Identifizierung nichtlinearer Koeffizientenfunktionen des reaktiven Transports durch poröse Medien unter Verwendung rekursiver und formfreier Ansätze

Der Naturwissenschaftlichen Fakultät
der Friedrich-Alexander-Universität Erlangen-Nürnberg
zur
Erlangung des Doktorgrades Dr. rer. nat

vorgelegt von
Michael Blume

aus Forchheim

Als Dissertation genehmigt von der Naturwissenschaftlichen Fakultät der Universität Erlangen-Nürnberg

Tag der mündlichen Prüfung:

Vorsitzender der
Promotionskommission:

Erstberichterstatter: Prof. Dr. P. Knabner

Zweitberichterstatter:

Drittberichterstatter:

Vorwort

An dieser Stelle möchte ich mich bei allen Personen bedanken, die mir bei der Erstellung der vorliegenden Arbeit hilfreich zur Seite standen. Zunächst ist hierfür selbstverständlich mein Doktorvater Prof. Dr. Peter Knabner zu nennen, der mir neben der Themenfindung und zahlreichen fachlichen Anregungen während der gesamten Zeit eine hervorragende Betreuung bot. Besonderen Dank gebührt auch Herrn Dr. Alexander Prechtel, der stets gerne und mit großem Interesse als Diskussionspartner bereitstand und mich unaufhaltsam motiviert und gefördert hat. Auch Herrn Florian Frank möchte ich für seine Unterstützung, insbesondere bei schwierigen Implementierungsfragestellungen, Dank aussprechen. Des Weiteren bedanke ich mich auch bei allen übrigen Mitarbeiterinnen und Mitarbeitern des Lehrstuhls für Angewandte Mathematik I der Friedrich-Alexander-Universität für das überaus freundliche und kollegiale Arbeitsklima. Auch allen Professoren und (z.T. ehemaligen) PhD-Studierenden, des, unter der Schirmherrschaft des Elitenetzwerkes Bayern geförderten, internationalen Doktorandenkollegs, gebührt Dank. Stellvertretend seien an dieser Stelle Prof. Dr. Günter Leugering und Prof. Dr. Hans-Josef Pesch zu nennen, die stets ein offenes Ohr für mich hatten und auch in fachlicher Hinsicht mehrfach Hilfestellungen gaben. Selbstverständlich möchte ich auch meinen Eltern für ihre unermüdliche Unterstützung einen außerordentlichen Dank aussprechen. Schließlich gilt meiner Frau ein ganz besonderer Dank. Ohne ihre selbstlose Hilfe, ihrem liebevollen Beistand und ihrem verständnisvollen Verzicht hätte ich niemals meine Arbeit in dieser Form abgeben können.

Erlangen, Mai 2011,
Michael Blume

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

Kapitel 1

Einleitung

Eine der wichtigsten Ressourcen zur Gewinnung unseres Trinkwassers ist das im Boden befindliche Grundwasser. Damit dies stets qualitativ hochwertig und schadstoffunbelastet ist, muss dafür gesorgt werden, dass es nicht mit verunreinigten oder kontaminierten Substanzen in Verbindung kommt. Insbesondere Altlasten, welche in Deutschland nach dem Bundes-Bodenschutzgesetz (BBodSchG) folgendermaßen definiert sind:

> *Altlasten im Sinne dieses Gesetzes sind*
>
> 1. *stillgelegte Abfallsbeseitigungsanlagen sowie sonstige Grundstücke, auf denen Abfälle behandelt, gelagert oder abgelagert worden sind (Altablagerungen), und*
>
> 2. *Grundstücke stillgelegter Anlagen und sonstige Grundstücke, auf denen mit umweltgefährdenden Stoffen umgegangen worden ist, ausgenommen Anlagen, deren Stilllegung einer Genehmigung nach dem Atomgesetz bedarf (Altstandorte),*
>
> *durch die schädliche Bodenveränderungen oder sonstige Gefahren für den einzelnen oder die Allgemeinheit hervorgerufen werden*[1],

stellen jedoch ein potentielles Risiko dar. Um die ausgehende Gefahr besser abschätzen zu können, werden im BBodSchG [13] und in der Bundes-Bodenschutz- und Altlastenverordnung (BBodSchV) [14] Sickerwasserprognosen eingeräumt, mit deren Hilfe der Schadstofftransport durch die ungesättigte Bodenzone ermittelt werden soll. Hierauf aufbauend wurde vom BMBF (Bundesministerium

[1] Bundes-Bodenschutzgesetzt, §2, Abs. (5)

für Bildung und Forschung) der Förderschwerpunkt *Sickerwasserprognose* eingerichtet (vgl. z.B. Rudek, Eberle [52]). U.a. beschäftigten sich Stieber, Kraßnitzer, Tiehm [57] im Rahmen dieser Maßnahme mit der Bedeutung biologischer Selbstreinigungsprozesse. Hierzu wurden experimentelle Messungen durchgeführt und die erzielten Daten ausgewertet. Insbesondere wurde das "biologische Abbauverhalten grundwassergängiger PAK während des Sickerwassertransports in der ungesättigten Bodenzone anhand von Laborexperimenten untersucht, um ein Testverfahren zur Ermittlung biologischer Abbauparameter zu entwickleln und anzuwenden."[2] Als PAK- (polyzyklische aromatische Kohlenwasserstoffe) Quellen standen hierbei ein kontaminierter Altlastboden und ein Bauschutt-Recyclingmaterial zur Verfügung. Die verwendete Versuchsanlage ist in Abbildung 1.1 (vgl. [57], Abbildung 1 und 2) dargestellt.

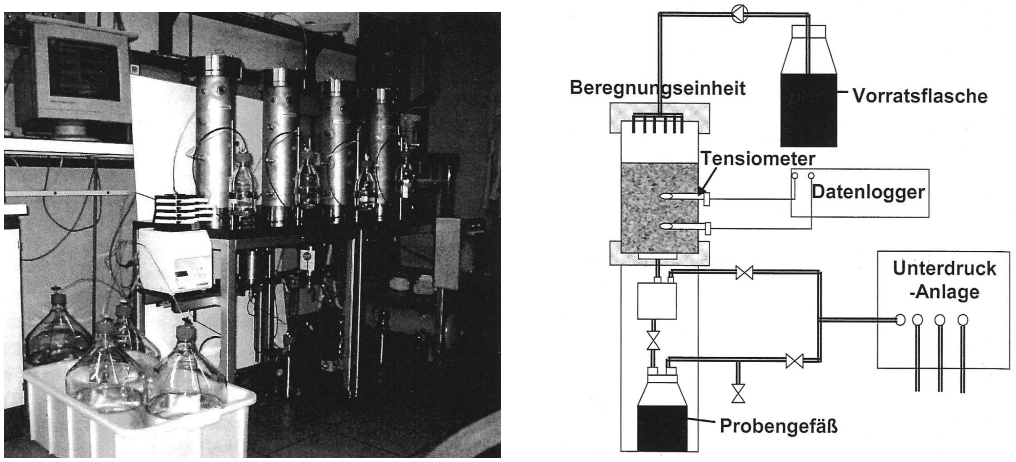

Abbildung 1.1: Versuchsanlage zur Durchführung von Sickerexperimenten

PAK treten in vielen Formen und Facetten auf. So sind sie beispielsweise ein "natürlicher Bestandteil von Kohle und Erdöl"[3]. Ebenso sind sie in "mit Steinkohleteer behandelte Produkte, z.B. teergebundener Asphalt aus der Zeit vor 1970, Teerpappe oder Teerimprägnierungen (für Telegrafenmasten oder Eisenbahnschwellen)"[3] nachzuweisen. "In Otto- und Dieselkraftstoff bzw. Heizöl findet man ebenfalls Spuren von PAK. ... In Gebäuden sind sie oftmals in Teer und pechhaltigen Klebstoffen und Farben unter Holzparkett, als Beschichtung von Trinkwasserleitungen sowie bei alten Fußbodenbelägen mit teerhaltigem Asphalt zu finden."[3] "Wegen Ihrer Persistenz, ihrer Toxizität und ihrer ubiquitären Verbreitung haben PAK eine große Bedeutung als Schadstoff in der Umwelt. Bereits in den

[2]Stieber, Kraßnitzer, Tiehm [57], Seite 111
[3]www.wikipedia.de, PAK-Vorkommen, Stand 10. September 2007

1980er Jahren hat die amerikanische Bundesumweltbehörde (USEPA) ... 16 Substanzen in die Liste der *Priority Pollutants* aufgenommen. Diese 16 EPA-PAK werden seitdem hauptsächlich und stellvertretend für die gesamte Stoffgruppe analysiert."[4]

Eine aussagekräftige Prognose der Schadstoffausbreitung wird, aufbauend auf empirisch ermittelten Daten, durch die numerische Simulation der reaktiven Transportprozesse ermöglicht. Da diese i.d.R. sehr komplex sind und (meist) ein aufwändig zu lösendes, gekoppeltes nichtlineares partielles Differentialgleichungssystem bilden, war die rasant anwachsende Leistungssteigerung handelsüblicher Computer äußerst förderlich. So konnte gerade in den letzten Jahren auch vertieft mit effektiven Methoden auf die Parameteridentifizierung eingegangen werden. Entsprechend entstanden zahlreiche Arbeiten (vgl. z.B. Bitterlich et al. [9], Bitterlich, Knabner [11] oder Geisel [27]), welche sich mit neuen Identifizierungsansätzen beschäftigten.

Die Problematik, die sich bei einer computergestützten Simulation unweigerlich ergibt, ist die Tatsache, dass die Qualität der erzielten Prognose stark abhängig von den zur Verfügung stehenden Messdaten und der zugrundegelegten Modellierung ist. Eine genaue Kenntnis über alle vorliegenden bodenspezifischen Eigenschaften ist daher unerlässlich. Dementsprechend müssen bei einer Simulation alle relevanten Reaktionen und Gegebenheiten, wie beispielsweise Sorptions- und Abbauprozesse sowie eine gesättigte/ungesättigte Strömung, berücksichtigt werden.

Zur Modellierung eines gesättigten/ungesättigten Fließverhaltens wird die sogenannte Richards-Gleichung herangezogen. Dieses Zweiphasenmodell, bestehend aus dem zugrundegelegten Fluid (oft auch nur als Wasser bezeichnet) und einem mit konstantem Druck angenommenen Gas (vereinfacht Luft), ermöglicht eine adäquate Berechnung des Fließgeschehens. Zur Beschreibung der hydraulischen Funktionen wird typischerweise die van Genuchten-Mualem-Parametrisierung gewählt. Diese ist aufgrund ihrer geschlossen-analytischen Formulierung weit verbreitet und beliebt. Jedoch birgt sie auch numerische Schwierigkeiten. So nimmt selbst die druckabhängige Formulierung für ungeeignet gewählte Parameterwerte ($1 < n < 2$) nahe dem gesättigten Bereich eine beliebig hohe Steigung an und führt folglich zu einem nicht differenzierbaren Übergang. Diesem Problem entgegenwirkend werden in dieser Arbeit zwei mögliche Regularisierungsansätze

[4] www.wikipedia.de, PAK als Umweltschadstoff, Stand 10. September 2007

vorgestellt, bei denen, entweder druck- oder sättigungsabhängig, der kritische Bereich durch eine $\mathcal{P}^2$- bzw. $\mathcal{P}^3$-Approximation ersetzt wird. Ein weiteres, wesentlich bedeutenderes Problem liegt bei der Identifizierung dieser Nichtlinearitäten vor. Neben der eigentlichen Schlechtgestelltheit der inversen Aufgabenstellung liefert die van Genuchten-Mualem-Parametrisierung für völlig unterschiedliche Parameterwerte (inbesondere α und n) stückweise äußerst ähnliche hydraulische Leitfähigkeitsfunktionen. Dies macht eine ausreichend gute Identifizierung, insbesondere unter Verwendung mit (Mess-)Störung behafteter Daten, oftmals sehr schwierig.

Im Rahmen dieser Arbeit wird ein modifizierter Identifizierungsansatz vorgestellt, mit dessen Hilfe die Bestimmung von Parameterwerten fixer Funktionalitäten, wie sie beispielsweise durch die van Genuchten-Mualem-Leitfähigkeit gegeben ist, besser gelingen soll. Hierbei wird das gewöhnliche OLS- (Ordinary Least Squares) Fehlerfunktional durch eine sensitivitätsabhängige Gewichtung rekursiv manipuliert und damit eine Art sensitivitätsbezogene Mittelung erreicht. Ein mathematischer Beweis zur Belegung der besseren Identifizierbarkeit konnte nicht erstellt werden. Dennoch zeigten die durchgeführten Berechnungsbeispiele, welche sowohl auf virtuellen als auch auf realen Messdaten basierten, äußerst vielversprechende Resultate. Daß dieser Ansatz nicht nur zur Bestimmung hydraulischer Funktionen geeignet ist, wurde auch durch die Untersuchung einer weiteren Problemstellung untermauert. So waren die, auf einer Monod-parametrisierten Simulation eines dreikomponentigen reaktiven Schadstoffabbaus basierenden, Identifizierungsversuche ebenfalls erfolgreich. Allerdings zeigten diese Berechnungen auch klar die Grenzen der Identifizierbarkeit auf. Neben dem verwendeten Versuchsaufbau, dem sogenannten experimentellen Design, wie er u.a. durch die Anzahl und die Art und Lage der diskreten Messstellen sowie der zeitlichen Abfolge der einfließenden Konzentrationswerte gegeben ist, spielt auch die zugrundegelegte Modellierung eine wesentliche Rolle. Selbst unter Anwendung von Mehrfachexperimente, bei denen, zeitlich versetzt, unterschiedliche Designs verwendet werden, kann bei ungeeignet gewählten Komponenten eine adäquate Identifzierung nicht gewährleistet werden.

Motiviert durch Identifizierungsschwierigkeiten, welche auf (im Sinne der Identifizierung) schlecht gestellte fixe Parametrisierungen zurückzuführen sind, ist insbesondere eine formfreie Approximation der gesuchten Nichtlinearität eine äußerst interessante Alternative. Hierbei wird statt einer vorgegebenen Parametrisierung, z.B. die nach van Genuchten-Mualem definierte hydraulische Leitfähigkeit oder eine auf Monod basierende biochemische Reaktionsrate, eine stückweise

polynomiale Funktion bestimmt. Abhängig von der Dimension der zu identifizierenden Nichtlinearität werden hierzu ein- bzw. mehrdimensionale B-Splines mit gewünschter Ordnung verwendet. In Bitterlich [8], Bitterlich, Knabner [10] und Igler [34] finden sich bereits ausführliche Untersuchungen zur formfreien Identifizierung von Sorptionsisothermen und der hydraulischen Funktionen. Auf diesen eindimensionalen Identifizierungsansätzen aufbauend, wird im Rahmen dieser Arbeit eine Erweiterung auf dreidimensionale Funktionalitäten vollzogen. Als Anwendungsbeispiel wurde eine dreikomponentige Redoxreaktionsgleichung gewählt, bei der die gesuchte Abbaurate durch den sogenannten Elektronen-Donator (dem Schadstoff selbst), einem Elektronen-Akzeptor (typischerweise Sauerstoff) und der benötigten Biomasse aufgespannt wird. Für die Implementierung in das universitätsinterne Softwaretool Richy wurden sowohl lokale als auch hierarchische trilineare Splines ausgewählt. Im Fall hierarchischer Basen stehen dabei neben einem vollen Gitter auch dünne Gitter, wie sie in Bungartz [15] vorgestellt werden, zur Verfügung. Unter Verwendung zahlreicher, optional zuschaltbarer, Identifizierungshilfen, wie beispielsweise ein monotones Ansteigen der gesuchten Reaktionsrate, fixierte Axialknoten zur Verringerung der Freiheitsgrade, eine linksgewichtete Diskretisierung für eine bessere Identifizierbarkeit oder einfach nur konvexitätsunterstützende Startwerte, werden mit Hilfe unterschiedlicher Berechnungsbeispiele schließlich die dargestellten Diskretisierungsansätze untereinander verglichen.

Kapitel 2

Mathematische Modellierung bodenphysikalischer Prozesse

2.1 Fluidtransport in porösen Medien

In den folgenden Abschnitten werden die grundlegenden Ansätze zur Modellierung des Fluidtransports im Aquifer vorgestellt. Für eine detaillierte Darstellung kann u.a. auf die klassischen Werke Bear [5], Freeze, Cherry [26] oder Kinzelbach, Rausch [37] zurückgegriffen werden.

2.1.1 Gesetz von Darcy

Die grundlegende Gleichung für die Beschreibung eines Fluidtransportes in einem porösen Medium (heterogenes Material mit Feststoffskelett und zusammenhängenden Hohl- bzw. Porenräumen) ist das **Gesetz von Darcy** (1856), welches ursprünglich eindimensional durch

$$q = -K \frac{\Delta H}{\Delta z}$$

mit

q	(Darcy-)Geschwindigkeit, mit der das Liquid strömt [L/T],
K	(gesättigte) hydraulische Leitfähigkeit [L/T],
H	Druckhöhe [L] und
Δz :	Abstand zu einer Referenzhöhe [L],

formuliert wurde. Da sowohl die molekulare Ebene als auch die Mikroskala ($\sim$ Größe einer Pore) für das System Boden nicht praktikabel bzw. durchführbar ist, werden die meisten Berechnungen von Fluidtransporten auf der nächst-

höheren Ebene, der Makroskala, durchgeführt. Diese entspricht etwa der Größenordnung, die durch übliche Laborsäulen ($100cm^3 \ldots 0.1m^3$) repräsentiert wird.

Da poröse Medien i.d.R. sehr heterogen sind und entsprechend auf mikroskopischen Skalen extreme Wechsel zwischen Festphase und Poren besitzen, wird im Folgenden die Existenz eines repräsentativen Elementar-Volumens vorausgesetzt, so dass alle charakteristischen Größen als konstant angesehen und gemittelt werden können. Verwendet man statt der Druckhöhe H den Druck des Fluides $p(\vec{x})$ im Punkt $\vec{x} \in \Omega$, Ω Gebiet in $\mathbb{R}^d$, $d \in \{1, 2, 3\}$, so lässt sich das Gesetz von Darcy auch mehrdimensional und in infinitesimaler Form durch

$$q(\vec{x}) = -K \nabla p(\vec{x})$$

angeben.

2.1.2 Darcy-Buckingham-Gleichung

Der (ungesättigte) Fluidtransport in porösen Medien wird durch Energiegradienten beschrieben, wobei von einer langsamen Bewegung ausgegangen wird, so dass die kinetische Energie vernachlässigt werden kann. Wird als Bezugsgröße das Gewicht gewählt, so entspricht die Gesamtenergiedichte dem (hydraulischen) Potential ψ_h mit Einheit $\left[\frac{Nm}{N}\right] = [m]$. Der Referenzzustand, dem die Energie Null zugeschrieben wird, ist mit $z = 0$ festgelegt und die z-Achse in Gravitationsrichtung (bzw. entgegen dieser) gewählt. Das hydraulische Potential lässt sich als Summe aus Matrixpotential ψ_m (Energiedichte, die durch alle Boden-Fluid-Wechselwirkungen hervorgerufen wird) und Gravitationspotential ψ_g (Quotient aus Lageenergie des Fluides in Höhe z und Gewichtskraft) beschreiben. Damit gilt $\psi_g = z$ (bzw. $\psi_g = -z$) und entsprechend $\psi_h = \psi_m + z$ (bzw. $\psi_h = \psi_m - z$).

Mit Hilfe der Entwicklung der Potentialtheorie wurde durch Buckingham (1907) die Gültigkeit der Darcy-Gleichung auf ungesättigte Medien (neben dem Fluid, typischerweise Wasser, liegt ein zweites Medium, z.B. Luft, vor) erweitert. Für imkompressible, isotherme und isotrope Medien gilt

$$\boldsymbol{q}(\vec{x}, t) = -K\big(\Theta(\vec{x}, t)\big) \nabla \big(\psi_m(\vec{x}, t) + z\big), \tag{2.1}$$

mit

$\boldsymbol{q}(\vec{x}, t)$	vektorwertige Fließgeschwindigkeit [L/T],
$\Theta(\vec{x}, t)$	volumetrischer Wassergehalt [-],
$K(\Theta)$	(ungesättigte) hydraulische Leitfähigkeit [L/T],

$\psi_h(\vec{x},t)$	hydraulisches Potential [L], $\psi_h = \psi_m + z$,
$\psi_m(\vec{x},t)$	Matrixpotential [L],
$x \in \Omega$	Ortsvariable [L$^{\mathrm{d}}$], $\Omega \subset \mathbb{R}^d$ Gebiet in $\mathbb{R}^d$, $d \in \{1,2,3\}$,
z	Abstand auf z-Achse zu einer Referenzhöhe [L] und
$t \in (0,T)$	Zeitvariable [T] mit Endzeitpunkt T.

Die Flussdichte verhält sich weiterhin proportional zum hydraulischen Gradienten, die hydraulische Leitfähigkeit ist allerdings nun abhängig vom Fluidgehalt.

2.1.3 Richards-Gleichung

Um die zeitliche Veränderung des Fluidgehaltes zu beschreiben, wird noch das Kontinuitätsgesetz benötigt. Dieses beruht auf der einfachen Tatsache, dass die Menge an Flüssigkeit, die in ein Medium hineinfließt auch herausfließen muss, oder aber der Inhalt ändert sich. Schlichter (1897) formulierte dies formal durch

$$-\frac{\partial}{\partial t}\Theta(\vec{x},t) = \nabla \cdot \boldsymbol{q}(\vec{x},t)\,.$$

Schließlich verknüpfte Richards das Kontinuitätsgesetz mit der Darcy-Buckingham-Gleichung (2.1) (vgl. Richards [50]). In der sogenannten Mischform, bei der Θ und K in Abhängigkeit vom Matrixpotential $\psi := \psi_m$ angegeben werden, folgt damit

$$\frac{\partial}{\partial t}\Theta\big(\psi(\vec{x},t)\big) - \nabla \cdot K\big(\psi(\vec{x},t)\big)\nabla\big(\psi(\vec{x},t)+z\big) = 0\,. \tag{2.2}$$

Im ungesättigten Bereich, charakterisiert durch $\psi < 0$, ist im Porenraum neben dem modellierten Fluid auch Luft zu finden. Die hydraulischen Funktionen sind dort (physikalisch motiviert) nichtnegativ und monoton steigend. Im gesättigten Bereich $\psi \geq 0$ sind sie konstant und durch $K(\psi) = K_{\mathrm{sat}}$, $\Theta(\psi) = \theta_{\mathrm{sat}}$, $K_{\mathrm{sat}}, \theta_{\mathrm{sat}} \in \mathbb{R}^+$ definiert.

Bemerkung 2.1
In Ladyzhenskaya et al. [40] wird gezeigt, dass (2.2) für hinreichend glatte $\Theta(\psi)$ und $K(\psi)$ und unter der Einschränkung, dass aufgrund von entsprechenden Anfangs- und Randbedingungen ein ungesättigtes Problem vorliegt, also stets $\psi \leq \underline{\psi} < 0$ gilt, eine eindeutig glatte Lösung besitzt. Sind allerdings sowohl gesättigte als auch ungesättigte Phasen zu erwarten, kann aufgrund der elliptisch-parabolischen Degeneration nur eindeutig schwache Lösung mit geringerer Regularität garantiert werden (siehe z.B. Alt, Luckhaus [4] oder Otto [44]).

2.1.4 van Genuchten-Mualem-Parametrisierung

Für eine numerische Simulation des ungesättigten Fluidtransports wird neben einer geeigneten Diskretisierung (vgl. z.B. Reeves, Duguid [49], Segol [56] oder Vauclin et al. [63]) eine Parametrisierung der hydraulischen Leitfähigkeit und des volumetrischen Wassergehaltes benötigt. Neben dem in Brooks, Corey [12] vorgestellten Ansatz hat sich vor allem die van Genuchten-Mualem-Parametrisierung (vgl. van Genuchten [61]) durchgesetzt. Hierbei wird die relative hydraulische Leitfähigkeit $K_{\text{rel}} = \frac{K}{K_{\text{sat}}}$, $K_{\text{sat}} \in \mathbb{R}^+$, in Abhängigkeit des Sättigungsgehaltes Φ durch

$$K_{\text{rel}}(\Phi) = \Phi^{\frac{1}{2}} \left[\int_0^{\Phi} \frac{1}{\psi(x)} dx \Big/ \int_0^1 \frac{1}{\psi(x)} dx \right]^2 \tag{2.3}$$

beschrieben (vgl. Mualem [43]). Mit Hilfe des im Boden vorherrschenden Wassergehaltes $\theta \in [\theta_{\text{res}}, \theta_{\text{sat}}]$, $\theta_{\text{res}}, \theta_{\text{sat}} \in \mathbb{R}_0^+$ $[-]$, $\theta_{\text{res}} < \theta_{\text{sat}}$, kann Φ als affine Abbildung

$$\Phi(\Theta) = \frac{\Theta - \theta_{\text{res}}}{\theta_{\text{sat}} - \theta_{\text{res}}}$$

beschrieben werden. Um Gleichung (2.3) lösen zu können wird noch ein funktionaler Zusammenhang zwischen Φ und der Druckhöhe $\psi \in \mathbb{R}^-$ benötigt (im gesättigten Fall gilt trivialerweise $\Phi = 1$ und $K = K_{\text{sat}}$). Bei der Modellierung nach van Genuchten wird hierfür der Ansatz

$$\Phi(\psi) = \frac{1}{\big(1 + (-\alpha\psi)^n\big)^m} \tag{2.4}$$

mit den Parametern $\alpha \in \mathbb{R}^+$ $[1/\text{M}]$, $n > 1$ $[-]$ und $m > 0$ $[-]$ herangezogen, welcher mit $m = 1$ u.a. bereits in Haverkamp [32], Ahuja, Swartzendruber [3] und Endelmann et al. [22] erfolgreich Anwendung fand. In van Genuchten [61] wird schließlich aufgezeigt, dass (2.3) mit (2.4) für alle ganzzahligen Werte $k = m - 1 + \frac{1}{n}$ integrierbar ist und dass für $k = 0$ und entsprechend $m = 1 - \frac{1}{n} \in (0, 1)$ die hydraulische Leitfähigkeit durch

$$K(\Phi) = K_{\text{sat}} \sqrt{\Phi(\psi)} \Big(1 - \big(1 - \Phi(\psi)^{\frac{1}{m}} \big)^m \Big)^2 \tag{2.5}$$

angegeben werden kann. Fortan wird (2.5) mit

$$\Theta(\Phi) = \theta_{\text{res}} + \big(\theta_{\text{sat}} - \theta_{\text{res}}\big) \Phi(\psi) \tag{2.6}$$

und

$$\Phi(\psi) = \frac{1}{\big(1 + (-\alpha\psi)^n\big)^m} \tag{2.7}$$

als **van Genuchten-Mualem-Parametrisierung** bezeichnet.

Bemerkung 2.2
Durch Einsetzen von (2.7) in (2.5) und (2.6) lassen sich die Nichtlinearitäten auch direkt als Abbildungen in Abhängigkeit von der Druckhöhe ψ angeben. Für den ungesättigten Bereich $\psi<0$ gilt dann

$$K(\psi) = K_{sat}\frac{\Big(1-(-\alpha\psi)^{n-1}\big(1+(-\alpha\psi)^n\big)^{\frac{1-n}{n}}\Big)^2}{\big(1+(-\alpha\psi)^n\big)^{\frac{n-1}{2n}}} \tag{2.8}$$

und

$$\Theta(\psi) = \theta_{res}+\frac{\theta_{sat}-\theta_{res}}{\big(1+(-\alpha\psi)^n\big)^{\frac{n-1}{n}}}\,. \tag{2.9}$$

Im gesättigten Fall $\psi\geq 0$ vereinfachen sich die Funktionen weiterhin zu

$$K(\psi)=K_{sat} \quad \textit{und} \quad \Theta(\psi)=\theta_{sat}\,.$$

Beweis:
Für den volumetrischen Wassergehalt ist die Aussage trivial. Die Korrektheit von (2.8) wird mit Lemma A.5, Anhang A, nachgewiesen.

Abbildung 2.1 zeigt für ausgewählte Parameterwerte n die nach van Genuchten parametrisierte relative Leitfähigkeit $K_{\mathrm{rel}}:=\frac{K}{K_{\mathrm{sat}}}$ in Abhängigkeit des Sättigungsgehalts bzw. der mit α gewichteten Druckhöhe sowie die druckabhängige Sättigungsfunktion Φ.

In Roth [51] finden sich die in Tabelle 2.1 exemplarisch aufgeführten Parametersätze für Sand, Schlick und Lehm. Da insbesondere α und n nicht direkt (durch experimentelle Messungen) ermittelt werden können und auch das Messen von θ_{res} ein völliges Austrocknen der Bodenprobe voraussetzt, wurden diese Parameterwerte, zumindest teilweise, mittels inverser Probleme bestimmt. Aufgrund der Schlechtgestelltheit dieser Problemstellungen (vgl. nächstes Kapitel) und der sehr allgemein gehaltenen Bodentypen sollten die angegebenen Werte jedoch nur als (mathematisch bestimmte) Schätzung verstanden werden. Eine etwas detailliertere Darstellung findet sich in Parker, Kool, van Genuchten [45]. Dort wurden die Parameterwerte K_{sat} und θ_{sat} experimentell gemessen und θ_{res}, α und n mittels unterschiedlicher Parameteridentifizierungsansätze an die Messdaten der durchgeführter Durchflussexperimente angepasst. Aufgrund er bereits angesprochenen Identifizierungsschwierigkeiten sind hierbei jedoch keine exakten Werte sondern lediglich die ermittelten Wertespannen angegeben.

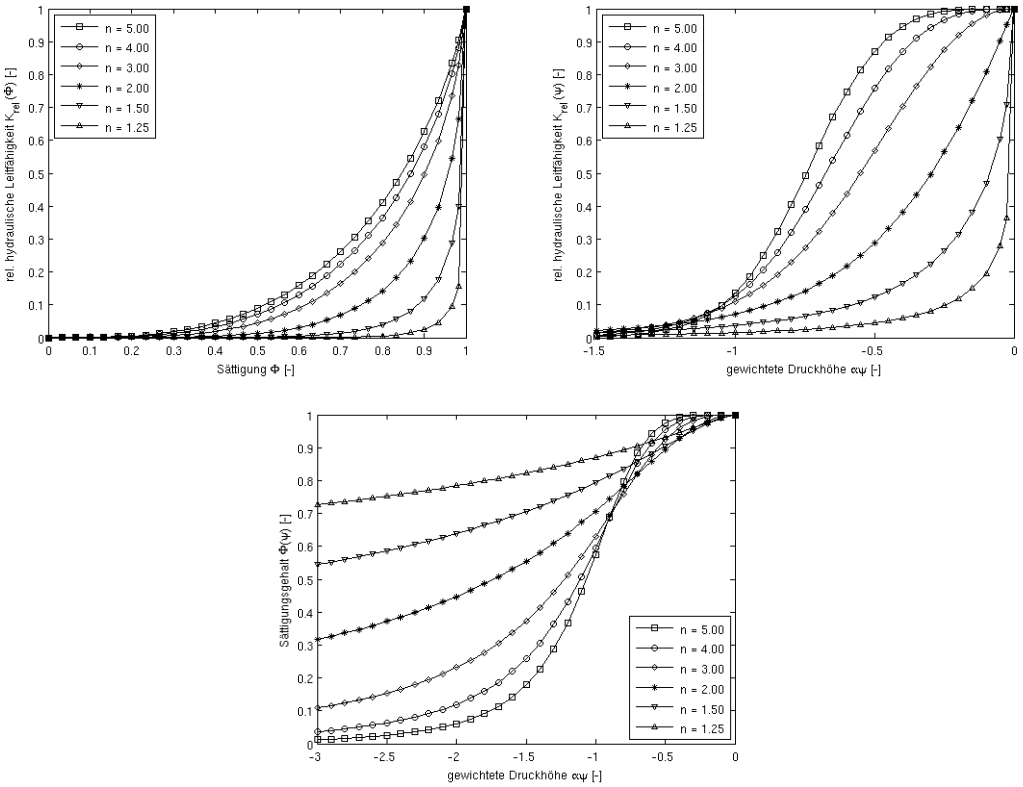

Abbildung 2.1: van Genuchten-Mualem-Parametrisierung

	θ_{res}	θ_{sat}	K_{sat}	α	n
	[–]	[–]	[cm s^{-1}]	[cm^{-1}]	[–]
Sand	0.03	0.32	$2.2{\cdot}10^{-3}$	$2.3{\cdot}10^{-2}$	4.17
Schlick	0.01	0.41	$1.0{\cdot}10^{-3}$	$7.0{\cdot}10^{-3}$	1.30
Lehm	0.00	0.43	$3.0{\cdot}10^{-4}$	$1.6{\cdot}10^{-2}$	1.25

Tabelle 2.1: van Genuchten-Mualem-Parametrisierung ausgewählter Bodenarten

2.1.5 Regularisierung

Bei der in Abschnitt 2.1.4 eingeführten van Genuchten-Mualem-Parametrisierung können bei ungeeignet gewählten Parameterwerten numerische Schwierigkeiten auftreten. Eine Ursache hierfür liegt im Ansteigen der in Gleichung (2.5) definierten hydraulischen Leitfähigkeitsfunktion $K(\Phi)$ nahe der Sättigung $\Phi = 1$. Die Ableitung nach der Sättigung Φ kann durch

$$K'(\Phi) = \frac{dK}{d\Phi}(\Phi) = \frac{K_{\text{sat}}}{2\sqrt{\Phi}}\Big(1-\big(1-\Phi^{\frac{1}{m}}\big)^{m}\Big)\Big(1+\big(5\Phi^{\frac{1}{m}}-1\big)\big(1-\Phi^{\frac{1}{m}}\big)^{m-1}\Big) \quad (2.10)$$

angegeben werden (vgl. Anhang A.1.1). Für fest gewählte $K_{\text{sat}} \in \mathbb{R}^+$, $n > 1$,

$m=1-\frac{1}{n}\in(0,1)$, gilt

$$\lim_{\Phi\searrow 1}\left(1-\Phi^{\frac{1}{m}}\right)^{m}=0^{+}\quad\text{und}\quad\lim_{\Phi\searrow 1}\left(1-\Phi^{\frac{1}{m}}\right)^{m-1}=+\infty$$

und damit schließlich

$$\lim_{\Phi\searrow 1}K'(\Phi)=+\infty\,.$$

Die hydraulische, vom Sättigungsgehalt abhängige Leitfähigkeitsfunktion $K(\Phi)$ steigt also nahe der Sättigung $\Phi=1$ für jeden nach dem Modell zulässigen Parametersatz, bestehend aus K_{sat}, n und α, beliebig stark an.

Betrachtet man die hydraulische Leitfähigkeitsfunktion direkt als Abbildung in Abhängigkeit der Druckhöhe ψ so wird in Anhang A.2.1 gezeigt, dass die entsprechende Ableitung durch

$$K'(\psi)=\frac{dK}{d\psi}(\psi)=\frac{1}{2}K_{\text{sat}}\sqrt{\Phi(\psi)}\,\frac{\alpha(n-1)(-\alpha\psi)^{n-1}}{1+(-\alpha\psi)^{n}}\left(1-\left(1-\Phi(\psi)^{\frac{1}{m}}\right)^{m}\right)\cdot\left[1-\left(1-\Phi(\psi)^{\frac{1}{m}}\right)^{m-1}\frac{(-\alpha\psi)^{n}-4}{1+(-\alpha\psi)^{n}}\right]$$

angegeben werden kann. Für fest gewählte $K_{\text{sat}},\alpha\in\mathbb{R}^{+}$, $n>1$, $m=1-\frac{1}{n}\in(0,1)$, folgt wegen

$$\lim_{\psi\searrow 0}\Phi(\psi)=\lim_{\psi\searrow 0}\frac{1}{\left(1+(-\alpha\psi)^{n}\right)^{m}}=1^{-}$$

auch hier

$$\lim_{\psi\searrow 0}\left(1-\Phi(\psi)^{\frac{1}{m}}\right)^{m}=0^{+}\quad\text{und}\quad\lim_{\psi\searrow 0}\left(1-\Phi(\psi)^{\frac{1}{m}}\right)^{m-1}=+\infty\,.$$

Zudem gilt wegen

$$\lim_{\psi\searrow 0}(-\alpha\psi)^{n}=\lim_{\psi\searrow 0}(-\alpha\psi)^{n-1}=0^{+}$$

auch

$$\lim_{\psi\searrow 0}\frac{(-\alpha\psi)^{n}-4}{1+(-\alpha\psi)^{n}}=-4\quad\text{und insbesondere}\quad\lim_{\psi\searrow 0}\frac{\alpha(n-1)(-\alpha\psi)^{n-1}}{1+(-\alpha\psi)^{n}}=0^{+}.$$

Bleibt der Grenzwert

$$\lim_{\psi\searrow 0}\frac{\alpha(n-1)(-\alpha\psi)^{n-1}}{1+(-\alpha\psi)^{n}}\left(1-\Phi(\psi)^{\frac{1}{m}}\right)^{m-1}\tag{2.11}$$

zu untersuchen. Durch Einsetzen der Sättigungsfunktion (2.7) gilt

$$\left(1-\Phi(\psi)^{\frac{1}{m}}\right)^{m-1} = \left(\frac{(-\alpha\psi)^n}{1+(-\alpha\psi)^n}\right)^{m-1} = \frac{\left(1+(-\alpha\psi)^n\right)^{\frac{1}{n}}}{(-\alpha\psi)^n}\,,$$

so dass (2.11) auch durch

$$\lim_{\psi\lneq 0} \frac{\alpha(n-1)(-\alpha\psi)^{n-2}}{\left(1+(-\alpha\psi)^n\right)^m}$$

angegeben werden kann. Hieraus folgt, dass (2.11) nur für $1<n<2$ bestimmt gegen $+\infty$ divergiert. Für $n=2$ konvergiert der Grenzwert hingegen gegen α und für $n>2$ gegen 0^+. Für die zu untersuchende Ableitung folgt damit schießlich

$$\lim_{\psi\lneq 0} K'(\psi) = \begin{cases} +\infty & , \text{ falls } 1<n<2\,, \\ 2\alpha K_{\text{sat}} & , \text{ falls } n=2\,, \\ 0^+ & , \text{ falls } n>2\,. \end{cases}$$

Im Folgenden werden zwei Regularisierungsansätze vorgestellt, bei denen die hydraulische Leitfähigkeitsfunktion im kritischen Bereich (nahe der Sättigung) durch ein Polynom ersetzt wird. Hierbei dient entweder der Sättigungsgrad oder die Druckhöhe als entscheidendes Regularisierungskriterium.

2.1.5.1 Sättigungsabhängige Regularisierung

Bei diesem Ansatz wird die hydraulische Leitfähigkeitsfunktion $K(\Phi)$ in (2.5) mittels eines Regularisierungsparameters $R_\Phi \in (0,1)$ für $\Phi \in (1-R_\Phi, 1]$ durch eine quadratische Funktion $k_{R_\Phi}(\Phi) \in \mathcal{P}^2$ approximiert. Es gilt

$$\begin{aligned} K_{R_\Phi}(\Phi) : [0,1] &\to [0, K_{\text{sat}}]\,, \\ \Phi &\mapsto \begin{cases} K(\Phi)\ , \text{ falls } \Phi \in [0, 1-R_\Phi]\,, \\ k_\Phi(\Phi)\ , \text{ falls } \Phi \in (1-R_\Phi, 1]\,. \end{cases} \end{aligned} \tag{2.12}$$

Die Freiheitsgrade von $k_\Phi(\Phi)$ werden dabei so gewählt, dass $K_{R_\Phi}(\Phi)$ (auch an der kritischen Stelle $\Phi = 1-R_\Phi$) stetig differenzierbar ist und $K_{R_\Phi}(1)=1$ gilt. Im Sinne einer einfachen Notation wird an dieser Stelle $R := R_\Phi$ und $k := k_{R_\Phi}$ gesetzt. Wird für $1-R \le \Phi \le 1$

$$z := \Phi - (1-R)\,, \quad z \in [0, R), \tag{2.13}$$

definiert, so muss für $k(z) := az^2 + bz + c$

$$\begin{aligned} k(0) &= c = K(1-R)\,, \\ k'(0) &= b = K'(1-R) \quad \text{und} \\ k(R) &= aR^2 + K'(1-R)R + K(1-R) = K_{\text{sat}} \end{aligned}$$

gefordert werden. Dies liefert

$$k(z) = \frac{K_{\text{sat}} - K(1-R) - K'(1-R)R}{R^2} z^2 + K'(1-R)z + K(1-R)$$

und damit schließlich

$$\begin{aligned} k(\Phi) = \frac{K_{\text{sat}} - K(1-R) - K'(1-R)R}{R^2} \big(\Phi - (1-R)\big)^2 \\ + K'(1-R)\big(\Phi - (1-R)\big) + K(1-R)\,. \end{aligned} \tag{2.14}$$

Definition 2.3
*Parametrisierung (2.12), einschießlich den Nichtlinearitäten (2.5) und (2.14), wird für $R \in (0,1)$ als **(sättigungsabhängige) ($\mathcal{P}^2$-)Regularisierung** der van Genuchten-Mualem-Leitfähigkeitsfunktion (2.5) bezeichnet.*

Abbildung 2.2 zeigt exemplarisch für ausgewählte n (und damit auch m) die (sättigungsabhängige) van Genuchten-Mualem-Parametrisierung der hydraulischen Leitfähigkeit (gepunktet) sowie die durch $K_{R,\text{rel}} := \frac{K_R}{K_{\text{sat}}}$ festgelegte relative $\mathcal{P}^2$-regulierte Leitfähigkeitsfunktion $K_{R,\text{rel}}(\Phi)$ (durchgezogen) mit Regularisierunggrad $R = 0.05$.

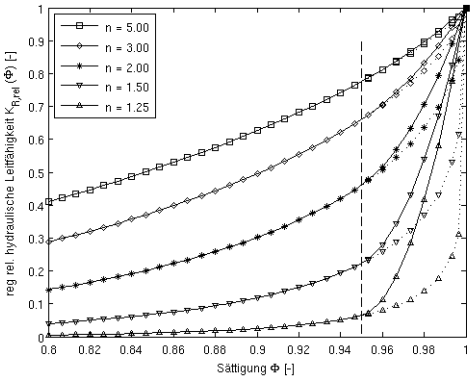

Abbildung 2.2: ($\mathcal{P}^2$-)Regulierte relative Leitfähigkeit $K_{R_\Phi,\text{rel}}(\Phi)$ mit $R_\Phi = 0.05$

Es sei an dieser Stelle angemerkt, dass der in Abbildung 2.2 und entsprechend in den nachfolgenden graphischen Ausgaben verwendete Regularisierungsgrad R

nur für eine bessere Darstellung groß gewählt wurde. Typischerweise wird ein kleinerer Parameter ($0<R\ll 0.05$) Verwendung finden (vgl. Tabelle 2.2).

Neben der strengen Monotonie und der Konkavität (vgl. Anhang A, Satz A.2) hat die van Genuchten-Mualem-Parametrisierung (2.5) die Eigenschaft, dass sich zwei mit unterschiedlichen Parameterwerten $m_1 \neq m_2$ definierte Funktionen auf $(0,1)$ nicht schneiden (vgl. Anhang A, Satz A.3). Die beiden nachfolgenden Sätze belegen, dass dies auch für die $\mathcal{P}^2$-regularisierte Leitfähigkeitsfunktion gilt.

Satz 2.4
Sei eine nach (2.5) definierte Leitfähigkeitsfunktion gegeben. Dann ist die zugehörige sättigungsabhängige ($\mathcal{P}^2$)-Regularisierung (2.12) mit beliebigem Regularisierungsgrad $R\in(0,1)$ streng monoton wachsend und konkav.

Beweis:
Nach Definition ist die regularisierte Leitfähigkeitsfunktion $K_R(\Phi)$ auf $(0,1)$ differenzierbar. Ihre Ableitung ist durch

$$K_R'(\Phi) = \begin{cases} K'(\Phi)\ , & \text{für } \Phi\in(0,1-R]\,, \\ k'(\Phi)\ , & \text{für } \Phi\in(1-R,1)\,, \end{cases}$$

mit der in (2.10) angegebenen Ableitung $K'(\Phi)$ und

$$k'(\Phi) = 2\frac{K_{\text{sat}}-K(1-R)-K'(1-R)R}{R^2}\big(\Phi-(1-R)\big)+K'(1-R)$$

gegeben. Unter Berücksichtigung der Monotonie und Konkavität der van Genuchten-Mualem-Parametrisierung (vgl. Satz A.2) folgt wegen

$$K'(1-R)>0 \quad \text{und} \quad K_{\text{sat}} > K(1-R)+K'(1-R)R$$

direkt $k'(\Phi)>0$, $\Phi\in[1-R,1)$, und damit die erste Behauptung. Die Konkavität von $K_R(\Phi)$ folgt aus der Konkavität von $K(\Phi)$, dem differenzierbaren Übergang

$$\lim_{\Phi\overset{>}{\to}1-R} k' = K'(1-R)$$

und der Tatsache, dass

$$k''(\Phi) = 2\,\frac{K_{\text{sat}}-K(1-R)-K'(1-R)R}{R^2} > 0$$

für alle $\Phi\in(1-R,1)$ gilt.

□

Satz 2.5
Seien zwei nach van Genuchten-Mualem definierte hydraulische Leitfähigkeitsfunktionen

$$K_i(\Phi) := K_{sat}\sqrt{\Phi}\Big(1-\big(1-\Phi^{\frac{1}{m_i}}\big)^{m_i}\Big)^2\,,\quad i=1,2\,,$$

mit unterschiedlichen Parameterwerten $m_1 \neq m_2$ und die für ein beliebig vorgegebenes $R\in(0,1)$ zugehörigen Regularisierungen

$$K_{R,i}(\Phi) : [0,1] \to [0,K_{sat}]\,,$$
$$\Phi \mapsto \begin{cases} K_i(\Phi)\,, & \textit{für } \Phi \in [0,1-R]\,, \\ k_i(\Phi)\,, & \textit{für } \Phi \in (1-R,1]\,, \end{cases}$$

mit

$$k_i(\Phi) := \frac{K_{sat}-K_i(1-R)-K_i'(1-R)R}{R^2}\big(\Phi-(1-R)\big)^2 + K_i'(1-R)\big(\Phi-(1-R)\big)+K_i(1-R)$$

gegeben. Dann gilt

$$K_{R,1}(\Phi) \neq K_{R,2}(\Phi) \quad \forall\,\Phi\in(0,1)\,.$$

Beweis:
Auf $(0,1-R]$ gilt $K_{R,i}\equiv K_i$, so dass wegen der Eindeutigkeit der van Genuchten-Mualem-Parametrisierung die Behauptung für $\Phi\in(0,1-R]$ sofort aus Satz A.3 folgt. Bleibt zu zeigen, dass $k_1(\Phi)\neq k_2(\Phi)$ für alle $\Phi_s\in(1–R,1)$ gilt. Angenommen es existiert ein $\Phi_s\in(1-R,1)$ mit

$$k_1(\Phi_s) = k_2(\Phi_s) \tag{2.15}$$

und es werden die Abbildungen k_i, $i=1,2$, analog zu (2.13) als Funktionen in Abhängigkeit von z betrachtet, dann folgt aus (2.15) für $z_s := \Phi_s-(1-R)\in(0,R)$ direkt

$$\frac{K_2(1-R)-K_1(1-R)+\big(K_2'(1-R)-K_1'(1-R)\big)R}{R^2}z_s^2 + \big(K_1'(1-R)-K_2'(1-R)\big)z_s + \big(K_1(1-R)-K_2(1-R)\big) = 0\,. \tag{2.16}$$

Da für beide Regularisierungen $k_i(R)=1$ gilt, kann (2.16) mittels Polynomdivision zu

$$\frac{K_2(1-R)-K_1(1-R)+\big(K_2'(1-R)-K_1'(1-R)\big)R}{R^2}z_s + \frac{K_2(1-R)-K_1(1-R)}{R} = 0$$

vereinfacht werden. Ohne Einschränkung der Allgemeinheit wird $m_1 > m_2$ vorausgesetzt. Ein Widerspruch zur Annahme (2.15) ist gefunden, falls hierfür

$$K_2(1-R)-K_1(1-R)+\big(K_2'(1-R)-K_1'(1-R)\big)R<0 \tag{2.17}$$

und damit wegen $K_1(1-R)>K_2(1-R)$ (vgl. Bemerkung A.4)

$$z_s=\frac{\big(K_1(1-R)-K_2(1-R)\big)R}{K_2(1-R)-K_1(1-R)+\big(K_2'(1-R)-K_1'(1-R)\big)R}<0$$

gezeigt werden kann. Durch Einsetzen der ersten Ableitungen

$$K_i'(\Phi)=\frac{K_{\text{sat}}}{2\sqrt{\Phi}}\Big(1-\big(1-\Phi^{\frac{1}{m_i}}\big)^{m_i}\Big)\Big(1+\big(5\Phi^{\frac{1}{m_i}}-1\big)\big(1-\Phi^{\frac{1}{m_i}}\big)^{m_i-1}\Big)\,,\quad i=1,2\,,$$

(vgl. Anhang A.1.1) lässt sich (2.17) mittels

$$\begin{aligned}T_i(R):=\,&2(1-R)\Big(1-\big(1-(1-R)^{\frac{1}{m_i}}\big)^{m_i}\Big)^2+R\Big(1-\big(1-(1-R)^{\frac{1}{m_i}}\big)^{m_i}\Big)\\&\cdot\Big(1+\big(5(1-R)^{\frac{1}{m_i}}-1\big)\big(1-(1-R)^{\frac{1}{m_i}}\big)^{m_i-1}\Big)\end{aligned}$$

auch durch

$$T_1(R)>T_2(R)$$

angeben. Durch einfaches Umstellen gilt

$$\begin{aligned}T_i(R)&=2-R+(3R-4)\big(1-(1-R)^{\frac{1}{m_i}}\big)^{m_i}+2(1-R)\big(1-(1-R)^{\frac{1}{m_i}}\big)^{2m_i}\\&\quad+R\big(5(1-R)^{\frac{1}{m_i}}-1\big)\Big(\big(1-(1-R)^{\frac{1}{m_i}}\big)^{m_i-1}-\big(1-(1-R)^{\frac{1}{m_i}}\big)^{2m_i-1}\Big)\\&=2-R+\big(1-(1-R)^{\frac{1}{m_i}}\big)^{m_i}\,\frac{2(2+R)(1-R)^{\frac{1}{m_i}}+R-2}{1-(1-R)^{\frac{1}{m_i}}}\\&\quad+\big(1-(1-R)^{\frac{1}{m_i}}\big)^{2m_i}\,\frac{2-R-(2+3R)(1-R)^{\frac{1}{m_i}}}{1-(1-R)^{\frac{1}{m_i}}}\end{aligned}$$

und mit

$$g_i(R):=-3-\frac{5}{2}R+\frac{4R}{1-(1-R)^{\frac{1}{m_i}}}$$

auch

$$\begin{aligned}T_i(R)=\,&2-R+\big(1-(1-R)^{\frac{1}{m_i}}\big)^{m_i}\Big(\frac{R}{2}-1+g_i(R)\Big)\\&+\big(1-(1-R)^{\frac{1}{m_i}}\big)^{2m_i}\Big(\frac{R}{2}-1-g_i(R)\Big)\end{aligned}$$

und damit

$$T_i(R) = 2-R+\left(1-(1-R)^{\frac{1}{m_i}}\right)^{2m_i}\left[\frac{1}{\left(1-(1-R)^{\frac{1}{m_i}}\right)^{m_i}}\right.$$
$$\left.\cdot\left(\left(\frac{R}{2}-1\right)\left(1+\left(1-(1-R)^{\frac{1}{m_i}}\right)^{m_i}\right)+g_i(R)\left(1-\left(1-(1-R)^{\frac{1}{m_i}}\right)^{m_i}\right)\right)\right].$$

Die Behauptung folgt schließlich aus der Tatsache, dass für alle $R\in(0,1)$

$$\left(1-(1-R)^{\frac{1}{m_1}}\right)^{m_1} < \left(1-(1-R)^{\frac{1}{m_2}}\right)^{m_2}$$

und $g_1(R)>g_2(R)$ gilt, sowie wegen

$$(1-R)^{\frac{1}{m_i}} < 1-R \quad\Leftrightarrow\quad \frac{R}{1-(1-R)^{\frac{1}{m_i}}} < 1 \quad\Leftrightarrow\quad -4+\frac{4R}{1-(1-R)^{\frac{1}{m_i}}} < 0$$

stets

$$\left(\frac{R}{2}-1\right)\left(1+\left(1-(1-R)^{\frac{1}{m_i}}\right)^{m_i}\right)+g_i(R)\left(1-\left(1-(1-R)^{\frac{1}{m_i}}\right)^{m_i}\right)$$
$$< \left(\frac{R}{2}-1+g(R)\right)\left(1-\left(1-(1-R)^{\frac{1}{m_i}}\right)^{m_i}\right)$$
$$< \frac{R}{2}-1+g(R) < -4+\frac{4R}{1-(1-R)^{\frac{1}{m_i}}} < 0$$

erfüllt ist.

□

Durch Einsetzen von (2.7) in (2.12) lässt sich die approximierte Leitfähigkeitsfunktion auch direkt in Abhängigkeit von ψ beschreiben. Abbildung 2.3 (links) zeigt hierzu exemplarisch, für ausgewählte n, die druckabhängige van Genuchten-Mualem-Parametrisierung (gepunktet) sowie die zugehörige relative ($\mathcal{P}^2$-)Regularisierung $K_{R,\mathrm{rel}}(\psi)=\frac{K_R(\Phi(\psi))}{K_{\mathrm{sat}}}$ (durchgezogen) für $R=0.05$. Zu beachten ist hierbei die unterschiedliche Lage der Regularisierungsgrenze (senkrechte Trennlinie). Diese wurde auf Basis der Sättigung (2.7) gewählt (vgl. waagrechte Linie in Abbildung 2.3 (rechts)) und ist demzufolge abhängig vom Parameter n.

Bemerkung 2.6

Die in Satz 2.5 nachgewiesene Eindeutigkeit der nach (2.12) definierten Regularisierung $K_R(\Phi)$ gilt i.A. nicht mehr für die zugehörige druckabhängige Formulierung $K_R(\psi)$. Vergleiche hierzu die in Abbildung 2.4 dargestellte relative ($\mathcal{P}^2$-)regulierte Leitfähigkeitsfunktion $K_{R,rel}$ für $R=0.1$. Da jedoch auch die ursprüngliche van Genuchten-Mualem-Parametrisierung $K(\psi)$ unter Verwendung

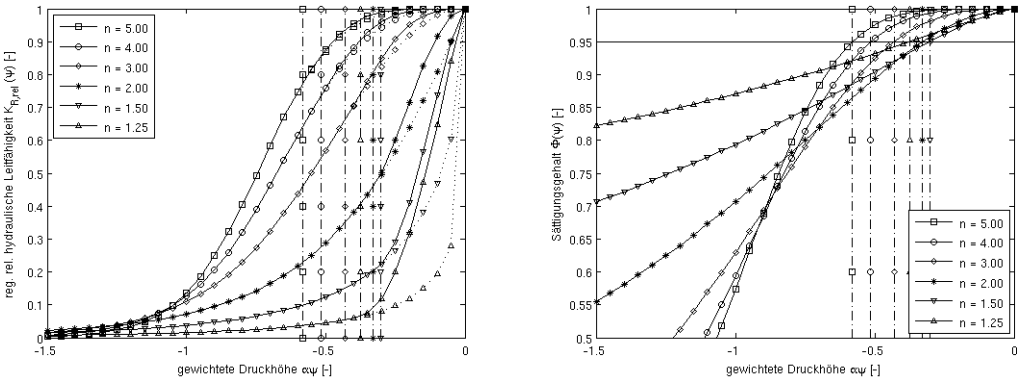

Abbildung 2.3: ($\mathcal{P}^2$-)Regulierte relative Leitfähigkeit $K_{R_\Phi,\mathrm{rel}}(\psi)$ mit $R_\Phi = 0.05$

verschiedener Parameterwerte n einen Schnittpunkt aufweist (vgl. Satz A.10, Anhang A) und vermutet wird, dass zusätzliche Schnittpunkte durch die Regularisierung nur für groß gewählte Regularisierungsparameter R auftreten, wird diesem Sachverhalt an dieser Stelle keine Bedeutung zugeordnet.

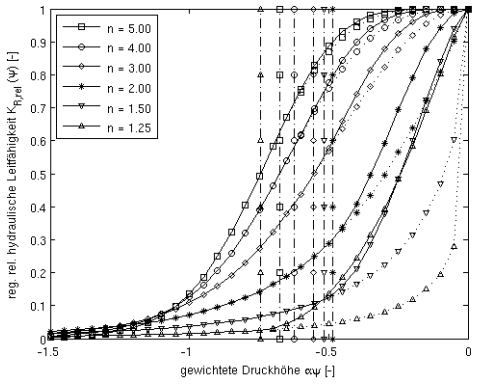

Abbildung 2.4: ($\mathcal{P}^2$-)Regulierte relative Leitfähigkeit $K_{R_\Phi,\mathrm{rel}}(\psi)$ mit $R_\Phi = 0.10$

Eine weitere numerische Schwierigkeit der druckabhängigen van Genuchten-Mualem-Parametrisierung ist für $n < 2$ neben dem beliebig starken Anwachsen der hydraulischen Leitfähigkeitsfunktion (2.8) nahe der Sättigung der unglatte Übergang zum gesättigen Bereich $\psi \geq 0$. Da dieser Sachverhalt bei der in (2.12) vorgestellten (sättigungsabhängigen) Regularisierung nicht berücksichtigt wurde, wird ein weiterer Approximationsansatz vorgestellt.

2.1.5.2 Druckabhängige Regularisierung

Im folgenden Abschnitt wird eine Regularisierung der hydraulischen Leitfähigkeit mit glatten Übergängen sowohl an der Regularisierungsgrenze als auch am Wechselpunkt gesättigt-/ungesättigter Bereich hergeleitet. Hierzu wird direkt die in (2.8) definierte, von der Druchhöhe abhängige, van Genuchten-Mualem-Leitfähigkeitsfunktion betrachtet und entsprechend für $\psi \in (R_\psi, 0)$, $R_\psi \in \mathbb{R}^-$, durch ein Polynom k_{R_ψ} approximiert. Es gilt

$$K_{R_\psi}(\psi) : \mathbb{R} \to (0, K_{\text{sat}}]\,, \qquad \psi \mapsto \begin{cases} K(\psi)\ , & \text{falls } \psi \in (-\infty, R_\psi]\,, \\ k_\psi(\psi)\ , & \text{falls } \psi \in (R_\psi, 0)\,, \\ K_{\text{sat}}\ , & \text{falls } \psi \geq 0\,. \end{cases} \tag{2.18}$$

Im Sinne einer einfachen Notation wird auch hier $R := R_\psi$ und $k := k_{R_\psi}$ eingeführt. Aufgrund der geforderten Bedingungen

$$k(R) = K(R)\,, \quad k'(R) = K'(R)\,, \quad k(0) = K_{\text{sat}} \quad \text{und} \quad k'(0) = 0 \tag{2.19}$$

ist ein Polynom dritten Grades $k(\psi) \in \mathcal{P}^3$ mittels Hermite-Interpolation (vgl. z.B. van Loan [62]) zu bestimmen. Unter Verwendung von

$$k(\psi) := a + b(\psi - R) + c(\psi - R)^2 + d(\psi - R)^2\psi$$

und entsprechend

$$k'(\psi) := b + 2c(\psi - R) + d\big(2(\psi - R)\psi + (\psi - R)^2\big)$$

folgt aus (2.19) direkt

$$a = K(R)\,, \quad b = K'(R)\,, \quad a - Rb + R^2c = K_{\text{sat}} \quad \text{und} \quad b - 2cR + R^2d = 0$$

und damit schließlich

$$\begin{aligned} a &= K(R)\,, \\ b &= K'(R)\,, \\ c &= \frac{K_{\text{sat}} - K(R) + K'(R)R}{R^2} \quad \text{und} \\ d &= \frac{K'(R)R + 2\big(K_{\text{sat}} - K(R)\big)}{R^3}\,. \end{aligned}$$

Die notwendige Ableitung $K'(R)$ kann gemäß Anhang A durch

$$K'(\psi) = \frac{dK}{d\psi}(\psi) = \frac{1}{2} K_{\text{sat}} \sqrt{\Phi(\psi)} \, \frac{\alpha(n-1)(-\alpha\psi)^{n-1}}{1+(-\alpha\psi)^n} \Big(1-\big(1-\Phi(\psi)^{\frac{1}{m}}\big)^m\Big) \cdot \left[1-\big(1-\Phi(\psi)^{\frac{1}{m}}\big)^{m-1} \frac{(-\alpha\psi)^n - 4}{1+(-\alpha\psi)^n}\right] \quad (2.20)$$

angegeben werden.

Definition 2.7
Parametrisierung (2.18), mit Nichtlinearitäten (2.8), (2.20) und

$$k(\psi) := K(R)+K'(R)(\psi-R)+\frac{K_{sat}-K(R)+K'(R)R}{R^2}(\psi-R)^2 + \frac{K'(R)R+2\big(K_{sat}-K(R)\big)}{R^3}(\psi-R)^2\psi \quad (2.21)$$

wird für $R \in \mathbb{R}^-$ als ***(druckabhängige) ($\mathcal{P}^3$-)Regularisierung*** *der van Genuchten-Mualem-Leitfähigkeitsfunktion $K(\psi)$ bezeichnet.*

Abbildung 2.5 zeigt exemplarisch für ausgewählte Parameterwerte n neben der relativen van Genuchten-Mualem-Leitfähigkeitsfunktion $K_{\text{rel}}(\psi)$ (gepunktet) die zum vorgegebenen Regularisierungsgrad $R=-\frac{1}{3\alpha}$ vorgestellte ($\mathcal{P}^3$-)regularisierte relative Leitfähigkeit $K_{R,\text{rel}}(\psi)$ (durchgezogen). Es ist deutlich zu erkennen, daß für $n>2$ der ursprüngliche Kurvenverlauf sehr gut approximiert wird. Für kritische Werte $n<2$ hingegen wird eine gute Näherung zu Gunsten eines glatten Übergangs aufgegeben. Es sei an dieser Stelle erneut darauf hingewiesen, dass R nur für eine gute Darstellung groß gewählt wurde. Typischerweise wird ein (betragsmäßig) deutlich kleinerer Wert Verwendung finden.

Satz 2.8
Sei eine nach (2.8) definierte Leitfähigkeitsfunktion gegeben. Dann ist die zugehörige drucksabhängige ($\mathcal{P}^3$)-Regularisierung für beliebigen Regularisierungsgrad $R\in\mathbb{R}^-$ streng monoton wachsend.

Beweis:
Aufgrund des strengen Anwachsens der van Genuchten-Mualem-Leitfähigkeitsfunktion $K(\psi)$ (vgl. Anhang A, Satz A.6) und der Stetigkeit von (2.18) genügt

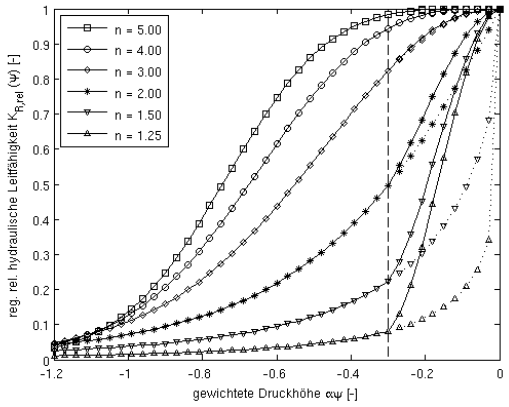

Abbildung 2.5: ($\mathcal{P}^3$-)Regulierte relative Leitfähigkeit $K_{R_\psi,\mathrm{rel}}(\psi)$ mit $R_\psi = -\frac{1}{3\alpha}$

es die Monotonie des kubischen Polynoms (2.21) für $\psi \in (R, 0)$ zu untersuchen. Es gilt

$$\begin{aligned} k'(\psi) &:= K'(R) + 2\frac{K_{\mathrm{sat}} - K(R) + K'(R)R}{R^2}(\psi - R) \\ &\quad + \frac{K'(R)R + 2\big(K_{\mathrm{sat}} - K(R)\big)}{R^3}(\psi - R)(3\psi - R) \\ &= K'(R) + \bigg(2\frac{K_{\mathrm{sat}} - K(R) + K'(R)R}{R^2} \\ &\quad + \frac{K'(R)R + 2\big(K_{\mathrm{sat}} - K(R)\big)}{R^3}(3\psi - R)\bigg)(\psi - R)\,. \end{aligned} \tag{2.22}$$

Für $n \leq 2$ kann aus der Konkavität von $K(\psi)$ (vgl. Anhang A, Satz A.7) direkt

$$\frac{K_{\mathrm{sat}} - K(R) + K'(R)R}{R^2} > 0 \quad \text{und} \quad \frac{K'(R)R + 2\big(K_{\mathrm{sat}} - K(R)\big)}{R^3} < 0 \tag{2.23}$$

und damit

$$\begin{aligned} k'(\psi) &> K'(R) + \bigg(2\frac{K_{\mathrm{sat}} - K(R) + K'(R)R}{R^2} - \frac{K'(R)R + 2\big(K_{\mathrm{sat}} - K(R)\big)}{R^2}\bigg)(\psi - R) \\ &> K'(R) + K'(R)(\psi - R) > K'(R) > 0 \end{aligned} \tag{2.24}$$

gefolgert werden. Ist $n > 2$ und der Regularisierungsgrad R so (betragsmäßig groß) gewählt, dass $K(\psi)$ bei $\psi = R$ lokal konkav ist, so kann analog zu (2.23) und (2.24) argumentiert werden. Bleibt also der konvexe Fall $n > 2$ und

$$T(R) := K_{\mathrm{sat}} - K(R) + K'(R)R < 0$$

zu untersuchen. Da hierfür aber (2.22) direkt durch

$$\begin{aligned} k'(\psi) &= K'(R)+\left[\frac{T(R)}{R^2}\left(2+\frac{3\psi-R}{R}\right)+\frac{K_{\text{sat}}-K(R)}{R^3}(3\psi-R)\right](\psi-R) \\ &> K'(R)+\left[\frac{T(R)}{R^2}-\frac{K_{\text{sat}}-K(R)}{R^2}\right](\psi-R) = K'(R)+\frac{K'(R)}{R}(\psi-R) \\ &> K'(R)-K'(R) = 0 \end{aligned}$$

abgeschätzt werden kann, folgt bereits die Behauptung.

□

Satz 2.9
Sei eine nach (2.8) definierte Leitfähigkeitsfunktion gegeben. Dann ist die zugehörige drucksabhängige ($\mathcal{P}^3$)-Regularisierung mit beliebigem Regularisierungsgrad $R\in\mathbb{R}^-$ genau dann konkav, wenn $1<n\leq 2$ gilt.

Beweis:
Sei zunächst der Fall $1<n\leq 2$ betrachet. Für $\psi\in(-\infty,R]$ kann hierfür wieder mit der Konkavität von (2.8) argumentiert werden. Da definitionsgemäß ein glatter Übergang $K'(R)=\lim_{\psi\nearrow 0} k'(\psi)$ vorliegt, bleibt das Vorzeichen von

$$k''(\psi) = 2\frac{K_{\text{sat}}-K(R)+K'(R)R}{R^2}+2\frac{K'(R)R+2\big(K_{\text{sat}}-K(R)\big)}{R^3}(3\psi-2)$$

für $\psi\in(R,0)$ zu untersuchen. Unter Verwendung von (2.23) folgt

$$k''(\psi) > 2\frac{K_{\text{sat}}-K(R)+K'(R)R}{R^2}-4\frac{K'(R)R+2\big(K_{\text{sat}}-K(R)\big)}{R^3}$$

und mit

$$T(R) := K_{\text{sat}}-K(R)+K'(R)R > 0$$

schließlich

$$k''(\psi) > \left(2-\frac{4}{R}\right)\frac{T(R)}{R^2}+\frac{8}{R^3}\big(K(R)-K_{\text{sat}}\big) > 0\,.$$

Für $n>2$ ist direkt aus $k'(R)>0$ und $\lim_{\psi\nearrow 0} k'(\psi)=0$ die Konvexität nahe der Sättigung ersichtlich.

□

Unabhängig von den beiden definierten Regularisierungsansätzen kann die Kirchhoff-Transformation als Regularisierung verwendet werden. Im folgenden Abschnitt wird diese kurz vorgestellt und mit ihr ein erster Vergleich der unterschiedlichen Parametrisierungen aufgezeigt.

2.1.5.3 Kirchhoff-Transformation

Für die Richards-Gleichung (2.2) lassen sich in der Literatur, neben der van Genuchten-Mualem-Parametrisierung (2.5), unterschiedliche funktionale Ansätze zur Beschreibung der hydraulischen Leitfähigkeit finden. Um diese besser vergleichen zu können empfiehlt es sich die sogenannte Kirchhoff-Transformation

$$\begin{aligned} \mathcal{K} : \mathbb{R} &\to \mathbb{R}\,, \\ \psi &\mapsto \int_0^{\psi} K\big(\Phi(s)\big) ds\,, \end{aligned}$$

durchzuführen (vgl. z.B. Alt, Luckhaus [4] oder auch Radu, Pop, Knabner [48]). Aus der Tatsache, dass $K\big(\Phi(s)\big)$ stets größer als Null ist, folgt die strenge Monotonie dieser Abbildung und damit die Invertierbarkeit. Somit kann unter Verwendung einer neuen Variablen

$$u := \mathcal{K}(\psi)$$

und

$$\begin{aligned} b(u) &:= \Phi \circ \mathcal{K}^{-1}(u)\,, \\ k\big(b(u)\big) &:= K \circ \Phi \circ \mathcal{K}^{-1}(u)\,, \end{aligned}$$

wegen

$$\mathcal{K}'(\psi) = K\big(\Phi(\psi)\big) \quad \text{und} \quad \nabla u = K\big(\Phi(\psi)\big)\nabla\psi$$

die Richardsgleichung (2.2) auch durch

$$\partial_t b(u) - \nabla \cdot \Big(\nabla u + k\big(b(u)\big) e_z\Big) = 0\,, \tag{2.25}$$

e_z entgegen der Gravitation gerichteter Einheitsvektor, beschrieben werden. Obwohl (2.25) weiterhin eine degenerierende PDE bleibt, kann aufgrund der nun linear auftretenden Diffusion eine (weitere) Regularisierung erwartet werden.

