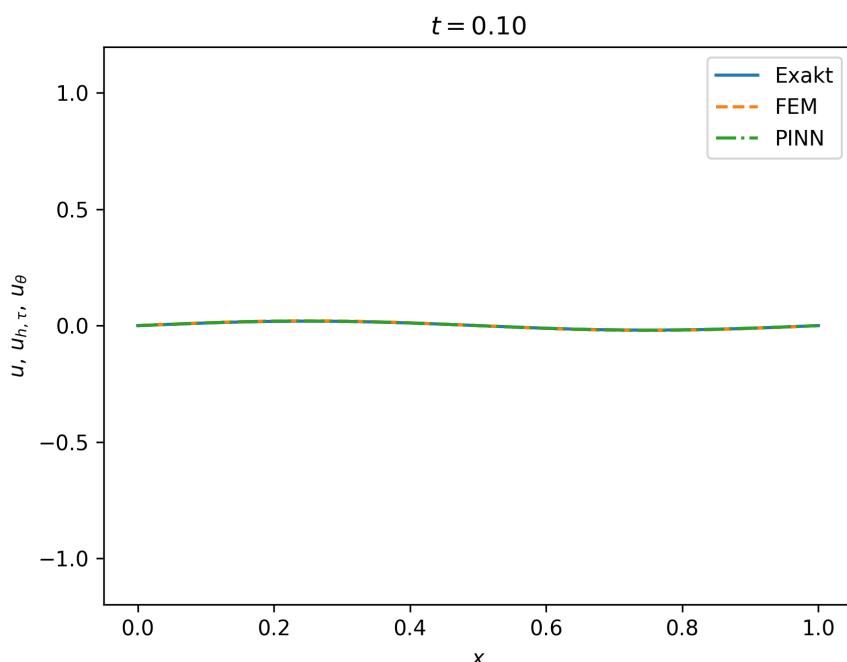


Abbildung 22: Plot der exakten Lösung und der Näherungslösungen der FEM und PINN mit exakten Anfangs- und Randdaten für feste Zeiten (a) $t = 0.02$ und (b) $t = 0.03$ des 1D-Testproblems (5.1).



(a)



(b)

Abbildung 23: Plot der exakten Lösung und der Näherungslösungen der FEM und PINN mit exakten Anfangs- und Randdaten für feste Zeiten (a) $t = 0.5$ und (b) $t = 0.1$ des 1D-Testproblems (5.1).

5.2 2D-Testproblem

In diesem Abschnitt betrachten wir die zweidimensionale Wärmeleitungsgleichung auf dem Gebiet $\Omega = (0, 1)^2$ mit Endzeitpunkt $T = 0.1$:

$$\partial_t u - \Delta u = f \quad \text{in } \Omega_T, \tag{5.2a}$$

$$u(\cdot, 0) = u_0 \quad \text{in } \Omega, \tag{5.2b}$$

$$u = g_D \quad \text{auf } \partial\Omega_D \times (0, T], \tag{5.2c}$$

$$\partial_\nu u = g_N \quad \text{auf } \partial\Omega_N \times (0, T], \tag{5.2d}$$

für disjunkte Teilmengen des Randes $\partial\Omega_D = \{0, 1\} \times (0, 1)$ und $\partial\Omega_N = [0, 1] \times \{0, 1\}$.

Definiere die Lösung

$$u(x, y, t) = e^{-4\pi^2 t} \cos(2\pi x) \cos(2\pi y).$$

Dann lässt sich leicht zeigen, dass die Daten (f, u_0, g_D, g_N) wie folgt gegeben sind:

$$\begin{aligned} f(x, y, t) &= 4\pi^2 e^{-4\pi^2 t} \cos(2\pi x) \cos(2\pi y), \\ u_0(x, y) &= \cos(2\pi x) \cos(2\pi y), \\ g_D(x, y, t) &= e^{-4\pi^2 t} \cos(2\pi y), \\ g_N(x, y, t) &\equiv 0. \end{aligned}$$

Seien $u_{h,\tau}$ und u_θ die Näherungslösungen von (5.2) nach der FEM und PINN. Erneut schränken wir uns in der FEM auf die Verwendung des impliziten Euler-Verfahrens ein. Dabei sind die Parameter der Verfahren in Tabellen 3 und 4 gegeben.

