Parameteridentifikation und Optimale Versuchsplanung bei instationären partiellen Differentialgleichungen

Zur Erlangung des akademischen Grades eines

DOKTORS DER NATURWISSENSCHAFTEN

von der Fakultät für Mathematik des

Karlsruher Instituts für Technologie (KIT)

genehmigte

DISSERTATION

von

Dipl.-Math. Andrea Nestler

aus

Neubrandenburg

Tag der mündlichen Prüfung: 28. November 2012

Referent: Prof. Dr. Vincent Heuveline
Korreferent: Prof. em. Dr. Götz Alefeld

Danksagung

Ich möchte mich an dieser Stelle bei allen herzlich bedanken, die mich bei der Erstellung der vorliegenden Arbeit unterstützt haben.

Mein Dank gilt insbesondere Herrn Professor Dr. Vincent Heuveline, der das Thema dieser Dissertation anregte und es mir ermöglichte, im Rahmen meiner Tätigkeit an seinem Lehrstuhl zu promovieren. Ich danke ihm dafür, dass er die wissenschaftliche Betreuung meiner Arbeit übernommen hat und diese durch kritische und inspirierende Fachdiskussionen bereicherte. Herrn Professor Dr. Götz Alefeld danke ich für die vielen wertvollen und hilfreichen Anregungen bei der Fertigstellung dieser Arbeit. Vielen Dank für die Unterstützung!

Ein besonderer Dank gilt Herrn Professor Dr. Jürgen Hubbuch und Frau Dipl.-Math. Anna Osberghaus vom Institut für Bio- und Lebensmitteltechnik (KIT) für die Bereitstellung eines mathematischen Modells zur Beschreibung eines präparativen Säulenchromatographieprozesses. Dieses Modell wurde verwendet um die in dieser Arbeit beschriebenen Verfahren der Parameteridentifikation und Optimalen Versuchsplanung numerisch umzusetzen.

Bei meinen Kollegen am Institut für Angewandte und Numerische Mathematik 4 bedanke ich mich für die angenehme und freundschaftliche Arbeitsatmosphäre. Sowohl bei inhaltlichen als auch methodischen Fragen standen diese mir immer mit Rat und Tat zur Seite und wussten mich in den richtigen Momenten zu motivieren. Insbesondere möchte ich mich an dieser Stelle bei Frau PD Gudrun Thäter bedanken!

Schließlich danke ich von ganzem Herzen meinem Ehemann Martin Nestler und meiner Mutter Roswitha Otzen sehr dafür, dass sie immer für mich da waren. Vielen Dank für die seelische und moralische Unterstützung und die unendliche Geduld, die ihr für mich aufgebracht habt. Euch beiden widme ich diese Arbeit!

Inhaltsverzeichnis

Notationsverzeichnis

Mengen, Räume und Parameter

Darstellung numerischer Lösungen

Darstellung von Vektoren, skalaren Größen und Iterierten

Differentialoperatoren

Kapitel 1

Einleitung und Motivation

1.1 Einführung

In vielen wissenschaftlichen Bereichen, wie z. B. in der Physik, Chemie, Biologie oder Sozialwissenschaften, sind mathematische Modelle zur Beschreibung von Systemen oder Prozessen von großer Bedeutung [19]. Da man daran interessiert ist mit solchen Modellen möglichst realitätsnahe Ergebnisse zu erzielen, wird oft eine Modellkalibrierung eingesetzt [64]. Bei einer Modellkalibrierung handelt es sich um eine Methode, mit der die Parameter eines mathematischen Modells mithilfe experimentell ermittelter Messdaten geschätzt werden um anschließend ein Modell zu erhalten, mit dem ein realer Prozess bestmöglich approximiert werden kann [64].

Wie in [17] erläutert, wird zur Schätzung von Modellparametern üblicherweise eine Parameteridentifikationsmethode genutzt. Da allerdings die Qualität der Parameteridentifikation stark davon abhängt, dass möglichst viele Experimente durchgeführt werden um Messfehler klein zu halten und zudem nicht jedes Experiment geeignete Messdaten liefert, ist es oft unabdingbar die Experimente im Vorhinein optimal zu planen. Dieses Vorgehen wird Optimale Versuchsplanung genannt. Mit ihr können Experimente konstruiert werden, mit denen Messdaten mit maximaler Information für die Parameteridentifikation gewonnen werden können. So kann die Genauigkeit beim Identifizieren der Parameter erhöht werden [7].

1.2 Inhalt

In Kapitel 2 dieser Arbeit wird anhand einer instationären partiellen Differentialgleichung (PDG) in allgemeiner Form beschrieben, wie ein unbekannter Modellparameter mit der Maximum-Likelihood-Methode geschätzt werden kann. Hierfür wird vorausgesetzt, dass die zur Parameterschätzung erforderlichen Messdaten einen Messfehler aufweisen, der normalverteilt mit Erwartungswert Null ist. Dann ermittelt der Maximum-Likelihood-Schätzer einen Modellparameter, für den die zugehörige Lösung einer instationären PDG minimalen Abstand zu den gemittelten Messreihen besitzt. Somit kann ein unbekannter Parameter durch Lösen eines nichtlinearen Optimierungsproblems geschätzt werden. Als Lösungsverfahren wird

- die Methode des steilsten Abstiegs,

- das Newton-Verfahren und
- das Gauss-Newton-Verfahren

beschrieben. Diese drei Verfahren werden in Kapitel 6 dieser Arbeit anhand eines nichtlinearen, instationären Problems eingesetzt um unbekannte Modellparameter numerisch zu ermitteln. Auf diese Weise kann gezeigt werden, dass alle drei Methoden zur Schätzung dieser Parameter ungeeignet sind, allerdings eine Kopplung dieser Verfahren eine gute Möglichkeit zur Parameterschätzung bieten.

In Kapitel 3 dieser Arbeit wird beschrieben, wie die Maximum-Likelihood-Schätzmethode durch Bestimmung einer optimalen Messstellenkonstellation optimiert werden kann: Es wird untersucht, wieviele Messstellen maximal benötigt werden und wo diese zu platzieren sind um eine Parameteridentifikation mit höchster Zuverlässigkeit durchführen zu können. Wie in [85] beschrieben, kann eine Optimierung dieser Schätzgenauigkeit durch eine Maximierung der Determinante der Fisher-Informationsmatrix erreicht werden. Eine auf diese Weise optimierte Messstellenkonfiguration wird D-optimales Design [85] genannt. Es wird untersucht, wie ein D-optimales Design mit minimaler Anzahl an Messstellen numerisch ermittelt werden kann. Da derzeitige Verfahren zur Bestimmung eines Designs diskret hergeleitet wurden und die Messstellen oft nur auf vorher festgelegten Diskretisierungspunkten liegen können, konnten nach dem derzeitigen Stand der Forschung noch keine Konvergenzaussagen getätigt werden. Aufgrund der Abhängigkeit des D-optimalen Designs von den Diskretisierungspunkten kann zudem nicht gewährleistet werden, dass eine globale Lösung gefunden werden kann. Viele Verfahren basieren zudem darauf, dass in jedem Iterationsschritt jeder Gitterpunkt „abgesucht“ wird, so dass das Abspeichern aller Lösungen in Raum und Zeit erforderlich ist und ein sehr hoher Rechenaufwand erwartet werden kann. Ein derartiges Lösungsverfahren zur iterativen Bestimmung eines D-optimalen Designs wurde in [85] hergeleitet und wird in Kapitel 3 wiedergegeben.

Da insbesondere die Konvergenz gegen eine kontinuierliche Lösung der zur Zeit verwendeten Lösungsverfahren bei einer „h-Verfeinerung“ nicht gewährleistet werden kann, wird in Kapitel 4 dieser Arbeit eine neuartige Vorgehensweise zur Bestimmung eines D-optimalen Designs vorgestellt, so dass Konvergenzaussagen getroffen werden können. Es werden diesbezüglich zwei Verfahren beschrieben, mit denen diese Vorgehensweise numerisch umgesetzt werden kann:

- ein primal-duales Innere-Punkte-Verfahren und
- eine Active-Set-Methode.

Für beide Verfahren werden Konvergenzaussagen formuliert. Dabei wird sich herausstellen, dass die Active-Set-Methode dem Innere-Punkte-Verfahren bei der Bestimmung eines D-optimalen Designs vorzuziehen ist, da neben einer gleichen Konvergenzordnung das zu lösende System wesentlich kleiner ausfällt.

Für die Berechnung eines D-optimalen Designs mit dem Innere-Punkte-Verfahren oder der Active-Set-Methode wird der Ortsgradient der Sensitivitäten benötigt. Da die Sensitivitäten als Finite-Elemente-Lösung einer partiellen Differentialgleichung üblicherweise H^1-Funktionen darstellen und folglich nur einmal schwach differenzierbar sind, ist dieser Gradient als L^2-Funktion auf Nullmengen nicht eindeutig definiert. So kann eine Punktauswertung des Gradienten nicht ohne Regularisierung durchgeführt werden,

obgleich dieses für die Umsetzung der Active-Set-Methode unabdingbar ist. Daher wird in Kapitel 4 gezeigt, wie eine solche Regularisierung mit der δ-Distribution durchgeführt werden kann. Das in Kapitel 3 beschriebene Standardverfahren und die in Kapitel 4 hergeleitete Active-Set-Methode werden dann in Kapitel 5 am Beispiel einer zweidimensionalen Wärmeleitungsgleichung numerisch umgesetzt und anschließend miteinander verglichen.

Nachdem beschrieben wurde, wie ein unbekannter Modellparameter bei einer instationären partiellen Differentialgleichung geschätzt und wie eine solche Schätzung durch Bestimmung einer D-optimalen Messstellenkonstellation optimiert werden kann, wird in Kapitel 6 dieser Arbeit eine Parameterschätzung und Optimale Versuchsplanung numerisch umgesetzt. Hierfür wird das mathematische Modell zur Beschreibung eines präparativen Säulenchromatographieprozesses zum Aufreinigen von Proteingemischen betrachtet. Wie in [29] und [65] erläutert, kann ein solches Modell durch ein nichtlineares System parabolischer und gewöhnlicher Differentialgleichungen dargestellt werden. Dieses System enthält Modellparameter wie zum Beispiel den Diffusionskoeffizienten, die Säulenkapazität, die Geschwindigkeit des Gemisches sowie die SMA-Parameter (steric-mass-action) mit

- dem Adsorbtionskoeffizienten,
- dem Desorbtionskoeffizienten,
- der Charakteristischen Ladung
- und dem Schirmungskoeffizienten

eines Proteins [13], [29]. Da insbesondere die SMA-Parameter für die Identifizierung von Proteinen genutzt werden, müssen diese besonders exakt ermittelt werden. Üblicherweise werden diese getrennt voneinander durch Vereinfachung der SMA-Gleichungen (die zeitliche Änderung wird Null gesetzt) geschätzt [29]. Da allerdings auf diese Weise Fehler beim Schätzen verursacht werden, wird in Kapitel 6 dieser Arbeit beschrieben, wie die SMA-Parameter ohne Vereinfachung durch Lösen eines nichtlinearen Optimierungsproblems mit der Maximum-Likelihood-Methode bestimmt werden können. Um die SMA-Parameter zusätzlich bestmöglich schätzen zu können wird anschließend erläutert, wie mit der Bestimmung eines D-optimalen Designs das Volumen des Konfidenzellipsoids eines Schätzers minimiert werden kann um die Zuverlässigkeit des Schätzers zu erhöhen.

In Vorbereitung dessen wird zunächst beschrieben, wie das gekoppelte System von gewöhnlichen und parabolischen Differentialgleichungen zur Beschreibung eines solchen Prozesses mit der Methode der Finiten-Elemente numerisch gelöst werden kann. Anschließend werden am Beispiel des Proteins Lysozym bei pH 7 die SMA-Parameter mit den in Kapitel 2 beschriebenen Lösungsverfahren geschätzt. Die auf diese Weise ermittelten Schätzwerte resultieren allerdings aus keiner guten Schätzung der SMA-Parameter: Einer der vier unbekannten Parameter konnte auf diese Weise gar nicht und die übrigen drei nicht besonders gut geschätzt werden. Da in der präparativen Säulenchromatographie die Messwerte üblicherweise am Ausgang der Säule erhoben werden, wird daher im Anschluss mit der in Kapitel 4 beschriebenen Active-Set-Methode ein D-optimales Design ermittelt um zu überprüfen, ob durch Verwendung dieses Designs die Zuverlässigkeit der Parameterschätzung erhöht werden kann. Diesbezüglich werden auf Basis dieses D-optimalen Designs neue Messdaten generiert um anschließend die SMA-Parameter neu zu schätzen.

Kapitel 2

Modellkalibrierung bei instationären partiellen Differentialgleichungen

In diesem Kapitel wird beschrieben, wie unbekannte Modellparameter bei instationären partiellen Differentialgleichungen geschätzt werden können. In Vorbereitung dessen wird eine allgemeine Form von instationären partiellen Differentialgleichung eingeführt, die einen Modellparameter $\boldsymbol{\theta} \in \mathbb{R}^m$ enthalten. Das Ziel ist, diesen Parameter mithilfe von experimentell ermittelten Messdaten zu schätzen, falls der Messfehler normalverteilt mit Erwartungswert Null ist. In diesem Fall sind die drei Schätzverfahren

- Methode der kleinsten Quadrate,
- Momentenmethode,
- Maximum-Likelihood-Schätzung

identisch [17]. Wie in [17] zudem beschrieben ist, sind diese Verfahren die am häufigst verwendeten um Parameter aus Messdaten zu approximieren.

Ist der Messfehler normalverteilt mit Erwartungswert Null, kann zudem mithilfe der sogenannten Fisher-Informationsmatrix [85] die Qualität dieser Schätzmethoden überprüft werden, da die Inverse der Fisher-Informationsmatrix die Kovarianzmatrix eines Schätzers approximiert [85]. Dieser Zusammenhang wird in Abschnitt 2.1.2 erläutert. Anschließend wird anhand der instationären partiellen Differentialgleichung in allgemeiner Form detailliert beschrieben, wie ein unbekannter Modellparameter $\boldsymbol{\theta}$ geschätzt werden kann.

Dieses und die folgenden Kapitel dienen zur Vorbereitung des Kapitels 6, in dem ein präparativer Säulenchromatographieprozess zur Aufreinigung von Proteingemischen betrachtet wird. Ein solcher Prozess kann durch ein gekoppeltes System von instationären partiellen und gewöhnlichen Differentialgleichungen dargestellt werden. Diese Gleichungen enthalten proteinspezifische Eigenschaften in Form von Parametern, die in Kapitel 6 mit der in diesem Kapitel beschriebenen Vorgehensweise geschätzt werden.

2.1 Einführung

In dieser Arbeit wird vorausgesetzt, dass $\Omega \subset \mathbb{R}^d$ für $d \in \{1,2,3\}$ ein beschränktes, einfach zusammenhängendes, offenes Gebiet mit genügend glattem Rand $\partial\Omega$ ist. Dann stellt

$$\left.\begin{array}{rcl} \partial_t u(t,\mathbf{x},\boldsymbol{\theta}) & = & F(t,\mathbf{x},u(t,\mathbf{x},\boldsymbol{\theta}),\nabla_{\mathbf{x}}u(t,\mathbf{x},\boldsymbol{\theta}),\Delta_{\mathbf{x}}u(t,\mathbf{x},\boldsymbol{\theta}),\boldsymbol{\theta}) \\ & =: & F(t,\mathbf{x},\boldsymbol{\theta}), \end{array}\right\} \tag{2.1}$$

eine allgemeine Form einer instationären partiellen Differentialgleichung dar, wobei

$$u(t,\mathbf{x},\boldsymbol{\theta}) \;\in\; \mathcal{C}^1(T;\mathcal{C}^2(\Omega;\mathcal{C}^2(\Theta;\mathbb{R})))$$

den Zustand darstellt mit $\mathbf{x} \in \Omega \subset \mathbb{R}^d$, $t \in T = (0,t_f)$, $\infty > t_f \in \mathbb{R}$ und $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^m$. Der Zustand $u(t,\mathbf{x},\boldsymbol{\theta})$ erfüllt zudem die Randbedingung

$$r(t,\mathbf{x},\boldsymbol{\theta}) \;:=\; r\big(t,\mathbf{x},u(t,\mathbf{x},\boldsymbol{\theta}),\nabla_{\mathbf{x}}u(t,\mathbf{x},\boldsymbol{\theta}),\boldsymbol{\theta}\big) \;=\; 0, \quad \mathbf{x} \in \partial\Omega, \tag{2.2}$$

sowie die Anfangsbedingung

$$u(0,\mathbf{x},\boldsymbol{\theta}) \;=\; u_0(\mathbf{x}), \quad \mathbf{x} \in \Omega. \tag{2.3}$$

Um den Banachraum $\mathcal{C}^1(T;\mathcal{C}^2(\Omega;\mathcal{C}^2(\Theta;\mathbb{R})))$ definieren zu können, wird zunächst der Raum der k-mal stetig differenzierbaren Funktionen beschrieben:

Definition 2.1.1. (Raum der k-mal stetig differenzierbaren Funktionen)
Sei $(X,\|\cdot\|_X)$ ein Banachraum. Mit $k \in \mathbb{N}_0$ wird der Banachraum

$$\mathcal{C}^k(X) \;:=\; \{v : X \to \mathbb{R} : v \text{ ist } k\text{-mal stetig differenzierbar in } X\}$$

versehen mit der Norm

$$\|v\|_{\mathcal{C}^k(X)} \;:=\; \max_{|\alpha|\le k} \|D^\alpha v\|_\infty$$

als Raum der k-mal stetig differenzierbaren Funktionen bezeichnet.

Der Raum $\mathcal{C}^1(T;\mathcal{C}^2(\Omega;\mathcal{C}^2(\Theta;\mathbb{R})))$ kann nun aus folgender Definition abgeleitet werden:

Definition 2.1.2.
Sei $T \subset \mathbb{R}^{d_1}$ und $\Omega \subset \mathbb{R}^{d_2}$ mit $d_1,d_2 \in \{1,2,3\}$. Dann ist

$$\mathcal{C}^1(T;\mathcal{C}^2(\Omega;\mathbb{R})) \;:=\; \{v \in \mathcal{C}^1(\Omega\times T) :\, v(t) \in \mathcal{C}^2(\Omega) \text{ für jedes fest gewählte } t \in T\}\ .$$

Bemerkung 2.1.1.
Sei $T \subset \mathbb{R}^{d_1}$ und $\Omega \subset \mathbb{R}^{d_2}$ mit $d_1,d_2 \in \{1,2,3\}$. Versehen mit der Norm

$$\|v\|_{\mathcal{C}^1(T;\mathcal{C}^2(\Omega;\mathbb{R}))} \;:=\; \max_{|\alpha_1|\le 2,\,|\alpha_2|\le 1} \|D_x^{\alpha_1} D_t^{\alpha_2} v\|_\infty$$

ist $\mathcal{C}^1(T;\mathcal{C}^2(\Omega;\mathbb{R}))$ ein Banachraum.

Wie man sieht ist $u(t, \mathbf{x}, \boldsymbol{\theta}) \in \mathcal{C}^1(T; \mathcal{C}^2(\Omega; \mathcal{C}^2(\Theta; \mathbb{R})))$ abhängig von einem Modellparameter $\boldsymbol{\theta} \in \Theta$. Falls dieser Parameter unbekannt ist, kann dieser mithilfe einer *Parameteridentifikation* geschätzt werden. Was den Parameterraum Θ anbelangt, kann dieser wie folgt gewählt werden:

(1) Konstante Parameter:

$$\Theta_1 := \{\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^\top \in \mathbb{R}^m : a_i \leq \theta_i \leq b_i,\, i = 1, \ldots, m\}, \tag{2.4}$$

mit den Schranken $a_i, b_i \in \mathbb{R}$,

(2) Parameter, die von $\mathbf{x} \in \Omega, t \in T$ und dem Zustand $u := u(t, \mathbf{x}, \boldsymbol{\theta})$ abhängen:

$$\Theta_2 := \{\boldsymbol{\theta} = (\theta_1(t, \mathbf{x}, u), \ldots, \theta_m(t, \mathbf{x}, u))^\top \in \mathbb{R}^m : a_i \leq \theta_i \leq b_i,\, i = 1, \ldots, m\},$$

mit den Schranken $a_i, b_i \in \mathbb{R}$,

(3) Parameter, die von $\mathbf{x} \in \Omega, t \in T, u := u(t, \mathbf{x}, \boldsymbol{\theta})$ und den θ_i abhängen:

$$\Theta_3 := \{\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^\top \in \mathbb{R}^m : \theta_i = g(t, \mathbf{x}, u, \theta_1, \ldots, \theta_m)\}.$$

In dieser Arbeit wird der erste Fall betrachtet. Es ist $\Theta := \Theta_1$.

Es wird nun ein kurzer Überblick über bekannte Schätzmethoden gegeben, die in [17] detailliert erläutert wurden. Anschließend wird beschrieben, dass mithilfe einer Optimalen Versuchsplanung die Genauigkeit dieser Schätzmethoden erhöht werden kann.

2.1.1 Schätzmethoden

Um einen Parameter $\boldsymbol{\theta} \in \Theta$ schätzen zu können, wird eine Parameteridentifikationsmethode eingesetzt. Wie in [17] beschrieben sind drei der wichtigsten Vertreter:

Die Methode der kleinsten Quadrate

Die Methode der kleinsten Quadrate wird insbesondere im Bereich der Regressionsanalyse eingesetzt [17]. Bei dieser Methode werden die unbekannten Modellparameter $\boldsymbol{\theta} \in \mathbb{R}^m$ durch einen Schätzer $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{z})$ mittels Lösen eines Optimierungsproblems

$$\hat{\boldsymbol{\theta}}(\mathbf{z}) \;=\; \arg\min_{\Theta} \sum_{i=1}^{n} (z_i - g_i(\boldsymbol{\theta}))^2 \tag{2.5}$$

approximiert [17], wobei der Vektor $\mathbf{z} = (z_1, \ldots, z_n)^\top$ die beobachteten Daten darstellt und $g_i(\boldsymbol{\theta})$ die Lösung einer Zustandsgleichung, die z_i approximieren soll. Eine Lösung $\hat{\boldsymbol{\theta}}(\mathbf{z})$ von (2.5) wird Kleinste-Quadrate-Schätzer (KQS) genannt.

Die Momentenmethode

Wie in [17] erläutert wird bei der Momentenmethode die Abhängigkeit zwischen dem Modellparameter $\boldsymbol{\theta} \in \mathbb{R}^m$ und den Momenten $m_i(\boldsymbol{\theta})$ ermittelt und in Form eines Gleichungssystems dargestellt. Seien $m_i(\boldsymbol{\theta}) := E(Y^i)$, $i = 1, \ldots, r$ die ersten r Momente einer generischen Zufallsvariablen Y. Dann kann $\boldsymbol{\theta}$ durch eine stetige Funktion g mittels $\boldsymbol{\theta} = g(m_1(\boldsymbol{\theta}), \ldots, m_r(\boldsymbol{\theta}))$ dargestellt werden. Bei der Momentenmethode hat der Schätzer $\hat{\boldsymbol{\theta}}$ die Form $\hat{\boldsymbol{\theta}} = g(\hat{m}_1, \ldots, \hat{m}_r)$, wobei $\hat{m}_k$ das k-te Stichprobenmoment darstellt.

Die Maximum-Likelihood-Schätzung

Die Maximum-Likelihood-Methode ist eine häufig verwendete Methode zur Bestimmung von Schätzern [17]. Angenommen, $E = \{Z = z\}$ sei ein beobachtetes Ereignis mit der Wahrscheinlichkeit $\mathbb{P}_{\boldsymbol{\theta}}(E)$. Da E theoretisch unter allen Modellparametern $\boldsymbol{\theta}$ möglich ist, wird bei der Maximum-Likelihood-Methode der Parameter ermittelt, für den die Wahrscheinlichkeit $\mathbb{P}_{\boldsymbol{\theta}}(E)$ maximal ist, d. h.

$$L(\hat{\boldsymbol{\theta}}(z), z) \;=\; \max\{L(\boldsymbol{\theta}, z) : \boldsymbol{\theta} \in \Theta\} \qquad \forall\, z \in \mathcal{Z}, \tag{2.6}$$

wobei $L : \Theta \times \mathcal{Z} \to \mathbb{R}^+$ mit $L(\boldsymbol{\theta}, z) = \mathbb{P}_{\boldsymbol{\theta}}(E)$ die Likelihood-Funktion von $\boldsymbol{\theta}$ für die Beobachtung z darstellt mit dem Stichprobenraum $\mathcal{Z}$. Eine Lösung $\hat{\boldsymbol{\theta}}(z)$ von (2.6) wird Maximum-Likelihood-Schätzer (MLS) genannt.

Bemerkung 2.1.2. *Wie in [17] erläutert, sind die beschriebenen Schätzmethoden identisch, wenn bei der Stichprobenerhebung für den Messfehler $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ gilt.*

2.1.2 Optimale Versuchsplanung zur Erhöhung der Zuverlässigkeit von Schätzmethoden

In [85] wurde beschrieben, dass ein unbekannter Modellparameter umso genauer durch einen Schätzer $\hat{\boldsymbol{\theta}}$ approximiert werden kann, wenn die zugehörige Kovarianzmatrix $\mathrm{cov}(\hat{\boldsymbol{\theta}})$ minimal bezüglich eines Optimalitätskriteriums $G : \mathbb{R}^{m\times m} \to \mathbb{R}$ ist. Die am häufigsten verwendeten Kriterien sind nach [85]

(1) das **D-Optimalitätskriterium** (2.7)

$$G(\mathrm{cov}\{\hat{\boldsymbol{\theta}}\}) \;=\; \ln\det(\mathrm{cov}\{\hat{\boldsymbol{\theta}}\}) \;:=\; \ln|\mathrm{cov}\{\hat{\boldsymbol{\theta}}\}|,$$

(2) das **E-Optimalitätskriterium** (2.8)

$$G(\mathrm{cov}\{\hat{\boldsymbol{\theta}}\}) \;=\; \lambda_{\max}(\mathrm{cov}\{\hat{\boldsymbol{\theta}}\}),$$

wobei $\lambda_{\max}$ den größten Eigenwert von $\mathrm{cov}\{\hat{\boldsymbol{\theta}}\}$ darstellt, (2.9)

(3) das **A-Optimalitätskriterium**

$$G(\mathrm{cov}\{\hat{\boldsymbol{\theta}}\}) \;=\; \mathrm{tr}(\mathrm{cov}\{\hat{\boldsymbol{\theta}}\})$$

mit der Spur $\mathrm{tr} : \mathbb{R}^{m\times m} \to \mathbb{R}$.

Dass die Minimierung der Kovarianzmatrix $\mathrm{cov}\{\hat{\boldsymbol{\theta}}\}$ bezüglich eines der Kriterien (2.7) - (2.9) die Zuverlässigkeit eines Schätzers $\hat{\boldsymbol{\theta}}$ erhöht, kann am zugehörigen Konfidenzellipsoid $\mathcal{K}$ erläutert werden. Ein Konfidenzellipsoid $\mathcal{K}$ ist ein Maß für die Präzision eines Schätzverfahrens [17]. Je kleiner das Ellipsoid ist, desto besser kann ein gesuchter Modellparameter geschätzt werden:

Das D-Optimalitätskriterium (2.7)

Die Minimierung der Determinante der Kovarianzmatrix ist äquivalent zur Minimierung des Volumens des Konfidenzellipsoids $\mathcal{K}$. Es gilt mit den Eigenwerten $\lambda_i \geq 0$ der Matrix $\mathrm{cov}\{\hat{\boldsymbol{\theta}}\}$ der Zusammenhang

$$|\mathrm{cov}\{\hat{\boldsymbol{\theta}}\}| \;=\; \prod_{i=1}^{m} \lambda_i \;=\; \prod_{i=1}^{m} c_i a_i^2, \tag{2.10}$$

wobei a_i für $i = 1, \ldots, m$ die Länge der i-ten Hauptachse des Konfidenzellipsoids $\mathcal{K}$ darstellt und $c_i > 0$ konstant ist [89]. Für das Volumen von $\mathcal{K}$ gilt

$$\text{vol}(\mathcal{K}) \;=\; \frac{4}{3}\pi \prod_{i=1}^{m} a_i.$$

Das E-Optimalitätskriterium (2.8)

Die Minimierung des größten Eigenwertes $\lambda_{\max}$ der Kovarianzmatrix $\text{cov}\{\hat{\boldsymbol{\theta}}\}$ ist äquivalent zur Minimierung der längsten Hauptachse des Konfidenzellipsoids. Mit den Hauptachsenlängen a_i und den Eigenwerten λ_i von $\text{cov}\{\hat{\boldsymbol{\theta}}\}$ gilt der Zusammenhang [89]

$$a_i = c_i' \sqrt{\lambda_i}, \qquad i = 1, \ldots, m,$$

wobei $c_i' = \frac{1}{\sqrt{c_i}}$ mit c_i aus (2.10).

Das A-Optimalitätskriterium (2.9)

Die Minimierung der Spur von der Kovarianzmatrix $\text{cov}\{\hat{\boldsymbol{\theta}}\}$ ist äquivalent zur Minimierung der durchschnittlichen Länge der Hauptachsen des Konfidenzellipsoids (in Abbildung 2.1 durch den Radius eines Kreises dargestellt).

Bemerkung 2.1.3. *Der Eigenvektor $\mathbf{v}^i \in \mathbb{R}^m$ zum Eigenwert λ_i für $i = 1, \ldots, m$ definiert die Richtung der i-ten Hauptachse des Konfidenzellipsoids $\mathcal{K}$.*

Der Zusammenhang zwischen der Kovarianzmatrix $\text{cov}\{\hat{\boldsymbol{\theta}}\}$ und dem zugehörigen Konfidenzellipsoid $\mathcal{K}$ kann wie folgt grafisch veranschaulicht werden:

Konfidenzellipsoid $\mathcal{K}$ mit den Optimalitätskriterien (2.7) - (2.9)

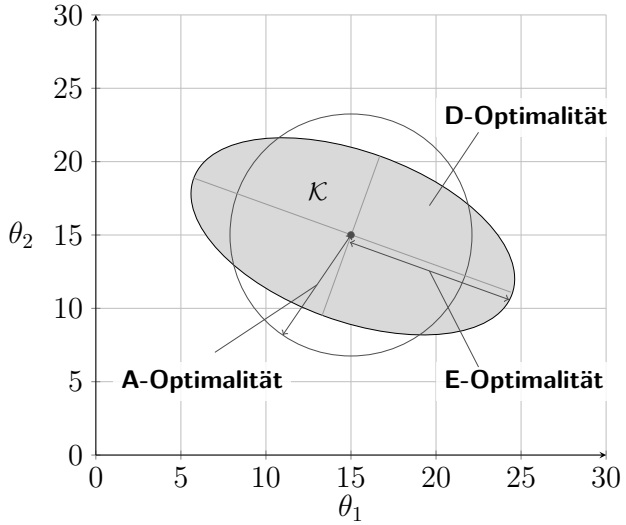


Abbildung 2.1: Das Konfidenzellipsoid $\mathcal{K}$ um einen Schätzwert $\hat{\boldsymbol{\theta}} = (15, 15)^\top$. In dessen Volumen befindet sich mit hoher Wahrscheinlichkeit der gesuchte Modellparameter $\boldsymbol{\theta}^* \in \Theta$. Je kleiner das Konfidenzellipsoid, desto genauer kann $\boldsymbol{\theta}^*$ durch $\hat{\boldsymbol{\theta}}$ approximiert werden.

In dieser Arbeit wird insbesondere das D-Optimalitätskriterium (2.7) untersucht und verwendet. In Kapitel 3 wird hierfür ein Lösungsverfahren aus dem Jahre 2005 vorgestellt, mit dem die Determinate der Kovarianzmatrix durch Bestimmung einer optimalen Messstellenkonstellation $\mathbf{x}^* = ({\mathbf{x}^{1,*}}^\top, \ldots, {\mathbf{x}^{\ell,*}}^\top)^\top$ minimiert wird. Die an den Messpositionen $\mathbf{x}^{i,*}, i \in I_\ell$ erhobenen Messdaten enthalten dann maximale Information, so dass eine Parameteridentifikation bestmöglich durchgeführt werden kann.