Im Fall der van Genuchten-Mualem-Leitfähigkeitsfunktion $K(\Phi)$ gilt entsprechend für den ungesättigten Bereich $\psi < 0$

$$\mathcal{K}(\psi) = \int_0^{\psi} K_{\text{sat}} \sqrt{\Phi(s)} \Big(1 - \big(1 - \Phi(s)^{\frac{1}{m}}\big)^m\Big)^2 ds \tag{2.26}$$

und $\mathcal{K}(\psi) = K_{\text{sat}}\psi$, sonst. Da (2.26) wohl nicht algebraisch bestimmt werden kann, wird die zusammengesetzte Trapezregel (vgl. z.B. Werner [64], Kapitel 4.2) zur

numerischen Berechnung verwendet. Hierfür sei für ein $\psi_{\min} \in \mathbb{R}^-$ das Intervall $[\psi_{\min}, 0]$ in ein diskretes, äquidistantes Gitter mit den Gitterpunkten

$$\psi_j := \frac{j}{N}\psi_{\min}\,, \quad j=0,\ldots,N,$$

zerlegt, so dass für $\psi \in [\psi_l, \psi_{l-1})$, $l \in \{1,\ldots,N\}$, wegen

$$\mathcal{K}(\psi) = \sum_{j=1}^{l-1} \int_{\psi_{j-1}}^{\psi_j} K\big(\Phi(s)\big)ds + \int_{\psi_{l-1}}^{\psi} K\big(\Phi(s)\big)ds$$

direkt

$$\mathcal{K}_h(\psi) := -\frac{\psi_{\min}}{N}\left[\frac{1}{2}K_{\text{sat}} + \sum_{j=1}^{l-2} K\big(\Phi(\psi_j)\big) + \frac{1}{2}K\big(\Phi(\psi_{l-1})\big)\right] + \frac{\psi-\psi_{l-1}}{2}\Big(K(\psi) + K(\psi_{l-1})\Big)$$

und damit

$$\mathcal{K}_h(\psi_l) = -\frac{\psi_{\min}}{N}\left[\frac{1}{2}K_{\text{sat}} + \sum_{j=1}^{l-1} K\big(\Phi(\psi_j)\big) + \frac{1}{2}K\big(\Phi(\psi_l)\big)\right]$$

gilt.

Abbildung 2.6 zeigt für ausgewählte Parameterwerte n neben der van Genuchten-Mualem-Leitfähigkeitsfunktion $K(\psi)$ die Graphen der diskreten Kirchhoff-Transformationen $\mathcal{K}_h(\psi)$ ($N = 10^4$, $\psi_{\min} = -\frac{3}{2\alpha}$). Es ist deutlich zu erkennen, dass im Gegensatz zur ursprünglichen Parametrisierung die Transformation einen glatten Übergang vom ungesättigten in den gesättigten Bereich aufweist.

Die Kirchhoff-Transformation der in (2.12) und (2.14) definierten ($\mathcal{P}^2$-)Regularisierung kann auf analoge Weise angegeben werden. Da nach Lemma A.1 die Sättigungsfunktion (2.7) streng monoton mit Wertebereich $(0, 1]$ ist, kann hierfür die durch $\Phi(\psi_{R_\Phi}) = 1 - R_\Phi$ eindeutig definierte Druckhöhe ψ_{R_Φ} als Regularisierungsgrad verwendet werden. Es gilt

$$\mathcal{K}_{R_\Phi}(\psi) = \begin{cases} K_{\text{sat}}\psi & , \text{ für } \quad \psi \geq 0\,, \\ \displaystyle\int_0^{\psi} k_\Phi\big(\Phi(s)\big)ds & , \text{ für } \quad \psi \in (\psi_{R_\Phi}, 0)\,, \\ \displaystyle\int_0^{\psi_{R_\Phi}} k_\Phi\big(\Phi(s)\big)ds + \int_{\psi_{R_\Phi}}^{\psi} K\big(\Phi(s)\big)ds & , \text{ für } \quad \psi \leq \psi_{R_\Phi}\,. \end{cases} \tag{2.27}$$

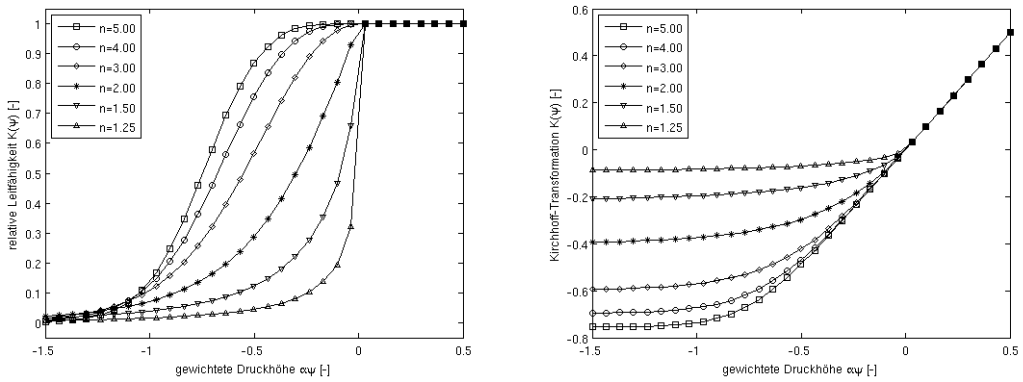

Abbildung 2.6: Kirchhoff-Transformation $\mathcal{K}(\psi)$ von $K(\psi)$

Für die mit (2.18) und (2.21) definierte ($\mathcal{P}^3$-)Regularisierung gilt entsprechend

$$\mathcal{K}_{R_\psi}(\psi) = \begin{cases} K_{\text{sat}}\psi & , \text{ für } \quad \psi \geq 0\,, \\ \displaystyle\int_0^{\psi} k_\psi\big(\Phi(s)\big)ds & , \text{ für } \quad \psi \in (R_\psi, 0)\,, \\ \displaystyle\int_0^{R_\psi} k_\psi\big(\Phi(s)\big)ds + \int_{R_\psi}^{\psi} K\big(\Phi(s)\big)ds & , \text{ für } \quad \psi \leq R_\psi\,. \end{cases} \tag{2.28}$$

Werden schießlich die in (2.27) und (2.28) definierten Transformationen mit (2.26) verglichen, so können qualitative Aussagen bzgl. der Regularisierungsfehler getroffen werden. Da sich $\mathcal{K}_{R_\Phi}$ und $\mathcal{K}_{R_\psi}$, bei angepassten Integralgrenzen $R_\psi = \psi_{R_\Phi}$, nur durch die Integrale über k_Φ bzw. k_ψ zu $\mathcal{K}$ unterscheiden, also

$$\frac{d\mathcal{K}}{d\psi} = \frac{d\mathcal{K}_{R_\Phi}}{d\psi} = \frac{d\mathcal{K}_{R_\psi}}{d\psi}$$

für alle $\psi \leq R_\psi$ und $\psi > 0$ gilt, genügt es die Regularisierungsfehler durch

$$r_{R_\psi} := \big\|K(\psi) - K_{R_\psi}(R)\big\|_{2,(R_\psi,0)} \quad \text{und} \quad r_{R_\Phi} := \big\|K(\psi) - K_{R_\Phi}(R)\big\|_{2,(R_\psi,0)} \tag{2.29}$$

zu beschreiben.

Tabelle 2.2 gibt für eine Auswahl unterschiedlich hoher Regularisierungsgrade R_ψ und ausgewählter Parameterwerte n die durch $R_\Phi = 1 - \Phi(R_\psi)$ festegelegten, auf der Sättigung basierenden, Regularisierungsgrade an. Die folgenden Tabellen 2.3 und 2.4 zeigen schließlich die zugehörigen, numerisch ermittelten ($N = 10^4$, $\psi_{\min} = -\frac{3}{2\alpha}$) Residuen (2.29). Wie zu erwarten war stellt die ($\mathcal{P}^3$-)Regularisierung (2.18) im Vergleich zur ($\mathcal{P}^2$-)Regularisierung (2.12) für kleine Parameterwerte n

eine schlechtere und für größere n eine (deutlich) bessere Approximation der van Genuchten-Mualem-Leitfähigkeit (2.5) dar. Überraschend ist jedoch, dass für (sehr) große n wiederum die ($\mathcal{P}^2$-)Regularisierung eine (geringfügig) bessere Näherung liefert. Ursache hierfür wird sein, dass die van Genuchten-Mualem-Parametrisierung für diese Wahl von n nahe der Sättigung annähernd konstant ist, so dass der quadratische Ansatz im Vergleich zur kubischen Approximation definitionsgemäß leichter angepasst werden kann. Zu überprüfen bleibt, inwieweit die zugrundegelegte van Genuchten-Mualem-Parametrisierung für extreme Parameterwerte (n sehr klein oder sehr groß) physikalisch motivierte Ergebnisse liefert (insbesondere die Nichtdifferenzierbarkeit an der Stelle $\psi = 0$ für $n < 2$). In Kapitel 4.1 werden schließlich auf sowohl virtuellen als auch realen Messdaten basierende Fallstudien durchgeführt und ein erneuter Vergleich angestrebt.

R_Φ	$n=5.0$	$n=4.0$	$n=3.0$	$n=2.0$	$n=1.5$	$n=1.25$
$R_\psi=-0.25$	$7.81\cdot10^{-4}$	$2.92\cdot10^{-3}$	$1.03\cdot10^{-2}$	$2.99\cdot10^{-2}$	$3.85\cdot10^{-2}$	$3.20\cdot10^{-2}$
$R_\psi=-0.10$	$8.00\cdot10^{-6}$	$7.50\cdot10^{-5}$	$6.66\cdot10^{-4}$	$4.96\cdot10^{-3}$	$1.03\cdot10^{-2}$	$1.09\cdot10^{-2}$
$R_\psi=-0.05$	$2.50\cdot10^{-7}$	$4.69\cdot10^{-6}$	$8.33\cdot10^{-5}$	$1.25\cdot10^{-3}$	$3.70\cdot10^{-3}$	$4.66\cdot10^{-3}$
$R_\psi=-0.02$	$2.56\cdot10^{-9}$	$1.20\cdot10^{-7}$	$5.33\cdot10^{-6}$	$2.00\cdot10^{-4}$	$9.41\cdot10^{-4}$	$1.50\cdot10^{-3}$
$R_\psi=-0.01$	$8.00\cdot10^{-11}$	$7.50\cdot10^{-9}$	$6.67\cdot10^{-7}$	$5.00\cdot10^{-5}$	$3.33\cdot10^{-4}$	$6.31\cdot10^{-4}$
$R_\psi=-0.001$	$8.88\cdot10^{-16}$	$7.50\cdot10^{-13}$	$6.67\cdot10^{-10}$	$5.00\cdot10^{-7}$	$1.05\cdot10^{-5}$	$3.56\cdot10^{-5}$

Tabelle 2.2: R_Φ unter Vorgabe ausgewählter R_ψ und n

$r_{R\Phi}$	$n=5.0$	$n=4.0$	$n=3.0$	$n=2.0$	$n=1.5$	$n=1.25$
$R_\psi=-0.25$	$4.07\cdot10^{-5}$	$2.50\cdot10^{-4}$	$1.73\cdot10^{-3}$	$1.42\cdot10^{-2}$	$4.34\cdot10^{-2}$	$7.29\cdot10^{-2}$
$R_\psi=-0.10$	$4.34\cdot10^{-7}$	$6.64\cdot10^{-6}$	$1.14\cdot10^{-4}$	$2.39\cdot10^{-3}$	$1.20\cdot10^{-2}$	$2.57\cdot10^{-2}$
$R_\psi=-0.05$	$1.37\cdot10^{-8}$	$4.20\cdot10^{-7}$	$1.44\cdot10^{-5}$	$6.06\cdot10^{-4}$	$4.36\cdot10^{-3}$	$1.13\cdot10^{-2}$
$R_\psi=-0.02$	$1.42\cdot10^{-10}$	$1.08\cdot10^{-8}$	$9.29\cdot10^{-7}$	$9.76\cdot10^{-5}$	$1.12\cdot10^{-3}$	$3.74\cdot10^{-3}$
$R_\psi=-0.01$	$1.44\cdot10^{-12}$	$6.79\cdot10^{-10}$	$1.16\cdot10^{-7}$	$2.44\cdot10^{-5}$	$4.00\cdot10^{-4}$	$1.60\cdot10^{-3}$
$R_\psi=-0.001$	$1.65\cdot10^{-18}$	$6.88\cdot10^{-14}$	$1.16\cdot10^{-10}$	$2.36\cdot10^{-7}$	$1.19\cdot10^{-5}$	$8.58\cdot10^{-5}$

Tabelle 2.3: Fehler der ($\mathcal{P}^2$-)Regularisierung

2.1.6 Anfangswerte und Randbedingungen

Damit die Richards-Gleichung (2.2) gelöst werden kann müssen geeignete Anfangswerte $\psi_0(x) = \psi(x, 0)$ für $x \in \Omega$ und Randbedingungen auf dem Rand $\partial\Omega\times(0, T)$ vorgegeben werden. Im Fall eines klassischen Säulenexperimentes mit manuell beeinflussbarem Druck werden Letztere typischerweise durch

r_{R_ψ}	$n=5.0$	$n=4.0$	$n=3.0$	$n=2.0$	$n=1.5$	$n=1.25$
$R_\psi=-0.25$	$8.28\cdot10^{-5}$	$4.23\cdot10^{-6}$	$7.89\cdot10^{-5}$	$1.23\cdot10^{-2}$	$5.52\cdot10^{-2}$	$1.03\cdot10^{-1}$
$R_\psi=-0.10$	$8.39\cdot10^{-7}$	$1.27\cdot10^{-7}$	$5.99\cdot10^{-7}$	$2.01\cdot10^{-3}$	$1.49\cdot10^{-2}$	$3.56\cdot10^{-2}$
$R_\psi=-0.05$	$2.59\cdot10^{-8}$	$4.58\cdot10^{-9}$	$1.58\cdot10^{-8}$	$5.06\cdot10^{-4}$	$5.41\cdot10^{-3}$	$1.56\cdot10^{-2}$
$R_\psi=-0.02$	$2.63\cdot10^{-10}$	$4.86\cdot10^{-11}$	$1.43\cdot10^{-10}$	$8.14\cdot10^{-5}$	$1.40\cdot10^{-3}$	$5.16\cdot10^{-3}$
$R_\psi=-0.01$	$8.19\cdot10^{-12}$	$1.53\cdot10^{-12}$	$4.28\cdot10^{-12}$	$2.04\cdot10^{-5}$	$4.98\cdot10^{-4}$	$2.22\cdot10^{-3}$
$R_\psi=-0.001$	$3.52\cdot10^{-17}$	$4.93\cdot10^{-18}$	$3.13\cdot10^{-16}$	$1.95\cdot10^{-7}$	$1.50\cdot10^{-5}$	$1.22\cdot10^{-4}$

Tabelle 2.4: Fehler der ($\mathcal{P}^3$-)Regularisierung

- **Dirichlet-Randbedingung:**

$$\psi(\vec{x},t) = g_D(\vec{x},t) \quad \text{auf} \quad \Gamma_D \times (0,T) \ \text{ und}$$

- **Homogene Fluss-Randbedingung:**

$$\boldsymbol{q}(\vec{x},t)\cdot\nu(x) = 0 \quad \text{auf} \quad \Gamma_F \times (0,T)$$

beschrieben. Dabei gibt $\Gamma_D \subset \partial\Omega$ den Bereich des Ausflusses und $\Gamma_F = \partial\Omega \setminus \Gamma_D$ den restlichen, isolierten Rand mit äußerer Einheitsnormalen ν an.

2.2 Reaktive Transportprozesse in porösen Medien

Im Folgenden wird die Modellierung des Transports biochemischer Substanzen in porösen Medien sowie ausgewählte Sorptions- und Abbauprozesse vorgestellt (vgl. z.B. Knabner [38] und Prechtel [46]).

2.2.1 Transportgleichung

Die grundlegende Gleichung zur Beschreibung des Transports einer mobilen Substanz c [M/L^3] ist durch die Massenerhaltung

$$\partial_t\big(\Theta(\vec{x},t)c(\vec{x},t)\big) + \nabla\cdot\boldsymbol{J}(\vec{x},t) = R(\vec{x},t) \tag{2.30}$$

mit

$\Theta(\vec{x},t)$	volumetrischer Wassergehalt [−],
$\boldsymbol{J}(\vec{x},t)$	Flussdichte [M/L$^{\text{d-1}}$T], $d\in\{1,2,3\}$,
$R(\vec{x},t)$	Rate der Quellen und Senken sowie reaktiver Prozesse [M/L$^{\text{d}}$T],
$x\in\Omega$	Ortsvariable [L$^{\text{d}}$], Ω Gebiet in R^d und
$t\in(0,T)$	Zeitvariable [T] mit Endzeitpunkt T

gegeben. Die Flussdichte $\boldsymbol{J}$ wird dabei durch das Ficksche Gesetz

$$\boldsymbol{J} = -\Theta(\vec{x},t)\boldsymbol{D}(\vec{x},t)\nabla c + \boldsymbol{q}(\vec{x},t)c(\vec{x},t)\,,$$

einem Diffusions/Dispersions-Tensor $\boldsymbol{D}$ [L^2/T] und dem spezifischen Fluss $\boldsymbol{q}$ [L/T] festgelegt. Auf die funktionale Beschreibung von $\boldsymbol{D}$ wird in dieser Arbeit nicht weiter eingegangen. Der Vollständigkeit halber ist jedoch angegeben, dass bei den in Kapitel 4 vorgestellten Fallstudien der Tensor durch

$$\boldsymbol{D} = \alpha_l(\vec{x},t)\boldsymbol{q}(\vec{x},t) + d\,\Theta(\vec{x},t)$$

mit gebietsbedingter Dispersionslänge α_l $[L]$ und stoffabhängiger Diffussion d [L^2/T] angegeben wurde. Der Wassergehalt Θ und der Fluss $\boldsymbol{q}$ sind im stationären Fall (zeitlich) konstant, im Allgemeinen jedoch nichtlinear und (meist) durch die in Abschnitt 2.1.3 vorgestellte Richards-Gleichung (2.2) bestimmt (vgl. z.B. Bear [5] und Evans [23]).

Bleibt die rechte Seite von Gleichung (2.30) anzugeben. Neben den im Gebiet Ω vorkommenden Quellen und Senken der betrachteten Substanz c wird R vorrangig durch reaktive Prozesse bestimmt. Im Folgenden werden hierzu ausgewählte Sorptions- (Anreicherung/Anlagerung eines Stoffes) und Abbauprozesse vorgestellt. Im einfachsten Fall (dies ist als Ausnahmesituation zu verstehen) ist R konstant. I.d.R. muss R jedoch nichtlinear in Ort und Zeit sowie abhängig von der Sorptionsrate und der Stoffkonzentration angenommen werden. Im Fall eines Multikomponentensystems ist R auch von anderen, an biochemischen Reaktionsprozessen mit c beteiligten Stoffen, abhängig. Im Sinne einer einfachen Notation wird daher in den folgenden Abschnitten nicht explizit die funktionale Abhängigkeit von R durch $R(.)$ angegeben.

2.2.2 Sorption

Prinzipiell wird der Sorptionsprozess in zwei Kategorien unterschieden. Bei der Absorption wird die untersuchte Substanz in einer Phase (hier das Liquid im Porenraum) und bei der Adsorption auf der Grenzfläche zweier Phasen (hier zwischen Liquid und Feststoffskelett) angereichert bzw. angelagert. Die Ablösung

eines bereits sorbierten Stoffes, dem sogenannten Absorpt bzw. Adsorpt, wird allgemein als Desorption bezeichnet. Als Ursache für diese Sorptionsvorgänge sind unterschiedliche chemische, physikalische und elektrostatische Wechselwirkungen zwischen dem Sorbent (Sorptionsmittel), dem Sorptiv (noch nicht sorbierter Stoff) und dem Sorbat (System aus Sorbent, sorbierten und noch nicht sorbierten Stoffen) anzusehen.

Nachfolgend werden kurz die Modellierungsansätze der Gleichgewichtsreaktion und der Reaktionskinetik (Ungleichgewichtsreaktion) vorgestellt.

2.2.2.1 Gleichgewichtssorption

Bei einer Gleichgewichtsreaktion wird angenommen, dass eine Änderung der Konzentration des gelösten Stoffes c *sofort*, also ohne Reaktionsverzögerung, eine Änderung der sorbierten Menge s bewirkt. Entsprechend stehen beide Stoffmengen über eine Isotherme

$$\phi\big(c(\vec{x},t)\big) = s(\vec{x},t)$$

im Gleichgewicht. Zur formalen Beschreibung von ϕ haben sich vorrangig folgende Parametrisierungen

Linear:	$\phi\big(c(\vec{x},t)\big) := K_d\, c(\vec{x},t)\,,$
Freundlich:	$\phi\big(c(\vec{x},t)\big) := K_d\, c^{\lambda}(\vec{x},t)\,,$
Langmuir:	$\phi\big(c(\vec{x},t)\big) := \dfrac{K_d\, c(\vec{x},t)}{1+\frac{K_d}{s_{\max}} c(\vec{x},t)}$ und
Freundlich-Langmuir:	$\phi\big(c(\vec{x},t)\big) := \dfrac{K_d\, c^{\lambda}(\vec{x},t)}{1+\frac{K_d}{s_{\max}}\, c^{\lambda}(\vec{x},t)}\,,$

mit $K_d > 0$ [L^3/M] bzw. $[\mathrm{L}^3/\mathrm{M}]^{\lambda}$, $\lambda > 0$ [$-$] und $s_{\max} > 0$ [M/M] durchgesetzt (vgl. z.B. Knabner [38] oder Fetter [24]). Gerade im Bezug auf die im Kapitel 3.1 vorgestellte Parameteridentifizierung ist neben diesen auch die im Kapitel 3.2 vorgestellte Spline-Approximation eine interessante Alternative (vgl. Igler, Knabner [35]).

Unter Verwendung der Lagerungsdichte $\varrho_b > 0$ [$\mathrm{M}/\mathrm{L}^{\mathrm{d}}$] und dem Massenanteil der Gleichgewichtssorptionsplätze $0 \le \varrho_\phi \le 1$ [$-$] kann die zugehörige Reaktionrate schließlich durch

$$R = -\varrho_b \partial_t \Big(\varrho_\phi \phi\big(c(\vec{x},t)\big)\Big) \tag{2.31}$$

angegeben werden.

2.2.2.2 Kinetische Sorption

Bei der kinetischen Sorption wird die Konzentration s, aufgrund einer berücksichtigten Reaktionsverzögerung, zeitlich entwickelt. Folglich steht die Stoffmenge c nicht mehr mit der von s im Gleichgewicht. Stattdessen ist die Differenz zwischen der Isotherme der kinetischen Reaktion φ und der sorbierten Konzentration s die treibende Kraft. Entsprechend gilt

$$R = -\varrho_b \partial_t \Big(\varrho_\varphi s\big(c(\vec{x},t)\big)\Big) \quad \text{und} \quad \partial_t s = r\Big(\varphi\big(c(\vec{x},t)\big) - s\Big) \tag{2.32}$$

mit Massenanteil der kinetischen Sorptionsplätze $0 \leq \varrho_\varphi \leq 1$ [−], Ratenparameter $r > 0$ [1/T] und der Lagerungsdichte ϱ_b. Für φ kann analog zu Gleichgewichtsisotherme ϕ eine lineare oder nichtlineare Parametrisierung verwendet werden.

2.2.3 Biologischer Abbau

Die Ausbreitung vieler organischer Schadstoffe, wie beispielsweise einige Radionuklide und PAKs (polyzyklische aromatische Kohlenwasserstoffe), wird vorwiegend durch mikrobielle Aktivitäten beeinflusst. Der biochemische Abbau dieser Substanzen wird durch im Boden vorkommende Mikroorganismen ermöglicht. Die beiden einfachsten Ansätze zur Beschreibung dieser Prozesse sind durch die Modellierungen 0ter und 1ter Ordnung gegeben. Hierbei wird je Zeiteinheit einfach eine konstante bzw. eine linear von der Schadstoffkonzentration c abhängige Menge abgebaut. Damit gilt

$$R = -\Theta(\vec{x},t)k_0,\ k_0 > 0\ [\mathrm{M/L^dT}], \quad \text{bzw.} \quad R = -\Theta(\vec{x},t)k_1 c(\vec{x},t),\ k_1 > 0\ [1/\mathrm{T}]. \tag{2.33}$$

(Für $k_0 < 0$ bzw. $k_1 < 0$ werden entsprechende Konzentrationsquellen in Ω beschrieben.) Zu beachten ist jedoch, dass bei einem Abbau 0ter Ordnung (unphysikalische) negative Konzentrationen auftreten können. Auch der lineare Ansatz ist nur eingeschränkt nutzbar, da naturgemäß eine maximale Abbaurate $\mu_{\max} > 0$ [1/T] den Prozess nach oben hin begrenzt.

Im folgenden Abschnitt wird entsprechend dieser Problematik die Modellierung auf eine begrenzte Abbaurate erweitert. Da diese jedoch nicht nur vom Schadstoff selbst, sondern auch von den anderen beteiligten Reaktionspartnern abhängt,

wird zudem eine mehrkomponentige Parametrisierung motiviert. Insbesondere wird das duale Monod-Modell vorgestellt (vgl. z.B. Prechtel [46] oder Alexander, Scow [1]).

2.2.3.1 Monod-Modell

Soll eine begrenzte Abbaurate modelliert werden, kann die, aus der Enzymkinetik bekannte, Michaelis-Menten-Gleichung

$$R = \Theta(\vec{x},t)\mu(\vec{x},t) \quad \text{mit} \quad \mu\big(c(\vec{x},t)\big) = \mu_{\max}\left(\frac{c(\vec{x},t)}{K_M + c(\vec{x},t)}\right), \tag{2.34}$$

$K_M > 0$ [M/L$^{\mathrm{d}}$] konstant, verwendet werden (vgl. z.B. Bisswanger [7]). Hierbei beschreibt K_M gerade die Konzentration, die notwendig ist, um eine Abbaurate $\mu = \frac{\mu_{\max}}{2}$ zu erhalten. Da in den meisten Fällen Redoxreaktionen (Elektronen werden von einem Reaktanten zu einem anderen übertragen) mit katalysierender Biomasse für den Abbau verantwortlich sind, kann (2.34) auch um c_B zu

$$\mu\big(c(\vec{x},t) = \mu_{\max}\left(\frac{c(\vec{x},t)}{K_M + c(\vec{x},t)}\right) c_B(\vec{x},t)$$

ergänzt werden.

Abbildung 2.7 zeigt den schematischen Verlauf der in (2.34) vorgestellten Abbaurate μ. Gut zu erkennen ist dabei der näherungsweise lineare Anstieg bei $c = 0$ sowie das assymptotische Verhalten $\mu \to \mu_{\max}$ für $c \to \infty$.

Treten toxische Effekte bei höheren Stoffkonzentrationen auf, kann des Weiteren (2.34) um einen Inhibitionsterm erweitert werden. In diesem Fall gilt schließlich

$$\mu\big(c(\vec{x},t)\big) = \mu_{\max}\left(\frac{c(\vec{x},t)}{K_M + c(\vec{x},t) + \frac{c^2(\vec{x},t)}{K_I}}\right) c_B(\vec{x},t)$$

mit Inhibitionskonstante $K_I > 0$ [M/L$^{\mathrm{d}}$].

Ist neben der Biomasse und einem elektronenabgebenden Schadstoff eine weitere, entsprechend elektronenaufnehmende Spezies (typischerweise Sauerstoff) an der biochemischen Reaktion beteiligt, so kann das sogenannte **duale Monod-Modell** zur Parametrisierung verwendet werden. Dem Elektronenaustausch folgeleistend wird der Schadstoff dann auch als Elektronendonator c_D und der elektronenaufnehmende Stoff als Elektronenakzeptor c_A bezeichnet. Schließlich gilt

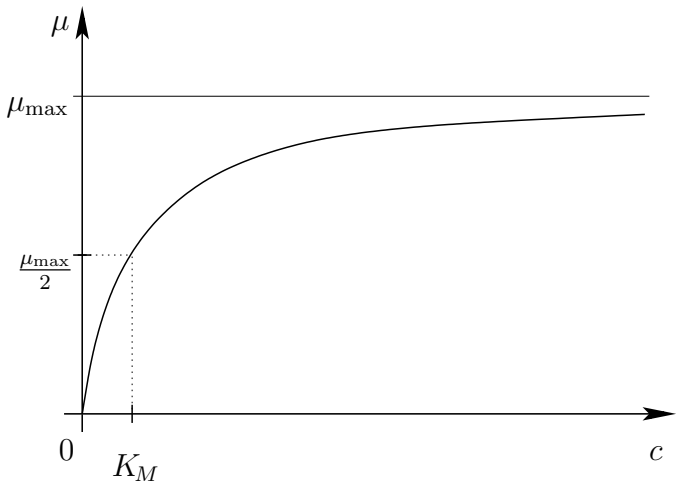

Abbildung 2.7: begrenzte Monod-Abbaurate

$$\mu(c_D, c_A, c_B) = \mu_{\max}\left(\frac{c_D}{K_{M_D}+c_D}\right)\left(\frac{c_A}{K_{M_A}+c_A}\right)c_B \tag{2.35}$$

bzw.

$$\mu(c_D, c_A, c_B) = \mu_{\max}\left(\frac{c_D}{K_{M_D}+c_D+\frac{c_D^2}{K_{ID}}}\right)\left(\frac{c_A}{K_{M_A}+c_A+\frac{c_A^2}{K_{IA}}}\right)c_B\,. \tag{2.36}$$

Auf die orts- und zeitabhägige Darstellung der Konzentrationen c_D, c_A, c_B wird im Sinne einer besseren Übersichtlichkeit an dieser Stelle verzichtet. Die Konstanten $K_{M_D}, K_{M_A}, K_{ID}, K_{ID}$ und $\mu_{\max}$ sind analog zu den vorangegangener Überlegungen motiviert. Schließlich können unter Verwendung der konstant angenommenen Feldfaktoren $\alpha_{A/D} > 0$ [M/M] und Y [M/M] sowie einem optionalen Strafterm für hohe Biomassekonzentrationen $1-\frac{c_B}{c_{B\max}}$ die Reaktionsraten für die beiden mobilen Substanzen durch

$$R_D = \Theta(\vec{x}, t)\mu(c_D, c_A, c_B) \quad \text{und} \quad R_A = \Theta(\vec{x}, t)\alpha_{A/D}\mu(c_D, c_A, c_B) \tag{2.37}$$

und für die immobile Biomasse durch

$$R_B = Y\left(1-\frac{c_B(\vec{x}, t)}{c_{B_{\max}}}\right)\mu(c_D, c_A, c_B) \tag{2.38}$$

angegeben werden.

Bemerkung 2.10
*Die Parameter K_{M_D}, K_{I_D}, K_{M_A}, K_{I_A}, μ_{max}, $c_{B_{max}}$, Y und $\alpha_{A/D}$ werden im Folgenden auch als **Monod**-Parameter bezeichnet.*

Es sei an dieser Stelle abschließend bemerkt, dass auch Modellierungsansätze für Systeme mit deutlich höherer Reaktantenzahl in der Literatur zu finden sind (vgl. z.B. Prechtel [46]). Da der Schwerpunkt dieser Arbeit in der zugehörigen Parameteridentifizierung gesetzt wurde (vgl. nachfolgende Kapitel) und bereits das vorliegende dreikomponentige System erfahrungsgemäß schwer identifizierbar ist, wird auf die Darstellung höherdimensionaler Modelle verzichtet.

2.2.3.2 Monotonie biologischer Abbauprozesse

Für biochemische Abbauprozesse mit (ausreichend) geringen Stoffkonzentrationen kann i.A. davon ausgegangen werden, dass sie monoton anwachsend sind. Dementsprechend wird durch eine Erhöhung der beteiligten Stoffmengen auch ein höherer (Schadstoff-)Abbau bewirkt, und umgekehrt. Sind allerdings auch höhere Konzentrationen zu erwarten, so müssen ggf. toxische und inhibierende Effekte berücksichtigt werden. Folglich wird die Zunahme für hohe Stoffmengen geringer ausfallen und möglicherweise sogar wieder abnehmen.

Die von Natur aus gegebenen Monotonieeigenschaften findet sich auch in den hier vorgestellten Modellgleichungen wieder. Während bei einem biologischen Abbau 0ter und 1ter Ordnung die Ableitung der Reaktionsgleichung nach der Konzentration trivialerweise kontant ist, bedarf es bei den Monod-Gleichungen einer kleinen Überprüfung.

Definition 2.11 (Monotonie im $\mathbb{R}^d$)
Ein Funktion $f: D \to \mathbb{R}$, $D \subset \mathbb{R}^d$, $d \in \mathbb{N}$, wird als ***streng monoton wachsend in $\vec{x}$****, $\vec{x} := (x_1, \ldots, x_d)^T \in D$, bezeichnet, falls für alle $\vec{y} := (y_1, \ldots, y_d)^T \in D$ mit $y_i \geq x_i$, $i = 1, \ldots, d$, und $\vec{y} \neq \vec{x}$ stets die Ungleichung $f(\vec{y}) > f(\vec{x})$ erfüllt ist. f wird als* ***streng monoton wachsend auf D*** *oder einfach als* ***streng monoton wachsend*** *bezeichnet, falls f in jedem Punkt $\vec{x} \in D$ monoton wachsend ist.*

Unter Verwendung des auf $\mathbb{R}^d$ definierten Monotoniebegriffes wird nachfolgend gezeigt, dass die ungehemmte Reaktionsrate (2.35) auf ganz $\left[\mathbb{R}_0^+\right]^3$ und die inhibierte Rate (2.36) auf dem Quader

$$Q := \left[0, \sqrt{K_{M_D} K_{I_D}}\right] \times \left[0, \sqrt{K_{M_A} K_{I_A}}\right] \times \mathbb{R}_0^+$$

streng monoton wachsend ist. Hierzu wird die zugrundegelegte Michaelis-Menten Gleichung

$$\mu(c) = \mu_{\max} \frac{c}{K_M + c} \quad \text{bzw.} \quad \mu(c) = \mu_{\max} \frac{c}{K_M + c + \frac{c^2}{K_I}}$$

nach c differenziert. Es gilt

$$\mu'(c) := \frac{d\mu}{dc}(c) = \mu_{\max}\frac{K_M}{(K_M+c)^2} > 0 \quad \forall\, c > 0 \tag{2.39}$$

bzw.

$$\mu'(c) = \mu_{\max}\frac{K_M - \frac{c^2}{K_I}}{\left(K_M+c+\frac{c^2}{K_I}\right)^2} > 0 \quad \forall\ 0<c<\sqrt{K_M K_I}\,.$$

Folgelich kann, ohne die Monotonie zu verletzen, die Konzentration c in (2.39) umso größer sein, je geringer die Inhibition (je größer der Wert von K_I) ausfällt. Wird allerdings der Grenzewert $c=\sqrt{K_M K_I}$ überschritten, so nimmt μ ab.

Da für das duale Monod-Modell entsprechend

$$\frac{\partial\mu}{\partial c_i}(c_D, c_A, c_B) = \mu_{\max}\frac{K_{M_i}}{(K_{M_i}+c_i)^2}\frac{c_j}{K_{M_j}+c_j}c_B\,, \quad i,j\in\{D,A\},\ i\neq j,$$

und

$$\frac{\partial\mu}{\partial c_B}(c_D, c_A, c_B) = \mu_{\max}\frac{c_D}{K_{M_D}+c_D}\frac{c_A}{K_{M_A}+c_A}$$

bzw.

$$\frac{\partial\mu}{\partial c_i}(c_D, c_A, c_B) = \mu_{\max}\frac{K_{M_i}-\frac{c_i^2}{K_{I_i}}}{\left(K_{M_i}+c_i+\frac{c_i^2}{K_{I_i}}\right)^2}\frac{c_j}{K_{M_j}+c_j+\frac{c_j^2}{K_{I_j}}}c_B\,, \quad i,j\in\{D,A\},\ i\neq j,$$

und

$$\frac{\partial\mu}{\partial c_B}(c_D, c_A, c_B) = \mu_{\max}\frac{c_D}{K_{M_D}+c_D+\frac{c_D^2}{K_{ID}}}\frac{c_A}{K_{M_A}+c_A+\frac{c_A^2}{K_{IA}}}\,.$$

gilt, kann diese Aussage schließlich auch auf das vorgestellte Dreikomkonentenmodell erweitert werden.

2.3 Vollständiges Differentialgleichungssystem

Die Transportgleichung (2.30) mit Gleichgewichts- und kinetischer Sorption (2.31) und (2.32), Abbau 0ter und 1ter Ordnung (2.33) sowie biologischem Abbau nach dem dualen Monod-Modell (2.36) liefert zusammenfassend folgendes nichtlineares

Modellsystem:

$$\begin{aligned}
\partial_t(\Theta c_D)-\nabla\cdot(\boldsymbol{D}_D\nabla c_D-\boldsymbol{q}c_D)+\varrho_b\partial_t\big(\varrho_\psi\psi_D(c_D)+\varrho_\varphi s_D(c_D)\big)\\
+\mu+\Theta k_{1_D}c_D+\Theta k_{0_D}=0\,,\\
\partial_t(\Theta c_A)-\nabla\cdot(\boldsymbol{D}_A\nabla c_A-\boldsymbol{q}c_A)+\varrho_b\partial_t\big(\varrho_\psi\psi_A(c_A)+\varrho_\varphi s_A(c_A)\big)\\
+\alpha_{A/D}\mu+\Theta k_{1_A}c_A+\Theta k_{0_A}=0\,,\\
\partial_t c_B-\frac{Y}{\Theta}\Big(1-\frac{c_B}{c_{B_{\max}}}\Big)\mu+k_{1_B}c_B+k_{0_B}=0\,,\quad\text{mit}
\end{aligned}$$

$$\mu=\Theta\mu_{\max}\left(\frac{c_D}{K_{M_D}+c_A+\frac{c_A^2}{K_{IA}}}\right)\left(\frac{c_A}{K_{M_A}+c_A+\frac{c_A^2}{K_{IA}}}\right)c_B\quad\text{und}$$

$$\begin{aligned}
\partial_t s_D&=r_D\big(\varphi_D(c_D)-s_D\big)\,,\\
\partial_t s_A&=r_A\big(\varphi_A(c_A)-s_A\big)\,.
\end{aligned}\tag{2.40}$$

Um dieses gekoppelte DGL-System lösen zu können sind geeignete Anfangswerte

$$u_0(\vec{x}):=u(\vec{x},0)\,,\ \ \vec{x}\in\Omega\,,$$

sowie konforme Randbedingungen

$$u(\vec{x},t)=g(\vec{x},t)\,,\ \ (\vec{x},t)\in\partial\Omega\times(0,T)\,,$$

$u=c_i(\vec{x},t)$, $i=\{D,A,B\}$ bzw. $u=s_j(\vec{x},t)$, $j=\{D,A\}$, festzulegen. Für Letzteres steht im Allgemeinen die Auswahl

- **Dirichlet-Randbedingung:**

$$u=g_1\quad\text{auf}\quad\Gamma_1\times(0,T)\,,$$

- **Neumann-Randbedingung:**

$$\boldsymbol{D}\nabla u\cdot\nu=g_2\quad\text{auf}\quad\Gamma_2\times(0,T)\,,$$

- **Fluss-Randbedingung:**

$$(\boldsymbol{D}\nabla u-\boldsymbol{q}u)\cdot\nu=g_3\quad\text{auf}\quad\Gamma_3\times(0,T)\,,$$

- **Gemischte Randbedingung:**

$$(\boldsymbol{D}\nabla u-\boldsymbol{q}u)\cdot\nu+\alpha_l c=g_4\quad\text{auf}\quad\Gamma_4\times(0,T)\,.$$

für eine disjunkte Zerlegung des Randes $\partial\Omega=\Gamma_1\cup\Gamma_2\cup\Gamma_3\cup\Gamma_4$ zur Verfügung.

Kapitel 3

Lösung inverser Probleme

Die in Kapitel 2 vorgestellten mathematischen Modelle zur Beschreibung bodenphysikalischer Reaktionsdynamiken enthalten eine Vielzahl unabhängiger Parameter. Diese können sowohl als konstanter Wert in vorgegebenen fixen Parametrisierungen wie beispielsweise die van Genuchten-Mualem-Leitfähigkeit (2.5) oder die duale Monod-Abbaurate (2.36) als auch als Nichtlinearität selbst (als Alternative zu einer fixen Parametrisierung) auftreten.

In den nachfolgenden Abschnitten wird, motiviert durch Igler [34], zunächst das Identifizierungsproblem allgemein formuliert. Anschließend wird neben der aus Bitterlich [8] bekannten formfreien 1D-Parametrisierung eine Erweiterung auf n-dimensionale formfreie Ansätze vorgestellt. Im Bezug auf die deutlich höhere Komplexität werden entsprechend Bungartz [15] schließlich dünne Gitter verwendet und diese um inhomogene Randwerte ergänzt.

3.1 Parameteridentifizierung

Der bei Vorgabe aller Parameter durchgeführte Lösungsprozess eines mathematischen Modells, die sogenannte Simulation des Experimentes, wird als das **direkte Problem** bezeichnet. Es wird durch eine nichtlineare Abbildung $\mathcal{A}$, den direkten Lösungsoperator, vom Parameterraum P in den Lösungsraum U,

$$\mathcal{A} : P \to U \,,$$

beschrieben. Ist eine direkte Messung einzelner Größen nicht möglich, so müssen diese durch indirekte Beobachtungen ausgewählter Experimente ermittelt werden. Dies geschieht durch Lösen des sogenannten **inversen Problems**, welches

dem Finden der Umkehrfunktion

$$\mathcal{A}^{-1} : U \to P$$

von $\mathcal{A}$ entspricht. Um der Begriffsbildung von Hadamard [30] zu folgen, wird eine mathematische Aufgabe als **korrekt gestellt** bezeichnet, falls die folgenden Bedingungen erfüllt sind:

1. es existiert eine Lösung,
2. die Lösung ist eindeutig bestimmt und
3. die Lösung hängt stetig von den Parametern ab.

Ist eine dieser Bedingungen verletzt, so ist das Problem entsprechend **schlecht gestellt**.

Im Folgenden seien P und U Banachräume. Wird nun angenommen, dass der direkte Lösungsoperator $\mathcal{A}$ auf ganz P definiert und stetig in den Metriken der Räume P und U ist, so ist das direkte Problem definitionsgemäß korrekt gestellt. Damit dies auch für das indirekte Problem gilt, muss noch die Bijektivität von $\mathcal{A}$ und die Stetigkeit von $\mathcal{A}^{-1}$ in den Metriken U und P gefordert werden. Dies ist jedoch i.A. nicht der Fall (vgl. z.B. Louis [42]). Letztendlich ist für inverse Probleme (fast immer) kennzeichnend, dass sie im obigen Sinne schlecht gestellt sind.

3.1.1 Problemformulierung

Zur Bestimmung gesuchter Parameter können ausgewählte Experimente durchgeführt werden. Aus zeitlichen und technischen Gründen wird sich die Anzahl dieser Versuche auf $\kappa \in \mathbb{N}$ beschränken, so dass alle zur Verfügung stehenden Informationen aus den Experimenten E_k, $k = 1, \ldots, \kappa$, gewonnen werden.

Im Allgemeinen wird bei der praktischen Durchführung von Experimenten nicht die komplette Lösung des direkten Problems sondern lediglich ein Teil davon messbar sein. Bei parabolischen Problemen, wie die in Kapitel 2 betrachteten Reaktions- und Transportprobleme, werden dies i.d.R. Messungen ausgewählter charakteristischer Modellgrößen auf einem Zeitintervall $[0, T]$ sein.

Im Folgenden wird davon ausgegangen, dass bei der Simulation der Experimente E_k, $k = 1, \ldots, \kappa$, die Beobachtungen $w_{k,i}$, $i = 1, ..., n_k$, mit $w_{k,i} \in W_{k,i}$ Banachraum,

durchgeführt werden. Da diese sowohl von der Lösung $u \in U$ selbst als auch von dem zu identifizierenden Parametersatz $p \in P$ abhängen können, lassen sich für $k = 1, \ldots, \kappa$ die folgenden Beobachtungsoperatoren

$$\begin{aligned} \mathcal{B}_k := \begin{pmatrix} \mathcal{B}_{k,1} \\ \vdots \\ \mathcal{B}_{k,n_k} \end{pmatrix} : U \times P \to \begin{pmatrix} W_{k,1} \\ \vdots \\ W_{k,n_k} \end{pmatrix} := W_k \,, \\ (u,p) \mapsto \begin{pmatrix} \mathcal{B}_{k,1}(u,p) \\ \vdots \\ \mathcal{B}_{k,n_k}(u,p) \end{pmatrix} = \mathcal{B}_k(u,p) =: w_k = \begin{pmatrix} w_{k,1} \\ \vdots \\ w_{k,n_k} \end{pmatrix} , \end{aligned} \tag{3.1}$$

definieren. Zusammenfassend wird ein beschränkter Beobachtungsoperator

$$\begin{aligned} \mathcal{B} := \begin{pmatrix} \mathcal{B}_1 \\ \vdots \\ \mathcal{B}_\kappa \end{pmatrix} : U \times P \to \begin{pmatrix} W_1 \\ \vdots \\ W_\kappa \end{pmatrix} := W, \\ (u,p) \mapsto \begin{pmatrix} \mathcal{B}_1(u,p) \\ \vdots \\ \mathcal{B}_\kappa(u,p) \end{pmatrix} = \mathcal{B}(u,p) =: w = \begin{pmatrix} w_1 \\ \vdots \\ w_\kappa \end{pmatrix} \end{aligned} \tag{3.2}$$

eingeführt, der, wie später noch gezeigt wird, unter den Voraussetzungen des Satzes über implizite Funktionen (Satz 3.6) mit Hilfe des direkten Lösungsoperator $\mathcal{A}$ auf einer Teilmenge des Parameterraums $\tilde{P} \subset P$ auch direkt als Abbildung in den Beobachtungsraum W

$$\begin{aligned} \mathcal{B} : \tilde{P} &\to W, \\ p &\mapsto w = \mathcal{B}\big(\mathcal{A}(p), p\big) \,, \end{aligned}$$

aufgefasst werden kann. Das Identifizierungsproblem beschreibt somit die Aufgabe zu einer vorgegebenen Beobachtung $w_0 \in W$ den Parametersatz $p_0 \in P$ zu finden, so dass

$$\mathcal{B}\big(\mathcal{A}(p_0), p_0\big) = w_0 \tag{3.3}$$

gilt. Es wird folgende Definition eingeführt.

Definition 3.1
*Ein Parameter $p \in P$ heißt **identifizierbar** aus der Beobachtung $w \in W$, falls der zugehörige Beobachtungsoperator $\mathcal{B}$, betrachtet als Abbildung vom Parameterraum P in den Beobachtungsraum W, injektiv ist.*

Somit hat sowohl der Raum P, die Wahl des Operators $\mathcal{B}$ (Art und Anzahl der Beobachtungen) als auch die Definition des Lösungsoperators $\mathcal{A}$ direkten Einfluss auf die Identifizierbarkeit eines Parameters.

Aufgrund von Modell- und Messfehlern wird in (3.3) für gewöhnlich statt der eigentlichen Beobachung $w_0 \in W$ lediglich eine gestörte experimentelle Messung

$$w_\varepsilon := \begin{pmatrix} (w_\varepsilon)_1 \\ \vdots \\ (w_\varepsilon)_\kappa \end{pmatrix} \in W, \quad (w_\varepsilon)_k := \begin{pmatrix} (w_\varepsilon)_{k,1} \\ \vdots \\ (w_\varepsilon)_{k,n_k} \end{pmatrix} \in W_i \,,$$

$(w_\varepsilon)_{k,i} \in W_{k,i}$, $k = 1, \ldots, \kappa$, $i = 1, \ldots, n_k$, mit

$$\|w_0 - w_\varepsilon\|_W \leq \varepsilon$$

gegeben sein. Da i.A. w_ε nicht im Bildraum des Beobachtungsoperators $\mathcal{B}$ enthalten ist ($\nexists\, p \in P$ mit $\mathcal{B}(\mathcal{A}(p), p) = w_\varepsilon$), wird ein Fehlerfunktion

$$\begin{aligned} \tilde{\mathcal{J}} : W &\to \mathbb{R} \,, \\ w &\mapsto \big\| \mathcal{B}\big(\mathcal{A}(p), p\big) - w_\varepsilon \big\|_P \,, \end{aligned}$$

definiert, welches den Grad der Abweichung der zum Parameter $p \in P$ gehörenden Beobachtung w von den mit Messfehlern behafteten Daten w_ε angibt. Unter Verwendung von

$$\begin{aligned} &\mathcal{J} : P \to \mathbb{R} \,, \\ &\quad p \mapsto \tilde{\mathcal{J}} \circ \mathcal{B}\big(\mathcal{A}(p), p\big) = \sum_{k=1}^{\kappa} \sum_{i=1}^{n_k} \tilde{\mathcal{J}}_{k,i} \circ \mathcal{B}_{k,i}\big(\mathcal{A}(p), p\big)\big(\mathcal{A}(p), p\big) \,, \end{aligned} \tag{3.4}$$

kann damit ein optimaler Parametersatz $\bar{p}$ durch

$$\mathcal{J}(\bar{p}) = \min_{p \in P} \mathcal{J}(p) \tag{3.5}$$

bestimmt werden.

3.1.2 Differentiation des Fehlerfunktionals

Das Identifizierungsproblem (3.5) entspricht einer nichtlinearen Optimierungsaufgabe, welche zur Formulierung notwendiger Optimalitätsbedingungen die Ableitung des Zielfunktionals nach den Parametern benötigt. Im folgenden Abschnitt

wird aufgezeigt, wie diese mit Hilfe eines adjungierten Problems effizient berechnet werden kann (vgl. u.A. Igler [34]).

Zunächst werden kurz verallgemeinerte Ableitungsbegriffe eingeführt (vgl. z.B. Alt [2]).

Definition 3.2
*Sei X Banachraum. Eine Teilmenge $M \subset X$ heißt **offen** in X, falls zu jedem Punkt $x \in M$ eine Kugel $B_r(x) \subset M$ mit $r := r(x) > 0$ existiert. Eine Menge U heißt **Umgebung** der Menge $V \subset X$, falls eine in X offene Menge W mit $V \subset W \subset U$ existiert.*

Definition 3.3 (Fréchet-Ableitung)
Seien X, Z Banachräume, $D \subseteq X$ eine Umgebung zu gegebenen $x_0 \in X$ und $\mathcal{F}: D \to Z$ eine Abbildung von D nach Z. Existiert ein Operator $\mathcal{F}' \in \mathcal{L}(X, Z)$, so dass

$$\lim_{\|h\|_X \to 0} \frac{\|\mathcal{F}(x_0+h) - \mathcal{F}(x_0) - \mathcal{F}'[x_0]h\|_Z}{\|h\|_X} = 0 \tag{3.6}$$

*erfüllt ist, so heißt $\mathcal{F}$ an der Stelle x_0 **Fréchet-differenzierbar** und $\mathcal{F}'[x_0]$ **Fréchet-Ableitung** von $\mathcal{F}$ an der Stelle x_0.*

Beispiel 3.4
Sei der Beobachtungsraum W wie in (3.1) und (3.2) unterteilt, $\|.\|_{W_{k,i}}$ die zugehörige Norm auf $W_{k,i}$, $k = 1, \ldots, \kappa, i = 1, \ldots, n_k$, und

$$\|.\|_W = \sum_{k=1}^{\kappa} \sum_{i=1}^{n_k} \|.\|_{W_{k,i}} \tag{3.7}$$

*die auf W definierte Norm. Ein Beispiel für das in (3.4) eingeführte Fehlerfunktional $\tilde{J}$ ist der **gewichtete OLS-Ansatz** welcher durch*

$$\begin{aligned} &\tilde{\mathcal{J}}_{k,i} : W_{k,i} \to \mathbb{R}, \\ &\qquad w_{k,i} \mapsto \Lambda_{k,i}(w_\varepsilon) \Big\| w_{k,i} - (w_\varepsilon)_{k,i} \Big\|^2_{W_{k,i}} \quad \forall\ k = 1, \ldots, \kappa,\ i = 1, \ldots, n_k, \end{aligned} \tag{3.8}$$

und

$$\begin{aligned} \tilde{\mathcal{J}} : W \to \mathbb{R}, \quad w \mapsto &\sum_{k=1}^{\kappa} \Lambda_k(E_k) \sum_{i=1}^{n_k} \tilde{\mathcal{J}}_{k,i}\big(w|_{W_{k,i}}\big) = \\ &\sum_{k=1}^{\kappa} \Lambda_k(E_k) \sum_{i=1}^{n_k} \Lambda_{k,i}(w_\varepsilon) \Big\| w_{k,i} - (w_\varepsilon)_{k,i} \Big\|^2_{W_{k,i}} \end{aligned} \tag{3.9}$$

mit den Gewichtungsfaktoren $\Lambda_k(E_k), \Lambda_{k,i}(w_{\varepsilon,ki}) \in \mathbb{R}_0^+,\ k=1,\dots,\kappa, i=1,\dots,n_k$, *gegeben ist. Sind dabei alle* $W_{k,i}$ *Hilberträume mit den entsprechenden Skalarprodukten* $(.,.)_{W_{k,i}}$, *dann sind* $\tilde{J}_{k,i}$ *Fréchet-differenzierbar auf* $W_{k,i}$,

$$\begin{aligned}\tilde{\mathcal{J}}'_{k,i} : W_{k,i} &\to \mathcal{L}(W_{k,i}, \mathbb{R})\,, \\ w_{k,i} &\mapsto \tilde{J}'_{k,i}[w_{k,i}]\,, \quad \forall\, k=1,\dots,\kappa,\ i=1,\dots,n_k\,,\end{aligned}$$

mit

$$\Big\langle \tilde{J}'_{k,i}[w_{k,i}]\,, v_{k,i} \Big\rangle = 2\Lambda_{k,i}(w_\varepsilon)\Big(w_{k,i} - (w_\varepsilon)_{k,i}, v_{k,i}\Big)_{W_{k,i}} \quad \forall\, v_{k,i} \in W_{k,i} \tag{3.10}$$

und entsprechend $\tilde{J}$ *Fréchet-differenzierbar auf* W,

$$\begin{aligned}\tilde{\mathcal{J}}' : W &\to \mathcal{L}(W, \mathbb{R})\,, \\ w &\mapsto \tilde{J}[w]\,,\end{aligned}$$

mit

$$\begin{aligned}\Big\langle \tilde{J}'[w]\,, v \Big\rangle &= \sum_{k=1}^{\kappa} \Lambda_k(E_k) \sum_{i=1}^{n_k} \Big\langle \tilde{J}'_{k,i}\big[w|_{W_{k,i}}\big]\,, v|_{W_{k,i}} \Big\rangle \\ &= 2\sum_{k=1}^{\kappa} \Lambda_k(E_k) \sum_{i=1}^{n_k} \Lambda_{k,i}(w_\varepsilon)\Big(w_{k,i} - (w_\varepsilon)_{k,i}\,, v_{k,i}\Big)_{W_{k,i}} \quad \forall\, v \in W\,.\end{aligned} \tag{3.11}$$

Beweis:
Sei das Fehlerfunktional nach (3.8) und (3.9) gegeben. Zu zeigen ist, dass $\tilde{J}$ mit der durch (3.10) und (3.11) definierten Abbildung $\tilde{J}'$ Gleichung (3.6) für jedes beliebig (aber fest) gewählte $w \in W$ erfüllt. Es gilt

$$\lim_{\|h\|_W \to 0} \frac{\big|\tilde{J}(w+h) - \tilde{J}(w) - \big\langle \tilde{J}'[w], h\big\rangle\big|}{\|h\|_W} = \tag{3.12}$$

$$\begin{aligned}= \lim_{\|h\|_W \to 0} \Bigg| &\sum_{k=1}^{\kappa} \Lambda_k(E_k) \sum_{i=1}^{n_k} \tilde{\mathcal{J}}_{k,i}\big((w+h)|_{W_{k,i}}\big) - \sum_{k=1}^{\kappa} \Lambda_k(E_k) \sum_{i=1}^{n_k} \tilde{\mathcal{J}}_{k,i}\big(w|_{W_{k,i}}\big) \\ &- \sum_{k=1}^{\kappa} \Lambda_k(E_k) \sum_{i=1}^{n_k} \Big\langle \tilde{\mathcal{J}}'_{k,i}\big[w|_{W_{k,i}}\big], h|_{W_{k,i}} \Big\rangle \Bigg| \Big/ \|h\|_W\end{aligned}$$

und mit (3.8) sowie $h_{k,i} := h|_{W_{k,i}}$

$$
\begin{aligned}
= \lim_{\|h\|_W \to 0} \Bigg| & \sum_{k=1}^{\kappa} \Lambda_k(E_k) \sum_{i=1}^{n_k} \Lambda_{k,i}(w_\varepsilon) \Big\| w_{k,i} + h_{k,i} - (w_\varepsilon)_{k,i} \Big\|_{W_{k,i}}^2 \\
& - \sum_{k=1}^{\kappa} \Lambda_k(E_k) \sum_{i=1}^{n_k} \Lambda_{k,i}(w_\varepsilon) \Big\| w_{k,i} - (w_\varepsilon)_{k,i} \Big\|_{W_{k,i}}^2 \\
& - 2 \sum_{k=1}^{\kappa} \Lambda_k(E_k) \sum_{i=1}^{n_k} \Lambda_{k,i}(w_\varepsilon) \Big(w_{k,i} - (w_\varepsilon)_{k,i} \,,\, h_{k,i} \Big)_{W_{k,i}} \Bigg| \Big/ \|h\|_W \,.
\end{aligned}
$$

Da für jedes $k = 1, \ldots, \kappa$ und jedes $i = 1, \ldots, n_k$

$$
\Big\| w_{k,i} + h_{k,i} - (w_\varepsilon)_{k,i} \Big\|_{W_{k,i}}^2 = \Big\| w_{k,i} - (w_\varepsilon)_{k,i} \Big\|_{W_{k,i}}^2 + 2 \Big(w_{k,i} - (w_\varepsilon)_{k,i} \,,\, h_{k,i} \Big)_{W_{k,i}} + \big\| h_{k,i} \big\|_{W_{k,i}}^2
$$

gilt, ist (3.12) äquivalent zu

$$
\lim_{\|h\|_W \to 0} \frac{\sum_{k=1}^{\kappa} \Lambda_k(E_k) \sum_{i=1}^{n_k} \Lambda_{k,i}(w_\varepsilon) \big\| h_{k,i} \big\|_{W_{k,i}}^2}{\|h\|_W} \,,
$$

so dass mit

$$
\Lambda_{\max} := \max_{k=1,\ldots,\kappa} \Big\{ \Lambda_k(E_k) \sum_{i=1}^{n_k} \Lambda_{k,i}(w_\varepsilon) \Big\}
$$

und der auf dem gesamten Beobachtungsraum W definierten Norm (3.7) bereits

$$
\begin{aligned}
0 &\leq \lim_{\|h\|_W \to 0} \frac{\big| \tilde{J}(w+h) - \tilde{J}(w) - \big\langle \tilde{J}'[w], h \big\rangle \big|}{\|h\|_W} \\
&\leq \lim_{\|h\|_W \to 0} \kappa \Lambda_{\max} \frac{\big\| h_{i,k} \big\|_{W_{i,k}}^2}{\|h\|_W} \leq \lim_{\|h\|_W \to 0} \kappa \Lambda_{\max} \big\| h \big\|_W = 0
\end{aligned}
$$

folgt.

□

Definition 3.5
Seien X, Y, Z Banachräume, $\mathcal{F}$ eine Abbildung der Form $\mathcal{F} : D \subseteq X \times Y \to Z$ mit $(x, y) \mapsto \mathcal{F}(x, y)$, $y_0 \in Y$ beliebig aber fest und $\mathcal{G}(x) := \mathcal{F}(x, y_0)$. Falls $\mathcal{G}$ Fréchet-differenzierbar im Punkt $x_0 \in X$ ist, heißt $\mathcal{F}$ an der Stelle (x_0, y_0) ***partiell Fréchet-differenzierbar*** *bzgl. Komponente x und*

$$
F_x[x_0, y_0] := \mathcal{G}'[x_0]
$$

partielle Fréchet-Ableitung *bzgl. Komponente x im Punkt (x_0, y_0). Ist entsprechend $x_0 \in X$ fest gewählt und $\mathcal{H}(y) := \mathcal{F}(x_0, y)$ Fréchet-differenzierbar im Punkt $y_0 \in Y$, so ist $\mathcal{F}$ im Punkt (x_0, y_0) Fréchet-differenzierbar bzgl. Komponente y und $\mathcal{F}_y := \mathcal{H}'[y_0]$ die zugehörige partielle Fréchet-Ableitung.*

Satz 3.6 (Satz über implizite Funktionen)
Seien X, Y, Z Banachräume, $Z^* = \mathcal{L}(Z, \mathbb{R})$ der duale Raum von Z und $\mathcal{F}$ eine Abbildung der Form $\mathcal{F} : X \times Y \to Z^*$, $(x, y) \mapsto \mathcal{F}(x, y)$. Weiter sei für einen Punkt $(x_0, y_0) \in X \times Y$ eine Umgebung $D \subseteq X \times Y$ gegeben, so dass die folgenden Bedingungen

1. *$\mathcal{F}$ ist stetig in D mit $\mathcal{F}(x_0, y_0) = 0$,*
2. *die partiellen Fréchet-Ableitungen $\mathcal{F}_x$ und $\mathcal{F}_y$ existieren in jedem Punkt $(x, y) \in D$ und sind stetig in D,*
3. *der inverse Operator*
$$\big(\mathcal{F}_x[x_0, y_0]\big)^{-1} : Z^* \to D$$
existiert und ist linear und beschränkt,

erfüllt sind. Dann existiert eine Umgebung $\tilde{Y} \subseteq Y$ von y_0 und ein Operator $\mathcal{G} : \tilde{Y} \to X$, so dass $\mathcal{F}(x, y) = 0$ direkt $x = \mathcal{G}(y)$ $\forall\, y \in \tilde{Y}$ impliziert, entsprechend

$$\mathcal{F}\big(\mathcal{G}(y), y\big) = 0 \quad \forall\, y \in Y$$

gilt und $\mathcal{G}$ Fréchet-differenzierbar in $\tilde{Y}$ mit

$$\mathcal{G}'[y] = -\Big(\mathcal{F}_x\big[\mathcal{G}(y), y\big]\Big)^{-1} \mathcal{F}_y\big[\mathcal{G}(y), y\big] \quad \forall\, y \in \tilde{Y}$$

ist.

Beweis:
Siehe z.B. Wouk [65], Chap. 12.4, Theorem 12.4.1 und Corollary 1.

□

Im Folgenden wird die Ableitung des Fehlerfunktionals (3.4) nach den Parametern berechnet. Sei hierzu das Modell des direkten Problems als nichtlineares Gleichungssystem (schwache Formulierung)

$$\mathcal{T}(u, p) = 0 \tag{3.13}$$

mit nichtlinearer Abbildung

$$\begin{aligned} \mathcal{T} &: U \times P \to V^*, \\ &(u, p) \mapsto \mathcal{T}(u, p)\,, \end{aligned} \tag{3.14}$$

gegeben. Sind die Voraussetzungen des Satzes für implizite Funktionen für $\mathcal{T}$ im Punkt $(u_0, p_0) \in U \times P$ erfüllt, so ist die Existenz einer Umgebung $\tilde{P} \subseteq P$ von p_0 mit

$$\mathcal{T}\big(\mathcal{A}(p), p\big) = 0 \quad \forall\, p \in \tilde{P}$$

und die Fréchet-Differenzierbarkeit des Operators $\mathcal{A}$ in $\tilde{P}$,

$$\mathcal{A}'[p] = -\Big(\mathcal{T}_u\big[\mathcal{A}(p),p\big]\Big)^{-1}\mathcal{T}_p\big[\mathcal{A}(p),p\big] \quad \forall\, p \in \tilde{P}\,,$$

gesichert. Ist zudem die Fréchet-Differenzierbarkeit des Beobachtungsoperators $\mathcal{B}$ und die der Abbildung $\tilde{\mathcal{J}}$ in $\tilde{P}$ vorausgesetzt/gegeben, so ist das zu untersuchende Fehlerfunktional $\mathcal{J}$ aus (3.4) wohldefiniert und ebenfalls auf $\tilde{P}$ Fréchet-differenzierbar. Für ein $p\in\tilde{P}$ gilt dann unter Anwendung der Kettenregel

$$\begin{aligned}
\mathcal{J}'[p] &= \tilde{\mathcal{J}}'\big[\mathcal{B}(u,p)\big]\frac{d}{dp}\mathcal{B}\big(\mathcal{A}(p),p\big)\\
&= \tilde{\mathcal{J}}'\big[\mathcal{B}(u,p)\big]\Big(\mathcal{B}_u\big[u,p\big]\mathcal{A}'[p]+\mathcal{B}_p\big[u,p\big]\Big)\\
&= \tilde{\mathcal{J}}'\big[\mathcal{B}(u,p)\big]\left(\mathcal{B}_p\big[u,p\big]-\mathcal{B}_u\big[u,p\big]\Big(\mathcal{T}_u\big[\mathcal{A}(p),p\big]\Big)^{-1}\mathcal{T}_p\big[\mathcal{A}(p),p\big]\right).
\end{aligned}$$

Um bei der Berechnung der Ableitung den inversen Operator zu vermeiden, empfiehlt es sich, dass nachfolgende adjungierte Problem zu lösen (vgl. z.B. Bitterlich [8] oder Igler [34]). Im Sinne einer besseren Lesbarkeit (und zur Verdeutlichung der Reellwertigkeit) wird für $x.\delta x\in X$

$$\big\langle x,\mathcal{F}\big\rangle_X := \mathcal{F}(x) \quad \text{und} \quad \big\langle \mathcal{F}'[x],\delta x\big\rangle_{X^*} := \mathcal{F}'[x]\delta x$$

verwendet.

Satz 3.7 (Adjungiertes Problem)
Gegeben seien das nach (3.4) definierte Fehlerfunktional $\mathcal{J}$ und der nach (3.14) festgelegte Operator $\mathcal{T}$. Seien die Voraussetzungen des Satzes über implizite Funktionen für $\mathcal{T}$ und einem Punkt $(u_0,p_0)\in U\times P$ erfüllt und $\tilde{P}$ eine hieraus folgernde Umgebung von p_0 mit entsprechenden Eigenschaften. Des Weiteren seien $\mathcal{B}$ und $\tilde{\mathcal{J}}$ auf $\tilde{P}$ Fréchet-differenzierbar, $p\in\tilde{P}$ und $u=\mathcal{A}(p)$. Dann ist, falls $\eta\in V$ für alle $\delta u\in U$ das adjungierte Problem

$$\Big\langle \eta,\mathcal{T}_u[u,p]\delta u\Big\rangle_V = \Big\langle \tilde{\mathcal{J}}'\big[\mathcal{B}(u,p)\big],\mathcal{B}_u[u,p]\delta u\Big\rangle_W$$

löst, die Ableitung des Fehlerfunktionals $\mathcal{J}$ für alle $\delta p\in P$ durch

$$\Big\langle \mathcal{J}'[p],\delta p\Big\rangle_{P^*} = \Big\langle \tilde{\mathcal{J}}'\big[\mathcal{B}(u,p)\big],\mathcal{B}_p[u,p]\delta p\Big\rangle_{W^*} - \Big\langle \eta,\mathcal{T}_p[u,p]\delta p\Big\rangle_{V^*}$$

gegeben.

Beweis:
vgl. Bitterlich [8], Satz 2.3, Punkt 2.

□

Bemerkung 3.8
Unter Verwendung der in (3.4) explizit angegebenen Aufspaltung des Fehlerfunktionals

$$\mathcal{J}(p) = \sum_{k=1}^{\kappa} \sum_{i=1}^{n_k} \tilde{\mathcal{J}}_{k,i} \circ \mathcal{B}_{k,i}\big(\mathcal{A}(p), p\big)\,,$$

kann die Ableitung entsprechend durch

$$\Big\langle \mathcal{J}'[p], \delta p \Big\rangle_{P^*} = \sum_{k=1}^{\kappa} \sum_{i=1}^{n_k} \left\langle \tilde{\mathcal{J}}'_{k,i}\big[\mathcal{B}_{k,i}(u,p)\big], \frac{\partial \mathcal{B}_{k,i}}{\partial p}[u,p]\delta p \right\rangle_{W^*} - \left\langle \eta, \frac{\partial \mathcal{T}}{\partial p}[u,p]\delta p \right\rangle_{V^*}$$

angegeben werden, falls $\eta \in V$ für alle $\delta u \in U$ das adjungierte Problem

$$\left\langle \eta, \frac{\partial \mathcal{T}}{\partial u}[u,p]\delta u \right\rangle_{V^*} = \sum_{k=1}^{\kappa} \sum_{i=1}^{n_k} \left\langle \tilde{\mathcal{J}}'_{k,i}\big[\mathcal{B}_{k,i}(u,p)\big], \frac{\partial \mathcal{B}_{k,i}}{\partial u}[u,p]\delta u \right\rangle_{W^*}$$

löst.

3.1.3 Diskrete Problemformulierung

Um eine numerische Lösung zu erhalten wird die kontinuierliche Modellgleichung (3.13) mittels eines geeigneten Diskretisierungsverfahrens in ein endlichdimensionales Problem überführt (vgl. z.B. Knabner [38] oder Quarteroni [47]). Es entsteht ein Gleichungssystem

$$\mathcal{T}_h(u_h, p_h) = 0 \tag{3.15}$$

mit nichtlinearer Abbildung

$$\begin{aligned} \mathcal{T}_h : \mathbb{R}^m \times \mathbb{R}^r &\to \mathbb{R}^m\,, \quad m, r \in \mathbb{N}\,, \\ (u_h, p_h) &\mapsto \mathcal{T}_h(u_h, p_h)\,. \end{aligned}$$

Sowohl die diskrete Lösung $u_h \in U_h \subset \mathbb{R}^m$ als auch der diskrete Parametersatz $p_h \in P_h \subset \mathbb{R}^r$ werden nun als reellwertige Vektoren angesehen. Analog zum kontinuierlichen Fall wird ein diskreter direkter Lösungsoperator

$$\begin{aligned} \mathcal{A}_h : \mathbb{R}^r &\to \mathbb{R}^m\,, \\ p_h &\mapsto u_h\,, \end{aligned}$$

und ein diskreter Beobachtungsoperator

$$\mathcal{B}_h := \begin{pmatrix} \big(\mathcal{B}_h\big)_1 \\ \vdots \\ \big(\mathcal{B}_h\big)_\kappa \end{pmatrix} : \mathbb{R}^m \times \mathbb{R}^r \to \begin{pmatrix} \big(W_h\big)_1 \\ \vdots \\ \big(W_h\big)_\kappa \end{pmatrix} := W_h = \mathbb{R}^n,$$

$$(u_h, p_h) \mapsto \begin{pmatrix} \big(\mathcal{B}_h\big)_1(u_h, p_h) \\ \vdots \\ \big(\mathcal{B}_h\big)_\kappa(u_h, p_h) \end{pmatrix} = \mathcal{B}(u_h, p_h) =: w_h = \begin{pmatrix} (w_h)_1 \\ \vdots \\ (w_h)_\kappa \end{pmatrix}$$

mit

$$\big(\mathcal{B}_h\big)_k := \begin{pmatrix} \big(\mathcal{B}_h\big)_{k,1} \\ \vdots \\ \big(\mathcal{B}_h\big)_{k,n_k} \end{pmatrix} : \mathbb{R}^m \times \mathbb{R}^r \to \begin{pmatrix} \big(W_h\big)_{k,1} \\ \vdots \\ \big(W_h\big)_{k,n_k} \end{pmatrix} := \big(W_h\big)_k,$$

$$(u_h, p_h) \mapsto \begin{pmatrix} \big(\mathcal{B}_h\big)_{k,1}(u_h, p_h) \\ \vdots \\ \big(\mathcal{B}_h\big)_{k,n_k}(u_h, p_h) \end{pmatrix} = \big(\mathcal{B}_h\big)_k(u_h, p_h) =: (w_h)_k = \begin{pmatrix} (w_h)_{k,1} \\ \vdots \\ (w_h)_{k,n_k} \end{pmatrix},$$

$k = 1, \ldots, \kappa$, $i = 1, \ldots, n_k$, eingeführt.