Tabelle 3: Parameter der FEM für 2D-Testproblem

Grad der Lagrange-Elemente	$k = 2$
Gitterweite	$h \approx 0.057$
Zeitschrittweite	$\tau = 0.001$

Dann ist in Abbildung 24 die Kostenfunktion von PINN für jeden Iterationsschritt zu sehen. Wir stellen fest, dass die Kostenfunktion im letzten Iterationsschritt deutlich größer ist als im 1D-Testproblem (5.1). Betrachten wir jetzt die Fehler im Ort der beiden Näherungslösungen für alle Zeiten in Abbildung 25, so erkennen wir, dass der Fehler der FEM für jeden Zeitpunkt kleiner ist als der Fehler von PINN. Zudem ist die Diskrepanz zwischen den beiden Fehlern größer als im 1D-Testproblem, siehe Abbildung 14. Dies ist der bereits erwähnten, höheren Kostenfunktion in Abbildung 24 zu verantworten. Außerdem sind in den Abbildungen 26 bis 31 die exakte und approximativen Lösungen dargestellt. Abschließend ist die Rechenzeit von PINN mit 188.25 s erneut deutlich höher als die von der FEM mit 3.83 s und folglich ist die FEM den PINN überlegen.

Tabelle 4: Parameter von PINN für 2D-Testproblem

Netzwerkstruktur	
Anzahl der Schichten	$L = 7$
Neuronen in jeder Schicht	$N = N_1 = \dots = N_{L-1} = 18$
Aktivierungsfunktion	$\sigma = \tanh$
Kostenfunktion	
Gewichte der Kostenfunktion	$\lambda_{\mathcal{H}} = \lambda_{\mathcal{I}} = \lambda_{\mathcal{B},D} = \lambda_{\mathcal{B},N} = 1$
Anzahl der Trainingsdaten in $\mathcal{T}_{\mathcal{H}}$	1000
Anzahl der Trainingsdaten in $\mathcal{T}_{\mathcal{I}}$	1000
Anzahl der Trainingsdaten in $\mathcal{T}_{\mathcal{B},D} \cup \mathcal{T}_{\mathcal{B},N}$	1000
Optimierung	
Verfahren	Adam
Gewichtsinitialisierung	Glorot
Lernrate	$\alpha = 0.01$
Anzahl der Iterationen	25000

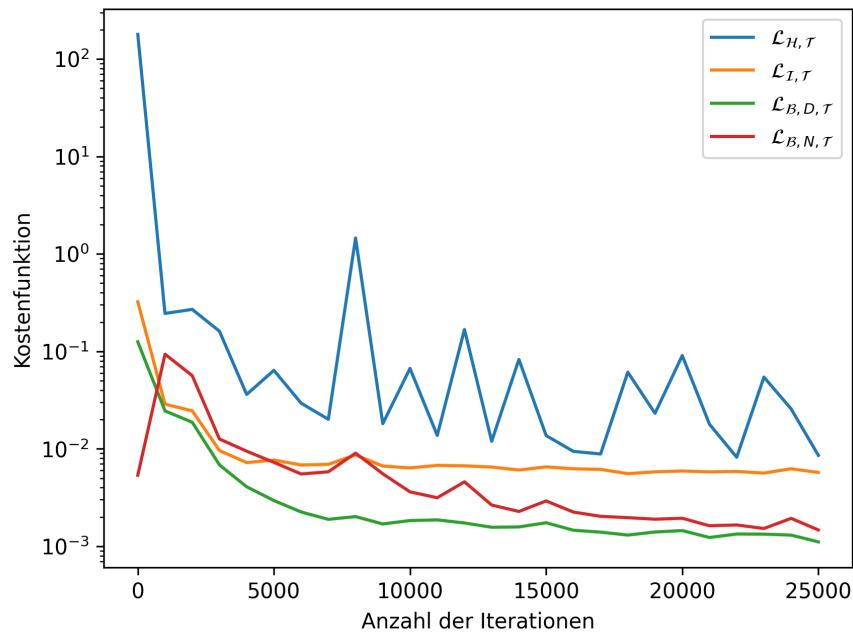
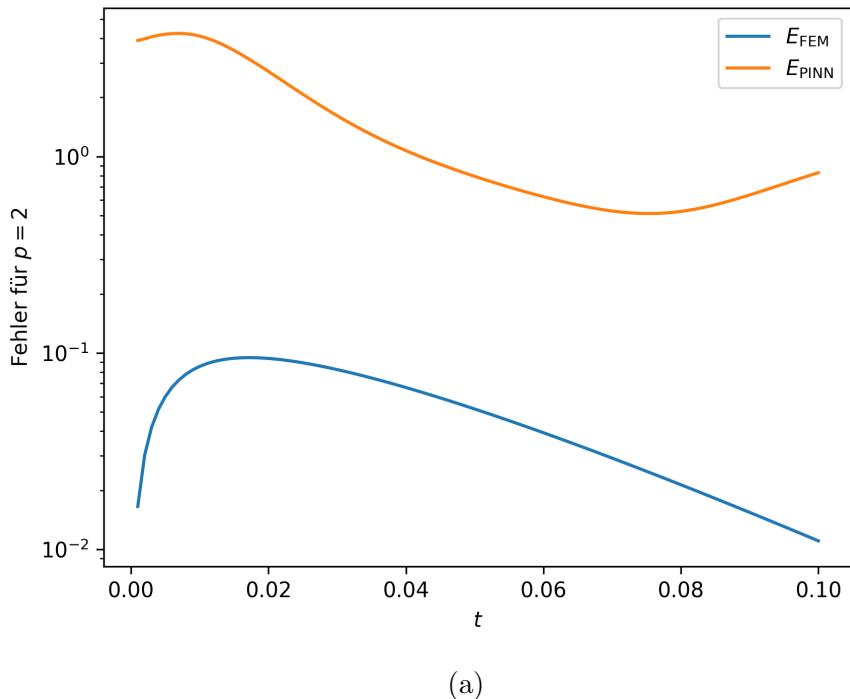
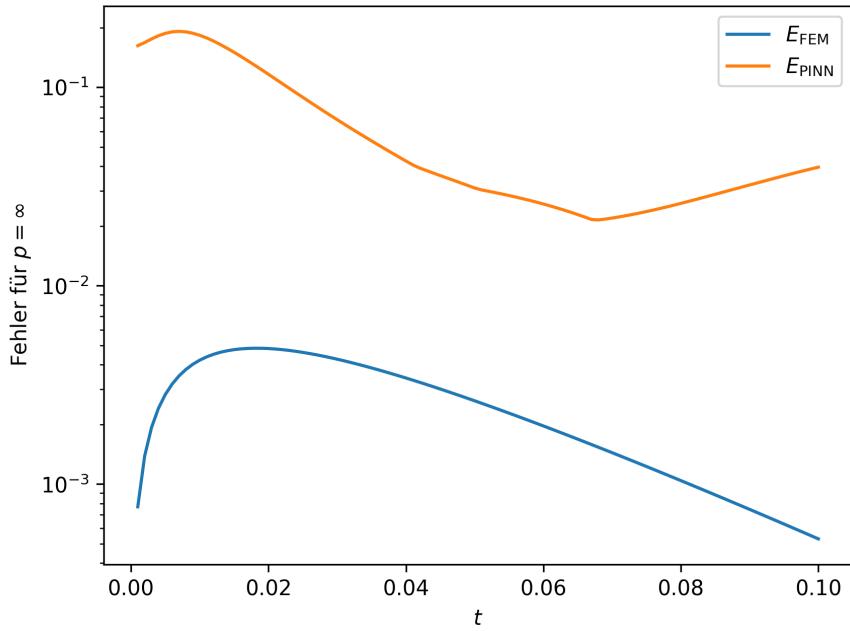