Ausgehend von der Bestimmung einer optimalen Messstellenkonstellation zur Minimierung der Kovarianzmatrix bezüglich (2.7) wird anschließend in Kapitel 4 eine neue Vorgehensweise vorgestellt um $\mathbf{x}^*$ zu bestimmen. Die Verfahren aus Kapitel 3 und 4 werden dann in Kapitel 5 am Beispiel einer zweidimensionalen Wärmeleitungsgleichung miteinander verglichen.

2.2 Parameteridentifikation

Eine Lösung $u(t, \mathbf{x}, \boldsymbol{\theta})$ der instationären partiellen Differentialgleichung (2.1) - (2.3) hängt von einem Parameter $\boldsymbol{\theta} \in \Theta$ ab, der mithilfe einer Parameteridentifikation durch einen Schätzer $\hat{\boldsymbol{\theta}}$ näherungsweise ermittelt werden kann. Wie eine solche Schätzung realisiert werden kann, wird im folgenden Abschnitt beschrieben. Zur übersichtlichen Darstellung sei

$$I_\ell \;:=\; \{i \in \mathbb{N} : \, 1 \leq i \leq \ell\}$$

die Indexmenge der Messpositionen, an denen die für die Parameterschätzung erforderlichen Messdaten erhoben werden. In dieser Arbeit werden Parameterschätzmethoden betrachtet für die die folgenden zwei Voraussetzungen gelten:

Voraussetzung 2.2.1.

1. *Für den Messfehler $\varepsilon(t, \mathbf{x}) = \varepsilon(t, \mathbf{x}, \hat{\boldsymbol{\theta}})$ wird vorausgesetzt, dass $\varepsilon(t, \mathbf{x}) \sim \mathcal{N}(0, \sigma^2)$ mit den Erwartungswerten*

   $$E\{\varepsilon(t, \mathbf{x})\} = 0 \quad \textit{und} \quad E\{\varepsilon(t, \mathbf{x}^i)\varepsilon(t', \mathbf{x}^j)\} = q(t, \mathbf{x}^i, \mathbf{x}^j)\delta(t - t') \tag{2.11}$$

   für alle $i, j \in I_\ell$ mit der Dirac-delta-Funktion δ.

2. *Zudem wird vorausgesetzt, dass der Messfehler räumlich unkorreliert ist, somit vereinfacht sich*

   $$q(t, \mathbf{x}^i, \mathbf{x}^j) \;:=\; \sigma^2 \delta_{ij}, \tag{2.12}$$

   so dass

   $$E\{\varepsilon(t, \mathbf{x}^i)\varepsilon(t', \mathbf{x}^j)\} \;=\; \sigma^2 \delta_{ij} \delta(t - t')$$

   mit der Kronecker-Delta-Funktion δ_{ij} gilt.

Wie in [17] beschrieben, sind die Schätzverfahren *Methode der kleinsten Quadrate*, *Momentenmethode* und *Maximum-Likelihood-Schätzung* unter der Voraussetzung 2.2.1 identisch. Der besseren Übersicht wegen wird in dieser Arbeit aber immer bei einem Schätzer von einem Maximum-Likelihood-Schätzer gesprochen um hervorzuheben, dass die Vorgehensweise der *Maximum-Likelihood-Schätzung* verfolgt wird. Trotzdem kann

gezeigt werden, dass die Vorgehensweisen der *Methode der kleinsten Quadrate* und der *Momentenmethode* auf das selbe zu lösende Optimierungsproblem führen.

Wie man unter der Voraussetzung 2.2.1 einen unbekannten Parameter $\boldsymbol{\theta}$ mit der Maximum-Likelihood-Methode schätzen kann, wird im folgenden Abschnitt zunächst allgemein und anschließend am Beispiel der eindimensionalen Wärmeleitungsgleichung erläutert.

2.2.1 Umsetzung einer Parameteridentifikationsmethode

Unter der Voraussetzung 2.2.1 kann ein Modellparameter $\boldsymbol{\theta}$ aus (2.1) - (2.3) wie folgt geschätzt werden:

Schritt 1: Experimentelle Bestimmung von Messdaten

Mithilfe von ℓ Sensoren werden zunächst an zuvor festgelegten Messpositionen $\mathbf{x}^i \in \bar{\Omega}$, $i \in I_\ell$ insgesamt M Messungen unternommen. Die Messwerte werden durch $z_i^j(t)$, $i \in I_\ell$, $j = 1, \ldots, M$ dargestellt. Die gemittelten Messwerte an der Messstelle $\mathbf{x}^i$ werden mit

$$z_i(t) \;:=\; \frac{1}{M}\sum_{j=1}^{M} z_i^j(t), \quad i \in I_\ell, \tag{2.13}$$

bezeichnet.

Schritt 2: Parameterschätzung durch Lösen eines Optimierungsproblems

Nachdem die Messreihen $z_i(t)$ an den Messpositionen $\mathbf{x}^i$ erhoben wurden, kann der unbekannte Parameter $\boldsymbol{\theta}^*$ durch Lösen eines Optimierungsproblems geschätzt werden: Für den Maximum-Likelihood-Schätzer $\hat{\boldsymbol{\theta}}$ gilt [17]

$$\hat{\boldsymbol{\theta}} \;=\; \arg\min_{\Theta} J(\boldsymbol{\theta}) \;=\; \frac{1}{2}\int_T \|\tilde{\mathbf{z}}(t) - \mathcal{P}u(\boldsymbol{\theta})\|^2_{Q^{-1}(t)}dt, \tag{2.14}$$

wobei $u(\boldsymbol{\theta}) := u(t, \mathbf{x}, \boldsymbol{\theta})$ eine Lösung der Zustandsgleichung (2.1) - (2.3) darstellt.

Die Matrix $Q(t) = [q(t, \mathbf{x}^i, \mathbf{x}^j)]_{i,j=1}^{\ell} \in \mathbb{R}^{\ell\times\ell}$ in (2.14) ist mit q aus (2.12) positiv definit und es gilt

$$\|\mathbf{v}\|^2_{Q^{-1}(t)} \;:=\; \mathbf{v}^\top Q^{-1}(t)\mathbf{v} \;\overset{(2.12)}{=}\; \frac{1}{\sigma^2}\mathbf{v}^\top\mathbf{v}, \quad \forall\, \mathbf{v} \in \mathbb{R}^\ell.$$

Der Vektor

$$\tilde{\mathbf{z}}(t) \;:=\; (z_1(t), \ldots, z_\ell(t))^\top$$

stellt mit $z_i(t)$ aus (2.13) die gemittelten Messreihen dar und es gilt

$$\mathcal{P}u(\boldsymbol{\theta}) := (u(t, \mathbf{x}^1, \boldsymbol{\theta}), \ldots, u(t, \mathbf{x}^\ell, \boldsymbol{\theta}))^\top.$$

Der Operator $\mathcal{P}$ ist demnach wie folgt definiert:

Definition 2.2.1. *Mit den fest gewählten Messpositionen* $\vec{\mathbf{x}} := ({\mathbf{x}^1}^\top, \ldots, {\mathbf{x}^\ell}^\top)^\top$ *ist*

$$\mathcal{P} : \mathcal{C}^1(T; \mathcal{C}^2(\Omega; \mathcal{C}^2(\Theta; \mathbb{R}))) \to \mathcal{C}^1(T)^\ell$$

eine Abbildung definiert durch

$$\mathcal{P}u(\boldsymbol{\theta}) := \mathcal{P}u(t, \mathbf{x}, \boldsymbol{\theta}) = (u(t, \mathbf{x}^1, \boldsymbol{\theta}), \ldots, u(t, \mathbf{x}^\ell, \boldsymbol{\theta}))^\top.$$

Da $\sigma^2 > 0$ ist, kann o. B. d. A. $\sigma^2 := 1$ gesetzt werden. Auf diese Weise erhält man das zu (2.14) äquivalente Optimierungsproblem

$$\min_{\Theta} J(\boldsymbol{\theta}) \quad = \quad \frac{1}{2} \int_T \|\tilde{\mathbf{z}}(t) - \mathcal{P}u(\boldsymbol{\theta})\|_2^2 \, dt. \tag{2.15}$$

Parameteridentifikation am Beispiel der eindimensionalen Wärmeleitungsgleichung

Anhand der eindimensionalen Wärmeleitungsgleichung wird in diesem Abschnitt gezeigt, wie die zuvor beschriebene Parameteridentifikation unter der Voraussetzung 2.2.1 durchgeführt werden kann. Dieses Beispiel ist dem Buch [85] entnommen.

Beispiel 2.2.1. (Die 1D-Wärmeleitungsgleichung)
Wie in [85] erläutert, beschreibt die parabolische Differentialgleichung

$$\partial_t u(t, x) \quad = \quad \theta \Delta_x u(t, x), \quad x \in \Omega := (0, 1), \quad t \in T := (0, t_f), \tag{2.16}$$

die zeitliche Änderung der Temperatur in einem Metalldraht. Ein solcher Draht kann über seinen Diffusionskoeffizienten $\theta > 0$ identifiziert werden. Es wird nun beschrieben, wie mithilfe einer Parameteridentifikation ein solcher Diffusionskoeffizient mithilfe von Messdaten geschätzt werden kann. Für die Datenerhebung werden die Experimente wie folgt durchgeführt:

<u>Schritt 1:</u> Experimentelle Bestimmung von Messdaten

Zu Beginn der Messungen hat der Metalldraht die Anfangstemperatur

$$u(0, x) \quad = \quad \sin(\pi x) \ [\text{in } ^\circ\text{C}] \quad \text{für } x \in \Omega. \tag{2.17}$$

Was die Temperatur am Rand betrifft, wird diese während des gesamten Experimentes so gesteuert, dass

$$u(t, 0) \ = \ u(t, 1) \ = \ 0 \ [\text{in } ^\circ\text{C}] \quad \text{für } t \in T, \tag{2.18}$$

gilt. Für die Datenerhebung wird als Messposition $x := 0.5$ gewählt. (In Kapitel 3.3 wird gezeigt, dass $x^* = 0.5$ die optimale Messstelle darstellt.) An der Position $x = 0.5$ wird mit einem Sensor die Temperatur 30 Sekunden lang gemessen. Nach der Durchführung von M Messungen, werden diese gemittelt und durch folgende Grafik dargestellt:

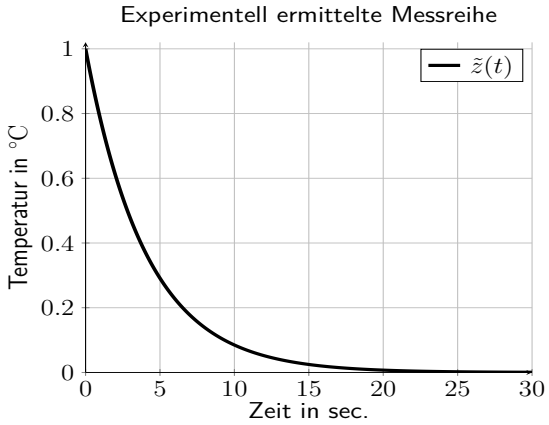


Abbildung 2.2: Durchschnitt der experimentell ermittelten Temperatur $\tilde{z}(t)$ eines Metalldrahtes an der Messposition $x = 0.5$.

Schritt 2: Parameterschätzung durch Lösen eines Optimierungsproblems

Die Wärmeleitungsgleichung (2.16) ist mit der Anfangsbedingung (2.17) und der Randbedingung (2.18) für jedes feste θ eindeutig lösbar [85]. Die analytische Lösung ist bekannt und lautet

$$u(t, x, \theta) \quad = \quad \exp(-\theta\pi^2 t)\sin(\pi x). \tag{2.19}$$

Diese wird in das Optimierungssystem (2.14) eingesetzt. Dann gilt für den Schätzer $\hat{\theta}$ mit der gemittelten Messreihe $\tilde{z}(t)$

$$\hat{\theta} \quad = \quad \arg\min_{\Theta} J(\theta) = \frac{1}{2}\int_T \|\tilde{z}(t) - \exp(-\theta\pi^2 t)\sin(\pi x)\|_2^2 \, dt.$$

In Kapitel 2.3 werden Verfahren beschrieben, mit denen eine optimale Lösung eines solchen Optimierungsproblems ermittelt werden kann. Es gilt: $\hat{\theta} \approx 0.025$.

Da im allgemeinen die analytische Lösung einer partiellen Differentialgleichung nicht ermittelt werden kann, wird die Gleichung (2.1) - (2.3) in dieser Arbeit numerisch betrachtet und mithilfe der Finiten-Elemente-Methode und dem Crank-Nicolson-Verfahren gelöst. In Vorbereitung dessen wird im folgenden Abschnitt die schwache Formulierung der Zustandsgleichung (2.1) - (2.3) aufgestellt.

2.2.2 Schwache Formulierung der Zustandsgleichung

Um einen unbekannten Modellparameter $\boldsymbol{\theta} \in \Theta$ mit der Maximum-Likelihood-Methode numerisch bestimmen zu können, wird eine Lösung der Zustandsgleichung (2.1) - (2.3) benötigt. Da im allgemeinen die analytische Lösung nicht ermittelt werden kann, wird diese in dieser Arbeit mit der Methode der finiten Elemente und mit dem Crank-Nicolson-Verfahren numerisch berechnet.

Die schwache Formulierung der Zustandsgleichung (2.1) - (2.3) erhält man, indem (2.1) - (2.3) zunächst mit einer beliebigen, aber festen Testfunktion $\varphi \in \mathcal{C}_c^\infty(\Omega)$ multipliziert und anschließend über Ω integriert wird. Der Raum der Testfunktionen ist dabei wie folgt definiert:

Definition 2.2.2. (Testfunktionen)
Sei $\Omega \subset \mathbb{R}^d$ offen. Dann ist

$$\mathcal{C}_c^\infty(\Omega) \quad := \quad \{v \in \mathcal{C}^\infty(\Omega) : \text{supp}(v) = \overline{\{\mathbf{x} : v(\mathbf{x}) \neq 0\}} \subset \Omega \text{ ist kompakt}\}$$

die Menge aller Funktionen $v \in \mathcal{C}^\infty(\Omega)$, deren Träger kompakte Teilmengen von Ω sind. Die Elemente von $\mathcal{C}_c^\infty(\Omega)$ heißen Testfunktionen.

In diesem Abschnitt wird vorausgesetzt, dass für jedes fest gewählte $\boldsymbol{\theta} \in \Theta$

$$u(t, \mathbf{x}, \boldsymbol{\theta}) \in \mathcal{C}^1(T; \mathcal{C}^2(\Omega; \mathbb{R}))$$

und mit $u(t)(\mathbf{x}) := u(t, \mathbf{x}, \boldsymbol{\theta})$ für jedes $t \in T$

$$u(t) \in \mathcal{C}^2(\Omega; \mathbb{R})$$

ist. Wie in [69] beschrieben, gilt dann für jedes fest gewählte $\boldsymbol{\theta} \in \Theta$ und jedes $t \in T$ mit $F(t)(\mathbf{x}) := F(t, \mathbf{x}, \boldsymbol{\theta})$ aus (2.1) und $r(t)(\mathbf{x}) := r(t, \mathbf{x}, \boldsymbol{\theta})$ aus (2.2) der Zusammenhang

$$\begin{aligned} (\partial_t u(t) - F(t), \varphi)_{L^2(\Omega)} &= 0, && (2.20)\\ (r(t), \varphi)_{L^2(\partial\Omega)} &= 0, && (2.21)\\ (u(0) - u_0, \varphi)_{L^2(\Omega)} &= 0, && (2.22) \end{aligned}$$

$\forall \varphi \in \mathcal{C}_c^\infty(\Omega)$ mit den Skalarprodukten

$$(v, \varphi)_{L^2(\Omega)} := \int_\Omega v\varphi d\mathbf{x} \quad \text{und} \quad (v, \varphi)_{L^2(\partial\Omega)} := \int_{\partial\Omega} v\varphi d\sigma.$$

Insbesondere muss demnach gefordert werden, dass

$$\partial_t u(t) - F(t) \in L^2(\Omega), \quad r(t) \in L^2(\partial\Omega) \quad \text{und} \quad u(0) - u_0 \in L^2(\Omega)$$

gilt, wobei $L^2(\Omega)$ und $L^2(\partial\Omega)$ wie folgt definiert sind:

Definition 2.2.3.
Sei $\Omega \subset \mathbb{R}^d$ offen. Dann ist mit $0 < p < \infty$

$$L^p(\Omega) \quad := \quad \{v : \Omega \to \mathbb{R} : v \text{ ist messbar und } \int_\Omega |v(x)|^p dx < \infty\}.$$

Eine Lösung von (2.20) - (2.22) wird schwache Lösung von (2.1) - (2.3) genannt. Die Zustandsgleichung in schwacher Form besitzt den Vorteil, dass mithilfe der ersten Greenschen Identität (durch partielle Integration) die zweite Ableitung in (2.20) verschwindet und nur noch die erste schwache Ableitung gefordert werden muss.

Definition 2.2.4. (Schwache Ableitung)
Sei $\Omega \subset \mathbb{R}^d$ offen. Mit der Menge aller lokal integrierbaren Funktionen

$$L^1_{loc}(\Omega) \;:=\; \{u : \Omega \to \mathbb{K} : \mathbb{1}_K u \in L^1(K) \;\; \forall K \subseteq \Omega \text{ ist kompakt}\}$$

ist für $u \in L^1_{loc}(\Omega)$ die Funktion $v \in L^1_{loc}(\Omega)$ die α-te schwache Ableitung von u, wenn

$$(u, \nabla_\alpha \varphi) \;=\; (-1)^\alpha (v, \varphi) \quad \forall \varphi \in \mathcal{C}_c^\infty(\Omega)$$

mit $\alpha \in \mathbb{N}_0$ gilt. Man schreibt $\nabla_\alpha u := v$.

Wendet man die erste Greensche Identität in Gleichung (2.20) an um die zweite partielle Ableitung verschwinden zu lassen und setzt man anschließend (2.21) in die auftretenden Randintegrale ein, erhält man eine partielle Differentialgleichung in schwacher Form. Diese wird im folgenden mit der Abbildung $a : V \times V \to \mathbb{R}$ dargestellt durch

$$\partial_t (u(t), \varphi)_{L^2(\Omega)} \;=\; a(u(t), \varphi), \tag{2.23}$$

mit der Anfangsbedingung

$$(u(0) - u_0, \varphi)_{L^2(\Omega)} \;=\; 0,$$

$\forall \varphi \in \mathcal{C}_c^\infty(\Omega)$. Da die Existenz der zweiten Ableitung nicht mehr gefordert werden muss, kann

$$V \;:=\; H^1(\Omega)$$

gesetzt werden, wobei der Sobolevraum $H^1(\Omega)$ wie folgt definiert ist:

Definition 2.2.5. (Sobolevraum)
Sei $\Omega \subset \mathbb{R}^d$ offen. Dann ist mit $k \in \mathbb{N}_0$

$$H^k(\Omega) \;:=\; \{v : \Omega \to \mathbb{K} : \exists \; \nabla_\alpha v \in L^1_{loc}(\Omega) \;\; \forall |\alpha| \leq k\}$$

der Raum, für den v alle schwachen Ableitungen der Ordnung $|\alpha| \leq k$ besitzt.
Versehen mit der Norm

$$\|v\|_{H^k} \;:=\; \left(\sum_{|\alpha| \leq k} (\nabla_\alpha v, \nabla_\alpha v) \right)^{\frac{1}{2}}$$

ist $H^k(\Omega)$ ein Banachraum.

Besitzt das Gebiet Ω zudem die sogenannte 1-Fortsetzungseigenschaft, kann als Testraum der Sobolevraum V gewählt werden. Die 1-Fortsetzungseigenschaft ist wie folgt definiert:

Definition 2.2.6. (m-Fortsetzungseigenschaft)
Eine offene Menge $\Omega \subseteq \mathbb{R}^d$ hat die m-Fortsetzungseigenschaft ($m \in \mathbb{N}$), wenn es für alle $k \in \{1, \ldots, m\}$ einen stetig linearen Fortsetzungsoperator

$$F_k : H^k(\Omega) \to H^k(\mathbb{R}^d)$$

gibt mit $F_k f_{|\Omega} = f$, $f \in H^k(\mathbb{R}^d)$.

Da der Raum der Testfunktionen $\mathcal{C}_c^\infty(\mathbb{R}^d)$ dicht in $H^1(\mathbb{R}^d)$ liegt [12], d. h.

$$H^1(\mathbb{R}^d) = \overline{\mathcal{C}_c^\infty(\mathbb{R}^d)}$$

und in dieser Arbeit die 1-Fortsetzungseigenschaft gefordert wird, kann $V \subset \mathcal{C}_c^\infty(\Omega)$ als Raum der Testfunktionen gewählt werden.

Die schwache Formulierung der Zustandsgleichung (2.20) - (2.22) lautet dann:

Sei $\boldsymbol{\theta} \in \Theta$ fest gewählt. Gesucht ist $u(t) \in V$, so dass

$$\partial_t (u(t), \varphi)_{L^2(\Omega)} \quad = \quad a(u(t), \varphi), \tag{2.24}$$

mit der Anfangsbedingung

$$(u(0), \varphi)_{L^2(\Omega)} \quad = \quad (u_0, \varphi)_{L^2(\Omega)}, \tag{2.25}$$

$\forall\, \varphi \in V$.

Es wird nun erläutert, wie das Parameteridentifikationsproblem (2.14) mit der klassischen Optimierungstheorie gelöst werden kann. Hierfür werden zunächst die Optimalitätsbedingungen erster und zweiter Ordnung wiedergegeben. Im Anschluss werden die Verfahren

- Methode des steilsten Abstiegs,
- Newton-Verfahren,
- Gauss-Newton-Verfahren

vorgestellt, mit denen das Optimierungsproblem (2.14) gelöst werden kann.

2.2.3 Optimalitätsbedingungen erster Ordnung

Wie in Kapitel 2.2.1 beschrieben, kann ein unbekannter Parameter $\boldsymbol{\theta}^* \in \Theta$ durch Lösen des Optimierungsproblems (2.15) geschätzt werden. Da im Vorhinein allerdings nicht bekannt ist, ob es sich bei (2.15) um ein konvexes Optimierungsproblem handelt, werden in diesem und im nächsten Abschnitt die notwendigen und hinreichenden Bedingungen dafür angegeben, dass eine Lösung $\hat{\boldsymbol{\theta}}$ von (2.15) ein lokales Minimum darstellt. Anschließend werden in Kapitel 2.3 Verfahren vorgestellt, mit denen das Parameterschätzproblem (2.15) gelöst werden kann.

Um die Optimalitätsbedingungen erster Ordnung (notwendige Bedingungen) zu erhalten, wird das zum Problem (2.15) zugehörige Lagrangefunktional $\mathcal{L} : \Theta \times \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$ benötigt [43]. Dieses lautet mit dem Parametergebiet Θ aus (2.4) und dem euklidischen Skalarprodukt $(\cdot,\cdot)_2$

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\lambda_a}, \boldsymbol{\lambda_b}) \quad = \quad \frac{1}{2} \int_T \|\tilde{\mathbf{z}}(t) - \mathcal{P}u(t, \mathbf{x}, \boldsymbol{\theta})\|_2^2 \, dt + (\mathbf{a} - \boldsymbol{\theta}, \boldsymbol{\lambda_a})_2 + (\boldsymbol{\theta} - \mathbf{b}, \boldsymbol{\lambda_b})_2. \tag{2.26}$$

Im Lagrangefunktional (2.26) ist der Operator $\mathcal{P}$ definiert wie in Definition 2.2.1. Da die Nebenbedingungen des Optimierungsproblems (2.15) lediglich aus der oberen und unteren Abschätzung

$$\mathbf{a} - \boldsymbol{\theta} \;\leq\; \mathbf{0} \quad \text{und} \quad \boldsymbol{\theta} - \mathbf{b} \;\leq\; \mathbf{0}$$

aus (2.4) bestehen, ist nach [43] das Problem (2.15) in jedem Punkt $\boldsymbol{\theta} \in \Theta$ regulär. Daher ist das Optimierungsproblem insbesondere in jedem lokalen Minimum $\bar{\boldsymbol{\theta}} \in \Theta$ regulär, so dass folgender Satz gilt:

Satz 2.2.1. (Satz von Kuhn-Tucker, [43])
Sei $\bar{\boldsymbol{\theta}} \in \Theta$ ein lokales Minimum von (2.15). Da (2.15) in $\bar{\boldsymbol{\theta}}$ regulär ist, existiert ein $(\bar{\boldsymbol{\lambda}}_{\boldsymbol{a}}, \bar{\boldsymbol{\lambda}}_{\boldsymbol{b}}) \in \mathbb{R}^m \times \mathbb{R}^m$ mit den folgenden Eigenschaften:

(i) $\bar{\boldsymbol{\lambda}}_{\boldsymbol{a}},\ \bar{\boldsymbol{\lambda}}_{\boldsymbol{b}} \geq \mathbf{0}$,

(ii) $(a_i - \bar{\theta}_i)\bar{\lambda}_{a,i} = 0$ *und* $(\bar{\theta}_i - b_i)\bar{\lambda}_{b,i} = 0$ *für alle* $i = 1, \ldots, m$,
(Bedingungen vom komplementären Schlupf),

(iii) $\nabla_{\boldsymbol{\theta}} \mathcal{L}(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\lambda}}_{\boldsymbol{a}}, \bar{\boldsymbol{\lambda}}_{\boldsymbol{b}}) = \mathbf{0}$.

Die Gleichungen $\nabla_{\boldsymbol{\theta}} \mathcal{L}(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\lambda}}_{\boldsymbol{a}}, \bar{\boldsymbol{\lambda}}_{\boldsymbol{b}}) = \mathbf{0}$ und die Bedingungen vom komplementären Schlupf aus Satz 2.2.1 stellen mit $(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\lambda}}_{\boldsymbol{a}}, \bar{\boldsymbol{\lambda}}_{\boldsymbol{b}}) \in \Theta \times \mathbb{R}^m_+ \times \mathbb{R}^m_+$ die Optimalitätsbedingungen erster Ordnung dar und werden KKT-Bedingungen genannt [43]. Mit dem Gradienten

$$\nabla_{\boldsymbol{\theta}} J(\bar{\boldsymbol{\theta}}) \quad = \quad -\int_T \left[\mathcal{P}_2 \nabla_{\boldsymbol{\theta}} u(t, \mathbf{x}, \bar{\boldsymbol{\theta}})\right] \left(\tilde{\mathbf{z}}(t) - \mathcal{P}u(t, \mathbf{x}, \bar{\boldsymbol{\theta}})\right) dt \tag{2.27}$$

lauten diese

$$\left.\begin{array}{rcl} \nabla_{\boldsymbol{\theta}} \mathcal{L}(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\lambda}}_{\boldsymbol{a}}, \bar{\boldsymbol{\lambda}}_{\boldsymbol{b}}) & = & -\displaystyle\int_T \left[\mathcal{P}_2 \nabla_{\boldsymbol{\theta}} u(t, \mathbf{x}, \bar{\boldsymbol{\theta}})\right] \left(\tilde{\mathbf{z}}(t) - \mathcal{P}u(t, \mathbf{x}, \bar{\boldsymbol{\theta}})\right) dt - \bar{\boldsymbol{\lambda}}_{\boldsymbol{a}} + \bar{\boldsymbol{\lambda}}_{\boldsymbol{b}} \\ & = & \mathbf{0}, \\ (\mathbf{a} - \bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\lambda}}_{\boldsymbol{a}})_2 & = & \mathbf{0}, \quad (\bar{\boldsymbol{\theta}} - \mathbf{b}, \bar{\boldsymbol{\lambda}}_{\boldsymbol{b}})_2 \ = \ \mathbf{0}, \quad \mathbf{a} \leq \bar{\boldsymbol{\theta}} \leq \mathbf{b}, \quad \bar{\boldsymbol{\lambda}}_{\boldsymbol{a}}, \bar{\boldsymbol{\lambda}}_{\boldsymbol{b}} \geq \mathbf{0}, \end{array}\right\} \tag{2.28}$$

wobei der Operator $\mathcal{P}_2$ wie folgt definiert ist:

Definition 2.2.7. *Sei $\boldsymbol{\theta} \in \Theta$ fest gewählt. Mit den fest gewählten Messpositionen $\vec{\mathbf{x}} = ({\mathbf{x}^1}^\top, \ldots, {\mathbf{x}^\ell}^\top)^\top$ ist*

$$\mathcal{P}_2 : \mathcal{C}^1(T; V)^{m \times 1} \to \mathcal{C}^1(T)^{m \times \ell}$$

definiert durch

$$\mathcal{P}_2 \nabla_{\boldsymbol{\theta}} u(t, \mathbf{x}, \boldsymbol{\theta}) \ := \ \begin{pmatrix} \partial_{\theta_1} u(t, \mathbf{x}^1, \boldsymbol{\theta}) & \ldots & \partial_{\theta_1} u(t, \mathbf{x}^\ell, \boldsymbol{\theta}) \\ & \ddots & \\ \partial_{\theta_m} u(t, \mathbf{x}^1, \boldsymbol{\theta}) & \ldots & \partial_{\theta_m} u(t, \mathbf{x}^\ell, \boldsymbol{\theta}) \end{pmatrix}.$$

Ein Punkt $(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\lambda}}_{\boldsymbol{a}}, \bar{\boldsymbol{\lambda}}_{\boldsymbol{b}}) \in \Theta \times \mathbb{R}^m_+ \times \mathbb{R}^m_+$ der die Bedingungen (2.28) erfüllt, wird KKT-Punkt genannt [43].

Da das Optimierungsproblem (2.15) in jedem Punkt regulär ist, stellt jeder lokale Minimalpunkt $\bar{\boldsymbol{\theta}} \in \Theta$ mit der zugehörigen dualen Lösung $(\bar{\boldsymbol{\lambda}}_{\boldsymbol{a}}, \bar{\boldsymbol{\lambda}}_{\boldsymbol{b}}) \in \mathbb{R}^m_+ \times \mathbb{R}^m_+$ einen KKT-Punkt von (2.28) dar. Allerdings ist nicht jeder KKT-Punkt ein lokaler Minimalpunkt.

In diesem Kapitel werden Verfahren beschrieben, mit denen ein lokales Minimum von (2.15) durch Lösen der KKT-Bedingungen (2.28) numerisch ermittelt werden kann. Um

sicher zu stellen, dass der primale Anteil eines ermittelten KKT-Punktes eine lokale Minimallösung von (2.15) darstellt, werden zusätzlich die Optimalitätsbedingungen zweiter Ordnung benötigt. Diese werden im nächsten Unterkapitel wiedergegeben.

Zunächst wird beschrieben, wie man die Sensitivitäten

$$w_i(t, \mathbf{x}, \boldsymbol{\theta}) \;:=\; \partial_{\theta_i} u(t, \mathbf{x}, \boldsymbol{\theta}) \tag{2.29}$$

für $i = 1, \ldots, m$ berechnen kann, da diese zum Lösen der KKT-Bedingungen (2.28) benötigt werden. Diese erhält man als Lösung der sogenannten Sensitivitätsgleichungen erster Ordnung, die durch partielle Ableitung der Zustandsgleichung (2.1) - (2.3) aufgestellt werden können.

Sensitivitätsgleichungen erster Ordnung für das allgemeine Modell (2.1) - (2.3)

Leitet man die instationäre partielle Differentialgleichung (2.1) - (2.3) partiell nach den Modellparametern θ_i für $i = 1, \ldots, m$ ab, erhält man die i-te Sensitivitätsgleichung erster Ordnung

$$\partial_t w_i(t) \;=\; \partial_u F(t, \mathbf{x}, \boldsymbol{\theta}) w_i(t) + \partial_{\theta_i} F(t, \mathbf{x}, \boldsymbol{\theta}), \quad x \in \Omega, \quad t \in T, \tag{2.30}$$

mit der Randbedingung

$$\partial_u r(t, \mathbf{x}, \boldsymbol{\theta}) w_i(t) + \partial_{\theta_i} r(t, \mathbf{x}, \boldsymbol{\theta}) \;=\; 0, \tag{2.31}$$

für $\mathbf{x} \in \partial\Omega$, $t \in T$ und der Anfangsbedingung

$$w_i(0) \;=\; 0, \tag{2.32}$$

wobei zur übersichtlichen Darstellung

$$u(t) := u(t, \mathbf{x}, \boldsymbol{\theta}) \quad \text{und} \quad w_i(t) := w_i(t, \mathbf{x}, \boldsymbol{\theta})$$

gesetzt wurde.