Bemerkung 3.9
Der im Abschnitt 3.1.2 vorgestellte Satz über implizite Funktionen (Satz 3.6) gilt auch im Endlichdimensionalen. Die vorkommenden (partiellen) Fréchet-Ableitungen sind im $\mathbb{R}^k$, $k \in \mathbb{N}$, definitionsgemäß äquivalent zu gewöhnlichen (partiellen) Ableitungen.

Sind die Voraussetzungen des Satzes 3.6 auch im diskreten Fall erfüllt, so existiert entsprechend eine nichtleere Teilmenge $\tilde{P}_h \subset P_h$, für die der diskrete Beobachtungsoperator direkt durch

$$\begin{aligned} \mathcal{B}_h : \mathbb{R}^r &\to \mathbb{R}^n, \\ p_h &\mapsto w_h = \mathcal{B}_h\big(\mathcal{A}_h(p_h), p_h\big)\,, \end{aligned}$$

dargestellt werden kann.

Unter der Annahme einer gestörten diskreten Messung $w_{h,\varepsilon} \in \mathbb{R}^n$,

$$\|w_{h,0} - w_{h,\varepsilon}\|_2 \leq \varepsilon_h\,,$$

$w_{h,0} \in \mathbb{R}^n$ ungestörte diskrete Beobachung, wird, analog zur kontinuierlichen Problemstellung, ein zu minimierendes diskretes Fehlerfunktional

$$\mathcal{J}_h : \mathbb{R}^r \to \mathbb{R},$$

$$p_h \mapsto \tilde{\mathcal{J}}_h \circ \mathcal{B}_h\big(\mathcal{A}_h(p_h), p_h\big) = \sum_{k=1}^{\kappa} \sum_{i=1}^{n_k} \big(\tilde{\mathcal{J}}_h\big)_{k,i} \circ \big(\mathcal{B}_h\big)_{k,i}\big(\mathcal{A}_h(p_h), p_h\big) \tag{3.16}$$

definiert. Um die zugehörige Ableitung $\mathcal{J}_h'$ effizient zu berechnen, empfiehlt es sich auch hier ein adjungiertes Problem zu lösen. (Der nachfolgende Satz stellt aufgrund des erweiterten Beobachtungsoperators eine (kleine) Erweiterung der in Bitterlich [8] vorgestellten Berechnung dar).

Satz 3.10 (Adjungiertes Problem)
Gegeben seien das in (3.16) definierte diskrete Fehlerfunktional $\mathcal{J}_h$ und der in (3.15) definierte diskrete Operator $\mathcal{T}_h$. Seien die Voraussetzungen des Satzes über implizite Funktionen (Satz 3.6) für $\mathcal{T}_h$ und einem Punkt $p_h \in P_h$ erfüllt und $\tilde{P}_h$ eine hieraus folgernde Umgebung von p_h mit entsprechenden Eigenschaften. Des Weiteren seien $\mathcal{B}_h$ und $\tilde{\mathcal{J}}_h$ auf $\tilde{P}_h$ differenzierbar und $u_h = \mathcal{A}_h(p_h)$. Wenn $\eta_h \in \mathbb{R}^m$ das adjungierte Problem

$$\eta_h^T \frac{\partial \mathcal{T}_h}{\partial u_h} = \sum_{k=1}^{\kappa} \sum_{i=1}^{n_k} \tilde{\mathcal{J}}'_{k,i} \frac{\partial (\mathcal{B}_h)_{k,i}}{\partial u_h}$$

löst, dann ist das diskrete Fehlerfunktional $\mathcal{J}_h : P_h \subset \mathbb{R}^r \to \mathbb{R}$ wohldefiniert, in $\tilde{P}_h$ differenzierbar und die Ableitung ist durch

$$\mathcal{J}_h' = \sum_{k=1}^{\kappa} \sum_{i=1}^{n_k} \big(\tilde{\mathcal{J}}_h\big)'_{k,i} \frac{\partial (\mathcal{B}_h)_{k,i}}{\partial p_h} - \eta_h^T \frac{\partial \mathcal{T}_h}{\partial p_h}$$

gegeben.

Beweis:
Die Wohldefiniertheit und Differenzierbarkeit ist bereits durch den Satz über implizite Funktionen belegt. Für die Ableitung gilt unter Anwendung der Kettenregel

$$\begin{aligned}
\mathcal{J}_h' &= \frac{d}{dp_h}\Big(\tilde{\mathcal{J}}_h \circ \mathcal{B}_h\big(\mathcal{A}_h(p_h), p_h\big)\Big) \\
&= \sum_{k=1}^{\kappa} \sum_{i=1}^{n_k} \big(\tilde{\mathcal{J}}_h\big)'_{k,i} \frac{d}{dp_h}\Big(\big(\mathcal{B}_h\big)_{k,i}\big(\mathcal{A}_h(p_h), p_h\big)\Big) \\
&= \sum_{k=1}^{\kappa} \sum_{i=1}^{n_k} \big(\tilde{\mathcal{J}}_h\big)'_{k,i} \frac{\partial (\mathcal{B}_h)_{k,i}}{\partial p_h} + \sum_{k=1}^{\kappa} \sum_{i=1}^{n_k} \big(\tilde{\mathcal{J}}_h\big)'_{k,i} \frac{\partial (\mathcal{B}_h)_{k,i}}{\partial u_h} \mathcal{A}_h'
\end{aligned}$$

und mit η_h schließlich

$$\mathcal{J}_h' = \sum_{k=1}^{\kappa} \sum_{i=1}^{n_k} \left(\tilde{\mathcal{J}}_h\right)'_{k,i} \frac{\partial (\mathcal{B}_h)_{k,i}}{\partial p_h} + \eta_h^T \frac{\partial \mathcal{T}_h}{\partial u_h} \mathcal{A}_h' \,. \tag{3.17}$$

Vollständiges Differenzieren der diskreten Modellgleichung $\mathcal{T}_h\big(\mathcal{A}_h(p_h, p_h)\big)$

$$0 = \frac{\partial \mathcal{T}_h}{\partial u_h} \mathcal{A}_h' + \frac{\partial \mathcal{T}_h}{\partial p_h}$$

liefert letztendlich durch Umstellen und Einsetzen in (3.17) die Behauptung.

□

Auf dem diskreten Beobachtungsraum kann, analog zu Beispiel 3.4, das (gewichtete) diskrete Fehlerfunktional

$$\tilde{\mathcal{J}}_h(w_h) = \sum_{k=1}^{\kappa} \Lambda_k(E_k) \sum_{i=1}^{n_k} \Lambda_{k,i}(w_{h,\varepsilon}) \sum_{j=1}^{m_{k,i}} \Big(\big(w_h\big)_{k,i,j} - \big(w_{h,\varepsilon}\big)_{k,i,j} \Big)^2 , \tag{3.18}$$

$m_{k,i} = \dim\big((W_h)_{k,i}\big)$, $\big(w_{h,\varepsilon}\big)_{k,i} := w_{h,\varepsilon}|_{(W_h)_{k,i}}$, definiert werden.

Bemerkung 3.11

- *Setzt man alle $\Lambda_k(w_{h,\varepsilon})$, $\Lambda_{k,i}(w_{h,\varepsilon})$ gleich eins, so ist (3.18) äquivalent zu dem (gewöhnlichen) diskreten OLS-Ansatz*

$$\tilde{\mathcal{J}}_h(w_h) = \sum_{j=1}^{n} \Big(\big(w_h\big)_{k,i,j} - \big(w_{h,\varepsilon}\big)_{k,i,j} \Big)^2 .$$

- *Das auf dem diskreten Beobachtungsraum operierende Fehlerfunktional (3.18) ist auf $\tilde{P}$ differenzierbar und die Ableitungen ist gegeben durch*

$$\tilde{\mathcal{J}}_h'(w_h) = 2 \sum_{k=1}^{\kappa} \Lambda_k(E_k) \sum_{i=1}^{n_k} \Lambda_{k,i}(w_{h,\varepsilon}) \sum_{j=1}^{m_{k,i}} \Big(\big(w_h\big)_{k,i,j} - \big(w_{h,\varepsilon}\big)_{k,i,j} \Big) .$$

Beweis:
Die Äquivalenz der Ansätze ist mit

$$n = \sum_{k=1}^{\kappa} \sum_{i=1}^{n_k} m_{k,i} \,.$$

trivialerweise erfüllt. Die Ableitung des diskreten Fehlerfunktionals kann analog zu (3.9) berechnet werden.

□

3.2 Formfreie Parametrisierung

Anstelle einer fest vorgegebenen Parametrisierung, wie sie beispielsweise unter Abschnitt 2.1.4 für die hydraulischen Funktionen vorgestellt wurde, kann auch eine **formfreie Parametrisierung** gewählt werden. Hierbei wird anstelle der Parameterwerte in den vordefinierten Abbildungen die Nichtlinearität selbst gesucht.

3.2.1 Nichtlinearitäten $f : \mathbb{R} \to \mathbb{R}$

In diesem Abschnitt werden die bereits in Bitterlich [8], Kap. 2.3 und 4.1 und Igler [34], Kap. 3.5 und 4.5 vorgestellten Untersuchungen bezüglich stetiger, skalarwertiger Funktionen $f \in C(I)$, $I \subset \mathbb{R}$ Intervall in $\mathbb{R}$, betrachtet. Die Notation wurde dabei auf die im Anschluß betrachteten Funktionen des $\mathbb{R}^d$ angepasst und im Fall hierarchischer Basen für eine beliegige (Start-)Untergliederung des Intervalls (bestehend aus den beiden Rand- und beispielsweise einem inneren Knoten) erweitert. Damit ist eine (z.B. linksseitige) Gewichtung der hierarchisch zu bestimmenden Knotenmenge möglich.

Um einen Multi-Level-Algorithmus zur Identifizierung einer gesuchten Funktion $f \in P$ herzuleiten wird für den (kontinuierlichen) Parameterraum P eine Folge endlichdimensionaler Unterräume $(P_r)_{r\in\mathbb{N}}$ mit

$$\dim(P_r) = r, \quad P_r \subset P_{r'} \text{ für } r < r'$$

und

$$\overline{\bigcup_{r\in\mathbb{N}} P_r} = P$$

definiert. Ist die Basis von P_r durch die Menge $\left\{\phi_{r,\nu}\right\}_{\nu=1}^{r}$ gegeben, so lässt sich jede Funktion $f_r \in P_r$ durch einen eindeutig bestimmbaren vektorwertigen Parameter $\vec{p}_r = (p_{r,1}, \ldots, p_{r,r})^T \in \mathbb{R}^r$ in der Form

$$f_r(x) = \sum_{\nu=1}^{r} p_{r,\nu}\phi_{r,\nu}(x)$$

darstellen. Folgelich kann jeder Unterraum P_r mit dem $\mathbb{R}^r$ identifiziert werden.

Im Folgenden wird zur Bestimmung geeigneter Basen ausgewählter Unterräume sowohl der traditionelle Ansatz lokaler Basen als auch die skalenweise Parametrisierung mit hierarchischen Basen vorgestellt.

3.2.1.1 Lokale Basen (B-Splines)

Die Wahl der Basis ist maßgeblich entscheidend für die Effektivität der anzuwendenden numerischen Interpolationsverfahren. Als numerisch geschickt haben sich die sogenannten B-Splines (vgl. für äquidistante Gitter z.B. Hämmerlin, Hoffmann [31]) herausgestellt, da in ihrem Fall das entsprechende Gleichungssystem entweder trivial ist oder eine einfach zu lösende Bandstruktur besitzt.

Sei im Folgenden das Intervall $[a,b]$ mit Hilfe der Knotenmenge

$$\chi_{\check{r}} := \{x_{\check{r},\upsilon}\}_{\upsilon=1}^{\check{r}}\,,\ \check{r}=r-k+1\,, \tag{3.19}$$

in

$$a = x_{\check{r},1} < \cdots < x_{\check{r},\check{r}} = b\,,$$

unterteilt, $B_{\chi_{\check{r}},k} := \left\{\phi_{\chi_{\check{r}},k,\nu}\right\}_{\nu=1}^{r}$ die auf $\chi_{\check{r}}$ definierten B-Splines k-ter Ordnung und $P^{\text{lok}}_{\chi_{\check{r}},k,r}=P_r$ der durch $B_{\chi_{\check{r}},k}$ aufgespannte Raum.

Da unter Verwendung von konstanten Splines die aufgespannten Funktionen in ihren Stützstellen Unstetigkeiten aufweisen und ihre Ableitungen in allen Stetigkeitspunkten verschwinden, sind lineare Splines

$$\phi_{\chi_{\check{r}},1,1}(x) := \tfrac{x_{\check{r},2}-x}{x_{\check{r},2}-x_{\check{r},1}} \text{ für } x\in[x_{\check{r},1},x_{\check{r},2}]\,,$$

$$\phi_{\chi_{\check{r}},1,\nu}(x) := \begin{cases} \tfrac{x-x_{\check{r},\nu-1}}{x_{\check{r},\nu}-x_{\check{r},\nu-1}} & \text{für } x\in[x_{\check{r},\nu-1},x_{\check{r},\nu}] \\ \tfrac{x_{\check{r},\nu+1}-x}{x_{\check{r},\nu+1}-x_{\check{r},\nu}} & \text{für } x\in[x_{\check{r},\nu},x_{\check{r},\nu+1}] \end{cases}, \quad \nu=2,\ldots,r-1,$$

$$\phi_{\chi_{\check{r}},1,r}(x) := \tfrac{x-x_{\check{r},\check{r}-1}}{x_{\check{r},\check{r}}-x_{\check{r},\check{r}-1}} \text{ für } x\in[x_{\check{r},\check{r}-1},x_{\check{r},\check{r}}]\,,$$

$\phi_{\chi_{\check{r}},1,\nu}(x)=0$, $\nu=1,\ldots,r$, sonst, die einfachste Wahl nichttrivialer Basisfunktionen. Abbildung 3.1 zeigt exemplarisch $\phi_{\chi_{\check{r}},1,\nu}(x)$ für ein $\nu\in\{2,\ldots,r-1\}$.

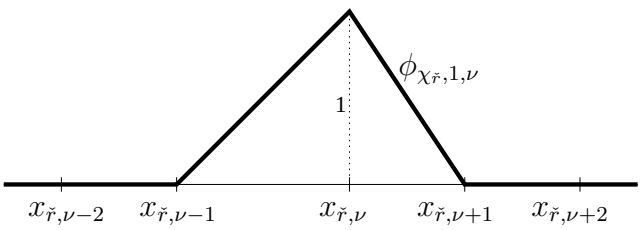

Abbildung 3.1: Linearer B-Spline

Aufgrund der Unabhängigkeit der gewählten Basisfunktionen lässt sich jede Funktion $f^{\text{lok}}_{\chi_{\check{r}},1,r}\in P^{\text{lok}}_{\chi_{\check{r}},1,r}$ auf $[a,b]$ eindeutig durch einen Parametersatz

$$\vec{p}^{\,\text{lok}}_{\chi_{\check{r}},1,r} := (p^{\text{lok}}_{\chi_{\check{r}},1,1},\ldots,p^{\text{lok}}_{\chi_{\check{r}},1,r})^T\in\mathbb{R}^r$$

in der Form

$$f^{\text{lok}}_{\chi_{\check{r}},1,r}(x) = \sum_{\nu=1}^{r} p^{\text{lok}}_{\chi_{\check{r}},1,\nu}\phi_{\chi_{\check{r}},1,\nu}(x)$$

darstellen. Sind für ein $f \in P$ die diskreten Funktionswerte $f(x_{\check{r},\nu})$, $\nu = 1,\ldots,\check{r}$, gegeben, so ist die Interpolationsaufgabe

$$f^{\text{lok}}_{\chi_{\check{r}},1,r}(x_{\check{r},\nu}) = f(x_{\check{r},\nu})\,, \quad \nu = 1,\ldots,\check{r},$$

wegen

$$f^{\text{lok}}_{\chi_{\check{r}},1,r}(x_{\check{r},\nu}) = \sum_{\mu=1}^{r} p^{\text{lok}}_{\chi_{\check{r}},1,\mu}\phi_{\chi_{\check{r}},1,\mu}(x_{\check{r},\nu}) = \sum_{\mu=1}^{r} p^{\text{lok}}_{\chi_{\check{r}},1,\mu}\delta_{\mu,\nu} = p^{\text{lok}}_{\chi_{\check{r}},1,\nu}$$

sofort durch

$$p^{\text{lok}}_{\chi_{\check{r}},1,\nu} = f(x_{\check{r},\nu})\,, \quad \nu = 1,\ldots,\check{r},$$

gelöst.

Da die Elemente des mit linearen Splines aufgespannten Raums in den Stützstellen nicht differenzierbar sind, empfiehlt es sich bei entsprechend höheren Differenzierbarkeitsansprüchen B-Splines höherer Ordnung zu verwenden. Zu beachten ist allerdings, dass hierbei zusätzliche Freiheitsgrade entstehen. So werden bei quadratischen B-Splines zu den in (3.19) vorgegebenen $\check{r}$ Stützstellen die folgenden $r = \check{r}+1$ Basisfunktionen

$$\phi_{\chi_{\check{r}},2,1}(x) := \tfrac{(x_{\check{r},2}-x)^2}{2(x_{\check{r},2}-x_{\check{r},1})^2} \text{ für } x \in [x_{\check{r},1}, x_{\check{r},2}]\,,$$

$$\phi_{\chi_{\check{r}},2,2}(x) := \begin{cases} \frac{(x_{\check{r},2}-x_{\check{r},1})^2+2(x_{\check{r},2}-x_{\check{r},1})(x-x_{\check{r},1})-2(x-x_{\check{r},1})^2}{2(x_{\check{r},2}-x_{\check{r},1})^2} & \text{für } x \in [x_{\check{r},1}, x_{\check{r},2}] \\ \frac{(x_{\check{r},3}-x)^2}{2(x_{\check{r},3}-x_{\check{r},2})^2} & \text{für } x \in [x_{\check{r},2}, x_{\check{r},3}] \end{cases},$$

für $\nu = 3,\ldots,r-2$

$$\phi_{\chi_{\check{r}},2,\nu}(x) := \begin{cases} \frac{(x-x_{\check{r},\nu-2})^2}{2(x_{\check{r},\nu-1}-x_{\check{r},\nu-2})^2} & \text{für } x \in [x_{\check{r},\nu-2}, x_{\check{r},\nu-1}] \\ \frac{(x_{\check{r},\nu}-x_{\check{r},\nu-1})^2+2(x_{\check{r},\nu}-x_{\check{r},\nu-1})(x-x_{\check{r},\nu-1})-2(x-x_{r,\nu-1})^2}{2(x_{\check{r},\nu}-x_{\check{r},\nu-1})^2} & \text{für } x \in [x_{\check{r},\nu-1}, x_{\check{r},\nu}] \\ \frac{(x_{\check{r},\nu+1}-x)^2}{2(x_{\check{r},\nu+1}-x_{\check{r},\nu})^2} & \text{für } x \in [x_{\check{r},\nu}, x_{\check{r},\nu+1}] \end{cases},$$

sowie

$$\phi_{\chi_{\check{r}},2,r-1}(x) := \begin{cases} \frac{(x-x_{\check{r},\check{r}-2})^2}{2(x_{\check{r},\check{r}-1}-x_{\check{r},\check{r}-2})^2} & \text{für } x \in [x_{\check{r},\check{r}-2}, x_{\check{r},\check{r}-1}] \\ \frac{(x_{\check{r},\check{r}}-x_{\check{r},\check{r}-1})^2+2(x_{\check{r},\check{r}}-x_{\check{r},\check{r}-1})(x-x_{\check{r},\check{r}-1})-2(x-x_{\check{r},\check{r}-1})^2}{2(x_{\check{r},\check{r}}-x_{\check{r},\check{r}-1})^2} & \text{für } x \in [x_{\check{r},\check{r}-1}, x_{\check{r},\check{r}}] \end{cases}$$

und

$$\phi_{\chi_{\check{r}},2,r}(x) := \frac{(x-x_{\check{r},\check{r}-1})^2}{2(x_{\check{r},\check{r}}-x_{\check{r},\check{r}-1})^2} \text{ für } x \in [x_{\check{r},\check{r}-1}, x_{\check{r},\check{r}}]$$

definiert und jede Funktion $f^{\text{lok}}_{\chi_{\check{r}},2,r} \in P^{\text{lok}}_{\chi_{\check{r}},2,r}$ mit Hilfe eines Parametervektors

$$\vec{p}^{\,\text{lok}}_{\chi_{\check{r}},2} := (p^{\text{lok}}_{\chi_{\check{r}},2,1}, \ldots, p^{\text{lok}}_{\chi_{\check{r}},2,r})^T \in \mathbb{R}^r$$

eindeutig in der Form

$$f^{\text{lok}}_{\chi_{\check{r}},2,r}(x) = \sum_{\nu=1}^{r} p^{\text{lok}}_{\chi_{\check{r}},2,\nu} \phi_{\chi_{\check{r}},2,\nu}(x)$$

darstellt. Abbildung 3.2 zeigt schematisch $\phi_{\chi_{\check{r}},2,\nu}(x)$, $\nu \in \{3, \ldots, r-2\}$.

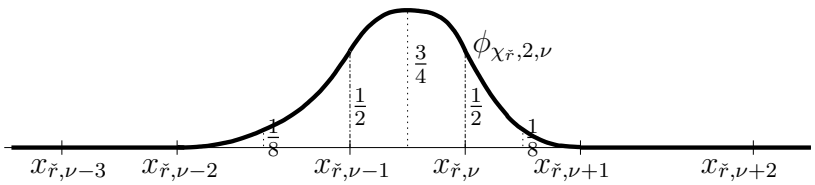

Abbildung 3.2: Quadratischer B-Spline

Die auf der Knotenmenge $\chi_{\check{r}}$ gegebene Interpolation

$$f^{\text{lok}}_{\chi_{\check{r}},2,r}(x_{\check{r},\nu}) = f(x_{\check{r},\nu}), \quad \nu = 1, \ldots, \check{r},$$

ist hingegen nicht eindeutig lösbar. Um dennoch Eindeutigkeit zu erlangen können beispielsweise die folgenden $r = \check{r}+1$ Interpolationsstützstellen

$$\begin{aligned} \xi_{r,1} &:= x_{r,1}, \\ \xi_{r,\mu} &:= \frac{1}{2}(x_{\check{r},\nu-1}+x_{\check{r},\nu}), \quad \mu = 2, \ldots, \check{r} \quad \text{und} \\ \xi_{r,r} &:= x_{\check{r},\check{r}} \end{aligned}$$

verwendet werden (vgl. Abbildung 3.3).

Abbildung 3.3: r+1 Interpolationsstützstellen

Damit gilt

$$f^{\text{lok}}_{\chi_{\check{r}},2,r}(\xi_{r,\mu}) = f(\xi_{r,\mu}), \quad \mu = 1, \ldots, r,$$

so dass entsprechend das Gleichungssystem

$$\frac{1}{8}\begin{pmatrix} 4 & 4 & & & \\ 1 & 6 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 6 & 1 \\ & & & 4 & 4 \end{pmatrix} \begin{pmatrix} p^{\text{lok}}_{2,r,1} \\ p^{\text{lok}}_{2,r,2} \\ \vdots \\ p^{\text{lok}}_{2,r,r-1} \\ p^{\text{lok}}_{2,r,r} \end{pmatrix} = \begin{pmatrix} f(\xi_{r,1}) \\ f(\xi_{r,2}) \\ \vdots \\ f(\xi_{r,r-1}) \\ f(\xi_{r,r}) \end{pmatrix}$$

gelöst werden muss.

Werden noch höhere Glattheitseigenschaften verlangt, so können auf analoge Weise auch Splines höherer Ordnung verwendet werden. Zu beachten ist hierbei jedoch stets, dass die zugehörige Interpolationsaufgabe eindeutig lösbar ist. Als Beispiel sei an dieser Stelle abschließend der Fall $k=3$ erwähnt, für den u.a. hermitsche Splines dritter Ordnung alle notwendigen Kriterien (insbesondere eine eindeutige Interpolation) erfüllen.

Wird im Sinne eines Multi-Level-Algorithmusses ein Übergang von P_{r_1} nach P_{r_2}, $r_2 > r_1$, vollzogen, so müssen in P_{r_2} geeignete Startwerte gewählt werden. Da die beiden Unterräume nicht zwingend gemeinsame Basisfunktionen besitzen (z.B. bei äquidistanter Unterteilung), ist entsprechend eine (weitere) Interpolation notwendig. Im Anhang B werden unterschiedliche (Verfeinerungs-)Strategien aufgezeigt, wie dies (auch unter Verwendung lokaler Basen) vermieden werden kann.

Ein entscheidender Nachteil der lokalen Basen ist die Tatsache, dass hochdimensionale Unterräume ausschließlich Splines mit sehr kleinen Trägern besitzen. Die Identifizierung globaler Eigenschaften der zu interpolierenden Funktion $f \in P$ könnten aufgrund der damit stark eingehenden lokalen Einflüße, wie beispielsweise vorliegende Messstörungen, erschwert werden.

3.2.1.2 Hierarchische Basen

Im Gegensatz zu den lokalen Basen, bei denen zwei unterschiedliche Unterräume nicht zwingend gemeinsame Basisfunktionen besitzen, basieren die hierarchischen Basen auf einer, durch den Skalenindex $s \in \mathbb{N}_0$ gegebenen, skalenweisen Parametrisierung. Hierbei wird beginnend mit $s = 0$ der Skalenindex sukzessive erhöht und der zugehörige Parameterraum um entsprechende Basisfunktionen zu einem nächsthöheren Ansatzraum erweitert.

Grundlage zur Berechnung der neuen Basisfunktionen ist dabei stets eine vorab zu definierende Skalierungsfunktion φ_k mit Träger $[0,1]$. In der vorliegenden Arbeit wird dies ein B-Spline der Ordnung k sein. Des Weiteren wird der zum Index $s=0$ gehörende Parameterraum $P^{\text{hier}}_{\chi_0,k,0} := P_{r_0}\left(=P^{\text{lok}}_{\chi_0,k,r_0}\right)$ durch die auf einer möglichst groben Unterteilung $\chi_0 := \left\{x_{0,\upsilon}\right\}_{\upsilon=1}^{\check{r}_0}$,

$$a = x_{0,1} < \cdots < x_{0,\check{r}_0} = b\,, \tag{3.20}$$

definierten lokalen Basis

$$\left\{\phi_{\chi_0,k,0,\nu}\right\}_{\nu=1}^{r_0}\,,$$

$$\phi_{\chi_0,k,0,\nu}(x) := \begin{cases} \varphi_k\left(\dfrac{1}{k+1}\left(k+\dfrac{x-x_{0,\upsilon}}{x_{0,\upsilon+1}-x_{0,\upsilon}}-(\nu-\upsilon)\right)\right), & \text{falls } x\in[x_{0,\upsilon},x_{0,\upsilon+1})\,, \\ \varphi_k\left(\dfrac{r_0-\nu+1}{k+1}\right) & \text{, falls } x=x_{0,\check{r}_0} \\ & \quad \wedge\, r_0-k+1\le\nu\le r_0\,, \\ 0 & \text{, sonst}\,, \end{cases} \tag{3.21}$$

der Dimension

$$r_0 = \check{r}_0+k-1$$

festgelegt.

Bemerkung 3.12
Die Knotenmenge χ_0 muss nicht zwingend aus $\check{r}_0=2$ Knotenpunkten bestehen. Um eine etwaige Gewichtung der hierarchisch erzeugten Splines zu erlangen, können auch (geringfügig) mehr nichtäquidistante Stützstellen verwendet werden.

Für die Skalen $s\ge 1$ kann unter Verwendung von

$$V_{\chi_0,k,s} := \bigoplus_{\iota=1}^{\check{r}_0-1} V_{\chi_0,k,s,\iota}$$

mit

$$V_{\chi_0,k,s,\iota} := \operatorname{span}\left\{\phi_{\chi_0,k,s,\iota,\nu}\right\}_{\nu=1}^{\tilde{r}_s}\,,\quad \iota=1,\ldots,\check{r}-1\,,\quad \tilde{r}_s := 2^{s-1}\,,$$

$$\phi_{\chi_0,k,\iota,\nu}(x) = \begin{cases} \varphi_k\left(\tilde{r}_s\dfrac{x-x_{0,\iota}}{x_{0,\iota+1}-x_{0,\iota}}-(\nu-1)\right) & \text{für } x\in(x_{0,\iota},x_{0,\iota+1}) \\ 0 & \text{sonst} \end{cases}, \quad \nu=1,\ldots,\tilde{r}_s\,, \tag{3.22}$$

der Parameterraum $P^{\text{hier}}_{\chi_0,k,s}$ sukzessiv als direkte Summe des vorrangegangenen Ansatzraums $P^{\text{hier}}_{\chi_0,k,s-1}$ und $V_{\chi_0,k,s}$ aufgespannt werden. Damit gilt

$$P^{\text{hier}}_{\chi_0,k,s} = P^{\text{hier}}_{\chi_0,k,s-1} \oplus V_{\chi_0,k,s} = P^{\text{hier}}_{\chi_0,k,0} \bigoplus_{l=1}^{s} V_{\chi_0,k,l} \ ,$$

so dass, aufgrund der Unabhängigkeit der Basisfunktionen, sich jede Funktion $f^{\text{hier}}_{\chi_0,k,s} \in P^{\text{hier}}_{\chi_0,k,s}$, $s \geq 0$, mit Hilfe eines Parametersatzes $\vec{p}^{\,\text{hier}}_{\chi_0,k,s} \in \mathbb{R}^{r_s}$,

$$r_s = r_0 + \sum_{l=1}^{s} (\check{r}_0 - 1) 2^{l-1} = r_0 + (\check{r}_0 - 1)(2^s - 1) \, ,$$

eindeutig durch

$$f^{\text{hier}}_{\chi_0,k,s}(x) =: \sum_{\nu=1}^{r_0} p^{\text{hier}}_{\chi_0,k,0,\nu} \phi_{\chi_0,k,0,\nu}(x) + \sum_{l=1}^{s} \sum_{\iota=1}^{\check{r}_0 - 1} \sum_{\nu=1}^{\check{r}_i} p^{\text{hier}}_{\chi_0,k,l,\iota,\nu} \, \phi_{\chi_0,k,l,\iota,\nu}(x) \qquad (3.23)$$

darstellen lässt.

In den Abbildungen 3.4 und 3.5 werden für $\check{r}_0 = 2$ bzw. $\check{r}_0 = 3$ jeweils unter Verwendung linearer B-Splines ($k = 1$) die Basisfunktionen für die Skalenindizes $s = 0, 1, 2$ dargestellt.

Analog zu den lokalen Basen müssen hinsichtlich einer eindeutigen Lösbarkeit der ausstehenden Interpolationsaufgabe im Fall

$$r_s > \check{r}_s = \check{r}_0 + \sum_{l=1}^{s} (\check{r}_0 - 1) 2^{l-1} = r_0 + (\check{r}_0 - 1)(2^s - 1) \quad \Leftrightarrow \quad r_0 > \check{r}_0$$

weitere Bedingungen eingeführt werden, so dass stets die Anzahl der Freiheitsgrade mit der Anzahl der Stützstellen übereinstimmt.

Im Folgenden seien die Stützstellen

$$x_{s,\upsilon} := \frac{1}{2} \big(x_{s-1,\upsilon_1} + x_{s-1,\upsilon_2} \big) \, ,$$
$$\begin{cases} \upsilon_1 = \frac{\upsilon}{2}, \ \upsilon_2 = \upsilon_1 + 1 \ , & \text{falls } \upsilon \text{ gerade} \\ \upsilon_1 = \upsilon_2 = \frac{\upsilon+1}{2} \quad , & \text{sonst} \end{cases} , \ \upsilon = 1, \ldots, \check{r}_s \, ,$$

eingeführt. Im Fall linearer B-Splines lässt sich damit, für die zum Level $s = 1$ gehörenden Koeffizienten, folgender Zusammenhang

$$p^{\text{hier}}_{\chi_0,1,1,\iota,1} = f(x_{1,2\iota}) - \frac{1}{2} \big(p^{\text{hier}}_{\chi_0,1,0,\iota} + p^{\text{hier}}_{\chi_0,1,0,\iota+1} \big) \, , \ \ \iota = 1, \ldots, \check{r}_0 - 1 \, ,$$

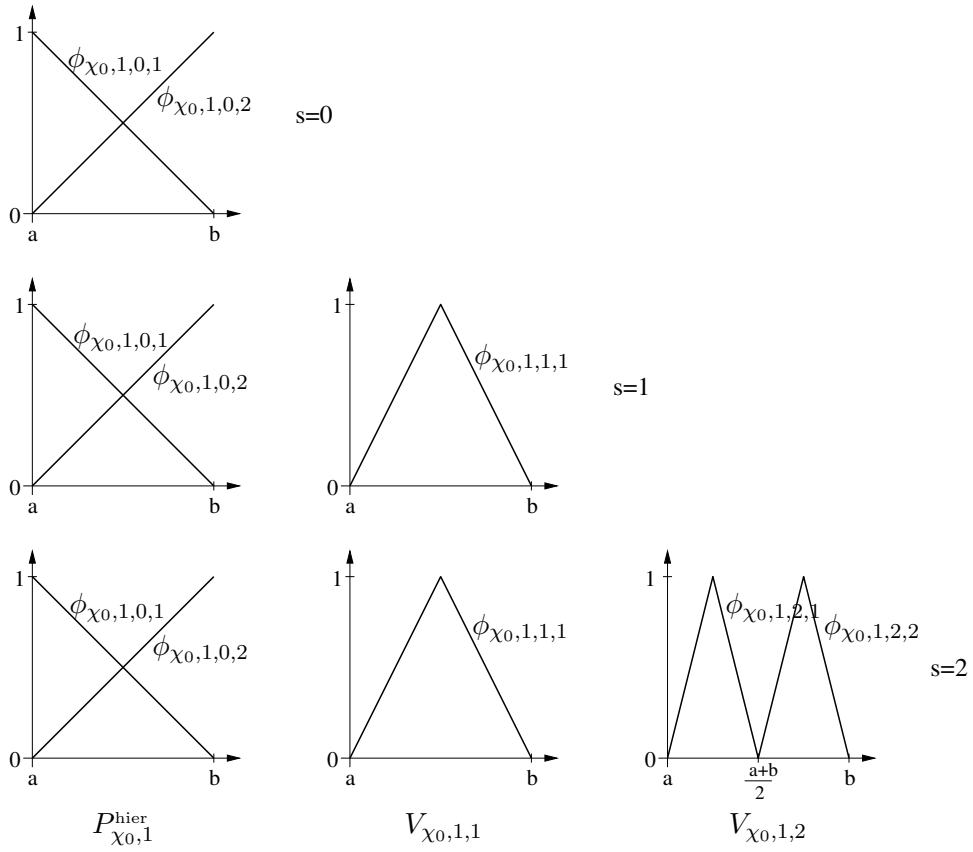

Abbildung 3.4: Lineare hierarchische Basen auf $\chi_0 = \{a, b\}$

angegeben. Vergleiche hierzu auch Abbildung 3.6. Da durch die vorangegangene Interpolation $p^{\text{hier}}_{\chi_0,1,0,\iota} = f(x_{0,\iota})$ und $p^{\text{hier}}_{\chi_0,1,0,\iota+1} = f(x_{0,\iota+1})$ gilt, beschreibt $p^{\text{hier}}_{\chi_0,1,1,\iota,1}$ gerade die an der Stelle $x_{1,2\iota}$ vorliegende Abweichung der im ersten Level generierten linearen Approximation zur gesuchten Nichtlinearität. Auf analoge Weise können rekursiv für $s \geq 2$ die entsprechenden Koeffizienten interpretiert werden. Es gilt

$$p^{\text{hier}}_{\chi_0,s,1,\iota,\nu} = f(x_{s,2^s(\iota-1)+2\nu}) - \frac{1}{2}\Big(f(x_{s-1,2^{s-1}(\iota-1)+\nu}) + f(x_{s-1,2^{s-1}(\iota-1)+\nu+1})\Big)$$

für alle $\iota = 1, \ldots, \check{r} - 1$, $\nu = 1, \ldots, \tilde{r}_s$. Unter Verwendung der durch die Träger der (symmetrischen) Basisfunktionen $\phi_{\chi_0,1,s,\iota,\nu}$, $\iota = 1, \ldots, \check{r}-1$, $\nu = 1, \ldots, \tilde{r}_s$, festgelegten Abstände

$$h_{\chi_0,s,\iota} := \frac{x_{0,\iota+1} - x_{0,\iota}}{2^s} = \frac{\big|\text{supp}\big(\phi_{\chi_0,1,s,\iota,\nu}\big)\big|}{2},$$

kann dies auch durch die in der Literatur üblicherweise Operatorschreibweise

$$p^{\text{hier}}_{\chi_0,1,s,\iota,\upsilon} = \big[-\tfrac{1}{2}\ \ 1\ \ -\tfrac{1}{2}\big]_{x_{s,2^s(\iota-1)+2\nu},h_{\chi_0,s,\iota}} f \tag{3.24}$$

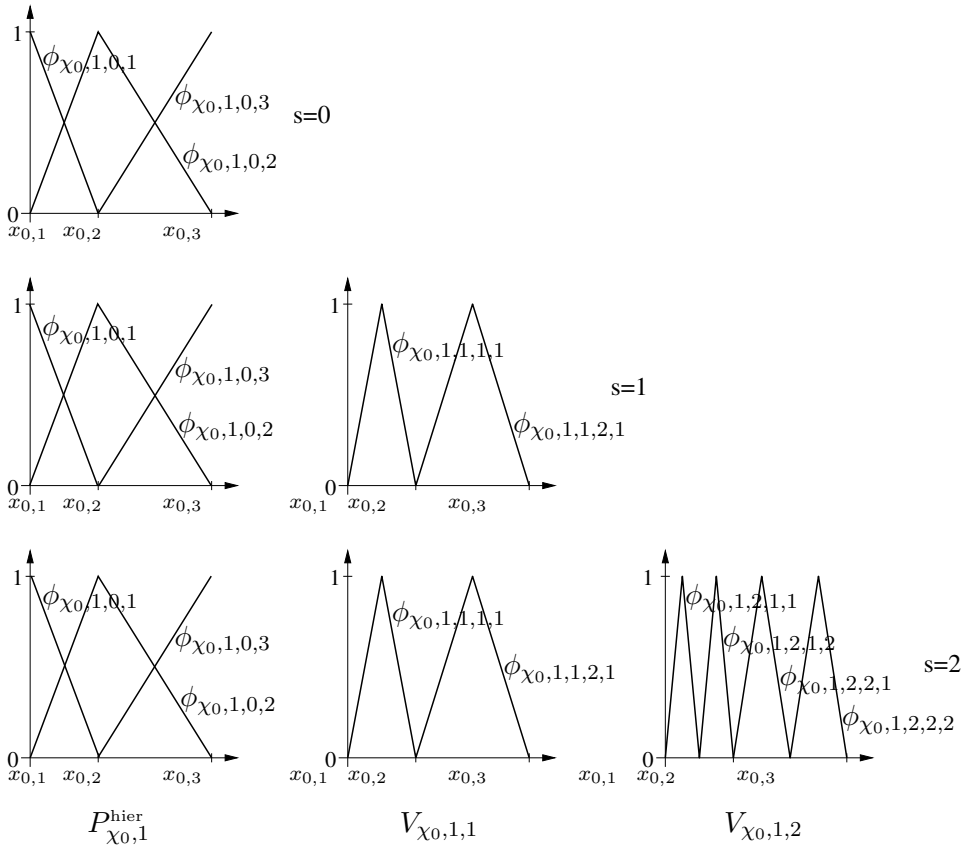

Abbildung 3.5: Lineare hierarchische Basen auf $\chi_0 = \{x_{0,1}, x_{0,2}, x_{0,3}\}$

angegeben werden. (vergleiche z.B. Bungartz [15], Kap. 2.1, oder Hackbush [29], Kap. 2.1).

Im folgenden Lemma wird eine weitere Darstellung der in (3.23) verwendeten Parameterkoeffizienten vorgestellt.

Lemma 3.13 (Integrale Darstellung der Koeffizienten)
Sei eine auf $I = [a, b]$ nach (3.20) definierte Knotenmenge $\chi_0 := \{x_{0,\upsilon}\}_{\upsilon=1}^{\check{r}_0}$ und eine Funktion $f \in P$ gegeben. Weiter seien durch die linearen Basissplines $\phi_{\chi_0,s,1,\iota,\nu}$, $\iota = 1, \ldots, \check{r}-1$, $\nu = 1, \ldots, 2^{s-1}$, der Parameterräume $P^{\text{hier}}_{\chi_0,1,s}$, $s \geq 1$, die Funktionen

$$\begin{aligned} \Psi_{\chi_0,1,s,\iota,\nu} &: \mathbb{R} \to \mathbb{R}\,, \\ x &\mapsto \frac{-h_{\chi_0,s,\iota}}{2}\, \phi_{\chi_0,1,s,\iota,\nu}(x)\,, \end{aligned} \tag{3.25}$$

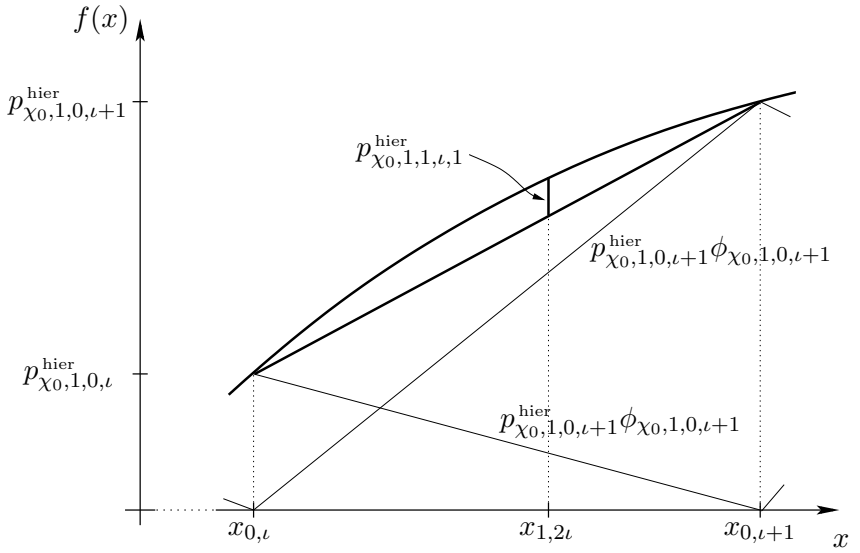

Abbildung 3.6: Interpretation der Parameterkoeffizienten $p_{\chi_0,1,1,\iota,1}$

definiert. Dann gilt für die nach (3.23) festgelegten Parameterkoeffizienten

$$p^{\text{hier}}_{\chi_0,1,s,\iota,\nu} = \int_I \Psi_{\chi_0,1,s,\iota,\nu}(x)\frac{d^2 f}{dx^2}(x)\,dx\,.$$

Beweis:
Nach Konstruktion der in (3.25) definierten Funktionen und unter Berücksichtigung von (3.24) gilt für $s \geq 1$

$$\begin{aligned}
\int_I \Psi_{\chi_0,1,s,\iota,\nu}(x)\frac{d^2 f}{dx^2}(x)\,dx &= \int_{x_{s-1,2^{s-1}(\iota-1)+\nu}}^{x_{s-1,2^{s-1}(\iota-1)+\nu+1}} \Psi_{\chi_0,1,s,\iota,\nu}(x)\frac{d^2 f}{dx^2}(x)\,dx \\
&\overset{P.I.}{=} \left[\Psi_{\chi_0,1,s,\iota,\nu}(x)\frac{df}{dx}(x)\right]_{x_{s-1,2^{s-1}(\iota-1)+\nu}}^{x_{s-1,2^{s-1}(\iota-1)+\nu+1}} \\
&\qquad - \int_{x_{s-1,2^{s-1}(\iota-1)+\nu}}^{x_{s-1,2^{s-1}(\iota-1)+\nu+1}} \frac{d\Psi_{\chi_0,1,s,\iota,\nu}}{dx}(x)\frac{df}{dx}(x)\,dx \\
&= \frac{1}{2}\int_{x_{s,2^s(\iota-1)+2\nu}-h_{\chi_0,s,\iota}}^{x_{s,2^s(\iota-1)+2\nu}} \frac{df}{dx}(x)\,dx \\
&\qquad - \frac{1}{2}\int_{x_{s,2^s(\iota-1)+2\nu}}^{x_{s,2^s(\iota-1)+2\nu}+h_{\chi_0,s,\iota}} \frac{df}{dx}(x)\,dx \\
&= \left[-\tfrac{1}{2}\ \ 1\ -\tfrac{1}{2}\right]_{x_{s,2^s(\iota-1)+2\nu},h_{\chi_s,\iota}} f = p_{\chi_0,1,s,\iota,\nu}\,.
\end{aligned}$$

□

Bemerkung 3.14
In Bungartz, Griebel [16], Kap. 3, findet sich ein vergleichbares Lemma, welches für den mehrdimensionalen Fall geführt wurde, sich jedoch auf $|\chi_0|=2$ beschränkt.

3.2.2 Allgemeine Nichtlinearitäten $\boldsymbol{f}: \mathbb{R}^d \to \mathbb{R}^d$

In diesem Abschnitt werden vektorwertige Funktionen mehrerer Veränderlicher betrachtet. Analog zum skalaren Fall werden zur Approximation in endlichen Parameterräumen mehrdimensionale lokale und hierarchische (multilineare) Basen eingeführt. Letztere werden zur Reduzierung der Variablenzahl neben einer vollen Diskretisierung auch auf sogenannten dünnen Gittern definiert.

Um einen Multi-Level-Algorithmus zur Identifizierung der gesuchten vektorwertigen Funktionen $\boldsymbol{f} \in P$ herleiten zu können, wird für jede Komponente P^j, $j=1,\ldots,d$, des (kontinuierlichen) Parameterraumes $P:=(P^1,\ldots,P^d)^T$ eine Folge endlichdimensionaler Unterräume $(P_r^j)_{r\in\mathbb{N}}$ mit

$$\dim(P_r^j) = r$$

und

$$P_{r_1}^j \subset P_{r_2}^j \quad \text{für } r_1 < r_2$$

definiert, so dass

$$\overline{\bigcup_{r\in\mathbb{N}} P_r^j} = P^j, \quad j=1,\ldots,d\,, \tag{3.26}$$

gilt. Da der Grundraum von P^j wiederum auf dem $\mathbb{R}^d$ lebt und i.A. die einzelnen Richtungen unterschiedlich stark diskretisiert sein können, kann verdeutlichend auch

$$P_{\vec{r}}^j := P_r^j\,, \quad \vec{r}=(r^1,\ldots,r^d)\in\mathbb{N}^d\,, \quad \prod_{i=1}^{d} r^i = r, \tag{3.27}$$

mit

$$P_{\vec{r}_1}^j \subset P_{\vec{r}_2}^j$$

für alle

$$\vec{r}_\eta := (r_\eta^1,\ldots,r_\eta^d)\in\mathbb{N}^d,\ \eta=1,2,\ \text{ mit } r_1^i \le r_2^i,\ i=1,\ldots,d,\ \vec{r}_1 \neq \vec{r}_2\,, \tag{3.28}$$

verwendet werden. Es sei an dieser Stelle bemerkt, dass die durch (3.27) festgelegte Wahl eines Unterraumes $P_{\vec{r}}^j = P_r^j$ keineswegs eindeutig ist, jedoch unabhängig von der gewählten Verfeinerungsstrategie wegen (3.28) stets nur ein $\vec{r}\in\mathbb{N}^d$ mit

$\dim(P_{\vec{r}}) = r$, $r \in \mathbb{N}$, exisitert.

Sind die Basen der Unterräume $P^j_{\vec{r}}$ durch $\left\{\phi_{\vec{r},\vec{\nu}}\right\}_{\vec{\nu}\in I_d(\vec{r})}$, $\phi_{\vec{r},\vec{\nu}} : \mathbb{R}^d \to \mathbb{R}$,

$$I_d(\vec{r}) := \left\{ \vec{\nu} := \begin{pmatrix} \nu^1 \\ \vdots \\ \nu^d \end{pmatrix} \in \mathbb{N}^d \;\middle|\; 1 \le \nu^i \le r^i \;\forall\, i = 1, \ldots, d \right\},$$

gegeben, dann lässt sich jede Komponente $f^j_{\vec{r}} \in P^j_{\vec{r}}$, $j = 1, \ldots, d$, einer Funktion $\boldsymbol{f}_{\vec{r}} := (f^1_{\vec{r}}, \ldots, f^d_{\vec{r}})^T \in P_{\vec{r}} := (P^1_{\vec{r}}, \ldots, P^d_{\vec{r}})^T$ durch einen eindeutig bestimmbaren vektorwertigen Parameter

$$\vec{p}^{\,j}_{\vec{r},\vec{\nu}} := \left(p^j_{\vec{r};1,\ldots,1}, \ldots, p^j_{\vec{r};r_1,\ldots,r_d}\right)^T \in \mathbb{R}^r$$

in der Form

$$f^j_{\vec{r}}(\vec{x}) = \sum_{\vec{\nu}\in I_d(\vec{r})} p^j_{\vec{r},\vec{\nu}}\, \phi_{\vec{r},\vec{\nu}}(\vec{x}) = \sum_{\nu_1=1}^{r_1} \cdots \sum_{\nu_d=1}^{r_d} p^j_{\vec{r},\nu_1,\ldots,\nu_d} \phi_{\vec{r},\nu_1,\ldots,\nu_d}(\vec{x})$$

darstellen. Auf diese Weise kann auch im mehrdimensionalen Fall jeder Unterraum $P^j_{\vec{r}}$ mit einem diskreten $\mathbb{R}^r$ identifiziert werden.

Im Folgenden sei ein beliebiger (aber fest gewählter) abgeschlossener ($\mathbb{R}^d$-)Quader

$$Q := [a^1, b^1] \times \ldots \times [a^d, b^d] \subset \mathbb{R}^d$$

als Definitionsbereich betrachtet. Des Weiteren wird angenommen, dass jede Komponente des vorliegenden Parameterraums $P := (P^1, \ldots, P^d)$ durch einen Sobolev-Raum

$$P^j := H^{m,p}(Q) = \left\{ f \in C(Q) \;\middle|\; D^\alpha f \in L^p(\mathring{Q}),\ |\alpha| \le m \right\}$$

mit Norm

$$\left\|f\right\|_{H^{m,p}(Q)} = \sum_{|\alpha|\le m} \left\|D^\alpha f\right\|_{L^p(Q)}, \quad \|.\|_{L^p(Q)} = \left(\int_Q |\,.\,|^p \,\mathrm{dx}\right)^{\frac{1}{p}},$$

gegeben ist. Um die Notationen zu vereinfachen, wird statt $\|.\|_{L^p(Q)}$ auch kurz $\|.\|_{Q,p}$ verwendet. Des Weiteren werden die auf der Maximum- und L^p-Norm basierenden Seminormen

$$|f|_{\alpha,\infty} := \|D^\alpha f\|_\infty \quad \text{und} \quad |f|_{\alpha,p} := \|D^\alpha f\|_p$$

ihre Anwendung finden.

Bemerkung 3.15
Für den Nachweis der Raum- und (Semi-)Normeigenschaften wird auf Alt [2], Kapitel 1, verwiesen.

3.2.2.1 Lokale Basen im $\mathbb{R}^d$

Im Fall lokaler Basen wird für ein $\vec{\check{r}} \in \mathbb{N}^d$, $\check{r}^i \geq 2$, $i=1,\ldots,d$, mit Hilfe einer (nicht zwingend äquidistanten) Koordinatenunterteilung

$$a^i = x_1^i < \ldots < x_{\check{r}^i}^i = b^i\,, \quad i=1,\ldots,d\,, \tag{3.29}$$

der vorgegebene Quader Q durch die Menge kantenparalleler Knoten

$$\chi_{\vec{r}} = \left\{\vec{x}_{\vec{r},\vec{v}}\right\}_{\vec{v}\in I_d(\vec{r})} := \left\{\vec{x}_{\vec{r},v^1,\ldots,v^d}\right\}_{\substack{v^i=1\\ i=1,\ldots,d}}^{\check{r}^i}\,, \tag{3.30}$$

$$\vec{x}_{\vec{r},\vec{v}} = \begin{pmatrix} x^1_{\vec{r},\vec{v}} \\ \vdots \\ x^d_{\vec{r},\vec{v}} \end{pmatrix} \in \mathbb{R}^d\,, \quad x^i_{\vec{r},\vec{v}} = x^i_{v^i}\,, \quad i=1,\ldots,d\,, \quad \forall\,\vec{v}\in I_d(\vec{r})\,,$$

lexikographisch unterteilt (vgl. Abbildung 3.7 für $\mathbb{R}^3$). Die Kantenlängen der damit entstandenen Teilquader $Q_{\chi_{\vec{r}},\vec{\iota}}$ sind entsprechend durch

$$q^i_{\chi_{\vec{r}},\vec{\iota}} := q^i_{\chi_{\vec{r}},\iota^i} = x^i_{\iota^i+1} - x^i_{\iota^i}\,, \quad i=1,\ldots,d\,, \quad \iota^i=1,\ldots,\check{r}^i-1\,,$$

gegeben.

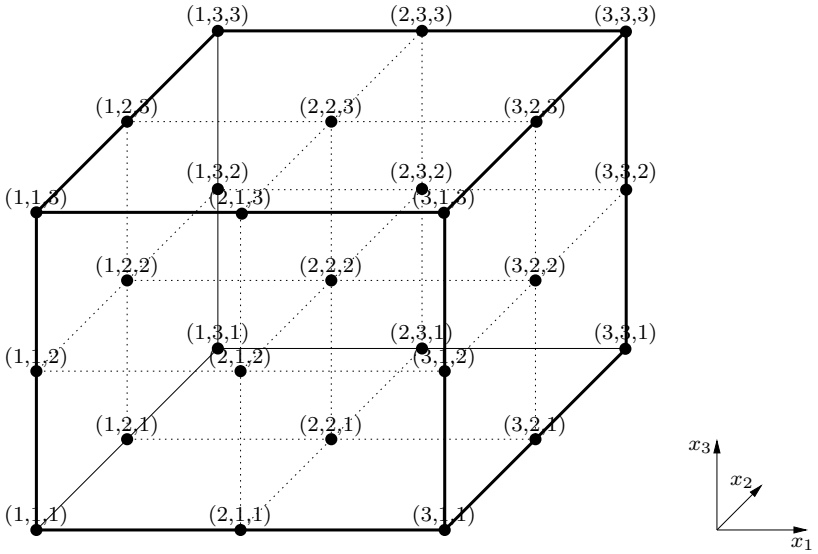

Abbildung 3.7: Beispiel einer 3D-lexikographisch sortierten Knotenmenge

Werden multidimensionale B-Spline der Ordnung k gewählt (vgl. für äquidistante Gitter ohne Randelemente z.B. Hämmerlin, Hoffmann [31]), so ist zu beachten, dass stets

$$r^i = \check{r}^i + k - 1\,, \quad i=1,\ldots,d\,, \tag{3.31}$$

erfüllt ist. Schließlich kann jeder Basisspline $\phi_{\chi_{\vec{r}},k,\vec{\nu}}$, $\vec{\nu}\in I_d(\vec{r})$, des Parameterraums $P^{\mathrm{lok},j}_{\chi_{\vec{r}},k,\vec{r}}:=P_r$ durch das Tensorprodukt

$$\phi_{\chi_{\vec{r}},k,\vec{\nu}} = \bigotimes_{i=1}^{d} \phi_{\chi_{\vec{r}^i},k,\nu^i} \tag{3.32}$$

der nach (3.21), auf der durch (3.29) festgelegten (eindimensionalen) Knotenmenge $\chi_{\vec{r}^i}$, definierten (skalaren) B-Splines beschrieben werden. Des Weiteren ist $f^{\mathrm{lok},j}_{\chi_{\vec{r}},k,\vec{r}}\in P^{\mathrm{lok},j}_{\chi_{\vec{r}},k,\vec{r}}$ wegen (3.31) mit $\vec{x}\in Q$ eindeutig durch einen Parametersatz

$$\vec{p}^{\,\mathrm{lok},j}_{\chi_{\vec{r}},k,\vec{r}} := \left(p^{\mathrm{lok},j}_{\chi_{\vec{r}},k,1,\ldots,1},\ldots,p^{\mathrm{lok},j}_{\chi_{\vec{r}},k,r^1,\ldots,r^d}\right)^T \in \mathbb{R}^r$$

in der Form

$$f^{\mathrm{lok},j}_{\chi_{\vec{r}},k,\vec{r}}(\vec{x}) = \sum_{\vec{\nu}\in I_d(\vec{r})} p^{\mathrm{lok},j}_{\chi_{\vec{r}},k,\vec{\nu}} \phi_{\chi_{\vec{r}},k,\vec{\nu}}(\vec{x})$$

darstellbar.

Analog zu den in Abschnitt 3.2.1.1 eingeführten linearen B-Splines sind im Mehrdimensionalen multilineare B-Splines die einfachste Wahl nichttrivialer Basisfunktionen. In ihrem Fall ist die Interpolationsaufgabe

$$f^{\mathrm{lok},j}_{\chi_{\vec{r}},1,\vec{r}}(\vec{x}_{\vec{r},\vec{v}}) = f^j(\vec{x}_{\chi_{\vec{r}},\vec{v}})\,,\quad \vec{v}\in I_d(\vec{r})=I_d(\vec{\tilde{r}})\,,\quad j=1,\ldots,d\,,$$

für ein $\boldsymbol{f}:=(f^1,\ldots,f^d)^T\in P$ wegen

$$f^{\mathrm{lok},j}_{\chi_{\vec{r}},1,\vec{r}}(\vec{x}_{\chi_{\vec{r}},\vec{v}}) = \sum_{\vec{\nu}\in I_d(\vec{r})} p^{\mathrm{lok},j}_{\chi_{\vec{r}},1,\vec{\nu}}\,\phi_{\chi_{\vec{r}},1,\vec{\nu}}(\vec{x}_{\chi_{\vec{r}},\vec{v}}) = \sum_{\vec{\nu}\in I_d(\vec{r})} p^{\mathrm{lok},j}_{\chi_{\vec{r}},1,\vec{\nu}}\,\delta_{\vec{\nu},\vec{v}}\,,$$

$$\delta_{\vec{\nu},\vec{v}} = \begin{cases} 1\,, & \text{falls } \vec{\nu}=\vec{v}\,,\\ 0\,, & \text{sonst}\,, \end{cases}$$

sofort durch

$$p^{\mathrm{lok},j}_{\chi_{\vec{r}},1,\vec{\nu}} = f^j(\vec{x}_{\chi_{\vec{r}},\vec{v}}),\quad \vec{\nu}\in I_d(\vec{r})\,,\quad j=1,\ldots,d\,,$$

gelöst.

Im Folgenden stellt sich die Frage, ob mögliche Monotonieeigenschaften der Nichtlinearität $f^j\in P^j$ auf $f^{\mathrm{lok},j}_{\chi_{\vec{r}},1,\vec{r}}(\vec{x}_{\chi,\vec{r}})$ übertragen werden. Dies ist im Gegensatz zum skalaren Fall nicht trivial, da die in (3.32) vorgestellten Splines für $k=1$ nur parallel der Koordinatenachsen linear sind. Entlang affiner Unterräume, welche nicht achsenparallel verlaufen, liegt ein nichtlineares Verhalten vor.

Aufgrund der einfachen Struktur lokaler Basen genügt es die Monotonie auf einem Teilquader nachzuweisen, welcher unter Berücksichtigung einer einfachen Skalierung als $[0,1]^d$ angenommen werden kann.

Lemma 3.16
Sei $f:[0,1]^d \to \mathbb{R}$, $d \in \mathbb{N}$, eine nach Definition 2.11 streng monoton wachsende Funktion und

$$p_{\vec{\iota}} := f(\vec{x}_{\vec{\iota}}) \;\; \textit{mit} \;\; \vec{x}_{\vec{\iota}} = (\iota_1, \ldots, \iota_d)^T$$

und

$$\vec{\iota} \in I(d) := \left\{ (\iota_1, \ldots, \iota_d) \in \mathbb{N}_0^d \,\middle|\, \iota_k \in \{0,1\},\, \forall k = 1, \ldots, d \right\}.$$

Dann ist die auf den Splines

$$\phi_{\vec{\iota}} : [0,1]^d \to \mathbb{R},$$
$$\vec{x} \mapsto \prod_{k=1}^{d} \begin{cases} 1-x_k\,, & \textit{falls } \iota_k = 0\,, \\ x_k\,, & \textit{sonst}\,, \end{cases}$$

$\vec{\iota} \in I(d)$, basierende Interpolierte

$$\tilde{f}(\vec{x}) = \sum_{\vec{\iota} \in I(d)} p_{\vec{\iota}}\, \phi_{\vec{\iota}}(\vec{x})$$

ebenfalls auf $[0,1]^d$ streng monoton wachsend.

Beweis:
Unabhängig von der Dimension folgt aus der Monotonie von f direkt

$$p_{\vec{\iota}} > p_{\vec{\nu}} \;\; \forall\, \vec{\iota}, \vec{\nu} \in I(d) \text{ mit } \iota_k \geq \nu_k,\; k = 1, \ldots, d, \text{ und } \vec{\iota} \neq \vec{\nu}\,.$$

Der Beweis wird im Folgenden zunächst neben dem trivialen eindimensionalen Fall nur für $d=2$ und den später relevanten Fall $d=3$ geführt. Erst im Anschluß wird die Aussage für eine beliebige Dimension bewiesen.

$d=1$:
Aufgrund der Linearität von $\tilde{f}$ und der Voraussetzung $\tilde{f}(0) = p_0 < p_1 = \tilde{f}(1)$ ist die Behauptung trivialerweise gegeben.

$d=2$:
Sei ein beliebig (aber fest) gewähltes $\vec{x} := (x_1, x_2)^T \in [0,1]^2$ gegeben. Dann gilt

$$\begin{aligned} \tilde{f}(\vec{x}) &= p_{00}(1-x_1)(1-x_2) + p_{10}x_1(1-x_2) + p_{01}(1-x_1)x_2 + p_{11}x_1x_2 \\ &= p_{00} + (p_{10}-p_{00})x_1 + (p_{01}-p_{00})x_2 + (p_{00}-p_{10}-p_{01}+p_{11})x_1x_2\,, \end{aligned}$$

und damit

$$\begin{aligned}\nabla\tilde{f}(\vec{x}) &= \begin{pmatrix} \underbrace{p_{10}-p_{00}}_{>0} + \underbrace{(p_{00}-p_{10}-p_{01}+p_{11})}_{?}\, x_2 \\ \underbrace{p_{01}-p_{00}}_{>0} + \underbrace{(p_{00}-p_{10}-p_{01}+p_{11})}_{?}\, x_1 \end{pmatrix} \\ &= \begin{pmatrix} \underbrace{(p_{10}-p_{00})}_{>0}(1-x_2) + \underbrace{(p_{11}-p_{01})}_{>0}\, x_2 \\ \underbrace{(p_{01}-p_{00})}_{>0}(1-x_1) + \underbrace{(p_{11}-p_{10})}_{>0}\, x_1 \end{pmatrix} . \end{aligned} \tag{3.33}$$

Hieraus folgt sofort $\frac{\partial\tilde{f}}{\partial x_i}(\vec{x}) > 0$, $i = 1, 2$, für alle $\vec{x} \in [0,1]^2$ und damit die Behauptung.

$d = 3$:
Sei nun $\vec{x} := (x_1, x_2, x_3)^T \in [0,1]^3$. Dann gilt

$$\begin{aligned}\tilde{f}(\vec{x}) = {} & p_{000} + (p_{100}-p_{000})x_1 + (p_{010}-p_{000})x_2 + (p_{001}-p_{000})x_3 \\ & + (p_{000}-p_{100}-p_{010}+p_{110})x_1x_2 \\ & + (p_{000}-p_{100}-p_{001}+p_{101})x_1x_3 \\ & + (p_{000}-p_{010}-p_{001}+p_{011})x_2x_3 \\ & + (-p_{000}+p_{100}+p_{010}-p_{110}+p_{001}-p_{101}-p_{011}+p_{111})x_1x_2x_3 .\end{aligned}$$

Aufgrund der Symmetrie genügt es das Vorzeichen der partiellen Ableitung nach x_3 zu untersuchen. Es gilt

$$\begin{aligned}\frac{\partial\tilde{f}}{\partial x_3}(\vec{x}) = {} & \underbrace{p_{001}-p_{000}}_{>0} + \underbrace{(p_{000}-p_{100}-p_{001}+p_{101})}_{?}\, x_1 + \underbrace{(p_{000}-p_{010}-p_{001}+p_{011})}_{?}\, x_2 \\ & + \underbrace{(-p_{000}+p_{100}+p_{010}-p_{110}+p_{001}-p_{101}-p_{011}+p_{111})}_{?}\, x_1x_2 \\ = {} & \underbrace{(p_{001}-p_{000})}_{>0}(1-x_1-x_2+x_1x_2) + \underbrace{(p_{101}-p_{100})}_{>0}\, x_1(1-x_2) \\ & + \underbrace{(p_{011}-p_{010})}_{>0}(1-x_1)x_2 + \underbrace{(p_{111}-p_{110})}_{>0}\, x_1x_2 .\end{aligned}$$

Wegen

$$1-x_1-x_2+x_1x_2 = (1-x_1)(1-x_2) \begin{cases} > 0 & \forall\, \vec{x} \in [0,1]^3 \text{ mit } x_1, x_2 \neq 1 \\ = 0 & \text{sonst} \end{cases} ,$$

folgt auch hier $\frac{\partial \tilde{f}}{\partial x_3}(\vec{x}) > 0$ für alle $\vec{x} \in [0,1]^3$.

$d \in \mathbb{N}$:
Führt man die Indexmenge

$$I(d,\vec{\iota}) := \left\{ \vec{\nu} \in I(d) \,\middle|\, \nu^i = 0 \text{ falls } \iota_i = 0,\ i = 1,\dots,d \right\}$$

ein, so lässt sich die Interpolierte durch

$$\tilde{f}(\vec{x}) = \sum_{i=0}^{d} \sum_{\substack{\vec{\iota} \in I(d) \\ |\vec{\iota}| = i}} \sum_{\vec{\nu} \in I(d,\vec{\iota})} (-1)^{|\vec{\iota}|+|\vec{\nu}|} p_{\vec{\nu}} \prod_{\substack{j \in \{1,\dots,d\} \\ \text{mit } \iota_j = 1}} x_j$$

angeben. Aufgrund der Symmetrie genügt es auch hier nur das Vorzeichen der partiellen Ableitung

$$\frac{\partial \tilde{f}}{\partial x_d}(\vec{x}) = \sum_{i=0}^{d-1} \sum_{\substack{\vec{\iota} \in I(d-1) \\ |\vec{\iota}| = i}} \sum_{\vec{\nu} \in I(d,(\vec{\iota},1))} (-1)^{|\vec{\iota}|+|\vec{\nu}|} p_{\vec{\nu}} \prod_{\substack{j \in \{1,\dots,d-1\} \\ \text{mit } \iota_j = 1}} x_j$$

zu untersuchen. Analog zu den vorangegangenen Überlegungen folgt damit bereits

$$\frac{\partial \tilde{f}}{\partial x_d}(\vec{x}) = \sum_{\vec{\nu} \in I(d-1)} (p_{(\vec{\nu},1)} - p_{(\vec{\nu},0)}) \left(\sum_{\substack{\vec{\iota} \in I(d-1) \\ \iota_j = 1 \text{ falls } \nu_j = 1}} (-1)^{|\vec{\iota}|+|\vec{\nu}|} \prod_{\substack{j \in \{1,\dots,d-1\} \\ \text{mit } \iota_j = 1}} x_j \right).$$

Bleibt das Vorzeichen von

$$\sum_{\substack{\vec{\iota} \in I(d-1) \\ \iota_j = 1 \text{ falls } \nu_j = 1}} (-1)^{|\vec{\iota}|+|\vec{\nu}|} \prod_{\substack{j \in \{1,\dots,d-1\} \\ \text{mit } \iota_j = 1}} x_j$$

für beliebige $\vec{\nu} \in I(d-1)$, $d \in \mathbb{N}$, induktiv zu untersuchen. Für $d = 2$ gilt

$$\sum_{\substack{\vec{\iota} \in I(d-1) \\ \iota_j = 1 \text{ falls } \nu_j = 1}} (-1)^{|\vec{\iota}|+|\vec{\nu}|} \prod_{\substack{j \in \{1,\dots,d-1\} \\ \text{mit } \iota_j = 1}} x_j = \begin{cases} 1 - x_1 & \text{falls } \nu = 0 \\ x_1 & \text{sonst} \end{cases}$$

und damit analog zu (3.33) das positive Vorzeichen der partiellen Ableitung.