Abbildung 24: Plot der einzelnen Komponenten der Kostenfunktion (4.3) von PINN für jeden Iterationsschritt des Optimierungsverfahrens.



(a)



(b)

Abbildung 25: Fehlerplots von FEM und PINN in Abhängigkeit von der Zeit für das 2D-Testproblem (5.2) für (a) $p = 2$ und (b) $p = \infty$.

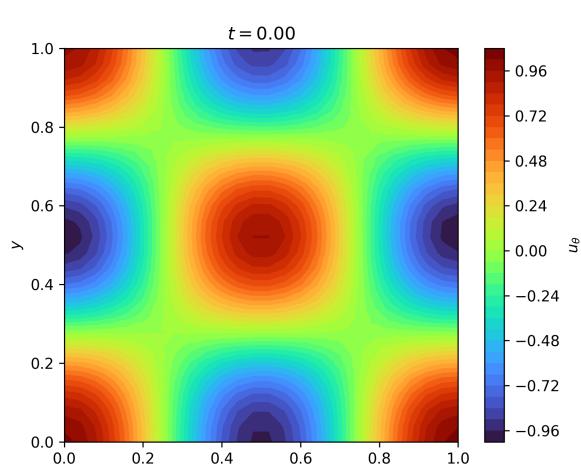
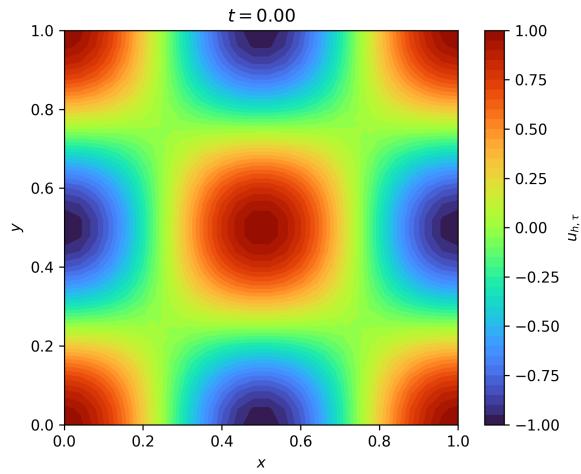
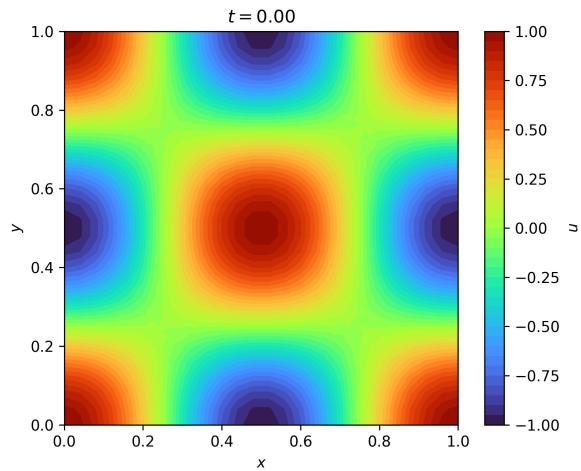
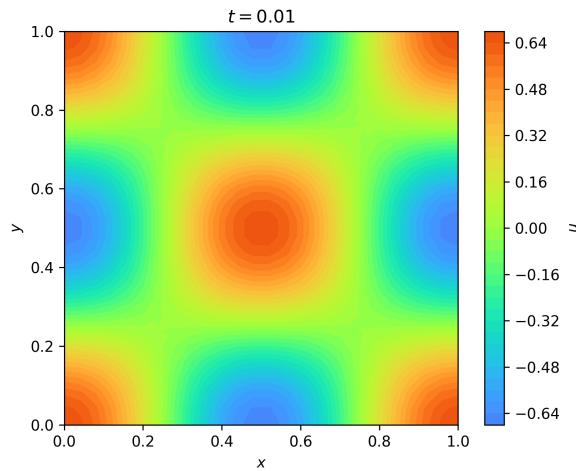
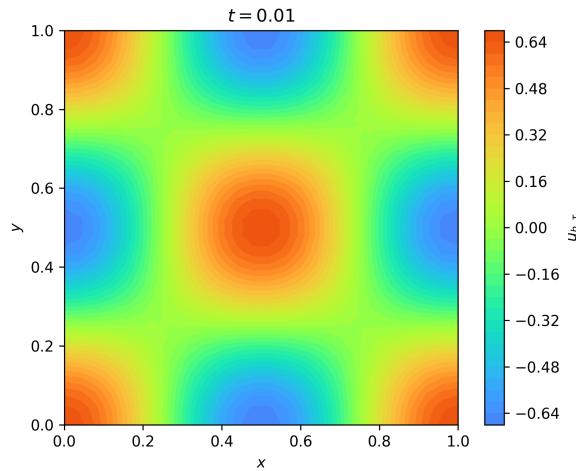


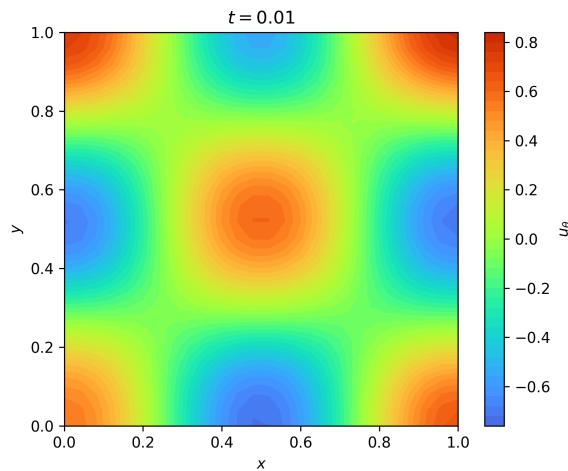
Abbildung 26: Plot der (a) exakten Lösung und der Näherungslösungen der (b) FEM und (c) PINN für das 2D-Testproblem (5.2) für $t = 0$.



(a)

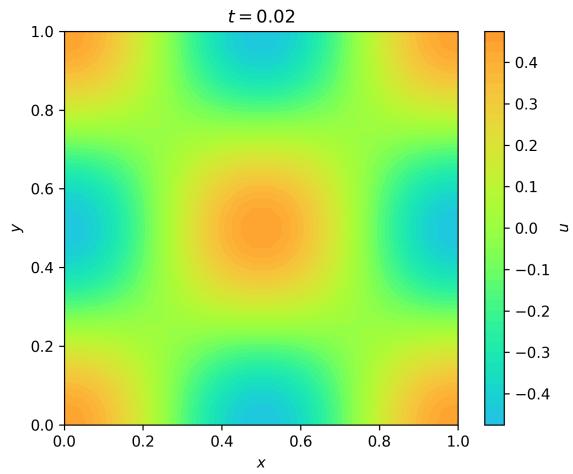


(b)