Die Sensitivitätsgleichungen erster Ordnung stellen für jedes $i = 1, \ldots, m$ jeweils selbst eine instationäre partielle Differentialgleichung dar, die in dieser Arbeit analog zur Zustandsgleichung (2.1) - (2.3) mit der Methode der Finiten Elemente gelöst werden. Um auch hier die Existenz der zweiten Ableitung nicht fordern zu müssen, werden die zugehörigen schwachen Formulierungen benötigt. Diese erhält man, indem die Zustandsgleichung in schwacher Form (2.24) - (2.25) nach $\theta_i, i = 1, \ldots, m$ partiell abgeleitet wird. Mit der Abbildung $a : V \times V \to \mathbb{R}$ aus (2.23) gilt:

Sei $\boldsymbol{\theta} \in \Theta$ fest gewählt und $u(t) \in V$ eine Lösung von (2.24) - (2.25). Gesucht ist $w_i(t) \in V$ für $i = 1, \ldots, m$, so dass die Gleichung

$$\partial_t (w_i(t), \varphi)_{L^2(\Omega)} \;=\; \partial_u a(u(t), \varphi) w_i(t) + \partial_{\theta_i} a(u(t), \varphi), \tag{2.33}$$

mit der Anfangsbedingung

$$(w_i(0), \varphi)_{L^2(\Omega)} \;=\; 0, \tag{2.34}$$

$\forall\, \varphi \in V$ erfüllt ist.

Bemerkung 2.2.2. *Die Gleichungen (2.33) hängen für alle $i = 1, \ldots, m$ linear von den Sensitivitäten $w_i(t)$ ab und sind bei fest gewähltem $\boldsymbol{\theta} \in \Theta$ und bekannter Lösung $u(t)$ voneinander unabhängig lösbar.*

2.2.4 Optimalitätsbedingungen zweiter Ordnung

Im Kapitel 2.3 werden Verfahren vorgestellt, mit denen das Optimierungsproblem (2.15) durch Bestimmung eines KKT-Punktes $\bar{\mathbf{w}} := (\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\lambda}}_{\boldsymbol{a}}, \bar{\boldsymbol{\lambda}}_{\boldsymbol{b}}) \in \Theta \times \mathbb{R}^m_+ \times \mathbb{R}^m_+$ gelöst werden kann. Der primale Anteil $\bar{\boldsymbol{\theta}}$ eines ermittelten KKT-Punktes $\bar{\mathbf{w}}$ entspricht dann entweder einem Minimal-, einem Maximal- oder einem Sattelpunkt des Parameterschätzproblems (2.15). Da die Optimalitätsbedingungen erster Ordnung notwendig, aber nicht hinreichend dafür sind, dass $\bar{\boldsymbol{\theta}}$ eine Minimallösung von (2.15) entspricht, wird in diesem Abschnitt zusätzlich die notwendige und hinreichende Optimalitätsbedingung zweiter Ordnung wiedergegeben. Diese sind dem Buch [43] entnommen.

Folgender Satz gibt die notwendige Bedingung zweiter Ordnung für das Optimierungsproblem (2.15) wieder.

Satz 2.2.3. (Notwendige Bedingung zweiter Ordnung, [43])
Sei $\bar{\boldsymbol{\theta}} \in \Theta$ ein lokales Minimum von Problem (2.15) und $(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\lambda}}_{\boldsymbol{a}}, \bar{\boldsymbol{\lambda}}_{\boldsymbol{b}}) \in \Theta \times \mathbb{R}^m_+ \times \mathbb{R}^m_+$ ein KKT-Punkt von (2.28). Dann gilt mit dem Lagrangefunktional aus (2.26)

$$\mathbf{s}^\top \nabla^2_{\boldsymbol{\theta}} \mathcal{L}(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\lambda}}_{\boldsymbol{a}}, \bar{\boldsymbol{\lambda}}_{\boldsymbol{b}})\, \mathbf{s} \;\geq\; 0 \quad \forall\, \mathbf{s} \in L(\mathcal{S}_1; \bar{\boldsymbol{\theta}}),$$

wobei $L(\mathcal{S}_1; \bar{\boldsymbol{\theta}})$ den linearisierten Kegel [43] von Θ in $\bar{\boldsymbol{\theta}}$ darstellt mit

$$\begin{aligned} \mathcal{S}_1 \;:=\; \{\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^\top \in \Theta :\;\; & \theta_i = a_i \;\; \forall\, i = 1, \ldots, m \text{ mit } \bar{\lambda}_{a,i} > 0, \\ & \theta_i = b_i \;\; \forall\, i = 1, \ldots, m \text{ mit } \bar{\lambda}_{b,i} > 0\}. \end{aligned}$$

Der Satz 2.2.3 besagt, dass die Hessematrix der Lagrangefunktion auf dem linearisierten Kegel $L(\mathcal{S}_1; \bar{\boldsymbol{\theta}})$ positiv semidefinit ist, wenn $\bar{\boldsymbol{\theta}} \in \Theta$ ein lokales Minimum von Problem (2.15) darstellt.

Die hinreichende Bedingung zweiter Ordnung lautet:

Satz 2.2.4. (Hinreichende Bedingung zweiter Ordnung, [43])
Sei $(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\lambda}}_{\boldsymbol{a}}, \bar{\boldsymbol{\lambda}}_{\boldsymbol{b}}) \in \Theta \times \mathbb{R}^m_+ \times \mathbb{R}^m_+$ ein KKT-Punkt von (2.28) und sei

$$\mathbf{s}^\top \nabla^2_{\boldsymbol{\theta}} \mathcal{L}(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\lambda}}_{\boldsymbol{a}}, \bar{\boldsymbol{\lambda}}_{\boldsymbol{b}})\, \mathbf{s} \;>\; 0 \quad \forall\, \mathbf{s} \in L(\mathcal{S}_1; \bar{\boldsymbol{\theta}}) \text{ mit } \mathbf{s} \neq 0. \tag{2.35}$$

Dann ist $\bar{\boldsymbol{\theta}} \in \Theta$ ein strikt lokales Minimum von (2.15).

Die Hessematrix des Lagrangefunktionals aus (2.26) entspricht der Hessematrix des Zielfunktionals $J(\bar{\boldsymbol{\theta}})$. Es gilt

$$\begin{aligned} \nabla^2_{\boldsymbol{\theta}} \mathcal{L}(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\lambda}}_{\boldsymbol{a}}, \bar{\boldsymbol{\lambda}}_{\boldsymbol{b}}) \;&=\; -\int_T [\mathcal{P}_3 \nabla^2_{\boldsymbol{\theta}} u(t, \mathbf{x}, \bar{\boldsymbol{\theta}})]\, (\tilde{\mathbf{z}}(t) - \mathcal{P} u(t, \mathbf{x}, \bar{\boldsymbol{\theta}}))\, dt \\ &\qquad + \int_T [\mathcal{P}_2 \nabla_{\boldsymbol{\theta}} u(t, \mathbf{x}, \bar{\boldsymbol{\theta}})]\, [\mathcal{P}_2 \nabla_{\boldsymbol{\theta}} u(t, \mathbf{x}, \bar{\boldsymbol{\theta}})]^\top\, dt \qquad (2.36) \\ &=\; \nabla^2_{\boldsymbol{\theta}} J(\bar{\boldsymbol{\theta}}), \end{aligned}$$

wobei der Operator $\mathcal{P}_3$ wie folgt definiert ist:

Definition 2.2.8. *Sei $\boldsymbol{\theta} \in \Theta$ fest gewählt. Mit den fest gewählten Messpositionen $\vec{\mathbf{x}} = ({\mathbf{x}^1}^\top, \ldots, {\mathbf{x}^\ell}^\top)^\top$ ist*

$$\mathcal{P}_3 : \mathcal{C}^1(T; V)^{m \times m} \to \mathcal{C}^1(T)^{m \times m \times \ell}$$

ein Operator definiert durch

$$\left[\mathcal{P}_3 \nabla^2_{\boldsymbol{\theta}} u(t, \mathbf{x}, \boldsymbol{\theta})\right]_{i,j,k} \;:=\; \partial_{\theta_i} \partial_{\theta_j} u(t, \mathbf{x}^k, \boldsymbol{\theta}).$$

Für die Berechnung der Hessematrix des Lagrangefunktionals werden die Sensitivitäten zweiter Ordnung

$$v_{ij}(t, \mathbf{x}, \boldsymbol{\theta}) \;:=\; \partial_{\theta_i} \partial_{\theta_j} u(t, \mathbf{x}, \boldsymbol{\theta})$$

für $i, j = 1, \ldots, m$ benötigt. Diese erhält man als Lösung der sogenannten Sensitivitätsgleichungen zweiter Ordnung.

Sensitivitätsgleichungen zweiter Ordnung für das allgemeine Modell (2.1) - (2.3)

In diesem Abschnitt werden die Sensitivitätsgleichungen zweiter Ordnung aufgestellt. Die Lösungen dieser Gleichungen werden benötigt um die Hessematrix des Lagrangefunktionals berechnen zu können. Die Sensitivitätsgleichungen zweiter Ordnung in starker Form erhält man durch partielle Ableitung der Gleichungen (2.30) - (2.32). Leitet man die i-te Sensitivitätsgleichung erster Ordnung nach θ_j für $j \in \{1, \ldots, m\}$ ab, gilt die partielle Differentialgleichung

$$\begin{aligned} \partial_t v_{ij}(t) \;&=\; \Big(\partial_u^2 F(t) w_j(t) + \partial_{\theta_j} \partial_u F(t)\Big) w_i(t) + \partial_u F(t) v_{ij}(t) \\ &\quad + \partial_{\theta_i} \partial_u F(t) w_j(t) + \partial_{\theta_j} \partial_{\theta_i} F(t) \end{aligned} \tag{2.37}$$

mit der Randbedingung

$$\begin{aligned} 0 \;&=\; \Big(\partial_u^2 r(t) w_j(t) + \partial_{\theta_j} \partial_u r(t)\Big) w_i(t) + \partial_u r(t) v_{ij}(t) \\ &\quad + \partial_{\theta_i} \partial_u r(t) w_j(t) + \partial_{\theta_j} \partial_{\theta_i} r(t) \end{aligned} \tag{2.38}$$

für $\mathbf{x} \in \partial\Omega$ und der Anfangsbedingung

$$v_{ij}(0) \;=\; 0. \tag{2.39}$$

Zur übersichtlichen Darstellung wurde

$$u(t) := u(t, \mathbf{x}, \boldsymbol{\theta}), \quad w_i(t) := w_i(t, \mathbf{x}, \boldsymbol{\theta}) \quad \text{und} \quad v_{ij}(t) := v_{ij}(t, \mathbf{x}, \boldsymbol{\theta})$$

gesetzt.

Wie die Zustandsgleichung und die Sensitivitätsgleichungen erster Ordnung werden die Sensitivitätsgleichungen zweiter Ordnung in dieser Arbeit mit der Methode der Finiten Elemente gelöst. Um auch hier die zweite Ableitung bezüglich der Ortsvariablen $\mathbf{x}$ nicht fordern zu müssen, wird die schwache Formulierung benötigt. Diese lautet mit der Abbildung $a : V \times V \to \mathbb{R}$ aus (2.23):

Sei $\boldsymbol{\theta} \in \Theta$ fest gewählt, $u(t) \in V$ eine Lösung von (2.24) - (2.25) und $w_i(t) \in V$ die i-te Lösung von (2.33) - (2.34). Gesucht ist $v_{ij}(t) \in V$ für $i, j = 1, \ldots, m$, so dass die Gleichungen

$$\begin{aligned} \partial_t \big(v_{ij}(t), \varphi\big)_{L^2(\Omega)} \;&=\; \Big(\partial_u^2 a\big(u(t), \varphi\big) w_j(t) w_i(t) + \partial_{\theta_j} \partial_u a\big(u(t), \varphi\big)\Big) w_i(t) \\ &\quad + \partial_u a\big(u(t), \varphi\big) v_{ij}(t) + \partial_{\theta_i} \partial_u a\big(u(t), \varphi\big) w_j(t) \\ &\quad + \partial_{\theta_j} \partial_{\theta_i} a\big(u(t), \varphi\big), \end{aligned} \tag{2.40}$$

mit den Anfangsbedingungen

$$(v_{ij}(0), \varphi)_{L^2(\Omega)} \quad = \quad 0, \tag{2.41}$$

$\forall\, \varphi \in V$ erfüllt ist.

Bemerkung 2.2.5. *Die Gleichungen (2.40) hängen für alle $i = 1, \ldots, m$ und $j = 1, \ldots, m$ linear von den Sensitivitäten $v_{ij}(t)$ ab und sind bei fest gewähltem $\boldsymbol{\theta} \in \Theta$ und bekannten Lösungen $u(t)$ und $w_i(t)$ für $i = 1, \ldots, m$ voneinander unabhängig lösbar.*

2.3 Optimierungsverfahren zur Parameteridentifikation

In Kapitel 2.2.1 wurde beschrieben, wie ein unbekannter Modellparameter $\boldsymbol{\theta}^* \in \Theta$ mit der Maximum-Likelihood-Methode geschätzt werden kann, wenn die Voraussetzung 2.2.1 erfüllt ist: Dann kann $\boldsymbol{\theta}^*$ durch Lösen eines Optimierungsproblems (2.14) approximiert werden. Es werden nun die Verfahren

- die Methode des steilsten Abstiegs,
- das Newton-Verfahren und
- das Gauss-Newton-Verfahren

betrachtet, mit denen das Optimierungsproblem (2.14) gelöst werden kann (siehe z. B. [43]).

Bei Parameteridentifikationsproblemen kann im allgemeinen die Menge Θ so gewählt werden, dass der gesuchte Modellparameter $\boldsymbol{\theta}^*$ im strikten Inneren Θ^0 von Θ liegt. Dann gilt für die Lösung $\hat{\boldsymbol{\theta}}$

$$\nabla_{\boldsymbol{\theta}} J(\hat{\boldsymbol{\theta}}) \quad = \quad 0, \tag{2.42}$$

mit dem Zielfunktional $J : \Theta \to \mathbb{R}$ aus (2.14). In dieser Arbeit wird vorausgesetzt, dass der unbekannte Parameter $\boldsymbol{\theta}^*$ derart eingegrenzt werden kann, dass $\boldsymbol{\theta}^* \in \Theta^0$ und folglich für den Schätzer $\hat{\boldsymbol{\theta}}$ (2.42) gilt.

2.3.1 Abstiegsverfahren

Wie in [43] beschrieben, ist der Gradient

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \quad = \quad -\int_T \left[\mathcal{P}_2 \nabla_{\boldsymbol{\theta}} u(t, \mathbf{x}, \boldsymbol{\theta})\right] \left(\tilde{\mathbf{z}}(t) - \mathcal{P} u(t, \mathbf{x}, \boldsymbol{\theta})\right) dt \tag{2.43}$$

ein Vektor, der in die Richtung des steilsten Anstiegs des Zielfunktionals J aus (2.14) im Punkt $\boldsymbol{\theta}$ zeigt. Somit zeigt der Vektor $-\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$ im Punkt $\boldsymbol{\theta}$ in die Richtung des steilsten Abstiegs. Ist der Gradient $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \neq \mathbf{0}$, dann existiert folglich immer eine Schrittweite $\alpha > 0$, so dass

$$J(\boldsymbol{\theta} - \alpha \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})) \; < \; J(\boldsymbol{\theta}).$$

Aufgrund dieser Eigenschaft kann durch sukzessives Ermitteln der Abstiegsrichtung $s := -\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}))$ ein lokales Minimum des Optimierungsproblems (2.14) wie folgt ermittelt werden:

Algorithmus 2.3.1. (Methode des steilsten Abstiegs, [43])

(S.0) *Initialisierung:*
Wähle ein $\boldsymbol{\theta}^0 \in \Theta$, $0 < \varepsilon \ll 1$ *und setze* $n := 0$.
Wähle ein $\alpha \in (0,1)$ *und* $\beta \in (0, 0.5)$ *für die Schrittweitenbestimmung.*

(S.1) *Lösen der Zustandsgleichung:*
Bestimme $u^n(t) := u(t; \boldsymbol{\theta}^n) \in V$, *so dass (2.24) - (2.25) für alle* $\varphi \in V$ *erfüllt ist.*

(S.2) *Lösen der Sensitivitätsgleichungen erster Ordnung:*
Setze $u^n(t)$ *in (2.33) ein und bestimme für jedes* $i = 1, \ldots, m$

$$w_i^n(t) := \partial_{\theta_i} u(t; \boldsymbol{\theta}^n) \in V,$$

so dass (2.33) - (2.34) für alle $\varphi \in V$ *erfüllt ist.*

(S.3) *Berechnung der Suchrichtung:*
Bestimme den Gradienten

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}^n) = \int_T [\mathcal{P}_2 \mathbf{w}^n(t)]^\top (\tilde{\mathbf{z}}(t) - \mathcal{P} u^n(t))\, dt$$

und setze $\mathbf{s}^n := -\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}^n)$.

(S.4) *Abbruchbedingung:*
Ist $\|\mathbf{s}^n\|_2 < \varepsilon$, *dann ist* $\boldsymbol{\theta}^n$ *ein stationärer Punkt von* $J(\boldsymbol{\theta})$. *Stop.*

(S.5) *Schrittweitenbestimmung (Armijo) und setzen der neuen Iterierten:*
Bestimme das kleinste $k \in \mathbb{N}_0$, *so dass*

$$J(\boldsymbol{\theta}^n + \alpha^k \mathbf{s}^n) \leq J(\boldsymbol{\theta}^n) - \beta \alpha^k \|\mathbf{s}^n\|_2^2.$$

Setze $\boldsymbol{\theta}^{n+1} := \boldsymbol{\theta}^n + \alpha^k \mathbf{s}^n$, $n := n + 1$ *und gehe zu (S.1).*

In Schritt (S.5) erzwingt der Summand $-\beta \alpha^k \|\mathbf{s}^n\|_2^2 < 0$ mit $\beta > 0$ und $\alpha^k > 0$ eine echte Verkleinerung von J. Es wird bewirkt, dass die Minimierung des Zielfunktionals im n-ten Iterationsschritt nicht zu klein ausfällt.

Wie in [43] beschrieben, fehlen Aussagen über die Konvergenzgeschwindigkeit der $J(\boldsymbol{\theta}^n)$ und der $\boldsymbol{\theta}^n$. Zudem kommt es bei der *Methode des steilsten Abstiegs* vor, dass der Gradient $\mathbf{s} = -\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$ und eine ideale Suchrichtung im Punkt $\boldsymbol{\theta}$ in der Nähe eines Minimums einen Winkel von nahezu 90° einschließen, so dass man sich in Richtung $\mathbf{s}$ nur sehr geringfügig dem gesuchten Minimum nähert. Diese Eigenschaft wird in Kapitel 6 bestätigt, wo numerisch die *Methode des steilsten Abstiegs* für ein Parameteridentifikationsproblem in der präparativen Säulenchromatographie umgesetzt wird: Sobald sich $\boldsymbol{\theta}^n$ in einer lokalen Umgebung des gesuchten Minimums befindet, müssen die Schrittweiten sehr klein gewählt werden um eine Reduktion im Zielfunktional zu erreichen.

Es stellt sich heraus, dass sich Algorithmus 2.3.1 gut eignet um in eine Nähe eines Minimums zu gelangen, für ein lokales Lösen allerdings zu viele Iterationsschritte benötigt werden. Um lokal einen Schätzer des Optimierungsproblems (2.14) numerisch zu ermitteln, bietet sich das Newton- und das Gauss-Newton-Verfahren an.

2.3.2 Newton-Verfahren

Befindet sich die Minimallösung $\hat{\boldsymbol{\theta}}$ des Optimierungsproblems (2.14) im strikten Inneren der Menge Θ, dann ist

$$\nabla_{\boldsymbol{\theta}} J(\hat{\boldsymbol{\theta}}) \;=\; 0. \tag{2.44}$$

In diesem Fall kann (2.44) als ein Nullstellenproblem aufgefasst werden, welches mit dem Newton-Verfahren [43] gelöst werden kann. Folgender Algorithmus beschreibt die Umsetzung dieser Methode um einen unbekannten Modellparameter durch Lösen des Optimierungsproblems (2.14) zu ermitteln.

Algorithmus 2.3.2. (Newton-Verfahren, [43])

(S.0) *Initialisierung:*
Wähle ein $\boldsymbol{\theta}^0 \in \Theta$, $0 < \varepsilon \ll 1$ *und setze* $n := 0$.
Wähle ein $\alpha \in (0,1)$ *und* $\beta \in (0, 0.5)$ *für die Schrittweitenbestimmung.*

(S.1) *Lösen der Zustandsgleichung:*
Bestimme $u^n(t) \;:=\; u(t;\boldsymbol{\theta}^n) \in V$, *so dass (2.24) - (2.25) für alle* $\varphi \in V$ *erfüllt ist.*

(S.2) *Lösen der Sensitivitätsgleichungen erster Ordnung:*
Setze $u^n(t)$ *in (2.33) ein und bestimme für jedes* $i = 1, \ldots, m$

$$w_i^n(t) \;:=\; \partial_{\theta_i} u(t;\boldsymbol{\theta}^n) \in V,$$

so dass (2.33) - (2.34) für alle $\varphi \in V$ *erfüllt ist.*

(S.3) *Lösen der Sensitivitätsgleichungen zweiter Ordnung:*
Setze $u^n(t)$ *und* $w_i^n(t)$ *für* $i = 1, \ldots, m$ *in (2.33) ein.*
Bestimme für jedes $i = 1, \ldots, m$ *und* $j = 1, \ldots, m$

$$v_{i,j}^n(t) \;:=\; \partial_{\theta_j} w_i(t;\boldsymbol{\theta}^n) \in V,$$

so dass (2.40) - (2.41) für alle $\varphi \in V$ *erfüllt ist.*

(S.4) *Berechnung der Suchrichtung:*
Bestimme den Gradienten $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}^n)$ *mit (2.27) und die Hessematrix* $\nabla^2_{\boldsymbol{\theta}} J(\boldsymbol{\theta}^n)$ *mit (2.36).*

(S.5) *Abbruchbedingung:*
Ist $\|\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}^n)\|_2 < \varepsilon$*:*
Ist $\nabla^2_{\boldsymbol{\theta}} J(\boldsymbol{\theta}^n)$ *positiv definit, dann ist* $\boldsymbol{\theta}^n$ *ein lokales Minimum. Stop.*
Wähle sonst ein neues $\boldsymbol{\theta}^0 \in \Theta$, *setze* $n := 0$ *und gehe zu (S.1).*
Sonst: Bestimme die Richtung $\mathbf{s}^n$ *als Lösung des Systems*

$$\nabla^2_{\boldsymbol{\theta}} J(\boldsymbol{\theta}^n)\mathbf{s}^n \;:=\; -\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}^n).$$

(S.6) *Schrittweitenbestimmung (Armijo) und setzen der neuen Iterierten:*
Bestimme das kleinste $k \in \mathbb{N}_0$, *so dass*

$$J(\boldsymbol{\theta}^n + \alpha^k \mathbf{s}^n) \;\leq\; J(\boldsymbol{\theta}^n) - \beta\alpha^k \|\mathbf{s}^n\|_2^2.$$

Setze $\boldsymbol{\theta}^{n+1} := \boldsymbol{\theta}^n + \alpha^k \mathbf{s}^n$, $n := n+1$ *und gehe zu (S.1).*

Bei der Realisierung von Algorithmus 2.3.2 muss beachtet werden, dass das Newton-Verfahren gegen ein Maximum oder einen Sattelpunkt $\hat{\boldsymbol{\theta}}$ des Optimierungsproblems (2.14) konvergieren kann [43]. Die Richtung

$$\mathbf{s}^n \;:=\; -\left(\nabla^2_{\boldsymbol{\theta}} J(\boldsymbol{\theta}^n)\right)^{-1} \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}^n)$$

stellt allerdings immer eine geeignete Abstiegsrichtung dar, wenn die Hessematrix

$$H(\boldsymbol{\theta}^n) \;:=\; \nabla^2_{\boldsymbol{\theta}} J(\boldsymbol{\theta}^n)$$

positiv definit ist [43]. Daher kann das Parameterschätzproblem (2.14) numerisch gelöst werden, indem zunächst der Algorithmus 2.3.1 (*Methode des steilsten Abstiegs*) umgesetzt wird, bis die Hessematrix $H(\boldsymbol{\theta}^n)$ positiv definit ist. Da $\boldsymbol{\theta}^n$ sich nun in einer konvexen Umgebung einer lokalen Minimallösung $\hat{\boldsymbol{\theta}}$ von (2.14) befindet, konvergiert das in Algorithmus 2.3.2 dargestellte Newton-Verfahren gegen $\hat{\boldsymbol{\theta}}$.

Bemerkung 2.3.1. *Wie in [43] beschrieben, ist im allgemeinen die in Schritt (S.6) beschriebene Schrittweitensteuerung notwendig um die globale Konvergenz des Verfahrens zu gewährleisten, auch wenn die Hessematrix $H(\boldsymbol{\theta}^n)$ positiv definit ist.*

2.3.3 Gauss-Newton-Verfahren

Es wird nun beschrieben, wie mit dem Gauss-Newton-Verfahren das Parameterschätzproblem

$$\min_{\Theta} J(\boldsymbol{\theta}) \quad \overset{(2.15)}{=} \quad \frac{1}{2}\int_T \|\tilde{\mathbf{z}}(t) - \mathcal{P}u(t,\mathbf{x},\boldsymbol{\theta})\|_2^2 \, dt \tag{2.45}$$

gelöst werden kann. Hierfür wird der folgende *Satz von Taylor* benötigt:

Satz 2.3.2. (Satz von Taylor im $\mathbb{R}^n$, [43])
Sei $\mathcal{M} \subset \mathbb{R}^n$ offen und $g : \mathcal{M} \to \mathbb{R}^k$ zweimal stetig differenzierbar. Seien ferner $\bar{\boldsymbol{\theta}} \in \mathcal{M}$ und $\delta > 0$ gegeben mit $\{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\|_\infty \le \delta\} \subset \mathcal{M}$. Dann gibt es ein $C = C(\delta) > 0$, so dass für alle $\mathbf{h} \in \mathcal{M}$ mit $\|\mathbf{h}\|_\infty < \delta$ die Abschätzung

$$g(\bar{\boldsymbol{\theta}} + \mathbf{h}) \;=\; g(\bar{\boldsymbol{\theta}}) + \nabla_{\boldsymbol{\theta}} g(\bar{\boldsymbol{\theta}})^\top \mathbf{h} + R(\mathbf{h}) \quad \textit{mit} \quad \|R(\mathbf{h})\|_\infty \le C\|\mathbf{h}\|_\infty^2$$

gilt.

Wie in [43] erläutert, wird beim Gauss-Newton-Verfahren zunächst die Funktion $\tilde{\mathbf{z}}(t) - \mathcal{P}u(t,\mathbf{x},\boldsymbol{\theta})$ in (2.45) durch die zugehörige Linearisierung approximiert. Da nach Satz 2.3.2 ein $C > 0$ existiert, so dass

$$\tilde{\mathbf{z}}(t) - \mathcal{P}u(t,\mathbf{x},\boldsymbol{\theta}) \;=\; \tilde{\mathbf{z}}(t) \;-\; \mathcal{P}u(t,\mathbf{x},\boldsymbol{\theta}+\mathbf{h}) \;+\; \left[\mathcal{P}_2 \nabla_{\boldsymbol{\theta}} u(t,\mathbf{x},\boldsymbol{\theta}+\mathbf{h})\right]^\top \mathbf{h} \;+\; R(\mathbf{h})$$

mit $\|R(\mathbf{h})\|_\infty \le C\|\mathbf{h}\|_\infty^2$ gilt, kann für jedes $\mathbf{h} \in \Theta$ mit $\|\mathbf{h}\|_\infty^2 \le \delta$

$$\tilde{\mathbf{z}}(t) - \mathcal{P}u(t,\mathbf{x},\boldsymbol{\theta}) \;:=\; \tilde{\mathbf{z}}(t) \;-\; \mathcal{P}u(t,\mathbf{x},\boldsymbol{\theta}+\mathbf{h}) \;+\; \left[\mathcal{P}_2 \nabla_{\boldsymbol{\theta}} u(t,\mathbf{x},\boldsymbol{\theta}+\mathbf{h})\right]^\top \mathbf{h},$$

gesetzt werden, wenn $\delta > 0$ hinreichend klein gewählt ist. Dadurch erhält man für fest gewähltes $\boldsymbol{\theta}^n \in \Theta$ mit $\|\boldsymbol{\theta}^n - \boldsymbol{\theta}^*\| \leq \delta$ das Schätzproblem

$$\min_{\boldsymbol{\theta}\in\Theta} \tilde{J}(\boldsymbol{\theta}, \boldsymbol{\theta}^n) \;=\; \frac{1}{2}\int_T \|\tilde{\mathbf{z}}(t) - \mathcal{P}u(t,\mathbf{x},\boldsymbol{\theta}^n) + \left[\mathcal{P}_2\nabla_{\boldsymbol{\theta}}u(t,\mathbf{x},\boldsymbol{\theta}^n)\right]^\top(\boldsymbol{\theta}-\boldsymbol{\theta}^n)\|_2^2\, dt, \quad (2.46)$$

wobei $\boldsymbol{\theta}^* \in \Theta$ eine optimale Lösung von (2.45) darstellt.

Das Gauss-Newton-Verfahren löst (2.45) durch Anwendung der Newton-Methode auf das linearisierte Schätzproblem (2.46). Mit dem Gradienten

$$\nabla_{\boldsymbol{\theta}}\tilde{J}(\boldsymbol{\theta}, \boldsymbol{\theta}^n) \;=\; \int_T \mathcal{P}_2\nabla_{\boldsymbol{\theta}}u(t,\mathbf{x},\boldsymbol{\theta}^n)\Big(\tilde{\mathbf{z}}(t) - \mathcal{P}u(t,\mathbf{x},\boldsymbol{\theta}^n) + \left[\mathcal{P}_2\nabla_{\boldsymbol{\theta}}u(t,\mathbf{x},\boldsymbol{\theta}^n)\right]^\top(\boldsymbol{\theta}-\boldsymbol{\theta}^n)\Big)\, dt$$

und der Hessematrix

$$\nabla^2_{\boldsymbol{\theta}}\tilde{J}(\boldsymbol{\theta}, \boldsymbol{\theta}^n) \;=\; \int_T \left[\mathcal{P}_2\nabla_{\boldsymbol{\theta}}u(t,\mathbf{x},\boldsymbol{\theta}^n)\right]\left[\mathcal{P}_2\nabla_{\boldsymbol{\theta}}u(t,\mathbf{x},\boldsymbol{\theta}^n)\right]^\top dt \quad (2.47)$$

kann das Gauss-Newton-Verfahren wie folgt umgesetzt werden.

Algorithmus 2.3.3. (Gauss-Newtonverfahren, [43])

(S.0) *Initialisierung:*
Wähle ein $\boldsymbol{\theta}^0 \in \Theta$, $\varepsilon > 0$ und setze $n := 0$.
Wähle ein $\alpha \in (0,1)$ und $\beta \in (0, 0.5)$ für die Schrittweitenbestimmung.

(S.1) *Lösen der Zustandsgleichung:*
Bestimme
$$u^n(t) \;:=\; u(t;\boldsymbol{\theta}^n) \in V,$$
so dass (2.24) - (2.25) für alle $\varphi \in V$ erfüllt ist.

(S.2) *Lösen der Sensitivitätsgleichungen erster Ordnung:*
Setze $u^n(t)$ in (2.33) ein und bestimme für jedes $i = 1, \ldots, m$
$$w_i^n(t) \;:=\; \partial_{\theta_i}u(t;\boldsymbol{\theta}^n) \in V,$$
so dass (2.33) - (2.34) für alle $\varphi \in V$ erfüllt ist.