Betrachte nun den Induktionsschritt $d \to d+1$. Da für $\nu_d = 0$

$$\begin{aligned}
&\sum_{\substack{\vec{\iota}\in I(d)\\ \iota_j=1 \text{ falls } \nu_j=1}} (-1)^{|\vec{\iota}|+|\vec{\nu}|} \prod_{\substack{j\in\{1,\ldots,d\}\\ \text{mit } \iota_j=1}} x_j \\
&\qquad = \sum_{\substack{\vec{\iota}\in I(d-1)\\ \iota_j=1 \text{ falls } \nu_j=1}} (-1)^{|\vec{\iota}|+|(\nu_1,\ldots,\nu_{d-1})|} \prod_{\substack{j\in\{1,\ldots,d-1\}\\ \text{mit } \iota_j=1}} x_j \\
&\qquad\quad + x_d \sum_{\substack{\vec{\iota}\in I(d-1)\\ \iota_j=1 \text{ falls } \nu_j=1}} (-1)^{|\vec{\iota}|+|(\nu_1,\ldots,\nu_{d-1})|+1} \prod_{\substack{j\in\{1,\ldots,d-1\}\\ \text{mit } \iota_j=1}} x_j \\
&\qquad = (1-x_d) \sum_{\substack{\vec{\iota}\in I(d-1)\\ \iota_j=1 \text{ falls } \nu_j=1}} (-1)^{|\vec{\iota}|+|(\nu_1,\ldots,\nu_{d-1})|} \prod_{\substack{j\in\{1,\ldots,d-1\}\\ \text{mit } \iota_j=1}} x_j
\end{aligned}$$

und für $\nu_d = 1$

$$\sum_{\substack{\vec{\iota}\in I(d)\\ \iota_j=1 \text{ falls } \nu_j=1}} (-1)^{|\vec{\iota}|+|\vec{\nu}|} \prod_{\substack{j\in\{1,\ldots,d\}\\ \iota_j=1}} x_j = x_d \sum_{\substack{\vec{\iota}\in I(d-1)\\ \iota_j=1 \text{ falls } \nu_j=1}} (-1)^{|\vec{\iota}|+|(\nu_1,\ldots,\nu_{d-1})|} \prod_{\substack{j\in\{1,\ldots,d-1\}\\ \iota_j=1}} x_j$$

gilt, folgt die Behauptung.

□

Nun wird noch untersucht, wieviele Ungleichungsnebenbedingungen auf dem gesamten Definitionsbereich zu erfüllen sind. Während dies für $d=1$ trivialerweise $\check{r}-1$ Bedingungen sind, bedarf es für $d \geq 2$ eines kleinen Hilfssatzes.

Lemma 3.17
Sei ein Quader

$$Q = [a^1, b^1] \times \ldots \times [a^d, b^d] \subset \mathbb{R}^d,\ d \in \mathbb{N},$$

für $\check{r} \in \mathbb{N}^d$, $\check{r}^i \geq 2$, $i = 1, \ldots, d$, mit einer nach (3.30) definierten Knotenmenge $\chi_{\vec{\check{r}}} = \left\{\vec{x}_{\check{r},\vec{v}}\right\}_{\vec{v}\in I_d(\vec{\check{r}})}$ unterteilt. Sei des Weiteren an jedem Knotenpunkt $\vec{x}_{\check{r},\vec{v}}$, $\vec{v} \in I_d(\vec{\check{r}})$, der Wert $p_{\vec{v}} \in \mathbb{R}$ vorgegeben. Dann erfordert die Überprüfung von $p_{\vec{v}_1} > p_{\vec{v}_2}$ für alle

$$\vec{v}_1, \vec{v}_2 \in I(d),\ v_1^k \geq v_2^k,\ k = 1, \ldots, d,\ \vec{v}_1 \neq \vec{v}_2, \tag{3.34}$$

die Auswertung von

$$M(d) = d \prod_{i=1}^{d} \check{r}^i - \sum_{j=1}^{d} \prod_{\substack{i=1\\ i\neq j}}^{d} \check{r}^i$$

Ungleichungen.

Beweis:
Zwei Knotenpunkte $\vec{x}_{\chi_{\vec{r}},\vec{v}_1}$ und $\vec{x}_{\chi_{\vec{r}},\vec{v}_2}$ mit $v_1^j = v_2^j - 1$ für ein $j \in \{1,\ldots,d\}$ und $v_1^i = v_2^i \;\forall\, i \in \{1,\ldots,d\}\backslash j$ werden als **direkte Nachbarn** bezeichnet. Da jeweils zwei direkte Nachbarn durch eine Kante eines Teilquaders verbunden sind, bedarf die Überprüfung von $p_{\vec{v}_1} > p_{\vec{v}_2}$ für alle direkten Nachbarn genau soviele Ungleichungen wie der unterteilte Quader (Teil-)Kanten besitzt. Ist $p_{\vec{v}_1} > p_{\vec{v}_2}$ für alle direkten Nachbarn erfüllt, so folgt für zwei beliebig nach (3.34) gewählte Knotenpunkte $\vec{x}_{\chi_{\vec{r}},\vec{v}_a}$ und $\vec{x}_{\chi_{\vec{r}},\vec{v}_b}$ wegen der Existenz endlich vieler direkter Nachbarn $x_{\check{r},\vec{v}_i}$, $i=1,\ldots,l$, mit

$$\vec{v_a} \leq \vec{v_1} \leq \ldots \leq \vec{v_l} \leq \vec{v_b}$$

und $\vec{x}_{\check{r},\vec{v}_a}, \vec{x}_{\check{r},\vec{v}_1}$ sowie $\vec{x}_{\check{r},\vec{v}_l}, \vec{x}_{\check{r},\vec{v}_b}$ ebenfalls direkte Nachbarn, bereits die Behauptung $p_{\vec{v}_a} > p_{\vec{v}_b}$. Bleibt also die Anzahl der (Teil-)Kanten zu berechnen. Der Beweis kann mit vollständiger Induktion geführt werden. Für $d=1$ ist die Aussage trivialerweise durch $M(1)=\check{r}-1$ gegeben. Für einen Quader im $\mathbb{R}^2$ (also ein Rechteck) gilt

$$M(2) = (\check{r}^{\,1}-1)\check{r}^{\,2} + (\check{r}^{\,2}-1)\check{r}^{\,1} = 2\check{r}^{\,1}\check{r}^{\,2} - \check{r}^{\,1} - \check{r}^{\,2}\,.$$

Bleibt die Betrachtung eines Quaders im $\mathbb{R}^{d+1}$. Unter Verwendung der Induktionsvoraussetzung folgt

$$\begin{aligned} M(d+1) &= \check{r}_{d+1} M(d) + (\check{r}_{d+1}-1)\prod_{i=1}^{d}\check{r}^{\,i} \\ &\overset{IV}{=} d\prod_{i=1}^{d+1}\check{r}^{\,i} - \sum_{j=1}^{d}\prod_{\substack{i=1\\ i\neq j}}^{d+1}\check{r}^{\,i} + (\check{r}_{d+1}-1)\prod_{i=1}^{d}\check{r}^{\,i} = (d+1)\prod_{i=1}^{d+1}\check{r}^{\,i} - \sum_{j=1}^{d+1}\prod_{\substack{i=1\\ i\neq j}}^{d+1}\check{r}^{\,i} \end{aligned}$$

und damit die Behauptung.

□

Bemerkung 3.18
Ist ein Quader mit $n := \check{r}^{\,1} = \ldots = \check{r}^{\,d}$ *gegeben, so folgt für* $d \geq 1$ *sofort*

$$M(d) = d\big(n^d - n^{d-1}\big)\,.$$

Tabelle 3.1 verdeutlicht die Entwicklung von $M(d)$ für den in Bemerkung 3.18 vorgestellten Fall.

Sind (in den Stützstellen) höhere Differenzierungseigenschaften notwendig, können, analog zum Eindimensionalen, B-Splines höherer Ordnung verwendet werden. Da sich die im nachfolgenden Kapitel untersuchten Fallstudien zur Identifizierung mehrdimensionaler Nichtlinearitäten jedoch auf trilineare B-Splines

	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$	$n=7$	$n=8$	$n=9$	$n=10$	$n=15$
$d=1$	1	2	3	4	5	6	7	8	9	14
$d=2$	4	12	24	40	60	85	112	144	180	420
$d=3$	12	54	144	300	540	882	1344	1944	2700	9450

Tabelle 3.1: Anzahl der Ungleichungsnebenbedingungen

beschränken, wird an dieser Stelle nicht weiter auf höhergradige Splines eingegangen.

3.2.2.2 Hierarchische Basen im $\mathbb{R}^d$

Auch im Mehrdimensionalen basieren hierarchische Basen auf einer, durch einen Skalenindex $s \in \mathbb{N}_0$ festgelegten, skalenweisen Parametrisierung. Da jedoch unterschiedliche Verfeinerungsstrategien möglich sind und auch die Dimensionalität des zugehörigen (Identifizierungs-)Problems (bei voller Diskretisierung) mit jedem Skalenschritt stark anwächst, werden die einzelnen Skalen weiter untergliedert. Hierzu wird wegen $Q \in \mathbb{R}^d$ die folgende, vom Skalenindex s abhängige, vektorwertige Indexmenge $\Upsilon_d^{\text{Strat}}(s)$ mit

$$\Upsilon_d^{\text{Strat}}(0) = \Upsilon_d(0) := \left\{\vec{0}\right\}$$

und

$$\Upsilon_d^{\text{Strat}}(s) \subseteq \Upsilon_d(s) := \left\{ \vec{\sigma} \in \mathbb{N}_0^d \,\middle|\, \sigma^i \leq s \;\forall\, i = 1, \ldots, d \right\},$$

$$\Upsilon_d^{\text{Strat}}(s_1) \cap \Upsilon_d^{\text{Strat}}(s_2) = \{\} \quad \forall\, s_1, s_2 \in \mathbb{N}_0,\; s_1 \neq s_2\,,$$

eingeführt. Die Unterteilung des Quaders Q erfolgt durch eine möglichst grobe, nach (3.30) festgelegte, kantenparallele Knotenmenge $\chi_0 := \chi_{\vec{r}_0} = \left\{\vec{x}_{\vec{r}_0,\vec{v}}\right\}_{\vec{v} \in I_d(\vec{r}_0)}$, $\vec{r} \in \mathbb{N}^d$, $\check{r}_0^{\,i} \geq 2$, $i = 1, \ldots, d$. Grundlage zur Berechnung der Basisfunktionen ist wieder eine vorab zu definierende Skalierungsfunktion φ_k^d mit Träger $[0,1]^d$, welche in dieser Arbeit als multidimensionaler B-Spline vorausgesetzt wird.

Der (unabhängig von der durch $\Upsilon_d^{\text{Strat}}(s)$ festgelegten Skalierungsstrategie) zu $s=0$ gehörende Parameterraum

$$P_{\chi_0,k,0}^{\text{hier},j} := P_{|\vec{r}_0|}^j \;\left(= P_{\chi_0,k,\vec{r}_0}^{\text{lok},j}\right)$$

der Dimension $|\vec{r}_0|$, $\vec{r}_0 \in \mathbb{N}^d$,

$$r_0^i = \check{r}_0^{\,i} + k - 1\,, \quad i = 1, \ldots, d\,,$$

ist durch die auf χ_0 definierte lokale Basis $\left\{\phi_{\chi_0,k,\vec{0},\vec{\nu}}\right\}_{\vec{\nu}\in I_d(\vec{r}_0)}$ aufgespannt. Die zugehörigen Basisfunktionen können entsprechend als Tensorprodukt

$$\phi_{\chi_0,k,\vec{0},\vec{\nu}} = \bigotimes_{i=1}^{d} \phi_{\chi_0^i,k,0,\nu^i}\,, \quad \vec{\nu}\in I_d(\vec{r}_0)\,,$$

der nach (3.21), auf

$$\chi_0^i = \left\{x_{0,\upsilon^i}^i\right\}_{\upsilon^i=1}^{\check{r}_0^i}\,, \quad x_{0,\upsilon^i}^i = a^i + \sum_{\iota^i=2}^{\upsilon^i} q_{\chi_0,\iota^i-1}^i\,, \tag{3.35}$$

definierten skalaren Basisfunktion beschrieben werden.

Für $s\geq 1$ wird jedem

$$\vec{\sigma}\in\Upsilon_{\mathrm{hier},d}^{\mathrm{Strat}}(s) := \left\{\vec{\sigma}\in\Upsilon_d^{\mathrm{Strat}}(s)\,\middle|\,\sigma^i\geq 1,\ i=1,\ldots,d\right\}$$

unter Verwendung von

$$\vec{\tilde{r}}_{\vec{\sigma}}\in\mathbb{N}^d\,, \quad \tilde{r}_{\vec{\sigma}}^{\,i} := 2^{\sigma^i-1}\,, \quad i=1,\ldots,d\,,$$

der Raum

$$W_{\chi_0,k,\vec{\sigma}}^{\mathrm{hier},j} := \bigoplus_{\vec{\iota}\in I_d(\vec{\tilde{r}}_0-\vec{1})} W_{\chi_0,k,\vec{\sigma},\vec{\iota}}^{\mathrm{hier},j}$$

mit

$$W_{\chi_0,k,\vec{\sigma},\vec{\iota}}^{\mathrm{hier},j} := \mathrm{span}\left\{\phi_{\chi_0,k,\vec{\sigma},\vec{\iota},\vec{\nu}}\right\}_{\vec{\nu}\in I_d(\vec{\tilde{r}}_{\vec{\sigma}})}\,, \tag{3.36}$$

und jedem

$$\vec{\sigma}\in\Upsilon_{\mathrm{twg},d}^{\mathrm{Strat}}(s) := \Upsilon_d^{\mathrm{Strat}}(s)\setminus\Upsilon_{\mathrm{hier},d}^{\mathrm{Strat}}(s)$$

unter Verwendung von

$$I_d^{\mathrm{twg}}(\vec{\sigma}) := \left\{\vec{\nu}\in\mathbb{N}^d\,\middle|\,\nu^i\leq(\check{r}_0^{\,i}-1)2^{\sigma^i-1}\text{ falls }\sigma^i\geq 1\text{ und }\nu^i\leq\check{r}_0^{\,i}+k-1\text{ sonst}\right\}$$

der Raum

$$W_{\chi_0,k,\vec{\sigma}}^{\mathrm{twg},j} = \left\{\phi_{\chi_0,k,\vec{\sigma},\vec{\nu}}\right\}_{\vec{\nu}\in I_d^{\mathrm{twg}}(\vec{\sigma})}\,, \tag{3.37}$$

zugeordnet. Die Unterscheidung der in (3.36) und (3.37) definierten Räume ist (auch in Bezug auf die ausstehende Implementierung) sinnvoll, da Basisfunktionen des $W_{\chi_0,k,\vec{\sigma},\vec{\iota}}^{\mathrm{twg},j}$, im Gegensatz zu denen des $W_{\chi_0,k,\vec{\sigma},\vec{\iota}}^{\mathrm{hier},j}$, in den Koordinatenrichtungen mit $\sigma^i=0$ globale, teilquaderübergreifende Trägerkompenten besitzen und

damit eingeschränkt auf $Q_{\chi_0,\vec{\iota}}$ entsprechende Randsplines darstellen. Die in (3.36) verwendeten Basisfunktionen können schließlich direkt als Tensorprodukt

$$\phi_{\chi_0,k,\vec{\sigma},\vec{\iota},\vec{\nu}} = \bigotimes_{i=1}^{d} \phi_{\chi_0^i,k,\sigma^i,\iota^i,\nu^i} \,, \quad \vec{\iota} \in I_d(\vec{\check{r}}_0 - \vec{1}) \,, \ \vec{\nu} \in I_d(\vec{\tilde{r}}_{\vec{\sigma}}) \,,$$

der auf χ_0^i nach (3.22) definierten skalarwertigen Basisfunktionen angegeben werden. Für die durch (3.37) definierten Splines

$$\phi_{\chi_0,k,\vec{\sigma},\vec{\nu}} = \bigotimes_{i=1}^{d} \phi_{\chi_0^i,k,\sigma^i,\nu^i} \,, \quad \vec{\nu} \in I_d^{\mathrm{twg}}(\vec{\sigma}) \,,$$

gilt hingegen (nur) für die Komponenten mit $\sigma^i \geq 1$

$$\phi_{\chi_0^i,k,\sigma^i,\nu^i}(x) = \begin{cases} \varphi_k\left(2^{\sigma^i-1}\dfrac{x^i - x^i_{0,\upsilon^i}}{x^i_{0,\upsilon^i+1} - x^i_{0,\upsilon^i}} - (\nu^i - 1)\right) & \text{für } x \in (x_{0,\upsilon^i}, x_{0,\upsilon^i+1}) \\ 0 & \text{sonst} \end{cases} ,$$

$\nu^i = 1, \ldots, 2^{\sigma^i-1}$, und sonst (in mindestens einer Koordinatenrichtung)

$$\phi_{\chi_0^i,k,\sigma^i,\nu^i}(x) := \begin{cases} \varphi_k\left(\dfrac{1}{k+1}\left(k + \dfrac{x^i - x^i_{0,\upsilon^i}}{x^i_{0,\upsilon^i+1} - x^i_{0,\upsilon^i}} - (\nu^i - \upsilon^i)\right)\right), \text{falls } x \in [x_{0,\upsilon^i}, x_{0,\upsilon^i+1}) \,, \\ \varphi_k\left(\dfrac{1 - \nu^i + \check{r}_0^i}{k+1}\right) \quad \text{, falls } x = x_{0,\check{r}_0^i} \\ \qquad\qquad\qquad\qquad \wedge\, r_0^i - k + 1 \leq \nu^i \leq r_0^i \,, \\ 0 \qquad\qquad\qquad\qquad \text{, sonst} \,. \end{cases}$$

Zusammenfassend kann der zu einem Skalenindex $s \geq 1$ festgelegte Parameterraum $P^{\mathrm{hier},j}_{\chi_0,k,s}$ als direkte Summe des vorangegangenen Ansatzraums $P^{\mathrm{hier},j}_{\chi_0,k,s-1}$ und den Räumen $W^{\mathrm{hier},j}_{\chi_0,k,\vec{\sigma}}$ und $W^{\mathrm{twg},j}_{\chi_0,k,\vec{\sigma}}$, $\vec{\sigma} \in \Upsilon_d^{\mathrm{Strat}}(s)$, aufgespannt werden. Es gilt

$$P^{\mathrm{hier},j}_{\chi_0,k,s} = P^{\mathrm{hier},j}_{\chi_0,k,s-1} \bigoplus_{\vec{\sigma} \in \Upsilon^{\mathrm{Strat}}_{\mathrm{hier},d}(s)} W^{\mathrm{hier},j}_{\chi_0,k,\vec{\sigma}} \bigoplus_{\vec{\sigma} \in \Upsilon^{\mathrm{Strat}}_{\mathrm{twg},d}(s)} W^{\mathrm{twg},j}_{\chi_0,k,\vec{\sigma}} \,.$$

Wird die Unterscheidung in (3.36) und (3.37) nicht explizit gewünscht, so kann unter Verwendung von

$$V^j_{\chi_0,k,\vec{\sigma}} := \begin{cases} W^{\mathrm{hier},j}_{\chi_0,k,\vec{\sigma}} \,, \text{ falls } \vec{\sigma} \in \Upsilon^{\mathrm{Strat}}_{\mathrm{hier},d}(s) \,, \\ W^{\mathrm{twg},j}_{\chi_0,k,\vec{\sigma}} \,, \text{ sonst} \,, \end{cases}$$

auch

$$P^{\mathrm{hier},j}_{\chi_0,k,s} = P^{\mathrm{hier},j}_{\chi_0,k,s-1} \bigoplus_{\vec{\sigma} \in \Upsilon^{\mathrm{Strat}}_{d}(s)} V^j_{\chi_0,k,\vec{\sigma}} = P^{\mathrm{hier},j}_{\chi_0,k,0} \bigoplus_{l=1}^{s} \bigoplus_{\vec{\sigma} \in \Upsilon^{\mathrm{Strat}}_{d}(l)} V^j_{\chi_0,k,\vec{\sigma}}$$

und mit

$$V^j_{\chi_0,k,s} := \bigoplus_{\vec{\sigma}\in\Upsilon^{\mathrm{Strat}}_d(s)} V^j_{\chi_0,k,\vec{\sigma}}$$

auch direkt

$$P^{\mathrm{hier},j}_{\chi_0,k,s} = P^{\mathrm{hier},j}_{\chi_0,k,s-1} \oplus V^j_{\chi_0,k,s} = P^{\mathrm{hier},j}_{\chi_0,k,0} \bigoplus_{l=1}^{s} V^j_{\chi_0,k,l}$$

angegeben werden. Mit

$$\dim\left(W^{\mathrm{hier},j}_{\chi_0,k,\vec{\sigma}}\right) = \prod_{i=1}^{d} (\check{r}^i_0 - 1) 2^{\sigma^i - 1} =: d^{\mathrm{hier}}_{\vec{\sigma}}$$

und

$$\dim\left(W^{\mathrm{twg},j}_{\chi_0,k,\vec{\sigma}}\right) = \prod_{i=1}^{d} \begin{cases} (\check{r}^i_0 - 1) 2^{\sigma^i - 1}, & \text{falls } \sigma^i \geq 1 \\ \check{r}^i_0, & \text{sonst} \end{cases} =: d^{\mathrm{twg}}_{\vec{\sigma}}$$

kann schließlich aufgrund der Unabhängigkeit aller Basisfunktionen jede Funktion $f^{\mathrm{hier},j}_{\chi_0,k,s} \in P^{\mathrm{hier},j}_{\chi_0,k,s}$ mit Hilfe eines Parametersatzes $\vec{p}^{\,\mathrm{hier},j}_{\chi_0,k,s} \in \mathbb{R}^{r_s}$ der Dimension

$$\begin{aligned} r_s &= |\vec{r}_0| + \sum_{l=1}^{s} \sum_{\vec{\sigma}\in\Upsilon^{\mathrm{Strat}}_d(l)} \dim\left(V^j_{\chi_0,k,\vec{\sigma}}\right) \\ &= |\vec{r}_0| + \sum_{l=1}^{s} \left(\sum_{\vec{\sigma}\in\Upsilon^{\mathrm{Strat}}_{\mathrm{hier},d}(l)} d^{\mathrm{hier}}_{\vec{\sigma}} + \sum_{\vec{\sigma}\in\Upsilon^{\mathrm{Strat}}_{\mathrm{twg},d}(l)} d^{\mathrm{twg}}_{\vec{\sigma}} \right) \end{aligned}$$

eindeutig durch

$$\begin{aligned} f^{\mathrm{hier},j}_{\chi_0,k,s}(x) =: & \sum_{\vec{\nu}\in I_d(\vec{r}-1)} p^{\mathrm{hier},j}_{\chi_0,k,\vec{0},\vec{\nu}}\, \phi_{\chi_0,k,\vec{0},\vec{\nu}}(x) \\ & + \sum_{l=1}^{s} \left(\sum_{\vec{\sigma}\in\Upsilon^{\mathrm{Strat}}_{\mathrm{hier},d}(l)} \sum_{\vec{\imath}\in I_d(\vec{r}_0-\vec{1})} \sum_{\vec{\nu}\in I_d(\vec{\tilde{r}}_{\vec{\sigma}})} p^{\mathrm{hier},j}_{\chi_0,k,\vec{\sigma},\vec{\imath},\vec{\nu}}\, \phi_{\chi_0,k,\vec{\sigma},\vec{\imath},\vec{\nu}}(x) \right. \\ & \qquad \left. + \sum_{\vec{\sigma}\in\Upsilon^{\mathrm{Strat}}_{\mathrm{twg},d}(l)} \sum_{\vec{\nu}\in I^{\mathrm{twg}}_d(\vec{\sigma})} p^{\mathrm{hier},j}_{\chi_0,k,\vec{\sigma},\vec{\nu}}\, \phi_{\chi_0,k,\vec{\sigma},\vec{\nu}}(x) \right) \end{aligned}$$

dargestellt werden.

3.2.2.2.1 Volle Diskretisierung

Bei der sogenannten vollen Diskretisierung wird für $s \geq 1$ sie Skalierungsstrategie

$$\Upsilon_d^{\text{Strat}}(s) = \Upsilon_d^{\text{voll}}(s) := \Upsilon_d(s) \backslash \Upsilon_d(s-1) \tag{3.38}$$

verwendet. Vergleiche hierzu die Abbildungen 3.8 und 3.9, in denen exemplarisch für $Q = [a,b]^2$ unter Verwendung der Minimalunterteilung $|\chi_0| = 4$ bzw. für eine nicht äquidistante Diskretisierung mit $|\chi_0| = 9$ die Träger der Basisfunktionen sowie die zugehörigen Stützstellen aufgezeigt werden. Dabei sind die einzelnen durch $\vec{\sigma} \in \Upsilon_d^{\text{voll}}(s)$, $s \leq 4$ bzw. $s \leq 3$, festgelegten Unterräume $V^j_{\chi_0,k,\vec{\sigma}}$ separat dargestellt und die zur gleichen Skalierungsebene s gehörenden Funktionsräume, zur besseren Übersicht, mit gleicher Intensität schattiert. Abbildung 3.10 motiviert schließlich noch die verwendete (Index-)Notation durch Angabe der in $V^j_{\chi_0,k,(1,2)}$ definierten Basisfunktionen.

Eine wesentliche Frage für die Interpolation einer (kontinuierlichen) vektorwertigen Nichtlinearität $\boldsymbol{f} := (f^1, \ldots, f^d)^T \in P$ ist der Einfluss der gewählten Skalierungsstrategie $\Upsilon_d^{\text{Strat}}(s)$ auf die zu ermittelnde Interpolierte und der damit anfallende Speicheraufwand.

Bemerkung 3.19

Im Folgenden werden sich alle Untersuchungen (entsprechend der durchgeführten Implementierung) auf multilineare Basisfunktionen ($k=1$) beschränken.

Im Fall der vollen Diskretisierung (3.38) ist die Frage nach dem Speicheraufwand für eine Komponente f^j sofort durch

$$\begin{aligned}
\left|P^{\text{voll},j}_{\chi_0,1,s}\right| &= \sum_{l=1}^{s} \sum_{\vec{\sigma} \in \Upsilon_d^{\text{voll}}(l)} \dim \left|V^j_{\chi_0,1,\vec{\sigma}}\right| = \sum_{\vec{\sigma} \in \Upsilon_d(s)} \dim \left|V^j_{\chi_0,1,\vec{\sigma}}\right| \\
&= \prod_{i=1}^{d} \left[\check{r}_0^{\,i} + (\check{r}_0^{\,i} - 1) \sum_{l=1}^{s} 2^{l-1} \right] = \prod_{i=1}^{d} \left[\check{r}_0^{\,i} + (\check{r}_0^{\,i} - 1)(2^s - 1) \right] \\
&= \prod_{i=1}^{d} \left[1 + 2^s (\check{r}_0^{\,i} - 1) \right] = \mathcal{O}\left(r_0 \cdot 2^{ds} \right)
\end{aligned} \tag{3.39}$$

beantwortet. Vergleiche hierzu auch die Abbildungen 3.11 und 3.12, welche die Stützstellen aller skalenweise definierter Unterräume $V^j_{\chi_0,1,\vec{\sigma}}$, $\vec{\sigma} \in \Upsilon_d^{\text{voll}}(s)$, $s \leq 4$ bzw. $s \leq 3$, für die bereits in den vorangegangenen Abbildungen vorgestellten $\mathbb{R}^2$-Gitter mit $|\chi_0| = 4$ bzw. $|\chi_0| = 9$ graphisch ausgeben.

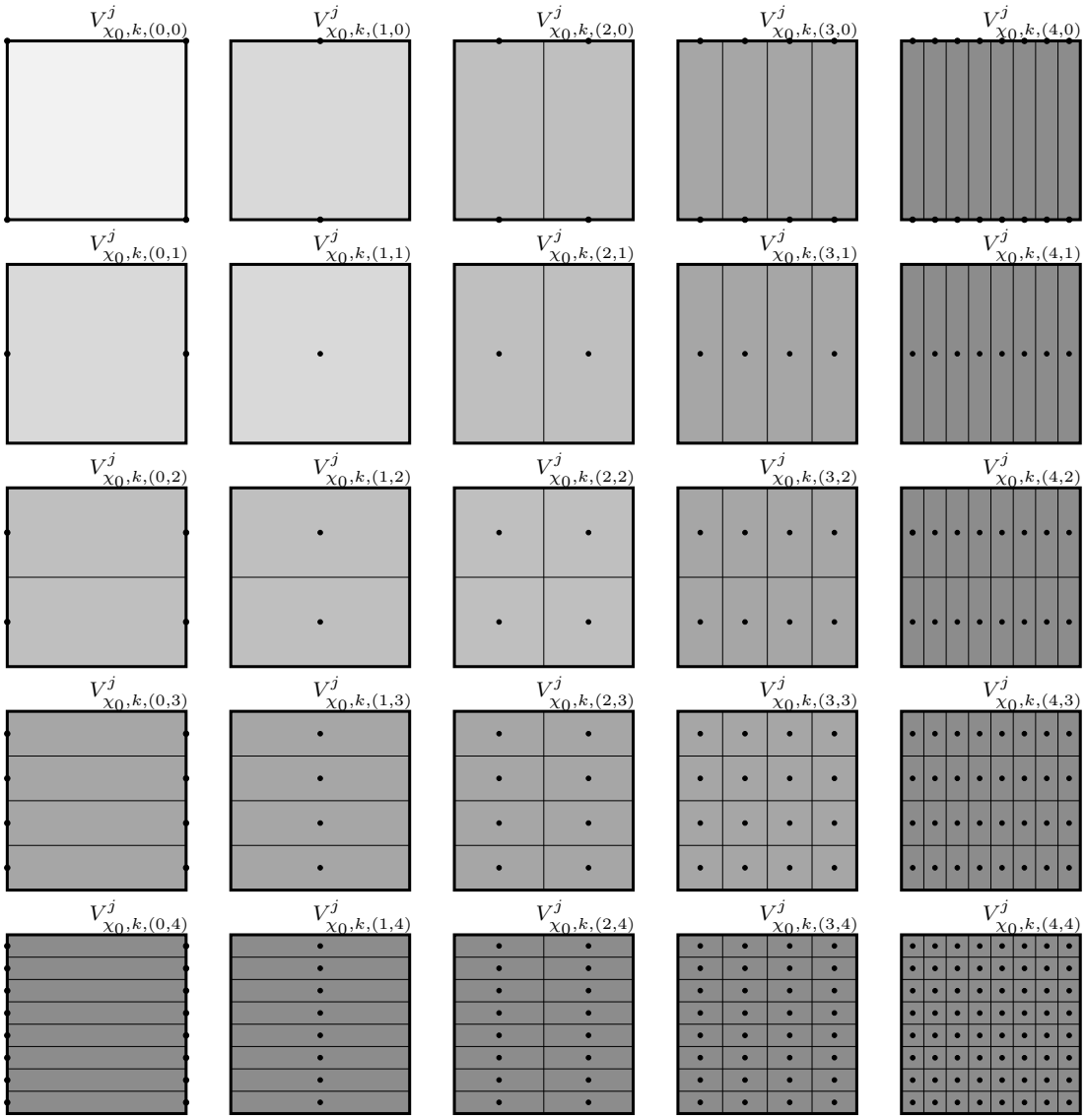

Abbildung 3.8: Volles $\mathbb{R}^2$-Gitter für $|\chi_0|=4$ bis Skalenindex $s=4$

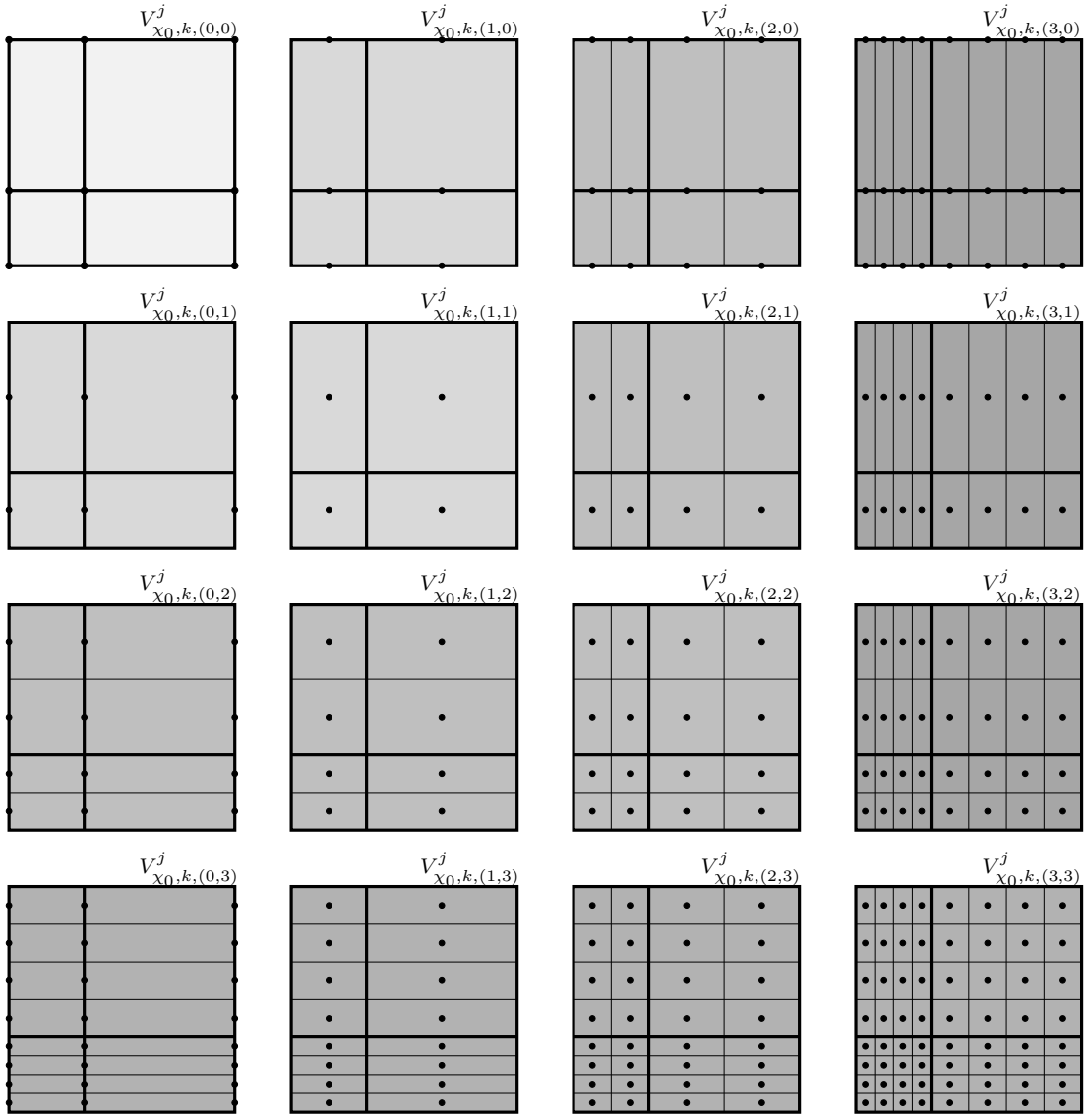

Abbildung 3.9: Volles $\mathbb{R}^2$-Gitter für $|\chi_0|=9$ bis Skalenindex $s=3$.

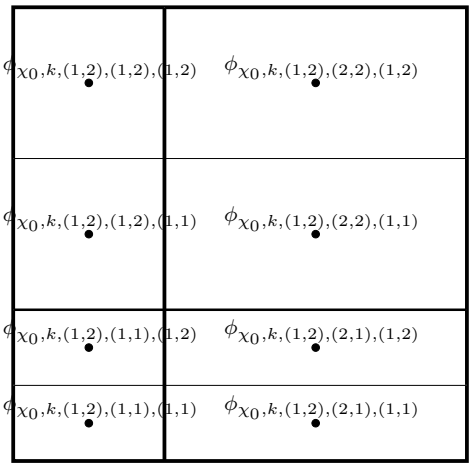

Abbildung 3.10: Basisfunktionen des Funktionsraums $V^j_{\chi_0,k,(1,2)}$

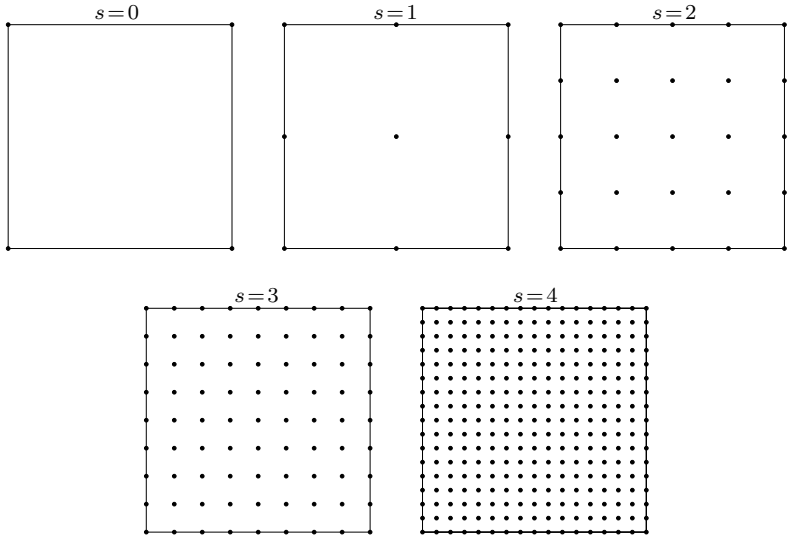

Abbildung 3.11: Stützstellen des vollen $\mathbb{R}^2-$Gitters, $|\chi_0|=4$, $s\leq 4$

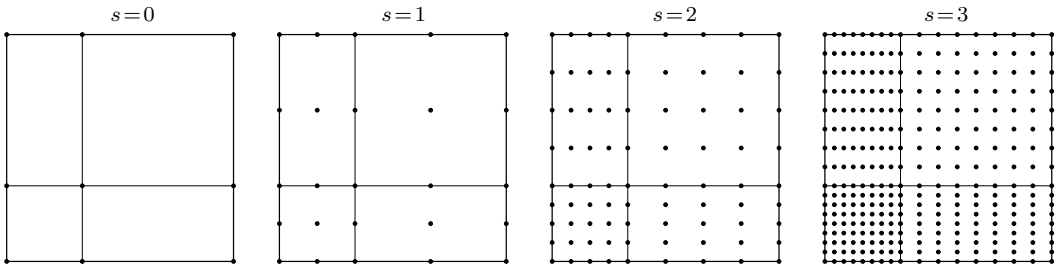

Abbildung 3.12: Stützstellen des vollen $\mathbb{R}^2-$Gitters, $|\chi_0|=9$, $s\leq 3$

Über die globale Bedeutung der Basisfunktionen gibt das folgende Lemma Auskunft.

Lemma 3.20 (Normen der Basisfunktionen)
Sei auf $Q=[a^1,b^1]\times\ldots\times[a^d,b^d]$, $d\geq 1$, *eine nach (3.30) definierte Knotenmenge* $\chi_0:=\chi_{\vec{r}_0}=\left\{\vec{x}_{\vec{r}_0,\vec{v}}\right\}_{\vec{v}\in I_d(\vec{r}_0)}$, $\vec{r}_0\in\mathbb{N}^d$, $\check{r}_0^i\geq 2$, $i=1,\ldots,d$, *mit*

$$Q_{\chi_0,\vec{\iota}}=\bigotimes_{i=1}^{d}\left[x^i_{\chi_0,\iota^i},x^i_{\chi_0,\iota^i+1}\right],\quad |Q_{\chi_0,\vec{\iota}}|=\prod_{i=1}^{d}q^i_{\chi_0,\iota^i}\,,$$

$\vec{\iota}\in I_d(\vec{\check{r}}_0-\vec{1})$, *gegeben. Dann gilt für beliebigen Skalenindex* $s\geq 0$, *unter Verwendung der multilinearen Skalierungsfunktion*

$$\varphi_1^d:[0,1]^d\to[0,1]\,,\quad \vec{x}:=(x^1,\ldots,x^d)^T\mapsto\bigotimes_{i=1}^{d}\varphi_1(x^i)\,,$$

$$\varphi_1(x^i)=\begin{cases}1-2|x^i-\frac{1}{2}|\,, & \textit{falls } x^i\in[0,1]\\ 0\qquad\quad, & \textit{sonst}\end{cases},$$

für $\phi_{\chi_0,1,\vec{\sigma},\vec{\iota},\vec{\nu}}\in W^{\mathrm{hier},j}_{\chi_0,1,\vec{\sigma},\vec{\iota},\vec{\nu}}$, $\vec{\sigma}\in\Upsilon^{\mathrm{voll},j}_{\mathrm{hier},d}(s)$, $\vec{\iota}\in I_d(\vec{\check{r}}_0-\vec{1})$, $\vec{\nu}\in I_d(\vec{\check{r}}_{\vec{\sigma}})$,

$$\begin{aligned}\left\|\phi_{\chi_0,1,\vec{\sigma},\vec{\iota},\vec{\nu}}\right\|_\infty &= 1\quad \textit{und}\\ \left\|\phi_{\chi_0,1,\vec{\sigma},\vec{\iota},\vec{\nu}}\right\|_p &= \left(\frac{2}{p+1}\right)^{\frac{d}{p}}\left(\frac{|Q_{\chi_0,\vec{\iota}}|}{2^{|\vec{\sigma}|_1}}\right)^{\frac{1}{p}},\quad 1\leq p<\infty\,,\end{aligned}$$

und für $\phi_{\chi_0,1,\vec{\sigma},\vec{\nu}}\in W^{\mathrm{twg},j}_{\chi_0,1,\vec{\sigma},\vec{\nu}}$, $\vec{\sigma}\in\Upsilon^{\mathrm{voll},j}_{\mathrm{twg},d}(s)$, $\vec{\nu}\in I_d^{\mathrm{twg}}(\vec{\sigma})$,

$$\begin{aligned}\left\|\phi_{\chi_0,1,\vec{\sigma},\vec{\nu}}\right\|_\infty &= 1\quad \textit{bzw.}\\ \left\|\phi_{\chi_0,1,\vec{\sigma},\vec{\nu}}\right\|_p &= \left(\frac{2}{p+1}\right)^{\frac{d}{p}}\left(\sum_{\substack{\vec{\iota}\in I_d(\vec{r}-\vec{1})\,mit\\ \mathrm{supp}(\phi_{\chi_0,1,\vec{\sigma},\vec{\nu}})\cap Q_{\chi_0,\vec{\iota}}\neq\{\}}}\frac{|Q_{\chi_0,\vec{\iota}}|}{2^{|\vec{\sigma}|_1}2^{n(\vec{\sigma})}}\right)^{\frac{1}{p}},\quad 1\leq p<\infty\,,\end{aligned}$$

mit

$$n(\vec{\sigma}):=d-\sum_{i=1}^{d}\mathrm{sgn}(\sigma^i)\,.$$

Beweis:
Die Aussagen bzgl. der Maximumnorm sind durch die Definition der Skalierungsfunktion trivial. Für die L^p-Norm der auf $W^{\mathrm{hier},j}_{\chi_0,k,\vec{\sigma},\vec{\iota},\vec{\nu}}$ definierten Basisfunktionen

$\phi_{\chi_0,k,\vec{\sigma},\vec{\iota},\vec{\nu}} \in W^{\mathrm{hier},j}_{\chi_0,k,\vec{\sigma},\vec{\iota},\vec{\nu}}$, $\vec{\sigma} \in \Upsilon^{\mathrm{voll},j}_{\mathrm{hier},d}$, $\vec{\iota} \in I_d(\vec{r}_0 - \vec{1})$, $\vec{\nu} \in I_d(\vec{\tilde{r}}_{\vec{\sigma}})$, gilt unter Verwendung der in (3.35) eingeführten Notation

$$\begin{aligned}
\left\|\phi_{\chi_0,1,\vec{\sigma},\vec{\iota},\vec{\nu}}\right\|_p &= \left(\int_Q \left|\phi_{\chi_0,1,\vec{\sigma},\vec{\iota},\vec{\nu}}(\vec{x})\right|^p d\vec{x}\right)^{\frac{1}{p}} \\
&= \left(\int_{Q_{\chi_0,\vec{\iota}}} \prod_{i=1}^{d} \varphi_1\left(2^{\sigma^i-1}\frac{x^i - x^i_{0,\iota^i}}{q^i_{\chi_0,\iota^i}} - (\nu^i - 1)\right)^p d\vec{x}\right)^{\frac{1}{p}} \\
&= \left(\prod_{i=1}^{d} \int_{x^i_{0,\iota^i}}^{x^i_{0,\iota^i+1}} \varphi_1\left(2^{\sigma^i-1}\frac{x^i - x^i_{0,\iota^i}}{q^i_{\chi_0,\iota^i}} - (\nu^i - 1)\right)^p dx^i\right)^{\frac{1}{p}}.
\end{aligned}$$

Aus der Symmetrie und dem kompakten Träger von φ_1 folgt mit

$$2^{\sigma^i-1}\frac{x^i - x^i_{0,\iota^i}}{q^i_{\chi_0,\iota^i}} - (\nu^i - 1) = \frac{1}{2} \quad \Leftrightarrow \quad x^i = x^i_{0,\iota^i} + \frac{2\nu^i - 1}{2^{\sigma^i}} q^i_{\chi_0,\iota^i} =: \tau^i_a$$

und

$$2^{\sigma^i-1}\frac{x^i - x^i_{0,\iota^i}}{q^i_{\chi_0,\iota^i}} - (\nu^i - 1) = 1 \quad \Leftrightarrow \quad x^i = x^i_{0,\iota^i} + \frac{\nu^i}{2^{\sigma^i-1}} q^i_{\chi_0,\iota^i} =: \tau^i_b$$

schließlich

$$\begin{aligned}
\left\|\phi_{\chi_0,1,\vec{\sigma},\vec{\iota},\vec{\nu}}\right\|_p &= \left(\prod_{i=1}^{d} 2\int_{\tau^i_a}^{\tau^i_b} \left(1 - 2\left(2^{\sigma^i-1}\frac{x^i - x^i_{0,\iota^i}}{q^i_{\chi_0,\iota^i}} - \nu^i + \frac{1}{2}\right)\right)^p dx^i\right)^{\frac{1}{p}} \\
&= \left(\prod_{i=1}^{d} 2\int_{\tau^i_a}^{\tau^i_b} \left(2\nu^i - 2^{\sigma^i}\frac{x^i - x^i_{0,\iota^i}}{q^i_{\chi_0,\iota^i}}\right)^p dx^i\right)^{\frac{1}{p}} \\
&= \left(\prod_{i=1}^{d} \frac{2}{p+1}\left[\left(2\nu^i - 2^{\sigma^i}\frac{x^i - x^i_{0,\iota^i}}{q^i_{\chi_0,\iota^i}}\right)^{p+1}\frac{q^i_{\chi_0,\iota^i}}{-2^{\sigma^i}}\right]_{\tau^i_a}^{\tau^i_b}\right)^{\frac{1}{p}} \\
&= \left(\frac{2}{p+1}\right)^{\frac{d}{p}}\left(\prod_{i=1}^{d}\frac{q^i_{\chi_0,\iota^i}}{2^{\sigma^i}}\right)^{\frac{1}{p}} = \left(\frac{2}{p+1}\right)^{\frac{d}{p}}\left(\frac{|Q_{\chi_0,\iota^i}|}{2^{|\vec{\sigma}|_1}}\right)^{\frac{1}{p}}.
\end{aligned}$$

Bleibt die L^p-Norm

$$\begin{aligned}\left\|\phi_{\chi_0,1,\vec{\sigma},\vec{\nu}}\right\|_p &= \left(\int_Q \left|\phi_{\chi_0,1,\vec{\sigma},\vec{\nu}}(\vec{x})\right|^p d\vec{x}\right)^{\frac{1}{p}} \\ &= \left(\sum_{\substack{\vec{\iota}\in I_d(\vec{r}-\vec{1})\text{mit} \\ \text{supp}(\phi_{\chi_0,1,\vec{\sigma},\vec{\nu}})\cap Q_{\chi_0,\vec{\iota}}\neq\{\}}} \int_{Q_{\chi_0\, vec\iota}} \left|\phi_{\chi_0,1,\vec{\sigma},\vec{\nu}}(\vec{x})\right|^p d\vec{x}\right)^{\frac{1}{p}} \\ &= \left(\sum_{\substack{\vec{\iota}\in I_d(\vec{r}-\vec{1})\text{mit} \\ \text{supp}(\phi_{\chi_0,1,\vec{\sigma},\vec{\nu}})\cap Q_{\chi_0,\vec{\iota}}\neq\{\}}} \prod_{i=1}^{d} \int_{x^i_{0,\iota^i}}^{x^i_{0,\iota^i+1}} \left|\phi_{\chi_0,1,\sigma^i,\nu^i}(\vec{x})\right|^p dx^i\right)^{\frac{1}{p}}\end{aligned}$$

der auf $W^{\text{twg},j}_{\chi_0,k,\vec{\sigma},\vec{\nu}}$ definierten Basisfunktionen $\phi_{\chi_0,k,\vec{\sigma},\vec{\nu}} \in W^{\text{twg},j}_{\chi_0,k,\vec{\sigma},\vec{\nu}}$, $\vec{\sigma} \in \Upsilon^{\text{voll},j}_{\text{twg},d}$, $\vec{\nu}\in I^{\text{twg}}_d(\vec{\sigma})$, zu bestimmen. Für die Komponenten mit $\sigma^i \geq 1$ kann analog zu oben argumentiert werden, so dass direkt

$$\int_{x^i_{0,\iota^i}}^{x^i_{0,\iota^i+1}} \left|\phi_{\chi_0,1,\sigma^i,\nu^i}(\vec{x})\right|^p dx^i = \frac{1}{p+1}\,\frac{q^i_{\chi_0,\iota^i}}{2^{\sigma^i-1}}$$

angegeben werden kann. Im Fall $\sigma^i = 0$ folgt

$$\begin{aligned}\int_{x^i_{0,\iota^i}}^{x^i_{0,\iota^i+1}} \left|\phi_{\chi_0,1,\sigma^i,\nu^i}(\vec{x})\right|^p dx^i &= \int_{x^i_{0,\iota^i}}^{x^i_{0,\iota^i+1}} \varphi_k\left(\frac{1}{2}\left(1+\frac{x^i-x^i_{0,\iota^i}}{q^i_{\chi_0,\iota^i}}-(\nu^i-\iota^i)\right)\right)^p \\ &= \int_{x^i_{0,\iota^i}}^{x^i_{0,\iota^i+1}} \left(1-\left|\frac{x^i-x^i_{0,\iota^i}}{q^i_{\chi_0,\iota^i}}-(\nu^i-\iota^i)\right|\right)^p\end{aligned}$$

und wegen $\nu^i-\iota^i\in\{0,1\}$ und

$$\left|\frac{x^i-x^i_{0,\iota^i}}{q^i_{\chi_0,\iota^i}}-(\nu^i-\iota^i)\right| = \begin{cases} \dfrac{x^i-x^i_{0,\iota^i}}{q^i_{\chi_0,\iota^i}} & \text{, falls } \nu^i=\iota^i \\ \dfrac{x^i_{0,\iota^i+1}-x^i}{q^i_{\chi_0,\iota^i}} & \text{, sonst}\end{cases}$$

auch

$$\begin{aligned}\int_{x^i_{0,\iota^i}}^{x^i_{0,\iota^i+1}} \left|\phi_{\chi_0,1,\sigma^i,\nu^i}(\vec{x})\right|^p dx^i &= \int_{x^i_{0,\iota^i}}^{x^i_{0,\iota^i+1}} \left(1-\frac{x^i-x^i_{0,\iota^i}}{q^i_{\chi_0,\iota^i}}\right)^p = \int_{x^i_{0,\iota^i}}^{x^i_{0,\iota^i+1}} \left(\frac{x^i_{0,\iota^i+1}-x^i}{q^i_{\chi_0,\iota^i}}\right)^p \\ &= \left[\frac{-1}{p+1}\left(\frac{x^i_{0,\iota^i+1}-x^i}{q^i_{\chi_0,\iota^i}}\right)^{p+1} q^i_{\chi_0,\iota^i}\right]_{x^i_{0,\iota^i}}^{x^i_{0,\iota^i+1}} = \frac{1}{p+1}\,q^i_{\chi_0,\iota^i}\end{aligned}$$

und damit die Behauptung

□

Bemerkung 3.21
Ein Beweis von Lemma 3.20 findet sich für $|\chi_0| = 2^d$ *und* $\vec{\sigma} \neq \Upsilon_{\mathrm{twg},d}^{\mathrm{voll},j}$ *z.B. auch in Bungartz [15], Lemma 2.3 und 4.1.*

Folgerung 3.22
Seien alle Voraussetzungen des Lemmas 3.20 erfüllt. Dann gilt für beliebiges $\vec{\sigma} \in \Upsilon_d^{\mathrm{voll}}(s)$, $s \geq 0$,

$$\left\| \phi_{\chi_0,1,\vec{\sigma},\vec{\iota}_1,\vec{\nu}_1} \right\|_p = \left\| \phi_{\chi_0,1,\vec{\sigma},\vec{\iota}_2,\vec{\nu}_2} \right\|_p \quad \forall\, \vec{\iota}_1, \vec{\iota}_2 \in I_d(\vec{\vec{r}}_0 - \vec{1}),\ \vec{\nu}_1, \vec{\nu}_2 \in I_d(\vec{\vec{r}}_{\vec{\sigma}})$$

bzw.

$$\left\| \phi_{\chi_0,1,\vec{\sigma},\vec{\nu}_1}\big|_{Q_{\chi_0,\vec{\iota}}} \right\|_p = \left\| \phi_{\chi_0,1,\vec{\sigma},\vec{\nu}_2}\big|_{Q_{\chi_0,\vec{\iota}}} \right\|_p \quad \forall\, \vec{\iota} \in I_d(\vec{\vec{r}}_0 - \vec{1}),\ \vec{\nu}_1, \vec{\nu}_2 \in I_d^{\mathrm{twg}}(\vec{\sigma})$$

mit $\mathrm{supp}(\phi_{\chi_0,1,\vec{\sigma},\vec{\nu}_l}) \cap Q_{\chi_0,\vec{\iota}} \neq \{\}$, $l = 1, 2$. *Weiter gilt für* $s_1, s_2 \geq 0$, $s_1 \neq s_2$,

$$\begin{aligned} &\max \left\{ \left\| \phi_{\chi_0,1,\vec{\sigma}} \right\|_p,\ \vec{\sigma}\ \textit{Basis des}\ V_{\chi_0,1,s_1}^j \right\} \\ &\qquad > \max \left\{ \left\| \phi_{\chi_0,1,\vec{\sigma}} \right\|_p,\ \vec{\sigma}\ \textit{Basis des}\ V_{\chi_0,1,s_2}^j \right\}. \end{aligned} \tag{3.40}$$

Beweis:
Direkte Konsequenz von Lemma 3.20.

□

Bemerkung 3.23
Durch (3.40) wurde formal gezeigt, dass (unter Verwendung eines vollen Gitters) $V_{\chi_0,1,s}^j$*, mit steigendem Skalenindex* s*, einen immer stärker lokalisierenden Einfluss besitzt. Folglich treten, im Gegensatz zu lokalen Basen, neben anfänglich sehr globalen Basisfunktionen nach und nach Funktionen mit immer kleineren Trägern auf.*

Ein (entscheidendes) Problem in der praktischen Umsetzung dieser Skalierungsstrategie ist die extrem schnell anwachsende Komplexität der entstehenden Optimierungsaufgabe. Insbesondere für höherdimensionale Aufgabenstellungen steigt die Dimension ins Uferlose (vgl. Gleichung (3.39)), so dass sofort der 1961 von Bellmann [6] verwendete Ausspruch *"Fluch der Dimensionalität"* in Erinnerung ist. Um diesem Effekt entgegenzuwirken wird im Folgenden der Ansatz dünner Gitter vorgestellt.

3.2.2.2.2 Dünne Gitter

Einen eleganten Ansatz zur Reduzierung der Dimensionalität des zugehörigen Optimierungsproblems liefern sogenannte dünne Gitter (vgl. z.B. Bungartz [15] oder Bungartz, Griebel [16]). Damit die Bedingung (3.26) nicht verletzt wird, müssen auch hier alle bereits für die volle Gitterdiskretisierung definierten Funktionräume $V^j_{\chi_0,k,\vec{\sigma}}$, $\vec{\sigma} \in \Upsilon_d(s)$, hierarchisch eingebunden werden. Der Unterschied liegt nun darin, dass teure Räume mit großem Freiheitsgrad sukzessive in höhere Level verschoben werden. Konkret lautet die Skalierungsstrategie

$$\Upsilon_d^{\text{Strat}}(s) = \Upsilon_d^{\text{dünn}}(s) := \left\{ \vec{\sigma} \in \Upsilon_d(s) \,\middle|\, |\vec{\sigma}|_1 = d+(s-1) \right\}.$$

Vergleiche hierzu auch die Abbildungen 3.13 und 3.14, in denen, exemplarisch für $d=2$ für die Minimalknotenmenge $|\chi_0|=4$ bzw. für eine nicht äquidistante Unterteilung mit $|\chi_0|=9$, die Träger aller Basisfunktionen sowie die zugehörigen Stützstellen für $s \leq 4$ bzw. $s \leq 3$ dargestellt werden. Die zur gleichen Skalierungsebene gehörenden Unterräume $V^j_{\chi_0,k,\vec{\sigma}}$, $\vec{\sigma} \in \Upsilon_d^{\text{dünn}}(s)$, sind wieder mit gleicher Intensität schattiert. In den Abbildungen 3.15 und 3.16 werden schließlich die Stützstellen aller skalenweise definierter Unterräume $V^j_{\chi_0,k,\vec{\sigma}}$, $\vec{\sigma} \in \Upsilon_d^{\text{dünn}}(s)$, $s \leq 4$, bzw. $s \leq 3$, in einer Domain ausgegeben. Es ergibt sich die typische Struktur dünner Gitter.

Die Anzahl der Parameter und damit die Dimensionalität dünner Gitter wird durch den folgenden Satz gegeben.

Satz 3.24

Sei auf $Q = [a^1, b^1] \times \ldots \times [a^d, b^d]$, $d \geq 2$, eine nach (3.30) definierte Knotenmenge $\chi_0 := \chi_{\vec{r}_0} = \left\{ \vec{x}_{\vec{r}_0,\vec{v}} \right\}_{\vec{v} \in I_d(\vec{r}_0)}$, $\vec{r}_0 \in \mathbb{N}^d$, $\check{r}_0^{\,i} \geq 2$, $i = 1, \ldots, d$, gegeben. Dann lässt sich die Dimension der auf dünnen Gittern erzeugten Parameterräume $P^{\text{dünn},j}_{\chi_0,1,s}$, $s \geq 1$, rekursiv mit Hilfe von

$$\left| P^{\text{dünn},j}_{\chi_0^{(1)},1,s} \right| = (\check{r}_0^1 - 1)(2^s - 1) + \check{r}_0^1 \,,$$

$$\left| P^{\text{dünn},j}_{\chi_0^{(i)},1,s} \right| = (2\check{r}_0^{\,i} - 1) \left| P^{\text{dünn},j}_{\chi_0^{(i-1)},1,s} \right| + (\check{r}_0^{\,i} - 1) \sum_{l=1}^{s-1} 2^l \left| P^{\text{dünn},j}_{\chi_0^{(i-1)},1,s-l} \right|, \quad i = 2, \ldots, d \,,$$

und

$$\chi_0^{(i)} := \left\{ \vec{x}_{(i)} \in \mathbb{R}^i \,\middle|\, \exists\, \vec{x} \in \chi_0 : x^j_{(i)} = x^j \; \forall j = 1, \ldots, i \right\}, \quad i = 1, \ldots, d \,,$$

durch

$$\left| P^{\text{dünn},j}_{\chi_0,1,s} \right| = \left| P^{\text{dünn},j}_{\chi_0^{(d)},1,s} \right|$$

angeben.

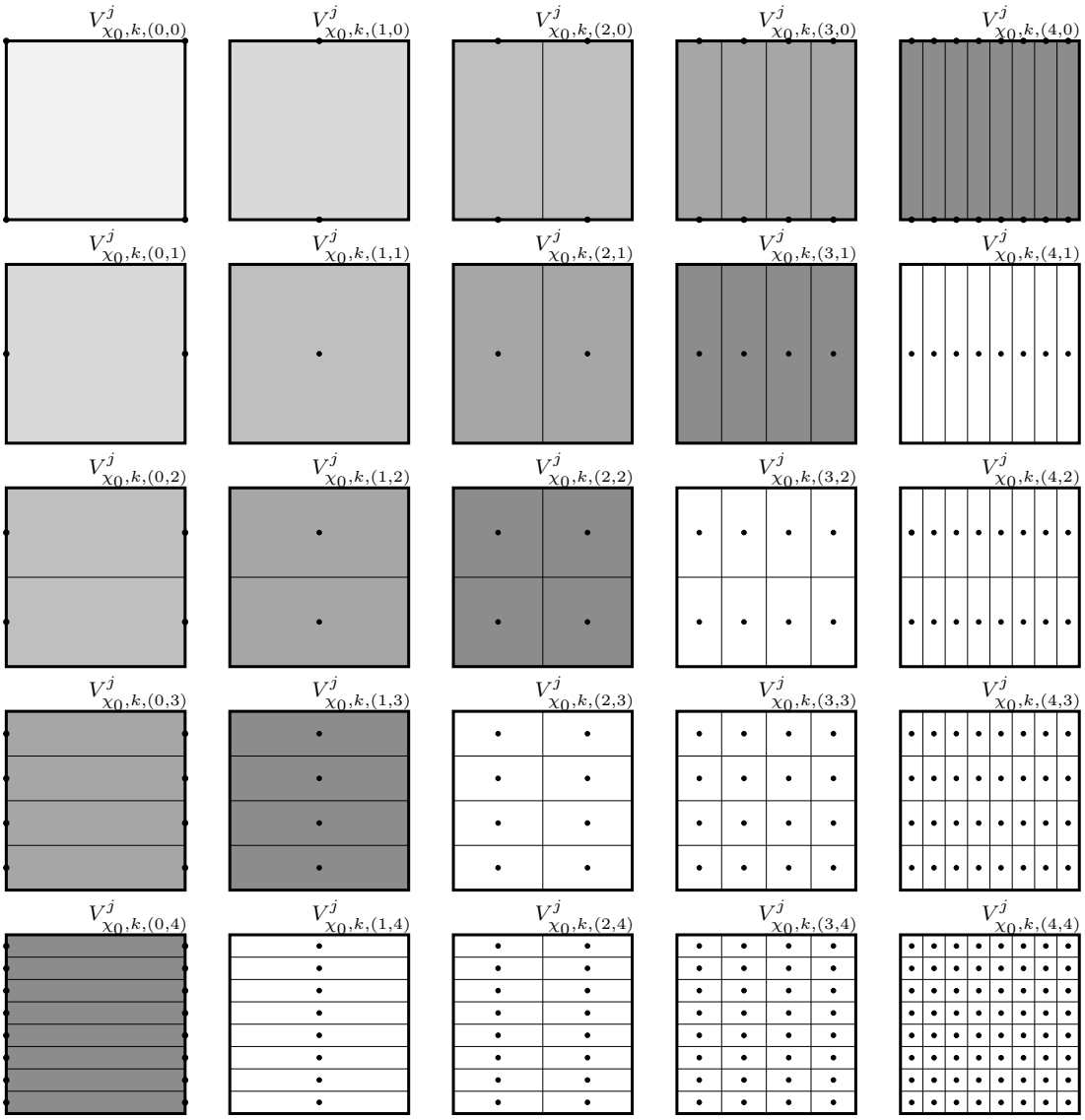

Abbildung 3.13: Dünnes $\mathbb{R}^2$-Gitter für $|\chi_0| = 4$ bis Skalenindex $s = 4$

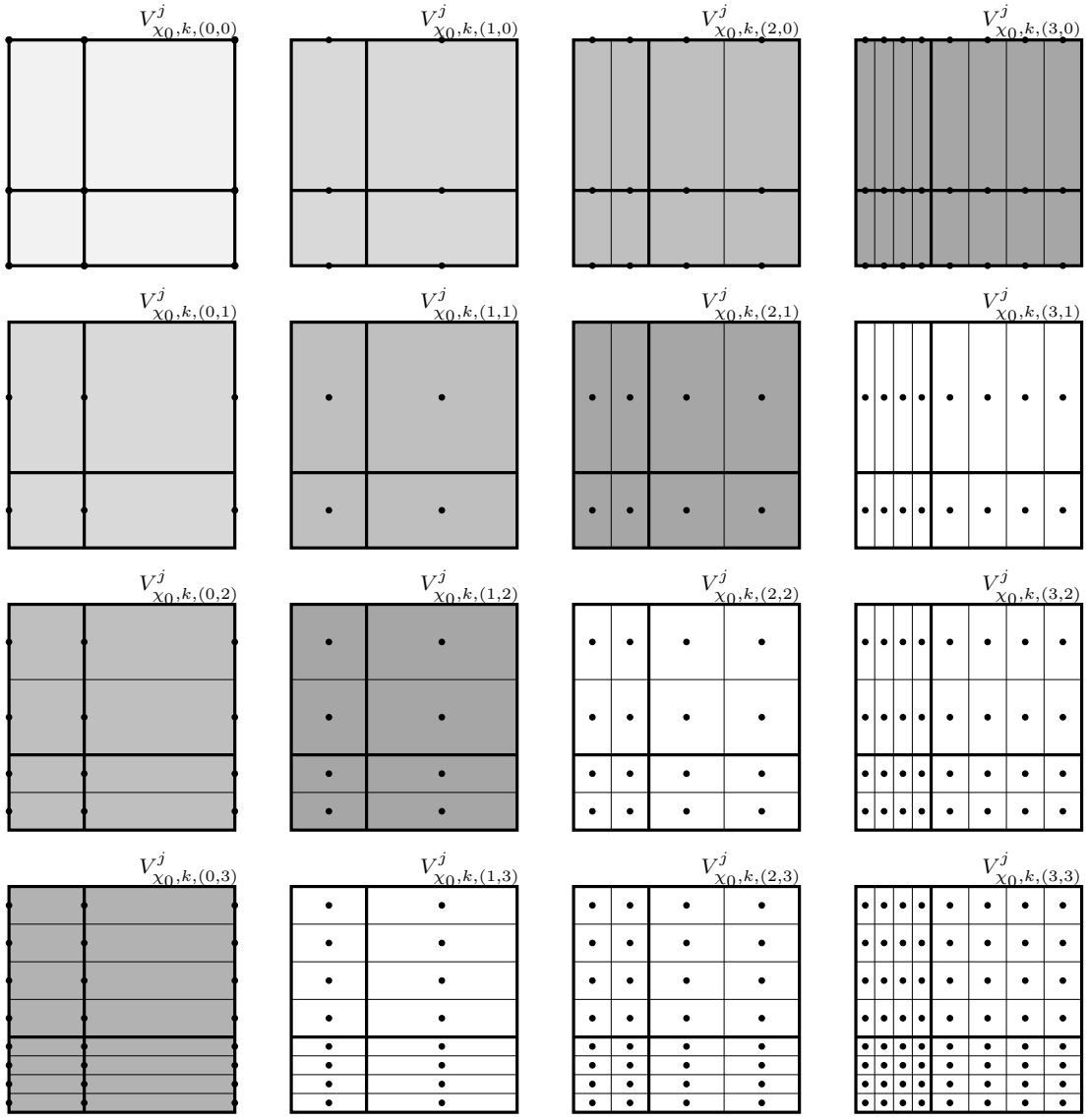

Abbildung 3.14: Dünnes $\mathbb{R}^2$-Gitter für $|\chi_0| = 9$ bis Skalenindex $s = 3$.

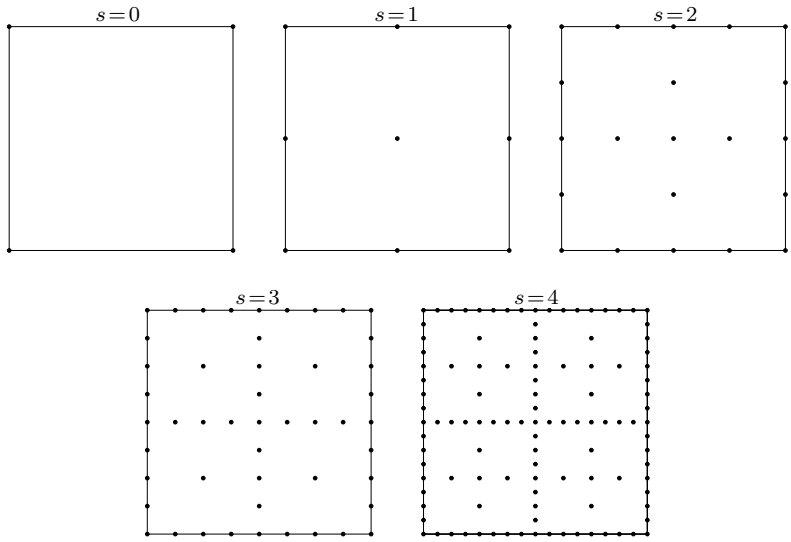

Abbildung 3.15: Stützstellen des vollen $\mathbb{R}^2-$Gitters, $|\chi_0|=4$, $s\leq 4$

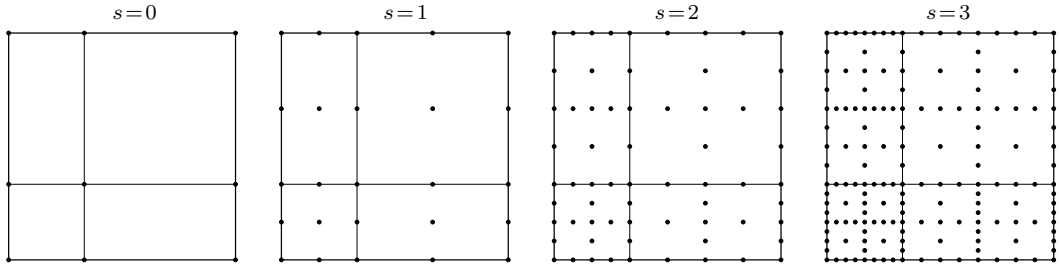

Abbildung 3.16: Stützstellen des vollen $\mathbb{R}^2-$Gitters, $|\chi_0|=9$, $s\leq 3$

Beweis:
Ein Beweis für $|\chi_0|=2^d$ findet sich in Bungartz [15], Satz 3.3. Der hier aufgeführte verallgemeinerte Fall $\check{r}_0^i \geq 2$, $i=1,\ldots,d$, wird, unter Berücksichtigung der Vorfaktoren $(2\check{r}_0^i-1)$ und $(\check{r}_0^i-1)$, auf analoge Weise bewiesen.

□

Die Dimension des zugehörigen Interpolationsproblems reduziert sich damit entscheidend. In den Tabellen 3.2 und 3.3 werden die ermittelten Dimensionalitäten sowohl dünner als auch voller Gitter für $d\in\{2,3,4\}$ für die Minimalunterteilung $|\chi_0|=2^d$ sowie den (nicht äquidistanten) Fall $|\chi_0|=3^d$ angegeben.