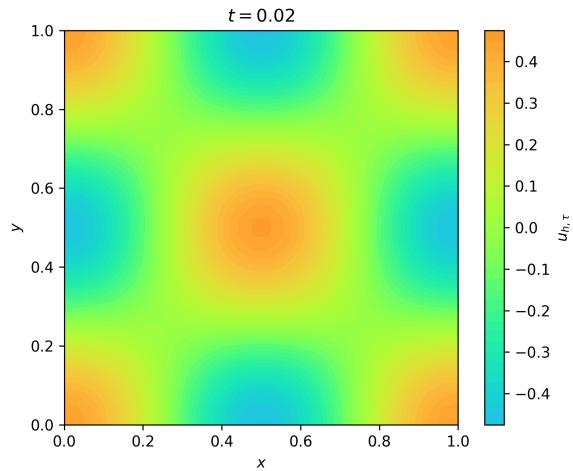


(c)

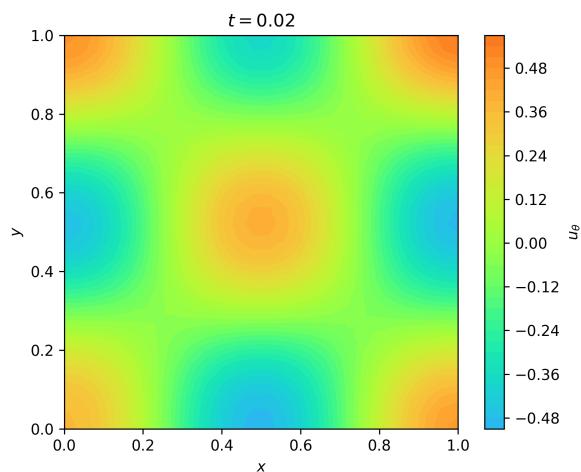
Abbildung 27: Plot der (a) exakten Lösung und der Näherungslösungen der (b) FEM und (c) PINN für das 2D-Testproblem (5.2) für $t = 0.01$.



(a)

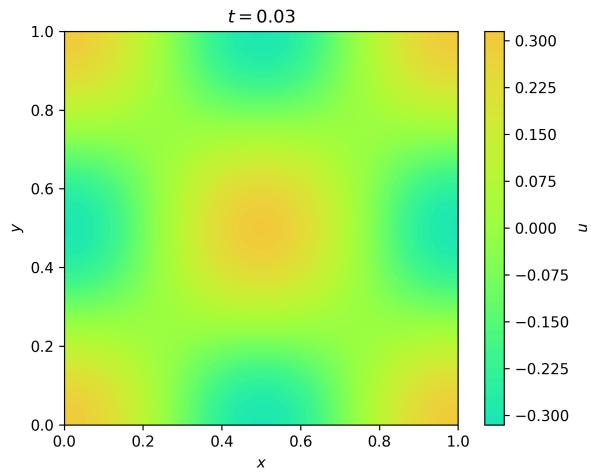


(b)

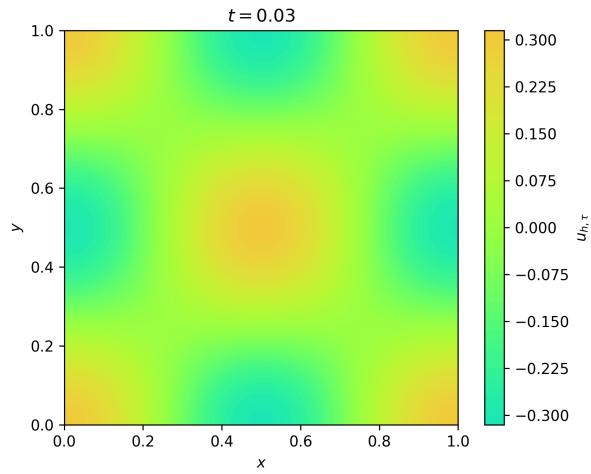


(c)

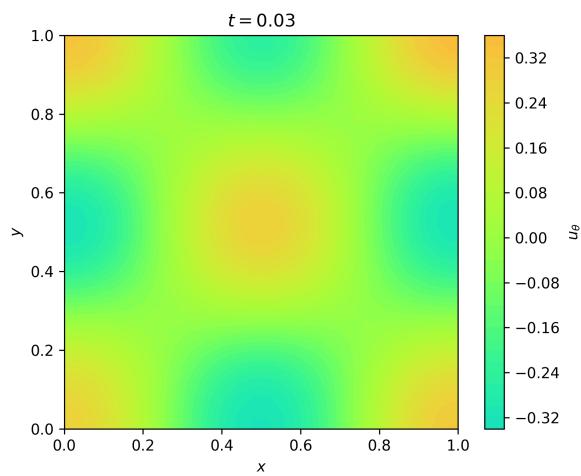
Abbildung 28: Plot der (a) exakten Lösung und der Näherungslösungen der (b) FEM und (c) PINN für das 2D-Testproblem (5.2) für $t = 0.02$.