(S.3) *Berechnung des Gradienten und der Fisher-Informationsmatrix:*
Bestimme den Gradienten
$$\nabla_{\boldsymbol{\theta}}J(\boldsymbol{\theta}^n, \boldsymbol{\theta}^n) \;=\; \int_T \left[\mathcal{P}_2\mathbf{w}^n(t)\right]^\top(\tilde{\mathbf{z}}(t) - \mathcal{P}u^n(t))\, dt$$
und die modifizierte Hessematrix
$$\nabla^2_{\boldsymbol{\theta}}\tilde{J}(\boldsymbol{\theta}^n, \boldsymbol{\theta}^n) \;=\; \int_T \left[\mathcal{P}_2\mathbf{w}^n(t)\right]\left[\mathcal{P}_2\mathbf{w}^n(t)\right]^\top dt.$$

(S.4) *Berechnung der Suchrichtung:*
Bestimme die Richtung $\mathbf{s}^n$ als Lösung des Systems
$$\nabla^2_{\boldsymbol{\theta}}\tilde{J}(\boldsymbol{\theta}^n, \boldsymbol{\theta}^n)\mathbf{s}^n \;:=\; -\nabla_{\boldsymbol{\theta}}J(\boldsymbol{\theta}^n, \boldsymbol{\theta}^n).$$

(S.5) *Schrittweitenbestimmung (Armijo) und setzen der neuen Iterierten:*
Bestimme das kleinste $k \in \mathbb{N}_0$, *so dass*

$$J(\boldsymbol{\theta}^n + \alpha^k \mathbf{s}^n, \boldsymbol{\theta}^n) \;\leq\; J(\boldsymbol{\theta}^n, \boldsymbol{\theta}^n) - \beta\alpha^k \|\mathbf{s}^n\|_2^2.$$

Setze $\boldsymbol{\theta}^{n+1} := \boldsymbol{\theta}^n + \alpha^k \mathbf{s}^n$, $n := n+1$ *und gehe zu (S.1).*

Bemerkung 2.3.3. *Im folgenden Kapitel 3 wird beschrieben, dass die Kovarianzmatrix eines Schätzers* $\hat{\boldsymbol{\theta}}$ *durch die Inverse der sogenannten Fisher-Informationsmatrix* M *approximiert werden kann. An dieser Stelle sei zu erwähnen, dass für die modifizierte Hessematrix* $\nabla^2_{\boldsymbol{\theta}} \tilde{J}(\boldsymbol{\theta})$ *aus Schritt (S.3)*

$$M \;=\; \nabla^2_{\boldsymbol{\theta}} \tilde{J}(\boldsymbol{\theta}, \boldsymbol{\theta}^n)$$

gilt.

Kapitel 3

Optimale Versuchsplanung bei instationären partiellen Differentialgleichungen

In Kapitel 2 wurde beschrieben, wie man mit einer Parameteridentifikationsmethode einen unbekannten Parameter $\boldsymbol{\theta} \in \Theta$ durch einen Schätzer $\hat{\boldsymbol{\theta}} := \hat{\boldsymbol{\theta}}(\boldsymbol{\theta})$ numerisch bestimmen kann. Hierfür wurde vorausgesetzt, dass $\hat{\boldsymbol{\theta}}$ erwartungstreu ist, d. h. es gilt

$$E\{\hat{\boldsymbol{\theta}}\} = \boldsymbol{\theta}.$$

In diesem Kapitel wird nun untersucht, wie gut der Schätzer $\hat{\boldsymbol{\theta}}$ den unbekannten Parameter approximiert. Ein Maß für die Genauigkeit eines Schätzers liefert dessen Kovarianzmatrix $\mathrm{cov}\{\hat{\boldsymbol{\theta}}\}$ (siehe z. B. [85]): Ein Parameter $\boldsymbol{\theta}$ kann durch $\hat{\boldsymbol{\theta}}$ umso genauer approximiert werden, wenn die zugehörige Kovarianzmatrix $\mathrm{cov}\{\hat{\boldsymbol{\theta}}\}$ minimal bezüglich eines Optimalitätskriteriums $G : \mathbb{R}^{m \times m} \to \mathbb{R}$ ist. Bewährt haben sich die Optimalitätskriterien:

(1) **D-Optimalitätskriterium** (Determinante) mit (3.1)
$$G(\mathrm{cov}\{\hat{\boldsymbol{\theta}}\}) = \ln \det(\mathrm{cov}\{\hat{\boldsymbol{\theta}}\}) := \ln |\mathrm{cov}\{\hat{\boldsymbol{\theta}}\}|,$$

(2) **E-Optimalitätskriterium** (größter Eigenwert von $\mathrm{cov}\{\hat{\boldsymbol{\theta}}\}$) mit (3.2)
$$G(\mathrm{cov}\{\hat{\boldsymbol{\theta}}\}) = \lambda_{\max}(\mathrm{cov}\{\hat{\boldsymbol{\theta}}\}),$$

(3) **A-Optimalitätskriterium** (Spur) mit (3.3)
$$G(\mathrm{cov}\{\hat{\boldsymbol{\theta}}\}) = \mathrm{tr}(\mathrm{cov}\{\hat{\boldsymbol{\theta}}\}).$$

Wie in Kapitel 2.1.2 beschrieben, ist die Minimierung von $\ln |\mathrm{cov}\{\hat{\boldsymbol{\theta}}\}|$ äquivalent zur Minimierung des Volumens des *Konfidenzellipsoids* $\mathcal{K}$ des Schätzers $\hat{\boldsymbol{\theta}}$. Ein Konfidenzellipsoid ist eine Menge die beschreibt, in welchen Bereichen der unbekannte Parameter $\boldsymbol{\theta}$ mit „großer“ Wahrscheinlichkeit liegt. Je kleiner diese Bereiche sind, desto genauer kann $\boldsymbol{\theta}$ bestimmt werden. Was das E-Optimalitätskriterium anbelangt, bewirkt eine Minimierung von $\lambda_{\max}(\mathrm{cov}\{\hat{\boldsymbol{\theta}}\})$ eine Minimierung der längsten Achse des Konfidenzellipsoids. Eine Minimierung mittels A-Optimalitätskriteriums minimiert den Durchschnitt der Varianzen des Schätzers $\hat{\boldsymbol{\theta}}$.

In dieser Arbeit werden wir uns auf das D-Optimalitätskriterium (3.1) konzentrieren. Das Ziel ist somit die Minimierung des Volumens des zu $\hat{\boldsymbol{\theta}}$ zugehörigen Konfidenzellipsoids. Zur Realisierung dessen wird die sogenannte *Fisher-Informationsmatrix* $M(\vec{\mathbf{x}})$ verwendet, die unter der Voraussetzung 2.2.1 die Inverse der Kovarianzmatrix $\mathrm{cov}\{\hat{\boldsymbol{\theta}}\}$ approximiert [33]. In diesem Kapitel gelte die Voraussetzung 2.2.1, so dass

$$M(\vec{\mathbf{x}}) \;\approx\; \mathrm{cov}\{\hat{\boldsymbol{\theta}}\}^{-1}$$

gilt.

Folglich ist eine Minimierung von $G(\mathrm{cov}\{\hat{\boldsymbol{\theta}}\}) = \ln|\mathrm{cov}\{\hat{\boldsymbol{\theta}}\}|$ äquivalent zur Maximierung von $G(M(\vec{\mathbf{x}})) = \ln|M(\vec{\mathbf{x}})|$. Wie in [33] beschrieben, ist $M(\vec{\mathbf{x}})$ generell einfacher zu bestimmen als die Kovarianzmatrix von $\hat{\boldsymbol{\theta}}$. Daher wird in dieser Arbeit die Fisher-Informationsmatrix verwendet und es wird gezeigt, wie diese mithilfe der Sensitivitäten erster Ordnung (2.29) berechnet werden kann.

Der Fokus dieser Arbeit liegt insbesondere auf der Bestimmung einer für die Parameteridentifikation benötigten optimalen Messstellenkonstellation

$$\vec{\mathbf{x}}^* = ({\mathbf{x}^{1,*}}^\top, \ldots, {\mathbf{x}^{\ell,*}}^\top)^\top,$$

unter der die Fisher-Informationsmatrix $M(\vec{\mathbf{x}}^*)$ maximal bezüglich (3.1) ist. Eine solche Konstellation kann durch Lösen eines nichtlinearen Optimierungsproblems ermittelt werden. Darius Ucinski [85] hat bereits 2005 eine Lösungsmethode beschrieben, mit der man

$$\vec{\mathbf{x}}^* = \arg\max\{\ln|M(\vec{\mathbf{x}})|\}$$

ermitteln kann. Diese Methode wird in diesem Kapitel vorgestellt. Die hierfür benötigten mathematischen Hintergründe sind weitestgehend dem Buch [85] entnommen.

Bevor in diesem Kapitel eine Lösungsmethode zur Bestimmung einer optimalen Messstellenkonstellation $\vec{\mathbf{x}}^*$ beschrieben wird, wird zunächst im folgenden Abschnitt auf die Fisher-Informationsmatrix $M(\vec{\mathbf{x}})$ eingegangen.

3.1 Fisher-Informationsmatrix

Wie in [85] beschrieben, ist die Fisher-Informationsmatrix zum Parameteridentifikationsproblem (2.14) eine Abbildung $M : \bar{\Omega}^\ell \to \mathbb{R}^{m\times m}$ definiert durch

$$M(\vec{\mathbf{x}}) \;=\; \frac{1}{\sigma^2}\sum_{i=1}^{\ell}\int_T \nabla_{\boldsymbol{\theta}} u(t,\mathbf{x}^i,\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}} u(t,\mathbf{x}^i,\boldsymbol{\theta})^\top dt \tag{3.4}$$

mit den Messstellen $\vec{\mathbf{x}} = ({\mathbf{x}^1}^\top, \ldots, {\mathbf{x}^\ell}^\top)^\top \in \bar{\Omega}^\ell$, wobei für jedes fest gewählte $\boldsymbol{\theta} \in \Theta$

$$u(t,\mathbf{x},\boldsymbol{\theta}) \;\in\; \mathcal{C}^1(T;V)$$

eine Lösung der Zustandsgleichung (2.24) - (2.25) und

$$\nabla_{\boldsymbol{\theta}} u(t,\mathbf{x},\boldsymbol{\theta}) \;\in\; \mathcal{C}^1(T;V)^m$$

die zugehörige Sensitivität erster Ordnung als Lösung von (2.33) - (2.34) darstellt. Hierfür wurde vorausgesetzt, dass für den Messfehler $\varepsilon(t, \mathbf{x})$ bei der Parameteridentifikation

$$\varepsilon(t, \mathbf{x}) \sim \mathcal{N}(0, \sigma^2)$$

gilt mit

$$E\{\varepsilon(t, \mathbf{x})\} = 0, \quad E\{\varepsilon(t, \mathbf{x}^i)\varepsilon(t', \mathbf{x}^j)\} = \sigma^2 \delta_{ij} \delta(t - t'),$$

wobei δ_{ij} die Kronecker-Delta- und δ die Dirac-Delta-Funktion darstellt (Voraussetzung 2.2.1). Die Standardabweichung wird mit σ bezeichnet.

Wie in [33] beschrieben, gilt unter der Voraussetzung 2.2.1 die *Cramér-Rao Schranke*

$$\text{cov}\{\hat{\boldsymbol{\theta}}\} \ \succeq \ M(\vec{\mathbf{x}})^{-1}, \tag{3.5}$$

mit der Kovarianzmatrix des Schätzers $\hat{\boldsymbol{\theta}}$

$$\text{cov}\{\hat{\boldsymbol{\theta}}\} \ = \ E\{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top\}$$

und der Fisher-Informationsmatrix $M(\vec{\mathbf{x}})$ aus (3.4). Der gesuchte, unbekannte Modellparameter wird mit $\boldsymbol{\theta}^*$ bezeichnet. Die Löwner-Halbordnung [73] $\succeq$ für symmetrische Matrizen aus (3.5) ist wie folgt definiert.

Definition 3.1.1. (Löwner-Halbordnung für symmetrische Matrizen)
Sei $n \in \mathbb{N}$. Dann gilt für die symmetrischen Matrizen $A \in \mathbb{R}^{n \times n}$ und $B \in \mathbb{R}^{n \times n}$

$$A \ \succeq \ B,$$

wenn $\mathbf{d}^\top A \mathbf{d} \geq \mathbf{d}^\top B \mathbf{d}$ für alle $\mathbf{d} \in \mathbb{R}^n$ gilt. In diesem Fall sind alle Eigenwerte von $A - B$ nicht negativ.

Unter gewissen Voraussetzungen gilt sogar Gleichheit, so dass

$$\text{cov}\{\hat{\boldsymbol{\theta}}\} \ = \ M(\vec{\mathbf{x}})^{-1}. \tag{3.6}$$

ist. Dieser Fall tritt zum Beispiel ein, wenn die Sensitivitäten $\nabla_{\boldsymbol{\theta}} u(t, \mathbf{x}, \boldsymbol{\theta})$ linear vom Parameter $\boldsymbol{\theta}$ abhängen. Dann existiert eine Funktion $\mathbf{y}(t, \mathbf{x}) \in \mathcal{C}^1(T; V)^m$, so dass

$$u(t, \mathbf{x}, \boldsymbol{\theta}) \ = \ \mathbf{y}(t, \mathbf{x})^\top \boldsymbol{\theta}$$

und folglich

$$\nabla_{\boldsymbol{\theta}} u(t, \mathbf{x}, \boldsymbol{\theta}) \ = \ \mathbf{y}(t, \mathbf{x})$$

gilt. Für diesen Fall wird nun gezeigt, dass der Maximum-Likelihood-Schätzer aus (2.14) erwartungstreu ist und (3.6) gilt.

Wie in Kapitel 2.2.1 beschrieben, kann ein unbekannter Modellparameter $\boldsymbol{\theta}^*$ durch einen Schätzer $\hat{\boldsymbol{\theta}}$ durch Lösen des Optimierungsproblems

$$\hat{\boldsymbol{\theta}} \ = \ \arg\min_{\Theta} J(\boldsymbol{\theta}) \ = \ \arg\min_{\Theta} \frac{1}{2\sigma^2} \int_T \|\tilde{\mathbf{z}}(t) - \mathcal{P} u(t, \mathbf{x}, \boldsymbol{\theta})\|_2^2 dt, \tag{3.7}$$

mit dem Operator $\mathcal{P}$ aus Definition (2.2.1) approximiert werden, wobei

$$\tilde{\mathbf{z}}(t) = (\tilde{z}_1(t), \ldots, \tilde{z}_\ell(t))^\top$$

den Vektor der gemittelten Messreihen darstellt und $\tilde{z}_i(t)$ für $i \in I_\ell$ die gemittelten Messreihen sind, die an der Messposition $\mathbf{x}^i \in \bar{\Omega}$ ermittelt wurden. Der Parameterraum Θ aus (2.4) sei so gewählt, dass sowohl $\boldsymbol{\theta}^*$ als auch $\hat{\boldsymbol{\theta}}$ im strikten Inneren von Θ liegen, d. h. es gilt $a_i < \theta_i^* < b_i$ und $a_i < \hat{\theta}_i < b_i$ für alle $i = 1, \ldots, m$. Dann gilt mit dem euklidischen Skalarprodukt $(\cdot,\cdot)_2$ für den Schätzer $\hat{\boldsymbol{\theta}}$ aus (3.7) die Optimalitätsbedingung erster Ordnung

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}} J(\hat{\boldsymbol{\theta}}) &= -\frac{1}{\sigma^2}\int_T [\mathcal{P}_2 \mathbf{y}(t,\mathbf{x})]\,(\tilde{\mathbf{z}}(t) - \mathcal{P}u(t,\mathbf{x},\hat{\boldsymbol{\theta}}))\,dt \\
&= -\frac{1}{\sigma^2}\sum_{i=1}^{\ell}\int_T \mathbf{y}(t,\mathbf{x}^i)(\tilde{z}_i(t) - u(t,\mathbf{x}^i,\hat{\boldsymbol{\theta}}))\,dt \\
&= -\frac{1}{\sigma^2}\sum_{i=1}^{\ell}\int_T \mathbf{y}(t,\mathbf{x}^i)(\tilde{z}_i(t) - \mathbf{y}(t,\mathbf{x}^i)^\top\hat{\boldsymbol{\theta}})\,dt \;=\; 0,
\end{aligned}$$

mit dem Operator $\mathcal{P}_2$ aus Definition (2.2.7) und folglich

$$M(\vec{\mathbf{x}})\,\hat{\boldsymbol{\theta}} \;=\; \frac{1}{\sigma^2}\sum_{i=1}^{\ell}\int_T \mathbf{y}(t,\mathbf{x}^i)\mathbf{y}(t,\mathbf{x}^i)^\top dt\,\hat{\boldsymbol{\theta}} \;=\; \frac{1}{\sigma^2}\sum_{i=1}^{\ell}\int_T \mathbf{y}(t,\mathbf{x}^i)\tilde{z}_i(t)\,dt.$$

Ist die Fisher-Informationsmatrix regulär, dann gilt für den Maximum-Likelihood-Schätzer

$$\hat{\boldsymbol{\theta}} \;=\; M(\vec{\mathbf{x}})^{-1}\sum_{i=1}^{\ell}\int_T \mathbf{y}(t,\mathbf{x}^i)\tilde{z}_i(t)\,dt.$$

Dieser Schätzer ist erwartungstreu, da für seinen Erwartungswert

$$\begin{aligned}
E\{\hat{\boldsymbol{\theta}}\} &= M(\vec{\mathbf{x}})^{-1}\sum_{i=1}^{\ell}\int_T \mathbf{y}(t,\mathbf{x}^i)E\{\tilde{z}_i(t)\}\,dt \;=\; M(\vec{\mathbf{x}})^{-1}\sum_{i=1}^{\ell}\int_T \mathbf{y}(t,\mathbf{x}^i)\mathbf{y}(t,\mathbf{x}^i)^\top\hat{\boldsymbol{\theta}}\,dt \\
&= M(\vec{\mathbf{x}})^{-1}M(\vec{\mathbf{x}})\,\hat{\boldsymbol{\theta}} \;=\; \hat{\boldsymbol{\theta}}
\end{aligned}$$

gilt. Zudem ist die Kovarianzmatrix $\mathrm{cov}\{\hat{\boldsymbol{\theta}}\}$ gleich der Inversen der Fisher-Informationsmatrix. Es gilt

$$\begin{aligned}
\mathrm{cov}\{\hat{\boldsymbol{\theta}}\} &= E\{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top\} \\
&= M(\vec{\mathbf{x}})^{-1}\Big(\frac{1}{\sigma^4}\sum_{i=1}^{\ell}\int_T\int_T \mathbf{y}(t,\mathbf{x}^i)^\top E\big\{[\mathcal{P}\varepsilon(t)]\,[\mathcal{P}\varepsilon(\tau)]^\top\big\}\mathbf{y}(\tau,\mathbf{x}^i)\,dtd\tau\Big)M(\vec{\mathbf{x}})^{-1} \\
&= M(\vec{\mathbf{x}})^{-1}M(\vec{\mathbf{x}})M(\vec{\mathbf{x}})^{-1} \;=\; M(\vec{\mathbf{x}})^{-1}.
\end{aligned}$$

Da im allgemeinen allerdings die Lösung $u(t,\mathbf{x},\boldsymbol{\theta})$ der Zustandsgleichung (2.24) - (2.25) nicht linear von $\boldsymbol{\theta}$ abhängt, wird in dieser Arbeit folgendes vorausgesetzt:

Voraussetzung 3.1.1. *Es ist bereits eine Näherung $\hat{\boldsymbol{\theta}}^1 \in \Theta$ des unbekannten Parameters $\boldsymbol{\theta}^* \in \Theta$ bekannt mit hinreichend kleiner Differenz $\|\hat{\boldsymbol{\theta}}^1 - \boldsymbol{\theta}^*\|_\infty^2$.*

Unter der Voraussetzung 3.1.1 und mit Satz 2.3.2 existiert ein $C > 0$, so dass

$$u(t,\mathbf{x},\boldsymbol{\theta}^*) \;=\; u(t,\mathbf{x},\hat{\boldsymbol{\theta}}^1) + \nabla_{\boldsymbol{\theta}}u(t,\mathbf{x},\hat{\boldsymbol{\theta}}^1)^\top(\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}^1) + R(\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}^1),$$

mit $\|R(\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}^1)\|_\infty \leq C\|\hat{\boldsymbol{\theta}}^1 - \boldsymbol{\theta}^*\|_\infty^2$. Da Voraussetzung 3.1.1 erfüllt ist, ist folglich $R(\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}^1) \approx 0$, so dass

$$u(t, \mathbf{x}, \boldsymbol{\theta}^*) \;\approx\; u(t, \mathbf{x}, \hat{\boldsymbol{\theta}}^1) + \nabla_{\boldsymbol{\theta}} u(t, \mathbf{x}, \hat{\boldsymbol{\theta}}^1)^\top (\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}^1)$$

gilt und somit eine Funktion $y(t, \mathbf{x}) \in \mathcal{C}^1(T; V)^m$ existiert mit $\nabla_{\boldsymbol{\theta}} u(t, \mathbf{x}, \hat{\boldsymbol{\theta}}) \approx y(t, \mathbf{x})\hat{\boldsymbol{\theta}}$. Dann gilt

$$\mathrm{cov}\{\hat{\boldsymbol{\theta}}\} \;\approx\; M(\vec{\mathbf{x}})^{-1}. \tag{3.8}$$

Bemerkung 3.1.1.

- *Die Zuordnung „≈“ zwischen zwei Funktionen $g : \mathbb{R}^m \to \mathbb{R}$ und $f : \mathbb{R}^m \to \mathbb{R}$*

  $$g(\mathbf{x}) \;\approx\; f(\mathbf{x})$$

  bedeutet in dieser Arbeit, dass die Differenz zwischen beiden Funktionen vernachlässigbar klein ist und $g(\mathbf{x})$ durch $f(\mathbf{x})$ approximiert werden kann.

- *Die Zuordnung „≈“ zwischen zwei Matrizen $A \in \mathbb{R}^{m \times m}$ und $B \in \mathbb{R}^{m \times m}$*

  $$A \;\approx\; B$$

  bedeutet in dieser Arbeit, dass $A_{ij} \approx B_{ij}$ für alle $i, j \in \{1, \ldots, m\}$ gilt.

Bemerkung 3.1.2. *Wie man eine gute Näherung $\hat{\boldsymbol{\theta}}^1 \in \Theta$ von $\boldsymbol{\theta}^* \in \Theta$ bestimmen kann, wird im folgenden Kapitel 3.2 erläutert.*

Um eine optimale Messstellenkonstellation $\vec{\mathbf{x}}^* = ({\mathbf{x}^{1,*}}^\top, \ldots, {\mathbf{x}^{\ell,*}}^\top)^\top \in \bar{\Omega}^\ell$ mit dem D-Optimalitätskriterium (3.1) zu erhalten, löst man das Optimierungsproblem

$$\vec{\mathbf{x}}^* \;=\; \arg\min\{-\ln|M(\vec{\mathbf{x}})|\} \tag{3.9}$$

mit der Fisher-Informationsmatrix

$$M(\vec{\mathbf{x}}) \;=\; \frac{1}{\sigma^2} \sum_{i=1}^{\ell} \int_T \nabla_{\boldsymbol{\theta}} u(t, \mathbf{x}^i, \hat{\boldsymbol{\theta}}^1) \nabla_{\boldsymbol{\theta}} u(t, \mathbf{x}^i, \hat{\boldsymbol{\theta}}^1)^\top dt. \tag{3.10}$$

Im Kapitel 3.3.3 wird ein Lösungsverfahren vorgestellt, mit dem das Optimierungsproblem (3.9) gelöst werden kann. Dieses ist dem Buch [85] entnommen. Zuvor wird im folgenden Abschnitt erläutert, wie man eine gute Näherung $\hat{\boldsymbol{\theta}}^1 \in \Theta$ des unbekannten Parameters $\boldsymbol{\theta}^* \in \Theta$ bestimmen kann, so dass Voraussetzung 3.1.1 gilt.

3.2 Sequentielle Versuchsplanung

In Kapitel 2.2 wurde beschrieben wie mithilfe von Messreihen ein unbekannter Modellparameter $\boldsymbol{\theta}^* \in \Theta$ geschätzt werden kann. Diese Messreihen erhält man, indem zunächst Messpositionen $\vec{\mathbf{x}} = ({\mathbf{x}^1}^\top, \ldots, {\mathbf{x}^\ell}^\top)^\top \in \bar{\Omega}^\ell$ festgelegt werden, an denen mithilfe von Sensoren experimentell die erforderlichen Daten erhoben werden. Da allerdings nicht jede Messreihe geeignet ist um einen Modellparameter $\boldsymbol{\theta}^*$ zu schätzen, sollte $\vec{\mathbf{x}}$ so gewählt werden, dass der Modellparameter $\boldsymbol{\theta}^* \in \Theta$ bestmöglich geschätzt werden kann. Dieses kann durch Lösen des Optimierungsproblems (3.9) realisiert werden. Das Problem ist allerdings:

- Um einen Modellparameter $\boldsymbol{\theta}^* \in \Theta$ mit der Maximum-Likelihood-Methode (2.14) schätzen zu können, müssen die zur Datenerhebung erforderlichen Messstellen $\vec{\mathbf{x}}$ a priori festgelegt werden.
- Um eine optimale Messstellenkonstellation $\vec{\mathbf{x}}^*$ (3.9) ermitteln zu können, wird eine gute Schätzung $\hat{\boldsymbol{\theta}}$ des Parameters $\boldsymbol{\theta}^* \in \Theta$ benötigt.

Da die Parameteridentifikationsmethode (2.14) von einer optimalen Messstellenkonstellation $\vec{\mathbf{x}}^* \in \bar{\Omega}^\ell$ abhängt, eine optimale Messstellenbestimmung allerdings nur mithilfe einer näherungsweisen, optimalen Schätzung des Parameters $\boldsymbol{\theta}^*$ realisiert werden kann, wird eine Umsetzung beider Optimierungsprobleme mittels einer äußeren Schleife vorgenommen. Dieses Vorgehen wird Sequentielle Versuchsplanung genannt und wird wie folgt realisiert.

Das Ziel ist eine optimale Schätzung eines Modellparameters $\boldsymbol{\theta}^* \in \Theta$. Um diesen zu bestimmen, wird mit $n := 0$ wie folgt vorgegangen:

(S.1) Zunächst werden $0 < \ell$ Messpositionen $\vec{\mathbf{x}}^0 = (\mathbf{x}^{1,0^\top}, \ldots, \mathbf{x}^{\ell,0^\top})^\top \in \bar{\Omega}^\ell$ festgelegt, an denen experimentell Messdaten erhoben werden.

(S.2) Auf Basis dieser Messreihen $\vec{\mathbf{x}}^n$ wird nun durch Lösen des Optimierungsproblems (3.9) eine n-te Schätzung $\boldsymbol{\theta}^n$ des Parameters $\boldsymbol{\theta}^* \in \Theta$ vorgenommen.

(S.3) Der Schätzwert $\boldsymbol{\theta}^n$ wird nun in die Fisher-Informationsmatrix $M(\vec{\mathbf{x}})$ aus (3.10) eingesetzt um anschließend eine optimale Messstellenkonstellation $\vec{\mathbf{x}}^{n+1}$ durch Lösen des Problems (3.9) zu ermitteln.

(S.4) Falls $\boldsymbol{\theta}^n$ noch keine gute Schätzung darstellt, wird $n := n + 1$ gesetzt und mit Schritt (S.2) fortgefahren.

Mithilfe einer solchen Schleife kann nach endlich vielen Schritten eine optimale Lösung $\hat{\boldsymbol{\theta}}$ des Schätzproblems (3.9) erwartet werden.

3.3 Lösungsverfahren zur Bestimmung eines D-optimalen Designs

Das Ziel ist die Minimierung der Kovarianzmatrix eines Schätzers $\hat{\boldsymbol{\theta}}$ bezüglich des D-Optimalitätskriteriums (3.1) um einen unbekannten Parameter $\boldsymbol{\theta}^*$ mit einer Parameteridentifikationsmethode so genau wie möglich approximieren zu können. In Kapitel 2 wurde bereits beschrieben, dass ein unbekannter Parameter durch Lösen eines Optimierungsproblems der Form

$$\min_{\Theta} J(\boldsymbol{\theta}) \quad = \quad \frac{1}{2}\frac{1}{\sigma^2}\int_T \|\tilde{\mathbf{z}}(t) - \mathcal{P}u(t, \mathbf{x}, \boldsymbol{\theta})\|_2^2 dt$$

approximiert werden kann, wobei $u(t, \mathbf{x}, \boldsymbol{\theta})$ eine Lösung von (2.33) - (2.34) darstellt. Es ist $\tilde{\mathbf{z}}(t) = (\tilde{z}_1(t), \ldots, \tilde{z}_\ell(t))^\top$ der Vektor der gemittelten Messreihen, wobei $\tilde{z}_i(t)$ die gemittelten Messwerte an der Messstelle $\mathbf{x}^i$ für $i \in I_\ell$ darstellen.

Unter der Voraussetzung 3.1.1 gilt (3.8), d. h.

$$\mathrm{cov}\{\hat{\boldsymbol{\theta}}\} \;\approx\; M(\vec{\mathbf{x}})^{-1}.$$

Infolgedessen kann die Kovarianzmatrix eines Schätzers $\hat{\boldsymbol{\theta}}$ bezüglich (3.1) minimiert werden, indem die Fisher-Informationsmatrix $M(\vec{\mathbf{x}})$ bezüglich (3.1) maximiert wird. In dieser Arbeit wird die Fisher-Informationsmatrix durch Bestimmung einer optimalen Messstellenkonfiguration $\vec{\mathbf{x}}^*$ „maximiert". Eine derartige Konfiguration $\vec{\mathbf{x}}^* \in \bar{\Omega}^\ell$ erhält man durch Lösen eines Optimierungsproblems der Form

$$\vec{\mathbf{x}}^* \;=\; \arg\min_{\vec{\mathbf{x}}\in\bar{\Omega}^\ell} \Psi\big(M(\vec{\mathbf{x}})\big) \tag{3.11}$$

mit der Fisher-Informationsmatrix

$$M(\vec{\mathbf{x}}) \;=\; \frac{1}{\sigma^2}\sum_{i=1}^{\ell}\int_T \nabla_{\boldsymbol{\theta}} u(t,\mathbf{x}^i,\hat{\boldsymbol{\theta}}^1)\nabla_{\boldsymbol{\theta}} u(t,\mathbf{x}^i,\hat{\boldsymbol{\theta}}^1)^\top dt \tag{3.12}$$

und der Abbildung $\Psi : \mathbb{R}^{m\times m} \to \mathbb{R}$ definiert durch

$$\Psi\big(M(\vec{\mathbf{x}})\big) \;:=\; -\ln|M(\vec{\mathbf{x}})|. \tag{3.13}$$

Anhand der eindimensionalen Wärmeleitungsgleichung aus Beispiel 2.2.1, bei der eine Parameteridentifikation bereits durchgeführt wurde, wird nun eine optimale Messstellenkonfiguration $\vec{\mathbf{x}}^* \in \bar{\Omega}^\ell$ bestimmt. Dieses Beispiel ist dem Buch [85] entnommen und enthält genau einen unbekannten, zu schätzenden Parameter θ, so dass $m = 1$ gilt. In diesem Kapitel wird in Satz 3.3.2 ausgesagt, dass eine optimale Messstellenkonfiguration $\vec{\mathbf{x}}^* \in \bar{\Omega}^\ell$ aus höchstens $\frac{m(m+1)}{2}$ Messstellen besteht. Da für die Wärmeleitungsgleichung aus Beispiel 2.2.1 $m = 1$ gilt, ist folglich $\ell := 1$, so dass lediglich eine Messposition gesucht wird.