Bleibt die Frage, inwieweit der Vorteil der verringerten Dimensionalität durch eine resultierende Vergrößerung des Approximationsfehlers kompensiert wird. Sei

$d=2$	$s=0$	$s=1$	$s=2$	$s=3$	$s=4$	$s=5$	$s=6$	$s=7$	$s=8$	$s=9$
voll	4	9	25	81	289	1089	4225	16641	66049	263169
dünn	4	9	21	49	113	257	577	1281	2817	6145

$d=3$	$s=0$	$s=1$	$s=2$	$s=3$	$s=4$	$s=5$	$s=6$
voll	8	27	125	729	4913	35937	274625
dünn	8	27	81	225	593	1505	3713

$d=4$	$s=0$	$s=1$	$s=2$	$s=3$	$s=4$	$s=5$
voll	16	81	625	6561	83521	1185921
dünn	16	81	297	945	2769	7681

Tabelle 3.2: Dimensionalität voller und dünner $\mathbb{R}^d$-Gitter für $|\chi_0|=2^d$

$d=2$	$s=0$	$s=1$	$s=2$	$s=3$	$s=4$	$s=5$	$s=6$	$s=7$	$s=8$
voll	9	25	81	289	1089	4225	16641	66049	263169
dünn	9	25	65	161	385	897	2049	4609	10241

$d=3$	$s=0$	$s=1$	$s=2$	$s=3$	$s=4$	$s=5$
voll	27	125	729	4913	35937	274625
dünn	27	125	425	1265	3489	9153

$d=4$	$s=0$	$s=1$	$s=2$	$s=3$	$s=4$
voll	81	625	6561	83521	1185921
dünn	81	625	2625	9025	27905

Tabelle 3.3: Dimensionalität voller und dünner $\mathbb{R}^d$-Gitter für $|\chi_0|=3^d$

hierzu der folgende Satz angegeben.

Satz 3.25 (Approximationsfehler)
Sei auf $Q=[a^1,b^1]\times\ldots\times[a^d,b^d]$, $d\geq 2$, *eine nach (3.30) definierte Knotenmenge* $\chi_0:=\chi_{\vec{r}_0}=\left\{\vec{x}_{\vec{r}_0,\vec{v}}\right\}_{\vec{v}\in I_d(\vec{r}_0)}$, $\vec{r}_0\in\mathbb{N}^d$, $r_0^i\geq 2$, $i=1,\ldots,d$, *mit*

$$Q_{\chi_0,\vec{\iota}}=\bigotimes_{i=1}^{d}\left[x^i_{\chi_0,\iota^i},x^i_{\chi_0,\iota^i+1}\right],\quad |Q_{\chi_0,\vec{\iota}}|=\prod_{i=1}^{d}q^i_{\chi_0,\iota^i},$$

$\vec{\iota}\in I_d(\vec{r}_0-\vec{1})$, *gegeben. Dann gilt für beliebigen Skalenindex* $s\geq 0$ *für die auf einer vollen Diskretisierung basierenden Parametrisierung*

$$\left\| f^j - f^{\text{voll},j}_{\chi_0,1,s} \right\|_{Q_{\chi_0,\vec{\iota}},\infty} \leq \frac{d\,|Q_{\chi_0,\vec{\iota}}|^2}{4^s\,6^d}\,\left|f^j\right|_{Q_{\chi_0,\vec{\iota}},2,\infty} \quad \textit{und}$$

$$\left\| f^j - f^{\text{voll},j}_{\chi_0,1,s} \right\|_{Q_{\chi_0,\vec{\iota}},2} \leq \frac{d\,|Q_{\chi_0,\vec{\iota}}|^2}{4^s\,9^d}\,\left|f^j\right|_{Q_{\chi_0,\vec{\iota}},2,2}$$

und unter Verwendung von

$$A(d,s) := \sum_{l=0}^{d-1} \binom{s+d-1}{l} = \frac{s^{d-1}}{(d-1)!} + \mathcal{O}(s^{d-2})$$

für eine auf dünnen Gittern definierte Parametrisierung

$$\left\| f^j - f^{\text{dünn},j}_{\chi_0,1,s} \right\|_{Q_{\chi_0,\vec{\iota}},\infty} \leq \frac{2\,|Q_{\chi_0,\vec{\iota}}|^2}{4^s\,8^d} A(d,s) \left|f^j\right|_{Q_{\chi_0,\vec{\iota}},2,\infty},$$

$$\left\| f^j - f^{\text{dünn},j}_{\chi_0,1,s} \right\|_{Q_{\chi_0,\vec{\iota}},2} \leq \frac{2\,|Q_{\chi_0,\vec{\iota}}|^2}{4^s\,12^d} A(d,s) \left|f^j\right|_{Q_{\chi_0,\vec{\iota}},2,2}.$$

Beweis:
Ein Beweis findet sich z.B. in Bungartz, Griebel [16], Kap. 3, Theorem 3.5 und 3.8.

□

Somit ist sichergestellt, dass trotz signifikanter Reduzierung der Dimension nur ein geringfügig höherer Approximationsfehler entsteht. In den Tabellen 3.4 - 3.6 werden für $d \in \{2,3,4\}$ entsprechend die oberen Schranken der relativen Fehler

$$\varepsilon_{\text{voll},\infty} := \frac{\left\| f^j - f^{\text{voll},j}_{\chi_0,1,s} \right\|_{Q_{\chi_0,\vec{\iota}},\infty}}{|Q_{\chi_0,\vec{\iota}}|^2 \left|f^j\right|_{Q_{\chi_0,\vec{\iota}},2,\infty}} \leq \frac{d}{4^s 6^d},$$

$$\varepsilon_{\text{voll},2} := \frac{\left\| f^j - f^{\text{voll},j}_{\chi_0,1,s} \right\|_{Q_{\chi_0,\vec{\iota}},2}}{|Q_{\chi_0,\vec{\iota}}|^2 \left|f^j\right|_{Q_{\chi_0,\vec{\iota}},2,2}} \leq \frac{d}{4^s 9^d}$$

und

$$\varepsilon_{\text{dünn},\infty} := \frac{\left\| f^j - f^{\text{dünn},j}_{\chi_0,1,s} \right\|_{Q_{\chi_0,\vec{\iota}},\infty}}{|Q_{\chi_0,\vec{\iota}}|^2 \left|f^j\right|_{Q_{\chi_0,\vec{\iota}},2,\infty}} \leq \frac{2A(d,s)}{4^s 8^d},$$

$$\varepsilon_{\text{dünn},2} := \frac{\left\| f^j - f^{\text{dünn},j}_{\chi_0,1,s} \right\|_{Q_{\chi_0,\vec{\iota}},2}}{|Q_{\chi_0,\vec{\iota}}|^2 \left|f^j\right|_{Q_{\chi_0,\vec{\iota}},2,2}} \leq \frac{2A(d,s)}{4^s 12^d}$$

beider Skalierungsstrategien gegenübergestellt. Wird eine maximale Fehlertoleranz vorgegeben, so kann hieraus leicht die notwendige Skalierungstiefe und die damit verbundene Anzahl an Interpolationsparametern ermittelt werden (vgl. Tabelle 3.7 - 3.12).

$d=2$		$s=1$	$s=2$	$s=3$	$s=4$	$s=5$	$s=6$	$s=7$
voll	$\varepsilon_{\text{voll},\infty}$	$1,4{\cdot}10^{-2}$	$3,5{\cdot}10^{-3}$	$8,7{\cdot}10^{-4}$	$2,2{\cdot}10^{-4}$	$5,4{\cdot}10^{-5}$	$1,4{\cdot}10^{-5}$	$3,4{\cdot}10^{-6}$
	$\varepsilon_{\text{voll},2}$	$6,2{\cdot}10^{-3}$	$1,5{\cdot}10^{-3}$	$3,9{\cdot}10^{-4}$	$9,6{\cdot}10^{-5}$	$2,4{\cdot}10^{-5}$	$6,0{\cdot}10^{-6}$	$1,5{\cdot}10^{-6}$
dünn	$\varepsilon_{\text{dünn},\infty}$	$2,3{\cdot}10^{-2}$	$7,8{\cdot}10^{-3}$	$2,4{\cdot}10^{-3}$	$7,3{\cdot}10^{-4}$	$2,1{\cdot}10^{-4}$	$6,1{\cdot}10^{-5}$	$1,7{\cdot}10^{-5}$
	$\varepsilon_{\text{dünn},2}$	$1,0{\cdot}10^{-2}$	$3,5{\cdot}10^{-3}$	$1,0{\cdot}10^{-3}$	$3,3{\cdot}10^{-4}$	$9,5{\cdot}10^{-5}$	$2,7{\cdot}10^{-5}$	$7,6{\cdot}10^{-6}$

Tabelle 3.4: Approximationsfehler voller und dünner $\mathbb{R}^2$-Gitter

$d=3$		$s=1$	$s=2$	$s=3$	$s=4$	$s=5$	$s=6$	$s=7$
voll	$\varepsilon_{\text{voll},\infty}$	$3,5{\cdot}10^{-3}$	$8,7{\cdot}10^{-4}$	$2,2{\cdot}10^{-4}$	$5,4{\cdot}10^{-5}$	$1,4{\cdot}10^{-5}$	$3,4{\cdot}10^{-6}$	$8,5{\cdot}10^{-7}$
	$\varepsilon_{\text{voll},2}$	$1,0{\cdot}10^{-3}$	$2,6{\cdot}10^{-4}$	$6,4{\cdot}10^{-5}$	$1,6{\cdot}10^{-5}$	$4,0{\cdot}10^{-6}$	$1,0{\cdot}10^{-6}$	$2,5{\cdot}10^{-7}$
dünn	$\varepsilon_{\text{dünn},\infty}$	$6,8{\cdot}10^{-3}$	$2,7{\cdot}10^{-3}$	$9,8{\cdot}10^{-4}$	$3,4{\cdot}10^{-4}$	$1,1{\cdot}10^{-4}$	$3,5{\cdot}10^{-5}$	$1,1{\cdot}10^{-5}$
	$\varepsilon_{\text{dünn},2}$	$2,0{\cdot}10^{-3}$	$8,0{\cdot}10^{-4}$	$2,9{\cdot}10^{-4}$	$9,9{\cdot}10^{-5}$	$3,3{\cdot}10^{-5}$	$1,0{\cdot}10^{-5}$	$3,2{\cdot}10^{-6}$

Tabelle 3.5: Approximationsfehler voller und dünner $\mathbb{R}^3$-Gitter

$d=4$		$s=1$	$s=2$	$s=3$	$s=4$	$s=5$	$s=6$	$s=7$
voll	$\varepsilon_{\text{voll},\infty}$	$7,7{\cdot}10^{-4}$	$1,9{\cdot}10^{-4}$	$4,8{\cdot}10^{-5}$	$1,2{\cdot}10^{-5}$	$3,0{\cdot}10^{-6}$	$7,5{\cdot}10^{-7}$	$1,9{\cdot}10^{-7}$
	$\varepsilon_{\text{voll},2}$	$1,5{\cdot}10^{-4}$	$3,8{\cdot}10^{-5}$	$9,5{\cdot}10^{-6}$	$2,3{\cdot}10^{-6}$	$6,0{\cdot}10^{-7}$	$1,5{\cdot}10^{-7}$	$3,7{\cdot}10^{-8}$
dünn	$\varepsilon_{\text{dünn},\infty}$	$1,8{\cdot}10^{-3}$	$7,9{\cdot}10^{-4}$	$3,2{\cdot}10^{-4}$	$1,2{\cdot}10^{-4}$	$4,4{\cdot}10^{-5}$	$1,5{\cdot}10^{-5}$	$5,2{\cdot}10^{-6}$
	$\varepsilon_{\text{dünn},2}$	$3,6{\cdot}10^{-4}$	$1,6{\cdot}10^{-4}$	$6,3{\cdot}10^{-5}$	$2,4{\cdot}10^{-5}$	$8,8{\cdot}10^{-6}$	$3,1{\cdot}10^{-6}$	$1,0{\cdot}10^{-6}$

Tabelle 3.6: Approximationsfehler voller und dünner $\mathbb{R}^4$-Gitter

3.2.3 Multi-Level Algorithmus

Seien $\kappa \in \mathbb{N}$ verschiedene Experimente mit $w_{k,i} \in W_{k,i}$, $k=1,\ldots,\kappa$, $i=1,\ldots,n_k$, Beobachtungen durchgeführt. Desweiteren seien M skalare oder vektorwertige, (kontinuierliche) Nichtlinearitäten $\boldsymbol{f}_l := (f_l^1,\ldots,f_l^{d_l})^T \in P^l = (P^{l,1},\ldots,P^{l,d_l})^T$, $l=1,\ldots,M$, $d_l \in \mathbb{N}$, gesucht. Dann besteht die Identifizierungsaufgabe darin, optimale Parametersätze $\vec{p}_{r_l}^{\,l,j_l} \in P_{r_l}^{l,j_l}$, $l=1,\ldots,M$, $j_l = 1,\ldots,d_l$, der zugehörigen Koeffizientenfunktionen $f_{r_l}^{l,j_l} \in P_{r_l}^{l,j_l}$ zu finden, so dass die simulierten Beobachtungen möglichst optimal an den gegebenen Messdaten angepasst sind. Somit

$d = 2$	volles Gitter		dünnes Gitter	
ε_∞	s	dim $P_{1,s}^{\text{voll},j}$	s	dim $P_{1,s}^{\text{dünn},j}$
$1,0 \cdot 10^{-3}$	3	81	4	113
$5,0 \cdot 10^{-4}$	4	289	5	257
$1,0 \cdot 10^{-4}$	5	1089	6	577
$5,0 \cdot 10^{-5}$	6	4225	7	1281
$1,0 \cdot 10^{-5}$	7	16641	8	2817
$5,0 \cdot 10^{-6}$	7	16641	8	2817
$1,0 \cdot 10^{-6}$	8	66049	10	13313
$5,0 \cdot 10^{-7}$	9	263169	10	13313

Tabelle 3.7: Notwendige Skalierungstiefe eines $\mathbb{R}^2$-Gitters unter Vorgabe von ε_∞

$d = 2$	volles Gitter		dünnes Gitter	
ε_2	s	dim $P_{1,s}^{\text{voll},j}$	s	dim $P_{1,s}^{\text{dünn},j}$
$1,0 \cdot 10^{-3}$	3	81	3	49
$5,0 \cdot 10^{-4}$	3	81	4	113
$1,0 \cdot 10^{-4}$	4	289	5	257
$5,0 \cdot 10^{-5}$	5	1089	6	577
$1,0 \cdot 10^{-5}$	6	4225	7	1281
$5,0 \cdot 10^{-6}$	7	16641	8	2817
$1,0 \cdot 10^{-6}$	8	66049	9	6145
$5,0 \cdot 10^{-7}$	8	66049	10	13313

Tabelle 3.8: Notwendige Skalierungstiefe eines $\mathbb{R}^2$-Gitters unter Vorgabe von ε_2

$d = 3$	volles Gitter		dünnes Gitter	
ε_∞	s	dim $P_{1,s}^{\text{voll},j}$	s	dim $P_{1,s}^{\text{dünn},j}$
$1,0 \cdot 10^{-3}$	2	125	3	225
$5,0 \cdot 10^{-4}$	3	729	4	593
$1,0 \cdot 10^{-4}$	4	4913	6	3713
$5,0 \cdot 10^{-5}$	5	35937	6	3713
$1,0 \cdot 10^{-5}$	6	274625	8	21249
$5,0 \cdot 10^{-6}$	6	274625	8	21249
$1,0 \cdot 10^{-6}$	7	2146689	9	49665

Tabelle 3.9: Notwendige Skalierungstiefe eines $\mathbb{R}^3$-Gitters unter Vorgabe von ε_∞

$d = 3$	volles Gitter		dünnes Gitter	
ε_2	s	dim $P_{1,s}^{\text{voll},j}$	s	dim $P_{1,s}^{\text{dünn},j}$
$1,0 \cdot 10^{-3}$	2	125	2	15
$5,0 \cdot 10^{-4}$	2	125	3	39
$1,0 \cdot 10^{-4}$	3	729	4	119
$5,0 \cdot 10^{-5}$	4	4913	5	359
$1,0 \cdot 10^{-5}$	5	35937	7	2823
$5,0 \cdot 10^{-6}$	5	35937	7	2823
$1,0 \cdot 10^{-6}$	7	2146689	8	21249

Tabelle 3.10: Notwendige Skalierungstiefe eines $\mathbb{R}^3$-Gitters unter Vorgabe von ε_2

$d = 4$	volles Gitter		dünnes Gitter	
ε_∞	s	dim $P_{1,s}^{\text{voll},j}$	s	dim $P_{1,s}^{\text{dünn},j}$
$1,0 \cdot 10^{-3}$	1	81	2	297
$5,0 \cdot 10^{-4}$	2	625	3	945
$1,0 \cdot 10^{-4}$	3	6561	5	7681
$5,0 \cdot 10^{-5}$	3	6561	5	7681
$1,0 \cdot 10^{-5}$	5	1185921	7	52912
$5,0 \cdot 10^{-6}$	5	1185921	8	133835
$1,0 \cdot 10^{-6}$	6	17850625	9	331669

Tabelle 3.11: Notwendige Skalierungstiefe eines $\mathbb{R}^4$-Gitters unter Vorgabe von ε_∞

$d = 4$	volles Gitter		dünnes Gitter	
ε_2	s	dim $P_{1,s}^{\text{voll},j}$	s	dim $P_{1,s}^{\text{dünn},j}$
$1,0 \cdot 10^{-3}$	1	81	1	81
$5,0 \cdot 10^{-4}$	1	81	1	81
$1,0 \cdot 10^{-4}$	2	625	3	945
$5,0 \cdot 10^{-5}$	2	625	4	2769
$1,0 \cdot 10^{-5}$	3	6561	5	7681
$5,0 \cdot 10^{-6}$	4	83521	6	20481
$1,0 \cdot 10^{-6}$	5	1185921	8	133835

Tabelle 3.12: Notwendige Skalierungstiefe eines $\mathbb{R}^4$-Gitters unter Vorgabe von ε_2

gilt es

$$\min_{\substack{\vec{p}_{r_l}^{\,l,j_l}\in P_{r_l}^{l,j_l}\\ l=1,\dots,M,\, j_l=1,\dots,d_l}} \mathcal{J}_h\big(\vec{p}_{r_1}^{\,1,1},\dots,\vec{p}_{r_1}^{\,1,d_1},\dots,\vec{p}_{r_M}^{\,M,1},\dots,\vec{p}_{r_M}^{\,M,d_M}\big) =$$

$$\min_{\substack{\vec{p}_{r_l}^{\,l,j_l}\in P_{r_l}^{l,j_l}\\ l=1,\dots,M,\, j_l=1,\dots,d_l}} \sum_{k=1}^{\kappa}\sum_{i=1}^{n_k} (\mathcal{J}_h)_{k,i}\big(\vec{p}_{r_1}^{\,1,1},\dots,\vec{p}_{r_1}^{\,1,d_1},\dots,\vec{p}_{r_M}^{\,M,1},\dots,\vec{p}_{r_M}^{\,M,d_M}\big) \tag{3.41}$$

zu lösen. Hierzu muss neben einem als Definitionsmenge geeignetes Gebiet $G\subset\mathbb{R}^d$ auch eine optimale Anzahl an Freiheitsgraden r_l, $l=1,\dots,M$, der diskreten Parameterräume $P_{r_l}^{l,j_l}$ gewählt werden. Das Gebiet G wird meist durch einen bereits durch die Experimente $w_{k,i}\in W_{k,i}$ festgelegten Quader $Q\in\mathbb{R}^d$ vorgeschrieben, so dass die entsprechenden Koeffizientenfunktionen lediglich dort zu bestimmen sind. Anders sieht es mit der Frage nach einer optimalen Anzahl an Freiheitsgraden aus. Bei wachsender Anzahl steigt die Komplexität des Optimierungsproblems und die Flexibilität der einzelnen Koeffizientenfunktionen. Dabei kann eine zu geringe Flexibilität zu einer unzureichenden Reproduktion der Messdaten und eine zu hohe (Flexibiliät) zu auf Messfehlern basierenden Effekten führen. Des Weiteren wird eine wachsende Anzahl von Freiheitsgraden die Konvergenzgeschwindigkeit des Optimierungsverfahrens verringern und ggf. zu einem Abbruch in einem Nebenminimum führen.

Aufgrund der oben genannten Aspekte und der von Grund auf hohen Dimension des Identifizierungsproblems ist eine möglichst gute Startschätzung unerlässlich. Daher wird das Problem zunächst mit einer möglichst kleinen Anzahl an Freiheitsgraden (i.d.R. vorgegeben durch den Grad der verwendeten Splineparametrisierung) gelöst und die Parameterdarstellung $\vec{p}_{r_{l,0}+\Delta r_l}^{\,l,j_l}\in P_{r_{l,0}+\Delta r_l}^{l,j_l}$ der ermittelten optimalen Parametervektoren $\vec{p}_{r_{l,0}}^{\,l,j_l}\in P_{r_{l,0}}^{l,j_l}$, $l=1,\dots,M$, $j_l=1,\dots,d_l$, als Startwerte für eine Optimierung in den Räumen $P_{r_{l,0}+\Delta r_l}^{l,j_l}$ herangezogen. Bei sukzessiver Fortsetzung dieser Vorgehensweise gelangt man schließlich zu dem nachfolgenden Multi-Level-Algorithmus wie er bereits (auf ähnliche Weise) in Bitterlich [8], Abschnitt 4.1.4, oder Igler [34], Sektion 4.5, vorgestellt wurde. Dabei können sowohl lokale als auch hierarchischer Basen verwendet werden. Der hierarchische Ansatz hat allerdings den Vorteil, dass die notwendige Interpolationsaufgabe, zum Erreichen der nächsthöheren Skala, durch eine einfache Null-Initialisierung der neuen Unbekannten und Beibehalten der alten Parameterwerte trivialerweise erfüllt ist. In diesem Fall sind Δr_l für alle $l=1,\dots,M$ durch den Skalenschritt $\Delta s=1$, $P_{r_l+\Delta r_l}^{l,j_l}=P_{s_l+\Delta s_l}^{\mathrm{hier},l,j_l}$, $r_l+\Delta r=\dim(P_{s_l+\Delta s_l}^{\mathrm{hier},l,j_l})$, vorgeben. Bei den lokalen Basen müssen die entsprechenden Δr_l, $l=1,\dots,M$, geeignet gewählt werden.

(Formfreie) Parameteridentifizierung — Multi-Level-Algorithmus

FOR $l=1,\dots,M$ DO

- Wähle möglichst geringe Startdimension $r_{l,0}$
- FOR $j_l=1,\dots,d_l$ DO
  - Wähle geeigneten Startwert $\vec{p}^{\,l,j_l}_{r_{l,0}} \in P^{l,j_l}_{r_{l,0}}$
- Setze $r_l \leftarrow r_{l,0}$

Wiederhole solange bis ein Abbruchkriterium erfüllt ist

- Löse das Optimierungsproblem

$$\min_{\substack{\vec{p}^{\,l,j_l}_{r_l} \in P^{l,j_l}_{r_l} \\ l=1,\dots,M,\, j_l=1,\dots,d_l}} \sum_{k=1}^{\kappa} \sum_{i=1}^{n_k} \left(\mathcal{J}_h\right)_{k,i} \left(\vec{p}^{\,1,1}_{r_1}, \dots, \vec{p}^{\,1,d_1}_{r_1}, \dots, \vec{p}^{\,M,1}_{r_M}, \dots, \vec{p}^{\,M,d_M}_{r_M}\right)$$

- FOR $l=1,\dots,M$ DO
  - Wähle möglichst kleines Δr_l
  - FOR $j_l=1,\dots,d_l$ DO
    - Interpoliere Optimalwert $\vec{p}^{\,l,j_l}_{r_l}$ zu $\vec{p}^{\,l,j_l}_{r_l+\Delta r_l} \in P^{l,j_l}_{r_l+\Delta r_l}$
  - Setze $r_l \leftarrow r_l + \Delta r_l$

FOR $l=1,\dots,M$, $j_l=1,\dots,d_l$ DO

- Gib optimalen Parameterwert $\vec{p}^{\,l,j_l}_{r_l}$ aus

Bleiben geeignete Abbruchkriterien festzulegen. Hierzu wird im Folgenden angenommen, dass das betrachtete diskrete Fehlerfunktional in (3.41) die Form (3.18) besitzt und dass der Datenfehler durch ε gegeben ist. Zunächst kann natürlich

trivialerweise eine maximale Anzahl rekursiver Schritte $r_{\max}$ gesetzt werden, so dass der Algorithmus abgebrochen wird, sobald $r_l \geq r_{\max}$ für alle $l \in \{1, \ldots, M\}$ erreicht wurde. Dieser Ansatz liefert jedoch in keinerlei Hinsicht eine optimale Anzahl an Durchläufen. Den Datenfehler ε berücksichtigend, wird in Isakov [36] oder auch Scherzer [53] das von Morozov vorgestellte Diskrepanzkriterium als Abbruchmaßstab verwendet. Dies besagt, dass solange fortgefahren werden soll, bis (erstmals) die L^2-Norm des Fehlerfunktionals $\mathcal{J}_h$ kleiner oder gleich dem vorliegenden (quadrierten) Datenfehler ε^2 ist, also bis

$$\begin{aligned}\mathcal{J}_h\big(\vec{p}^{\,1,1}_{r_1^*}, \ldots, \vec{p}^{\,1,d_1}_{r_1^*}, \ldots, \vec{p}^{\,M,1}_{r_M^*}, \ldots, \vec{p}^{\,M,d_M}_{r_M^*}\big) \\ \leq \varepsilon^2 < \mathcal{J}_h\big(\vec{p}^{\,1,1}_{r_1}, \ldots, \vec{p}^{\,1,d_1}_{r_1}, \ldots, \vec{p}^{\,M,1}_{r_M}, \ldots, \vec{p}^{\,M,d_M}_{r_M}\big)\end{aligned}$$

für $r_l^* > r_l$, $l = 1, \ldots, M$, gilt. Eine Abwandlung dieses Abbruchkriteriums ist in Bitterlich [8], Kapitel 4.1.4, zu finden. Dieses beruht auf der Tatsache, dass i.A.

$$\lim_{\substack{r_l \to \infty \\ l=1,\ldots,M}} \frac{\mathcal{J}_h\big(\vec{p}^{\,1,1}_{r_1}, \ldots, \vec{p}^{\,1,d_1}_{r_1}, \ldots, \vec{p}^{\,M,1}_{r_M}, \ldots, \vec{p}^{\,M,d_M}_{r_M}\big)}{\mathcal{J}_h\big(\vec{p}^{\,1,1}_{r_1+\Delta r_1}, \ldots, \vec{p}^{\,1,d_1}_{r_1+\Delta r_1}, \ldots, \vec{p}^{\,M,1}_{r_M+\Delta r_M}, \ldots, \vec{p}^{\,M,d_M}_{r_M+\Delta r_M}\big)} \overset{>}{\to} 1$$

gilt, so dass der Algorithmus gestoppt werden kann, sobald

$$1 - \frac{\mathcal{J}_h\big(\vec{p}^{\,1,1}_{r_1+\Delta r_1}, \ldots, \vec{p}^{\,1,d_1}_{r_1+\Delta r_1}, \ldots, \vec{p}^{\,M,1}_{r_M+\Delta r_M}, \ldots, \vec{p}^{\,M,d_M}_{r_M+\Delta r_M}\big)}{\mathcal{J}_h\big(\vec{p}^{\,1,1}_{r_1}, \ldots, \vec{p}^{\,1,d_1}_{r_1}, \ldots, \vec{p}^{\,M,1}_{r_M}, \ldots, \vec{p}^{\,M,d_M}_{r_M}\big)} < \varepsilon^2 \mu_{\text{tol}}$$

für ein vorab festgelegtes $\mu_{\text{tol}} > 0$ nicht mehr erfüllt ist.

Kapitel 4

Numerische Resultate

In den nachfolgenden Abschnitten werden unterschiedliche, sowohl auf virtuellen als auch auf realen Messdaten basierende, Fallstudien betrachtet. Konkret werden in Sektion 4.1 die beiden im Kapitel 2.1.5 vorgestellten Regularisierungsansätze der van Genuchten-Mualem-Parametrisierung der hydraulischen Leitfähigkeit untersucht und miteinander verglichen. In Abschnitt 4.2 wird ein rekursiver Identifizierungsansatz vorgestellt, mit dessen Hilfe die Koeffizienten fixer Nichtlinearitäten durch adaptive Gewichtung des Residuums besser identifiziert werden können. Die aufgeführten Berechnungsbeispiele basieren dabei auf der Identifizierung der durch die van Genuchten-Mualem-Parametrisierungen (2.5) und (2.6) bzw. den Monod-Reaktionsraten (2.37) und (2.38) festgelegten Koeffizienten. In Sektion 4.3 werden schließlich die in Kapitel 3.2.2 vorgestellten formfreien Ansätze zur Identifizierung dreidimensionaler Reaktionsraten angewendet und sowohl untereinander als auch mit entsprechenden Resultaten fixer Parameteridentifizierungen verglichen. Des Weiteren werden optionale Einschränkungen, wie beispielsweise vorgegebene Monotonieeigenschaften, eine lineare dritte Komponente, fest vorgegebene Achsenwerte oder auch eine nichtäquidistante Diskretisierung, vorgestellt und ihre Notwendigkeit für eine erfolgreiche Identifizierung diskutiert.

4.1 Regularisierung der van Genuchten-Mualem-Leitfähigkeitsfunktion

Im Folgenden werden zur Simulation unterschiedlicher Säulenexperimente sowohl die durch (2.5) und (2.7) festgelegte van Genuchten-Mualem-Parametrisierung der hydraulischen Leitfähigkeit als auch die beiden in den Abschnitten 2.1.5.1 und 2.1.5.2 definierten sättigungs- bzw. druckabhängigen Regularisierungen verwen-

det und entsprechend miteinander verglichen. Als Berechnungsgrundlage dient eine eindimensional betrachtete, $L = 40$cm hohe, Laborsäule mit entgegen der Schwerkraft gerichteter Ortsachse z. An dieser Stelle sei bemerkt, dass bei ausreichend homogenem Erdreich und nicht zu klein gewähltem Laborsäulendurchmesser sowohl Randeffekte als auch Auswirkungen in x- und y-Richtung vernachlässigt werden können. Entsprechend stellt, unter den genannten Voraussetzungen, eine eindimensionale Betrachtung des Fließgeschehens keine signifikante Einschränkung dar.

Zu Beginn des Experiments ($t = 0$) ist die Bodenprobe vollständig mit einem Fluid, im Weiteren auch als Wasser bezeichnet, gesättigt. Für $0 < t < T$ wird an der Unterkante der Säule Feuchtigkeit entzogen und damit die Druck- bzw. Sättigungsverteilung verändert. Da die Laborsäule hermetisch verschlossen sein soll, hat der entstehende Unterdruck Auswirkungen auf alle Bereiche der vorliegenden Bodenprobe. Aus diesem Grund werden Druck- und Sättigungsdaten an unterschiedlichen Messstellen $h_1 = 5$cm, $h_2 = 15$cm, $h_3 = 25$cm und $h_4 = 35$cm gesammelt. Vergleiche hierzu auch Abbildung 4.1.

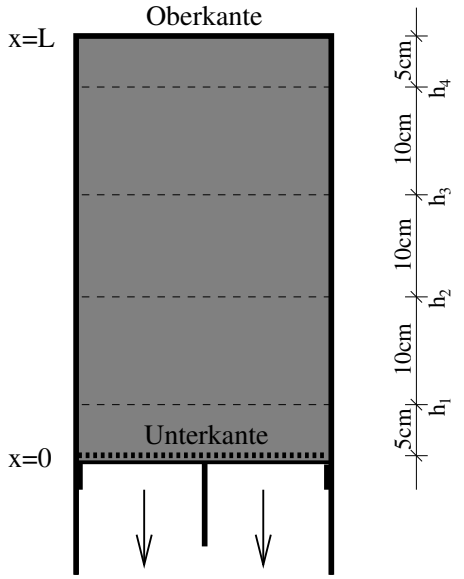

Abbildung 4.1: Laborsäule mit Drucksteuerung an der Bodenplatte

Um den Einfluss einer Regularisierung auf die Lösung des direkten Problems möglichst realitätsnah zu untersuchen, werden im Folgenden neben völlig unterschiedlichen Parameterwerten n, auf realen Messungen basierende Bodendaten herangezogen und die vermeintlich exakte Lösung, welche durch numeri-

sches Lösen der Richards-Gleichung (2.2) unter Verwendung der van Genuchten-Mualem-Parametrisierung gegeben sein soll, mit den Resultaten des regularisierten Ansatzes verglichen. Sei hierfür der betrachtete Zeithorizont $[0,T]$ durch $N+1$ diskrete Zeitpunkte $t_j \in [0,T]$ mit $t_0 = 0$, $t_{j-1} < t_j < t_{j+1}$, $j = 1,\ldots,N-1$, und $t_N = T$ gegeben. Bezeichnen $\psi(h_i,t_j)$ und $\Phi(h_i,t_j)$ die zu t_j, $j=1,\ldots,N$, an den Messstellen h_i, $i = 1,\ldots,4$, numerisch ermittelten Druck- und Sättigungswerte des, auf der van Genuchten-Mualem-Parametrisierung basierenden, Modellansatzes und $\psi_R(h_i,t_j)$ und $\Phi_R(h_i,t_j)$ die entsprechenden Werte unter Verwendung der betrachteten Regularisierung, dann wird im Folgenden der (relative) diskrete Regularisierungsfehler abhängig vom Regularisierungsgrad R durch

$$E(R) = \frac{1}{8}\left(\sum_{i=1}^{4} E_{\Phi,h_i}(R) + \sum_{i=1}^{4} E_{\psi,h_i}(R) \right) \tag{4.1}$$

mit

$$E_{\Phi,h_i}(R) = \sum_{j=1}^{N} \left|\Phi(h_i,t_j) - \Phi_R(h_i,t_j)\right| \Big/ \sum_{j=1}^{N} \Phi(h_i,t_j)\ , \quad i=1,\ldots,4\,,$$

und

$$E_{\psi,h_i}(R) = \sum_{j=1}^{N} \left|\psi(h_i,t_j) - \psi_R(h_i,t_j)\right| \Big/ \Big| \sum_{j=1}^{N} \psi(h_i,t_j)\Big|\ , \quad i=1,\ldots,4\,,$$

verwendet.

Bei den folgenden Fallstudien wurden die van Genuchten Parameter (inbesondere n) so gewählt, dass beim Lösen des direkten Problems (ohne Verwendung einer Regularisierung) numerische Schwierigkeiten auftreten. Diese machen sich durch signifikant längere Rechenzeiten bis hin zum Abbruch des Verfahrens bemerkbar. Die physikalischen Parameter wurden entsprechend realer Bodenproben auf $K_{\text{sat}} = 1.0 \cdot 10^{-3}\,\frac{\text{cm}}{\text{min}}$, $\theta_{\text{sat}} = 0.4$ und $\theta_{\text{res}} = 0.1$ gesetzt (vgl. z.B. van Genuchten [61], Kool, Parker, van Genuchten [39]). Zudem wurde $\alpha = 1.0 \cdot 10^{-2}\,\frac{1}{\text{cm}}$ und $n \in \{1.05, 1.1, 1.25, 1.5\}$ gewählt.

Im Folgenden wird angenommen, dass der initiale, an der Unterkante vorherrschende Druck $\psi(0,0) = 40$ (gemessen in cm) bis zum (End-)Zeitpunkt $T = 1440$ (gemessen in Minuten) linear auf $\psi(0,T) = 0$ verringert wird. Im Falle eines realen Experiments geschieht dies durch ein computergesteuertes Abpumpen des bereits durch die poröse Bodenplatte versickerten und in einem separaten Auffangbehälter angesammelten Fluids. Das Erdreich wird dabei durch eine luft- und

wasserdurchlässige Membran, typischerweise eine poröse Keramikscheibe, an der ursprünglichen Position gehalten.

Zur numerischen Lösung der Richards-Gleichung sind noch entsprechende Randbedingungen zu definieren. Aufgrund des abgedichteten Deckels ergibt sich hierfür an der rechten Seite des betrachteten Gebiets

$$\frac{\partial \psi}{\partial z}(L,t) = 0$$

und an der linken Seite entsprechend dem linearen Druckabfall

$$\psi(0,t) = -2,\bar{7}\cdot 10^{-2}t + 40\,.$$

In den nachfolgenden Tabellen 4.1 und 4.2 wird, abhängig von der gewählten Regularisierung (druck- oder sättigungsabhängig) und dem zugehörigen Regularisierungsgrad R_Φ bzw. R_ψ, der numerische Aufwand zur Lösung des zugehörigen diskreten Gleichungssystems für ausgewählte Parameterwerte n aufgezeigt. Als zugrundegelegte Vergleichswerte dienen hierbei die beiden Größen **tot** (Gesamtanzahl aller zur Lösung notwendiger Newton-Iterationen) und **max** (maximale Anzahl an Newton-Iterationen in einem Zeitschritt). Als Zeitschrittweite wurde $\Delta t = 20\,\text{min}$ gewählt, so dass in der Summe 72 Gleichungssysteme zu lösen waren. Um etwaige Einflüsse der gewählten Diskretisierung (weitestgehend) auszuschließen, wurde das Gebiet Ω sehr fein in $N = 500$ äquidistante Knotenpunkte unterteilt. Schließlich wurde als Abbruchkriterium n.A. die maximale Newton-Iterationszahl auf $i_{\max} = 15$ gesetzt (i.d.R genügen deutlich weniger Iterationsschritte).

Im Fall der sättigungsabhängigen $\mathcal{P}^2$-Regularisierung fällt auf, dass für sehr kleine Parameterwerte n ($n = 1.05$ oder $n = 1.10$) bereits ein sehr kleiner Regularisierungsgrad $R_\Phi = 1.0 \cdot 10^{-5}$ oder $R_\Phi = 1.0 \cdot 10^{-6}$ zu einer abbruchfreien Berechnung führte. Für etwas größere Parameterwerte ($n = 1.25$ und $n = 1.50$) kam es zu keinem Abbruch. Dennoch lieferte auch hier die $\mathcal{P}^2$-Regularisierung i.d.R. einen Geschwindigkeitsvorteil. Für $n \geq 2$ sind keine Unterschiede bzgl. des Berechnungsaufwandes zu finden. Dies war auch nicht zu erwarten, da hier, im Gegensatz zu $1 < n < 2$, die Ableitung $K'(\psi)$ nach oben beschränkt ist.

Anders sieht es bei der $\mathcal{P}^3$-Regularisierung aus. Diese ist definitionsgemäß nur für $n > 2$ als Approximation der van Genuchten-Mualem-Leitfähigkeitsfunktion geeignet. Neben den Approximationsschwierigkeiten dieser Regularisierung für $n \leq 2$ treten jedoch auch weitere numerische Schwierigkeiten auf. Diese machen sich

	$n=1.05$		$n=1.10$		$n=1.25$		$n=1.50$		$n=2.00$		$n=3.00$	
R_Φ	tot	max	tot	max	tot	max	tot	max	tot	max	tot	max
vG	-	-	-	-	156	5	136	11	106	2	81	2
10^{-6}	-	-	215	11	161	7	128	3	106	2	81	2
10^{-5}	217	6	202	7	155	4	125	3	106	2	81	2
10^{-4}	215	4	201	4	148	4	124	2	106	2	81	2
0.001	198	3	182	3	146	3	97	2	106	2	81	2
0.002	137	3	141	3	125	3	97	2	106	2	81	2
0.005	80	2	88	3	85	3	98	2	106	2	81	2
0.10	75	2	81	3	88	3	98	2	106	2	81	2

Tabelle 4.1: Performanz der $\mathcal{P}^2$-Regularisierung

	$n=1.05$		$n=1.10$		$n=1.25$		$n=1.50$		$n=2.00$		$n=3.00$	
R_ψ	tot	max	tot	max	tot	max	tot	max	tot	max	tot	max
vG	-	-	-	-	156	5	136	11	106	2	81	2
-0.1	-	-	-	-	-	-	-	-	106	2	81	2
-0.2	-	-	-	-	-	-	-	-	106	2	81	2
-0.5	-	-	-	-	-	-	173	3	106	2	81	2
-1.0	-	-	349	4	232	4	162	3	106	2	81	2
-2.0	515	10	361	7	216	4	151	3	106	2	81	2
-5.0	513	12	139	7	184	4	140	2	106	2	81	2
-10.0	338	8	213	5	138	3	113	2	106	2	81	2

Tabelle 4.2: Performanz der $\mathcal{P}^3$-Regularisierung

vor allem bei einem zu klein gewählten Regularisierungsgrad R_ψ bemerkbar. Insbesondere fiel auf, dass das Newton-Verfahren für $|R_\psi| \leq -1$ nicht konvergierte. Je größer der Parameterwert n gewählt wurde, desto weniger trat dieses Problem auf. Dennoch war der Rechenaufwand im Vergleich zum unregularisierten Gleichungssystem höher. Folglich ist die $\mathcal{P}^3$-Regularisierung auch bzgl. der Behebung numerischer Schwierigkeiten (für $n < 2$) äußerst kontraproduktiv. Bleibt letztendlich die Frage, ob es überhaupt eine sinnvolle Verwendung für diese Regularisierung gibt. Hierzu muss überlegt werden, inwieweit die nichtdifferenzierbare Stelle der druckabhängigen (nicht sättigungsabhängigen) van Genuchten-Mualem-Leitfähigkeitsfunktion (2.8) für $n \leq 2$ physikalisch motiviert ist. Sollte dies nicht der Fall sein, so hat die $\mathcal{P}^3$-Regularisierung neben den aufgeführten Nachteilen, den Vorteil eines differenzierbaren Übergangs. Andernfalls sollte von diesem Ansatz Abstand genommen werden.

Neben den bereits geführten Untersuchungen bzgl. des Rechenaufwandes interes-

siert vor allem die Frage nach dem Approximationsfehler beider Ansätze. Hierzu wird im Folgenden, abhängig von der verwendeten Regularisierung, der diskrete Regularisierungsfehler (4.1) verwendet. Kann (aufgrund eines Programmabbruchs) keine unregularisierte Lösung berechnet werden, so wird alternativ die, mit dem kleinsten Regularisierunsgrad R_Φ vorliegende, Lösung der $\mathcal{P}^2$-Regularisierung herangezogen. In den Tabellen 4.3 und 4.4 finden sich hierzu die numerisch ermittelten Resultate.

	$E(R_\Phi)$					
R_Φ	$n=1.05$	$n=1.10$	$n=1.25$	$n=1.5$	$n=2.0$	$n=3.0$
10^{-6}	-	Ref.-Dat.	$7.5\cdot10^{-6}$	$9.0\cdot10^{-5}$	$3.6\cdot10^{-7}$	$4.3\cdot10^{-8}$
10^{-5}	Ref.-Dat.	$1.4\cdot10^{-4}$	$2.0\cdot10^{-4}$	$7.8\cdot10^{-5}$	$3.5\cdot10^{-6}$	$4.2\cdot10^{-7}$
10^{-4}	$2.5\cdot10^{-3}$	$2.6\cdot10^{-3}$	$2.1\cdot10^{-3}$	$2.2\cdot10^{-4}$	$3.5\cdot10^{-5}$	$4.0\cdot10^{-6}$
0.001	0.0725	0.0544	0.0320	$2.9\cdot10^{-3}$	$3.4\cdot10^{-4}$	$3.4\cdot10^{-5}$
0.002	0.1934	0.1420	0.0741	$6.1\cdot10^{-3}$	$6.7\cdot10^{-4}$	$6.0\cdot10^{-5}$
0.005	0.5576	0.4718	0.2253	0.0162	$1.6\cdot10^{-3}$	$1.2\cdot10^{-4}$
0.010	0.7733	0.9321	0.4902	0.0322	$2.7\cdot10^{-3}$	$1.7\cdot10^{-4}$

Tabelle 4.3: Regularisierungsfehler der $\mathcal{P}^2$-Regularisierung

	$E(R_\psi)$					
R_ψ	$n=1.05$	$n=1.10$	$n=1.25$	$n=1.5$	$n=2.0$	$n=3.0$
-0.1	-	-	-	-	$1.4\cdot10^{-7}$	$5.6\cdot10^{-12}$
-0.2	-	-	-	-	$5.6\cdot10^{-7}$	$1.8\cdot10^{-11}$
-0.5	-	-	-	$3.0\cdot10^{-4}$	$3.6\cdot10^{-6}$	$1.3\cdot10^{-10}$
-1.0	-	0.0309	0.0244	$1.0\cdot10^{-3}$	$1.4\cdot10^{-5}$	$5.5\cdot10^{-10}$
-2.0	0.0552	0.0855	0.0679	$3.3\cdot10^{-3}$	$5.8\cdot10^{-5}$	$2.3\cdot10^{-9}$
-5.0	0.2082	0.3216	0.2633	0.0143	$3.6\cdot10^{-4}$	$2.0\cdot10^{-8}$
-10.0	0.4865	0.7669	0.6586	0.0399	$1.3\cdot10^{-3}$	$1.9\cdot10^{-7}$

Tabelle 4.4: Regularisierungsfehler der $\mathcal{P}^2$-Regularisierung

Unabhängig von der Art der verwendeten Regularisierung wird der Regularisierungsfehler umso kleiner, je größer der Parameterwert n ist. Für große n ($n \geq 2$) liefern damit beide Ansätze sehr gute Approximationswerte. Im Fall $n < 2$ bietet, insbesondere für sehr kleine Parameterwerte, (wie erwartet) nur die $\mathcal{P}^2$-Regularisierung mit klein gewählten $R_\Phi \leq 1.0 \cdot 10^{-4}$ einen vertretbar geringen Fehler. Sowohl die $\mathcal{P}^3$-Regularisierung als auch ein zu groß gewählter Regularisierungsgrad R_Φ führen zu einer schlechten Approximation.

Zusammenfassend stellt die vorgestellte $\mathcal{P}^2$-Regularisierung mit Regularisierungsgrad $R_\Phi = 1.0 \cdot 10^{-5}$ oder $R_\Phi = 1.0 \cdot 10^{-4}$ eine attraktive Approximation der (unregularisierten) van-Genuchten-Mualem-Leitfähigkeitsfunktion dar. Der Aufwand zum Lösen des auf der Richards-Gleichung basierenden Gleichungssystems wird verringert, ohne dass die notwendige Rechengenauigkeit darunter leidet.

4.2 Rekursive Parameteridentifizierung

Dieser Abschnitt beschäftigt sich mit der Identifizierung festgelegter Parametrisierungen. Inbesondere wird ein rekursiver Ansatz vorgestellt, mit dessen Hilfe die unbekannten Parameter besser bestimmt werden können. Als numerische Fallstudien dienen neben einem virtuellen Beispiel, bei dem die optimale Lösung bekannt ist und entsprechend qualitative Aussagen bzgl. der neuen Methoden getroffen werden können, auch eine auf experimentellen Messungen basierende Simulation einer realen Laborsäule. Die zu identifizierenden Nichtlinearitäten sind dabei durch die van Genuchten-Mualem-Parametrisierungen (2.5) und (2.6) bzw. den Monod-Reaktionsraten (2.37) und (2.38) gegeben, jedoch in keiner Weise auf diese beschränkt.

4.2.1 Rekursiv gewichtetes Residuum

Erste Parameteridentifizierungen ungesättigter Flussprobleme finden sich in Zachmann, DuChateau, Klute [66], Dane, Hruska [18] und Hornung [33]. Bereits hier wurde die Schlechtgestelltheit des zugehörigen Problems erkannt und auf die entsprechenden Schwierigkeiten bei der Bestimmung der unbekannten Koeffizienten hingewiesen. Insbesondere die fehlende Eindeutigkeit der inversen Lösung motivierte die Verwendung von Druck- (vgl. z.B. Toorman, Wierenga, Hills [58] oder Eching, Hopmans [20]) und Sättigungswerten (vgl. z.B. van Dam, Stricker, Verhoef [59]) an unterschiedlichen Messstellen. Spätere Untersuchungen beschäftigten sich detaillierter mit dem experimentellen Setup. So wurde vor allem erkannt, dass eine sogenannte Multi-Step-Vorgehensweise, bei der der vorherrschende Druck in mehreren kleinen Schritten verändert wird, deutlich bessere Ergebnisse lieferte als Experimente mit nur einer großen Druckänderung. Insbesondere bei der van Genuchten-Mualem-Parametrisierung konnten damit erstmals die hydraulischen Parameter ausreichend gut identifiziert werden (vgl. Eching, Hopmans, Wendroth [21], van Dam, Stricker, Droogers [60], Crescimanno, Iovino [17] und Zurmühl [67]). In Durner, Schultze, Zurmühl [19] wurde schließlich eine Zusammenfassung

der bisherigen Forschungsarbeiten zur inversen Modellierung bei Säulenflussexperimenten bereitgestellt.

Wie bereits im Kapitel 3.1 erläutert, wird bei der inversen Modellierung typischerweise ein (gewichtetes) Fehlerfunktional der Art

$$\mathcal{J}_h(p) = \sum_{k=1}^{\kappa} \Lambda_k(E_k) \sum_{i=1}^{n_k} \Lambda_{k,i}(w_{h,\varepsilon}) \sum_{j=1}^{m_{k,i}} \Big(\big(w_h\big)_{k,i,j}(p) - (w_{h,\varepsilon})_{k,i,j} \Big)^2 \tag{4.2}$$

verwendet. Dabei beschreibt $\kappa \in \mathbb{N}$ die Anzahl an durchgeführten Experimenten, n_k, $k = 1, \ldots \kappa$, die jeweilige Anzahl an unterschiedlichen Messungen und $m_{k,i}$, $k = 1, \ldots \kappa$, $i = 1, \ldots, n_k$, die zugehörige Anzahl an diskreten Messzeitpunkten. Durch $\Lambda_k(E_k)$, $k = 1, \ldots, \kappa$ und $\Lambda_{k,i}(w_{h,\varepsilon})$, $k = 1, \ldots, \kappa$, $i = 1, \ldots, n_k$, ist eine Gewichtung der einzelnen Experimente und Messungen möglich, so dass explizit deren Einfluss auf das zu minimierende Zielfunktional manipuliert werden kann. Im einfachsten Falls sind alle Faktoren gleich eins und (4.2) in nach Bemerkung 3.11 als gewöhnliches **Ordinary Least Quares (OLS)**-Problem

$$\mathcal{J}_h(w_h) = \sum_{k=1}^{\kappa} \sum_{i=1}^{n_k} \sum_{j=1}^{m_{k,i}} \Big(\big(w_h\big)_{k,i,j}(p) - (w_{h,\varepsilon})_{k,i,j} \Big)^2 =: \sum_{j=1}^{n} \Big(\big(w_h\big)_{k,i,j}(p) - (w_{h,\varepsilon})_{k,i,j} \Big)^2$$

darstellbar. Nach Gribb [28] und Kool, Parker, van Genuchten [39] ist es jedoch vorteilhaft, die einzelnen Messungs- bzw. Beobachtungsdaten mit dem quadrierten Kehrbruch der jeweiligen Mittelwerte zu gewichten, so dass

$$\Lambda_{k,i}(w_{h,\varepsilon}) = \left(\frac{1}{m_{k,i}} \sum_{j=1}^{m_{k,i}} (w_{h,\varepsilon})_{k,i,j} \right)^{-2}, \tag{4.3}$$

$k = 1, \ldots, \kappa$, $i = 1, \ldots, n_k$, gilt. Falls mehrere Experimente durchgeführt werden, also $\kappa > 1$ ist, können zusätzlich (optional) noch die Gewichtungsfaktoren $\Lambda_k(E_k)$ verwendet werden. Dies kann gerade dann genutzt werden, wenn den einzelnen Versuchsaufbauten aufgrund vorhandener Störeinflüsse oder anderer Beweggründe eine unterschiedlich starke Bedeutung zugesprochen wird. Im Folgenden wird hierauf jedoch nicht weiter eingegangen.

Trotz aller bisherigen Bemühungen bzgl. des experimentellen Designs und der vorgeschlagenen Gewichtungen terminiert das verwendete Minimierungsverfahren meist nur in einem lokalen Minimum. Die ermittelten Optimalwerte sind oftmals weit entfernt von globalen Eigenschaften und entsprechend unbrauchbar. Ursachen hierfür gibt es viele. Neben der Schlechtgestelltheit des zugehörigen inversen

Problems liegt eine hohe Sensitivität in Datenfehlern und dem Versuchsaufbau vor. Dies hat zur Folge, dass unterschiedliche Parameterwerte zu äußerst ähnlichen Ergebnissen, z.B. in Form von Durchbruchskurven, führen können.

Nachfolgend wird eine weitere Möglichkeit der Gewichtung vorgestellt und numerisch untersucht. Es sei vorab erwähnt, dass dieser Ansatz zwar durch die Sensitivitätsanalyse motiviert ist, sein Erfolg jedoch in keiner Weise mathematisch bewiesen werden konnte. Die Untersuchungen beziehen sich daher ausschließlich auf numerische Berechnungen. Um dennoch eine möglichst aussagekräftige Bewertung zu erlangen, werden sowohl auf ungestörten und gestörten virtuellen Experimenten als auch auf realen Messungen basierende Identifizierungsbeispiele herangezogen. Die erzielten Resultate waren dabei äußerst vielversprechend, so dass trotz fehlender Beweisführung diese neue, rekursive Minimierungsmethode an dieser Stelle vorgestellt wird.

In den bereits vorgestellten Überlegungen wurden mögliche Einflüsse durch unterschiedliche Experimente und Messungen durch die Gewichtungsfaktoren $\Lambda_k(E_k)$, $k=1,\dots,\kappa$ und $\Lambda_{k,i}(w_{h,\varepsilon})$, $k=1,\dots,\kappa$, $i=1,\dots,n_k$, berücksichtigt. Unbeachtet blieben jedoch bisher vollständig die Sensitivitäten der einzelnen Messstellen und ihre Auswirkung auf die erzielte Lösung. Im Folgenden wird daher versucht eine geeignete Gewichtung des Residuums zu finden, um das zugehörige System bzgl. dieser Sensitivitäten auszugleichen und entsprechende Konvergenzvorteile zu erzielen. Sei hierzu angenommen, dass für alle Beobachtungen $w_{k,i}\in L^2(D_{k,i})$, $k=1,\dots,\kappa$, $i=1,\dots,n_k$, gilt. $D_{k,i}$ beschreibt dabei gerade diejenige Teilmenge des Raum/Zeit-Quaders $\Omega\times[0,T]$, auf der die jeweilige Beobachtung definiert ist. Gesucht sind damit Gewichtungsfunktionen $\lambda_{k,i}(p)\in L^\infty$, $k=1,\dots,\kappa$, $i=1,\dots,n_k$, so dass das (kontinuierliche) Fehlerfunktional

$$\mathcal{J}(p)=\sum_{k=1}^{\kappa}\Lambda_k(E_k)\sum_{i=1}^{n_k}\Lambda_{k,i}(w_\varepsilon)\Big\|\lambda_{k,i}(p)\Big(w_{k,i,j}(p)-(w_\varepsilon)_{k,i,j}\Big)\Big\|_2^2 \tag{4.4}$$

zum globalen, oder zumindest zu einem besseren, (brauchbaren) lokalen Minimum führt. Doch wie sind die Gewichtungen $\lambda_{k,i}$ zu wählen und wie kann mit der in (4.4) vorliegenden, ungeeignet hohen Nichtlinearität umgegangen werden? Beide Fragen werden im Folgenden für den diskreten Fall beantwortet. Zunächst seien hierfür die Sensitivitätsmatrizen

$$S_{k,i}=D_p\Big((\mathcal{B}_h)_{k,i}\big((\mathcal{A}_h)_{k,i}(p),p\big)\Big)\in\mathbb{R}^{m_{k,i}\times r}\,,\quad p\in\mathbb{P}^r\,,$$

betrachtet. Da i.A. $m_{k,i}\gg r$ (inbesondere $m_{k,i}\neq r$) gilt, besitzt $S_{k,i}$ keine her-

kömmliche Inverse. Mit Hilfe der folgenden Definition kann jedoch Abhilfe geschaffen werden.

Definition 4.1
Sei $A \in \mathbb{R}^{m\times n}$ *und* $b \in \mathbb{R}^n$ *gegeben. Dann ist*

- *die* ***pseudonormale Lösung*** *definiert durch* $x_b^+ \in L(b)$ *mit* $\left\|x_b^+\right\|_2 < \left\|x\right\|_2$ *für alle*

$$x \in L(b) := \left\{ \tilde{x} \in \mathbb{R}^n \;\middle|\; \|A\tilde{x} - b\|_2 = \inf_{x\in\mathbb{R}^n} \|Ax - b\|_2 \right\} \quad \textit{und}$$

- *die* ***Pseudoinverse von A*** *definiert durch die lineare Abbildung*

$$\begin{aligned} A^+ \colon \mathbb{R}^m &\to \mathbb{R}^n, \\ b &\mapsto A^+(b) := (A^T A)^{-1} A^T b = x_b^+ \,. \end{aligned}$$

Analog zu den Überlegungen bzgl. unterschiedlich großer Messdatenmittelwerte wird nun jede einzelne Messwertabweichung separat durch den parameterwertabhängigen Faktor

$$(\lambda_{k,i})_j(p) := \sqrt{\sum_{l=1}^{r} \left(\left(S_{k,i}^+\right)_{lj}(p) \right)^2}, \;\; j = 1, \ldots, m_{k,i} \,, \tag{4.5}$$

gewichtet. Da jeder Eintrag der j-ten Zeile, $j \in 1, \ldots, m_{k,i}$, von $S_{k,i}$ die Sensitivität des j-ten Messwertes von einem der r Parameterwerte angibt, enthält die j-te Spalte von $S_{k,i}^+ \in \mathbb{R}^{r\times m_{k,i}}$ eine Art Umkehrwerte dieser Sensitivitäten. Da wiederum $(\lambda_{k,i})_j(p)$ durch die Quadratwurzel der Spaltensummennorm der quadrierten Einträge der pseudoinversen Matrix definiert ist, werden entsprechend durch diese neuen Gewichtungen die diskreten Messstellen bzgl. ihrer Sensitivitäten gemittelt.

Bleibt das Problem mit der hohen Nichtlinearität des verwendeten Fehlerfunktionals. Hierzu wird im Folgenden eine rekursive Approximation verwendet. Statt einer kontinuierlich vom Parametersatz p abhängigen Gewichtung $(\lambda_{k,i})(p)$ wird lediglich $(\lambda_{k,i})(p_0)$, mit initialer Parameterwertschätzung p_0, herangezogen, so dass das diskrete Fehlerfunktional

$$\mathcal{J}_h(p; p_0) = \sum_{k=1}^{\kappa} \Lambda_k(E_k) \sum_{i=1}^{n_k} \Lambda_{k,i}(w_{h,\varepsilon}) \sum_{j=1}^{m_{k,i}} (\lambda_{k,i})_j(p_0) \Big(\big(w_h\big)_{k,i,j}(p) - \big(w_{h,\varepsilon}\big)_{k,i,j} \Big)^2 \tag{4.6}$$

zu lösen ist. Sobald das zugehörige (lokale) Minimum

$$p_{\text{opt}} = \arg\min_{p \in P} \mathcal{J}_h(p; p_0)$$

ermittelt wurde, wird eine neue Gewichtung $(\lambda_{k,i})(p_{\text{opt}})$ berechnet, entsprechend ein neues Fehlerfunktional bestimmt und der Optimierungsprozess erneut gestartet. Auf diese Weise entsteht das nachfolgende, im Nassi-Shneiderman-Diagramm dargestellte rekursive Optimierungsverfahren.

Rekursive Parameteridentifizierung — Algorithmus

Setze rek = 0

Wähle initiale Parameterschätzung p_0

Wiederhole solange bis ein Abbruchkriterium erfüllt ist

FOR $k = 1, \ldots, \kappa$, $i = 1, \ldots, n_k$, DO

Berechne Sensitivitätsmatrix $S_{k,i} = D_p\Big((\mathcal{B}_h)_{k,i}\big((\mathcal{A}_h)_{k,i}(p_{\text{rek}}), p_{\text{rek}} \big) \Big)$

Berechne pseudoinverse Matrix $S_{k,i}^{+}$ von $S_{k,i}$

FOR $j = 1, \ldots, m_{k,i}$ DO

Berechne Gewichtungsfaktor $(\lambda_{k,i})_j(p_{\text{rek}}) = \sqrt{\sum_{l=1}^{r} \Big(\big(S_{k,i}^{+}\big)_{lj}(p) \Big)^2}$

Berechne (lokales) Optimum $p_{\text{opt}} = \arg\min_{p \in P} \mathcal{J}_h(p; p_{\text{rek}})$

Überpüfe alle Abbruchkriterien. Ist eines erfüllt?

Ja: Beende den Optimierungsalgorithmus

Nein: Erhöhe rek ← rek + 1

Nein: Setze $p_{\text{rek}} = p_{\text{opt}}$

Es sei an dieser Stelle nochmals betont, dass weder die Konvergenz dieses Algorithmusses noch der Erhalt geeigneter Minima mathematisch bewiesen wurde. In den nachfolgenden Abschnitten werden allerdings anhand unterschiedlicher Identifizierungsprobleme, welche sowohl auf virtuellen (mit ungestörten und störungsbehafteten Daten) als auch auf realen Messdaten basieren, die gewünschten Konvergenzeigenschaften heuristisch aufgezeigt. Insbesondere unter abwechselnder Verwendung des gewichteten und ungewichteten Fehlerfunktionals wurde in allen untersuchten Fallstudien die Lage des jeweils ermittelten Optimums verbessert. Ob tatsächlich auch das globale Minimum ermittelt wird, ist jedoch auch weiterhin noch von vielen Faktoren, wie beispielsweise die zur Verfügung gestellten Messdaten und das experimentelle Setup, abhängig.

4.2.2 Virtuelles Experiment

Diese erste Fallstudie basiert auf einem virtuellen Transport- und Schadstoffabbau-Experiment, welches mit Hilfe des in Abschnitt 2.2.3.1 vorgestellten dualen Monod-Modells simuliert wird. Als Grundlage dient eine eindimensional betrachtete Laborsäule mit Höhe $L = 20\text{cm}$. Die Ortsachse soll in Gravitationsrichtung verlaufen und ihren Ursprung an der Oberkante der Bodenprobe haben. Zu Beginn des Experiments ($t_0 = 0$) sei der als homogen betrachtete Boden vollständig mit einem Fluid, im Folgenden auch vereinfachend als Wasser bezeichnet, gesättigt ($\theta = 0.33$) und ein kontaminationsfreier (spezifischer) Fluss $\boldsymbol{q}(x,0) = 1.0\,\frac{\text{cm}}{\text{d}}$ in Richtung der Schwerkraft angelegt. Des Weiteren liegen eine schadstoffabbauende immobile Biomasse der Konzentration $c_B(x,0) = 3.0 \cdot 10^{-3}\,\frac{\mu\text{g}}{\text{cm}^3}$ und der zum biologischen Abbau notwendige mobile Elektronen-Akzeptor $c_A(x,0) = 2.0\,\frac{\mu\text{g}}{\text{cm}^3}$ gleichmäßig im Boden vor. Der Schadstoff selbst wird erst für $t > 0$ dem einfließenden Wasser und damit dem Boden zugeführt, so dass eingangs die Konzentration $c_D(x,0) = 0.0\,\frac{\mu\text{g}}{\text{cm}^3}$ vorliegt. Die Diffusionskoeffizienten der beiden mobilen Spezies sind mit $D_D = 7.4 \cdot 10^{-5}\,\frac{\text{cm}^2}{\text{d}}$ und $D_D = 2.0 \cdot 10^{-7}\,\frac{\text{cm}^2}{\text{d}}$ festgelegt.

Um ein möglichst gutes Identifzierungsresultat zu erzielen, erfolgt die Einspeisung des mobilen Elektronen-Donators während eines (simulierten) 75-tägigen Experiments ($t_{end} = 75\text{d}$) mit unterschiedlichen Intensitätsstufen. So wird zeitlich versetzt zwischen unkontaminierten Zuflusszyklen der Schadstoff derart beigemischt, dass zeitweise die anliegenden Konzentrationen in Höhe von $c_D(0,t) = 1.0\,\frac{\mu\text{g}}{\text{cm}^3}$, $c_D(0,t) = 1.5\,\frac{\mu\text{g}}{\text{cm}^3}$ bzw. $c_D(0,t) = 2.0\,\frac{\mu\text{g}}{\text{cm}^3}$ vorliegen. Des Weiteren wird auch die Konzentration des Elektronen-Akzeptors im Zufluss manipuliert, so dass der ursprüngliche Wert $c_A(0,t) = 2.0\,\frac{\mu\text{g}}{\text{cm}^3}$ im Laufe der Simulation auf $c_A(0,t) = 3.0\,\frac{\mu\text{g}}{\text{cm}^3}$

angehoben wird. Die genaue zeitliche Abfolge beider Konzentrationswerte kann in Abbildung 4.2 nachgesehen werden. Schließlich wird, basierend auf den Monod-Parametern wie sie in Schirmer [54] angegeben sind, das Säulenexperiment numerisch simuliert und der ermittelte Ausfluss des Schadstoffes als auch der des Elektronen-Akzeptors an der Unterkante der (virtuellen) Laborsäule als zeitabhängige Funktionen, den sogenannten Durchbruchskurven $g_D(t) := c_D(20,t)$ und $g_A(t) := c_A(20,t)$, $t \in [0,75]$, erfasst und für die späteren Identifizierungsaufgaben als ungestörte oder auch mit (künstlichen) Messfehlern behaftete Messdaten zur Verfügung gestellt. Die zur Simulation verwendete Zeitschrittweite wurde dabei auf $\Delta t = 0.2$d gesetzt, so dass in der Summe $m = 376$ diskrete (Mess-)Zeitpunkte vorliegen. In Abbildung 4.3 finden sich neben den numerisch ermittelten (exakten) Durchbruchskurven jeweils exemplarisch zwei unterschiedlich stark gestörte Datensätze, wie sie im Folgenden auch zur Identifizierung der gesuchten Parameterwerte verwendet werden.

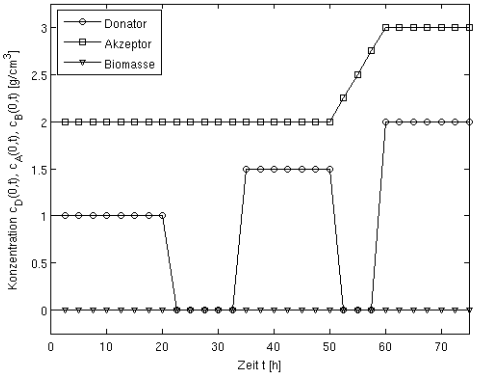

Abbildung 4.2: Einflusskonzentrationen der Monod-Modellkomponenten

Unter Verwendung des, im vorangegangenen Abschnitt definierten, adaptiven Optimierungsverfahrens ist das zu minimierende diskrete Fehlerfunktional durch

$$\mathcal{J}_h(p;p_0) = \Lambda_D \sum_{j=1}^{m} (\lambda_D)_j(p_0)\Big(c_D\big(L,(j-1)\Delta t\big) - g_D((j-1)\Delta t)\Big)^2 + \Lambda_A \sum_{j=1}^{m} (\lambda_A)_j(p_0)\Big(c_A\big(L,(j-1)\Delta t\big) - g_A((j-1)\Delta t)\Big)^2$$

mit den in (4.3) und (4.5) definierten Gewichtungsfaktoren gegeben. Damit ein fairer Vergleich der vorgestellten, adaptiven Optimierungsmethode mit dem ursprünglichen OLS-Problem möglich ist, werden beide Ansätze rekursiv gestar-

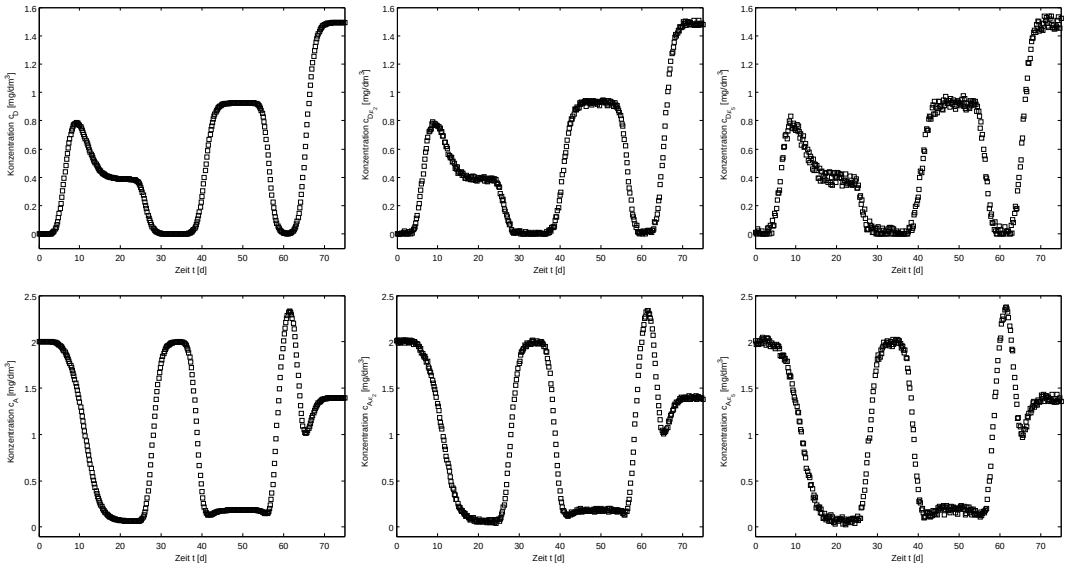

Abbildung 4.3: Durchbruchskurven (ungestörte und mit Messfehlern behaftet)

tet. Im erstgenannten Fall werden die durch die pseudoinverse Sensitivitätsmatrix festgelegten Gewichtungsfaktoren verwendet, im gewöhnlichen OLS-Ansatz nicht. Interessanterweise führt auch ein mehrfacher Durchlauf des ungewichteten OLS-Ansatzes ebenfalls zu besseren Identifizierungsresultaten. Dieses ist auf die verwendeten Abbruchkriterien zurückzuführen, welche insbesondere durch

- maximale Anzahl an SQP-Iterationen $MaxIt = 50$,
- maximale Anzahl an Funktionsauswertungen während einer Liniensuche $MaxFunc = 10$,
- vorgegebene Berechnungsgenauigkeit $StopCrit = 1.0 \cdot 10^{-10}$

gegeben sind.