(a)

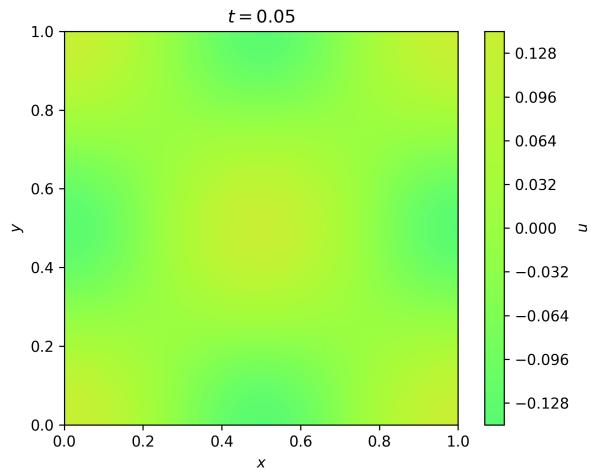


(b)

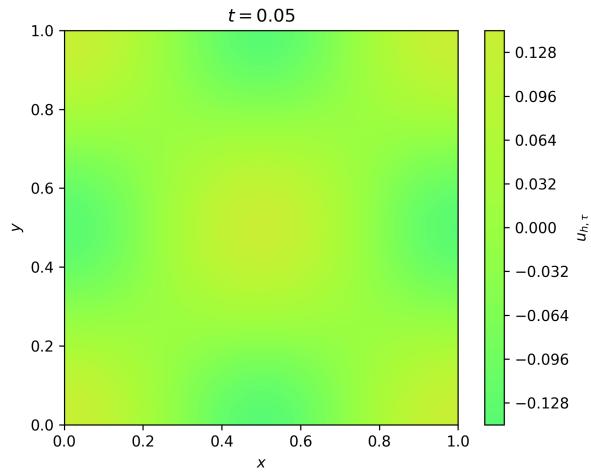


(c)

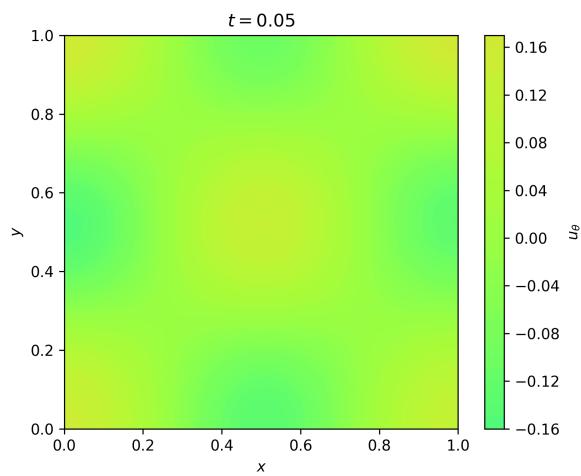
Abbildung 29: Plot der (a) exakten Lösung und der Näherungslösungen der (b) FEM und (c) PINN für das 2D-Testproblem (5.2) für $t = 0.03$.



(a)



(b)



(c)

Abbildung 30: Plot der (a) exakten Lösung und der Näherungslösungen der (b) FEM und (c) PINN für das 2D-Testproblem (5.2) für $t = 0.05$.