Beispiel 3.3.1. (Die 1D-Wärmeleitungsgleichung)
Anhand der eindimensionalen Wärmeleitungsgleichung aus Beispiel (2.2.1) wird gezeigt, wie eine optimale Messstelle $x^* \in \bar{\Omega}$ bestimmt werden kann. Mit $\Omega := (0,1)$ und $T := (0,t_f), 0 < t_f < \infty$ ist die Wärmeleitungsgleichung gegeben durch

$$\partial_t u(t,x,\theta) \;=\; \theta\Delta_x u(t,x,\theta), \quad x\in\Omega, \quad t\in T,$$

mit der homogenen Randbedingung

$$u(t,0,\theta) \;=\; u(t,1,\theta) \;=\; 0, \quad \textit{für } t\in T,$$

und der Anfangsbedingung

$$u(0,x,\theta) \;=\; \sin(\pi x), \quad \textit{für } x\in\Omega.$$

Die analytische Lösung für diese Gleichung ist bekannt und lautet

$$u(t,x,\theta) \;=\; \exp(-\theta\pi^2 t)\sin(\pi x).$$

Um eine optimale Messstelle $x^* \in \bar{\Omega}$ zu finden, wird die Fisher-Informationsmatrix $M(x)$ benötigt. Mit der partiellen Ableitung $\partial_\theta u(t,x,\theta) = -\pi^2 t\exp(-\theta\pi^2 t)\sin(\pi x)$ gilt

$$\begin{aligned} M(x) \;&=\; \frac{1}{\sigma^2}\int_0^{t_f}\big(-\pi^2 t\exp(-\theta\pi^2 t)\sin(\pi x)\big)^2 dt \\ &=\; \frac{\pi^4}{\sigma^2}\int_0^{t_f}\big(t\exp(-\theta\pi^2 t)\big)^2 dt\sin^2(\pi x) \;=\; C\sin^2(\pi x), \end{aligned}$$

mit der Konstanten $C = \frac{\pi^4}{\sigma^2} \int_0^{t_f} \left(t \exp(-\theta \pi^2 t)\right)^2 dt > 0$. *Das Zielfunktional*

$$\Psi\big(M(x)\big) \;=\; -\ln\big(C \sin^2(\pi x)\big)$$

ist genau dann minimal, wenn $\tilde{\Psi}(x) = -\sin^2(\pi x)$ *minimal ist und* $\tilde{\Psi}$ *ist genau in* $x^* = 0.5 \in \bar{\Omega}$ *minimal. Folgende Abbildung zeigt die Funktion* $\tilde{\Psi}(x)$ *mit dem Minimum* $x^* = 0.5$*:*

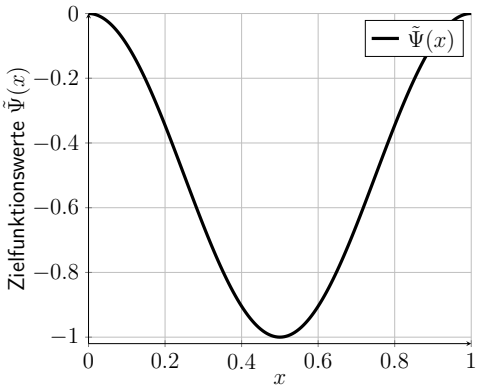


Abbildung 3.1: Die Funktion $\tilde{\Psi}(x)$. Die Fisher-Information $\Psi\big(M(x)\big)$ ist genau dann minimal, wenn $\tilde{\Psi}(x)$ minimal ist. Man sieht, dass sich das Minimum in $x^* = 0.5$ befindet.

Es wird nun ein Verfahren vorgestellt, mit dem eine optimale Messstellenkonstellation $\vec{\mathbf{x}}^* \in \bar{\Omega}^\ell$ numerisch ermittelt werden kann. In Vorbereitung dessen werden zunächst die Begriffe eines *exakten Designs*, *approximativen Designs* und *stetigen Designs* definiert.

3.3.1 Vom exakten zum stetigen Design

Die Fisher-Informationsmatrix $M(\vec{\mathbf{x}}) \in \mathbb{R}^{m \times m}$ aus (3.12) kann in die Summe von Teilmatrizen zerlegt werden. Es gilt

$$M(\vec{\mathbf{x}}) \;=\; \frac{1}{\sigma^2} \sum_{i=1}^{\ell} \int_T \nabla_{\boldsymbol{\theta}} u(t, \mathbf{x}^i, \hat{\boldsymbol{\theta}}^1)\, \nabla_{\boldsymbol{\theta}} u(t, \mathbf{x}^i, \hat{\boldsymbol{\theta}}^1)^\top dt \;=\; \sum_{i=1}^{\ell} \Upsilon(\mathbf{x}^i),$$

mit

$$\Upsilon(\mathbf{x}^i) \;=\; \frac{1}{\sigma^2} \int_T \nabla_{\boldsymbol{\theta}} u(t, \mathbf{x}^i, \hat{\boldsymbol{\theta}}^1)\, \nabla_{\boldsymbol{\theta}} u(t, \mathbf{x}^i, \hat{\boldsymbol{\theta}}^1)^\top dt.$$

Die Matrix $M(\vec{\mathbf{x}})$ ist demnach die Summe der Informationsmatrizen für individuelle Messstellen. Bevor beschrieben wird, wie man eine optimale Messstellenkonstellation $\vec{\mathbf{x}}^*$ numerisch ermitteln kann, werden die Begriffe *exaktes Design*, *approximatives Design* und *stetiges Design* eingeführt. Ein Design ist eine Matrixdarstellung eines optimalen $\vec{\mathbf{x}}^*$ und wird als Hilfsmittel für die Optimierung von (3.11) verwendet.

Exaktes Design

Das Ziel ist die Bestimmung einer optimalen Konfiguration von Messstellen

$$\vec{\mathbf{x}}^* \;=\; (\mathbf{x}^{1,*^\top}, \ldots, \mathbf{x}^{N_\ell,*^\top})^\top \in \bar{\Omega}^{N_\ell},$$

so dass das Zielfunktional $\Psi(M(\vec{\mathbf{x}}))$ in $\vec{\mathbf{x}}^*$ minimal ist. Erfahrungsgemäß kann ein solches $\vec{\mathbf{x}}^*$ mehrfach gleiche Messstellen beinhalten, d. h. $\mathbf{x}^i = \mathbf{x}^j$ für $i \neq j$. Um dieses zu vermeiden, fasst ein *exaktes Design* mehrfache Messstellen zusammen und gewichtet diese entsprechend:

Ein *exaktes Design* $\boldsymbol{\xi}_\ell$ ist eine Matrix

$$\boldsymbol{\xi}_\ell \;=\; \left\{ \begin{array}{cccc} \mathbf{x}^1, & \mathbf{x}^2, & \ldots, & \mathbf{x}^\ell \\ p_1, & p_2, & \ldots, & p_\ell \end{array} \right\}, \tag{3.14}$$

mit $\mathbf{x}^i \neq \mathbf{x}^j$ für $i \neq j$, wobei die prozentuale Gewichtung von $\mathbf{x}^i$ für $i \in I_\ell$ in (3.14) durch $p_i = \dfrac{r_i}{N_\ell} \in [0,1]$ mit $r_i \in \mathbb{N}$ und $\sum_{i=1}^{\ell} r_i = N_\ell$ dargestellt wird.

Definition 3.3.1. (Designpunkt)
Eine Messstelle $\mathbf{x}^i \in \bar{\Omega}$ *aus einem Design* $\boldsymbol{\xi}_\ell$ *wird* Designpunkt *genannt.*

Das exakte Design (3.14) kann interpretiert werden als eine diskrete Wahrscheinlichkeitsverteilung für die Designpunkte $\mathbf{x}^1, \ldots, \mathbf{x}^\ell$. Die Fisher-Informationsmatrix kann nun mit (3.14) umformuliert werden zu

$$M(\boldsymbol{\xi}_\ell) \;=\; \frac{1}{\sigma^2} \sum_{i=1}^{\ell} p_i \int_T \nabla_{\boldsymbol{\theta}} u(t, \mathbf{x}^i, \hat{\boldsymbol{\theta}}^1)\, \nabla_{\boldsymbol{\theta}} u(t, \mathbf{x}^i, \hat{\boldsymbol{\theta}}^1)^\top dt \;=\; \sum_{i=1}^{\ell} p_i \Upsilon(\mathbf{x}^i).$$

Um ein optimales *exaktes Design* $\boldsymbol{\xi}_\ell^*$ bezüglich (3.1) zu ermitteln, löst man das Optimierungsproblem

$$\boldsymbol{\xi}_\ell^* \;=\; \arg\min \Psi(M(\boldsymbol{\xi}_\ell)), \tag{3.15}$$

unter den Nebenbedingungen

$$\mathbf{x}^i \neq \mathbf{x}^j \text{ für } i \neq j, \quad p_i = \frac{r_i}{N_\ell} \text{ mit } r_i \in \mathbb{N} \text{ und } \sum_{i=1}^{\ell} r_i = N_\ell. \tag{3.16}$$

Hierbei handelt es sich um ein gemischt-ganzzahliges Optimierungsproblem, welches zum Beispiel mit einem Branch-and-Bound-Verfahren [62] gelöst werden kann. Da das Lösen gemischt-ganzzahliger Probleme sehr kostspielig ist, wird das Optimierungsproblem (3.15) mit (3.16) in dieser Arbeit nicht gelöst. Um die Forderung an ganzzahlige Lösungen vermeiden zu können, wird nun das sogenannte *approximative Design* eingeführt.

Approximatives Design

Bei einem *approximativen Design* werden die Gewichte p_i für $i \in I_\ell$ als reellwertige Größen aus dem Intervall $[0,1]$ betrachtet mit der Forderung, dass $\sum_{i=1}^{\ell} p_i = 1$ gilt. Ein *approximatives Design* $\boldsymbol{\xi}$ hat die Darstellung

$$\boldsymbol{\xi} \;=\; \left\{ \begin{array}{cccc} \mathbf{x}^1, & \mathbf{x}^2, & \ldots, & \mathbf{x}^\ell; \\ p_1, & p_2, & \ldots, & p_\ell \end{array} \quad \sum_{i=1}^{\ell} p_i = 1 \right\}, \tag{3.17}$$

wobei $\mathbf{x}^i \neq \mathbf{x}^j$ für $i \neq j$ gilt. Die Zahl $p_i \in [0,1]$ gibt wie beim exakten Design die prozentuale Gewichtung einer Messstelle $\mathbf{x}^i \in \bar{\Omega}$ an.

Um ein optimales *approximatives Design* $\boldsymbol{\xi}^*$ bezüglich (3.1) zu ermitteln, wird das Optimierungsproblem

$$\boldsymbol{\xi}_\ell^* \;=\; \arg\min \Psi(M(\boldsymbol{\xi}_\ell)),$$

unter den Nebenbedingungen

$$\mathbf{x}^i \in \bar{\Omega}, \quad \mathbf{x}^i \neq \mathbf{x}^j \text{ für } i \neq j, \quad p_i \in [0,1], \quad \sum_{i=1}^{\ell} p_i = 1$$

gelöst.

Bemerkung 3.3.1. *Im Vergleich zu den exakten Designs ist die Menge aller approximativen Designs konvex. Somit gilt für zwei beliebige approximative Designs* $\boldsymbol{\xi}^1$ *und* $\boldsymbol{\xi}^2$ *der Form (3.17), dass* $\tilde{\boldsymbol{\xi}} = (1-\alpha)\boldsymbol{\xi}^1 + \alpha\boldsymbol{\xi}^2$ *für jedes* $\alpha \in [0,1]$ *wieder ein approximatives Design ist.*
Mit

$$\boldsymbol{\xi}^1 \;=\; \left\{ \begin{array}{cccc} \mathbf{x}^1, & \mathbf{x}^2, & \ldots, & \mathbf{x}^\ell \\ p_1^1, & p_2^1, & \ldots, & p_\ell^1 \end{array} \right\} \quad \textit{und} \quad \boldsymbol{\xi}^2 \;=\; \left\{ \begin{array}{cccc} \mathbf{x}^1, & \mathbf{x}^2, & \ldots, & \mathbf{x}^\ell \\ p_1^2, & p_2^2, & \ldots, & p_\ell^2 \end{array} \right\}$$

ist

$$\tilde{\boldsymbol{\xi}} \;=\; \left\{ \begin{array}{ccc} \mathbf{x}^1, & \mathbf{x}^2, & \ldots, \quad \mathbf{x}^\ell \\ (1-\alpha)p_1^1 + \alpha p_1^2, & (1-\alpha)p_2^1 + \alpha p_2^2, & \ldots, \quad (1-\alpha)p_\ell^1 + \alpha p_\ell^2 \end{array} \right\}.$$

Fasst man die *approximativen Designs* als ein Wahrscheinlichkeitsmaß über Ω auf, erhält man das *stetige Design*.

Stetiges Design

Wie in [85] beschrieben, kann die Klasse der *approximativen Designs* zur Klasse der sogenannten *stetigen Designs* erweitert werden. Hierfür wird $\boldsymbol{\xi}$ aufgefasst als ein Wahrscheinlichkeitsmaß über Ω, welches absolut stetig bezüglich des Lebesguemaßes ist und die Bedingung

$$\int_\Omega \boldsymbol{\xi}(d\mathbf{x}) = 1 \tag{3.18}$$

erfüllt (Lebesgue-Stieltjes-Integral [85]). Mit dem *stetigen Design* (3.18) kann nun die Fisher-Informationsmatrix umformuliert werden zu

$$M(\boldsymbol{\xi}) \;=\; \int_\Omega \Upsilon(\mathbf{x})\boldsymbol{\xi}(d\mathbf{x}) \tag{3.19}$$

mit

$$\Upsilon(\mathbf{x}) = \frac{1}{\sigma^2}\int_T \nabla_{\boldsymbol{\theta}} u(t,\mathbf{x},\hat{\boldsymbol{\theta}}^1)\,\nabla_{\boldsymbol{\theta}} u(t,\mathbf{x},\hat{\boldsymbol{\theta}}^1)^\top\,dt.$$

Definition 3.3.2. (D-optimales Design)
Ein stetiges Design $\boldsymbol{\xi}^*$ *für dass*

$$\boldsymbol{\xi}^* \;=\; \arg\min \Psi(M(\boldsymbol{\xi}))$$

mit $M(\boldsymbol{\xi})$ *aus (3.19) gilt, wird* D-optimales Design *genannt.*

Auf Basis der Klasse von *stetigen Designs* wird in diesem Kapitel gezeigt, wie man ein D-optimales Design $\boldsymbol{\xi}^*$ numerisch ermitteln kann. Die hierfür benötigte Theorie ist weitestgehend dem Buch [85] entnommen. Folgende Notationen werden benötigt. Es ist

- $\Xi(\Omega)$ die Menge aller Wahrscheinlichkeitsmaße $\boldsymbol{\xi}$ auf Ω,
- $\mathfrak{M}(\Omega)$ die Menge aller zulässigen Informationsmatrizen

$$\mathfrak{M}(\Omega) = \{M(\boldsymbol{\xi}) : \boldsymbol{\xi} \in \Xi(\Omega)\},$$

- $\boldsymbol{\xi}^* = \arg\min_{\boldsymbol{\xi}\in\Xi(\Omega)} \Psi(M(\boldsymbol{\xi}))$ ein D-optimales Design.

3.3.2 Eigenschaften eines D-optimalen Designs

In Kapitel 3.3.3 wird ein Verfahren vorgestellt, mit dem ein D-optimales Design $\boldsymbol{\xi}^* \in \Xi(\Omega)$ numerisch ermittelt werden kann. In Vorbereitung dessen werden in diesem Abschnitt wichtige Eigenschaften eines D-optimalen Designs $\boldsymbol{\xi}^* \in \Xi(\Omega)$ wiedergegeben, die für dieses Verfahren benötigt werden. Die hierfür verwendeten Sätze sind dem Buch [85] entnommen.

Das Ziel ist die Bestimmung eines D-optimalen Designs $\boldsymbol{\xi}^* \in \Xi(\Omega)$, so dass mit der Fisher-Informationsmatrix $M(\boldsymbol{\xi})$ aus (3.19) und der Abbildung $\Psi : \mathbb{R}^{m\times m} \to \mathbb{R}$ aus (3.13)

$$\boldsymbol{\xi}^* \;=\; \arg\min_{\boldsymbol{\xi}\in\Xi(\Omega)} \Psi(M(\boldsymbol{\xi})) \tag{3.20}$$

gilt. In diesem Kapitel gelten folgende Voraussetzungen:

Voraussetzung 3.3.1.

(V1) *Die Menge $\bar{\Omega}$ ist kompakt.*

(V2) *Für jedes fest gewählte $\boldsymbol{\theta} \in \Theta$ gilt für die Sensitivitäten erster Ordnung*

$$\nabla_{\boldsymbol{\theta}} u(t, \mathbf{x}, \boldsymbol{\theta}) \;\in\; \mathcal{C}^1(T; V)^m.$$

Um das Optimierungsproblem (3.20) lösen zu können, ist es sinnvoll im Vorhinein zu wissen, wieviele Designpunkte $\mathbf{x}^i$ in einem gesuchten D-optimalen Design $\boldsymbol{\xi}^*$ enthalten sind. Das die Anzahl ℓ der D-optimalen Designpunkte nach oben beschränkt ist, wird im folgenden Abschnitt erläutert.

Obere Schranke für die Anzahl der D-optimalen Designpunkte

In [85] wurde bewiesen, dass die Anzahl $0 < \ell$ der D-optimalen Designpunkte $\mathbf{x}^{i,*}$ nach oben beschränkt ist und von der Anzahl $0 < m$ der zu schätzenden Modellparameter θ_i abhängt. Folgender Satz gibt diesen Zusammenhang wieder:

Satz 3.3.2. ([85])
Unter der Voraussetzung 3.3.1 beinhaltet ein D-optimales Design $\boldsymbol{\xi}^*$ höchstens $\frac{m(m+1)}{2}$ Designpunkte $\mathbf{x}^i$, d. h. es ist

$$\ell \;\leq\; \frac{m(m+1)}{2},$$

wobei $0 < m \in \mathbb{N}$ die Anzahl der unbekannten, zu schätzenden Parameter θ_i für $i = 1, \ldots, m$ darstellt.

Wie in [85] beschrieben, kann der Satz 3.3.2 mit dem Satz von Carathéodory bewiesen werden.

An dieser Stelle sei zu erwähnen, dass es stetige Designs $\boldsymbol{\xi} \in \Xi(\Omega)$ gibt, die mehr als ℓ Designpunkte beinhaltet, für die aber $M(\boldsymbol{\xi}) = M(\boldsymbol{\xi}^*)$ gilt, d. h. sie besitzen die selbe Fisher-Information wie ein D-optimales Design $\boldsymbol{\xi}^*$. Ein D-optimales Design ist demnach ein Design mit minimaler Anzahl an Designpunkten bei minimaler Fisher-Information.

Richtungsableitung

Das Verfahren zur Bestimmung eines D-optimalen Designs, welches in Kapitel 3.3.3 vorgestellt wird, ist ein iteratives Verfahren, welches ein Optimum von (3.20) mithilfe der einseitigen Richtungsableitung ermittelt. Die einseitige Richtungsableitung einer Funktion $\Psi : \mathbb{R}^{m\times m} \to \mathbb{R}$ an der Stelle $M(\boldsymbol{\xi}) \in \mathfrak{M}(\Omega)$ in Richtung $H \in \mathfrak{M}(\Omega)$ ist definiert durch

$$D^+\Psi\big(M(\boldsymbol{\xi})\big)H \quad := \quad \lim_{\substack{\alpha\to 0\\ \alpha>0}} \frac{\Psi\big(M(\boldsymbol{\xi}) + \alpha H\big) - \Psi\big(M(\boldsymbol{\xi})\big)}{\alpha}.$$

Mit Ψ aus (3.13) und der Richtung $H = M(\bar{\boldsymbol{\xi}}) - M(\boldsymbol{\xi})$ mit den Designs $\boldsymbol{\xi} \in \mathfrak{M}(\Omega)$ und $\bar{\boldsymbol{\xi}} \in \mathfrak{M}(\Omega)$ erhält man als einseitig Richtungsableitung

$$\begin{aligned} D^+\Psi\big(M(\boldsymbol{\xi})\big)\big[M(\bar{\boldsymbol{\xi}}) - M(\boldsymbol{\xi})\big] \quad &= \quad \lim_{\substack{\alpha\to 0\\ \alpha>0}} \frac{\Psi\big(M(\boldsymbol{\xi}) + \alpha(M(\bar{\boldsymbol{\xi}}) - M(\boldsymbol{\xi}))\big) - \Psi\big(M(\boldsymbol{\xi})\big)}{\alpha} \\ &=: \quad \int_\Omega \psi(\mathbf{x}, \boldsymbol{\xi})\, \bar{\boldsymbol{\xi}}(d\mathbf{x}), \end{aligned} \tag{3.21}$$

mit dem Funktional

$$\psi(\mathbf{x}, \boldsymbol{\xi}) \quad = \quad m - \frac{1}{t_f} \int_0^{t_f} \nabla_{\boldsymbol{\theta}} u(t, \mathbf{x}, \boldsymbol{\theta})^\top M(\boldsymbol{\xi})^{-1} \nabla_{\boldsymbol{\theta}} u(t, \mathbf{x}, \boldsymbol{\theta})\, dt. \tag{3.22}$$

Die ausführliche Herleitung dieser Richtungsableitung für das D-Optimalitätskriterium (3.1) kann in [85] nachgelesen werden.

Mit dem Funktional $\psi(\mathbf{x}, \boldsymbol{\xi})$ aus (3.22) wird nun das sogenannte Äquivalenztheorem für D-Optimale Designs wiedergegeben.

Äquivalenztheorem für D-Optimale Designs

Folgendes Äquivalenztheorem für D-Optimale Designs ist von großer Relevanz bei der Bestimmung von $\boldsymbol{\xi}^*$ mit dem Algorithmus 3.3.1, der im folgenden Kapitel 3.3.3 vorgestellt wird. Es gilt mit

$$\Phi(\mathbf{x}, \boldsymbol{\xi}) \quad = \quad \frac{1}{t_f} \int_0^{t_f} \nabla_{\boldsymbol{\theta}} u(t, \mathbf{x}, \boldsymbol{\theta})^\top M(\boldsymbol{\xi})^{-1} \nabla_{\boldsymbol{\theta}} u(t, \mathbf{x}, \boldsymbol{\theta})\, dt \tag{3.23}$$

und demnach $\psi(\mathbf{x}, \boldsymbol{\xi}) = m - \Phi(\mathbf{x}, \boldsymbol{\xi})$ das folgende Theorem:

Theorem 3.3.1. (Äquivalenztheorem für D-Optimale Designs, [85])
Unter der Voraussetzung 3.3.1 sind die folgenden Bedingungen äquivalent:

(*i*) *Das Design* $\boldsymbol{\xi}^*$ *minimiert* $\Psi\big(M(\boldsymbol{\xi})\big) = -\ln|M(\boldsymbol{\xi})|$.

(ii) *Das Design* $\boldsymbol{\xi}^*$ *minimiert* $\max_{\mathbf{x}\in\Omega} \Phi(\mathbf{x}, \boldsymbol{\xi})$.

(iii) *Es ist* $\max_{\mathbf{x}\in\Omega} \Phi(\mathbf{x}, \boldsymbol{\xi}^*) = m$.

Mithilfe dieser Äquivalenzbedingungen für D-Optimale Designs wird nun ein Verfahren beschrieben, mit dem numerisch eine Lösung $\boldsymbol{\xi}^*$ des Optimierungsproblems (3.20) ermittelt werden kann.

3.3.3 Algorithmus zur Bestimmung eines D-optimalen Designs

In diesem Abschnitt wird ein numerisches Verfahren zur Bestimmung eines D-optimalen Designs $\boldsymbol{\xi}^* \in \Xi(\Omega)$ vorgestellt, welches im Buch [85] ausführlich hergeleitet wurde. Hierbei handelt es sich um ein Verfahren, welches ausgehend von einem Startdesign $\boldsymbol{\xi}^0 \in \Xi(\Omega)$ mit $l := \frac{m(m+1)}{2}$ Designpunkten sukzessive in jedem Iterationsschritt einen weiteren Designpunkt $\tilde{\mathbf{x}}$ sucht, so dass mit dem Zielfunktional Ψ aus (3.13), dem Einpunkt-Design

$$\delta_{\tilde{\mathbf{x}}} \;:=\; \begin{Bmatrix} \tilde{\mathbf{x}} \\ 1 \end{Bmatrix}$$

und einer Schrittweite $\alpha \in (0,1)$

$$\Psi\big(M((1-\alpha)\boldsymbol{\xi}^n + \alpha\delta_{\tilde{\mathbf{x}}})\big) \;\leq\; \Psi\big(M(\boldsymbol{\xi}^n)\big)$$

gilt. Das Design nach dem n-ten Iterationsschritt ist dann $\boldsymbol{\xi}^{n+1} := (1-\alpha)\boldsymbol{\xi}^n + \alpha\delta_{\tilde{\mathbf{x}}}$.

Folgender Algorithmus beschreibt, wie mithilfe der Funktion $\psi(\mathbf{x}, \boldsymbol{\xi})$ aus (3.22) für ein Design $\boldsymbol{\xi}$ ein Designpunkt $\tilde{\mathbf{x}}$ ermittelt werden kann, so dass mit

$$\boldsymbol{\xi}^+ \;=\; (1-\alpha)\boldsymbol{\xi} + \alpha\delta_{\tilde{\mathbf{x}}}$$

das Design $\boldsymbol{\xi}^+ - \boldsymbol{\xi}$ die Richtung des steilsten Abstiegs im Zielfunktional darstellt.

Algorithmus 3.3.1. (Iterative Bestimmung eines D-optimalen Designs, [85])

(S.0) *Initialisierung:*
Wähle ein $\boldsymbol{\xi}^0 \in \Xi(\Omega)$ *mit* $l := \frac{m(m+1)}{2}$, *so dass* $\Psi\big(M(\boldsymbol{\xi}^0)\big) < \infty$.
Wähle ein $0 < \varepsilon \ll 1$ *und setze* $n := 0$.

(S.1) *Berechnung der Suchrichtung:*
Bestimme $\tilde{\mathbf{x}} := \arg\max_{\mathbf{x}\in\bar{\Omega}} \Phi(\mathbf{x}, \boldsymbol{\xi}^n)$.

(S.2) *Abbruchbedingung:*
Ist $\Phi(\tilde{\mathbf{x}}, \boldsymbol{\xi}^n) < m + \varepsilon$, *dann ist* $\boldsymbol{\xi}^n$ *das gesuchte D-optimale Design. Stop.*
Sonst gehe zu (S.3).

(S.3) *Schrittweitenbestimmung:*
Bestimme

$$\alpha^n := \arg\min_{\alpha\in(0,1)} \Psi\big((1-\alpha)M(\boldsymbol{\xi}^n) + \alpha\Upsilon(\tilde{\mathbf{x}})\big)$$

und setze $\boldsymbol{\xi}^{n+1} := (1-\alpha^n)\boldsymbol{\xi}^n + \alpha^n\delta_{\tilde{\mathbf{x}}}$, $n := n+1$ *und gehe zu (S.1).*

Ist der in Schritt (S.1) ermittelte Designpunkt $\tilde{\mathbf{x}} \in \bar{\Omega}$ noch nicht im Design $\boldsymbol{\xi}^n$ enthalten, wird dieser in das neue Design $\boldsymbol{\xi}^{n+1}$ so hinzugefügt, dass $\boldsymbol{\xi}^{n+1}$ genau einen Designpunkt mehr enthält als $\boldsymbol{\xi}^n$. Mit

$$\boldsymbol{\xi}^n = \left\{ \begin{array}{cccc} \mathbf{x}^{1,n}, & \mathbf{x}^{2,n}, & \ldots, & \mathbf{x}^{\ell,n} \\ p_1^n, & p_2^n, & \ldots, & p_\ell^n \end{array} \right\}, \; \delta_{\tilde{\mathbf{x}}} = \left\{ \begin{array}{c} \tilde{\mathbf{x}} \\ 1 \end{array} \right\} \quad \text{und} \quad \boldsymbol{\xi}^{n+1} := (1-\alpha^n)\boldsymbol{\xi}^n + \alpha^n \delta_{\tilde{\mathbf{x}}}$$

ist

$$\boldsymbol{\xi}^{n+1} = \left\{ \begin{array}{cccc} \mathbf{x}^{1,n}, & \mathbf{x}^{2,n}, & \ldots, & \mathbf{x}^{\ell,n}, & \tilde{\mathbf{x}} \\ (1-\alpha^n)p_1^n, & (1-\alpha^n)p_2^n, & \ldots, & (1-\alpha^n)p_\ell^n, & \alpha^n \end{array} \right\},$$

falls $\tilde{\mathbf{x}} \neq \mathbf{x}^i$ für alle $i \in I_\ell$ gilt.

Folglich kann es vorkommen, dass ein Design $\boldsymbol{\xi}^n$ in Algorithmus 3.3.1 mehr als $l = \frac{m(m+1)}{2}$ Designpunkte enthält. Da in Algorithmus 3.3.1 zu keinem Zeitpunkt wieder ein Designpunkt aus einem Design $\boldsymbol{\xi}^n$ herausgenommen wird, scheint dieses ein Widerspruch zu Satz 3.3.2 zu sein, da dann $\boldsymbol{\xi}^*$ mehr als ℓ Designpunkte enthält. Tatsächlich kann man aber beobachten, dass die Gewichte p_i^n für $n \to \infty$ gegen Null konvergieren, wenn der zugehörige Designpunkt $\mathbf{x}^i$ nicht im D-optimalen Design $\boldsymbol{\xi}^*$ enthalten ist. Daher kann bei der Umsetzung von Algorithmus 3.3.1 ein Designpunkt $\mathbf{x}^i$ jederzeit entfernt werden, wenn das Gewicht p_i^n unter eine vorgegebene Schranke $0 < \varepsilon_p \ll 1$ fällt. Auf diese Weise kann ein D-optimales Design $\boldsymbol{\xi}^*$ in kleinerer Anzahl an Iterationsschritten ermittelt werden [85].

Cluster-Problem

Wie in [85] beschrieben, kann bei der Realisierung von Algorithmus 3.3.1 zur Bestimmung eines D-optimalen Designs das sogenannte *Cluster-Problem* erwartet werden. Das *Cluster-Problem* tritt in Schritt (S1) von Algorithmus 3.3.1 auf, wo durch Bestimmung von

$$\tilde{\mathbf{x}} \; := \; \arg\max_{\mathbf{x} \in \bar{\Omega}} \Phi(\mathbf{x}, \boldsymbol{\xi}^n)$$

der Designpunkt $\tilde{\mathbf{x}}$ ermittelt wird, der eine Minimierung im Zielfunktional $\Psi(M(\boldsymbol{\xi}))$ bewirkt. Ein solcher Punkt wird entweder in das Design $\boldsymbol{\xi}^n$ hinzugenommen (falls dieser noch nicht enthalten ist) oder das zugehörige Gewicht p_i^n in $\boldsymbol{\xi}^n$ wird vergrößert.

Erfahrungsgemäß kommt es allerdings vor, dass ein in Schritt (S.1) ermittelter Designpunkt $\tilde{\mathbf{x}}$ zwar noch nicht in einem Design $\boldsymbol{\xi}^n$ enthalten ist, es aber einen Designpunkt $\mathbf{x}^i$ gibt mit $\|\mathbf{x}^i - \tilde{\mathbf{x}}\|_2^2 < \varepsilon_x$, $0 < \varepsilon_x \ll 1$, d. h. $\tilde{\mathbf{x}}$ liegt in einer kleinen Umgebung von $\mathbf{x}^i$.

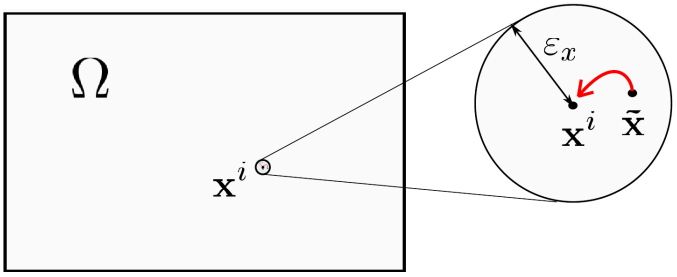


Abbildung 3.2: Befindet sich ein mit Algorithmus 3.3.1 ermittelter Designpunkt $\tilde{\mathbf{x}}$ in einer kleinen Umgebung um einen bereits existierenden Designpunktes $\mathbf{x}^i$, dann muss $\tilde{\mathbf{x}} := \mathbf{x}^i$ gesetzt werden um die Abbruchbedingung (S.2) zu erreichen.

In diesem Fall müssen nach [85] alle Designpunkte in einer Umgebung von $\mathbf{x}^i$ aggregiert und gleich $\mathbf{x}^i$ gesetzt werden, da sich sonst in einer Umgebung von $\mathbf{x}^i$ viele Designpunkte häufen können, was zur Folge hat, dass die Abbruchbedingung (S.2) nie erreicht werden könnte.