In einem ersten Identifizierungsansatz sollen zunächst die fünf Monod-Parameter K_{M_D}, K_{I_D}, K_{M_A}, $\mu_{\max}$ und Y, im Übrigen auch durch p^l, $l = 1, \ldots, 5$, bezeichnet, bestimmt werden. Als Startwerte wird jeweils völlig unbedarft und willkürlich $p^l = 1.0$, $l = 1, \ldots, 5$, (mit entsprechender Einheit) verwendet. Auch die Zulässigkeitsbereiche der einzelnen Parameter sind mit $p^l \in [0.0\ 5.0]$, $l = 1, \ldots, 5$, so gewählt, dass die erzielten Identifizierungsresultate möglichst nicht (direkt) durch künstlich erzeugte Einschränkungen vereinfacht bzw. verfälscht werden. In den Tabellen 4.5 und 4.6 finden sich die ermittelten Ergebnisse. In beiden Fällen wird

das globale Minimum gefunden, doch der notwendige Rechenaufwand unterscheidet sich signifikant. Während der gewöhnliche OLS-Ansatz 93 rekursive Aufrufe und 662 SQP-Iterationen benötigt, sind es bei dem neuen, adaptiv gewichteten Verfahren nur 50 Aufrufe mit 372 Iterationen. Des Weiteren ist zu bemerken, dass trotz der geringeren Rechenzeit die Genauigkeit von $1.5{\cdot}10^{-9}$ auf $4.8{\cdot}10^{-11}$ angehoben wurde (in beiden Fällen wurde das ungewichtete Fehlerfunktional angegeben). Alternativ zur Auswertung des Residuums ist jeweils auch der gemittelte relative Fehler

$$\varepsilon_{\text{rel}} := \frac{1}{5}\sum_{l=1}^{5}\frac{\left|p^l_{\text{opt}} - p^l_{\text{exakt}}\right|}{p^l_{\text{exakt}}}$$

mit den erzielten Optimalwerten p^l_{opt}, $l=1,\ldots,5$, und den (in diesem virtuellen Beispiel) vorliegenden exakten Parameterwerten p^l_{exakt}, $l=1,\ldots,5$, angegeben. Auffällig ist, dass die ermittelten Minima des ungewichteten OLS-Ansatzes zu Beginn näher am globalen Optimum sind, jedoch nach 24 Durchläufen das adaptive Verfahren bessere Resultate lieferte. Schließlich terminieren beide Ansätze mit $\varepsilon_{\text{rel}} = 1.0{\cdot}10^{-4}$ bzw. $\varepsilon_{\text{rel}} = 1.2{\cdot}10^{-5}$.

	exakt	start	OLS	Rek 5	Rek 10	Rek 20	Rek 30	Rek 40	Rek ≥ 50
K_{MD}	0.79	1.0	0.7029	0.2514	0.4712	0.7392	0.7898	0.78999	0.78998
K_{ID}	0.50	1.0	2.6177	1.3693	0.8548	0.5320	0.5001	0.50001	0.50001
K_{MA}	0.10	1.0	0.6647	0.3108	0.1014	0.0988	0.0999	0.09999	0.10000
μmax	4.13	1.0	2.6359	2.3460	2.8093	3.9320	4.1290	4.12993	4.12993
Y	0.52	1.0	2.6652	0.5420	0.5473	0.5217	0.5200	0.52000	0.52000
Funk			55.577	4.5448	0.1993	$5.7{\cdot}10^{-4}$	$7.1{\cdot}10^{-7}$	$4.4{\cdot}10^{-9}$	$4.8{\cdot}10^{-11}$
ε_{rel}			2.8960	1.0005	0.2999	0.0343	$3.4{\cdot}10^{-4}$	$3.0{\cdot}10^{-5}$	$1.2{\cdot}10^{-5}$
Σ Iterat			6	39	70	175	248	328	372

Tabelle 4.5: WOLS-Identifizierung: 5 Parameter, keine Störung

	exakt	start	OLS	Rek 5	Rek 10	Rek 20	Rek 30	Rek 50	Rek ≥ 93
K_{MD}	0.79	1.0	0.7029	0.3360	0.6015	0.7770	0.7811	0.78859	0.78989
K_{ID}	0.50	1.0	2.6177	1.0447	0.6532	0.5082	0.5055	0.50089	0.50007
K_{MA}	0.10	1.0	0.6647	0.1046	0.0913	0.0991	0.0995	0.09989	0.09999
μmax	4.13	1.0	2.6359	2.3870	3.3549	4.0758	4.0935	4.12400	4.12956
Y	0.52	1.0	2.6652	0.5619	0.5315	0.5206	0.5204	0.52007	0.52001
Funk			55.577	0.6377	$1.1{\cdot}10^{-2}$	$2.4{\cdot}10^{-5}$	$1.0{\cdot}10^{-5}$	$3.4{\cdot}10^{-7}$	$1.5{\cdot}10^{-9}$
ε_{rel}			2.8960	0.4425	0.1684	0.0112	$7.4{\cdot}10^{-3}$	$3.4{\cdot}10^{-3}$	$1.0{\cdot}10^{-4}$
Σ Iterat			6	31	92	169	238	374	652

Tabelle 4.6: OLS-Identifizierung: 5 Parameter, keine Störung

Ergänzend wird in den Abbildungen 4.4.a und 4.4.b die rekursive Veränderung

des, auf dem adaptiv gewichteten Ansatz basierenden, (ungewichteten) Residuums und des relativen Fehlers durch semilogarithmische Graphen dargestellt. Neben kleineren Identifzierungsplateaus mit geringen Schwankungen des jeweils betrachteten Funktionswertes ist vor allem die kontinuierliche Abnahme beider Graphen deutlich erkennbar. Dies bestätigt insbesondere die Vermutung, dass die durch (4.3) definierte Gewichtung sinnvoll gewählt ist und keinem zufälligen Muster nachgeht.

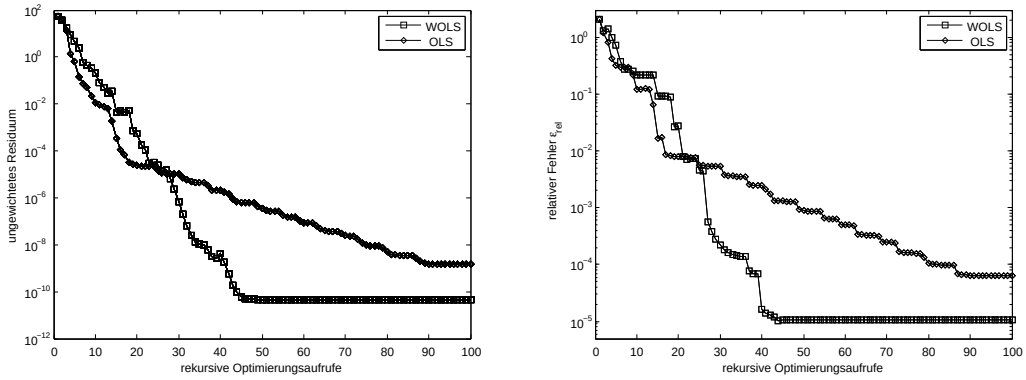

Abbildung 4.4: Rekursive Verbesserung des ungestörten Identifizierungsproblems

In den nachfolgenden Tabellen 4.7 und 4.8 bzw. 4.9 und 4.10 finden sich die ermittelten Identifizierungsresultate beider Ansätze unter Verwendung gestörter Messdaten. Dabei wurde jeder einzelne Messwert $(g_i)_j$, $i \in \{D, A\}$, $j = 1, \ldots, 376$, mit einer zufälligen Störung $|\varepsilon_{i,j}| \in [0, 0.02]$ bzw. $|\varepsilon_{i,j}| \in [0, 0.05]$ behaftet, so dass

$$(g_{\varepsilon,i})_j = \max\left\{(g_i)_j + \varepsilon_{i,j}, 0\right\}, \quad i \in \{D, A\}, \quad j = 1, \ldots, 376, \tag{4.7}$$

für die Identifizierungsaufgabe zur Verfügung stand. Wieder zeigt sich, dass der gewöhnliche OLS-Ansatz nur in den ersten Durchläufen bessere Ergebnisse liefert und am Ende auch hier das adaptive Verfahren näher am globalen Minimum terminiert. Mehrfache Identifizierungsversuche mit jeweils (auf gleichem Niveau) zufällig gestörten Messdaten zeigten jedoch, dass die ermittelten Identifizierungsresultate, unabhängig vom gewählten Ansatz (OLS und WOLS), stark abhängig von der Störung sind. Je größer die verwendete Störung war, desto unterschiedlicher waren auch die erzielten Identifizierungsresultate, so dass mit steigender Messtoleranz mehr und mehr der Zufall über die Qualität des Ergebnisses entschied. Schließlich wurde festgestellt, dass der gewichtete Ansatz im Vergleich zur ungewichteten OLS-Identifizierung, umso eher ein besseres Ergebnis liefert, je kleiner die Störung ist. Bei zu großen Messschwankungen konnte kein signifikanter Unterschied festgestellt werden. Sowohl das gewichtete als auch ungewichtete

Fehlerfunktional führte in diesen Fällen teilweise zu völlig unbrauchbaren Resultaten.

	exakt	start	OLS	Rek 3	Rek 6	Rek 9	Rek 12	Rek ≥ 14
K_{MD}	0.79	1.0	0.7020	0.3537	0.2914	0.4263	0.7893	0.7893
K_{ID}	0.50	1.0	2.6155	2.1866	1.1782	0.8746	0.5009	0.5009
K_{MA}	0.10	1.0	0.6637	0.6260	0.2626	0.0911	0.1055	0.1055
μmax	4.13	1.0	2.6344	2.5370	2.4793	2.7104	4.1444	4.1444
Y	0.52	1.0	2.6604	0.6434	0.5424	0.5363	0.5190	0.5190
Funk			55.560	17.085	3.2707	0.3394	0.2163	0.2163
rel. Err			2.8915	1.9617	0.8113	0.3347	0.0126	0.0126
Σ Iterat			6	24	52	69	125	127

Tabelle 4.7: WOLS-Identifizierung: 5 Parameter, 0.02 max. Random-Störung

	exakt	start	OLS	Rek 3	Rek 6	Rek 9	Rek 12	Rek 15	Rek ≥ 20
K_{MD}	0.79	1.0	0.7020	0.5173	0.3381	0.4343	0.6170	0.6128	0.7057
K_{ID}	0.50	1.0	2.6155	1.9894	1.2333	0.8859	0.6416	0.6409	0.5599
K_{MA}	0.10	1.0	0.6637	0.2247	0.0600	0.0733	0.0860	0.0890	0.0957
μmax	4.13	1.0	2.6344	2.2057	2.1800	2.6757	3.3953	3.4011	3.7799
Y	0.52	1.0	2.6604	0.8264	0.5661	0.5425	0.5303	0.5295	0.5242
Funk			55.560	13.548	0.4319	0.2562	0.2212	0.2204	0.2158
rel. Err			2.8915	1.1252	0.5999	0.3769	0.1680	0.1622	0.0725
Σ Iterat			6	19	43	73	134	154	188

Tabelle 4.8: OLS-Identifizierung: 5 Parameter, 0.02 max. Random-Störung

	exakt	start	OLS	Rek 4	Rek 8	Rek 12	Rek 16	Rek 20	Rek ≥ 23
K_{MD}	0.79	1.0	0.7033	0.2421	0.4760	0.5752	0.6217	0.6833	0.6833
K_{ID}	0.50	1.0	2.5985	1.6434	0.8574	0.6894	0.6420	0.5862	0.5862
K_{MA}	0.10	1.0	0.6662	0.4195	0.0890	0.0861	0.0889	0.0908	0.0911
μmax	4.13	1.0	2.6292	2.2638	2.7675	3.2209	3.4053	3.6474	3.6474
Y	0.52	1.0	2.6086	0.6560	0.5583	0.5344	0.5306	0.5271	0.5274
Funk			56.270	10.562	1.5073	1.3451	1.3409	1.3389	1.3386
ε_{rel}			2.8697	1.3777	0.3252	0.2075	0.1608	0.1060	0.1055
Σ Iterat			6	31	67	95	118	150	156

Tabelle 4.9: WOLS-Identifizierung: 5 Parameter, 0.05 max. Random-Störung

Auf die angedeuteten Identifizierungsschwierigkeiten und die Schlechtgestelltheit des Problems Bezug nehmend werden im Folgenden neben den Durchbruchskurven $g_D(t) = c_D(20, t)$ und $g_A(t) = c_A(20, t)$ auch die numerisch bestimmten Konzentrationen $c_i(z_k, t)$, $i \in \{D, A\}$, an drei weiteren Messstellen $z_k = 5k\,\text{cm}$, $k = 1, 2, 3$, herangezogen, so dass in der Summe acht unterschiedliche Messungen

	exakt	start	OLS	Rek 4	Rek 8	Rek 12	Rek 16	Rek $\geq$ 18
K_{MD}	0.79	1.0	0.7033	0.2750	0.3795	0.4679	0.6190	0.6190
K_{ID}	0.50	1.0	2.5985	1.3689	1.0206	0.8465	0.6460	0.6460
K_{MA}	0.10	1.0	0.6662	0.1011	0.0756	0.0717	0.0870	0.0871
μmax	4.13	1.0	2.6292	2.0875	2.4610	2.7690	3.3843	3.3843
Y	0.52	1.0	2.6086	0.5746	0.5503	0.5441	0.5319	0.5320
Funk			56.270	1.9135	1.4131	1.3610	1.3408	1.3408
ε_{rel}			2.8697	0.6001	0.4534	0.3519	0.1684	0.1682
Σ Iterat			6	28	58	85	127	132

Tabelle 4.10: OLS-Identifizierung: 5 Parameter, 0.05 max. Random-Störung

zur Verfügung stehen. In den Tabellen 4.11 und 4.12 sind die erzielten Identifizierungsresultate unter Verwendung der ungestörten Datensätze dargestellt. Wie zu erwarten war, konvergieren beide Verfahren etwas schneller. Auch das Residuum hat sich trotz der zusätzlich zu berücksichtigenden Messwerte verringert, was sich schließlich in einem deutlich geringen relativen Fehler ε_{rel} widerspiegelt. Zu bemerken bleibt, dass auch bei diesem Identifizierungsproblem zunächst der gewöhnliche OLS-Ansatz bessere Resultate liefert, bis, wie auch in den vorangegangenen Beispielen, das gewichtete Verfahren dominiert.

	exakt	start	OLS	Rek 5	Rek 10	Rek 15	Rek 20	Rek 30	Rek $\geq$ 41
K_{MD}	0.79	1.0	0.2725	0.4318	0.5508	0.7527	0.7908	0.78997	0.79000
K_{ID}	0.50	1.0	1.2699	0.8956	0.7000	0.5235	0.4995	0.50002	0.50000
K_{MA}	0.10	1.0	0.1173	0.0749	0.0918	0.0986	0.1000	0.10000	0.10000
μmax	4.13	1.0	2.1668	2.6753	3.1964	3.9850	4.1330	4.12987	4.13001
Y	0.52	1.0	0.5343	0.5378	0.5297	0.5209	0.5200	0.52000	0.52000
Funk			1.8597	0.1787	$5.2{\cdot}10^{-2}$	$1.2{\cdot}10^{-3}$	$1.6{\cdot}10^{-6}$	$7.0{\cdot}10^{-10}$	$2.6{\cdot}10^{-12}$
rel. Err			0.5741	0.3764	0.2059	0.0290	$5.5{\cdot}10^{-4}$	$2.2{\cdot}10^{-5}$	$4.8{\cdot}10^{-7}$
Σ Iterat			16	37	66	145	180	239	289

Tabelle 4.11: WOLS-Identifizierung: 5 Parameter, 8 Messungen ohne Störung

	exakt	start	OLS	Rek 5	Rek 10	Rek 20	Rek 30	Rek 41	Rek $\geq$ 50
K_{MD}	0.79	1.0	0.2725	0.4887	0.7793	0.7895	0.7899	0.78997	0.78999
K_{ID}	0.50	1.0	1.2699	0.7793	0.5064	0.5003	0.5001	0.50002	0.50001
K_{MA}	0.10	1.0	0.1173	0.0805	0.0992	0.0999	0.1000	0.10000	0.10000
μmax	4.13	1.0	2.1668	2.9405	4.0873	4.1278	4.1294	4.12987	4.12995
Y	0.52	1.0	0.5343	0.5346	0.5205	0.5200	0.5200	0.52000	0.52000
Funk			1.8597	$8.5{\cdot}10^{-2}$	$7.7{\cdot}10^{-5}$	$2.8{\cdot}10^{-7}$	$1.1{\cdot}10^{-8}$	$4.6{\cdot}10^{-10}$	$6.3{\cdot}10^{-11}$
rel. Err			0.5741	0.2902	0.0091	$5.5{\cdot}10^{-4}$	$9.4{\cdot}10^{-5}$	$2.2{\cdot}10^{-5}$	$8.9{\cdot}10^{-6}$
Σ Iterat			16	46	127	202	267	331	374

Tabelle 4.12: OLS-Identifizierung: 5 Parameter, 8 Messungen ohne Störung

Unberücksichtigt blieb bisher die Tatsache, dass in den vorangegangenen Identifizierungsbeispielen nur fünf der insgesamt sieben Monod-Parameter zu bestimmen waren. Aus diesem Grund wird daher im Folgenden das bisherige Identifizierungsproblem um die fehlenden Parameter $\alpha_{A/D}$ und K_{I_A} erweitert. Unter Verwendung des in Schirmer [54] angegebenen Parameterwerts $K_{I_A} = 91.7\,\frac{\mu g}{cm^3}$ zeigt sich jedoch, welche Einschränkung durch die bisher verwendeten Zulässigkeitsbereiche gegeben war. Zwar liegen alle bisherigen Parameterwerte $K_{M_D} = 0.79\,\frac{\mu g}{cm^3}$, $K_{I_D} = 0.5\,\frac{\mu g}{cm^3}$, $K_{M_A} = 0.1\,\frac{\mu g}{cm^3}$, $\mu_{\max} = 4.13\,\frac{1}{d}$, $Y = 0.52$ und auch $\alpha_{A/D} = 3.16$ in [0.0 5.0], aber nicht der Wert von K_{I_A}. Da gerade die Parameterwerte K_{M_D}, K_{I_D}, K_{M_A}, K_{I_A} nicht direkt bestimmbar sind, müssten daher eigentlich deutlich größere Zulässigkeitsbereiche verwendet werden. Hierbei ergibt sich jedoch das Problem, dass selbst unter exakter Vorgabe der Feldfaktoren Y und $\alpha_{A/D}$ sowie der maximalen Abbaurate $\mu_{\max}$ (z.B. durch separat durchgeführte reale Experimente) das zugehörige Identifizierungsproblem, ob nun als OLS oder WOLS ausgeführt, mit den verbleibenden vier Parametern und einen jeweils angenommenen Zulässigkeitsbereich von [0.0 100.0] zu keinem brauchbaren Ergebnis führt. Um dennoch weitere Untersuchungen bzgl. des gewichteten Ansatzes durchführen zu können, werden im Folgenden die einzelnen Parameter, wider des eigentlichen Identifizierungsgedankens, unterschieden. Als Zulässigkeitsbereich für K_{I_A} wird [0.0 100.0] und für alle anderen Parameter [0.0 5.0] gewählt. Die Startwerte sind weiterhin für alle Parameterwerte mit 1.0 festgelegt. In Tabelle 4.13 finden sich schließlich die entsprechenden Identifizierungsergebnisse unter Verwendung des adaptiv gewichteten Fehlerfunktionals. Es zeigt sich, dass lediglich $\mu_{\max}$, Y und $\alpha_{A/D}$ adäquat bestimmt wurden. Die restlichen Parameter sind weit ab vom globalen Optimum, was sich auch in einem großen Residuum und einem sehr hohen relativen Fehler ε_{rel} widerspiegelt. Da der ungewichtete OLS-Ansatz (in diesem Fall) bereits nach dem ersten Durchlauf terminiert, wurde hierfür keine separate Tabelle angegeben. Die ermittelten Parameterwerte finden sich daher in der ersten Spalte der bereits vorgestellten Tabelle. Es zeigt sich, dass auch diese Werte unbrauchbar sind.

In einem letzten, direkt auf Monod-Daten basierenden Beispiel wird das im vorangegangenen Abschnitt vorgestellte Identifizierungsproblem leicht abgeändert, indem für den Parameter K_{I_A} statt 1.0 nun 20.0 als Startwert verwendet wird. In den Tabellen 4.14 und 4.15 finden sich die erzielten Ergebnisse. Trotz der besseren initialen Schätzung enden beide Verfahren weiterhin in äußerst unbrauchbaren lokalen Minima. Überraschenderweise terminiert jedoch der gewöhnliche OLS-Ansatz deutlich näher am globalen Minimum als das adaptive Verfahren. Als Ursache kann vor allem die hohe Komplexität der Identifizierungsaufgabe angesehen werden. Wie bereits zuvor für stark gestörte Probleme angedeutet, regiert

	exakt	start	OLS	Rek 2	Rek 4	Rek 6	Rek 8	Rek ≥ 11
K_{MD}	0.79	1.0	0.4827	0.4969	0.5037	0.5014	0.5269	0.5269
K_{ID}	0.50	1.0	0.8251	0.8393	0.7971	0.7929	0.7686	0.7696
K_{MA}	0.10	1.0	0.2853	0.2909	0.2929	0.2912	0.2750	0.2751
K_{IA}	91.7	1.0	6.2198	6.2325	6.9316	6.9331	7.1054	7.1058
μmax	4.13	1.0	3.9454	3.9490	3.9920	3.9896	4.0407	4.0408
Y	0.52	1.0	0.5222	0.5281	0.5190	0.5202	0.5192	0.5203
$\alpha_{A/D}$	3.16	1.0	3.1489	3.1552	3.1452	3.1440	3.1473	3.1478
Funk			0.3475	0.3923	0.3553	0.3437	0.3081	0.3087
rel. Err			0.5538	0.5645	0.5500	0.5467	0.5100	0.5103
Σ Iterat			38	43	58	68	82	87

Tabelle 4.13: WOLS-Identifizierung: 7 Parameter, 8 Messungen ohne Störung

auch hier der Zufall über die Qualität der erzielten Identifizierungsresultate. Zudem kommt, dass die Gewichtungsfaktoren auf sehr schlechten Parameterwerten basieren. Dies führte bereits in allen vorangegangenen Beispielen in den ersten rekursiven Durchläufen zu schlechteren Ergebnissen. Erst nachdem etwas bessere Parameterwerte vorlagen, wurde das Defizit zum gewöhnlichen OLS-Ansatz mehr als aufgehoben. Da in diesem Fall jedoch keine ausreichend guten Parameterwerte gefunden wurden, blieb der Erfolg der adaptiven Gewichtung aus.

	exakt	start	OLS	Rek 2	Rek 3	Rek 4	Rek 6	Rek ≥ 8
K_{MD}	0.79	1.0	0.7775	0.1699	0.2046	0.1863	0.3063	0.3138
K_{ID}	0.50	1.0	2.5867	1.9270	1.6104	1.8996	1.3546	1.2117
K_{MA}	0.50	1.0	0.5472	0.4213	0.2655	0.0839	0.1138	0.1045
K_{IA}	0.10	20.0	19.993	19.353	19.051	18.412	19.037	18.411
μmax	4.13	1.0	2.4813	2.1158	2.1056	1.8261	2.3214	2.4314
Y	0.52	1.0	2.4143	0.5022	0.6640	0.5550	0.5573	0.5411
$\alpha_{A/D}$	0.52	1.0	3.3390	3.0584	3.2251	3.1856	3.1094	3.1483
Funk			86.111	21.960	7.7176	2.2866	0.7296	0.3257
rel. Err			1.9346	1.1707	0.8853	0.7367	0.5396	0.4751
Σ Iterat			5	14	21	30	44	54

Tabelle 4.14: WOLS-Identifizierung II: 7 Parameter, 8 Messungen ohne Störung

Neben der Komplexität und der Schlechtgestelltheit des Identifizierungsproblems haben die angedeuteten Identifizierungsschwierigkeiten eine weitere Ursache. An den Einheiten der Parameter lässt sich erkennen, dass K_{IA} einer sehr hohen Konzentration entspricht. Da jedoch c_D und c_A deutlich niedrigere Konzentrationswerte annehmen, hat der inhibierende Koeffizient K_{IA} kaum Einfluss auf das vorherrschende Geschehen. Folgelich ist das verwendete (virtuelle) Experiment eher ungeeignet um diesen speziellen Parameterwert zu identifizieren. Bleibt die Frage, inwieweit dennoch eine erfolgversprechende Anwendung des rekursiv gewichteten

	exakt	start	OLS	Rek 3	Rek 6	Rek 9	Rek 12	Rek 15	Rek ≥ 17
K_{MD}	0.79	1.0	0.7775	0.5342	0.4020	0.5058	0.5957	0.6376	0.6401
K_{ID}	0.50	1.0	2.5867	1.9367	0.9380	0.7657	0.6591	0.6189	0.6183
K_{MA}	0.10	1.0	0.5472	0.1519	0.1278	0.1316	0.1417	0.1452	0.1447
K_{IA}	0.50	20.0	19.993	19.538	19.483	19.498	19.546	19.591	19.593
μmax	4.13	1.0	2.4813	2.1876	2.8489	3.2875	3.6744	3.8592	3.8598
Y	0.52	1.0	2.4143	0.7487	0.5353	0.5284	0.5245	0.5221	0.5225
$\alpha_{A/D}$	0.52	1.0	3.3390	3.2932	3.1504	3.1525	3.1552	3.1552	3.1569
Funk			86.111	14.338	0.1555	$6.6{\cdot}10^{-2}$	$4.4{\cdot}10^{-2}$	$4.2{\cdot}10^{-2}$	$4.1{\cdot}10^{-2}$
rel. Err			1.9346	0.7793	0.3965	0.3167	0.2698	0.2486	0.2473
Σ Iterat			5	21	63	88	111	127	138

Tabelle 4.15: OLS-Identifizierung II: 7 Parameter, 8 Messungen ohne Störung

Fehlerfunktionals möglich ist. Da die Schwäche dieses Verfahrens scheinbar im Umgang mit zu schlechten Parameterwerten liegt, ist es naheliegend vorab eine bestmögliche Lösung durch den gewöhnlichen OLS-Ansatz zu bestimmen. Erst im Anschluß wird fortan das gewichtete Fehlerfunktional angesetzt. Im Gegensatz zu großen linearen Gleichungssystemen, welche (annähernd) mühelos durch geeignete Verfahren gelöst werden können, sind die vorliegenden Optimierungsprobleme an nichtlineare partielle Differentialgleichungen gekoppelt. Dies führt dazu, dass auch das zugehörige Fehlerfunktional (stark) nichtlineare Eigenschaften annimmt. Die aufgeführten Identifizierungsbeispiele zeigten, dass das vorzeitige, jeweils rekursive, Terminieren in einem lokalen Minimum meist durch ein Scheitern der Liniensuche verursacht wurde. Auch aus diesem Grund scheint die Gewichtung hilfreich zu sein, da hiermit das Fehlerfunktional derart manipuliert wird, dass der Optimierungsalgorithmus die zuvor erreichte Extremalstelle wieder verlassen kann. Dass hierfür keine beliebige Gewichtung herangezogen werden kann, ist selbstredend, da sonst eine (kontinuierliche) Verbesserung der Parameterwerte mehr als fragwürdig wäre. Die bereits vorgestellten Identifizierungsbeispiele zeigten jedoch (ohne mathematischen Beweis) eindrucksvoll, dass mit den auf der pseudoinversen Sensitivitätsmatrix basierenden Gewichtungsfaktoren (4.5) eine geeignete Wahl getroffen wurde. Um auch das gewichtete Fehlerfunktional, nach Terminierung des WOLS-Ansatzes in einem lokalen Minimum, zu manipulieren, wird im Anschluß erneut das ungewichtete Funktional herangezogen. Auf diese Weise werden im Folgenden beide Ansätze abwechselnd angewandt, bis keine Verbesserung der Parameterwerte mehr festgestellt bzw. erwartet wird oder der Algorithmus entgültig terminiert. Tabelle 4.16 zeigt die auf diese Weise ermittelten Identifizierungsresultate. In der ersten Spalte ('1. Opt') findet sich das bereits in Tabelle 4.15 vorgestellte lokale Optimimum des gewöhnlichen OLS-Ansatzes. Die zweite Spalte enthält die ermittelten Parameterwerte des direkt im

Anschluß durchgeführten WOLS-Verfahrens. Schließlich werden in den darauffolgenden Spalten die Resultate späterer Durchläufe dargestellt. Deutlich erkennbar ist die kontinuierliche Verbesserung aller Parameterwerte, was sich auch im Residuum und im relativen Fehler widerspiegelt. Selbst der, mit dem vorliegenden Experiment, als nicht identifizierbar geltende Parameter K_{IA} konnte von 19.593 auf 64.757 angehoben werden. Dies entspricht zwar immer noch einer relativen Abweichung von 29.4%, doch in der Summe hat sich der Mittelwert der relativen Fehler aller Parameterwerte um über 76% von 0.2473 auf 0.0582 verringert.

			1. Opt	2. Opt	25. Opt	50. Opt	75. Opt	100. Opt	140. Opt
	exakt	start	OLS	WOLS	OLS	WOLS	OLS	WOLS	OLS
K_{MD}	0.79	1.0	0.6401	0.6719	0.7112	0.7455	0.7506	0.7636	0.7695
K_{ID}	0.50	1.0	0.6183	0.5940	0.5559	0.5316	0.5262	0.5181	0.5141
K_{MA}	0.10	1.0	0.1447	0.1438	0.1206	0.1118	0.1094	0.1059	0.1049
K_{IA}	91.7	20.0	19.593	21.024	33.610	46.263	50.966	60.665	64.757
μmax	4.13	1.0	3.8598	3.9636	3.9812	4.0454	4.0533	4.0735	4.0880
Y	0.52	1.0	0.5225	0.5210	0.5212	0.5206	0.5206	0.5204	0.5203
$\alpha_{A/D}$	3.16	1.0	3.1569	3.1567	3.1585	3.1590	3.1593	3.1596	3.1597
Funk			$4.1{\cdot}10^{-2}$	$3.7{\cdot}10^{-2}$	$1.0{\cdot}10^{-2}$	$3.7{\cdot}10^{-3}$	$2.3{\cdot}10^{-3}$	$1.0{\cdot}10^{-3}$	$6.9{\cdot}10^{-4}$
rel. Err			0.2473	0.2271	0.1557	0.1079	0.0943	0.0688	0.0582
Σ Rek			17	30	188	316	420	525	668
Σ Iterat			138	217	1119	1892	2491	3141	4010

Tabelle 4.16: OLS/WOLS-Identifizierung

Die rekursiven Veränderungen des Residuums und des relativen Fehlers können auch graphisch in den Abbildungen 4.5.a und 4.5.b nachgeprüft werden. Es fällt deutlich auf, dass eine gewisse Kontinuität in der Verringerung dieser beiden Werte vorhanden ist. Zwar gibt es immer wieder kleinere Identifizierungsplateaus, in welchen das Residuum aufgrund der wechselnden Anwendung beider Verfahren kurzzeitig oszilliert und der relative Fehler stagniert, doch das Gesamtbild verdeutlicht die bereits weiter oben angesprochene Verbesserung der erzielten Parameterwerte. Ein unerwartetes, auf zufälligen Resultaten basierendes Verhalten ist nicht zu erkennen.

Abschließend wird der Vollständigkeit halber diese wechselnde OLS/WOLS-Vorgehensweise (abgekürzt auch mit O/W bezeichnet) auf die vorangegangenen, gestörten Identifizierungsbeispiele angewandt. Bereits wenige Optimierungsaufrufe ergaben hierbei (nochmals) bessere Identifizierungsresultate, welche in Tabelle 4.17 aufgezeigt sind. Die angegebenen Werte sind dabei nicht als endgültige Ergebnisse zu verstehen. Vielmehr wurde das Verfahren abgebrochen, nachdem sich ein oszillierendes Verhalten einstellte, bei dem abwechselnd qualitativ mehr

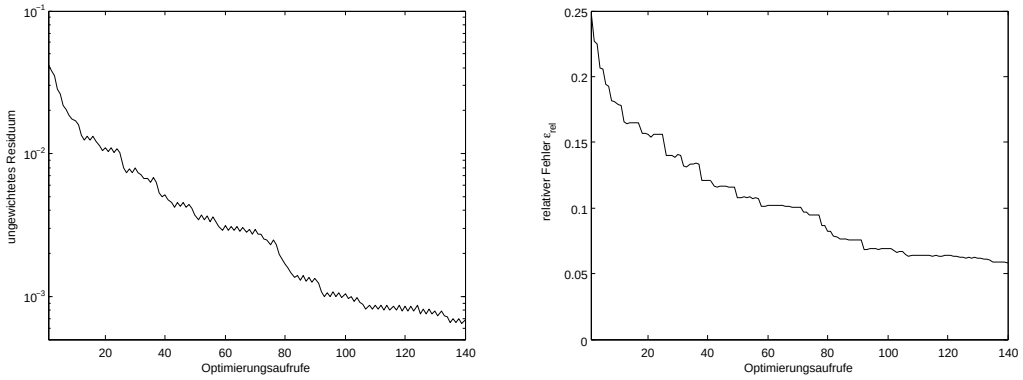

Abbildung 4.5: Rekursive OLS/WOLS-Identifizierung

oder weniger gleichwertige Resultate auftraten.

		Störung, max. 0.02				Störung, max. 0.05			
		Tab. 4.7		Tab. 4.8		Tab. 4.9		Tab. 4.10	
	exakt	WOLS	O/W	OLS	O/W	WOLS	O/W	WOLS	O/W
K_{MD}	0.79	0.7893	0.7924	0.7057	0.7818	0.6833	0.6936	0.6190	0.6985
K_{ID}	0.50	0.5009	0.4996	0.5599	0.5063	0.5862	0.5781	0.6460	0.5747
K_{MA}	0.10	0.1055	0.1018	0.0957	0.1012	0.0911	0.0917	0.0871	0.0929
μmax	4.13	4.1444	4.1432	3.7799	4.0987	3.6474	3.6876	3.3843	3.7069
Y	0.52	0.5190	0.5193	0.5242	0.5198	0.5274	0.5270	0.5320	0.5271
Funk		0.2163	0.2148	0.2158	0.2148	1.3386	1.3384	1.3408	1.3381
rel. Err		0.0126	0.0053	0.0725	0.0086	0.1055	0.0964	0.1682	0.0905
# Opt.		1	2	1	9	1	5	1	4

Tabelle 4.17: Gestörte OLS/WOLS-Identifizierung

Die bisherigen Resultate sind sehr vielversprechend. Da jedoch für alle vorgestellten Berechnungen virtuelle Messdaten zur Verfügung standen, die direkt durch eine auf einer Monod-Parametrisierung basierende Simulation beruhten, bleibt die Frage, inwieweit auch andere Reaktionsraten mit Hilfe dieser vorgegebenen Nichtlinearitäten identifiziert werden können. Im Folgenden werden daher funktionale Reaktionsraten erzeugt, welche der Monod-Parametrisierung ähneln, aber nicht exakt von dieser dargestellt werden können. Sei hierzu zunächst die in (2.34) definierte Funktion

$$f_c(x) = \frac{x}{c+x} \tag{4.8}$$

betrachtet. Diese hat auf einem begrenzten Definitionsbereich bei geeigneter, von c abhängiger, Wahl der Konstanten $c_1, c_3, c_3 \in \mathbb{R}^+$ ein vergleichbares Monotonie-

und Krümmungsverhalten wie

$$g_c(x) = \sqrt{-c_1x^2+c_2x+c_3}-\sqrt{c_3}\,. \tag{4.9}$$

Nachfolgend wird daher exemplarisch $f_{0.5}$ auf $[0,2]$ und $f_{0.25}$ auf $[0,3]$ durch

$$g_{0.5}(x) := \sqrt{-\frac{4}{25}x^2+\frac{4}{5}x+\frac{1}{25}}-\frac{1}{5} \quad \text{und} \quad g_{0.25}(x) := \sqrt{-\frac{3}{40}x^2+\frac{11}{20}x+\frac{1}{225}}-\frac{1}{15}$$

approximiert. Zur Bestimmung der Koeffizienten c_i, $i=1,2,3$, wurden hierbei neben $g_c(0) = f_c(0)$ und $g_c'(0) = f_c'(0)$ die Bedingungen $g_{0.5}(2) = f_{0.5}(2)$ und $g_{0.25}'(2) = f_{0.25}'(2)$ bzw. $g_{0.25}(3) = f_{0.25}(3)$ und $g_{0.25}'(3) = f_{0.25}'(3)$ angesetzt. Die zugehörige L^2-Norm der jeweiligen Abweichung wurde mittels Matlab (näherungsweise) bestimmt. Es gilt

$$\left\|f_{0.5}-g_{0.5}\right\|_{2,[0,2]} = 0.0578 \quad \text{und} \quad \left\|f_{0.25}-g_{0.25}\right\|_{2,[0,3]} = 0.2268\,.$$

Es fällt auf, dass (4.8) für $c=0.5$ relativ gut durch (4.9) abgebildet wird, wohingegen die Näherung für $c=0.25$, aufgrund der stärkenen Krümmung, deutlich schlechter ist. Da für kleinere c-Werte eine noch größere Krümmung zu erwarten ist, wird an dieser Stelle auf eine entsprechende Approximation verzichtet.

Nachfolgend wird zum Lösen des vorliegenden Differentialgleichungssystems statt der Monod-Parametrisierung (2.35) die Abbaurate

$$\mu_{\text{approx}}(c_D,c_A,c_B) = \mu_{\max}\cdot g_{0.5}(c_D)\cdot g_{0.25}(c_A)\cdot c_B \tag{4.10}$$

verwendet und die ermittelten Konzentrationswerte $c_i(z_k,t)$, $i\in\{D,A\}$, an den diskreten (Mess-)Stellen $z_k=5k$, $k=0,1,2,3$, als Messdaten für das im Anschluß stattfindende Identifizierungsproblem zur Verfügung gestellt. Die rekursiv ermittelten Optimalwerte lauten $\mu_{\max}=7.85$, $K_{M_D}=1.10$, $K_{I_D}=9.95$, $K_{M_A}=1.01$ und $K_{I_A}=11.34$. In den Abbildungen 4.6.a und 4.6.b finden sich neben den bereits definierten $f_{0.5}$ und $g_{0.5}$ bzw. $f_{0.25}$ und $g_{0.25}$ die zugehörigen, durch den Optimierungsprozess festgelegten, Funktionsterme

$$f_{\text{opt},i}(c_i) = \frac{c_i}{K_{M_i}+c_i+\frac{c_i^2}{K_{I_i}}}\ , \quad i\in\{A,D\}\,.$$

Auf den ersten Blick scheint der Parameteridentifizierungsprozess vollkommen gescheitert zu sein. Wird jedoch statt einer separaten Betrachtung der einzelnen Funktionskomponenten die gesamte Reaktionsrate (4.10) für einen Vergleich mit (2.35) herangezogen, so zeigt sich, dass diese sehr wohl adäquat identifiziert werden konnte. Dies spiegelt sich sowohl in der L^2-Norm

$$\left\|\mu-\mu_{\text{approx}}\right\|_{2,[0,2]\times[0,3]\times[0,0.1]} = 4.1\cdot10^{-3} \tag{4.11}$$

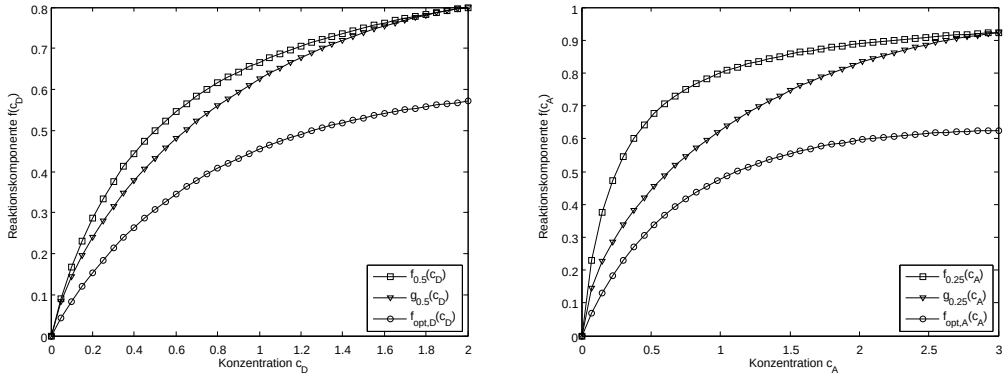

Abbildung 4.6: Identifizierung mit modifizierter Reaktionsrate

als auch im erzielten (ungewichteten) Residuum $\mathcal{J}_h = 5.4 \cdot 10^{-1}$ des zugehörigen Minimierungsproblems wieder. Der Preis, der hierfür jedoch bezahlt werden musste, liegt in einem äußerst schlecht ermittelten Parameterwert $\mu_{\max}$. Um das Defizit der beiden Komponenten f_D und f_A auszugleichen, nahm dieser statt 4.13 einen beinahe doppelt so hohen Wert an. Da allerdings $\mu_{\max}$ die maximale Abbaurate beschreibt und damit physikalisch motiviert ist, kann dies nicht ohne Weiteres akzeptiert werden. Trotz klein ausgefallener Norm (4.11) wird damit das vorgestellte Identifizierungsproblem als ungelöst eingestuft. Dies bekräftigt auch ein weiterer Identifizierungsansatz, bei dem nur die vier Koeffizienten K_{MD}, K_{ID}, K_{MA} und K_{IA} zu bestimmen waren und $\mu_{\max} = 4.13$ fest vorgegeben wurde. Unabhängig von der Größe der Zulässigkeitsbereiche der beiden inhibierenden Parameter K_{ID} und K_{IA}, terminierte das Verfahren stets mit maximalen Werten für diese beiden Unbekannten. Zusammenfassend musste damit festgestellt werden, dass, wie erwartet, die gesuchte Reaktionsrate (4.10) unter Vorgabe einer fixen Monod-Parametrisierung nicht ausreichend identifiziert werden kann.

4.2.3 Entwässerung einer Laborsäule

In dieser Fallstudie werden bei einer numerischen Entwässerungssimulation die hydraulischen Funktionen mittels Identifizierung der van Genuchten-Parameter bestimmt. Das zugrundeliegende Experiment wurde an der Bauhaus-Universität Weimar von Dr. Yvonne Lins und Prof. Dr. Tom Schanz durchgeführt. Hierbei wurde bei einer 67.0 cm hohen, hermetisch verschlossenen, zylinderförmigen ($d = 30.5$ cm) Laborsäule, welche bis auf 13.0 cm vollständig mit fluidgesättigten Huston Sand (reiner, enggestufter Sand, Korngröße 0.1-1.0 mm) gefüllt war (die Höhe der Sandprobe betrug damit $L = 54.0$ cm), an der Unterseite Feuchtigkeit

mit konstanter Flußrate $\boldsymbol{q} = 26.73\,\frac{\text{ml}}{\text{min}}$ abgepumpt. Während der gesamten Dehydrierung ($t_f = 429\,\text{min}$) wurde der Porendruck $\psi_\varepsilon(h_i, t)$ mit Hilfe von Tensiometern sowie der Wassergehalt $\theta_\varepsilon(h_i, t)$ mittels TDR- (Time Domain Reflectrometry) Sensoren an fünf unterschiedlichen Ebenen h_i, $i = 1, \ldots, 5$, und in Zeitabständen von $\Delta t = 3\,\text{min}$ gemessen. Unter Einführung einer entgegen der Gravitationsrichtung orientierten Ortsachse mit Ursprung an der Unterkante der Sandprobe, sind die diskreten Messebenen durch $h_1 = 46.0\,\text{cm}$, $h_2 = 37.0\,\text{cm}$, $h_3 = 26.8\,\text{cm}$, $h_4 = 17.5\,\text{cm}$ und $h_5 = 7.5\,\text{cm}$ gegeben. Eine schematische Darstellung des verwendeten Versuchsaufbaus findet sich in Abbildung 4.7 (vgl. Lins e.a. [41], Figure 1).

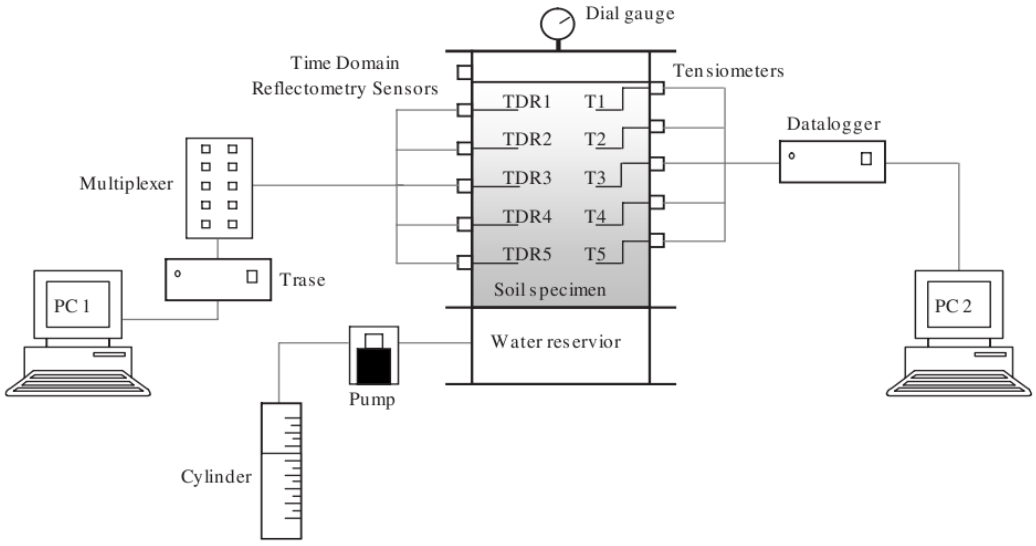

Abbildung 4.7: Weimar-Experiment: Schematischer Versuchsaufbau

Aufgrund des großen Innendurchmessers können Randeffekte vernachlässigt werden, so dass für die numerische Simulation eine eindimensionale Betrachtung gerechtfertigt ist. Die angelegte Flussrate wird entsprechend mit $q = -6.1 \cdot 10^{-4}\,\frac{\text{cm}}{\text{s}}$ angenommen. Da an der Oberkante keinerlei Messdaten vorliegen, wird das zu simulierende Gebiet auf das Intervall $[0, h_1]$ beschränkt. Damit können die zur Berechnung notwendigen Anfangswerte durch $\psi(z, 0) = 54 - z\,\text{cm}$, $z \in (0.0, 46.0)$, und die Randbedingungen durch $\psi(h_1, t) = \psi_\varepsilon(h_1, t)$ und $\frac{\partial}{\partial z}\psi(0, t) = q$, $t \in (0, 429)$, festgelegt werden. Als grundlegende Modellgleichung wird die Richards-Gleichung (2.2) mit der van Genuchten-Mualem-Parametrisierung (2.5)-(2.7) herangezogen und geeignet diskretisiert (vgl. z.B. Schneid [55]). Als Messdaten stehen dem zugehörigen Identifizierungsproblem schließlich die experimentell bestimmten Porendruck- und Wassergehaltsmessungen an den verbleibenden vier Ebenene h_i, $i = 2, \ldots, 5$,

zur Verfügung (vgl. hierzu auch Abbildung 4.8).

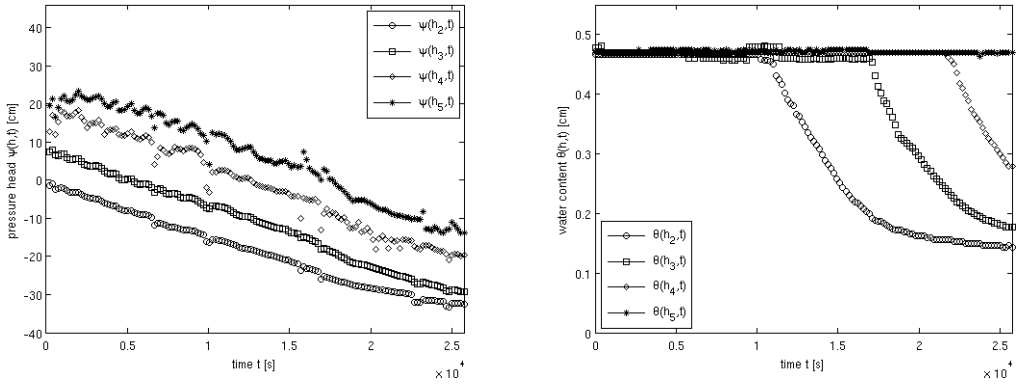

Abbildung 4.8: Weimar-Experiment: Druck- und Sättigungsdaten

Neben den Porendruck- und Wassergehaltsmessungen wurden bei der vorliegenden Sandprobe auch die (relativ einfach zu ermittelnden) Werte der maximalen hydraulischen Leitfähigkeit $K_{\text{sat}} = 2.75 \cdot 10^{-4} \frac{\text{m}}{\text{s}}$ und des maximalen Wassergehalts $\theta_{\text{sat}} = 0.47$ experimentell bestimmt. Entsprechend kann das zugehörige Identifizierungsproblem um diese beiden Parameterwerte reduziert werden, so dass statt dem eigentlichen Parametervektor $p_h \in P_h \subset \mathbb{R}^5$ ($K_{\text{sat}}, \theta_{\text{sat}}, \theta_{\text{res}}$, α und n) nur noch die Werte für θ_{res}, α und n zu identifizieren sind. Um die Identifizierungsresultate nicht durch zu klein gewählte Zulässigkeitsbereiche künstlich einzuschränken, wurde für die folgenden Berechnungen $\theta_{\text{res}} \in [0.0, 30.0]$, $\alpha \in [0.001, 0.1]$ und $n \in [1.1, 10.0]$ festgelegt. Die Anfangswerte sind entsprechend erster Schätzungen auf $\theta_{\text{res}} = 0.05$, $\alpha = 0.02$ und $n = 3.0$ gesetzt.

Im Folgenden wurden für das Identifizierungsproblem zunächst nur die (mit (4.3) gewichteten) Porendruckdaten $\psi_\varepsilon(h_i, t)$, $i = 2, \ldots, 5$, herangezogen. Unter Verwendung des gewöhnlichen OLS-Ansatzes konnten hierbei die Parameter $\theta_{\text{res}} = 0.0747$, $n = 6.3132$ und $\alpha = 0.0395$ bestimmt werden. Auch der rekursiv gewichtete Identifizierungsalgorithmus terminierte bereits nach drei Durchläufen mit sehr ähnlichem Ergebnis $\theta_{\text{res}} = 0.0728$, $n = 6.4254$ und $\alpha = 0.0383$. Da ein Vergleich mit exakten Optimalwerten (wie im Fall der vorangegangenen virtuellen Experimente) nicht möglich ist, bleibt nur eine allgemeine Analyse der erzielten Identifizierungsresultate. Nachträglich durchgeführte Messexperimente an einer vergleichbaren, unverdichteten Quartz-Sandprobe zeigten, dass der residuale Wassergehalt mit $\theta_{\text{res}} = 0.05$ (genauer $\theta_{\text{res}} \in [0.045, 0.05]$) angenommen werden kann (vgl. Lins et al.

[41], Figure 8.a). Entsprechend muss bei dem hier vorliegenden Identifizierungsproblem festgestellt werden, dass der ermittelte Wert für θ_{res} deutlich zu hoch ist. Was die beiden (unphysikalischen) Parameter n und α angeht, so sind diese äußerst schwer zu überprüfen. Auffällig ist allerdings, dass beide Werte überraschend hoch ausgefallen sind. Ein Vergleich mit Tabelle 2.1 zeigt jedoch, dass für Sand erheblich höhere Werte als bei anderen Bodenproben angenommen werden kann. Da insbesondere die Korngröße, eine etwaige Verdichtung und andere spezifische Eigenschaften des dort zugrundegelegten Sandes nicht bekannt sind und sich auch der residuale Wassergehalt zu dem hier vorliegenden Wert unterscheidet, bleibt lediglich festzustellen, dass die identifizierten Parameterwerte n und α durchaus im möglich Bereich liegen.

Um ein aussagekräftigeres Identifizierungsergebnis zu erzielen und um die bereits vorliegenden Werte besser bewerten zu können, werden nachfolgend zu den verwendeten Porendruckdaten $\psi_\varepsilon(h_i, t)$, $i = 2, \ldots, 5$, auch die vorliegenden Wassergehaltsmessungen $\theta_\varepsilon(h_i, t)$, $i = 2, \ldots, 5$, berücksichtig. Die ermittelten Parameterwerte finden sich in Tabelle 4.18.

	IC	OLS	Rec 2	Rec 4	Rec 6	Rec 8	Rec 10	Rec $\geq$13
θ_{res}	0.05	0.0100	0.0464	0.0648	0.0614	0.0474	0.0452	0.04719
n	3.00	10.000	8.8828	8.3171	8.4159	8.1725	7.7763	6.94320
α	0.02	0.0418	0.0423	0.0433	0.0436	0.0436	0.0414	0.04104
F		$8.91 \cdot 10^2$	$8.96 \cdot 10^2$	$1.02 \cdot 10^3$	$1.02 \cdot 10^3$	$9.53 \cdot 10^2$	$8.66 \cdot 10^2$	$8.63 \cdot 10^2$

Tabelle 4.18: Weimar-Experiment: Rekursiv ermittelte Parameterwerte

Es zeigt sich, dass der gewöhnliche OLS-Ansatz mit dem nun komplexeren Fehlerfunktional überfordert ist. Die rekursiv gewichtete Identifizierung hingegen terminiert mit vertretbarem Aufwand (13 Aufrufe) in einem sehr vielversprechenden Minimum. So liegt der identifizierte residuale Wassergehalt $\theta_{\text{res}} = 0.047$ exakt in dem erwarteten Wertebereich. Zudem nehmen die beiden Parameterwerte n und α wieder die bereits in der vorangegangenen Identifizierung ermittelte Größenordnung an. Da auch ein weiterer Identifizierungsversuch mit eingeschränktem Zulässigkeitsbereich $n \in [1.1, 5.0]$ den identifizierten Parameterwert mit $n = 5.0$ größtmöglich bestätigt, kann davon ausgegangen werden, dass tatsächlich die ermittelten Werte zutreffend sind. In Abbildung 4.9 sind abschließend die simulierten Porenruck- und Wassergehaltsdaten dargestellt.

Unabhängig von den erzielten Parameterwerten besitzt auch dieses rekursive Identifizierungsproblem die bereits im vorangegangenen Abschnitt (insbesondere

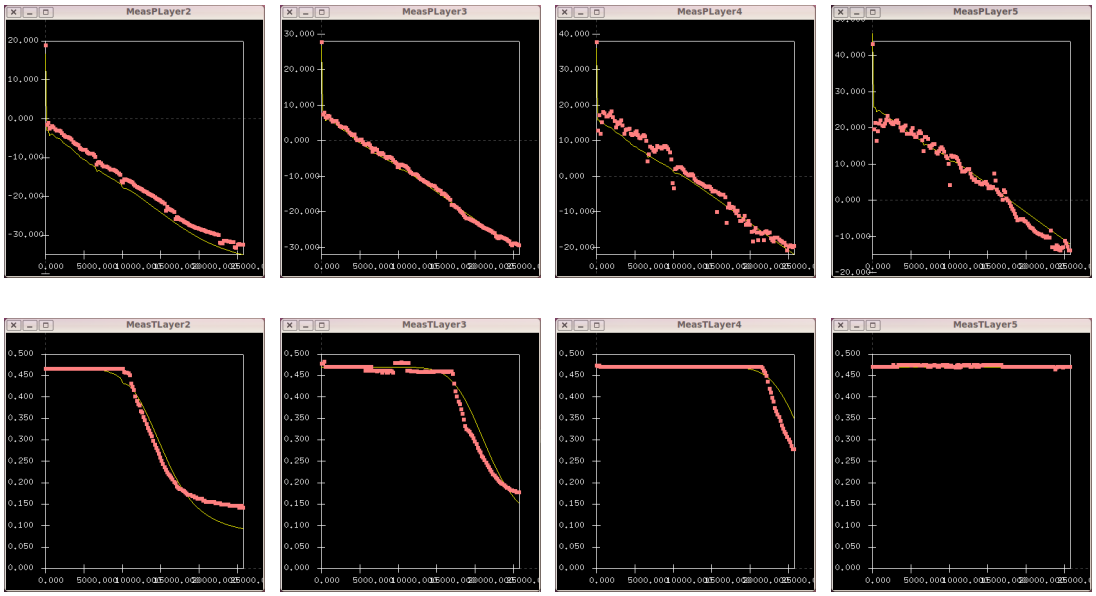

Abbildung 4.9: Simulierte Porendruck- und Wassergehaltsdaten

unter Verwendung gestörter Messdaten) festgestellte Entwicklung des zu minimierenden Residuums. Auch hier nimmt das (gewichtungsbereinigte) Fehlerfunktional zunächst etwas höhere Werte an, bis sich nach und nach ein Ausgleich einstellt und schließlich die optimale Lösung erreicht wird. Damit wird auch mit dieser Fallstudie (nochmals) bestätigt, dass das alleinige Betrachten des Residuums nicht ausreichend ist, um die Güte der ermittelten Identifizierungsresultate vollständig zu bewerten.

4.3 Formfreie Identifizierung der biochemischen Abbaurate

In diesem Abschnitt wird die dreikomponentige Abbaurate eines Schadstoffes, welche von der Konzentration der Biomasse, des Elektronen-Akzeptors (typischerweise Sauerstoff) und des Elektronen-Donators (der Schadstoff selbst) abhängt, mit Hilfe der beiden, im Kapitel 3.2.2 vorgestellten, formfreien Ansätze identifiziert. Als Berechnungsgrundlage dienen die bereits im Abschnitt 4.2.2 verwendeten und zum Teil auf Schirmer [54] basierenden bodenspezifischen Daten. Zudem wird das experimentelle Setup des vorangegangenen virtuellen Experimentes

vollständig übernommen. Damit stehen neben den Durchbruchskurven $c_D(20,t)$ und $c_A(20,t)$ auch wieder die Messwerte $c_i(h,t)$, $i \in \{D,A\}$, $h = 5k\text{cm}$, $k = 1,2,3$, zur Verfügung.

Um die Komplexität des formfreien Identifizierungsproblems zu reduzieren, wird statt der vektorwertigen Nichtlinearität $\boldsymbol{f} : \mathbb{R}^3 \to \mathbb{R}^3$ nachfolgend

$$\boldsymbol{f} =: \begin{pmatrix} f \\ \alpha_{A/D} f \\ Y f \end{pmatrix} \quad \text{mit } \alpha_{A/D}, Y \in \mathbb{R}^+$$

verwendet, so dass lediglich die Reaktionsrate $f : \mathbb{R}^3 \to \mathbb{R}$ und die beiden Feldfaktoren $\alpha_{A/D}$ und Y zu bestimmen sind. Dies stellt keine Einschränkung des Identifizierungsproblems dar, da die Änderung der beteiligten Konzentrationen durch biochemische Prozesse gekoppelt sind und entsprechend (unter Kenntnis der exakten Reaktionskinetiken), mittels geeigneter Zusammenfassung der stöchiometrischen Koeffizienten, gegenseitig dargestellt werden können. Zudem wird angenommen, dass im Fall des verwendeten Experimentes (mit niedriger Biomassenkonzentration) ein proportionales Verhalten zwischen c_D und f vorliegt. Folglich hat die dritte Komponente von f auch nur einen linearen Einfluss auf die zu bestimmende Abbaurate. Schließlich ist für die hier zugrundegelegte Problemestellung bekannt, dass die Reaktionsrate auf den Koordinatenebenen den Wert Null annimmt. Dies hat zur Folge, dass alle Splines mit mindestens einer Komponente gleich Null, mit Null gewichtet werden und damit nicht identifiziert werden müssen. In Tabelle 4.19 ist die zugehörige Anzahl der Freiheitsgrade (DOF) unter Verwendung von lokalen und hierarchischen trilinearen Basen für die ersten sechs (Diskretisierungs-)Stufen $n = (2,2,2), \ldots, (33,33,2)$ bzw. hierarchischen Skalen $s = 0, \ldots, 5$, angegeben. Im Fall hierarchischer Basisfunktionen wird zudem unterschieden, ob ein volles oder dünnes Gitter zugrunde gelegt ist. Ein Vergleich mit den dreifachen Werten der in Tabelle 3.2.b angegebenen Anzahl an Unbekannten zeigt eindrucksvoll, inwieweit die vorgestellten Vereinfachungen die Dimensionalität des Identifizierungsproblems reduzieren.

	$s=0$	$s=1$	$s=2$	$s=3$	$s=4$	$s=5$
	$(2,2,2)$	$(3,3,2)$	$(5,5,2)$	$(9,9,2)$	$(17,17,2)$	$(33,33,2)$
lok/voll	1	4	16	64	256	1024
dünn	1	4	12	32	80	192
Y,$\alpha_{A/D}$	2	2	2	2	2	2

Tabelle 4.19: Formfreie Identifizierung - Anzahl der Freiheitsgrade

Als (ungestörte) Messdaten werden im Folgenden zunächst die, unter Verwen-

dung der Monod-Parametrisierung (2.35), mit $\mu_{\max} = 4.13\,\frac{1}{\mathrm{d}}$, $K_{M_D} = 0.79\,\frac{\mu\mathrm{g}}{\mathrm{cm}^3}$, $K_{M_A} = 0.1\,\frac{\mu\mathrm{g}}{\mathrm{cm}^3}$, $Y = 0.52$ und $\alpha_{A/D} = 3.16$, simulierten Konzentrationen $c_i(h,t)$, $i \in \{D, A\}$, $h = 5k\mathrm{cm}$, $k = 0, \ldots, 3$, angesetzt. Entsprechend ist die zu identifizierende Abbaurate monoton steigend, so dass eine Einschränkung auf monotone Funktionen den Identifizierungsprozess weiter stabilisieren wird. An dieser Stelle sei bemerkt, dass nach Lemma 3.16 die Überprüfung der Monotonieeigenschaften an den diskreten Gewichtungspunkten ausreicht, um die Monotonie der gesamten, stückweise trilinearen Nichtlinearität nachzuweisen. Da für das vorliegende Identifizierungsproblem die auf den Koordinatenebenen liegenden Gewichtungspunkte ausgeschlossen wurden, sind für jede Stützstelle, unabhängig vom Basistyp, genau drei Ungleichungen zu erfüllen. Trotz gleicher Anzahl an Nebenbedingungen ist jedoch die Überprüfung im Fall hierarchischer Basen zeitlich aufwändiger, da nicht einfach nur die Gewichtung der beiden benachbarten Knoten verglichen werden kann, sondern vielmehr eine Interpolation aller Splines für jede einzelne Stützstelle ermittelt werden muss.

Analog zu den ersten, im Abschnitt 4.2.2 vorgestellten, rekursiven Identifizierungsansätzen werden auch hier zunächst die beiden (physikalisch motivierten) Feldfaktoren Y und $\alpha_{A/D}$ exakt vorgegeben. Damit bleibt die Gewichtung jeder einzelnen, im Folgenden stets trilinearen, Basisfunktion mit Hilfe des im Kapitel 3.2.3 eingeführten Multi-Level-Algorithmusses zu bestimmen. In Tabelle 4.20 sind die erzielten Identifizierungsresultate unter Verwendung lokaler Basen angegeben. Um ein möglichst gutes Ergebnis zu erlangen, wurde hierbei die Anzahl an Diskretisierungsknoten bei jeder hierarchischen Verfeinerung (auf Kosten der Berechnungszeit) jeweils nur in einer Richtung um eins erhöht. Die verwendeten Stützstellen wurden dabei stets erneut äquidistant angeordnet. Da es sich um ein virtuelles Experiment handelt, bei dem entsprechend die exakte Reaktionsrate μ_{exakt} bekannt ist, wurde neben dem jeweiligen Freiheitsgrad DOF und dem auf (3.18) basierenden Residuum F auch die L^2-Norm

$$\left\| \mu_{\mathrm{exakt}} - \mu_{\mathrm{opt}} \right\|_{2,[0,2]\times[0,3]\times[0,0.1]}$$

über der Abweichung der simulierten Rate μ_{opt} von μ_{exakt} angegeben. Wie zu erwarten war, wird das zu minimierende Residuum, aufgrund der immer höher werdenden Flexibilität, mit jedem hierarchischen Durchlauf (zum Teil) deutlich verringert. Geringfügige Ausnahmen, welche durch ein unverändertes lokales Minimum und der (hier) notwendigen Skaleninterpolation zu erklären sind, stellten sich nur selten ein. Wird direkt die zu identifizierende Reaktionrate bewertet, so konnte ihre Abweichung zur exakten Rate auf $9.0 \cdot 10^{-3}$ (gemessen in der L^2-Norm) reduziert werden. Qualitativ steht dies zwar in keinem Verhältnis zu den

im vorangegangenen Abschnitt erzielten Identifizierungsresultaten, doch unter Berücksichtigung, dass hier die verwendete Parametrisierung völlig unabhängig von der, für die virtuelle Erstellung der Messdaten, verwendete Reaktionsrate ist, sind die identifizierten Ratenwerte durchaus zufriedenstellend.

	n							
	$(2,2,2)$	$(3,2,2)$	$(3,3,2)$	$(4,3,2)$	$(4,4,2)$	$(5,4,2)$	$(5,5,2)$	$(6,5,2)$
F	832.72	175.55	69.345	61.516	31.724	27.322	12.829	11.008
DOF	1	2	4	6	9	12	16	20
$\|\,.\,\|_2$	$4.1{\cdot}10^{-2}$	$4.6{\cdot}10^{-2}$	$4.1{\cdot}10^{-2}$	$3.4{\cdot}10^{-2}$	$2.1{\cdot}10^{-2}$	$2.0{\cdot}10^{-2}$	$1.5{\cdot}10^{-2}$	$1.5{\cdot}10^{-2}$

	n						
	$(6,6,2)$	$(7,6,2)$	$(7,7,2)$	$(8,7,2)$	$(8,8,2)$	$(9,8,2)$	$(9,9,2)$
F	5.8376	6.0113	5.1085	5.4488	5.6823	2.9513	2.6290
DOF	25	30	36	42	49	56	64
$\|\,.\,\|_2$	$1.3{\cdot}10^{-2}$	$1.2{\cdot}10^{-2}$	$1.1{\cdot}10^{-2}$	$1.1{\cdot}10^{-2}$	$1.1{\cdot}10^{-2}$	$9.3{\cdot}10^{-3}$	$9.0{\cdot}10^{-3}$

Tabelle 4.20: FF-Identifizierung: Lokale Basen, äquidistante Unterteilung

In den beiden nachfolgenden Tabellen 4.21 und 4.22 sind entsprechend der vorangegangenen Überlegungen die erzielten Resultate unter Verwendung hierarchischer Basen, welche sowohl auf einem vollen als auch dünnen Gitter definiert sind, aufgezeigt. Um auch hier die hierarchische Verfeinerungsschrittweite möglichst gering zu halten, wurde, analog zu der in Kapitel 3.2.2.2 eingeführten Nomenklatur, jede Skala s mittels der Blöcke $\vec{\sigma} \in \Upsilon_3^{\mathrm{Strat}}(s)$ unterteilt. Ein Vergleich zeigt, dass, wie es die Theorie auch vorgibt, die Verwendung des vollen Gitters nachteilig ist. Ein qualitativer Unterschied zwischen lokalen und auf einem dünnen Gitter definierten hierarchischen Basen konnte bei niedriger bis mittlerer Dimensionalität (mit diesem Identifizierungsproblem) nicht festgestellt werden. Erst bei hohen Skalen wird mit lokalen Basen eine geringfügig bessere Reaktionsrate ermittelt. Ein weiterer Unterschied liegt in der benötigten Simulationszeit vor. Während bei lokalen Basen bei jeder Funktionsauswertung stets nur acht interpolierte Splinewerte eingehen, wächst bei hierarchischen Basen der Aufwand mit größer werdender Skala unentwegt an. Dies macht sich vor allem bei hohen Skalen bemerkbar, so dass die zur Verfügung stehende Berechnungszeit keine weiteren Durchläufe zulässt. Zudem kommt, dass selbst unter Verwendung einer weiteren Skala keine nennenswerte Verbesserung der bereits erzielten Resultate festgestellt werden konnte.

Sind bessere Identifizierungsresultate notwendig, so müssen dem zugrundegelegten Simulationsproblem weitere Informationen zur Verfügung gestellt werden. Im Fall der vorliegenden Reaktionsrate kann hierfür beispielsweise die Konvexität

	$s=0$	$s=1$			$s=2$				
	$(0,0,0)$	$(1,0,0)$	$(0,1,0)$	$(1,1,0)$	$(2,0,0)$	$(2,1,0)$	$(0,2,0)$	$(1,2,0)$	$(2,2,0)$
F	832.72	175.55	86.169	86.169	69.256	59.567	59.567	31.388	15.192
DOF	1	2	3	4	6	8	10	12	16
$\|\,.\,\|_2$	$4.1{\cdot}10^{-2}$	$4.6{\cdot}10^{-2}$	$5.4{\cdot}10^{-2}$	$5.4{\cdot}10^{-2}$	$3.9{\cdot}10^{-2}$	$3.5{\cdot}10^{-2}$	$3.5{\cdot}10^{-2}$	$2.1{\cdot}10^{-2}$	$1.5{\cdot}10^{-2}$

	$s=3$						
	$(3,0,0)$	$(3,1,0)$	$(3,2,0)$	$(0,3,0)$	$(1,3,0)$	$(2,3,0)$	$(3,3,0)$
F	14.289	14.122	14.122	13.695	13.668	13.667	13.667
DOF	20	24	32	36	40	48	64
$\|\,.\,\|_2$	$1.5{\cdot}10^{-2}$	$1.5{\cdot}10^{-2}$	$1.5{\cdot}10^{-2}$	$1.5{\cdot}10^{-2}$	$1.5{\cdot}10^{-2}$	$1.5{\cdot}10^{-2}$	$1.5{\cdot}10^{-2}$

Tabelle 4.21: FF-Identifizierung: Hierarchische Basen, volles Gitter

	$s=0$	$s=1$		$s=2$		
	$(0,0,0)$	$(1,0,0)$	$(0,1,0)$	$(2,0,0)$	$(1,1,0)$	$(0,2,0)$
F	832.72	175.55	86.169	86.169	67.748	59.598
DOF	1	2	3	5	6	8
$\|\,.\,\|_2$	$4.1{\cdot}10^{-2}$	$4.6{\cdot}10^{-2}$	$5.4{\cdot}10^{-2}$	$5.4{\cdot}10^{-2}$	$4.2{\cdot}10^{-2}$	$3.5{\cdot}10^{-2}$

	$s=3$				$s=4$				
	$(3,0,0)$	$(2,1,0)$	$(1,2,0)$	$(0,3,0)$	$(4,0,0)$	$(3,1,0)$	$(2,2,0)$	$(1,3,0)$	$(0,4,0)$
F	32.730	32.730	32.730	24.991	23.594	23.594	6.9092	6.4455	6.4455
DOF	12	14	16	20	28	32	36	40	48
$\|\,.\,\|_2$	$2.1{\cdot}10^{-2}$	$2.1{\cdot}10^{-2}$	$2.1{\cdot}10^{-2}$	$1.7{\cdot}10^{-2}$	$1.8{\cdot}10^{-2}$	$1.8{\cdot}10^{-2}$	$1.0{\cdot}10^{-2}$	$9.9{\cdot}10^{-3}$	$9.9{\cdot}10^{-3}$

Tabelle 4.22: FF-Identifizierung: Hierarchische Basen, dünnes Gitter

berücksichtigt und eine Gewichtung auf niedrige Konzentrationswerte angesetzt werden. Im Folgenden wird daher u.A. eine gewichtete Interpolation mit Gewichtungsfaktor $\lambda = 0.6$ verwendet (vgl. Anhang B.5.2). Dies gewährleistet in jedem hierarchischen Zwischenschritt eine lokal konvexe Startschätzung der neuen Knotenpunkte, ohne das eigentliche Identifizierungsproblem durch weitere Ungleichungen einzuschränken. Des Weiteren wird eine linksseitige Gewichtung der verwendeten Stützstellen herangezogen. Konkret wird unter Verwendung lokaler Basen die Gewichtung *WL I-4* und für hierarchische Basen *W-2/0.33* gewählt (vgl. Anhang B.4.1.2 bzw. B.4.1.3). In den Tabellen 4.23-4.25 finden sich die erzielten Resultate. Wie zu erwarten war, ist, unabhängig vom gewählten Basistyp, eine deutliche Verbesserung der zu identifizierenden Reaktionsrate festzustellen. Die beste Approximation wird, wie auch im vorangegangenen Beispiel, mit den lokalen Basen erzielt. Dies ist u.A. auf die linksgewichtete Diskretisierung *WL I-4* zurückzuführen. Während bei hierarchischen Basen, aufgrund der vorgegebenen Struktur, nur bedingt auf die Knotenunterteilung eingegangen werden kann, sind unter Verwendung lokaler Basen kaum Grenzen gesetzt. So konnten in diesem Fall bereits mit der geringen axialen Stützstellenzahl von $n^i = 6$, $i = D, A$, entscheidende Knotenpunkte nahe der linken Zulässigkeitsbereiche definiert werden.