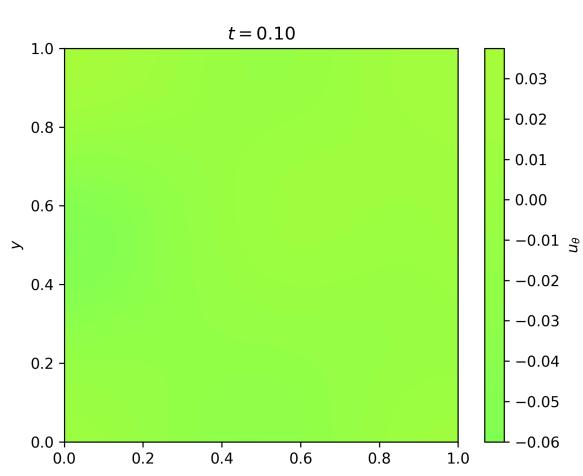
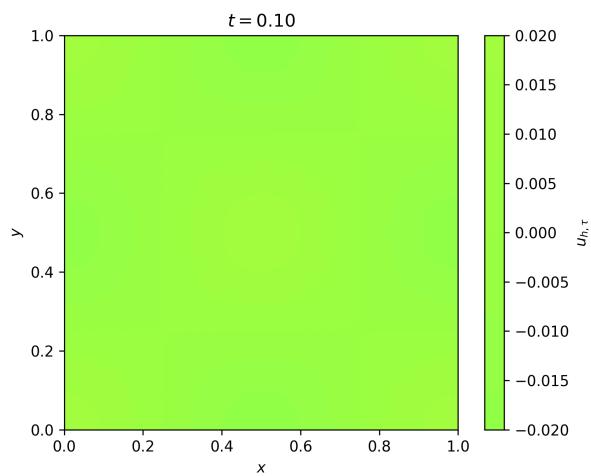
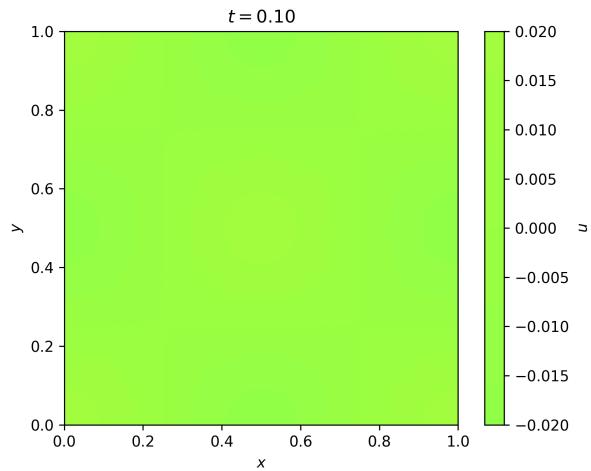


Abbildung 31: Plot der (a) exakten Lösung und der Näherungslösungen der (b) FEM und (c) PINN für das 2D-Testproblem (5.2) für $t = 0.1$.

6 Fazit

In dieser Arbeit sind klassische Verfahren (FEM) und Verfahren aus dem Bereich des maschinellen Lernens (PINN) auf die näherungsweise Berechnung der Wärmeleitung angewendet worden. Wohingegen die klassische Herangehensweise in den Abschnitten 5.1 und 5.2 unter verhältnismäßig geringen Rechenaufwand gute Näherungen liefern, fordern neuronale Netze deutlich mehr Rechenaufwand im Training für ähnliche Ergebnisse. Außerdem werden in PINN neben der PDE auch die Anfangs- und Randbedingungen lediglich approximativ erfüllt. Diese Problematik kann gelöst werden durch die exakte Einbettung dieser Bedingungen wie in Abschnitt 5.1. Jedoch darf nicht unerwähnt bleiben, dass dieser Ansatz nicht für beliebige (komplexe) Gebiete und Anfangs- bzw. Randbedingungen möglich ist. Eine weitere Schwierigkeit von PINN stellt die Suche von passenden Netzwerkstrukturen und anderen relevanten Parametern für den Trainingsprozess dar. Diese ist aktuell noch mit einer Vervielfachung des Rechenaufwandes und der Notwendigkeit einer Feinjustierung des Benutzers verbunden. Obendrein kann die Güte der Approximation in der klassischen Methode durch eine Reduktion der Gitter- und Zeitschrittweite unter steigendem Rechenaufwand verbessert werden. Ähnliche Konvergenzaussagen für PINN sind noch offene Probleme in der Wissenschaft. Abseits den genannten Nachteilen sind PINN aber nicht vom „Fluch der Dimensionalität“ betroffen. Wohingegen die Anzahl der Trainingsdaten in PINN unabhängig von der Wahl der Raumdimension ist, siehe Monte-Carlo-Integration in Abschnitt 4.2, sieht es bei der FEM schon anders aus. So stellen wir trotz Betrachtung von nur zwei Testproblemen fest, dass für steigende Raumdimensionen der Rechenaufwand in der FEM aufgrund der steigenden Anzahl an Gitterpunkten exponentiell wächst. Ebenfalls sind PINN deutlich flexibler einsetzbar im Gegensatz zu FEM. Tatsächlich sind nach [Lu+20, Abschnitte 2.6 und 2.7] PINN zum Lösen von Integro-Differentialgleichungen oder inversen Problemen geeignet. Zusammenfassend lässt sich sagen, dass trotz der aktuellen Unterlegenheit gegenüber klassischen Methoden (hier in der Wärmeleitung) PINN neue Möglichkeiten zum Lösen von PDEs und anderen verwandten Problemstellungen darstellen und insbesondere für hochdimensionale Probleme von Bedeutung sind.