3.3.4 Komplexität

In diesem Abschnitt wird die Komplexität von Algorithmus 3.3.1 wiedergegeben, wenn für ein fest gewähltes $\boldsymbol{\theta} \in \Theta$ die Fisher-Informationsmatrix $M(\boldsymbol{\xi})$ mithilfe der Finite-Elemente-Lösung der Zustandsgleichung (2.24) - (2.25)

$$u(t, \mathbf{x}, \boldsymbol{\theta}) \ \in \ H^1(T; V)$$

und den Sensitivitäten erster Ordnung als Finite-Elemente-Lösung von (2.33) - (2.34)

$$\nabla_{\boldsymbol{\theta}} u(t, \mathbf{x}, \boldsymbol{\theta}) \ \in \ H^1(T; V)^m$$

berechnet wird. Die für die Diskretisierung verwendete zeitliche Schrittweite wird mit Δt und die räumliche Schrittweite mit h bezeichnet. In diesem Fall liegt ein D-optimaler Designpunkt $\mathbf{x}^{i,*}$ für $i \in I_\ell$, der mit Algorithmus 3.3.1 ermittelt wurde, immer auf einem Gitterpunkt der diskreten Menge Ω_h (siehe z. B. [85]).

Folgende Tabelle gibt einen Überblick über die Komplexität der Schritte (S.0) - (S.3) von Algorithmus 3.3.1 in Abhängigkeit von

- der Anzahl der unbekannten, zu schätzenden Parameter $m > 0$,
- der Schrittweite in der Zeit $\Delta t > 0$,
- der Schrittweite im Ort $h > 0$,
- der Dimension $d > 0$ von Ω,
- und der Schrittweite $\Delta\alpha > 0$,

wieder. Es ist

	Anzahl der erforderlichen Operationen	Speicheraufwand
(S.0)	$\mathcal{O}(1)$	$\mathcal{O}(m^2)$
(S.1)	$\mathcal{O}(m^4(\Delta t)^{-1}h^{-d})$	$\mathcal{O}(m^2)$
(S.2)	$\mathcal{O}(1)$	$\mathcal{O}(1)$
(S.3)	$\mathcal{O}(m^2(\Delta\alpha)^{-1})$	$\mathcal{O}(1)$

Tabelle 3.1: Komplexität der Lösungsmethode zur Bestimmung eines D-optimalen Designs.

Kapitel 4

Primal-dualer Ansatz zur Bestimmung eines D-optimalen Designs

In diesem Kapitel wird eine neue Vorgehensweise zur Bestimmung eines D-optimalen Designs hergeleitet, die darin besteht, das zu lösende Problem aus Kapitel 3 als ein kontinuierliches Optimierungsproblem aufzufassen. Dieses kann zum Beispiel mit einem primal-dualen Innere-Punkte-Verfahren oder einer Active-Set-Methode gelöst werden.

In diesem Abschnitt wird zunächst vorausgesetzt, dass für jedes fest gewählte $\boldsymbol{\theta} \in \Theta$ die Sensitivitäten $\nabla_{\boldsymbol{\theta}} u(t, \mathbf{x}, \boldsymbol{\theta})$ in der Ortsvariablen $\mathbf{x}$ einmal stetig differenzierbar sind, d. h.

$$\nabla_{\boldsymbol{\theta}} u(t, \mathbf{x}, \boldsymbol{\theta}) \in \mathcal{C}^1(T; \mathcal{C}^1(\Omega; \mathbb{R}^m)). \tag{4.1}$$

4.1 Das kontinuierliche Optimierungsproblem

Wie in Kapitel 3 beschrieben, hat ein stetiges Design die Darstellung

$$\xi \;=\; \left\{ \begin{array}{cccc} \mathbf{x}^1, & \mathbf{x}^2, & \ldots, & \mathbf{x}^\ell \\ p_1, & p_2, & \ldots, & p_\ell \end{array} \right\}.$$

Von dieser Darstellung werden wir uns in diesem Kapitel trennen und betrachten von nun an die gesuchten Größen $\mathbf{x} = ({\mathbf{x}^1}^\top, \ldots, {\mathbf{x}^\ell}^\top)^\top$ und $\mathbf{p} = (p_1, \ldots, p_\ell)^\top$ separat voneinander. Somit erhält man das zu lösende, nichtlineare Optimierungsproblem

$$\min_{\mathbf{x},\mathbf{p}} J(\mathbf{x}, \mathbf{p}) \;=\; -\ln |M(\mathbf{x}, \mathbf{p})| \tag{4.2}$$

mit der Fisher-Informationsmatrix

$$M(\mathbf{x}, \mathbf{p}) \;=\; \sum_{i=1}^{\ell} p_i \int_T \nabla_{\boldsymbol{\theta}} u(t, \mathbf{x}^i, \boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} u(t, \mathbf{x}^i, \boldsymbol{\theta})^\top dt, \tag{4.3}$$

unter den Nebenbedingungen

$$\begin{aligned}\mathbf{p} &= (p_1,\ldots,p_\ell)^\top \in \mathbb{R}^\ell, \quad p_i \geq 0, \quad \sum_{i=1}^{\ell} p_i = 1,\\ \mathbf{x} &= ({\mathbf{x}^1}^\top,\ldots,{\mathbf{x}^\ell}^\top)^\top \in \mathbb{R}^{d\ell}, \quad \mathbf{a} \leq \mathbf{x}^i \leq \mathbf{b}, \quad \mathbf{a},\mathbf{b},\mathbf{x}^i \in \mathbb{R}^d.\end{aligned}$$

Zur übersichtlicheren Darstellung sei im folgenden

$$I_\ell := \{i \in \mathbb{N} : 1 \leq i \leq \ell\}. \tag{4.4}$$

Mit dieser Notation kann der zulässige Bereich des Optimierungsproblems (4.2) folgendermaßen dargestellt werden.

Definition 4.1.1. (Zulässiger Bereich) *Der zulässige Bereich des Optimierungsproblems (4.2) wird im folgenden durch die Menge*

$$\left.\begin{aligned}\mathcal{S} := \{(\mathbf{x},\mathbf{p}) \in \mathbb{R}^{(d+1)\ell} : \mathbf{f}^i(\mathbf{x},\mathbf{p}) &:= \mathbf{a} - \mathbf{x}^i \leq \mathbf{0} \ \textit{für}\ i \in I_\ell,\\ \mathbf{f}^{d\ell+i}(\mathbf{x},\mathbf{p}) &:= \mathbf{x}^i - \mathbf{b} \leq \mathbf{0} \ \textit{für}\ i \in I_\ell,\\ g_i(\mathbf{x},\mathbf{p}) &:= -p_i \leq 0 \ \textit{für}\ i \in I_\ell,\\ g_{\ell+1}(\mathbf{x},\mathbf{p}) &:= \sum_{j=1}^{\ell} p_j - 1 = 0\}\\ = \{(\mathbf{x},\mathbf{p}) \in \bar{\Omega}^\ell \times [0,1]^\ell : &\sum_{i=1}^{\ell} p_i = 1\}\end{aligned}\right\} \tag{4.5}$$

beschrieben, wobei die Menge $\bar{\Omega}$ das durch die Vektoren $\mathbf{a} \in \mathbb{R}^d$ und $\mathbf{b} \in \mathbb{R}^d$ begrenzte Hyperrechteck darstellt.

Da die konvexe Menge $\bar{\Omega}^\ell \times [0,1]^\ell$ abgeschlossen und die Gleichheitsbedingung $\sum_{i=1}^{\ell} p_i = 1$ affin ist, ist der zulässige Bereich $\mathcal{S}$ abgeschlossen und konvex [43]. Das Zielfunktional $J(\mathbf{x},\mathbf{p})$ ist im allgemeinen nicht konvex. Daher kann die Theorie der konvexen Optimierung nicht verwendet werden.

4.1.1 Regularitätsbedingung

Ein D-optimales Design kann ermittelt werden, indem das zum Optimierungsproblem (4.2) zugehörige KKT-System aufgestellt und gelöst wird. Ein auf diese Weise ermittelter KKT-Punkt enthält dann sowohl den primalen, als auch den dualen Anteil einer lokalen Lösung von (4.2). Um das Optimierungsproblem mithilfe der KKT-Bedingungen lösen zu können, muss zunächst überprüft werden, ob (4.2) in jedem lokalen Minimum regulär ist. Wie in [43] beschrieben, kann eine solche Regularität mit der Bedingung von Robinson nachgewiesen werden. Die Regularitätsbedingung von Robinson ist nach [43] erfüllt, wenn die Gradienten

- $\nabla f_k^j(\bar{\mathbf{x}},\bar{\mathbf{p}})$ für $j \in \{i \in I_\ell : f_k^i(\bar{\mathbf{x}},\bar{\mathbf{p}}) = 0\}$, $k \in K_d$,
- $\nabla f_k^{d\ell+j}(\bar{\mathbf{x}},\bar{\mathbf{p}})$ für $j \in \{i \in I_\ell : f_k^{d\ell+1}(\bar{\mathbf{x}},\bar{\mathbf{p}}) = 0\}$, $k \in K_d$,

und

- $\nabla g_j(\bar{\mathbf{x}}, \bar{\mathbf{p}})$ für $j \in \{i \in I_\ell \cup \{\ell+1\} : \; g_i(\bar{\mathbf{x}}, \bar{\mathbf{p}}) = 0\}$

linear unabhängig sind. Dann kann der Satz von Kuhn-Tucker für nichtkonvexe Optimierungsprobleme angewendet werden.

Satz 4.1.1. *Unter der Voraussetzung (4.1) ist das Optimierungsproblem (4.2) in jedem lokalen Minimum* $(\bar{\mathbf{x}}, \bar{\mathbf{p}}) \in \mathcal{S}$ *regulär.*

Beweis:
Um die Regularität in einem lokalen Optimum mit der Bedingung von Robinson zu zeigen wird der Vektor $\mathbf{e}^i \in \mathbb{R}^{(d+1)\ell}$ mit den Komponenten $e^i_j := \delta_{ij}$ eingeführt. Dann gilt für alle $k \in K_d$ und $(\mathbf{x}, \mathbf{p}) \in \mathcal{S}$

$$\left.\begin{array}{rcll} \nabla f^i_k(\mathbf{x}, \mathbf{p}) & = & -\mathbf{e}^{d\cdot(i-1)+k} & \text{für } i \in I_\ell, \\ \nabla f^{d\ell+i}_k(\mathbf{x}, \mathbf{p}) & = & \mathbf{e}^{d\cdot(i-1)+k} & \text{für } i \in I_\ell, \\ \nabla g_i(\mathbf{x}, \mathbf{p}) & = & -\mathbf{e}^{2d\ell+i} & \text{für } i \in I_\ell, \\ \nabla g_{\ell+1}(\mathbf{x}, \mathbf{p}) & = & (\underbrace{0, \ldots, 0}_{d\ell}, \underbrace{1, \ldots, 1}_{\ell})^\top. & \end{array}\right\} \tag{4.6}$$

Für die Vektoren aus (4.6) gilt

- Die Gradienten $\nabla f^i_k(\mathbf{x}, \mathbf{p})$ mit $i \in I_\ell$, $k \in K_d$ und $\nabla g_j(\mathbf{x}, \mathbf{p})$ mit $j \in I_\ell \cup \{\ell+1\}$ sind stets linear unabhängig.
- Die Gradienten $\nabla f^{d\ell+i}_k(\mathbf{x}, \mathbf{p})$ mit $i \in I_\ell$, $k \in K_d$ und $\nabla g_j(\mathbf{x}, \mathbf{p})$ mit $j \in I_\ell \cup \{\ell+1\}$ sind stets linear unabhängig.
- Die Gradienten $\nabla f^i_k(\mathbf{x}, \mathbf{p})$ und $\nabla f^{d\ell+i}_k(\mathbf{x}, \mathbf{p})$ sind für alle $i \in I_\ell, k \in K_d$ linear abhängig, allerdings ist stets $f^i_k(\mathbf{x}, \mathbf{p}) \neq 0$ oder $f^{d\ell+i}_k(\mathbf{x}, \mathbf{p}) \neq 0$, da $a_k < b_k$ für alle $k \in K_d$ vorausgesetzt wurde.
- Als einzige Gleichheitsbedingung ist $g_{\ell+1}(\mathbf{x}, \mathbf{p}) = 0$. Mit

  $$\sum_{i=1}^{\ell} (-\nabla g_i(\mathbf{x}, \mathbf{p})) = \nabla g_{\ell+1}(\mathbf{x}, \mathbf{p})$$

  ist folglich die Regularitätsbedingung von Robinson erfüllt, wenn mindestens ein $i \in I_\ell$ existiert mit $p_i \neq 0$. Da $\sum_{j=1}^{\ell} p_j = 1$ gilt, existiert immer ein solches $i \in I_\ell$.

Somit ist die Regularitätsbedingung von Robinson in jedem Punkt $(\mathbf{x}, \mathbf{p}) \in \mathcal{S}$ erfüllt, so dass folglich die Regularität in jedem lokalen Optimum $(\bar{\mathbf{x}}, \bar{\mathbf{p}}) \in \mathcal{S}$ gewährleistet ist.

□

Da die Regularitätsbedingung von Robinson erfüllt ist, kann der Satz von Kuhn-Tucker angewendet werden.

Satz 4.1.2. (Satz von Kuhn-Tucker für das Optimierungsproblem (4.2), [43])
Unter der Voraussetzung (4.1) sei $(\bar{\mathbf{x}}, \bar{\mathbf{p}}) \in \mathcal{S}$ *eine lokale Optimallösung von (4.2). Da das Problem (4.2) in* $(\bar{\mathbf{x}}, \bar{\mathbf{p}})$ *regulär ist, existiert ein* $(\bar{\boldsymbol{\lambda}}, \bar{\mu}) \in \mathbb{R}^{(2d+1)\ell} \times \mathbb{R}$ *mit* $\bar{\boldsymbol{\lambda}} := (\bar{\boldsymbol{\lambda}}_{\boldsymbol{a}}{}^\top, \bar{\boldsymbol{\lambda}}_{\boldsymbol{b}}{}^\top, \bar{\boldsymbol{\lambda}}_{\boldsymbol{p}}{}^\top)^\top \in \mathbb{R}^{d\ell} \times \mathbb{R}^{d\ell} \times \mathbb{R}^{\ell}$*, so dass*

(i) $\bar{\boldsymbol{\lambda}}_{\boldsymbol{a}},\ \bar{\boldsymbol{\lambda}}_{\boldsymbol{b}},\ \bar{\boldsymbol{\lambda}}_{\boldsymbol{p}} \geq 0$,

(ii) $\bar{\lambda}^i_{a,k} \cdot f^i_k(\bar{\mathbf{x}}, \bar{\mathbf{p}}) = 0,\ \bar{\lambda}^i_{b,k} \cdot f^{d\ell+i}_k(\bar{\mathbf{x}}, \bar{\mathbf{p}}) = 0,\ \bar{\lambda}_{p,i} \cdot g_i(\bar{\mathbf{x}}, \bar{\mathbf{p}}) = 0,\ \bar{\mu} \cdot g_{\ell+1}(\bar{\mathbf{x}}, \bar{\mathbf{p}}) = 0$ *für* $i \in I_\ell,\ k \in K_d$ *(Bedingungen vom komplementären Schlupf),*

(iii) $\nabla_{\mathbf{x},\mathbf{p}}\, \mathcal{L}(\bar{\mathbf{x}}, \bar{\mathbf{p}}, \bar{\boldsymbol{\lambda}}, \bar{\mu}) = 0$, *wobei*

$$\begin{aligned} \mathcal{L}(\mathbf{x}, \mathbf{p}, \boldsymbol{\lambda}, \mu) \ &:= \ -\ln|M(\mathbf{x}, \mathbf{p})| + \sum_{i=1}^{\ell}\sum_{k=1}^{d}\Big(\lambda^i_{a,k}\, f^i_k(\mathbf{x}, \mathbf{p}) + \lambda^i_{b,k}\, f^{d\ell+i}_k(\mathbf{x}, \mathbf{p})\Big) \\ &\quad + \sum_{i=1}^{\ell} \lambda_{p,i}\, g_i(\mathbf{x}, \mathbf{p}) + \mu\, g_{\ell+1}(\mathbf{x}, \mathbf{p}) \end{aligned} \tag{4.7}$$

die Lagrangefunktion von (4.2) darstellt.

4.1.2 Existenz einer Lösung

Satz 4.1.3. *Unter der Voraussetzung (4.1) besitzt das Optimierungsproblem (4.2) eine globale Lösung* $(\mathbf{x}^*, \mathbf{p}^*) \in \mathcal{S}$.

Beweis:

- Mit (4.1) sind die Sensitivitäten $\partial_{\theta_i} u(t, \mathbf{x}, \boldsymbol{\theta})$ für jedes feste $\boldsymbol{\theta} \in \Theta$ in $\mathbf{x}$ stetig. Da die Komposition von stetigen Funktionen ebenfalls stetig ist, folgt die Stetigkeit der Fisher-Informationsmatrix $M(\mathbf{x}, \mathbf{p})$ aus (4.3) bezüglich $\mathbf{x}$.
- Da $M(\mathbf{x}, \mathbf{p})$ linear von $\mathbf{p}$ abhängt, folgt die Stetigkeit von $\mathbf{p} \mapsto M(\mathbf{x}, \mathbf{p})$.

Da die Determinante eine stetige Abbildung darstellt, folgt die Stetigkeit von $M(\mathbf{x}, \mathbf{p}) \mapsto J(\mathbf{x}, \mathbf{p}) = -\ln|M(\mathbf{x}, \mathbf{p})|$. Da der zulässige Bereich $\mathcal{S}$ kompakt ist, folgt die Beschränktheit des Zielfunktionals $J(\mathbf{x}, \mathbf{p})$ auf $\mathcal{S}$, so dass $\min\limits_{\mathbf{x},\mathbf{p}} J(\mathbf{x}, \mathbf{p}) = C_J > -\infty$ mit einer Konstanten $C_J \in \mathbb{R}$ gilt.

□

4.1.3 Optimalitätsbedingungen erster Ordnung

Die Gleichungen $\nabla_{\mathbf{x},\mathbf{p}}\, \mathcal{L}(\bar{\mathbf{x}}, \bar{\mathbf{p}}, \bar{\boldsymbol{\lambda}}, \bar{\mu}) = 0$ und die Bedingungen vom komplementären Schlupf aus Satz 4.1.2 stellen mit $(\bar{\mathbf{x}}, \bar{\mathbf{p}}, \bar{\boldsymbol{\lambda}}, \bar{\mu}) \in \mathcal{W}$ die Optimalitätsbedingungen erster Ordnung dar und werden KKT-Bedingungen genannt, wobei

$$\mathcal{W} \ := \ \bar{\Omega}^\ell \times [0,1]^\ell \times \mathbb{R}_+^{(2d+1)\ell} \times \mathbb{R} \tag{4.8}$$

den Raum der zulässigen KKT-Punkte darstellt. Mit den Gradienten aus (4.6) lauten die KKT-Bedingungen für alle $i \in I_\ell,\ k \in K_d$ und dem euklidischen Skalarprodukt $\langle \cdot, \cdot \rangle$

$$\left.\begin{aligned} \partial_{x^i_k} \mathcal{L}(\bar{\mathbf{x}}, \bar{\mathbf{p}}, \bar{\boldsymbol{\lambda}}, \bar{\mu}) \ &= \ \partial_{x^i_k} J(\bar{\mathbf{x}}, \bar{\mathbf{p}}) - \bar{\lambda}^i_{a,k} + \bar{\lambda}^i_{b,k} \ = \ 0, \\ \partial_{p_i} \mathcal{L}(\bar{\mathbf{x}}, \bar{\mathbf{p}}, \bar{\boldsymbol{\lambda}}, \bar{\mu}) \ &= \ \partial_{p_i} J(\bar{\mathbf{x}}, \bar{\mathbf{p}}) - \bar{\lambda}_{p,i} + \bar{\mu} \ = \ 0, \\ \big\langle (\bar{\lambda}^i_{a,k}, \bar{\lambda}^i_{b,k}, \bar{\lambda}_{p,i}), (a_k - \bar{x}^i_k, \bar{x}^i_k - b_k, -\bar{p}_i) \big\rangle \ &= \ 0, \\ a_k \leq \bar{x}^i_k \leq b_k, \quad \bar{p}_i \geq 0, \quad \textstyle\sum_{i=1}^{\ell} \bar{p}_i = 1, \quad \bar{\lambda}^i_{a,k}, \bar{\lambda}^i_{b,k}, \bar{\lambda}_{p,i} \geq 0, \quad \bar{\mu} \in \mathbb{R}. \end{aligned}\right\} \tag{4.9}$$

Die im KKT-System (4.9) auftretenden partiellen Ableitungen $\partial_{x_k^i} J(\mathbf{x}, \mathbf{p})$ und $\partial_{p_i} J(\mathbf{x}, \mathbf{p})$ werden mithilfe der Kettenregel wie folgt bestimmt.

Satz 4.1.4. *Unter der Voraussetzung (4.1) ist*

$$\textit{(i)}\ \partial_{x_k^i} J(\mathbf{x}, \mathbf{p}) \ = \ -2p_i \int_T \partial_{x_k^i} \left[\nabla_{\boldsymbol{\theta}} u(t, \mathbf{x}^i, \boldsymbol{\theta})\right]^\top M(\mathbf{x}, \mathbf{p})^{-1} \nabla_{\boldsymbol{\theta}} u(t, \mathbf{x}^i, \boldsymbol{\theta}) dt,$$

$$\textit{(ii)}\ \partial_{p_i} J(\mathbf{x}, \mathbf{p}) \ = \ -\int_T \nabla_{\boldsymbol{\theta}} u(t, \mathbf{x}^i, \boldsymbol{\theta})^\top M(\mathbf{x}, \mathbf{p})^{-1} \nabla_{\boldsymbol{\theta}} u(t, \mathbf{x}^i, \boldsymbol{\theta}) dt,$$

$\forall\ i \in I_\ell,\ k \in K_d.$

Beweis:
Für den Beweis sei $w(t, \mathbf{x}^i) := \nabla_{\boldsymbol{\theta}} u(t, \mathbf{x}^i, \boldsymbol{\theta})$, wobei $\boldsymbol{\theta} \in \Theta$ als fest gewählt wird.

(i) Es ist

$$\partial_{x_k^i} J(\mathbf{x}, \mathbf{p}) \overset{(4.2)}{=} -\partial_{x_k^i} \ln |M(\mathbf{x}, \mathbf{p})| \overset{[85]}{=} -\mathsf{tr}\left\{M(\mathbf{x}, \mathbf{p})^{-1}\, \partial_{x_k^i} M(\mathbf{x}, \mathbf{p})\right\},$$

$\forall\, i \in I_\ell$ und $k \in K_d$. Mit $M(\mathbf{x}, \mathbf{p}) \in \mathbb{R}^{m \times m}$ ist $\partial_{x_k^i} M(\mathbf{x}, \mathbf{p}) \in \mathbb{R}^{m \times m}$. Dabei gilt für $r, s = 1, \ldots, m$

$$\left[\partial_{x_k^i} M(\mathbf{x}, \mathbf{p})\right]_{r,s} \ = \ p_i \left(\int_T \partial_{x_k^i} w_r(t, \mathbf{x}^i)\, w_s(t, \mathbf{x}^i) \ + \ w_r(t, \mathbf{x}^i)\, \partial_{x_k^i} w_s(t, \mathbf{x}^i)\, dt\right),$$

also

$$\partial_{x_k^i} M(\mathbf{x}, \mathbf{p}) \ = \ p_i \left(\int_T \partial_{x_k^i} w(t, \mathbf{x}^i)\, w(t, \mathbf{x}^i)^\top + w(t, \mathbf{x}^i)\, \partial_{x_k^i} w(t, \mathbf{x}^i)^\top\, dt\right). \quad (4.10)$$

Somit ist

$$\begin{aligned} \partial_{x_k^i} J(\mathbf{x}, \mathbf{p}) \ = \ & -p_i \left(\int_T \partial_{x_k^i} w(t, \mathbf{x}^i)^\top\, M(\mathbf{x}, \mathbf{p})^{-1}\, w(t, \mathbf{x}^i)\, dt \right. \\ & \left. + \int_T w(t, \mathbf{x}^i)^\top\, M(\mathbf{x}, \mathbf{p})^{-1}\, \partial_{x_k^i} w(t, \mathbf{x}^i)\, dt\right), \end{aligned}$$

$\forall\, i \in I_\ell$ und $k \in K_d$.

(ii) Es ist

$$\begin{aligned} \partial_{p_i} J(\mathbf{x}, \mathbf{p}) \ & \overset{(4.2)}{=} \ -\partial_{p_i} \ln |M(\mathbf{x}, \mathbf{p})| \overset{[85]}{=} -\mathsf{tr}\left\{M(\mathbf{x}, \mathbf{p})^{-1}\, \partial_{p_i} M(\mathbf{x}, \mathbf{p})\right\} \\ & = \ -\mathsf{tr}\left\{M(\mathbf{x}, \mathbf{p})^{-1} \int_T w(t, \mathbf{x}^i)\, w(t, \mathbf{x}^i)^T\, dt\right\} \\ & = \ -\int_T w(t, \mathbf{x}^i)^\top\, M(\mathbf{x}, \mathbf{p})^{-1}\, w(t, \mathbf{x}^i)\, dt \end{aligned}$$

$\forall\, i \in I_\ell.$ □

4.1.4 Optimalitätsbedingungen zweiter Ordnung

Da die Konvexität des Optimierungsproblems (4.2) im allgemeinen nicht erfüllt ist, sind die Optimalitätsbedingungen erster Ordnung notwendig, aber nicht hinreichend: Es ist zwar jeder lokale Minimalpunkt vom Optimierungsproblem (4.2) ein KKT-Punkt, es können aber KKT-Punkte auftreten, die ein lokales Maximum oder einen Sattelpunkt darstellen. Mit der Optimalitätsbedingung zweiter Ordnung erhält man notwendige und hinreichende Bedingungen für ein lokales Minimum. Diese werden im folgenden Abschnitt erläutert. Da die Existenz der zweiten Ableitung des Zielfunktionals nachfolgend gefordert wird, wird nun vorausgesetzt, dass für jedes feste $\boldsymbol{\theta} \in \Theta$ die Sensitivitäten $\nabla_{\boldsymbol{\theta}} u(t, \mathbf{x}, \boldsymbol{\theta})$ in der Ortsvariablen $\mathbf{x}$ zweimal stetig differenzierbar sind, d. h.

$$\nabla_{\boldsymbol{\theta}} u(t, \mathbf{x}, \boldsymbol{\theta}) \in \mathcal{C}^1(T; \mathcal{C}^2(\Omega; \mathbb{R}^m)). \tag{4.11}$$

Folgender Satz gibt die notwendigen Bedingungen zweiter Ordnung für das Optimierungsproblem (4.2) wieder.

Satz 4.1.5. (Notwendige Bedingungen zweiter Ordnung, [43])
Unter der Voraussetzung (4.11) sei $(\bar{\mathbf{x}}, \bar{\mathbf{p}}) \in \mathcal{S}$ ein lokales Minimum von Problem (4.2) und $(\bar{\mathbf{x}}, \bar{\mathbf{p}}, \bar{\boldsymbol{\lambda}}, \bar{\mu}) \in \mathcal{W}$ ein KKT-Punkt von (4.9). Dann gilt

$$\mathbf{s}^\top \nabla^2_{\mathbf{x},\mathbf{p}} \mathcal{L}(\bar{\mathbf{x}}, \bar{\mathbf{p}}, \bar{\boldsymbol{\lambda}}, \bar{\mu})\, \mathbf{s} \;\geq\; 0 \quad \forall\, \mathbf{s} \in T(\mathcal{S}_1; (\bar{\mathbf{x}}, \bar{\mathbf{p}})),$$

wobei $T(\mathcal{S}_1; (\bar{\mathbf{x}}, \bar{\mathbf{p}}))$ den Tangentialkegel [43] von $\mathcal{S}_1$ in $(\bar{\mathbf{x}}, \bar{\mathbf{p}})$ darstellt mit

$$\begin{aligned} \mathcal{S}_1 \;:=\; \{(\mathbf{x}, \mathbf{p}) \in \mathcal{S} :\; & f_k^i(\mathbf{x}, \mathbf{p}) = 0 \;\; \forall\, i \in I_\ell \text{ und } k \in K_d \text{ mit } \lambda^i_{a,k} > 0, \\ & f_k^{d\ell+i}(\mathbf{x}, \mathbf{p}) = 0 \;\; \forall\, i \in I_\ell \text{ und } k \in K_d \text{ mit } \lambda^i_{b,k} > 0, \\ & g_i(\mathbf{x}, \mathbf{p}) = 0 \;\; \forall\, i \in I_\ell \text{ mit } \lambda_{p,i} > 0\}. \end{aligned}$$

Bemerkung 4.1.6. *Die Menge $\mathcal{S}_1$ beschreibt die Punkte $(\mathbf{x}, \mathbf{p}) \in \mathcal{S}$, für die*

$$\mathcal{L}(\mathbf{x}, \mathbf{p}, \bar{\boldsymbol{\lambda}}, \bar{\mu}) = J(\mathbf{x}, \mathbf{p})$$

gilt, da mit $i \in I_\ell$ und $k \in K_d$ entweder

$$\left\{ \begin{matrix} \bar{\lambda}^i_{a,k} = 0 \\ \bar{\lambda}^i_{b,k} = 0 \\ \bar{\lambda}_{p,i} = 0 \end{matrix} \right\} \text{ oder } \left\{ \begin{matrix} f_i^k(\mathbf{x}, \mathbf{p}) = 0 \\ f_k^{d\ell+i}(\mathbf{x}, \mathbf{p}) = 0 \\ g_i(\mathbf{x}, \mathbf{p}) = 0 \end{matrix} \right\}$$

gilt.

Da die Nebenbedingungen des Optimierungsproblems (4.2) affin sind, ist der Tangentialkegel $T(\mathcal{S}_1; (\bar{\mathbf{x}}, \bar{\mathbf{p}}))$ gleich dem linearisierten Kegel $L(\mathcal{S}_1; (\bar{\mathbf{x}}, \bar{\mathbf{p}}))$ [43] von $\mathcal{S}_1$ in $(\bar{\mathbf{x}}, \bar{\mathbf{p}})$. Da der linearisierte Kegel $L(\mathcal{S}_1; (\bar{\mathbf{x}}, \bar{\mathbf{p}}))$ ein Polyeder darstellt, ist dieser konvex, wenn $(\bar{\mathbf{x}}, \bar{\mathbf{p}})$ ein lokales Minimum von (4.2) darstellt. Die Hessematrix ist dann in $(\bar{\mathbf{x}}, \bar{\mathbf{p}})$ auf $\mathcal{S}_1$ positiv semidefinit.

Die hinreichenden Bedingungen zweiter Ordnung werden durch den folgenden Satz wiedergegeben.

Satz 4.1.7. (Hinreichende Bedingungen zweiter Ordnung, [43])
Unter der Voraussetzung (4.11) sei $(\bar{\mathbf{x}}, \bar{\mathbf{p}}, \bar{\boldsymbol{\lambda}}, \bar{\mu}) \in \mathcal{W}$ *ein KKT-Punkt von (4.9) und es gelte*

$$\mathbf{s}^\top \nabla^2_{\mathbf{x},\mathbf{p}} \mathcal{L}(\bar{\mathbf{x}}, \bar{\mathbf{p}}, \bar{\boldsymbol{\lambda}}, \bar{\mu})\, \mathbf{s} \;>\; 0 \quad \forall\, \mathbf{s} \in T(\mathcal{S}_1; (\bar{\mathbf{x}}, \bar{\mathbf{p}})) \text{ mit } \mathbf{s} \neq 0. \tag{4.12}$$

Dann ist $(\bar{\mathbf{x}}, \bar{\mathbf{p}}) \in \mathcal{S}$ *ein strikt lokales Minimum von (4.2).*

Der primale Anteil des KKT-Punktes $(\bar{\mathbf{x}}, \bar{\mathbf{p}}, \bar{\boldsymbol{\lambda}}, \bar{\mu}) \in \mathcal{W}$ ist ein strikt lokales Minimum $(\bar{\mathbf{x}}, \bar{\mathbf{p}}) \in S$, wenn die Hessematrix $\nabla^2_{\mathbf{x},\mathbf{p}} \mathcal{L}(\bar{\mathbf{x}}, \bar{\mathbf{p}}, \bar{\boldsymbol{\lambda}}, \bar{\mu})$ auf $\mathcal{W}_1$ positiv definit ist.