Überraschend ist jedoch, dass die Verwendung voller Gitter scheinbar eine etwas bessere Lösung (bezogen auf F) liefert als die auf den dünnen Gittern basierende Berechnung. Wird allerdings die Qualität der ermittelten Reaktionsraten richtigerweise anhand der relevanten L^2-Norm gemessen, so zeigt sich, dass trotz alledem die dünnen Gitter zu bevorzugen sind. Zu bemerken ist an dieser Stelle jedoch, dass aufgrund der hohen Komplexität der zugehörigen Identifizierungsaufgabe nur sehr geringe Skalenwerte angenommen werden können. Demzufolge kommt der entscheidende Dimensionsvorteil dünner Gitter bei den vorliegenden Identifizierungsaufgaben kaum zu tragen. Zudem kommt, dass bei dünnen Gittern durch die verwendeten Blöcke $\vec{\sigma}=(4,0,0)$ und $\vec{\sigma}=(0,4,0)$ trotz geringerem DOF eine sehr feine Diskretisierung vorliegt, welche zu einer höheren Flexibilität und damit u.U. zu einem größeren Identifizierungsaufwand führen.

	n							
	$(2,2,2)$	$(3,2,2)$	$(3,3,2)$	$(4,3,2)$	$(4,4,2)$	$(5,4,2)$	$(5,5,2)$	$(6,5,2)$
F	832.72	175.55	69.323	57.453	13.249	12.659	1.3394	1.2832
DOF	1	2	4	6	9	12	16	20
$\|\,.\,\|_2$	$4.1{\cdot}10^{-2}$	$4.6{\cdot}10^{-2}$	$4.0{\cdot}10^{-2}$	$3.3{\cdot}10^{-2}$	$1.5{\cdot}10^{-2}$	$1.4{\cdot}10^{-2}$	$7.2{\cdot}10^{-3}$	$7.4{\cdot}10^{-3}$

	n						
	$(6,6,2)$	$(7,6,2)$	$(7,7,2)$	$(8,7,2)$	$(8,8,2)$	$(9,8,2)$	$(9,9,2)$
F	0.2647	0.1476	0.1118	0.1085	0.0952	0.0737	0.0584
DOF	25	30	36	42	49	56	64
$\|\,.\,\|_2$	$4.6{\cdot}10^{-3}$	$4.0{\cdot}10^{-3}$	$3.8{\cdot}10^{-3}$	$3.7{\cdot}10^{-3}$	$3.8{\cdot}10^{-3}$	$3.8{\cdot}10^{-3}$	$3.9{\cdot}10^{-3}$

Tabelle 4.23: FF-Identifizierung: Lokale Basen, linksgewichtet und konvex

	$s=0$	$s=1$			$s=2$				
	$(0,0,0)$	$(1,0,0)$	$(0,1,0)$	$(1,1,0)$	$(2,0,0)$	$(2,1,0)$	$(0,2,0)$	$(1,2,0)$	$(2,2,0)$
F	832.72	97.495	70.820	68.050	35.055	34.048	21.817	6.9948	1.7733
DOF	1	2	3	4	6	8	10	12	16
$\|\,.\,\|_2$	$4.1{\cdot}10^{-2}$	$5.2{\cdot}10^{-2}$	$5.2{\cdot}10^{-2}$	$5.2{\cdot}10^{-2}$	$2.1{\cdot}10^{-2}$	$2.1{\cdot}10^{-2}$	$1.7{\cdot}10^{-2}$	$8.9{\cdot}10^{-3}$	$6.3{\cdot}10^{-3}$

	$s=3$						
	$(3,0,0)$	$(3,1,0)$	$(3,2,0)$	$(0,3,0)$	$(1,3,0)$	$(2,3,0)$	$(3,3,0)$
F	1.2719	1.1895	1.1049	0.7809	2.0160	1.9422	1.5449
DOF	20	24	32	36	40	48	64
$\|\,.\,\|_2$	$6.5{\cdot}10^{-3}$	$6.4{\cdot}10^{-3}$	$6.4{\cdot}10^{-3}$	$5.3{\cdot}10^{-3}$	$8.6{\cdot}10^{-3}$	$8.8{\cdot}10^{-3}$	$8.4{\cdot}10^{-3}$

Tabelle 4.24: FF-Identifizierung: Hierarchische Basen, gewichtet, volles Gitter

In Abbildung 4.10 ist schließlich die hierarchische Entwicklung der zu identifizierenden Reaktionsrate (eingeschränkt) graphisch dargestellt. Da ein vollständiger Funktionsplot aufgrund des dreidimensionalen Definitionsbereiches nicht möglich ist, jedoch die dritte Komponente nur linear eingeht, wurde die ermittelte Ab-

	$s=0$	$s=1$		$s=2$		
	$(0,0,0)$	$(1,0,0)$	$(0,1,0)$	$(2,0,0)$	$(1,1,0)$	$(0,2,0)$
F	832.72	97.495	70.820	55.638	53.553	31.994
DOF	1	2	3	5	6	8
$\|\,.\,\|_2$	$4.1{\cdot}10^{-2}$	$5.2{\cdot}10^{-2}$	$5.2{\cdot}10^{-2}$	$3.7{\cdot}10^{-2}$	$3.8{\cdot}10^{-2}$	$2.1{\cdot}10^{-2}$

	$s=3$				$s=4$				
	$(3,0,0)$	$(2,1,0)$	$(1,2,0)$	$(0,3,0)$	$(4,0,0)$	$(3,1,0)$	$(2,2,0)$	$(1,3,0)$	$(0,4,0)$
F	10.077	10.016	2.7798	1.6281	1.7552	2.1904	2.3190	2.2076	2.1837
DOF	12	14	16	20	28	32	36	40	48
$\|\,.\,\|_2$	$1.1{\cdot}10^{-2}$	$1.1{\cdot}10^{-2}$	$6.9{\cdot}10^{-3}$	$6.2{\cdot}10^{-3}$	$6.3{\cdot}10^{-3}$	$6.5{\cdot}10^{-3}$	$6.5{\cdot}10^{-3}$	$6.3{\cdot}10^{-3}$	$6.1{\cdot}10^{-3}$

Tabelle 4.25: FF-Identifizierung: Hierarchische Basen, gewichtet, dünnes Gitter

baurate mit der, zur Simulation angenommenen, maximalen Biomassenkonzentration $c_B = 0.10\,\frac{\mu g}{cm^3}$ herangezogen und entsprechend $\mu(c_D, c_A, 0.10)$ ausgegeben. Als Berechnungsgrundlage wurde das auf lokalen Basen basierende Identifizierungsproblem gewählt, für welches bereits in Tabelle 4.23 die erzielten Resultate tabellarisch angegeben wurden. Es ist deutlich erkennbar, dass, beginnend mit einer äußerst groben Diskretisierung, die gesuchte Rate nach und nach immer besser approximiert wird. Auffällig ist auch, dass selbst das Krümmungsverhalten, welches sich in den beiden Koordinatenrichtungen unterschiedlich verhält, ausreichend gut bestimmt wurde. Erst bei einer (für die formfreie 3D-Identifizierung) feinen Unterteilung nimmt die Qualität der ermittelten Nichtlinearität wieder etwas ab. Da bisher noch keine künstlich erzeugten Messfehler eingebunden waren, ist die Ursache in den nur eingeschränkt vorliegenden Messdaten zu sehen. Für eine fehlerfreie Identifizierung müssten die gemessenen Konzentrationswerte an jeder Stelle/Höhe der Laborsäule vorliegen. Da dies technisch nicht möglich ist, muss entsprechend mit qualitativen Abstrichen gerechnet werden. Zusammenfassend wird bemerkt, dass die Anzahl der Freiheitsgrade, wie es bereits in Bitterlich [8] bei der formfreien Identifizierung eindimensionaler Nichtlinearitäten festgestellt wurde, nicht beliebig hoch gewählt werden sollte, da sonst mehr und mehr unerwünschte Stör- und Messeffekte die Charakteristik der Lösung verfälschen. Im Fall der hier vorliegenden 3D-Approximierung stellt dies jedoch keine weitere Einschränkung dar, da das zugehörige Idenfizierungsproblem bereits durch die hohe Komplexität auf eine grobe Diskretisierung beschränkt ist.

Die vorgestellten Identifizierungsresultate sind sehr vielversprechend. Allerdings ist zu berücksichtigen, dass die bisherigen Problemstellungen möglichst einfach gehalten wurden. So kann im Allgemeinen nicht davon ausgegangen werden, dass spezielle Messexperimente zur Bestimmung der beiden Feldfaktoren Y und $\alpha_{A/D}$ durchgeführt wurden. Folglich müssen diese, falls ihre Werte nicht exakt vorliegen,

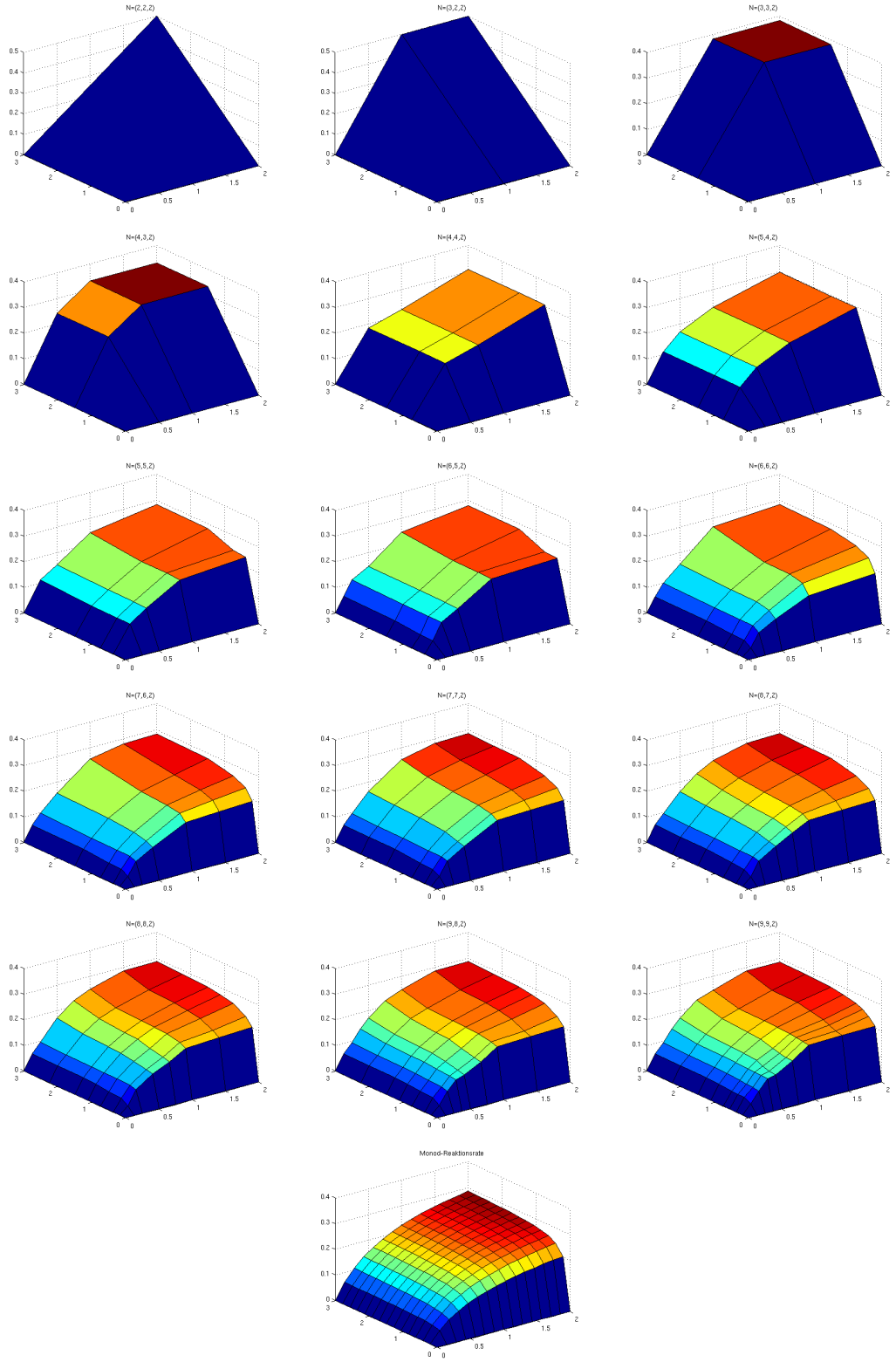

Abbildung 4.10: Hierarchische Entwicklung der identifizierten Reaktionsrate

mit in den Identifizierungsprozess eingebunden werden. Leider zeigt sich, dass damit die Komplexität des zugehörigen Identifizierungsproblems, trotz sehr geringer Erhöhung der Freiheitsgrade (um zwei), signifikant angehoben wird. Dies spiegelte sich auch in den durchgeführten Berechnungsbeispielen wider, so dass alle Identifizierungsversuche mit ernüchterndem Ergebnis endeten. Auch der im vorangegangenen Abschnitt vorgestellte rekursive Identifizierungsalgorithmus schaffte kaum Abhilfe. Im Gegensatz zu einer, auf einer fixen Parametrisierung basierenden, Identifizierung, ist hierfür vorrangig der enorme Rechenaufwand verantwortlich. Während bei einer groben Diskretisierung ein rekursiver Aufruf noch problemlos möglich ist, verbietet die begrenzt zur Verfügung stehende Rechenzeit gerade bei den relevanten Skalen ein mehrfaches Aufrufen. Unabhängig von der Wahl der zugrundegelegten Basisfunktionen und des verwendeten Optimierungsalgorithmusses wurde eine adäquate Lösung nur erreicht, wenn die Zulässigkeitsbereiche der beiden Parameter äußerst eingeschränkt vorgegeben wurden. Konkret war eine maximale Abweichung von höchstens 5% bzw. $Y \in [0.494, 0.546]$ und $\alpha_{A/D} \in [3.002, 3.318]$ erforderlich, um eine brauchbare Lösung zu erzielen. Da dies nicht ohne genaue Kenntnis der beiden Feldfaktoren möglich ist, bleiben vorangehende Untersuchungen und Messungen bzgl. dieser Faktoren unerlässlich.

Des Weiteren wurde bei den bisherigen Berechnungsbeispielen die Monotonie der gesuchten Reaktionsrate vorausgesetzt. Wird auf diese Einschränkung verzichtet um ggf. auch nichtmonotone Nichtlinearitäten identifizieren zu können, so steigt auch hier die Flexibilität des Identifizierungsproblems drastisch an. Zwar sind die ermittelten Residuen unabhängig vom Basistyp vertretbar gering, doch unter genauer Betrachtung zeigt sich, dass das gesuchte Steigungs- und Krümmungsverhalten keineswegs identifiziert werden konnte. Zusammenfassend muss damit leider festgestellt werden, dass, zumindest unter Verwendung der hier vorliegenden Experimente, eine Identifizierung nichtmonotoner Reaktionsraten mit dem vorgestellten formfreien 3D-Ansatz nicht möglich ist.

Bleibt die Empfindlichkeit auf Messstörungen zu untersuchen. Hierzu werden in den folgenden Berechnungsbeispielen wieder die beiden Feldfaktoren Y und $\alpha_{A/D}$ exakt vorgegeben und ein monotones Steigungsverhalten verlangt. Die vorgegebenen Störungen werden, wie auch bei den Untersuchungen des rekursiven Identifizierungsansatzes, mittels der in (4.7) definierten zufallsgenerierten Gewichtung unter Verwendung von $|\varepsilon_{i,j}| \in [0, 0.02]$ bzw. $|\varepsilon_{i,j}| \in [0, 0.05]$ gestört. In Tabelle 4.26 finden sich die erzielten Identifizierungsresultate. Es zeigt sich, dass die formfreie Identifizierung auf geringe Störungen unempfindlich reagiert. Erst bei höheren Messtoleranzen nimmt die Qualität der ermittelten Reaktionsrate,

insbesondere unter Verwendung eines vollen Gitters, signifikant ab. Auffällig ist jedoch, dass das auf dünnen Gittern basierende Identifizierungsproblem, selbst bei $|\varepsilon_{i,j}| \in [0, 0.05]$ die gesuchte Nichtlinearität sehr gut identifiziert. Hierfür ist vor allem die stabilisierende hierarchische Splinestruktur verantwortlich. Mit Hilfe der dünnen Gitter kann diese ohne den Nachteil eines hohen DOF sehr gut ausgenutzt werden. Selbst die lokalen Basen, welche bei den ungestörten Problemen die erste Wahl darstellten, liefern trotz vergleichbaren Fehlerfunktionals eine deutlich höhere Abweichung zur gesuchten Reaktionsrate.

	lokale Basen		volles Gitter		dünnes Gitter	
Level	$n=(9,9,2)$		$s=3,\ \vec{\sigma}=(3,3,0)$		$s=4,\ \vec{\sigma}=(0,4,0)$	
DOF	64		64		48	
$\lvert\varepsilon_{i,j}\rvert$	0.02	0.05	0.02	0.05	0.02	0.05
F	1.0095	5.9898	2.3881	13.975	1.6707	6.0483
$\lVert . \rVert_2$	$3.9{\cdot}10^{-3}$	$5.6{\cdot}10^{-3}$	$7.0{\cdot}10^{-3}$	$1.8{\cdot}10^{-2}$	$4.9{\cdot}10^{-3}$	$2.9{\cdot}10^{-3}$

Tabelle 4.26: Formfreie Identifizierung unter Verwendung gestörter Messdaten

In Abbildung 4.11 sind schließlich die zugehörigen Reaktionsraten $\mu(c_D, c_A, 0.10)$ (unter Verwendung der maximalen Biomassenkonzentration $c_B = 0.10\ \frac{\mu g}{cm^3}$) des jeweils höchsten Diskretisierungslevels graphisch ausgegeben. Auch hier ist erkennbar, dass, wie bereits weiter oben festgestellt, insbesondere bei starken Messstörungen die Verwendung dünner Gitter zu bevorzugen ist. Diese erreicht, im Gegensatz zu den beiden anderen Spline-Varienten, auch eine adäquate Identifizierung in dem (durch das vorliegende Experiment) schwer zu bestimmenden Definitionsbereich hoher Konzentrationen. Ebenso ist deutlich der Nachteil der lokalen Basen erkennbar. Im Gegensatz zu den hierarchischen Basen können sich dort, aufgrund der eingeschränkten Träger, lokale Unebenheiten viel deutlicher ausprägen.

Abschließend wurde noch das bereits im Abschnitt 4.2.2 vorgestellte Schadstoffabbauproblem mit der in (4.10) definierten und entsprechend zu bestimmenden Reaktionsrate untersucht. Im Gegensatz zur gescheiterten Identifizierung unter Verwendung der fixen Monod-Parametrisierung, zeigte sich, dass diese mit dem formfreien Ansatz sehr wohl adäquat ($\|.\|_2 = 3.3{\cdot}10^{-3}$) bestimmt werden kann.

Zusammenfassend lässt sich feststellen, dass die formfreie 3D-Identifizierung eine ernstzunehmende Alternative zur Identifizierung fixer Parametrisierungen darstellt. Inbesondere bei mit (Mess-)Störungen behafteter Daten oder komplexen chemischen Reaktionen, welche nicht vollständig durch eine fixe Parametrisierung darstellbar sind, ist die Verwendung eines formfreien Ansatzes vorteilhaft.

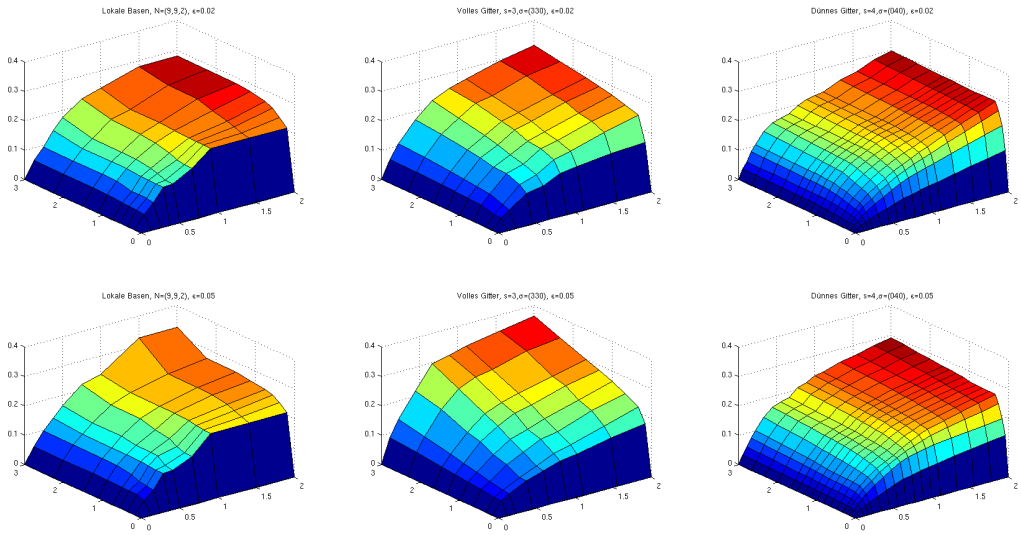

Abbildung 4.11: Formfreie Identifizierung unter Verwendung gestörter Messdaten

Jedoch ist zu beachten, dass die gesuchte Rate nicht zu stark nichtlinear sein darf, da ansonsten eine splinebasierende Approximation mittels weniger Stützstellen unzureichend wäre. Ebenso sollten grundlegende Eigenschaften, wie ein vorliegendes Monotonie- und/oder Krümmungsverhalten bekannt sein, da ansonsten, aufgrund der hohen Flexibilität, trotz geringem Residuum, das Identifizierungsergebnis möglicherweise unbrauchbar sein könnte.

Anhang A

Untersuchungen der van Genuchten-Mualem-Parametrisierung

In den folgenden Abschnitten werden für die im Kapitel 2.1.4 vorgestellte hydraulische Leitfähigkeitsfunktion

$$K(\Phi) = K_{\text{sat}}\sqrt{\Phi}\Big(1-\big(1-\Phi^{\frac{1}{m}}\big)^{m}\Big)^{2}, \quad K_{\text{sat}}\in\mathbb{R}^{+}, \tag{A.1}$$

mit

$$\Phi(\psi) = \frac{1}{\big(1+(-\alpha\psi)^{n}\big)^{m}} \tag{A.2}$$

und

$$m = 1-\frac{1}{n}, \quad n>1, \quad \alpha\in\mathbb{R}^{+}, \tag{A.3}$$

die ersten beiden Ableitungen nach der Sättigung Φ vorgestellt und strenge Monotonie, Konkavität und Eindeutigkeit nachgewiesen. Im Anschluß wird die Äquivalenz zu

$$K(\psi) = K_{\text{sat}}\frac{\Big(1-(-\alpha\psi)^{n-1}\big(1+(-\alpha\psi)^{n}\big)^{\frac{1-n}{n}}\Big)^{2}}{\big(1+(-\alpha\psi)^{n}\big)^{\frac{n-1}{2n}}}$$

aufgezeigt und $K(\psi)$ auf entsprechende Eigenschaften überprüft.

A.1 Sättigungsabhängige Leitfähigkeit $K(\Phi)$

A.1.1 Erste und zweite Ableitung nach Φ

Die erste Ableitung der in (A.1) definierten hydraulischen Leitfähigkeit nach der Sättigung Φ lässt sich angeben durch

$$\begin{aligned}\frac{dK}{d\Phi}(\Phi) &= K_{\text{sat}}\frac{1}{2\sqrt{\Phi}}\Big(1-\big(1-\Phi^{\frac{1}{m}}\big)^m\Big)^2+\\ &\qquad 2K_{\text{sat}}\sqrt{\Phi}\Big(1-\big(1-\Phi^{\frac{1}{m}}\big)^m\Big)(-m)\big(1-\Phi^{\frac{1}{m}}\big)^{m-1}\big(-\tfrac{1}{m}\big)\Phi^{\frac{1}{m}-1}\\ &= \frac{K_{\text{sat}}}{2\sqrt{\Phi}}\Big(1-\big(1-\Phi^{\frac{1}{m}}\big)^m\Big)\Big(1-\big(1-\Phi^{\frac{1}{m}}\big)^m+4\big(1-\Phi^{\frac{1}{m}}\big)^{m-1}\Phi^{\frac{1}{m}}\Big)\\ &= \frac{K_{\text{sat}}}{2\sqrt{\Phi}}\Big(1-\big(1-\Phi^{\frac{1}{m}}\big)^m\Big)\Big(1+\big(5\Phi^{\frac{1}{m}}-1\big)\big(1-\Phi^{\frac{1}{m}}\big)^{m-1}\Big). \qquad \text{(A.4)}\end{aligned}$$

Entsprechend gilt für die zweite Ableitung

$$\begin{aligned}\frac{d^2K}{d\Phi^2}(\Phi) = \frac{-K_{\text{sat}}}{4\Phi^{\frac{3}{2}}}&\Big(1-\big(1-\Phi^{\frac{1}{m}}\big)^m\Big)\Big(1+\big(5\Phi^{\frac{1}{m}}-1\big)\big(1-\Phi^{\frac{1}{m}}\big)^{m-1}\Big)+\\ \frac{K_{\text{sat}}}{2\sqrt{\Phi}}&\big(1-\Phi^{\frac{1}{m}}\big)^{m-1}\Phi^{\frac{1}{m}-1}\Big(1+\big(5\Phi^{\frac{1}{m}}-1\big)\big(1-\Phi^{\frac{1}{m}}\big)^{m-1}\Big)+\\ \frac{K_{\text{sat}}}{2\sqrt{\Phi}}&\Big(1-\big(1-\Phi^{\frac{1}{m}}\big)^m\Big)\Big(\tfrac{5}{m}\Phi^{\frac{1}{m}-1}\big(1-\Phi^{\frac{1}{m}}\big)^{m-1}-\\ &\tfrac{m-1}{m}\big(5\Phi^{\frac{1}{m}}-1\big)\big(1-\Phi^{\frac{1}{m}}\big)^{m-2}\Phi^{\frac{1}{m}-1}\Big).\end{aligned}$$

Mit

$$2\big(1-\Phi^{\frac{1}{m}}\big)^{m-1}\Phi^{\frac{1}{m}}-1+\big(1-\Phi^{\frac{1}{m}}\big)^m = \big(1-\Phi^{\frac{1}{m}}\big)^{m-1}\big(1+\Phi^{\frac{1}{m}}\big)-1$$

und

$$\begin{aligned}5\Phi^{\frac{1}{m}}\big(1-\Phi^{\frac{1}{m}}\big)^{m-1}-(m-1)\big(5\Phi^{\frac{1}{m}}-1\big)\big(1-\Phi^{\frac{1}{m}}\big)^{m-2}\Phi^{\frac{1}{m}} &=\\ \big(1-\Phi^{\frac{1}{m}}\big)^{m-2}\Big(5\big(1-\Phi^{\frac{1}{m}}\big)-(m-1)\big(5\Phi^{\frac{1}{m}}-1\big)\Big)\Phi^{\frac{1}{m}} &=\\ \big(1-\Phi^{\frac{1}{m}}\big)^{m-2}\big(4+m-5m\Phi^{\frac{1}{m}}\big)\Phi^{\frac{1}{m}}&\end{aligned}$$

folgt schließlich

$$\begin{aligned}\frac{d^2K}{d\Phi^2}(\Phi) = \frac{K_{\text{sat}}}{4\Phi^{\frac{3}{2}}}\Big[&\Big(\big(1-\Phi^{\frac{1}{m}}\big)^{m-1}\big(1+\Phi^{\frac{1}{m}}\big)-1\Big)\Big(1+\big(5\Phi^{\frac{1}{m}}-1\big)\big(1-\Phi^{\frac{1}{m}}\big)^{m-1}\Big)+\\ &2\Big(1-\big(1-\Phi^{\frac{1}{m}}\big)^m\Big)\big(1-\Phi^{\frac{1}{m}}\big)^{m-2}\big(\tfrac{4}{m}+1-5\Phi^{\frac{1}{m}}\big)\Phi^{\frac{1}{m}}\Big]. \qquad \text{(A.5)}\end{aligned}$$

A.1.2 Monotonie und Konkavität

Lemma A.1
Die Sättigungsfunktion (A.2) ist auf $(-\infty, 0]$ für alle zulässigen Parametersätze (A.3) streng monoton wachsend.

Beweis:
Eine hinreichende Bedingung liefert die erste Ableitung. Es gilt

$$\frac{d\Phi}{d\psi}(\psi) = -m\big(1+(-\alpha\psi)^n\big)^{-m-1} n(-\alpha\psi)^{n-1}(-\alpha)$$
$$= \frac{\alpha(n-1)(-\alpha\psi)^{n-1}}{\big(1+(-\alpha\psi)^n\big)^{m+1}} = \frac{\alpha(n-1)(-\alpha\psi)^{n-1}}{1+(-\alpha\psi)^n}\Phi(\psi) > 0 \quad \text{(A.6)}$$

und damit bereits die Behauptung.

□

Satz A.2
Die nach (A.1) definierte (sättigungsabhängige) hydraulische Leitfähigkeit ist für $m \in (0,1)$ streng monoton steigend und konkav.

Beweis:
Nach (A.4) kann die erste Ableitung von (A.1) nach der Sättigung Φ durch

$$K'(\Phi) = \frac{K_{\text{sat}}}{2\sqrt{\Phi}}\Big(1-\big(1-\Phi^{\frac{1}{m}}\big)^m\Big)\Big(1+\big(5\Phi^{\frac{1}{m}}-1\big)\big(1-\Phi^{\frac{1}{m}}\big)^{m-1}\Big)$$

angegeben werden. Da für $\Phi, m \in (0,1)$, sowohl

$$\big(1-\Phi^{\frac{1}{m}}\big)^{m-1} > 1 \quad \text{(A.7)}$$

und damit

$$1 + \big(5\Phi^{\frac{1}{m}}-1\big)\big(1-\Phi^{\frac{1}{m}}\big)^{m-1} \geq 5\Phi^{\frac{1}{m}} > 0 \quad \text{(A.8)}$$

als auch

$$0 < \big(1-\Phi^{\frac{1}{m}}\big)^m < 1 \quad \text{(A.9)}$$

gilt, folgt bereits die erste Behauptung bezüglich der Monotonie. Die zweite Ableitung lässt sich nach (A.5) durch

$$\frac{d^2K}{d\Phi^2}(\Phi) = \frac{K_{\text{sat}}}{4\Phi^{\frac{3}{2}}}\Bigg[\Big(\big(1-\Phi^{\frac{1}{m}}\big)^{m-1}\big(1+\Phi^{\frac{1}{m}}\big)-1\Big)\Big(1+\big(5\Phi^{\frac{1}{m}}-1\big)\big(1-\Phi^{\frac{1}{m}}\big)^{m-1}\Big)+$$
$$2\Big(1-\big(1-\Phi^{\frac{1}{m}}\big)^m\Big)\big(1-\Phi^{\frac{1}{m}}\big)^{m-2}\big(\tfrac{4}{m}+1-5\Phi^{\frac{1}{m}}\big)\Phi^{\frac{1}{m}}\Bigg]$$

angeben. Da neben den Ungleichungen (A.7)-(A.9) auch

$$\left(1-\Phi^{\frac{1}{m}}\right)^{m-2} > 1$$

und

$$\tfrac{4}{m}+1-5\Phi^{\frac{1}{m}} > 5-5\Phi^{\frac{1}{m}} > 0$$

gilt, folgt

$$\frac{d^2K}{d\Phi^2}(\Phi) > 0$$

und damit entsprechend die zweite Behauptung.

□

A.1.3 Eindeutigkeit

Satz A.3
Seien zwei van Genuchten-Mualem-Leitfähigkeitsfunktionen der Form

$$K_i(\Phi) := K_{sat}\sqrt{\Phi}\left(1-\left(1-\Phi^{\frac{1}{m_i}}\right)^{m_i}\right)^2,\ i=1,2\,, \tag{A.10}$$

mit $m_i \in (0,1)$, $m_1 \neq m_2$ *und* $K_{sat} \in \mathbb{R}^+$ *gegeben. Dann gilt für alle* $\Phi \in (0,1)$

$$K_1(\Phi) \neq K_2(\Phi)\,.$$

Beweis:
Angenommen es exisitiert ein $\Phi_s \in (0,1)$ mit

$$K_1(\Phi_s) = K_2(\Phi_s)\,,$$

dann folgt wegen $K_{\text{sat}} \in \mathbb{R}^+$ direkt

$$\left(1-\Phi_s^{\frac{1}{m_1}}\right)^{m_1} = \left(1-\Phi_s^{\frac{1}{m_2}}\right)^{m_2}.$$

Ohne Einschränkung der Allgemeinheit kann $m_1 > m_2$ angenommen werden, so dass wegen $\Phi_s \in (0,1)$ und $m_i \in (0,1)$, $i=1,2$, auch

$$1-\Phi_s^{\frac{1}{m_i}} \in (0,1) \tag{A.11}$$

gilt und entsprechend

$$1-\Phi_s^{\frac{1}{m_1}} < 1-\Phi_s^{\frac{1}{m_2}}$$

folgt. Hieraus folgt aber unter Beachtung von (A.11) bereits der Widerspruch

$$\left(1-\Phi_s^{\frac{1}{m_1}}\right)^{m_1} < \left(1-\Phi_s^{\frac{1}{m_1}}\right)^{m_2} < \left(1-\Phi_s^{\frac{1}{m_2}}\right)^{m_2}. \tag{A.12}$$

□

Bemerkung A.4
Aus (A.12) folgt direkt, dass für zwei nach (A.10) mit $1>m_1>m_2>0$ definierten hydraulischen Leitfähigkeitsfunktionen

$$K_1(\Phi) > K_2(\Phi) \quad \forall\, \Phi \in (0,1)$$

gilt.

A.2 Druckabhängige Leitfähigkeit $K(\psi)$

Lemma A.5
Die in (A.1) mit (A.2) definierte Parametrisierung ist äquivalent zu

$$K(\psi) = K_{sat}\frac{\Big(1-(-\alpha\psi)^{n-1}\big(1+(-\alpha\psi)^n\big)^{\frac{1-n}{n}}\Big)^2}{\big(1+(-\alpha\psi)^n\big)^{\frac{n-1}{2n}}}\,. \tag{A.13}$$

Beweis:
Durch Einsetzen folgt direkt

$$K(\psi) = K_{\mathrm{sat}}\frac{\Big(1-\big(1-\frac{1}{1+(-\alpha\psi)^n}\big)^m\Big)^2}{\big(1+(-\alpha\psi)^n\big)^{\frac{m}{2}}} = \frac{\left(1-\left(\frac{(-\alpha\psi)^n}{1+(-\alpha\psi)^n}\right)^m\right)^2}{\big(1+(-\alpha\psi)^n\big)^{\frac{m}{2}}}\,.$$

Da $m=\frac{n-1}{n}$ gilt, folgt wegen

$$\left(\frac{(-\alpha\psi)^n}{1+(-\alpha\psi)^n}\right)^{\frac{n-1}{n}} = (-\alpha\psi)^{n-1}\big(1+(-\alpha\psi)^n\big)^{\frac{1-n}{n}}$$

bereits die Behauptung.

□

A.2.1 Erste und zweite Ableitung nach ψ

Aufgrund der in Abschnitt A.2 nachgewiesenen Äquivalenz kann sowohl (A.13) als auch (A.1) mit (A.2) zur Berechnung der Ableitung verwendet werden. Unter Anwendung von (A.6) folgt

$$\Big(\Phi(\psi)^{\frac{1}{m}}\Big)' = \frac{1}{m}\Phi(\psi)^{\frac{1}{m}-1}\Phi'(\psi) = \frac{\alpha n(-\alpha\psi)^{n-1}}{1+(-\alpha\psi)^n}\Phi(\psi)^{\frac{1}{m}} = \frac{\alpha n(-\alpha\psi)^{n-1}}{\big(1+(-\alpha\psi)^n\big)^2}$$

und damit

$$\begin{aligned}\frac{dK}{d\psi}(\psi) = {} & K_{\text{sat}}\frac{\Phi(\psi)}{2\sqrt{\Phi(\psi)}}\frac{\alpha(n-1)(-\alpha\psi)^{n-1}}{1+(-\alpha\psi)^n}\Big(1-\big(1-\Phi(\psi)^{\frac{1}{m}}\big)^m\Big)^2 + \\ & 2K_{\text{sat}}\sqrt{\Phi(\psi)}\Big(1-\big(1-\Phi(\psi)^{\frac{1}{m}}\big)^m\Big)(-m)\big(1-\Phi(\psi)^{\frac{1}{m}}\big)^{m-1}\frac{n(-\alpha\psi)^{n-1}(-\alpha)}{\big(1+(-\alpha\psi)^n\big)^2} \\ = {} & \frac{1}{2}K_{\text{sat}}\sqrt{\Phi(\psi)}\,\frac{\alpha(-\alpha\psi)^{n-1}}{1+(-\alpha\psi)^n}\Big(1-\big(1-\Phi(\psi)^{\frac{1}{m}}\big)^m\Big) \\ & \cdot\left[(n-1)\Big(1-\big(1-\Phi(\psi)^{\frac{1}{m}}\big)^m\Big) + \big(1-\Phi(\psi)^{\frac{1}{m}}\big)^{m-1}\frac{4mn}{1+(-\alpha\psi)^n}\right].\end{aligned}$$

Durch weiteres Umstellen erhält man schließlich

$$\begin{aligned} = {} & \frac{1}{2}K_{\text{sat}}\sqrt{\Phi(\psi)}\,\frac{\alpha(n-1)(-\alpha\psi)^{n-1}}{1+(-\alpha\psi)^n}\Big(1-\big(1-\Phi(\psi)^{\frac{1}{m}}\big)^m\Big) \\ & \cdot\left[1-\big(1-\Phi(\psi)^{\frac{1}{m}}\big)^{m-1}\left(1-\Phi(\psi)^{\frac{1}{m}}-\frac{4}{1+(-\alpha\psi)^n}\right)\right] \\ = {} & \frac{1}{2}K_{\text{sat}}\sqrt{\Phi(\psi)}\,\frac{\alpha(n-1)(-\alpha\psi)^{n-1}}{1+(-\alpha\psi)^n}\Big(1-\big(1-\Phi(\psi)^{\frac{1}{m}}\big)^m\Big) \\ & \cdot\left[1-\big(1-\Phi(\psi)^{\frac{1}{m}}\big)^{m-1}\frac{(-\alpha\psi)^n-4}{1+(-\alpha\psi)^n}\right]. \qquad \text{(A.14)}\end{aligned}$$

Um die zweite Ableitung angeben zu können empfiehlt es sich vorab

$$\begin{aligned}\left(\frac{\alpha(n-1)(-\alpha\psi)^{n-1}}{1+(-\alpha\psi)^n}\right)' = {} & \\ = {} & \alpha(n-1)\frac{(n-1)(-\alpha\psi)^{n-2}(-\alpha)\big(1+(-\alpha\psi)^n\big)-(-\alpha\psi)^{n-1}n(-\alpha\psi)^{n-1}(-\alpha)}{\big(1+(-\alpha\psi)^n\big)^2} \\ = {} & \alpha^2(1-n)\frac{(-\alpha\psi)^{n-2}\Big((n-1)\big(1+(-\alpha\psi)^n\big)-n(-\alpha\psi)^n\Big)}{\big(1+(-\alpha\psi)^n\big)^2} \\ = {} & \alpha^2(1-n)\frac{(-\alpha\psi)^{n-2}\Big(n-1-(-\alpha\psi)^n\Big)}{\big(1+(-\alpha\psi)^n\big)^2}\end{aligned}$$

und

$$\begin{aligned}\left(\frac{(-\alpha\psi)^n-4}{1+(-\alpha\psi)^n}\right)' &= \frac{n(-\alpha\psi)^{n-1}(-\alpha)\big(1+(-\alpha\psi)^n\big)-\big((-\alpha\psi)^n-4\big)n(-\alpha\psi)^{n-1}(-\alpha)}{\big(1+(-\alpha\psi)^n\big)^2} \\ &= \frac{-5\alpha n(-\alpha\psi)^{n-1}}{\big(1+(-\alpha\psi)^n\big)^2}\end{aligned}$$

zu berechnen. Damit gilt

$$\begin{aligned}\frac{d^2K}{d\psi^2}(\psi) = \frac{K_{\text{sat}}\Phi(\psi)}{4\sqrt{\Phi(\psi)}}&\left(\frac{\alpha(n-1)(-\alpha\psi)^{n-1}}{1+(-\alpha\psi)^n}\right)^2\left(1-\left(1-\Phi(\psi)^{\frac{1}{m}}\right)^m\right)\\ &\cdot\left[1-\left(1-\Phi(\psi)^{\frac{1}{m}}\right)^{m-1}\frac{(-\alpha\psi)^n-4}{1+(-\alpha\psi)^n}\right]\\ +\frac{1}{2}K_{\text{sat}}\sqrt{\Phi(\psi)}\,\alpha^2(1-&n)\,\frac{(-\alpha\psi)^{n-2}\left(n-1-(-\alpha\psi)^n\right)}{\left(1+(-\alpha\psi)^n\right)^2}\left(1-\left(1-\Phi(\psi)^{\frac{1}{m}}\right)^m\right)\\ &\cdot\left[1-\left(1-\Phi(\psi)^{\frac{1}{m}}\right)^{m-1}\frac{(-\alpha\psi)^n-4}{1+(-\alpha\psi)^n}\right]\\ +\frac{1}{2}K_{\text{sat}}\sqrt{\Phi(\psi)}&\frac{\alpha(n-1)(-\alpha\psi)^{n-1}}{1+(-\alpha\psi)^n}\,(-m)\left(1-\Phi(\psi)^{\frac{1}{m}}\right)^{m-1}\frac{n(-\alpha\psi)^{n-1}(-\alpha)}{\left(1+(-\alpha\psi)^n\right)^2}\\ &\cdot\left[1-\left(1-\Phi(\psi)^{\frac{1}{m}}\right)^{m-1}\frac{(-\alpha\psi)^n-4}{1+(-\alpha\psi)^n}\right]\\ +\frac{1}{2}K_{\text{sat}}\sqrt{\Phi(\psi)}&\,\frac{\alpha(n-1)(-\alpha\psi)^{n-1}}{1+(-\alpha\psi)^n}\left(1-\left(1-\Phi(\psi)^{\frac{1}{m}}\right)^m\right)\\ &\cdot\left[(1-m)\left(1-\Phi(\psi)^{\frac{1}{m}}\right)^{m-2}\frac{n(-\alpha\psi)^{n-1}(-\alpha)}{\left(1+(-\alpha\psi)^n\right)^2}\,\frac{(-\alpha\psi)^n-4}{1+(-\alpha\psi)^n}+\right.\\ &\left.+\left(1-\Phi(\psi)^{\frac{1}{m}}\right)^{m-1}\frac{5\alpha n(-\alpha\psi)^{n-1}}{\left(1+(-\alpha\psi)^n\right)^2}\right].\end{aligned}$$

Durch Ausklammern folgt

$$\begin{aligned}\frac{d^2K}{d\psi^2}(\psi) = \frac{K_{\text{sat}}\sqrt{\Phi(\psi)}(n-1)\alpha^2(-\alpha\psi)^{n-2}}{2\left(1+(-\alpha\psi)^n\right)^2}&\left(\left[\left(1-n+\frac{1}{2}(n+1)(-\alpha\psi)^n\right)\right.\right.\\ \cdot\left(1-\left(1-\Phi(\psi)^{\frac{1}{m}}\right)^m\right)+\left(1-\Phi(\psi)^{\frac{1}{m}}\right)^{m-1}&\left.\frac{mn(-\alpha\psi)^n}{1+(-\alpha\psi)^n}\right]\left[1-\left(1-\Phi(\psi)^{\frac{1}{m}}\right)^{m-1}\right.\\ \cdot\frac{(-\alpha\psi)^n-4}{1+(-\alpha\psi)^n}\Big]+\left(1-\left(1-\Phi(\psi)^{\frac{1}{m}}\right)^m\right)&\frac{n(-\alpha\psi)^n}{1+(-\alpha\psi)^n}\left(1-\Phi(\psi)^{\frac{1}{m}}\right)^{m-2}\\ &\left.\cdot\left[5\left(1-\Phi(\psi)^{\frac{1}{m}}\right)+(m-1)\frac{(-\alpha\psi)^n-4}{1+(-\alpha\psi)^n}\right]\right)\end{aligned}$$

und wegen

$$1-\Phi(\psi)^{\frac{1}{m}} = 1-\frac{1}{1+(-\alpha\psi)^n} = \frac{(-\alpha\psi)^n}{1+(-\alpha\psi)^n}$$

und entsprechend

$$\left(1-\Phi(\psi)^{\frac{1}{m}}\right)^m = \left(\frac{(-\alpha\psi)^n}{1+(-\alpha\psi)^n}\right)^m = \frac{(-\alpha\psi)^{n-1}}{\left(1+(-\alpha\psi)^n\right)^m},$$

sowie

$$\left(1-\Phi(\psi)^{\frac{1}{m}}\right)^{m-1} = \frac{(-\alpha\psi)^{-1}}{\left(1+(-\alpha\psi)^n\right)^{-\frac{1}{n}}}$$

und

$$\left(1-\Phi(\psi)^{\frac{1}{m}}\right)^{m-2} = \frac{(-\alpha\psi)^{-n-1}}{\left(1+(-\alpha\psi)^n\right)^{-\frac{1}{n}-1}}$$

auch

$$\begin{aligned}5\left(1-\Phi(\psi)^{\frac{1}{m}}\right)+(m-1)\frac{(-\alpha\psi)^n-4}{1+(-\alpha\psi)^n} &= \frac{5(-\alpha\psi)^n+(m-1)\left((-\alpha\psi)^n-4\right)}{1+(-\alpha\psi)^n}\\ &= \frac{(4+m)(-\alpha\psi)^n+4(1-m)}{1+(-\alpha\psi)^n}\end{aligned}$$

und damit schließlich

$$\begin{aligned}\frac{d^2K}{d\psi^2}(\psi) = \frac{K_{\text{sat}}(n-1)\alpha^2(-\alpha\psi)^{n-2}}{2\left(1+(-\alpha\psi)^n\right)^{\frac{5n-1}{2n}}}\Bigg(&\left[\left(1-n+\frac{1}{2}(n+1)(-\alpha\psi)^n\right)\right.\\ &\cdot\left(1-\frac{(-\alpha\psi)^{n-1}}{\left(1+(-\alpha\psi)^n\right)^m}\right)+\left.\frac{(n-1)(-\alpha\psi)^{n-1}}{\left(1+(-\alpha\psi)^n\right)^m}\right]\left[1-\frac{(-\alpha\psi)^{n-1}-4(-\alpha\psi)^{-1}}{\left(1+(-\alpha\psi)^n\right)^m}\right]\\ &+\left[1-\frac{(-\alpha\psi)^{n-1}}{\left(1+(-\alpha\psi)^n\right)^m}\right]\frac{(5n-1)(-\alpha\psi)^{n-1}+4(-\alpha\psi)^{-1}}{\left(1+(-\alpha\psi)^n\right)^m}\Bigg).\end{aligned} \tag{A.15}$$

A.2.2 Monotonie und Konkavität

Satz A.6

Die nach (A.13) mit $K_{sat}, \alpha \in \mathbb{R}^+$ und $n>1$ definierte druckabhängige Leitfähigkeitsfunktion ist auf $(-\infty, 0]$ streng monoton wachsend.

Beweis:

Da alle relevanten Terme in (A.14) für $\psi \in (\infty, 0)$ positiv sind, folgt bereits die Behauptung.

□

Satz A.7
Die nach (A.13) mit $K_{sat}, \alpha \in \mathbb{R}^+$ definierte druckabhängige Leitfähigkeitsfunktion ist für

$$-\frac{1}{\alpha}\left(\frac{2(5n-1)}{n+1}\right)^{\frac{1}{n}} \leq \psi < 0$$

konkav genau dann, wenn $1 < n \leq 2$ gilt.

Beweis:
Um Aussagen bzgl. der Konkavität von (A.13) zu treffen, wird das Vorzeichen der zweiten Ableitung untersucht. Durch Einsetzen der Sättigungsfunktion (A.2) in (A.15) und Ausmultiplizieren aller inneren Klammern folgt

$$\begin{aligned}\frac{d^2K}{d\psi^2}(\psi) = \frac{K_{\text{sat}}(n-1)\alpha^2(-\alpha\psi)^{n-2}}{2\left(1+(-\alpha\psi)^n\right)^{\frac{5n-1}{2n}}}\Bigg(&1-n+4(2-n)\Phi(\psi)(-\alpha\psi)^{-1}\\ &+4(2n-3)\Phi^2(\psi)(-\alpha\psi)^{n-2}+2(5n-1)\Phi(\psi)(-\alpha\psi)^{n-1}\\ &+\frac{1}{2}(n+1)(-\alpha\psi)^n-(9n-1)\Phi^2(\psi)(-\alpha\psi)^{2n-2}\\ &-(n+1)\Phi(\psi)(-\alpha\psi)^{2n-1}+\frac{1}{2}(n+1)\Phi^2(\psi)(-\alpha\psi)^{3n-2}\Bigg). \quad \text{(A.16)}\end{aligned}$$

Unabhängig von $n > 1$ gilt für $\psi < 0$

$$\frac{K_{\text{sat}}(n-1)\alpha^2(-\alpha\psi)^{n-2}}{2\left(1+(-\alpha\psi)^n\right)^{\frac{5n-1}{2n}}} > 0 \quad \text{(A.17)}$$

und $\Phi(\psi) \in (0,1)$. Da für (fest gewähltes) $n > 2$ mit Ausnahme von

$$\lim_{(-\alpha\psi)\overset{>}{\to}0} 4(2-n)(-\alpha\psi)^{-1} = -\infty$$

alle ψ-abhängigen Terme für $\psi \overset{<}{\to} 0$ gegen Null konvergieren, gilt

$$\lim_{(-\alpha\psi)\overset{>}{\to}0} \frac{d^2K}{d\psi^2}(\psi) < 0$$

und damit bereits die erste Behauptung. Bleibt die Konkavität für $1 < n \leq 2$ nachzuweisen. Da (A.17) für alle $n > 1$ und $\psi < 0$ gilt, genügt es weiterhin das Vorzeichen der in (A.16) aufgeführten Klammer zu untersuchen. Durch einfaches Umstellen lässt sich zeigen, dass sich diese äquivalent zu

$$
\begin{aligned}
f(\psi) := 1-n+\frac{1}{2}(n+1)(-\alpha\psi)^n \\
+\left(4(2-n)(-\alpha\psi)^{-1}+2(5n-1)(-\alpha\psi)^{n-1}-(n+1)(-\alpha\psi)^{2n-1}\right)\Phi(\psi) \\
+\left(4(2n-3)(-\alpha\psi)^{n-2}-(9n-1)(-\alpha\psi)^{2n-2}+\frac{1}{2}(n+1)(-\alpha\psi)^{3n-2}\right)\Phi^2(\psi)
\end{aligned}
\tag{A.18}
$$

verhält. Im Folgenden sei zunächst $0<-\alpha\psi\leq 1$ betrachtet. Wegen

$$
1<\left(1+(-\alpha\psi)^n\right)^{2m}=\frac{\left(1+(-\alpha\psi)^n\right)^2}{\left(1+(-\alpha\psi)^n\right)^{\frac{2}{n}}}<\frac{\left(1+(-\alpha\psi)^n\right)^2}{1+(-\alpha\psi)^n}<1+(-\alpha\psi)^n
$$

und

$$
\begin{aligned}
&4\underbrace{(2-n)}_{>0}(-\alpha\psi)^{-1}+2(5n-1)(-\alpha\psi)^{n-1}-(n+1)(-\alpha\psi)^{2n-1}> \\
&\qquad\Big(4(2-n)+2(5n-1)-(n+1)\Big)(-\alpha\psi)^{2n-1}=5(n+1)(-\alpha\psi)^{2n-1}>0
\end{aligned}
$$

folgt aus (A.18) die Ungleichung

$$
\begin{aligned}
f(\psi)>\Bigg((1-n)\Big(1+(-\alpha\psi)^n\Big)+\frac{1}{2}(n+1)(-\alpha\psi)^n \\
+4(2-n)(-\alpha\psi)^{-1}+2(5n-1)(-\alpha\psi)^{n-1}-(n+1)(-\alpha\psi)^{2n-1} \\
+4(2n-3)(-\alpha\psi)^{n-2}-(9n-1)(-\alpha\psi)^{2n-2}+\frac{1}{2}(n+1)(-\alpha\psi)^{3n-2}\Bigg)\Phi^2(\psi)\,.
\end{aligned}
$$

Da

$$
4(2-n)(-\alpha\psi)^{-1}+4(2n-3)(-\alpha\psi)^{n-2}>4(n-1)(-\alpha\psi)^{n-2}>4(n-1)\,,
$$

$$
\frac{1}{2}(n+1)(-\alpha\psi)^n-\frac{1}{2}(n+1)(-\alpha\psi)^{2n-1}>0
$$

und

$$
2(5n-1)(-\alpha\psi)^{n-1}-(9n-1)(-\alpha\psi)^{2n-2}>(n-1)(-\alpha\psi)^{2n-2}>0
$$

gilt, folgt

$$
\begin{aligned}
f(\psi)&>\left(3(n-1)+(1-n)(-\alpha\psi)^n-\frac{1}{2}(n+1)(-\alpha\psi)^{2n-1}+\frac{1}{2}(n+1)(-\alpha\psi)^{3n-2}\right)\Phi^2(\psi) \\
&>\left(2(n-1)-\frac{1}{2}(n+1)(-\alpha\psi)^{2n-1}+\frac{1}{2}(n+1)(-\alpha\psi)^{3n-2}\right)\Phi^2(\psi)\,.
\end{aligned}
$$

Unter Verwendung von $x := -\alpha\psi \in (0,\infty)$ und

$$g(x) := 2(n-1) - \frac{1}{2}(n+1)x^{2n-1} + \frac{1}{2}(n+1)x^{3n-2}$$

lässt sich zeigen, dass wegen

$$\begin{aligned}\frac{dg}{dx}(x) = -\frac{1}{2}(n+1)\Big((2n-1)x^{2n-2} - (3n-2)x^{3n-3}\Big) = 0 \quad &\Leftrightarrow \\ \Big((2n-1) - (3n-2)x^{n-1}\Big)x^{2n-2} = 0 \quad &\Leftrightarrow \\ x_{\text{ext}} = \left(\frac{2n-1}{3n-2}\right)^{\frac{1}{n-1}} &\end{aligned}$$

und

$$\frac{dg}{dx}(x) < 0 \;\; \forall\, x \in (0, x_{\text{ext}})\,, \quad \frac{dg}{dx}(x) > 0 \;\; \forall\, x \in (x_{\text{ext}}, \infty)$$

sowie

$$\begin{aligned} g(x_{\text{ext}}) &= 2(n-1) - \frac{1}{2}(n+1)\left(1 - \frac{2n-1}{3n-2}\right)\left(\frac{2n-1}{3n-2}\right)^{\frac{2n-1}{n-1}} \\ &= 2(n-1) - \frac{1}{2}(n+1)\,\underbrace{\frac{n-1}{3n-2}}_{>1}\left(\underbrace{\frac{2n-1}{3n-2}}_{<1}\right)^{\frac{2n-1}{n-1}} \\ &> \frac{2(2n-2)(3n-2) - (n+1)(n-1)}{2(3n-2)} \\ &= \frac{11n^2-20n+9}{2(3n-2)} = \frac{(11n-9)(n-1)}{2(3n-2)} > 0 \end{aligned}$$

schießlich auch $f(\psi) > 0$ und damit die Behauptung für $0 < -\alpha\psi \leq 1$ gilt. Bleibt $-\alpha\psi > 1$ zu untersuchen. Analog zu den vorangegangenen Überlegungen gilt

$$4(2-n)(-\alpha\psi)^{-1}\Phi(\psi) + 4(2n-3)(-\alpha\psi)^{n-2}\Phi^2(\psi) > 4(n-1)(-\alpha\psi)^{-1}\Phi^2(\psi)\,,$$

und damit

$$\begin{aligned} f(\psi) > 1 - n + \frac{1}{2}(n+1)(-\alpha\psi)^n + \Big(2(5n-1)(-\alpha\psi)^{n-1} - (n+1)(-\alpha\psi)^{2n-1}\Big)\Phi(\psi) \\ + \left(4(n-1)(-\alpha\psi)^{-1} - (9n-1)(-\alpha\psi)^{2n-2} + \frac{1}{2}(n+1)(-\alpha\psi)^{3n-2}\right)\Phi^2(\psi)\,. \end{aligned}$$

Mit der Einschränkung

$$\begin{aligned} 2(5n-1)(-\alpha\psi)^{n-1} - (n+1)(-\alpha\psi)^{2n-1} \geq 0 \quad &\Leftrightarrow \\ 2(5n-1) - (n+1)(-\alpha\psi)^n \geq 0 \quad \Leftrightarrow \quad -\alpha\psi \leq \left(\frac{2(5n-1)}{n+1}\right)^{\frac{1}{n}} & \end{aligned}$$

und unter Anwendung von

$$\Big(1+(-\alpha\psi)^n\Big)^m > (-\alpha\psi)^{nm} = (-\alpha\psi)^{n-1}$$

gilt

$$f(\psi) > 1-n+\frac{1}{2}(n+1)(-\alpha\psi)^n+\Bigg(4(n-1)(-\alpha\psi)^{-1} + (n-1)(-\alpha\psi)^{2n-2}-\frac{1}{2}(n+1)(-\alpha\psi)^{3n-2}\Bigg)\Phi^2(\psi)\,,$$

so dass wegen

$$1-n+\frac{1}{2}(n+1)\underbrace{(-\alpha\psi)}_{>1}^{n} > 0$$

und

$$\Big(1+(-\alpha\psi)^n\Big)^{2m} > (-\alpha\psi)^{2nm} = (-\alpha\psi)^{2n-2}$$

schließlich

$$f(\psi) > 4(n-1)(-\alpha\psi)^{-1}\Phi^2(\psi) > 0$$

und damit die Behauptung folgt.

□

Bemerkung A.8
Satz A.7 lässt sich auf $\psi\in\mathbb{R}^-$ erweitern, doch (vermutlich) nicht algebraisch beweisen.

A.2.3 (Eingeschränkte) Eindeutigkeit

Satz A.9
Seien zwei van Genuchten-Mualem-Leitfähigkeitsfunktionen der Form

$$K_i(\psi) = K_{sat}\frac{\Big(1-(-\alpha\psi)^{n_i-1}\big(1+(-\alpha\psi)^{n_i}\big)^{\frac{1-n_i}{n_i}}\Big)^2}{\big(1+(-\alpha\psi)^{n_i}\big)^{\frac{n_i-1}{2n_i}}}\,,\quad i=1,2\,,$$

mit $n_1>n_2>1$ und $K_{sat},\alpha\in\mathbb{R}^+$ gegeben. Dann gilt für alle $\psi\in[-\frac{1}{\alpha},0)$

$$K_1(\psi) > K_2(\psi)\,.$$

Beweis:
Durch einfaches Umstellen gilt

$$K_i(\psi) = \left(K_{\text{sat}} \frac{\left(1 - (-\alpha\psi)^{n_i-1}\left(1+(-\alpha\psi)^{n_i}\right)^{\frac{1-n_i}{n_i}}\right)^{\frac{4n_i}{n_i-1}}}{1+(-\alpha\psi)^{n_i}} \right)^{\frac{n_i-1}{2n_i}}, \quad i=1,2\,,$$

und wegen

$$(-\alpha\psi)^{n_i-1}\left(1+(-\alpha\psi)^{n_i}\right)^{\frac{1-n_i}{n_i}} = \left(\frac{1}{(-\alpha\psi)^{n_i}}\left(1+(-\alpha\psi)^{n_i}\right)\right)^{\frac{1-n_i}{n_i}} = \frac{1}{\left(\frac{1}{(-\alpha\psi)^{n_i}}+1\right)^{\frac{n_i-1}{n_i}}} \tag{A.19}$$

auch

$$K_i(\psi) = \left(K_{\text{sat}} \frac{\left(1 - \frac{1}{\left(\frac{1}{(-\alpha\psi)^{n_i}}+1\right)^{\frac{n_i-1}{n_i}}}\right)^{\frac{4n_i}{n_i-1}}}{1+(-\alpha\psi)^{n_i}} \right)^{\frac{n_i-1}{2n_i}}, \quad i=1,2\,.$$

Ohne Einschränkung der Allgemeinheit kann $n_1 > n_2$ angenommen werden. Da hierfür

$$1+(-\alpha\psi)^{n_1} < 1+(-\alpha\psi)^{n_2} \qquad \text{sowie} \qquad \frac{1}{(-\alpha\psi)^{n_1}}+1 > \frac{1}{(-\alpha\psi)^{n_2}}+1 > 1$$

und wegen

$$\frac{n_1-1}{n_1} > \frac{n_2-1}{n_2}$$

auch

$$1-\frac{1}{\left(\frac{1}{(-\alpha\psi)^{n_1}}+1\right)^{\frac{n_1-1}{n_1}}} > 1-\frac{1}{\left(\frac{1}{(-\alpha\psi)^{n_2}}+1\right)^{\frac{n_2-1}{n_2}}}$$

gilt, folgt aufgrund

$$\frac{4n_1}{n_1-1} > \frac{4n_2}{n_2-1} \qquad \text{und} \qquad \frac{n_1-1}{2n_1} > \frac{n_2-1}{2n_2}$$

bereits die Behauptung.

□

Satz A.10
Seien zwei van Genuchten-Mualem-Leitfähigkeitsfunktionen der Form

$$K_i(\psi) = K_{sat}\frac{\Big(1-(-\alpha\psi)^{n_i-1}\big(1+(-\alpha\psi)^{n_i}\big)^{\frac{1-n_i}{n_i}}\Big)^2}{\big(1+(-\alpha\psi)^{n_i}\big)^{\frac{n_i-1}{2n_i}}}\,,\quad i=1,2\,, \tag{A.20}$$

mit $n_1>n_2>1$ *und* $K_{sat},\alpha\in\mathbb{R}^+$ *gegeben. Dann existiert ein* $\psi_s\in(-\frac{2}{\alpha},-\frac{1}{\alpha})$ *mit*

$$K_1(\psi_s) = K_2(\psi_s)\,.$$

Beweis:
Aufgrund der Stetigkeit von (A.20) kann die Aussage mit dem Zwischenwertsatz bewiesen werden. Aus Satz A.9 ist hierfür bereits

$$K_1\Big(-\frac{1}{\alpha}\Big) - K_2\Big(-\frac{1}{\alpha}\Big) > 0$$

bekannt. Bleibt die Ungleichung für $\psi=-\frac{2}{\alpha}$ zu untersuchen. Unter Verwendung von (A.19) gilt

$$K_i\Big(-\frac{2}{\alpha}\Big) = K_{\text{sat}}\frac{\left(1-\Big(\frac{1}{\frac{1}{2^{n_i}}+1}\Big)^{\frac{n_i-1}{n_i}}\right)^2}{\big(1+2^{n_i}\big)^{\frac{n_i-1}{2n_i}}}\,,\quad i=1,2\,,$$

und wegen

$$\big(1+2^{n_1}\big)^{\frac{n_1-1}{2n_1}} > \big(1+2^{n_2}\big)^{\frac{n_2-1}{2n_2}}$$

sowie

$$\frac{1}{\frac{1}{2^{n_1}}+1} > \frac{1}{\frac{1}{2^{n_2}}+1} \qquad \text{und} \qquad \frac{n_1-1}{n_1} > \frac{n_2-1}{n_2}$$

auch

$$0 < 1-\Big(\frac{1}{\frac{1}{2^{n_1}}+1}\Big)^{\frac{n_1-1}{n_1}} < 1-\Big(\frac{1}{\frac{1}{2^{n_2}}+1}\Big)^{\frac{n_2-1}{n_2}}\,.$$

Hieraus folgt schließlich

$$K_1\Big(-\frac{2}{\alpha}\Big) - K_2\Big(-\frac{2}{\alpha}\Big) < 0$$

und damit die Behauptung.

□

Bemerkung A.11
Es wird vermutet, dass zwei durch (A.20) mit $n_1\neq n_2$ *definierte hydraulische Leitfähigkeitsfunktionen auf ganz* $\mathbb{R}^-$ *nur den bereits nachgewiesenen Schnittpunkt besitzen. Leider lässt sich dieser (vermutlich) i.A. weder algebraisch bestimmen noch die Existenz weiterer Schnittpunkte widerlegen.*

Anhang B

Implementierungen

In diesem Abschnitt werden die (wichtigsten) durchgeführten Implementierungsarbeiten in **Richy1D**, einem modularen, plattformunabhängigen Simulationstool, welches vom Lehrstuhl für Angewandte Mathematik I der Friedrich-Alexander Universität Erlangen-Nürnberg zur Verfügung gestellt wurde, kurz aufgezeigt. Mit Hilfe dieses umfangreichen Programms lässt sich eine Vielzahl hydrologischer Fragestellungen betrachten und numerisch berechnen (vgl. Abbildung B.1). Die genaue Funktionalität und eine detaillierte Darstellung aller Modellierungsansätze kann im online abrufbaren Manual nachgeschlagen werden. Über implementierungsspezifische Datenstrukturen (Speicherung, Menuführung, etc.) gibt u.A. Frank [25] Auskunft.

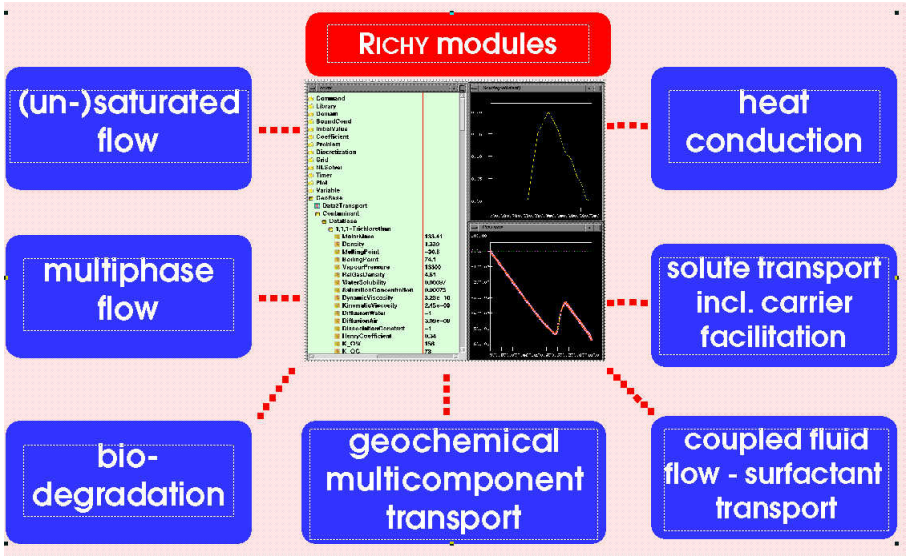

Abbildung B.1: Richy1D - Eingebundene Simulationsmodule

B.1 Regularisierung der vG-Leitfähigkeit

Um bei ungeeignet gewählten Parameterwerten die auftretenden numerischen Schwierigkeiten der nach van Genuchten-Mualem parametrisierten Leitfähigkeit (2.5) zu umgehen, wurden in Richy die beiden im Kapitel 2.1.5 vorgestellten Regularisierungsansätze implementiert. Damit stehen dem User nun optional sowohl die sättigungs- als auch druckabhängige Regularisierung (2.12) und (2.18) zur Verfügung. Der jeweilige Regularisierungsgrad kann dabei beliebig, d.h. $R_\Phi \in [0, 1]$ bzw. $R_\psi \leq 0$, gewählt werden. In Abbildung (B.2) ist ein entsprechender Bildschirmausschnitt angegeben, auf dem, neben anderen Parameterwerten, die notwendigen Einstellungsmöglichkeiten für eine Regularisierung ersichtlich sind.