Literatur

- [AD18] Ergün Akgün und Metin Demir. “Modeling Course Achievements of Elementary Education Teacher Candidates with Artificial Neural Networks”. In: *International Journal of Assessment Tools in Education* 5 (Jan. 2018). DOI: [10.21449/ijate.444073](https://doi.org/10.21449/ijate.444073).
- [AU18] Wolfgang Arendt und Karsten Urban. *Partielle Differenzialgleichungen*. Springer Spektrum, Berlin, Heidelberg, 2018. DOI: [10.1007/978-3-662-58322-7](https://doi.org/10.1007/978-3-662-58322-7).
- [Baz+10] Yuri Bazilevs u. a. “Residual-Based Variational Multiscale Theory of LES Turbulence Modeling”. In: Springer, Dordrecht, Okt. 2010, S. 3–18. DOI: [10.1007/978-90-481-9809-2_1](https://doi.org/10.1007/978-90-481-9809-2_1).
- [Cal20] Ovidiu Calin. *Deep Learning Architectures*. Bd. 1. Springer, Cham, 2020. DOI: [10.1007/978-3-030-36721-3](https://doi.org/10.1007/978-3-030-36721-3).
- [Cor16] Michael Cortis. “Numerical Modelling of Braided Fibres for Reinforced Concrete”. Diss. Nov. 2016. DOI: [10.13140/RG.2.2.19232.79364](https://doi.org/10.13140/RG.2.2.19232.79364).
- [DLM10] Claudia D’Ambrosio, Andrea Lodi und Silvano Martello. “Piecewise linear approximation of functions of two variables in MILP models”. In: *Oper. Res. Lett.* 38 (2010), S. 39–46.
- [Dzi10] Gerhard Dziuk. *Theorie und Numerik partieller Differentialgleichungen*. De Gruyter, 2010. DOI: [10.1515/9783110214819](https://doi.org/10.1515/9783110214819).
- [EG04] Alexandre Ern und Jean-Luc Guermond. *Theory and Practice of Finite Elements*. Springer, New York, NY, 2004. DOI: [10.1007/978-1-4757-4355-5](https://doi.org/10.1007/978-1-4757-4355-5).
- [ES00] Michael Evans und Timothy Swartz. *Approximating Integrals via Monte Carlo and Deterministic Methods*. Oxford University Press, 2000. URL: <https://EconPapers.repec.org/RePEc:oxp:obooks:9780198502784> (besucht am 23.04.2022).
- [Eva10] Lawrence C. Evans. *Partial differential equations*. Bd. 2. University of California, Berkeley, 2010.
- [Fre08] Pascal Frey. *The numerical simulation of complex PDE problems*. Vorlesungsskript. 2008. URL: https://www.1j11.math.upmc.fr/~frey/cours/UdC/ma691/ma691_ch7.pdf (besucht am 23.04.2022).
- [Joh] Prof. Dr. Volker John. *Numerische Mathematik III*. Vorlesungsskript. URL: https://www.wias-berlin.de/people/john/LEHRE/TH_NUM_PDE/th_num_pde_10.pdf (besucht am 23.04.2022).

- [Lu+20] Lu Lu u.a. *DeepXDE: A deep learning library for solving differential equations*. Preprint. 2020. arXiv: 1907.04502 [cs.LG].
- [Lu+21] Lu Lu u.a. *Physics-informed neural networks with hard constraints for inverse design*. Preprint. 2021. arXiv: 2102.04626 [cs.LG].
- [Mee19] Remco van der Meer. *Solving Partial Differential Equations with Neural Networks*. PhD thesis. Delft University of Technology, 2019. URL: <https://repository.tudelft.nl/islandora/object/uuid:c77e1bcc-7212-4234-af34-6586b628ab1c/datastream/OBJ/download> (besucht am 23.04.2022).
- [SB05] Josef Stoer und Roland Bulirsch. *Numerische Mathematik 2. Eine Einführung — unter Berücksichtigung von Vorlesungen von F.L. Bauer*. 5. Aufl. Springer, Berlin, Heidelberg, 2005. DOI: 10.1007/b137272.
- [Tho04] Vidar Thomée. *Galerkin Finite Element Methods for Parabolic Problems*. Springer, Berlin, Heidelberg, 2004. DOI: 10.1007/3-540-33122-0.
- [Wik22] Wikipedia contributors. *Mixed boundary condition — Wikipedia, The Free Encyclopedia*. [Online; accessed 23-April-2022]. 2022. URL: https://en.wikipedia.org/w/index.php?title=Mixed_boundary_condition&oldid=1073916912 (besucht am 23.04.2022).
- [Zü13] Francesco Züger. *Solution of Non-Homogeneous Dirichlet Problems with FEM*. Master thesis. Universität Zürich, 2013. URL: <https://www.math.uzh.ch/li/index.php?file&key1=25297> (besucht am 23.04.2022).

Erklärung der Urheberschaft

Ich erkläre hiermit an Eides statt, dass ich die vorliegende Arbeit ohne Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht. Die Arbeit wurde bisher in gleicher oder ähnlicher Form in keiner anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.

Ort, Datum

Unterschrift