In den folgenden beiden Abschnitten wird beschrieben, wie man einen KKT-Punkt des Optimierungsproblems (4.2) durch numerisches Lösen des KKT-Systems (4.9) bestimmen kann, wenn (4.11) erfüllt ist und somit die Sensitivitäten bezüglich des Ortes zweimal stetig differenzierbar sind.

Das System (4.9) enthält sowohl Gleichungen, als auch Ungleichungen. Wären nur Gleichungen gegeben, könnte man dieses System als ein Nullstellenproblem auffassen und zum Beispiel mit der Newtonmethode lösen. Da dieses aber nicht der Fall ist, werden nun zwei Möglichkeiten vorgestellt, mit denen ein KKT-Punkt des Systems (4.9) bestimmt werden kann. Diese beiden Verfahren lauten

- Primal-duales Innere-Punkte-Verfahren,
- Active-Set-Methode.

4.2 Innere-Punkte-Verfahren

Eine Methode mit der ein lokales Minimum des Optimierungsproblems (4.2) unter der Voraussetzung (4.11) bestimmt werden kann, ist das *primal-duale Innere-Punkte-Verfahren* [43], [88]. In diesem Abschnitt wird ein solches Verfahren hergeleitet, welches global linear gegen eine lokale Minimallösung von (4.2) konvergiert.

Ein Innere-Punkte-Verfahren wird dadurch charakterisiert, dass eine Optimallösung ausgehend vom strikten Inneren des zulässigen Bereichs gesucht wird. Um dieses im folgenden zu gewährleisten, wird das sogenannte Logarithmische-Barriere-Verfahren verwendet. Es handelt sich hierbei um ein Innere-Punkte-Verfahren, bei dem das Zielfunktional $J(\mathbf{x}, \mathbf{p}) = -\ln|M(\mathbf{x}, \mathbf{p})|$ durch gewichtete Strafterme erweitert wird. Anstelle des Ursprungsproblems (4.2) wird eine Folge von Barriereproblemen der Form

$$(BP_\alpha) \qquad \min_{\mathbf{x},\mathbf{p}} J_\alpha(\mathbf{x}, \mathbf{p}) := J(\mathbf{x}, \mathbf{p}) + \alpha \cdot B(\mathbf{x}, \mathbf{p}) \quad \text{u.d.N.} \quad \sum_{i=1}^{\ell} p_i = 1,$$

gelöst, wobei

$$B(\mathbf{x}, \mathbf{p}) = -\sum_{k=1}^{d}\sum_{i=1}^{\ell} \ln(x_k^i - a_k) \;-\; \sum_{k=1}^{d}\sum_{i=1}^{\ell} \ln(b_k - x_k^i) \;-\; \sum_{i=1}^{\ell} \ln(p_i)$$

die logarithmische Barrierefunktion darstellt, $\alpha \in \mathbb{R}$, $\alpha > 0$. Zusammengefasst erhält

man die durch $\alpha > 0$ parametrisierte Schar von Optimierungsproblemen

$$\begin{aligned} \min_{\mathbf{x},\mathbf{p}} J_\alpha(\mathbf{x},\mathbf{p}) \quad &:= \quad -\ln|M(\mathbf{x},\mathbf{p})| \\ &\quad -\alpha\cdot\left(\sum_{k=1}^{d}\sum_{i=1}^{\ell}\ln(x_k^i-a_k)+\sum_{k=1}^{d}\sum_{i=1}^{\ell}\ln(b_k-x_k^i)+\sum_{i=1}^{\ell}\ln(p_i)\right) \end{aligned} \tag{4.13}$$

u.d.N. $\displaystyle\sum_{i=1}^{\ell} p_i = 1, \quad \mathbf{a} < \mathbf{x}^i < \mathbf{b}, \quad p_i > 0 \quad \forall\, i \in I_\ell,$

die für $\alpha \to 0$ gegen das Ursprungsproblem (4.2) konvergiert.

Bemerkung 4.2.1. *Der zulässige Bereich der Barriereprobleme (4.13) entspricht für beliebige $\alpha > 0$ dem strikten Inneren von $\mathcal{S}$ aus (4.5) und wird im folgenden durch die Menge*

$$\mathcal{S}^0 \quad := \quad \{(\mathbf{x},\mathbf{p}) \in \mathcal{S} : \mathbf{a} < \mathbf{x}^i < \mathbf{b},\, p_i > 0,\, \sum_{i=1}^{\ell} p_i = 1\} \tag{4.14}$$

beschrieben. Barriere-Verfahren bestrafen die $(\mathbf{x},\mathbf{p}) \in \mathcal{S}^0$, die sich dem Rand $\partial\mathcal{S}^0$ nähern. Daher sind für jedes $\alpha > 0$ alle Lösungen von (4.13) strikt innere Punkte des zulässigen Bereichs $\mathcal{S}$ aus (4.5).

Folgender Satz besagt, unter welcher Voraussetzung die Optimierungsprobleme (4.13) eine globale Minimallösung besitzen.

Satz 4.2.2. *Unter der Voraussetzung (4.11) besitzt das Optimierungsproblem (4.13) für jedes $\alpha > 0$ eine globale Minimallösung.*

Beweis:
Nach Satz 4.1.3 besitzt das Optimierungsproblem (4.2) unter der Voraussetzung (4.1) eine globale Lösung $(\mathbf{x}^*,\mathbf{p}^*) \in \mathcal{S}$, folglich auch unter der Voraussetzung (4.11). Somit existiert eine Konstante $C_J \in \mathbb{R}$, so dass $\min\limits_{\mathbf{x},\mathbf{p}} J(\mathbf{x},\mathbf{p}) = C_J > -\infty$.

Das Problem (4.13) besitzt genau dann eine globale Minimallösung, wenn die Barriereanteile der logarithmischen Barrierefunktion $B(\mathbf{x},\mathbf{p})$ ebenfalls nach unten beschränkt sind:
Mit $\mathbf{a} < \mathbf{x}^i < \mathbf{b}$ ist $a_k < x_k^i < b_k$ für alle $i \in I_\ell$ und $k \in K_d$. Daraus folgt, dass

$$\ln(x_k^i - a_k) < \ln(b_k - a_k) \le \ln(\|\mathbf{b}-\mathbf{a}\|_\infty)$$

und

$$\ln(b_k - x_k^i) < \ln(b_k - a_k) \le \ln(\|\mathbf{b}-\mathbf{a}\|_\infty).$$

Somit gilt

$$-\alpha\sum_{i=1}^{\ell}\sum_{k=1}^{d}\Big(\ln(x_k^i-a_k)-\ln(b_k-x_k^i)\Big) \;>\; -2\alpha\ell d\ln(\|\mathbf{b}-\mathbf{a}\|_\infty).$$

Mit $p_i \in (0,1)$ ist zudem $\ln(p_i) < \ln(1) = 0$ für alle $i \in I_\ell$. Daraus folgt

$$-\alpha\sum_{i=1}^{\ell}\ln(p_i) \;>\; 0.$$

Somit ist

$$J_\alpha(\mathbf{x},\mathbf{p}) \;\geq\; J(\mathbf{x},\mathbf{p}) - 2\ell d\alpha \ln(\|\mathbf{b}-\mathbf{a}\|_\infty) \;=\; -C_\alpha \;>\; -\infty$$

mit einer Konstanten $C_\alpha > 0$. Das Zielfunktional $J_\alpha(\mathbf{x},\mathbf{p})$ ist demnach nach unten beschränkt. □

Regularität in einer lokalen Lösung

Bei einem primal-dualen Innere-Punkte-Verfahren werden zur Bestimmung einer lokalen Lösung die zu den Barriereproblemen (4.13) zugehörigen KKT-Systeme gelöst. Um die Existenz eines KKT-Punktes zu gewährleisten muss zunächst überprüft werden, ob für jedes $\alpha > 0$ die Barriereprobleme in einem lokalen Minimum regulär sind. Wie in Kapitel 4.1.1 kann auch hier die Regularität mit der Bedingung von Robinson überprüft werden. Da $\mathcal{S}^0 \subset \mathcal{S}$ mit $\mathcal{S}$ aus (4.5), kann analog zu Kapitel 4.1.1 gezeigt werden, dass die Barriereprobleme (4.13) für jedes $\alpha > 0$ in jedem Punkt $(\mathbf{x},\mathbf{p}) \in \mathcal{S}^0$ regulär sind, insbesondere also in jedem lokalen Minimum $(\bar{\mathbf{x}},\bar{\mathbf{p}})$ [43].

Optimalitätsbedingungen erster Ordnung

Da die Regularitätsbedingung von Robinson erfüllt ist, kann der Satz von Kuhn-Tucker angewendet werden. Er besagt:

Satz 4.2.3. (Satz von Kuhn-Tucker für das Optimierungsproblem (4.13), [43])
Unter der Voraussetzung (4.1) sei $(\bar{\mathbf{x}},\bar{\mathbf{p}}) \in \mathcal{S}^0$ für festes $\alpha > 0$ eine lokale Optimallösung von (4.13). Da das Problem (4.13) in $(\bar{\mathbf{x}},\bar{\mathbf{p}})$ regulär ist, existiert ein $\bar{\mu} \in \mathbb{R}$, so dass

(i) $\bar{\mu} \cdot \left(\sum_{i=1}^{\ell} \bar{p}_i - 1 \right) = 0$ *(Bedingung vom komplementären Schlupf),*

(ii) $\nabla_{\mathbf{x},\mathbf{p}} \mathcal{L}_\alpha(\bar{\mathbf{x}},\bar{\mathbf{p}},\bar{\mu}) = 0$, *wobei*

$$\begin{aligned}\mathcal{L}_\alpha(\mathbf{x},\mathbf{p},\mu) \;:=\; & -\ln|M(\mathbf{x},\mathbf{p})| \\ & -\alpha \cdot \left(\sum_{k=1}^{d}\sum_{i=1}^{\ell} \ln(x_k^i - a_k) + \sum_{k=1}^{d}\sum_{i=1}^{\ell} \ln(b_k - x_k^i) + \sum_{i=1}^{\ell} \ln(p_i) \right) \\ & +\mu \cdot \left(\sum_{i=1}^{\ell} p_i - 1 \right)\end{aligned}$$

die Lagrangefunktion von (4.13) darstellt.

Die Gleichungen $\nabla_{\mathbf{x},\mathbf{p}} \mathcal{L}_\alpha(\bar{\mathbf{x}},\bar{\mathbf{p}},\bar{\mu}) = 0$ und die Bedingung vom komplementären Schlupf aus Satz 4.2.3 stellen mit $(\bar{\mathbf{x}},\bar{\mathbf{p}},\bar{\mu}) \in \mathcal{S}^0 \times \mathbb{R}$ die zu den Barriereproblemen (4.13) zugehörigen KKT-Bedingungen dar. Ausformuliert lauten diese

$$\left.\begin{aligned} \partial_{x_k^i}\mathcal{L}_\alpha(\bar{\mathbf{x}},\bar{\mathbf{p}},\bar{\mu}) &= \partial_{x_k^i} J(\bar{\mathbf{x}},\bar{\mathbf{p}}) - \frac{\alpha}{\bar{x}_k^i - a_k} + \frac{\alpha}{b_k - \bar{x}_k^i} &&= 0, \\ \partial_{p_i}\mathcal{L}_\alpha(\bar{\mathbf{x}},\bar{\mathbf{p}},\bar{\mu}) &= \partial_{p_i} J(\bar{\mathbf{x}},\bar{\mathbf{p}}) - \frac{\alpha}{\bar{p}_i} + \bar{\mu} &&= 0, \\ & \sum_{i=1}^{\ell} \bar{p}_i = 1, \quad a_k < \bar{x}_k^i < b_k, \qquad \bar{p}_i &&> 0. \end{aligned}\right\} \tag{4.15}$$

Eine Lösung $(\bar{\mathbf{x}}, \bar{\mathbf{p}}, \bar{\mu}) \in \mathcal{S}^0 \times \mathbb{R}$ des KKT-Systems (4.15) wird KKT-Punkt genannt.

Durch Einsetzen der Variablen

$$\bar{z}^i_{a,k} := \frac{\alpha}{\bar{x}^i_k - a_k}, \quad \bar{z}^i_{b,k} := \frac{\alpha}{b_k - \bar{x}^i_k}, \quad \bar{z}_{p,i} := \frac{\alpha}{\bar{p}_i}$$

für $i \in I_\ell$ und $k \in K_d$ in (4.15) erhält man mit

$$\bar{\mathbf{z}}^i_{\mathbf{a}} = (\bar{z}^i_{a,1}, \ldots, \bar{z}^i_{a,d})^\top, \quad \bar{\mathbf{z}}^i_{\mathbf{b}} = (\bar{z}^i_{b,1}, \ldots, \bar{z}^i_{b,d})^\top$$

und

$$\bar{\mathbf{z}}_{\mathbf{a}} = \begin{pmatrix} \bar{\mathbf{z}}^1_{\mathbf{a}} \\ \vdots \\ \bar{\mathbf{z}}^\ell_{\mathbf{a}} \end{pmatrix}, \quad \bar{\mathbf{z}}_{\mathbf{b}} = \begin{pmatrix} \bar{\mathbf{z}}^1_{\mathbf{b}} \\ \vdots \\ \bar{\mathbf{z}}^\ell_{\mathbf{b}} \end{pmatrix}, \quad \bar{\mathbf{z}}^i_{\mathbf{a}}, \bar{\mathbf{z}}^i_{\mathbf{b}} \in \mathbb{R}^d_+, \quad \bar{\mathbf{z}}_{\mathbf{p}} = \begin{pmatrix} \bar{z}_{p,1} \\ \vdots \\ \bar{z}_{p,\ell} \end{pmatrix}, \quad \bar{z}_{p,i} \in \mathbb{R}_+$$

sowie $\bar{\mathbf{z}} = (\bar{\mathbf{z}}^\top_{\mathbf{a}}, \bar{\mathbf{z}}^\top_{\mathbf{b}}, \bar{\mathbf{z}}^\top_{\mathbf{p}})^\top$ die sogenannten *gestörten KKT-Bedingungen*: Für alle $i \in I_\ell$ und $k \in K_d$ gilt

$$\left.\begin{aligned}
&\partial_{x^i_k} \mathfrak{L}(\bar{\mathbf{x}}, \bar{\mathbf{p}}, \bar{z}, \bar{\mu}) \;=\; \partial_{x^i_k} J(\bar{\mathbf{x}}, \bar{\mathbf{p}}) - \bar{z}^i_{a,k} + \bar{z}^i_{b,k} \;=\; 0, \\
&\partial_{p_i} \mathfrak{L}(\bar{\mathbf{x}}, \bar{\mathbf{p}}, \bar{z}, \bar{\mu}) \;=\; \partial_{p_i} J(\bar{\mathbf{x}}, \bar{\mathbf{p}}) - \bar{z}_{p,i} + \bar{\mu} \;=\; 0, \\
&\sum_{i=1}^{\ell} \bar{p}_i = 1, \quad a_k < \bar{x}^i_k < b_k, \quad \bar{p}_i > 0, \\
&(\bar{x}^i_k - a_k)\bar{z}^i_{a,k} = \alpha, \quad (b_k - \bar{x}^i_k)\bar{z}^i_{b,k} = \alpha, \quad \bar{p}_i \bar{z}_{p,i} = \alpha, \\
&\bar{z}^i_{a,k} > 0, \quad \bar{z}^i_{b,k} > 0, \quad \bar{z}_{p,i} > 0,
\end{aligned}\right\} \tag{4.16}$$

wobei der primale Anteil $(\bar{\mathbf{x}}, \bar{\mathbf{p}}) \in \mathcal{S}^0$ entweder ein lokales Minimum, lokales Maximum oder einen Sattelpunkt von (4.13) darstellt. Für eine Lösung von (4.16) gilt $(\bar{\mathbf{x}}, \bar{\mathbf{p}}, \bar{\mathbf{z}}, \bar{\mu}) \in \mathcal{W}^0$, wobei $\mathcal{W}^0$ das strikte Innere der Menge $\mathcal{W}$ aus (4.8) darstellt mit

$$\mathcal{W}^0 \;:=\; \Omega^\ell \times (0,1)^\ell \times \mathbb{R}_+^{(2d+1)\ell} \times \mathbb{R}. \tag{4.17}$$

Definition 4.2.1. *Die durch $\alpha > 0$ parametrisierte Lösungsmenge*

$$\bar{\mathbf{w}}(\alpha) \;=\; (\bar{\mathbf{x}}(\alpha), \bar{\mathbf{p}}(\alpha), \bar{\mathbf{z}}(\alpha), \bar{\mu}(\alpha)) \in \mathcal{W}_0 \tag{4.18}$$

von (4.16) heißt zentraler Pfad.

Stetigkeit des zentralen Pfades

Bei einem Innere-Punkte-Verfahren wird die Lösung eines Optimierungsproblems aus dem strikten Inneren des zulässigen Bereichs $\mathcal{S}$ heraus iterativ bestimmt. Zur Realisierung dessen löst man eine Folge von Optimierungsproblemen der Gestalt (4.13), wobei $\alpha > 0$ einen Barriereparameter darstellt. Je größer α ist, umso weiter liegt die Lösung $\mathbf{w}(\alpha) \in \mathcal{W}^0$ vom Rand $\partial\mathcal{S}$ entfernt. Um zu gewährleisten, dass die durch $\alpha > 0$ parametrisierten Lösungen $\mathbf{w}(\alpha) \in \mathcal{W}^0$ von (4.13) für $\alpha \to 0$ gegen eine Lösung des Ursprungsproblems (4.2) konvergieren, wird die Stetigkeit des zentralen Pfades (4.18) gefordert.

Satz 4.2.4. (Stetigkeit des zentralen Pfades)
Unter der Voraussetzung (4.1) sei $(\bar{\mathbf{x}}(\alpha), \bar{\mathbf{p}}(\alpha), \bar{\mathbf{z}}(\alpha), \bar{\mu}(\alpha)) \in \mathcal{W}_0$ *eine Lösung von (4.16). Wenn* $(\bar{\mathbf{x}}(\alpha), \bar{\mathbf{p}}(\alpha)) \to (\bar{\mathbf{x}}(\tilde{\alpha}), \bar{\mathbf{p}}(\tilde{\alpha}))$ *für* $\alpha \to \tilde{\alpha}$*, dann*

$$(\bar{\mathbf{x}}(\alpha), \bar{\mathbf{p}}(\alpha), \bar{\mathbf{z}}(\alpha), \bar{\mu}(\alpha)) \overset{\alpha\to\tilde{\alpha}}{\longrightarrow} (\bar{\mathbf{x}}(\tilde{\alpha}), \bar{\mathbf{p}}(\tilde{\alpha}), \bar{\mathbf{z}}(\tilde{\alpha}), \bar{\mu}(\tilde{\alpha}))$$

in $\mathcal{W}^0$.

Beweis:
Für den Beweis sei $w(t, \mathbf{x}^i) := \nabla_{\boldsymbol{\theta}} u(t, \mathbf{x}^i, \boldsymbol{\theta})$, wobei $\boldsymbol{\theta} \in \Theta$ als fest gewählt vorausgesetzt wird.

(i) Zunächst wird gezeigt, dass mit $(\bar{\mathbf{x}}(\alpha), \bar{\mathbf{p}}(\alpha)) \overset{\alpha\to\tilde{\alpha}}{\longrightarrow} (\bar{\mathbf{x}}(\tilde{\alpha}), \bar{\mathbf{p}}(\tilde{\alpha}))$ für alle $i \in I_\ell$ und $k \in K_d$ gilt

$$\partial_{x_k^i} J(\bar{\mathbf{x}}(\alpha), \bar{\mathbf{p}}(\alpha)) \overset{\alpha\to\tilde{\alpha}}{\longrightarrow} \partial_{x_k^i} J(\bar{\mathbf{x}}(\tilde{\alpha}), \bar{\mathbf{p}}(\tilde{\alpha})). \tag{4.19}$$

Nach Satz 4.1.4 gilt für alle $i \in I_\ell$ und $k \in K_d$

$$\partial_{x_k^i} J(\mathbf{x}, \mathbf{p}) \;=\; -2p_i \int_T \partial_{x_k^i} w(t, \mathbf{x}^i)^\top M(\mathbf{x}, \mathbf{p})^{-1} w(t, \mathbf{x}^i) dt.$$

Da $w(t, \mathbf{x}^i) \in \mathcal{C}^1(T; \mathcal{C}^1(\Omega; \mathbb{R}^m))$ gilt, folgt $\partial_{x_k^i} w(t, \mathbf{x}^i) \in \mathcal{C}^1(T; \mathcal{C}^0(\Omega; \mathbb{R}^m))$. Die Elemente der Fisher-Informationsmatrix $M(\mathbf{x}, \mathbf{p})$ sind in $(\mathbf{x}, \mathbf{p})$ stetig, so dass die Stetigkeit der Minoren von $M(\mathbf{x}, \mathbf{p})$ [27] und somit auch die Stetigkeit der Inversen $M(\mathbf{x}, \mathbf{p})^{-1}$ in $(\mathbf{x}, \mathbf{p})$ gewährleistet ist. Als Komposition von stetigen Funktionen ist somit $\partial_{x_k^i} J(\mathbf{x}, \mathbf{p})$ stetig in $(\mathbf{x}, \mathbf{p})$.
Es gilt für $\alpha, \tilde{\alpha} > 0$ und alle $i \in I_\ell$ und $k \in K_d$

$$\begin{aligned}
0 \;&\overset{(4.16)}{=}\; \lim_{\alpha\to\tilde{\alpha}} \Big(\partial_{x_k^i} J(\bar{\mathbf{x}}(\alpha), \bar{\mathbf{p}}(\alpha)) - z_{a,k}^i(\alpha) + z_{b,k}^i(\alpha) \\
&\qquad\qquad - \partial_{x_k^i} J(\bar{\mathbf{x}}(\tilde{\alpha}), \bar{\mathbf{p}}(\tilde{\alpha})) + z_{a,k}^i(\tilde{\alpha}) - z_{b,k}^i(\tilde{\alpha}) \Big) \\
&\overset{(4.19)}{=}\; \lim_{\alpha\to\tilde{\alpha}} \Big(-z_{a,k}^i(\alpha) + z_{b,k}^i(\alpha) + z_{a,k}^i(\tilde{\alpha}) - z_{b,k}^i(\tilde{\alpha}) \Big) \\
\iff \quad &\lim_{\alpha\to\tilde{\alpha}} z_{a,k}^i(\alpha) - z_{a,k}^i(\tilde{\alpha}) \;=\; \lim_{\alpha\to\tilde{\alpha}} z_{b,k}^i(\alpha) - z_{b,k}^i(\tilde{\alpha})
\end{aligned} \tag{4.20}$$

Mit $a_k < x_k^i(\alpha) < b_k$ ist entweder $x_k^i(\alpha) - a_k > \dfrac{b_k - a_k}{2}$, $b_k - x_k^i(\alpha) \geq \dfrac{b_k - a_k}{2}$ oder $x_k^i(\alpha) - a_k = b_k - x_k^i(\alpha) = \dfrac{b_k - a_k}{2}$:
(1) Ist $x_k^i(\alpha) - a_k > \dfrac{b_k - a_k}{2}$, existiert immer ein $\tilde{\alpha}_0 > 0$, so dass $x_k^i(\tilde{\alpha}) - a_k \geq \dfrac{b_k - a_k}{2}$ für alle $\tilde{\alpha} \in \{\beta > 0 : \; |\alpha - \beta| < |\alpha - \tilde{\alpha}_0|\}$. Dann ist

$$\begin{aligned}
|z_{a,k}^i(\alpha) - z_{a,k}^i(\tilde{\alpha})| \;&\overset{(4.16)}{=}\; \left| \frac{\alpha}{x_k^i(\alpha) - a_k} - \frac{\tilde{\alpha}}{x_k^i(\tilde{\alpha}) - a_k} \right| \\
= \; &\left| \frac{\alpha(x_k^i(\tilde{\alpha}) - a_k) - \tilde{\alpha}(x_k^i(\alpha) - a_k)}{(x_k^i(\alpha) - a_k)(x_k^i(\tilde{\alpha}) - a_k)} \right| \;=\; \left| \frac{\alpha\, x_k^i(\tilde{\alpha}) - \tilde{\alpha}\, x_k^i(\alpha) + a_k(\tilde{\alpha} - \alpha)}{(x_k^i(\alpha) - a_k)(x_k^i(\tilde{\alpha}) - a_k)} \right| \\
\leq \; &\left| \frac{4(\alpha\, x_k^i(\tilde{\alpha}) - \tilde{\alpha}\, x_k^i(\alpha))}{(b_k - a_k)^2} + \frac{4a_k(\tilde{\alpha} - \alpha)}{(b_k - a_k)^2} \right| \;\overset{\tilde{\alpha}\to\alpha}{\longrightarrow}\; 0.
\end{aligned}$$

Mit (4.20) folgt $|z^i_{b,k}(\alpha) - z^i_{b,k}(\tilde{\alpha})| \overset{\tilde{\alpha}\to\alpha}{\longrightarrow} 0$.

(2) Ist $b_k - x^i_k(\alpha) > \dfrac{b_k - a_k}{2}$, existiert immer ein $\tilde{\alpha}_0 > 0$, so dass $b_k - x^i_k(\tilde{\alpha}) \geq \dfrac{b_k - a_k}{2}$ für alle $\tilde{\alpha} > 0$ mit $|\alpha - \tilde{\alpha}| < |\alpha - \tilde{\alpha}_0|$. Dann ist

$$\begin{aligned} |z^i_{b,k}(\alpha) - z^i_{b,k}(\tilde{\alpha})| &\overset{(4.16)}{=} \left| \frac{\alpha}{b_k - x^i_k(\alpha)} - \frac{\tilde{\alpha}}{b_k - x^i_k(\tilde{\alpha})} \right| \\ &\leq \left| \frac{4(\tilde{\alpha}\, x^i_k(\alpha) - \alpha\, x^i_k(\tilde{\alpha}))}{(b_k - a_k)^2} + \frac{4a_k(\alpha - \tilde{\alpha}}{(b_k - a_k)^2} \right| \overset{\tilde{\alpha}\to\alpha}{\longrightarrow} 0. \end{aligned}$$

Mit (4.20) folgt $|z^i_{a,k}(\alpha) - z^i_{a,k}(\tilde{\alpha})| \overset{\tilde{\alpha}\to\alpha}{\longrightarrow} 0$.

(3) Mit $x^i_k(\alpha) - a_k = \dfrac{b_k - a_k}{2}$ ist

$$|z^i_{a,k}(\alpha) - z^i_{a,k}(\tilde{\alpha})| \overset{(4.16)}{=} \left| \frac{2\alpha}{b_k - a_k} - \frac{\tilde{\alpha}}{x^i_k(\tilde{\alpha}) - a_k} \right| \overset{\tilde{\alpha}\to\alpha}{\longrightarrow} 0,$$

da mit $x(\tilde{\alpha}) \overset{\tilde{\alpha}\to\alpha}{\longrightarrow} x(\alpha)$

$$\lim_{\tilde{\alpha}\to\alpha} \frac{\tilde{\alpha}}{x^i_k(\tilde{\alpha}) - a_k} = \frac{\alpha}{x^i_k(\alpha) - a_k} = \frac{2\alpha}{b_k - a_k}$$

gilt. Mit (4.20) folgt $|z^i_{b,k}(\alpha) - z^i_{b,k}(\tilde{\alpha})| \overset{\tilde{\alpha}\to\alpha}{\longrightarrow} 0$.

(ii) Analog zu (4.19) kann gezeigt werden, dass mit $(\bar{\mathbf{x}}(\alpha), \bar{\mathbf{p}}(\alpha)) \overset{\alpha\to\tilde{\alpha}}{\longrightarrow} (\bar{\mathbf{x}}(\tilde{\alpha}), \bar{\mathbf{p}}(\tilde{\alpha}))$ für alle $i \in I_\ell$ gilt

$$\partial_{p_i} J(\bar{\mathbf{x}}(\alpha), \bar{\mathbf{p}}(\alpha)) \overset{\alpha\to\tilde{\alpha}}{\longrightarrow} \partial_{p_i} J(\bar{\mathbf{x}}(\tilde{\alpha}), \bar{\mathbf{p}}(\tilde{\alpha}). \tag{4.21}$$

Mit $\alpha, \tilde{\alpha} > 0$ und alle $i \in I_\ell$ gilt dann

$$\begin{aligned} 0 &\overset{(4.16)}{=} \lim_{\alpha\to\tilde{\alpha}} \Big(\partial_{p_i} J(\bar{\mathbf{x}}(\alpha), \bar{\mathbf{p}}(\alpha)) - z_{p,i}(\alpha) + \mu(\alpha) \\ &\qquad\qquad - \partial_{p_i} J(\bar{\mathbf{x}}(\tilde{\alpha}), \bar{\mathbf{p}}(\tilde{\alpha})) + z_{p,i}(\tilde{\alpha}) - \mu(\tilde{\alpha}) \Big) \\ &\overset{(4.21)}{=} \lim_{\alpha\to\tilde{\alpha}} \Big(-z_{p,i}(\alpha) + \mu(\alpha) + z_{p,i}(\tilde{\alpha}) - \mu(\tilde{\alpha}) \Big) \\ \iff \quad & \lim_{\alpha\to\tilde{\alpha}} z_{p,i}(\alpha) - z_{p,i}(\tilde{\alpha}) = \lim_{\alpha\to\tilde{\alpha}} \mu(\alpha) - \mu(\tilde{\alpha}). \end{aligned} \tag{4.22}$$

Zudem ist

$$|z_{p,i}(\alpha) - z_{p,i}(\tilde{\alpha})| \overset{(4.16)}{=} \left| \frac{\alpha}{p_i(\alpha)} - \frac{\tilde{\alpha}}{p_i(\tilde{\alpha})} \right| = \left| \frac{\alpha\, p_i(\tilde{\alpha}) - \tilde{\alpha}\, p_i(\alpha)}{p_i(\alpha)\, p_i(\tilde{\alpha})} \right| \overset{\tilde{\alpha}\to\alpha}{\longrightarrow} 0.$$

Mit (4.22) folgt $|\mu(\alpha) - \mu(\tilde{\alpha})| \overset{\tilde{\alpha}\to\alpha}{\longrightarrow} 0$. □

Es wird nun gezeigt, dass für jedes fest gewählte $\mathbf{x} \in \bar{\Omega}^\ell$ das Optimierungsproblem (4.2) bezüglich $\mathbf{p}$ konvex ist. Dann kann gezeigt werden, dass aus $\bar{\mathbf{x}}(\alpha) \to \bar{\mathbf{x}}(\tilde{\alpha})$ folgt, dass $\bar{\mathbf{p}}(\alpha) \to \bar{\mathbf{p}}(\tilde{\alpha})$ gilt.

Satz 4.2.5. *Unter der Voraussetzung (4.1) ist das Optimierungsproblem (4.2) für jedes feste* $\tilde{\mathbf{x}} \in \bar{\Omega}^\ell$ *konvex.*

Beweis:
Nach [66] ist die Abbildung $M(\mathbf{x}, \mathbf{p}) \mapsto -\ln|M(\mathbf{x}, \mathbf{p})|$ stetig und konvex.
Sei

$$f(\mathbf{x}^i) := \int_T \nabla_{\boldsymbol{\theta}} u(t, \mathbf{x}^i, \boldsymbol{\theta}) \, \nabla_{\boldsymbol{\theta}} u(t, \mathbf{x}^i, \boldsymbol{\theta})^\top dt,$$

dann gilt für fest gewähltes $\tilde{\mathbf{x}} \in \bar{\Omega}$ und beliebigem $\beta \in [0, 1]$

$$\begin{aligned} M(\tilde{\mathbf{x}}, \beta\,\mathbf{p}^1 + (1-\beta)\,\mathbf{p}^2) &\overset{(4.3)}{=} \sum_{i=1}^{\ell} (\beta\, p_i^1 + (1-\beta)\, p_i^2)\, f(\tilde{\mathbf{x}}^{\mathbf{i}}) \\ = \beta \sum_{i=1}^{\ell} p_i^1\, f(\tilde{\mathbf{x}}^i) + (1-\beta) \sum_{i=1}^{\ell} p_i^2\, f(\tilde{\mathbf{x}}^i) &= \beta\, M(\tilde{\mathbf{x}}, \mathbf{p}^1) + (1-\beta)\, M(\tilde{\mathbf{x}}, \mathbf{p}^2). \end{aligned}$$

Da die Komposition zweier konvexer Funktionen wieder konvex ist, folgt die Konvexität von

$$\mathbf{p} \mapsto J(\tilde{\mathbf{x}}, \mathbf{p}) = -\ln|M(\tilde{\mathbf{x}}, \mathbf{p})|.$$

□

Satz 4.2.6. *Unter der Voraussetzung (4.1) sei* $(\bar{\mathbf{x}}(\alpha), \bar{\mathbf{p}}(\alpha), \bar{\mathbf{z}}(\alpha), \bar{\mu}(\alpha)) \in \mathcal{W}_0$ *eine Lösung von (4.16). Wenn* $\bar{\mathbf{x}}(\alpha) \to \bar{\mathbf{x}}(\tilde{\alpha})$ *für* $\alpha \to \tilde{\alpha}$*, dann*

$$(\bar{\mathbf{x}}(\alpha), \bar{\mathbf{p}}(\alpha), \bar{\mathbf{z}}(\alpha), \bar{\mu}(\alpha)) \overset{\alpha \to \tilde{\alpha}}{\longrightarrow} (\bar{\mathbf{x}}(\tilde{\alpha}), \bar{\mathbf{p}}(\tilde{\alpha}), \bar{\mathbf{z}}(\tilde{\alpha}), \bar{\mu}(\tilde{\alpha}))$$

in $\mathcal{W}^0$.