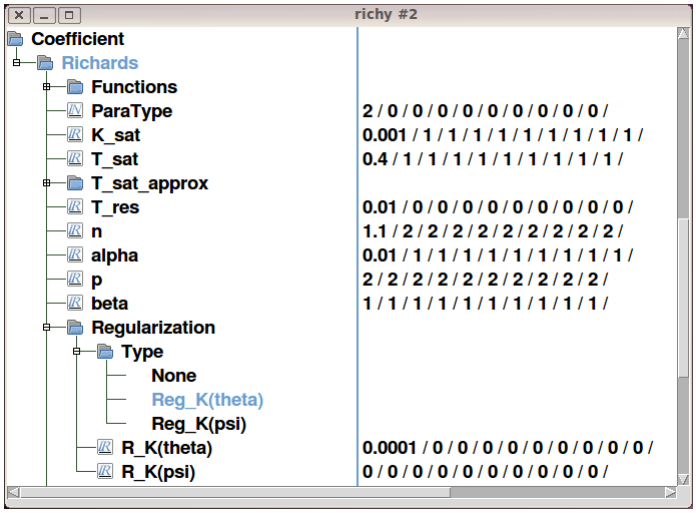

Abbildung B.2: Richy-Screenshot: Regularisierung vG-Leitfähigkeitsfunktion

B.2 Gedämpfte Liniensuche

Insbesondere bei der Liniensuche des in Richy eingebundenen SQP-Optimierungsverfahrens *NLPQL* kam es vereinzelt vor, dass ein ausgewählter Parametersatz zu numerischen Schwierigkeiten bis hin zum Abbruch der Berechnung führte. Zwar sollte bei einer sinnvollen Wahl der Zulässigkeitsbereiche stets eine Lösung existieren, doch gerade bei einer fixen Parametrisierung mit nicht vollständig unabhängigen Koeffizienten lassen sich ungeeignet kombinierte Parameterwerte nur schwer vermeiden. Grund für einen Abbruch war stets ein Scheitern/Nichtkonvergieren des verwendeten Newton-Verfahrens während der zugehörigen Simulation des ent-

standenen Differentialgleichungssystems. Um dennoch ein Fortführen des Identifizierungsprozesses zu gewährleisten, wurde eine Art gedämpfte Liniensuche eingebunden. Hierbei wird, ausschließlich im Fall einer gescheiterten Berechnung, die vorab ermittelte Schrittweite um einen Faktor $\zeta \in (0,1)$ verkürzt und entsprechend ein neuer Knotenpunkt herangezogen. Unter Berücksichtigung der Armijo-Goldsteinbedingung findet der Algorithmus auf diese Weise (u.U. nach mehrfachem Aufruf) stets wieder zu einem sinnvollen Datensatz zurück. Es sollte jedoch beachtet werden, dass ζ nicht zu klein gewählt wird, da sonst die erlangten Parameterwerte zu weit von dem zuvor als optimal angenommenen Parametersatz abweichen. Ein zu groß gewählter Faktor ist ebenfalls nicht ideal, da sonst die gedämpfte Liniensuche mehrfach eingreifen muss, bevor ein brauchbarer Datensatz ermittelt wird. Beide Argumente berücksichtigend, wurde der Dämpfungsfaktor schließlich mit $\zeta = 0.5$ festgelegt.

B.3 Rekursive Parameteridentifizierung

Die in den Kapiteln 4.2.2 und 4.2.3 vorgestellten Identifizierungsresultate basieren auf einer rekursiven Parameteridentifizierung. Hierzu wird das zu minimierende (diskrete) Fehlerfunktional (4.6) mit den auf der pseudoinversen Sensitivitätsmatrix basierenden Faktoren (4.5) gewichtet. Um schließlich den in Abschnitt 4.2.1 vorgestellten rekursiven Identifizierungsalgorithmus in Richy einzubinden, mussten neben der eigentlichen Berechnung der Gewichtungsfaktoren $\lambda(p)$ entsprechende Schleifen in den Optimierungsprozess implementiert werden. Hierbei wurde explizit sichergestellt, dass auch weiterhin alle bisher angebotenen Funktionalitäten, wie beispielsweise die Verwendung unterschiedlicher Minimierer (u.A. wird SQP, BundleTrust und GGPRV angeboten), eine direkte oder FD-approximierte Gradientenberechnung oder einfach nur das Speichern der Identifizierungsresultate, benutzbar bleiben. Dem User steht damit optional und völlig uneingeschränkt die rekursive Identifizierung zur Verfügung.

In Abbildung B.3 ist ein Bildschirmausschnitt angegeben, auf dem alle, speziell für eine rekursive Identifizierung, relevanten Optionen ersichtlich sind. So kann neben den in (4.5) definierten Gewichtungen auch eine normierte Variante

$$(\lambda_{k,i})_j(p) := \frac{1}{\|S^+\|_F} \sqrt{\sum_{l=1}^{r} \left(\left(S^+_{k,i}\right)_{lj}(p) \right)^2}, \quad j = 1, \ldots, m_{k,i},$$

mit Frobenius Norm $\|.\|_F$, gewählt werden. Insbesondere bei sehr hohen Sensiti-

vitätswerten, wie sie u.A. bei einer formfreien Parametrisierung auftreten können, nimmt damit das zu minimierende Fehlerfunktional nicht unnötig große Werte an. Um einen korrekten Vergleich mit den Resultaten eines ungewichteten Identifizierungsproblems zu ermöglichen, wird neben dem zugehörigen gewichteten Fehlerfunktional auch das ungewichtete Residuum ausgegeben. Schließlich lassen sich die pseudoinverse Sensitivitätsmatrix und die berechneten Gewichtungsfaktoren optional abspeichern.

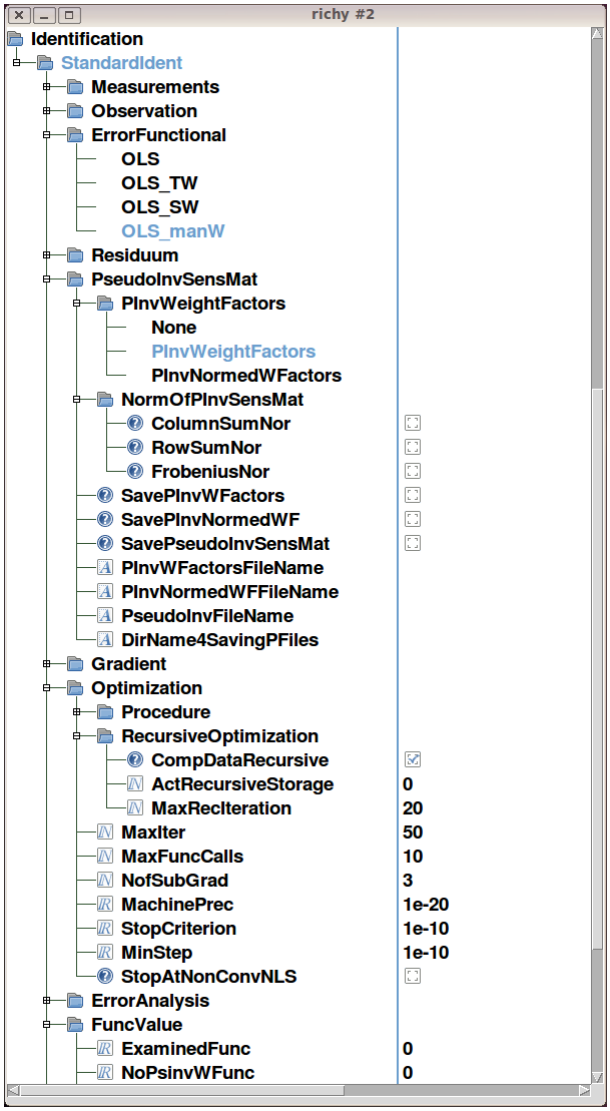

Abbildung B.3: Richy-Screenshot: Rekursive Identifizierung

B.4 Mehrdimensionaler formfreier Ansatz

Den Schwerpunkt der durchgeführten Implementierungsarbeiten stellt zweifelsohne das Einbinden der formfreien Identifizierung dar. Hierzu entstanden in den unterschiedlichsten Systembereichen mehrere tausend Zeilen Code. Neben der Erweiterung der Funktionsklasse für biochemische Reaktionsraten mussten vor allem das Richy-Kernel, das Identifizierungsmodul sowie das Optimierungsinterface umfangreich erweitert werden. Hierbei wurde rigoros darauf geachtet, dass sich die Implementierungen nicht nur auf die in dieser Arbeit zu untersuchende Reaktionsrate anwenden lassen. Vielmehr besteht die Möglichkeit auch in anderen Anwendungsbereichen beliebige dreidimensionale Nichtlinearitäten zu erzeugen und mittels der nun zur Verfügung stehenden Tools formfrei zu identifizieren. Ebenfalls wurde darauf geachtet, dass die bereits vorhandenen Funktionalitäten (z.B. den Optimierungsalgorithmus oder die Gradientenberechnung betreffend) auch weiterhin (auch unter Verwendung der neuen Parametrisierung) genutzt werden können. Dem Benutzer steht damit ein flexibles Werkzeug mit zahlreichen Optionen, welche nachfolgend noch aufgezeigt werden, zur Verfügung.

Abbildung B.4 zeigt für die zu identifizierende formfreie Abbaurate die zugehörige Bildschirmausgabe. Da Richy eine Unterteilung der zugrundegelegten Domain (eine eindimensional betrachtete Laborsäule) in bis zu zehn Subdomains erlaubt, kann auch die Diskretisierung der formfreien Parametrisierung für jedes einzelne Teilgebiet separat angegeben werden. Da jedoch bereits eine homogene Bodenprobe ausreichend Schwierigkeiten für die vorgestellten Identifizierungsprobleme mit sich bringt, wird im Folgenden von einer separaten Unterteilung in mehrere unterschiedliche Subdomains abgesehen. Als Basisfunktionen stehen sowohl lokale als auch hierarchische Basen zur Verfügung. Bisher sind diese auf trilineare Funktionen beschränkt, doch ohne größeren Aufwand können auch höherdimensionale Splines eingebunden werden. Schließlich stehen dem User hilfreiche Einlese- und Speicherfunktionen zur Verfügung, so dass im Fall einer feinen Diskretisierung die zugehörigen Funktionswerte nicht per Hand eingetragen bzw. ausgelesen werden müssen. Insbesondere für die graphische Ausgabe der ermittelten Reaktionsraten, wie sie beispielsweise in Abbildung 4.10 dargestellt sind, können damit alle relevanten Daten elegant an Matlab übergeben werden.

Im Fall einer formfreien Identifizierung der biochemischen Reaktionsrate kann entschieden werden, ob neben der Abbaurate des Schadstoffes auch die Feldfaktoren Y und $\alpha_{A/D}$ (zur Bestimmung der Reaktionsraten R_A und R_B) zu bestimmen sind, oder ob, beispielsweise durch vorab durchgeführte experimentelle Messun-

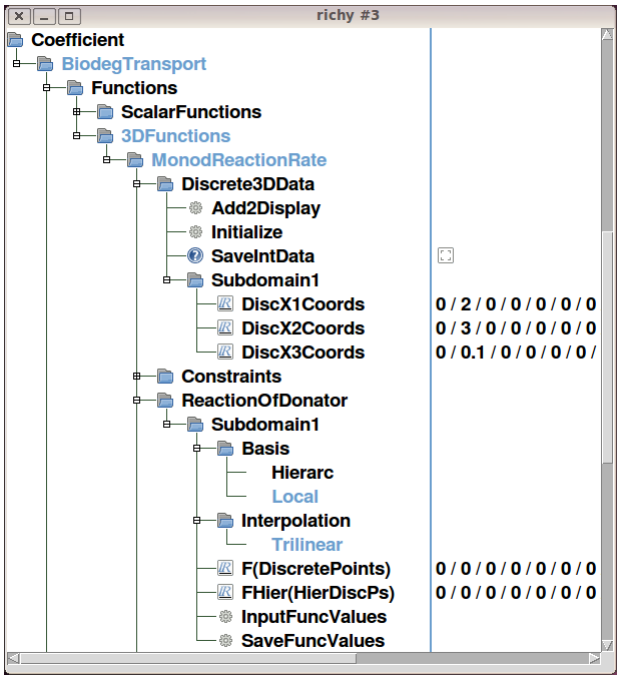

Abbildung B.4: Richy-Screenshot: Formfreie 3D-Abbaurate

gen, diese bereits durch konkrete Werte vorliegen. Selbst die zu bestimmende formfreie Abbaurate R_D lässt sich deaktivieren, so dass, je nach aktiv gewählten Parametern, ein ein- bzw. zweidimensionales Identifizierungsproblem zum "Nachbessern" der erzielten Feldfaktoren generiert werden kann. Des Weiteren wird sowohl für die Funktionswerte der zu identifizierenden Nichtlinearität als auch für die diskreten Parameterwerte (hier Y und $\alpha_{A/D}$) jeweils separat ein geeigneter Zulässigkeitsbereich gewählt. Schließlich kann noch die Monotonie der gesuchten Reaktionsrate vorgegeben werden. Dies schränkt zwar den Identifizierungsprozess aufgrund zahlreicher Ungleichungsnebenbedingungen kravierend ein, jedoch ermöglicht es oftmals erst das Berechnen eines brauchbares Identifizierungsresultats.

Nachfolgend finden sich alle in Richy implementierten Optionen für die dreidimensionale, formfreie Identifizierung. Beginnend mit unterschiedlichen, frei wählbaren hierarchischen Diskretisierungs- und Verfeinerungsstrategien, sowie optional eingeschränkten Axialknoten und einer linearen dritten Komponente hin zu einer gewichteten Interpolation zur Unterstützung des konvexen Reaktionsverhaltens.

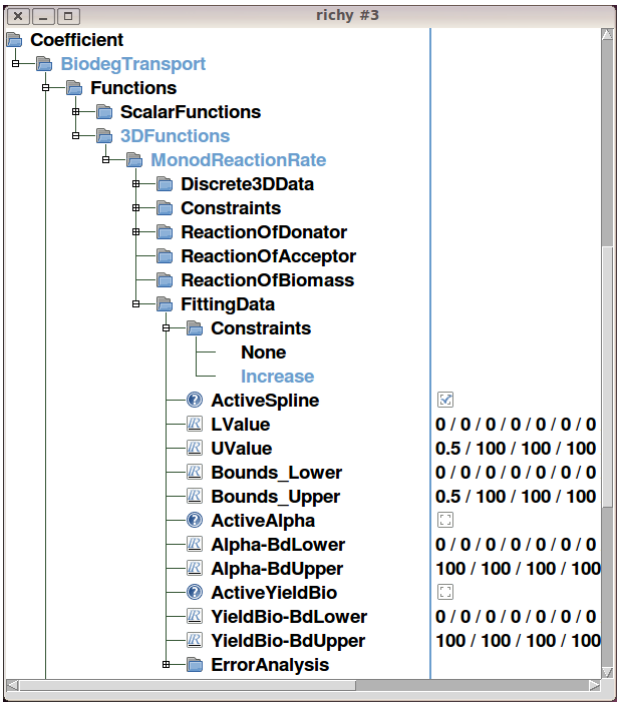

Abbildung B.5: Richy-Screenshot: Eingeschränkte FF-Identifizierung

B.4.1 Hierarchische Diskretisierungsstrategien

Für die Diskretisierung stehen unterschiedliche Strategien zur Verfügung. Während die Abfolge der hierarchisch einzubindenden Stützstellen vorab festgelegt werden muss, können zusätzliche Optionen gewählt werden, welche die Dimensionalität des Identifizierungsproblems deutlich senken.

B.4.1.1 Verfeinerungsabfolge

Der Benutzer muss, abhängig vom Typ der verwendeten Basisfunktionen, eine grundlegende Verfeinerungsstrategie wählen. Diese gibt konkret an, welche Knoten bzw. Splines in den einzelnen hierarchischen Diskretisierungsschritten dem Identifizierungsproblem zugefügt werden sollen. Im Fall lokaler Basen besteht die Möglichkeit neben der sogenannten *LevelByLevel*-Diskretisierung, bei der abwechselnd eine axiale Richtung um eine Stützstelle erweitert wird, die *AxisByAxis*-Diskretisierung zu wählen, so dass alternativ zunächst nur eine Richtung bis zu einer maximalen Knotenzahl erweitert wird. Beginnend mit einer möglichst groben Startdiskretisierung bilden sich damit die zugehörigen Freiheitsgrade wie

folgt

$$\begin{aligned}\vec{n} = (2,2,2) \to (3,2,2) \to &(3,3,2) \to (3,3,3) \to (4,3,3) \to (4,4,3) \\ &\to (4,4,4) \to (5,4,4) \ldots \to (n^1_{\max}, n^2_{\max}, n^3_{\max})\end{aligned}$$

bzw.

$$\begin{aligned}\vec{n} =&(2,2,2) \to (3,2,2) \to (4,2,2) \to \ldots \to (n^1_{\max},2,2) \\ &\to (n^1_{\max},3,2) \to\to (n^1_{\max},4,2) \to \ldots \to (n^1_{\max},n^2_{\max},2) \\ &\to (n^1_{\max},n^2_{\max},3) \to\to (n^1_{\max},n^2_{\max},4) \to \ldots \to (n^1_{\max},n^2_{\max},n^3_{\max})\,.\end{aligned}$$

Im Fall hierarchischer Basen muss vorab entschieden werden, ob die einzubindenden Splines auf vollen oder dünnen Gittern (vgl. Kapitel 3.2.2.2) definiert sein sollen. Zusätzlich kann zwischen einer vollen Schrittweite und einer (klar zu favorisierenden) blockweisen Erweiterung, wie sie durch einzelne $\vec{\sigma} \in \Upsilon_d^{\text{Strat}}(s)$ gegeben ist, gewählt werden. In Abbildung B.6 findet sich die entsprechende Richy-Bildschirmausgabe.

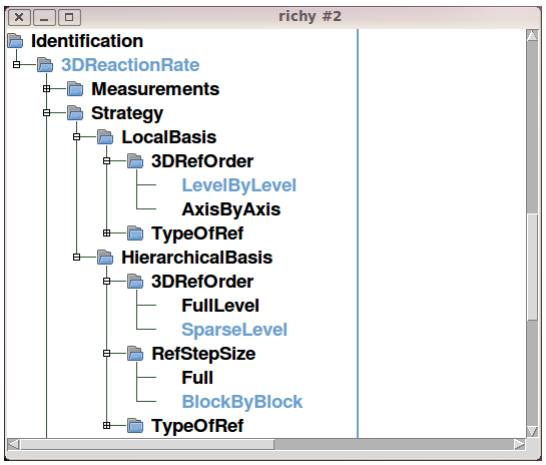

Abbildung B.6: Richy-Screenshot: Formfreie Verfeinerungsstrategien

B.4.1.2 Diskretisierungsstrategien für lokale Basen

Neben der Verfeinerungsabfolge ist vor allem die Art der Knotenunterteilung für das Identifizierungsresultat entscheidend. Im Fall lokaler Basen kann zwischen einer vollständig äquidistanten, einer links-dyadischen und zweier unterschiedlich

links-gewichteten Diskretisierungen gewählt werden. Während bei der äquidistanten Unterteilung kanonischerweise das Intervall $[a^i, b^i]$ der i-ten Koordinatenrichtung durch

$$x_k^i = \frac{1}{n^i}\Big((n^i-k)a^i+kb^i\Big)\,,\quad k=0,\ldots,n^i\,, \tag{B.1}$$

in n^i-1 gleichgroße Teilintervalle aufgeteilt wird, können im Fall der linksdyadischen Diskretisierung die Stützstellen mit Hilfe von

$$s=\begin{cases} 0 & ,\ \text{falls } n^i=2\,,\\ \lceil \log_2(n^i-1)\rceil & ,\ \text{sonst}\,,\end{cases}$$

durch

$$x_k^i = \begin{cases} \frac{1}{2^s}\Big((2^s-k)a^i+kb^i\Big) & ,\ \text{falls } k\le k':=2(n^i-2^{s-1}-1)\,,\\ \frac{1}{2^{s-1}}\Big(\big(2^{s-1}-\big(k-\frac{k'}{2}\big)\big)a^i+\big(k-\frac{k'}{2}\big)b^i\Big) & ,\ \text{sonst}\,,\end{cases}$$

$k=0,\ldots,n^i$, angegeben werden. Vergleiche hierzu auch Abbildung B.7. Entsprechend wird auch hier eine äqudistante Unterteilung angestebt, jedoch teilweise mit linkseitiger Gewichtung.

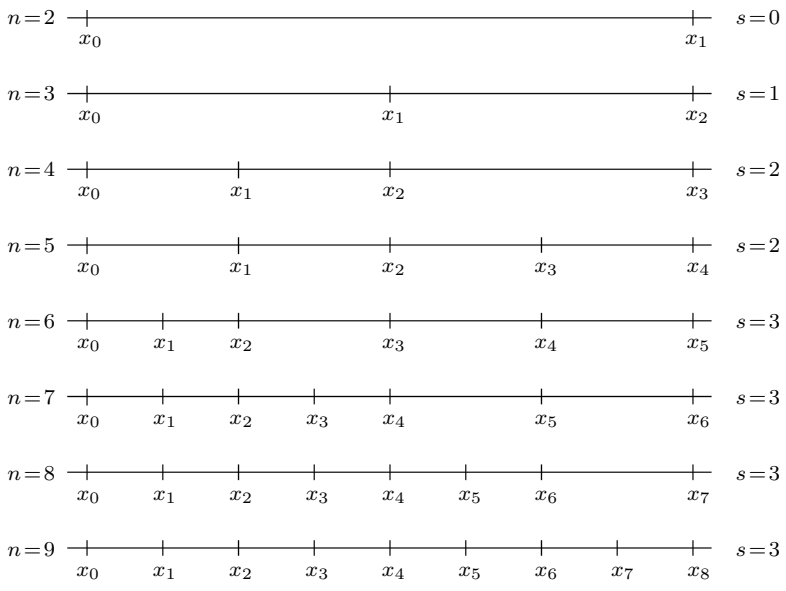

Abbildung B.7: Linksdyadische Unterteilung

Soll vollends auf eine äquidistante Unterteilung verzichtet werden, kann eine der beiden nachfolgenden linksgewichteten Diskretisierungsstrategien gewählt werden.

B.4.1.2.1 *WL I-l*

Um die Diskretisierungsvorschrift *WL I* der Tiefe $l \in \mathbb{N}$ explizit angeben zu können, wird vorab

$$s = \begin{cases} 0 & , \text{ falls } n^i = 2\,, \\ 1 & , \text{ falls } 2 < n^i \leq 2+l\,, \\ s'+1 & , \text{ falls } 2^{s'-1}(2+l) < n^i \leq 2^{s'}(2+l)\,,\ s' \in \mathbb{N},\ s' \geq 1\,, \end{cases}$$

definiert. Liegt nur eine geringe Knotenzahl n^i mit $s \leq 1$ vor, so können die Stützstellen direkt durch $x_0^i = a^i$,

$$x_k^i = \frac{1}{2^{n^i-1-k}}\left((2^{n^i-1-k}-1)a^i + b^i\right)\,,\quad k = 1, \ldots, n^i-2\,,$$

und $x_{n^i-1}^i = b^i$ beschrieben werden. Wird jedoch eine feinere Diskretisierung benötigt, so sind die einzelnen Werte nur noch rekursiv bestimmbar. In diesem Fall werden die Stützstellen zunächst nur für den ersten Level $\tilde{s} = 1$ formuliert. Es gilt $x_{1,0} = a^i$,

$$x_{1,k}^i = \frac{1}{2^{l+1-k}}\left((2^{l+1-k}-1)a^i + b^i\right)\,,\quad k = 1, \ldots, l\,,$$

und $x_{1,l+1}^i = b^i$. Für die folgenden Level $\tilde{s} = 2, \ldots,$ s–1, kann damit jeweils sukzessiv

$$x_{\tilde{s},k}^i = \begin{cases} x_{\tilde{s}-1,\frac{k}{2}}^i & , \text{ falls } k \text{ gerade}\,, \\ \frac{1}{2}\left(x_{\tilde{s}-1,\frac{k-1}{2}}^i + x_{\tilde{s}-1,\frac{k+1}{2}}^i\right) & , \text{ sonst}\,, \end{cases}$$

berechnet werden. Um schließlich die endgültigen Stützstellen angeben zu können, wird noch

$$n_{\tilde{s}}^i = \begin{cases} 2+l & , \text{ falls } s = 2\,, \\ 2+l+(2^{s-2}-1)(l+1) & , \text{ sonst}\,, \end{cases}$$

benötigt. Damit gilt für $k \leq k' := n_{\tilde{s}}^i + 2^{s-2}(l+1) - n^i$ direkt $x_k^i = x_{s-1,k}^i$ und sonst

$$x_k^i = \begin{cases} x_{s-1,\frac{k+k'}{2}}^i & , \text{ falls } k+k' \text{ gerade}\,, \\ \frac{1}{2}\left(x_{s-1,\frac{k+k'-1}{2}}^i + x_{s-1,\frac{k+k'+1}{2}}^i\right) & , \text{ sonst}\,. \end{cases}$$

Um einen graphischen Überblick zu erlangen, sind in Abbildung B.8 die ersten Schritte der Diskretisierungsstrategie *WL I-3* schematisch dargestellt.

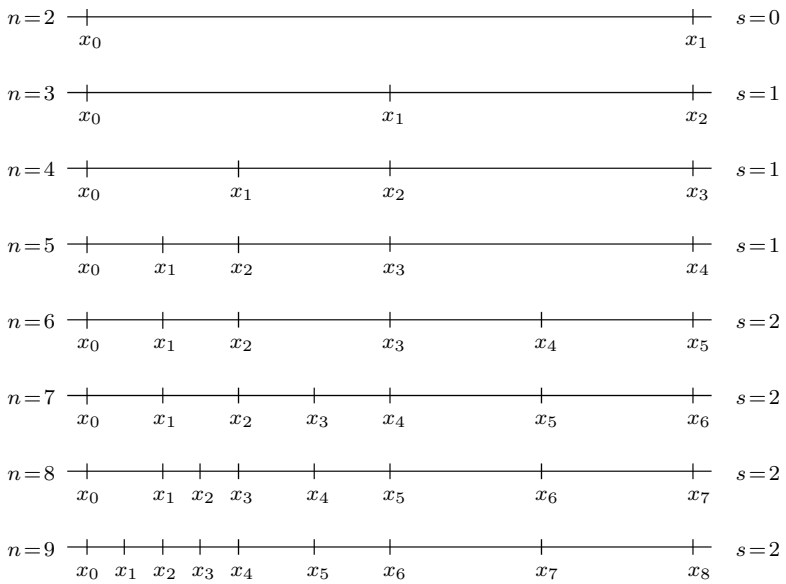

Abbildung B.8: Linksgewichtete Diskretisierung *WL I-3*

B.4.1.2.2 *WL II-l*

Im Fall der linksgewichteten Diskretisierungsstrategie *WL II-l* können ebenfalls die Stützstellen für eine geringe Knotenzahl $n^i \leq 2+l$ direkt angegeben werden. Es gilt $x_0^i = a^i$,

$$x_k^i = \frac{1}{l+1}\Big((l+1-k)a^i + kb^i\Big)\,, \quad k=1,\ldots,n^i-2,$$

und $x_{n^i-1}^i = b^i$. Für $n^i > 2+l$ werden entsprechend für $s \in \mathbb{N}$, $s \geq 2$, mit

$$2+2^{s-2}l < n^i \leq 2+2^{s-1}l$$

die Knotenpunkte rekursiv ermittelt. Hierzu wird für $\tilde{s}=1$ das Intervall äquidistant unterteilt. Es gilt

$$x_{1,k}^i = \frac{1}{l+1}\Big((l+1-k)a^i + kb^i\Big)\,, \quad k=0,\ldots,l+1\,.$$

Für die folgenden Level $\tilde{s}=2,\ldots,s-1$ kann rekursiv

$$x_{\tilde{s},k}^i = \begin{cases} x_{\tilde{s}-1,\frac{k}{2}}^i & , \text{ falls } k \leq 2l \text{ und gerade}\,, \\ \frac{1}{2}\big(x_{\tilde{s}-1,\frac{k-1}{2}}^i + x_{\tilde{s}-1,\frac{k+1}{2}}^i\big) & , \text{ falls } k < 2l \text{ und ungerade}\,, \\ x_{\tilde{s}-1,k-l}^i & , \text{ sonst}\,, \end{cases}$$

berechnet werden. Schließlich wird noch $l' = n^i - 2 - (s-1)l$ definiert, um die entgültigen Stützstellen durch

$$x_k^i = \begin{cases} x^i_{s-1,\frac{k}{2}} & , \text{ falls } k \leq 2l' \text{ und gerade}, \\ \frac{1}{2}\left(x^i_{s-1,\frac{k-1}{2}} + x^i_{s-1,\frac{k+1}{2}}\right) & , \text{ falls } k < 2l' \text{ und ungerade}, \\ x^i_{s-1,k-l'} & , \text{ sonst}, \end{cases}$$

angeben zu können. Abschließend findet sich in Abbildung B.9 eine schematische Darstellung der Diskretisierungsstrategie *WL II-2*.

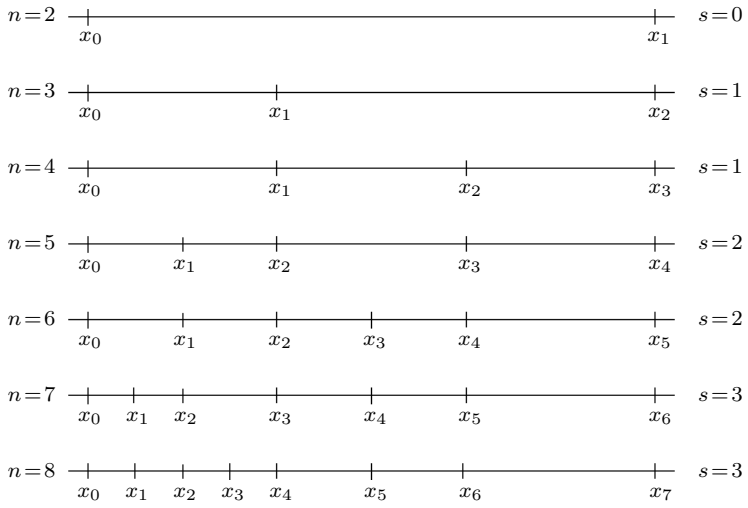

Abbildung B.9: Linksgewichtete Diskretisierung *WL II-2*

B.4.1.3 Diskretisierungsstrategien für hierarchische Basen

Trotz der einschränkenden Struktur hierarchischer Basen stehen auch hier zwei unterschiedliche Diskretisierungsstrategien zur Verfügung. Neben der äquidistanten Unterteilung, wie sie bereits für lokale Basen in (B.1) definiert ist, kann optional eine linksgewichtete Anordnung der Stützstellen mittels der Funktion *W-l/f* gewählt werden. Hierbei wird das Intervall $[a^i, b^i]$, unter Berücksichtigung von $l \in \{1, 2\}$ und $d \in (0, 1)$, skalenweise unterteilt. Für $s = 1$ kann, unabhängig von l, direkt $x^i_{1,0} = a^i$, $x^i_{1,1} = (1-f)a^i + fb^i$ und $x^i_{1,2} = b^i$ angegeben werden. Für $s = 2$ und $l = 2$ ist eine weitere Gewichtung mit

$$x_{2,k} = \begin{cases} x^i_{1,\frac{k}{2}} & , \text{ falls } k \text{ gerade}, \\ (1-f)x^i_{1,\frac{k-1}{2}} + fx^i_{1,\frac{k+1}{2}} & , \text{ sonst}, \end{cases}$$

zu beachten. Für $s=2$ und $l=1$ sowie für $s\geq 3$ (unabhängig von l) lassen sich schließlich die Stützstellen rekursiv durch

$$x_{s,k} = \begin{cases} x^i_{s-1,\frac{k}{2}} & , \text{ falls } k \text{ gerade}, \\ \frac{1}{2}\left(x^i_{s-1,\frac{k-1}{2}} + x^i_{s-1,\frac{k+1}{2}}\right) & , \text{ sonst}, \end{cases}$$

beschreiben. Abschließend werden die zu verwendenden Knoten x^i_k, $k=1,\ldots,n^i$, direkt durch die ermittelten Werte der höchsten Skala definiert.

Zur Verdeutlichung des vorgestellten Diskretisierungsansatzes werden in Abbildung B.10 die beiden Strategien *W-1/0.33* und *W-2/0.33* schematisch dargestellt. In Abbildung B.11 findet sich schließlich ein Richy-Screenshot, welcher die zugehörigen Einstellmöglichkeiten für beide Basistypen aufzeigt.

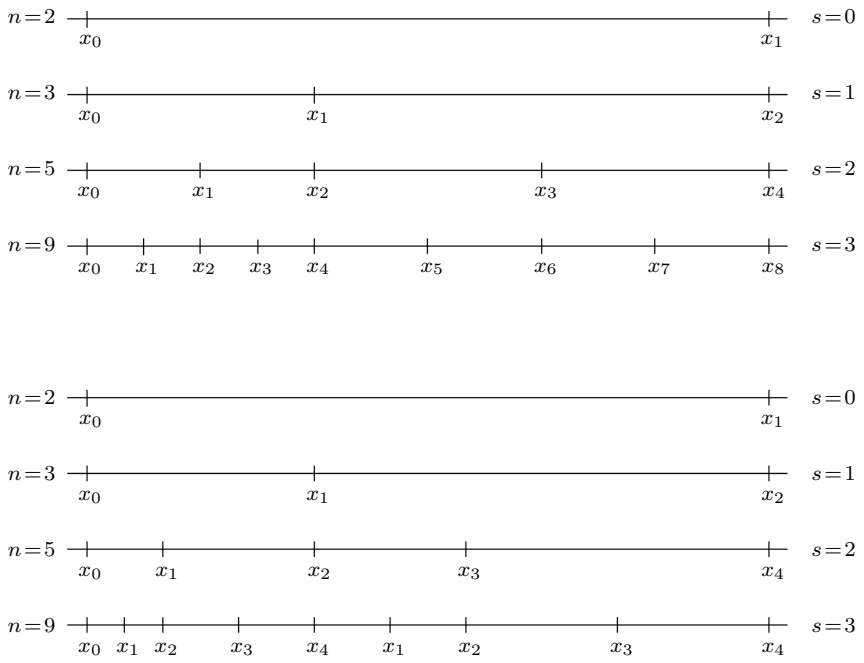

Abbildung B.10: Linksgewichtete Diskretisierung *W-1/0.33* und *W-2/0.33*

B.4.2 Reduzierte Knoten

Um die Dimensionalität nicht unnötig hoch zu halten, können für die hier zu identifizierende Abbaurate, unabhängig vom verwendeten Basistyp, alle Splines, mit Stützstelle auf einer der Koordinatenebenen, mit Null gewichtet und entsprechend aus dem Identifzierungsprozess genommen werden. Ebenso kann das hier vorliegende bzw. angenommene lineare Verhalten der dritten Komponente (doppelte

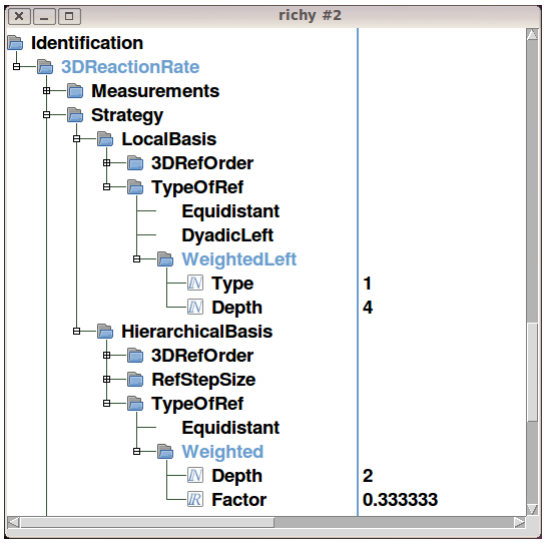

Abbildung B.11: Richy-Screenshot: Diskretisierungsstrategien

Biomassenkonzentration erzeugt auch einen doppelt so hohen Schadstoffabbau) berücksichtigt werden, so dass in x^3-Richtung stets die gröbste Diskretisierung, bestehend aus lediglich zwei Stützstellen, für eine adäquate Approximation ausreicht. Ein Vergleich der in den Tabellen 3.2.b und 4.19 angegebenen Freiheitsgrade gibt den gewonnenen Dimensionsvorteil konkret an. In Abbildung B.12 lassen sich schließlich die diesbezüglichen Einstellmöglichkeiten in der Richy-Oberfläche erkennen.

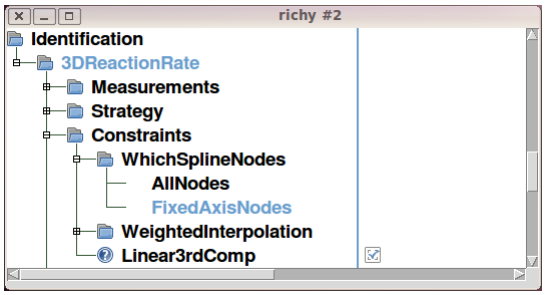

Abbildung B.12: Richy-Screenshot: Fixe Axialknoten und lineare 3. Komponente

B.5 Interpolation im $\mathbb{R}^3$

Grundlegend für die Auswertung einer formfreien Parametrisierung p ist das Interpolieren von Funktionswerten ausgewählter Stützstellen. Im Folgenden wird daher die, in Richy implementierte, trilineare Interpolation vorgestellt.

B.5.1 Trilineare Interpolation

Betrachtet wird ein achsenparalleler Quader

$$Q := \left[a^1, b^1\right] \times \left[a^2, b^2\right] \times \left[a^3, b^3\right] \subset \mathbb{R}^3$$

mit einer nach (3.30) vorgegebenen Knotenmenge. Da zu einem beliebig gewählten $\vec{x} \in Q$ (mindestens) ein Teilquader

$$Q_{i_1 i_2 i_3} := \left[k^1_{i_1}, k^1_{i_1+1}\right] \times \left[k^2_{i_2}, k^2_{i_2+1}\right] \times \left[k^3_{i_3}, k^3_{i_3+1}\right]$$

mit $\vec{x} \in Q_{i_1 i_2 i_3}$ existiert und sich im Fall $\vec{x} \in \partial Q_{i_1 i_2 i_3} \cap \partial Q_{i_4 i_5 i_6}$ die Interpolation auf die entsprechende (Schnitt-)Fläche bzw. Kante beschränkt, genügt es zur Auswertung der vorgegebenen Parametrisierung eine (noch zu definierende) Funktion $f : Q_{i_1 i_2 i_3} \to \mathbb{R} \in \mathcal{L}(Q_{i_1 i_2 i_3})$ heranzuziehen. Ist hierbei $Q_{i_1 i_2 i_3}$ durch lexikographisch angeordnete Knoten $\vec{x}_{000}, \ldots, \vec{x}_{111}$ (der Einfachheit halber sind die Indizes unabhängig von i_1, i_2, i_3 gewählt) festgelegt und gilt für die dort vorliegenden Funktionswerte

$$f_{j_1 j_2 j_3} := f(\vec{x}_{j_1 j_2 j_3}) = p(\vec{x}_{j_1 j_2 j_3}), \;\; j_l \in \{0, 1\}, \;\; j, l \in \{0, 1\},$$

so kann der gesuchte Funktionswert durch $f(\vec{x})$ und damit mittels einer trilinearen Interpolation bestimmt werden. Wird hierzu die Achsenparallelität von $Q_{i_1 i_2 i_3}$ ausgenutzt und

$$\Delta x^1 := x^1_{100} - x^1_{000}, \quad \Delta x^2 := x^2_{010} - x^2_{000}, \quad \Delta x^3 := x^3_{001} - x^3_{000},$$

sowie

$$\lambda^l(x^l) := \frac{x^l - x^l_{000}}{\Delta x^l}, \quad l = 1, \ldots, 3, \tag{B.2}$$

definiert, so lässt sich diese durch die drei nachfolgenden Schritte berechnen.

1. Interpolation in x^3-Richtung:

$$f_{00}(x^3) := \left(1 - \lambda^3(x^3)\right) f_{000} + \lambda^3(x^3) f_{001},$$

$$f_{10}(x^3) := \left(1 - \lambda^3(x^3)\right) f_{100} + \lambda^3(x^3) f_{101},$$

$$f_{01}(x^3) := \left(1 - \lambda^3(x^3)\right) f_{010} + \lambda^3(x^3) f_{011},$$

$$f_{11}(x^3) := \left(1 - \lambda^3(x^3)\right) f_{110} + \lambda^3(x^3) f_{111}.$$

2. Interpolation in x^2-Richtung:

$$f_0(x^2, x^3) := \big(1-\lambda^2(x^2)\big) f_{00}(x^3) + \lambda^2(x^2) f_{01}(x^3)\,,$$
$$f_1(x^2, x^3) := \big(1-\lambda^2(x^2)\big) f_{10}(x^3) + \lambda^2(x^2) f_{11}(x^3)\,.$$

3. Interpolation in x^1-Richtung:

$$f(\vec{x}) = \big(1-\lambda^1(x^1)\big) f_0(x^2, x^3) + \lambda^1(x^1) f_1(x^2, x^3)\,.$$

Vergleiche hierzu auch Abbildung B.13, in der die beschriebenen Teilschritte graphisch motiviert werden. Es ist sofort erkennbar, dass die im ersten Schritt berechneten Werte $f_{00}, \ldots, f_{11}$ Interpolationen an den Knoten $\vec{x}_{00}, \ldots, \vec{x}_{11}$ darstellen. Da sich diese jedoch allesamt auf einer Ebene befinden, verbleibt eine bilineare Interpolationsaufgabe. Da der zweite Schritt wiederum das Problem um eine Dimension reduziert, sind im letzten Schritt nur noch die Werte f_0 und f_1 auf $[\vec{x}_0, \vec{x}_1]$ zu interpolieren.

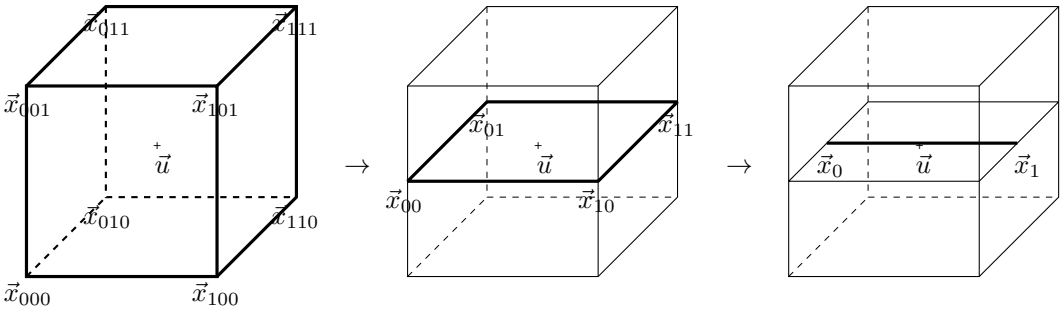

Abbildung B.13: Trilineare Interpolation in $Q_{i_1 i_2 i_3}$

Zusammenfassend kann die trilineare Interpolationsaufgabe auch direkt durch

$$f(\vec{x}) = \sum_{i_1, i_2, i_3 = 0}^{1} \lambda_{i_1 i_2 i_3}(\vec{x}) f_{i_1 i_2 i_3}$$

mit

$$\lambda_{i_1 i_2 i_3}(\vec{x}) := \prod_{l=1}^{3} \lambda_{l, i_l}(x^l)\,, \quad \lambda_{l, i_l}(x^l) := \begin{cases} 1-\lambda^l(x^l)\,, & \text{falls } i_l = 0\,, \\ \lambda^l(x^l) & \text{, sonst}\,, \end{cases}$$

$l = 1, \ldots, 3$, berechnet werden.

Abschließend sei bemerkt, dass selbst eine Funktionsauswertung außerhalb des definierten Zulässigkeitsbereichs, also für ein $\vec{x} \notin Q$, bei der durchgeführten Implementierung berücksichtigt wurde. In diesem Fall wird

$$\vec{u} := \arg\min_{\vec{y} \in Q} \left\| \vec{x} - \vec{u} \right\|$$

(auf triviale Weise) gelöst und $\vec{u}$ alternativ für die trilineare Interpolation zur Verfügung gestellt. Anschaulich entspricht diese Vorgehensweise einer konstanten Erweiterung der ansonsten lediglich auf Q definierten Parametrisierung p. Insbesondere bei am Rand von Q sehr flach verlaufenden Nichtlinearitäten kann damit der Zulässigkeitsbereich etwas reduziert und entsprechend die Diskretisierung in den relevanten Bereichen feiner gestaltet werden.

B.5.2 Gewichtete Interpolation

Für die hierarchisch zu entwickelnde Diskretisierung müssen zu Beginn eines jeden Verfeinerungsschrittes neue Stützstellen eingebunden werden. Je nach Strategie werden diese nach und nach entweder äqudistant, linksdyadisch oder linksgewichtet angeordnet und mit einem Startwert belegt. Für gewöhnlich werden diese initialen Startschätzungen entsprechend der jeweiligen Position zwischen den direkten Nachbarn mittels einer trilinearen Interpolation errechnet. Um ein (vermutetes) Krümmungsverhalten als ergänzende Funktionseigenschaft zu nutzen, ohne den Identifizierungsprozess explizit auf entsprechende Funktionsklassen einzuschränken, kann alternativ eine gewichtete Interpolation genutzt werden. Hierbei bleibt die Position der neu hinzugekommenen Knoten unverändert. Lediglich zur Berechnung der initialen Startwerte wird eine andere, angepasste Stelle angenommen. Konkret bedeutet dies für einen neuen Knoten $\vec{x} =: (x^1, \ldots, x^3)$ mit seinen (bis zu) acht Nachbarknoten $\vec{x}_{000}, \ldots, \vec{x}_{111}$, dass, statt den in (B.2) definierten Interpolationsgewichtungen $\lambda^l(x^l)$, $l = 1, \ldots, 3$, nun

$$\lambda^l_{\text{mod}} := \begin{cases} 2\lambda_{\text{gwInt}}\lambda^l & , \text{ falls } \lambda^l_{\text{gwInt}} \leq \frac{1}{2} , \\ 2(1-\lambda_{\text{gwInt}})\lambda^l + 2(\lambda_{\text{gwInt}} - \frac{1}{2}) & , \text{ sonst} , \end{cases}$$

Verwendung findet. Folglich wird entsprechend für $0 < \lambda_{\text{gwInt}} < \frac{1}{2}$ ein konkaves und für $\frac{1}{2} < \lambda_{\text{gwInt}} < 1$ ein konvexes Kurvenverhalten iniziiert. Für $\lambda_{\text{gwInt}} = \frac{1}{2}$ bleibt die gewichtete Interpolation ohne Wirkung, so dass weiterhin eine gewöhnliche trilineare Interpolation durchgeführt wird. In Abbildung B.14 findet sich die entsprechende Richy-Bildschirmausgabe.

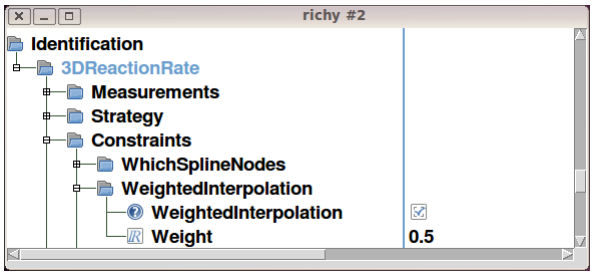

Abbildung B.14: Richy-Screenshot: Gewichtete Interpolation

Anhang C

Zusammenfassung

Die vorliegende Arbeit beschäftigt sich mit der numerischen Identifizierung gesuchter Koeffizientenfunktionen bodenphysikalischer Modellierungsansätze. Konkret wurde sowohl die hydraulische Leitfähigkeit als auch eine dreikomponentige biochemische Abbaurate mit zwei unterschiedlichen Identifizierungsverfahren untersucht.

Das erste und einleitende Kapitel dient der Vorstellung und Motivation der bearbeiteten Thematik. Hierbei wird das zugrundegelegte Problem zur Prognose von Schadstofftransport und Abbauprozessen anhand des BMBF-Förderschwerpunktes *Sickerwasserprognose* aufgezeigt. Zudem wird die Notwendigkeit der numerischen Simulation (un-)gesättigter Fließverhalten durch poröse Medien und reaktiver Transportprozesse verdeutlicht. Schließlich wird kurz auf die Problematik ausgewählter Parametrisierungen eingegangen und die verwendeten Identifizierungsprozesse angedeutet.

Das zweite Kapitel ist ausführlich der Modellierung bodenspezifischer Problemstellungen gewidmet. Hier wird neben der grundlegenden Richards-Gleichung auch die allgemeine Transportgleichung mit nichtlinearem Diffusions/Dispersions-Tensor und variablem Fluss aufgezeigt. Zur Beschreibung der hydraulischen Funktionen wird die van Genuchten-Mualem-Parametrisierung vorgestellt und auf die damit verbundenen numerischen Schwierigkeiten, welche bei ungeeignet gewählten Parameterwerten entstehen können, eingegangen. Motiviert durch diese Problematik, werden zudem zwei mögliche Regularisierungsansätze definiert. Diese basieren auf einer sättigungs- und druckabhängigen $\mathcal{P}^2$- bzw. $\mathcal{P}^3$-Approximation der hydraulischen Leitfähigkeitsfunktion nahe des gesättigten Bereichs. Ein aussagekräftiger Vergleich beider Ansätze wird schließlich mit Hilfe der Kirchhoff-Transformation vollzogen. Des Weiteren wird in diesem Kapitel neben der Model-

lierung der Gleichgewichts- und kinetischen Sorption insbesondere der biochemische Abbau mittels des dreikomponentigen dualen Monod-Modells vorgestellt und dessen Monotonieeigenschaften untersucht. Zusammenfassend wird abschließend das zugehörige Differentialgleichungssystem einschließlich möglicher Randbedingungen dargestellt.

Im dritten Kapitel wird zunächst, in Analogie zu Bitterlich [8], das inverse Problem allgemein diskutiert. Hierzu werden entsprechende Lösungs- und Beobachtungsoperatoren definiert und ein (un)gewichtetes Fehlerfunktional eingeführt. Neben grundlegenden Differentiationsfragestellungen wird der, auf die Problemstellung angepasste, Satz über implizite Funktionen aufgezeigt und schließlich das adjungierte Problem vorgestellt. Der zweite Teil dieses Kapitels beschäftigt sich konkret mit der formfreien Parametrisierung und der hierauf basierenden Identifizierung. Hierbei werden zunächst skalare Nichtlinearitäten einer Unbekannten untersucht und mit Hilfe eines hierarchischen Ansatzes approximiert. Für die notwendigen Basisfunktionen stehen zwei unterschiedliche Klassen zur Verfügung. Zum Einen sind lokale Basen definiert, deren Träger stets ein gleiches Größenniveau aufweisen, so dass einfache Strukturen, wie beispielsweise eine äquidistante Unterteilung, sehr einfach realisiert werden können. Zum Anderen werden aber auch sogenannte hierarchische Basen vorgestellt, die mit Hilfe einer vorab definierten Skalierungsfunktion abhängig vom aktuellen Level bestimmt werden. Damit wird gewährleistet, dass die Struktur der einzelnen Basisfunktionen gleich bleibt, ihr Träger jedoch mit steigender Skala immer kleiner wird. Entsprechend setzt sich der diskrete, skalenabhängige Parameterraum aus völlig unterschiedlichen Basisfunktionen zusammen. In einem weiteren Abschnitt dieses Kapitels werden die vorangegangenen Überlegungen auf (vektorwertige) Nichtlinearitäten mehrerer Veränderlicher übertragen. Während sich bei den lokalen Basen, bis auf die Notation und die deutlich höhere Komplexität, bei der mehrdimensionalen Vorgehensweise nicht viel ändert, stehen bei den hierarchischen Basen nun mehrere Verfeinerungsstrategien zur Verfügung. Neben einer verallgemeinerten Notation sind hierbei insbesondere die in Bungartz [15] vorgestellten, auf sogenannten vollen und dünnen Gittern definierten, Basisfunktionen aufgezeigt. Schließlich wird für einen vorab festgelegten, achsenparallelen Definitionsbereich, unter Berücksichtigung nicht verschwindender Randwerte, die Dimensionalität (Anzahl der Freiheitsgrade im Fall einer Identifizierung) sowie der zugehörige Diskretisierungsfehler der einzelnen Basistypen miteinander vergleichen. Abschließend wird der bereits in Igler [34] vorgestellte hierachische Identifizierungsalgorithmus, angepasst auf die hier verwendete Notation, angegeben.

Das vierte und letzte Kapitel ist der Untersuchung numerischer Fallstudien gewidmet. Beginnend mit den im Kapitel 2 vorgestellten Regularisierungsansätzen der van Genuchten-Mualem-Parametrisierung werden diese in einem ersten Abschnitt numerisch untersucht. Hierbei wird mittels simuliertem Unterdruck eine realitätsnahe Dehydrierung einer virtuellen Laborsäule berechnet und die erzielten Resultate unter Verwendung unterschiedlich starker Regularisierungsgrade miteinander verglichen. Schließlich wird festgestellt, dass bereits die sättigungsabhängige $\mathcal{P}^2$-Regularisierung mit klein gewähltem Regularisierungsgrad $R \approx 10^{-4}$ einen signifikanten Performanzvorteil besitzt, ohne zu sehr von der ursprünglichen van Genuchten-Mualem-Leitfähigkeit abzuweichen. Im zweiten Abschnitt dieses Kapitels wird zunächst, zur Verbesserung der auf fixen Parametrisierungen basierenden Identifizierung, ein rekursiver Identifizierungsalgorithmus vorgestellt. Hierzu wird, abhängig von der Sensitivität jedes einzelnen diskreten Messpunktes von den jeweiligen Parameterwerten, eine rekursive Gewichtung des Residuums definiert. Mit Hilfe dieser Modifikation soll es dem Minimierungsverfahren gelingen aus bereits erreichten lokalen Minima zu entfliehen und nach und nach zu immer besseren Extremalstellen zu gelangen. Ein mathematischer Beweis zur Bestätigung dieses Verhaltens konnte nicht formuliert werden. Dennoch belegen die in diesem Abschnitt vorgestellten Berechnungsbeispiele eindrucksvoll die vermutete Entwicklung. Neben einem virtuellen Säulenexperiment zur Beschreibung eines, auf der Monod-Parametrisierung definierten, Schadstoffabbauprozesses wurden auch die hydraulischen Funktionen einer (realen) Sandprobe mittels Identifizierung der zugehörigen van Genuchten-Mualem-Parameter bestimmt. In beiden Fällen konnte durch den rekursiven Identifizierungsalgorithmus eine signifikante Verbesserung der erzielten Parameterwerte erzielt werden. Der letzte Abschnitt dieses Kapitels beschäftigt sich ausgiebig mit der formfreien Parameteridentifizierung des bereits vorgestellten dreikomponentigen Schadstoffabbauprozesses. Hierbei werden sowohl lokale als auch auf vollen und dünnen Gittern definierte hierarchische Basen verwendet und die erzielten Resultate miteinander verglichen. Um die Komplexität und die Dimensionalität der zugehörigen Identifizierungsprobleme möglichst gering zu halten, ohne auf die notwendige Flexibilität der approximierenden Nichtlinearität zu verzichten, sind optional zahlreiche Einschränkungen, wie beispielsweise fixierte Axialknoten, ein monotones Steigungsverhalten oder einfach nur eine weitere Untergliederung der einzelnen Skalen verwendet worden. Ein Vergleich mit der vorangegangenen fixen Parameteridentifzierung zeigt deutlich, dass insbesondere unter Vorgabe gestörter bzw. nicht auf der zugrundegelegten Parametrisierung ermittelter Daten, wie sie vor allem durch reale Messungen gegeben sind, die formfreie Identifizierung klare Vorteile aufweist. Dennoch wurde auch festgestellt, dass sehr wohl klare Vorgaben an die

gesuchte Nichtlinearität bekannt sein müssen, da sonst eine adäquate Identifizierung der gesuchten Reaktionsrate fraglich ist.

Im Anhang findet sich schließlich neben der Überprüfung einfacher Monotonie- und Krümmungseigenschaften der van Genuchten-Mualem-Parametrisierung alle relevanten Implementierungsarbeiten kurz aufgezeigt. Diese umfassen neben den bereits angesprochenen Regularisierungen, der rekursiven Parameteridentifizierung und den dreidimensionalen formfreien Parametrisierungs- und Identifizierungsansätzen, zahlreiche optionale Identifizierungshilfen, welche die Identifizierbarkeit entscheidend beeinflussen können. Zusammenfassend steht dem Benutzer damit ein vollständig skriptgesteuertes Tool zur Verfügung, welches alle hier vorgestellten und untersuchten Problemstellungen ermöglicht.

Literaturverzeichnis

[1] M. Alexander, K.M. Scow. *Kinetics of biodegradation in soil*, in B.L. Sawhney, K. Brown. *Reactions and Movement of Organic Chemicals in Soils*. Soil Science Society of America, Madison, 1989.

[2] H.W. Alt. *Lineare Funktionalanalysis*. Springer Verlag, Berlin, 2006.

[3] L.R. Ahuja, D. Swartzendruber. *An improved form of soil-water diffusivity function*. Soil Science Society of America, Medison, Proc. 36:9-14, 1972.

[4] H.W. Alt, S. Luckhaus. *Quasilinear Elliptic-Parabolic Differential Equations*. Math. Z., 184 (1983), pp. 311-341.

[5] J. Bear. *Dynamics of fluids in Porous Media*. Dover Publications, New York, 1972.

[6] R. Bellmann. *Adaptive Control Processes: A Guided Tour*. Princton University Press, 1961.

[7] H. Bisswanger. *Theorie und Methoden der Enzymkinetik*. Verlag Chemie, Weinheim, 1979.

[8] S. Bitterlich. *Identifizierung der hydraulischen Funktionen poröser Medien unter Verwendung formfreier Ansätze*. Dissertation, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany, 2003.

[9] S. Bitterlich, W. Durner, S.C. Iden, P. Knabner. *Inverse Estimation of the Unsaturated Soil Hydraulic Properties from Column Outflow Experiments Using Form-Free Parametrizations*, Vadose Zone Journal, pp. 1-22, 2004.

[10] S. Bitterlich, P. Knabner. *Adaptive and Formfree Identification of Nonlinearities in Fluid Flow from Column Experiments*. American Mathematical Society, authors Z. Chen, R.E. Ewing, *Contemporary Mathematics*, Volume 295, 2002.

[11] S. Bitterlich, P. Knabner. *Numerical Methods for the Determination of Material Properties in Soil Science*, First International Conference *Inverse Problems: Modeling and Simulation*, July 14-21, Fethiye-Turkey, 2002.

[12] R.H. Brooks, A.T. Corey. *Hydraulic properties of porous media.* Hydrology Paper no. 3, Civil Engineering Dep., Colorado State University, Fort Collins, Colorado, 1964.

[13] Bundes-Bodenschutzgesetz (BBodSchG): *Gesetz zum Schutz vor schädlichen Bodenveränderungen und zur Sanierung von Altlasten.* Bundesgesetzblatt 1998, Teil I, Nr. 16, 502-510.

[14] Bundes-Bodenschutz- und Altlastenverordnung (BBodSchV). Bundesgesetzblatt 1999, Teil 1, Nr. 36, 1554-1582.

[15] H.-J. Bungartz. *Dünne Gitter und deren Anwendung bei der adaptiven Lösung der dreidimensionalen Poisson-Gleichung.* Dissertation, Institut für Informatik, Technische Universität München, 1992.

[16] H.-J. Bungartz, M. Griebel. *Sparse grids.* Acta Numerica (2004), Seite 147-269, Cambridge University Press, 2004.

[17] G. Crescimanno, M. Iovino. *Parameter estimation by inverse method based on one-step and multi-step outflow experiments.* Geoderma, 68:257-277, 1995.

[18] J.H. Dane, S. Hruska. *In-situ determination of soil hydraulic properties during drainage*, Soil Science of America, Journal, 47:619-624, 1983.

[19] W. Durner, B. Schultze, T. Zurmühl. *State-of-the-Art in Inverse Modeling of Inflow/Outflow Experiments*, in M.Th. van Genuchten, F.J. Leij, L. Wu (editors), *Proceeding International Workshop on Characterization and Measurement of the Hydraulic Properties of Unsaturated Porous Media*, October 22-24, 1997, University of California, Riverside, CA, 1999.

[20] S.O. Eching, J.W. Hopmans. *Optimization of hydraulic functions from transient outflow and soil water pressure date.* Soil Science Scociety of America, Journal, 57:1167-1175, 1993.

[21] S.O. Eching, J.W. Hopmans, O. Wendroth. *Unsaturated hydraulic conductivity from transient multistep outflow and soil water pressure data.* Soil Science Society of America, Journal, 58:687-695, 1994.

[22] F.J. Endelmann, G.E.P. Box, J.R. Boyle, R.R. Hughes, D.R. Keeney, M.L. Northrup, P.G. Saffigna. *The mathematical modeling of soil-water-nitrogen phenomena*. EDFB-IBP-74-8, Oak Ridge National Laboratory, Oak Ridge, 1974.

[23] L.C. Evans. *Partial Differential Equations*, Graduate Studies in Mathematics, Volume 19, American Mathematical Society, Providence, Rhode Island, 1998.

[24] C.W. Fetter. *Contaminant Hydrogeology*. Prentice Hall, Upper Saddle River, second edition, 1999.

[25] F. Frank. *Hydrochemical Multi-Component transport - Mineral Dissolution and Precipitation with Consideration of Porosity-Changes in Variably-Saturated Porous Media*, Diploma Thesis, University of Erlangen-Nürnberg, 2008.

[26] R.A. Freeze, J.A. Cherry. *Groundwater*. Prentice-Hall, New York, 1979.

[27] M. Geisel. *Ein inverses Problem für die degeneriert parabolische Richardsgleichung*. Dissertation, Johannes Gutenberg-Universität Mainz, 2003.

[28] M.M. Gribb. *Parameter estimation for determining hydraulic properties of a fine sand from transient flow measurements*, Water Resources Research, 32(7):1965-1974, 1996.

[29] W. Hackbusch. *Multigrid Methods and Aplications*, Springer Verlag, Berlin, Heidelberg, 1985.

[30] J. Hadamard. *Lectures on the Cauchy Problem in Linear Partial Differential Equations*, Yale University Press, New Haven, 1923.

[31] G. Hämmerlin, K.H. Hoffmann. *Numerische Mathematik*. Springer Verlag, Berlin, 1991.

[32] R. Haverkamp, M. Vaucin, J. Touma, P.J. Wierenga, G. Vachaud. *A Comparison of Numerical Simulation Models For One-Dimensional Infiltration*. Soil Science Society of America, J. 41, 1977.

[33] U. Hornung. *Identification of non-linear physical parameters from input-output experiments*, in P. Deyuflhard, E. Haier (editors), *Workshop on Numerical Treatments of Inverse Problems in Differential and Integral Equations*, 227-237, Birkhäuser, Boston, 1983.

[34] B.A. Igler. *Identification of Nonlinear Coefficient Functions in Reactive Transport through Porous Media*. PhD thesis, Friedrich-Alexander Universität Erlangen-Nürnberg, 1998.

[35] B. Igler, P. Knabner. *Identification of nonlinear sorption isotherms by soil column breakthrough experiments*. Physics and Chemistry of the Earth, 23(2):215-219, 1998.

[36] V. Isakov. *Inverse Problems for Partial Differential Equations*. Applied Mathematical Sciences 127. Springer Verlag, New York, 1998.

[37] W. Kinzelbach, R. Rausch. *Grundwassermodellierung*. Borntraeger, Berlin, 1995.

[38] P. Knabner. *Mathematische Modelle für Transport und Sorption gelöster Stoffe in porösen Medien*. P. Lang, Frankurt, 1991.

[39] J.N. Kool, J.C. Parker, M.Th. van Genuchten. *Parameter estimation for unsaturated flow and transport models - A review*. Journal Hydrology, 91:255-293, 1987.

[40] A. Ladyzhenskaya, V.A. Solonnikov, N.N. Uralceva. *Linear and Quasi-linear Equations of Parabolic Typ*. Transl. Math. Monographs 23, AMS, RI, 1969.

[41] Y. Lins, T. Schanz, D.G. Fredlund. *Modified Pressure Plate Apparatus and Columns Testing Device for Measuring SWCC of Sand*, Geotechnical Testing Journal, Vol. 32, No. 5, 2009.

[42] A.K. Louis. *Inverse und schlecht gestellte Probleme*. Teubner Verlag, Stuttgart 1989.

[43] Y. Mualem. *A new model for predicting the hydraulic conductivity of unsaturated porous media*. Water Resources Research. 12:513-522, 1976.

[44] F. Otto. *L^1-contraction and uniqueness for unstationary saturated-unsaturated porous media flow*. Adv. Math. Sci. Appl., 7(2):537-553, 1997.

[45] J.C. Parker, J.B. Kool, M.T. van Genuchten. *Determining Soil Hydraulic Properties from One-step Outflow Experiments by Paramter Estimation II. Experimental Studies*. Soil Science Society of America, Proc. 56(4):1042-1050, 1985.

[46] A. Prechtel. *Modelling and Efficient Numerical Solution of Hydrogeochemical Multicomponent Transport Problems by Process-Preserving Decoupling Techniques*. Dissertation, Friedrich-Alexander University Erlangen-Nuremberg, Germany, 2005.

[47] A. Quarteroni. *Numerical Models for Differential Problems - Modeling, Simulation and Applications*. Springer Verlag, Berlin, 4. Edition, 2009.

[48] F. Radu, I.S. Pop, P. Knabner. *Order of convergence estimates for an euler implicit, mixed finite element discretization of Richards' equation*. Siam Journal of Numerical Analysis, Vol. 42, Issue 4, Seite 1452-1478(2004).

[49] M. Reeves, J.O.Duguid. *Water movement through saturated-undersaturated porous media: A finite-element galerkin model*. ORNL-4927, Oak Ride National Laboratory, Oak Ridge, Tennesses, 1975.

[50] L.A. Richards. *Capillary conduction of liquids through porous mediums*. PhD-Thesis, Cornell University, 1931.

[51] K. Roth. *Soil Physics - Lecture Notes*. Institute of Environmental Physics, University of Heidelberg, Germany, 2006.

[52] R. Rudek, H. Eberle. *Der Förderschwerpunkt "Sickerwasserprognose" des Bundesministeriums für Bildung und Forschung - Ein Überlick"* in *altlasten spektrum*, Nr. 6, 294-304, 2001.

[53] O. Scherzer. *An iterative multi level algorithm for solving nonlinear ill-posed problems*. Numerische Mathematik, 80:579-600, 1998.

[54] M. Schirmer. *Investigation of Multiscale Biodegradation Processes: A Modelling Approach*. PhD-Thesis, University of Waterloo, Canada, 1998.

[55] E. Schneid. *Hybrid-Gemischte Finite-Elemente-Diskretisierung der Richards-Gleichung*. PhD-Thesis, Friedrich-Alexander Universität Erlangen-Nürnberg, 2000.

[56] G. Segol. *A three-dimensional galerkin finite element model for the analysis of contaminant transport in variably saturated porous media*. User's guide, Dep. of Earth Sciences, University of Waterloo, Canada.

[57] M. Stieber, S. Kraßnitzer, A. Tiehm. *Bedeutung biologischer Selbstreinigung in der ungesättigten Bodenzone für die Sickerwasserprognose - Teil 1: Modell-Säulenexperimente zur Elimination von PAK* in *altlasten spektrum*, Nr. 3, 111-118, 2007.

[58] A.F. Toorman, P.J. Wierenga, R.G. Hills. *Parameter estimation of hyraulic properties from one-step outflow data*. Water Resources Research, 28:3021-3028, 1992.

[59] J.C. van Dam, J.N.M. Stricker, A. Verhoef. *An evaluation of the one-step outflow method*, in M.Th. van Genuchten, F.J. Leij, L.J. Lund (editors) Proceeding International Workshop, *Indirect Methods for Estimating the Hydraulic Properties of Unsaturated Soils*. 633-644, University of California, Riverside, CA, 1992.

[60] J.C. van Dam, J.N.M. Stricker, P. Droogers. *Inverse method to determine soil hydraulic functions from multistep outflow experiments*. Soil Science Society of America, Journal, 58:647-652, 1994.

[61] M.T. van Genuchten. *A closed form equation for predicting the hydraulic conductivity of unsaturated soils*. Soil Science Society of America, Madison, Vol. 44, 1980, Seite 892-898.

[62] C.F. van Loan. *Introduction to Scientific Computing: a Matrix-Vector Approach using Matlab*, 2nd edition, Prentice Hall, 2000.

[63] M. Vauklin, R. Haverkamp, G Vauchaud. *Résolution numérique d'une equation de diffusion non linéaire*. Presses Universitaires des Grenoble, 1979.

[64] J. Werner. *Numerische Mathematik 1*, 1. Auflage, Vieweg Verlag, Braunschweig/Wiesbaden, 1992.

[65] A. Wouk. *A course of applied functional analysis*. John Wiley and Sons, New York, 1979.

[66] D.W. Zachmann, P.C. DuChateau, A. Klute. *Simultaneous approximation of water capacity and soil hydraulic conductivity by parameter identification*, Soil Science, 134:157-163, 1982.

[67] T. Zurmühl. *Evaluation of different boundary conditions for independent determination of hydraulic parameters using outflow methods*, in P. DuChateau, J. Gottlieb (editors), *Parameter identification and Inverse Problems in Hydrology, Geology and Ecology*. Water Science and Technology Libary, 23:165-184, Kluwer, Dordrecht, 1996.

Curriculum Vitae

Persönliche Daten

Name:	Michael Blume
Email:	michael.blume@elitenetzwerk.de
Geburtsdatum, -ort:	23.09.1975, Forchheim

Ausbildungsdaten

Schule:

-07/1991	Volksschule Stammbach, Quali,
09/1991-07/1994	Berufsschule Naila, Quabi,
09/1994-07/1999	Berufsoberschule Bayreuth, '95 mittlere Reife, '97 Fachhochschulreife, '99 Hochschulreife

Beruf:

08/1991-08/1994	Firma OFRA in Stammbach Ausbildung zum staatlich geprüften Bauzeichner

Studium:

10/1999-09/2005	Universität Bayreuth 1. Studiengang: Wirtschaftsmathematik, Vordiplom 2. Studiengang: Mathematik, Diplom

Promotion:

11/2005-05/2011	Universität Erlangen Lehrstuhl für Angewandte Mathematik I, Department Mathematik

Hochschulanstellungen

11/2005-12/2005	Universität Erlangen BMBF-Projekt "Modellierung des reaktiven Transports von Schadstoffen in der (un-)gesättigten Bodenzone zur Prognose der natürlichen Selbstreinigung"
01/2006-04/2009	Universität Erlangen Promotion im internationalen Doktorandenkolleg "Identification, Optimization and Control with Applications in Modern Technologies" des Elitenetzwerkes Bayern
05/2009-03/2011	Universität Münster Betreuung universitärer und praxisorientierter Praktika mit (bio-)medizinischer Orientierung