Beweis:
Nach Satz 4.2.5 ist das Optimierungsproblem (4.2) für jedes fest gewählte $\tilde{\mathbf{x}} \in \bar{\Omega}^\ell$ konvex, d. h. es existiert genau ein $\tilde{\mathbf{p}} \in [0, 1]^\ell$, so dass

$$\tilde{\mathbf{p}} \;=\; \arg\min_{\mathbf{p}} J(\tilde{\mathbf{x}}, \mathbf{p}) \quad \text{u.d.N.} \quad p_i \geq 0 \;\; \forall\, i \in I_\ell, \quad \sum_{j=1}^{\ell} p_j = 1$$

ist.

Die logarithmische Barrierefunktion $B(\mathbf{x}, \mathbf{p})$ aus (4.13) ist für jedes $\alpha > 0$ konvex und somit insbesondere für jedes fest gewählte $\tilde{\mathbf{x}} \in \bar{\Omega}^\ell$. Daraus folgt die Konvexität des Optimierungsproblems

$$\min_p J_\alpha(\tilde{\mathbf{x}}, \mathbf{p}) \;=\; J(\tilde{\mathbf{x}}, \mathbf{p}) + \alpha \cdot B(\tilde{\mathbf{x}}, \mathbf{p}) \quad \text{u.d.N.} \quad p_i > 0 \;\; \forall\, i \in I_\ell, \quad \sum_{j=1}^{\ell} p_j = 1.$$

Somit existiert für jedes $\tilde{\mathbf{x}}(\alpha)$ genau ein $\tilde{\mathbf{p}}(\alpha)$, so dass

$$\tilde{\mathbf{p}}(\alpha) \;=\; \arg\min_{\mathbf{p}} J(\tilde{\mathbf{x}}(\alpha), \mathbf{p}(\alpha)) \quad \text{u.d.N.} \quad p_i(\alpha) > 0 \;\; \forall\, i \in I_\ell, \quad \sum_{j=1}^{\ell} p_j(\alpha) = 1.$$

Aufgrund der Stetigkeit der Bedingungen (4.16) folgt aus $\bar{\mathbf{x}}(\alpha) \overset{\alpha\to\tilde{\alpha}}{\longrightarrow} \bar{\mathbf{x}}(\tilde{\alpha})$, dass $\bar{\mathbf{p}}(\alpha) \overset{\alpha\to\tilde{\alpha}}{\longrightarrow} \bar{\mathbf{p}}(\tilde{\alpha})$. Mit Satz 4.2.4 folgt die Behauptung.

□

Nachdem die Stetigkeit des zentralen Pfades gezeigt wurde, wird nun das primal-duale Newtonsystem aufgestellt mit dem die gestörten KKT-Bedingungen (4.16) gelöst werden können. Anschließend wird beschrieben, wie mithilfe einer Pfadverfolgungsmethode ein lokales Minimum des Optimierungsproblems (4.2) bestimmt werden kann.

4.2.1 Die primal-duale Newtonmethode

Die Barriereterme stellen sicher, dass $\mathbf{a} < \mathbf{x}^i < \mathbf{b}$ und $p_i > 0$ gewährleistet ist. Und da $\alpha > 0$ ist, sind zwangsläufig auch $z^i_{a,k} > 0$, $z^i_{b,k} > 0$ und $z_{p,i} > 0$. Daher sind die Ungleichungen in (4.16) automatisch erfüllt. Um eine Lösung $(\bar{\mathbf{x}}, \bar{\mathbf{p}}, \bar{\mathbf{z}}, \bar{\mu})$ von (4.16) zu finden, werden im folgenden nur noch die Gleichungen betrachtet. Somit kann das Gleichungssystem (4.16) mithilfe einer *primal-dualen Newtonmethode* gelöst werden. Hierfür wird es auf das Nullstellenproblem

$$F_\alpha(\mathbf{w}) = \begin{pmatrix} \nabla_{\mathbf{x}}\,\mathfrak{L}(\mathbf{w}) \\ \nabla_{\mathbf{p}}\,\mathfrak{L}(\mathbf{w}) \\ \sum_{i=1}^{\ell} p_i - 1 \\ {\mathbf{z_a}}^\top I_3\,(\mathbf{x} - \mathbf{u_a}) - \alpha\mathbf{e} \\ {\mathbf{z_b}}^\top I_3\,(\mathbf{u_b} - \mathbf{x}) - \alpha\mathbf{e} \\ {\mathbf{z_p}}^\top I_3\,\mathbf{p} - \alpha\mathbf{e} \end{pmatrix} \overset{!}{=} \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ 0 \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}$$

umformuliert, wobei $\mathbf{w} = (\mathbf{x}, \mathbf{p}, \mathbf{z}, \mu) \in \mathcal{W}^0$, $I_3 \in \mathbb{R}^{d\ell\times d\ell\times d\ell}$ die $(d\ell)^3$-Einheitsmatrix und $\mathbf{e} = (1, \ldots, 1)^\top$ den Einsvektor darstellt,

$$\mathbf{u_a} := (\mathbf{a}^\top, \ldots, \mathbf{a}^\top)^\top \in \mathbb{R}^{d\ell}, \quad \mathbf{u_b} := (\mathbf{b}^\top, \ldots, \mathbf{b}^\top)^\top \in \mathbb{R}^{d\ell}$$

gilt. Die Anwendung der Newtonmethode auf dieses System liefert mit den Diagonalmatrizen

$$Z_a := I_3\mathbf{z_a},\ Z_b := I_3\mathbf{z_b},\ Z_p := I_3\mathbf{z_p},\ U_a := I_3(\mathbf{x}-\mathbf{u_a}),\ U_b := I_3(\mathbf{u_b}-\mathbf{x}),\ U_p := I_3\mathbf{p}$$

das *primal-duale Newtonsystem* $DF_\alpha(\mathbf{w})\mathbf{s} = -F_\alpha(\mathbf{w})$ mit $DF_\alpha(\mathbf{w}) := \nabla_{\mathbf{w}}F_\alpha(\mathbf{w})$, also

$$\begin{pmatrix} \nabla^2_{\mathbf{xx}}\,\mathfrak{L}(\mathbf{w}) & \nabla^2_{\mathbf{xp}}\,\mathfrak{L}(\mathbf{w}) & 0 & -I & I & 0 \\ \nabla^2_{\mathbf{px}}\,\mathfrak{L}(\mathbf{w}) & \nabla^2_{\mathbf{pp}}\,\mathfrak{L}(\mathbf{w}) & \mathbf{e} & 0 & 0 & -I \\ 0 & \mathbf{e}^\top & 0 & 0 & 0 & 0 \\ Z_a & 0 & 0 & U_a & 0 & 0 \\ -Z_b & 0 & 0 & 0 & U_b & 0 \\ 0 & Z_p & 0 & 0 & 0 & U_p \end{pmatrix} \begin{pmatrix} \mathbf{s}_x \\ \mathbf{s}_p \\ s_\mu \\ \mathbf{s}_{\mathbf{z_a}} \\ \mathbf{s}_{\mathbf{z_b}} \\ \mathbf{s}_{\mathbf{z_p}} \end{pmatrix} = - \begin{pmatrix} \nabla_{\mathbf{x}}\,\mathfrak{L}(\mathbf{w}) \\ \nabla_{\mathbf{p}}\,\mathfrak{L}(\mathbf{w}) \\ \sum_{i=1}^{\ell} p_i - 1 \\ {\mathbf{z_a}}^\top U_a - \alpha\mathbf{e} \\ {\mathbf{z_b}}^\top U_b - \alpha\mathbf{e} \\ {\mathbf{z_p}}^\top U_p - \alpha\mathbf{e} \end{pmatrix}. \quad (4.23)$$

Wie in [47] beschrieben, konvergiert für jedes fest gewählte $\alpha > 0$

- das klassische Newtonverfahren lokal quadratisch und
- ein gedämpftes Newtonverfahren global linear

gegen eine Lösung

$$(\bar{\mathbf{x}}_\alpha, \bar{\mathbf{p}}_\alpha) = \arg\min_{\mathbf{x},\mathbf{p}} J_\alpha(\mathbf{x}, \mathbf{p}).$$

Der Punkt $(\bar{\mathbf{x}}_\alpha, \bar{\mathbf{p}}_\alpha) \in \mathcal{S}^0$ liegt auf einem zentralen Pfad (4.18), der nach Satz 4.2.4 stetig ist. Daher „mündet" dieser Pfad für $\alpha \to 0$ in einem lokalen Minimum des Optimierungsproblems (4.2).

Ein Innere-Punkte-Verfahren kann demnach umgesetzt werden, indem eine Nullfolge $\{\alpha^n\}_{n=0}^{\infty}$ mit $\alpha^{n+1} \leq \alpha^n, n \in \mathbb{N}_0$ gewählt wird und sukzessive für jedes α^n das zugehörige Barriereproblem (4.13) gelöst wird. Für das n-te Barriereproblem $(n \geq 1)$ wählt man als Startiterierte den Optimalwert des $(n-1)$-ten Barriereproblems, d. h. $\left((\mathbf{x}^n)^0, (\mathbf{p}^n)^0\right) := \left((\mathbf{x}^{n-1})^*, (\mathbf{p}^{n-1})^*\right)$. Wählt man die Differenz $\alpha^n - \alpha^{n+1}$ hinreichend klein, liegen $\left((\mathbf{x}^n)^*, (\mathbf{p}^n)^*\right)$ und $\left((\mathbf{x}^{n+1})^*, (\mathbf{p}^{n+1})^*\right)$ auf dem selben Pfad und für hinreichend kleines α^n ist eine lokale Minimallösung des Ursprungsproblems (4.2) gefunden.

Für $\alpha^n \gg 0$ ist es allerdings nicht notwendig, die Barriereprobleme exakt zu lösen. Daher bietet sich eine sogenannte *Pfadverfolgungsmethode* an. Pfadverfolgungsmethoden sind Iterationsverfahren, die sich am zentralen Pfad (4.18) orientieren. Da mehr als ein zentraler Pfad existieren kann, wird bei einer Pfadverfolgungsmethode ein lokales Minimum von (4.2) innerhalb einer *Umgebung des zentralen Pfades* gesucht. Dadurch wird gewährleistet, dass die Iterierten $(\mathbf{x}^n, \mathbf{p}^n, \mathbf{z}^n, \mu^n) \in \mathcal{W}^0$ entlang eines zentralen Pfades bestimmt werden.

Definition 4.2.2 (Umgebung des zentralen Pfades). *Die Menge*

$$\begin{aligned} N(\alpha) := \{&(\mathbf{x}, \mathbf{p}, \mathbf{z_a}, \mathbf{z_b}, \mathbf{z_p}, \mu) \in \mathcal{W}^0 : \\ &(x_k^i - a_k) z_{a,k}^i \geq \gamma\alpha,\ (b_k - x_k^i) z_{b,k}^i \geq \gamma\alpha,\ p_i z_{p,i} \geq \gamma\alpha, \\ &z_{a,k}^i|_{\{x_k^i > \frac{(b_k+a_k)}{2}\}} \leq \frac{2\alpha}{b_k - a_k},\ z_{b,k}^i|_{\{x_k^i < \frac{(b_k+a_k)}{2}\}} \leq \frac{2\alpha}{b_k - a_k}, \\ &z_{p,i}|_{\{p_i > \frac{1}{2}\}} \leq 2\alpha, \\ &\min\{z_{a,k}^i|_{\{x_k^i = \frac{(b_k+a_k)}{2}\}}, z_{b,k}^i|_{\{x_k^i = \frac{(b_k+a_k)}{2}\}}\} \leq \frac{2\alpha}{b_k - a_k}\Big\} \end{aligned} \tag{4.24}$$

$\gamma \in (0,1)$, $k \in K_d$, $i \in I_\ell$ *heißt* Umgebung des zentralen Pfades. *Jede Lösung von (4.16) liegt in* $N(\alpha)$.

Um zu gewährleisten, dass stets $(\mathbf{x}^n, \mathbf{p}^n, \mathbf{z}^n, \mu^n) \in N(\alpha)$ ist, wird nach jedem Iterationsschritt eine Projektion auf die Menge $N(\alpha)$ vorgenommen, d. h.

$$(\mathbf{x}^n, \mathbf{p}^n, \mathbf{z}^n, \mu^n) := \mathcal{P}_\alpha(\mathbf{x}^n, \mathbf{p}^n, \mathbf{z}^n, \mu^n)$$

mit

$$\begin{aligned} \mathcal{P}_\alpha(\mathbf{x}^n, \mathbf{p}^n, \mathbf{z}^n, \mu^n) \ &= \ \arg\min_{(\mathbf{x},\mathbf{p},\mathbf{z},\mu)\in\mathcal{W}^0} \|(\mathbf{x}^n, \mathbf{p}^n, \mathbf{z}^n, \mu^n) - (\mathbf{x}, \mathbf{p}, \mathbf{z}, \mu)\|_2^2, \\ \text{u.d.N.} \quad &(\mathbf{x}, \mathbf{p}, \mathbf{z}, \mu) \in N(\alpha). \end{aligned}$$

Ist $\alpha > 0$ hinreichend klein, ist aufgrund der Stetigkeit des zentralen Pfades eine genäherte lokale Minimallösung des Ursprungsproblems (4.2) gefunden. In Algorithmus 4.2.1 wird beschrieben, wie eine Minimallösung gefunden werden kann.

Bevor Algorithmus 4.2.1 vorgestellt wird, wird zunächst gezeigt, dass die Operatornorm $\|\cdot\|$ der Inversen von $DF_\alpha(\mathbf{w})$ durch eine positive Konstante nach oben beschränkt ist, die vom Barriereparameter $\alpha > 0$ abhängt. Die Beweisidee ist der Veröffentlichung [88] entnommen und wurde in dieser Arbeit für optimale Designprobleme erweitert.

Satz 4.2.7. *Unter der Voraussetzung (4.1) gilt für die Inverse von $DF_\alpha(\mathbf{w})$ mit $\alpha > 0$ und $\mathbf{w} \in N(\alpha)$*

$$\|DF_\alpha(\mathbf{w})^{-1}\| \;\leq\; \frac{C}{\sqrt{\alpha}},$$

mit einer Konstanten $C > 0$.

Beweis:
Sei

$$S(\mathbf{w}) \;=\; \begin{pmatrix} I & & & & & & \\ & I & & & & & \\ & & I & & & & \\ & & & (U_a + Z_a)^{-1} & & & \\ & & & & (U_b + Z_b)^{-1} & & \\ & & & & & (U_p + Z_p)^{-1} & \end{pmatrix}, \tag{4.25}$$

dann ist die Matrix $S(\mathbf{w})DF_\alpha(\mathbf{w})$ unabhängig von α. Aufgrund der Stetigkeit von $S(\mathbf{w})$ und $DF_\alpha(\mathbf{w})$, ist $S(\mathbf{w})DF_\alpha(\mathbf{w})$ ebenfalls stetig. Da sich die Inverse einer Matrix mittels ihrer Minoren [27] berechnen lässt, die für $S(\mathbf{w})DF_\alpha(\mathbf{w})$ wiederum stetig sind, folgt die Stetigkeit von $(S(\mathbf{w})DF_\alpha(\mathbf{w}))^{-1}$. Daher ist

$$\|(S(\mathbf{w})DF_\alpha(\mathbf{w}))^{-1}\| \leq \tilde{C}$$

mit einer Konstanten $\tilde{C} \geq 0$.

Laut Definition 4.2.2 gilt mit $\gamma \in (0,1)$

$$\begin{aligned} (x_k^i - a_k) + z_{a,k}^i &\geq 2\sqrt{(x_k^i - a_k)z_{a,k}^i} &&\geq 2\sqrt{\gamma\alpha}, \\ (b_k - x_k^i) + z_{b,k}^i &\geq 2\sqrt{(b_k - x_k^i)z_{b,k}^i} &&\geq 2\sqrt{\gamma\alpha}, \\ p_i + z_{p,i} &\geq 2\sqrt{p_i z_{p,i}} &&\geq 2\sqrt{\gamma\alpha}. \end{aligned}$$

Daher gilt für die Diagonalmatrix $S(\mathbf{w})$

$$\|S(\mathbf{w})\| \leq \frac{1}{\min\{1, 2\sqrt{\gamma\alpha}\}}$$

und folglich

$$\begin{aligned} \|DF_\alpha(\mathbf{w})^{-1}\| &= \|(S(\mathbf{w})DF_\alpha(\mathbf{w}))^{-1}S(\mathbf{w})\| \;\leq\; \|(S(\mathbf{w})DF_\alpha(\mathbf{w}))^{-1}\|\|S(\mathbf{w})\| \\ &\leq \frac{\tilde{C}}{\min\{1, 2\sqrt{\gamma\alpha}\}} \;=\; \frac{C}{\sqrt{\alpha}}, \end{aligned} \tag{4.26}$$

mit einer Konstanten $C > 0$.

□

Folgender Algorithmus basiert auf eine *Pfadverfolgungsmethode* und entstand in Anlehnung eines Innere-Punkte-Verfahrens aus [88].

Algorithmus 4.2.1. (Primal-duales Innere-Punkte-Verfahren zur Bestimmung eines lokalen D-optimalen Designs)

(S.0) *Initialisierung:*
Wähle ein $\alpha^0 > 0$, $\sigma_{\min} \in (0,1)$, $C_0 > 0$ *und* $\varepsilon > 0$ *hinreichend klein.*
Wähle ein $\gamma \in (0,1)$ *für* $N(\alpha)$ *aus (4.24) und setze* $n := 0$.
Wähle ein $\mathbf{w}^0 \in N(\alpha^0)$, *so dass* $\|F_{\alpha^0}(\mathbf{w}^0)\|_2 < C_0\alpha^0$.

(S.1) *Berechnung einer Suchrichtung:*
Bestimme die Lösung $\mathbf{s}^n$ *des Newtonsystems*

$$DF_{\alpha^n}(\mathbf{w}^n)\mathbf{s}^n \;=\; -F_{\alpha^n}(\mathbf{w}^n).$$

(S.2) *Berechnung der neuen Iterierten:*
Setze $\mathbf{w}^{n+1} := P_\alpha(\mathbf{w}^n + \mathbf{s}^n)$.
Wähle ein $\sigma^n \in [\sigma_{\min}, 1]$ *und setze* $\alpha^{n+1} := \sigma^n\alpha^n$.

(S.3) *Abbruchbedingung:*
Ist $\alpha^n \le \varepsilon$: *Ein KKT-Punkt von (4.2) ist gefunden. Stop.*
Sonst setze $n := n+1$ *und gehe zu (S.1).*

Im folgenden Abschnitt wird gezeigt, dass Algorithmus 4.2.1 global linear gegen einen KKT-Punkt von (4.2) konvergiert.

4.2.2 Konvergenz

In diesem Abschnitt wird angenommen, dass für den Barriereparameter $\alpha^0 > 0$ aus Algorithmus 4.2.1 bereits eine Näherung einer globalen Lösung $\mathbf{w}^*(\alpha^0) \in \mathcal{W}^0$ des Barriereproblems (4.13) näherungsweise ermittelt wurde. Auf dem zentralen Pfad, auf dem $\mathbf{w}^*(\alpha^0)$ liegt, befindet sich für jedes $\alpha \in (0, \alpha^0]$ eine globale Lösung des zugehörigen Barriereproblems (4.13). Dieser Pfad wird im folgenden durch $\mathbf{w}^*(\alpha), \alpha > 0$, dargestellt. Es gilt

$$\lim_{\alpha \to 0} \mathbf{w}^*(\alpha) = \mathbf{w}^*$$

mit einer globalen Lösung $\mathbf{w}^* \in \mathcal{W}$ des Optimierungsproblems (4.2).

Für fest gewähltes $\alpha > 0$ konvergiert das Newtonverfahren lokal quadratisch gegen den *zentralen Pfad* $\mathbf{w}^*(\alpha)$.

Bemerkung 4.2.8. *Die für dieses Unterkapitel verwendeten Beweisideen lehnen sich an die Veröffentlichung [88] und wurden in dieser Arbeit für optimale Designprobleme erweitert.*

Satz 4.2.9. *Unter der Voraussetzung (4.11) sei* $\alpha^0 > 0$ *und* $\varepsilon > 0$ *fest. Dann gilt für alle* $\alpha \in (0, \alpha^0]$ *und* $\mathbf{w}^n \in \{\mathbf{w} \in N(\alpha) : \|\mathbf{w} - \mathbf{w}^*(\alpha)\|_2 \le \varepsilon\}$

(i) $\|\mathbf{w}^{n+1} - \mathbf{w}^*(\alpha)\|_2 \le \dfrac{C}{\sqrt{\alpha}}\|\mathbf{w}^n - \mathbf{w}^*(\alpha)\|_2^2,$

(ii) $\|P_\alpha(\mathbf{w}^{n+1}) - \mathbf{w}^*(\alpha)\|_2 \le \dfrac{2C}{\sqrt{\alpha}}\|\mathbf{w}^n - \mathbf{w}^*(\alpha)\|_2^2,$

wobei $\mathbf{w}^{n+1} \in \mathcal{W}^0$ *die Lösung des Newtonsystems*

$$DF_\alpha(\mathbf{w}^n)(\mathbf{w}^{n+1} - \mathbf{w}^n) = -F_\alpha(\mathbf{w}^n) \tag{4.27}$$

aus (4.23) und $P_\alpha(\mathbf{w}^{n+1})$ *dessen Projektion auf die Umgebung* $N(\alpha)$ *darstellt.*

Beweis:

(i) Es gilt mit $\mathbf{w}^*_\alpha := \mathbf{w}^*(\alpha)$

$$F_\alpha(\mathbf{w}^*_\alpha) \;=\; F_\alpha(\mathbf{w}^n) + DF_\alpha(\mathbf{w}^n)(\mathbf{w}^*_\alpha - \mathbf{w}^n) + \frac{1}{2}(\mathbf{w}^*_\alpha - \mathbf{w}^n)^\top H(\boldsymbol{\xi})(\mathbf{w}^*_\alpha - \mathbf{w}^n)$$

mit einem $\boldsymbol{\xi} = \mathbf{w}^*_\alpha + \beta(\mathbf{w}^n - \mathbf{w}^*_\alpha)$, $\beta \in [-1, 1]$.
Mit

$$DF_\alpha(\mathbf{w}^n)(\mathbf{w}^{n+1} - \mathbf{w}^n) = -F_\alpha(\mathbf{w}^n) \;\text{ und }\; DF_\alpha(\mathbf{w}^n)(\mathbf{w}^*_\alpha - \mathbf{w}^*_\alpha) = -F_\alpha(\mathbf{w}^*_\alpha)$$

ist

$$\begin{aligned} DF_\alpha(\mathbf{w}^n)(\mathbf{w}^{n+1} - \mathbf{w}^*_\alpha) &= F_\alpha(\mathbf{w}^*_\alpha) - F_\alpha(\mathbf{w}^n) - DF_\alpha(\mathbf{w}^n)(\mathbf{w}^*_\alpha - \mathbf{w}^n) \\ &= \frac{1}{2}(\mathbf{w}^*_\alpha - \mathbf{w}^n)^\top H(\boldsymbol{\xi})(\mathbf{w}^*_\alpha - \mathbf{w}^n). \end{aligned}$$

Somit ist

$$\begin{aligned} \|\mathbf{w}^{n+1} - \mathbf{w}^*_\alpha\|_2 &\leq \tilde{C}\|DF_\alpha(\mathbf{w}^n)^{-1}\|\|\mathbf{w}^*_\alpha - \mathbf{w}^n\|_2^2 \\ &\overset{(4.26)}{\leq} \frac{C'}{\sqrt{\alpha}}\|\mathbf{w}^*_\alpha - \mathbf{w}^n\|_2^2 \end{aligned} \tag{4.28}$$

mit den Konstanten $\tilde{C} > 0$ und $C' > 0$.

(ii) Aus

$$\|P_\alpha(\mathbf{w}^{n+1}) - \mathbf{w}^{n+1}\|_2 \;\leq\; \|\mathbf{w}^*_\alpha - \mathbf{w}^{n+1}\|_2$$

folgt

$$\begin{aligned} \|P_\alpha(\mathbf{w}^{n+1}) - \mathbf{w}^*_\alpha\|_2 &\leq \|P_\alpha(\mathbf{w}^{n+1}) - \mathbf{w}^{n+1}\|_2 + \|\mathbf{w}^{n+1} - \mathbf{w}^*_\alpha\|_2 \\ &\leq 2\|\mathbf{w}^{n+1} - \mathbf{w}^*_\alpha\|_2 \overset{(4.28)}{\leq} \frac{2C'}{\sqrt{\alpha}}\|\mathbf{w}^*_\alpha - \mathbf{w}^n\|_2^2. \end{aligned}$$

□

Satz 4.2.10. (Global lineare Konvergenz)
Unter der Voraussetzung (4.11) sei $\alpha^0 > 0$ *und* $\varepsilon > 0$ *fest. Dann existiert ein* $\tilde{\varepsilon} > 0$, *so dass der Algorithmus 4.2.1 für jeden Startwert* $\mathbf{w}^0 \in \{\mathbf{w} \in N(\alpha^0) : \|\mathbf{w} - \mathbf{w}^*(\alpha)\|_2 \leq \tilde{\varepsilon}\}$ *eine Folge* $\mathbf{w}^n \in N(\alpha^n)$ *generiert mit den Eigenschaften*

(i) $\|\mathbf{w}^n - \mathbf{w}^*(\alpha^n)\|_2 \leq C\sqrt{\alpha^n}$,

(ii) $\|\mathbf{w}^n - \mathbf{w}^*\|_2 \leq C\sqrt{\alpha^n} + L\,\alpha^n$,

mit $C, L > 0$ *und* $\mathbf{w}^* = \lim_{\alpha \to 0} \mathbf{w}^*(\alpha)$.

Beweis:

(i) Sei $\alpha \in (0, \alpha^0]$ fest, aber beliebig. Dann gibt es nach Satz 4.2.9 eine Konstante $C > 0$, so dass für jedes $\mathbf{w}^n \in \{\mathbf{w} \in N(\alpha) : \|\mathbf{w} - \mathbf{w}^*(\alpha)\|_2 \leq \varepsilon\}$ die Abschätzung

$$\|P_\alpha(\mathbf{w}^{n+1}) - \mathbf{w}^*(\alpha)\|_2 \;\leq\; \frac{2C}{\sqrt{\alpha}}\|\mathbf{w}^n - \mathbf{w}^*(\alpha)\|_2^2$$

gilt, wobei $\mathbf{w}^{n+1} \in \mathcal{W}^0$ die Lösung des Newtonsystems (4.27) und $P_\alpha(\mathbf{w}^{n+1})$ dessen Projektion auf die Umgebung $N(\alpha)$ darstellt.
Es wird nun $\tau \in (0,1)$ derart gewählt, so dass

$$\tilde{\varepsilon} \;:=\; \frac{\tau\sqrt{\alpha}}{2C} \;\leq\; \varepsilon$$

gilt. Dann gilt für alle $\mathbf{w}^n \in \{\mathbf{w} \in N(\alpha) : \|\mathbf{w} - \mathbf{w}^*(\alpha)\|_2 \leq \tilde{\varepsilon}\}$

$$\|P_\alpha(\mathbf{w}^{n+1}) - \mathbf{w}^*(\alpha)\|_2 \;\leq\; \tau\|\mathbf{w}^n - \mathbf{w}^*(\alpha)\|_2 \;\leq\; \tau\tilde{\varepsilon} \;=\; \frac{\tau^2\sqrt{\alpha}}{2C}.$$

Mit $\sigma \in (0,1)$ ist

$$\begin{aligned}\|P_\alpha(\mathbf{w}^{n+1}) - \mathbf{w}^*(\sigma\alpha)\|_2 &\leq \|P_\alpha(\mathbf{w}^{n+1}) - \mathbf{w}^*(\alpha)\|_2 + \|\mathbf{w}^*(\alpha) - \mathbf{w}^*(\sigma\alpha)\|_2\\ &\leq \frac{\tau^2\sqrt{\alpha}}{2C} + L|\alpha - \sigma\alpha|\\ &= \frac{\tau^2\sqrt{\alpha}}{2C} + L(1-\sigma)\alpha.\end{aligned}$$

Um zu gewährleisten, dass wiederum $\|P_\alpha(\mathbf{w}^{n+1}) - \mathbf{w}^*(\sigma\alpha)\|_2 \leq \tilde{\varepsilon}$ erfüllt ist, wird $\sigma \in (0,1)$ derart gewählt, dass $\frac{\tau^2\sqrt{\alpha}}{2C} + L(1-\sigma)\alpha \;\leq\; \frac{\tau\sqrt{\sigma\alpha}}{2C}$, beziehungsweise

$$\frac{\tau^2}{2C} + L(1-\sigma)\sqrt{\alpha} \;\leq\; \frac{\tau\sqrt{\sigma}}{2C} \tag{4.29}$$

gilt. Diese Abschätzung ist erfüllt für alle $\sigma \in [\bar{\sigma}_{\min}, 1)$ mit

$$\bar{\sigma}_{\min} := \left(-\frac{\tau}{4CL\sqrt{\alpha^0}} + \sqrt{\frac{\tau^2}{16C^2L^2\alpha^0} + \frac{\tau^2}{2CL\sqrt{\alpha^0}} + 1}\right)^2, \tag{4.30}$$

wobei $\bar{\sigma}_{\min}$ aus (4.30) die Lösung der Gleichung

$$\frac{\tau^2}{2C} + L(1-\bar{\sigma}_{\min})\sqrt{\alpha^0} \;=\; \frac{\tau\sqrt{\bar{\sigma}_{\min}}}{2C}$$

darstellt mit $\bar{\sigma}_{\min} \in (0,1)$. Da in (4.30) die Abschätzung $\frac{\tau^2}{2CL\sqrt{\alpha^0}} + 1 > 0$ gilt,

ist stets $\bar{\sigma}_{\min} > 0$ und

$$\begin{aligned}
\bar{\sigma}_{\min} &= \left(-\frac{\tau}{4CL\sqrt{\alpha^0}} + \sqrt{\frac{\tau^2}{16C^2L^2\alpha^0} + \frac{\tau^2}{2CL\sqrt{\alpha^0}} + 1}\right)^2 \\
&\overset{\tau^2<\tau}{<} \left(-\frac{\tau}{4CL\sqrt{\alpha^0}} + \sqrt{\frac{\tau^2}{16C^2L^2\alpha^0} + \frac{\tau}{2CL\sqrt{\alpha^0}} + 1}\right)^2 \\
&= \left(-\frac{\tau}{4CL\sqrt{\alpha^0}} + \sqrt{\left(\frac{\tau}{4CL\sqrt{\alpha^0}} + 1\right)^2}\right)^2 \\
&= 1.
\end{aligned}$$

Somit bestimmt der Algorithmus 4.2.1 mit $\sigma_{\min} \in [\bar{\sigma}_{\min}, 1)$ eine Folge $\mathbf{w}^n$ mit $\mathbf{w}^{n+1} := P_\alpha(\mathbf{w}^n + \mathbf{s}^n)$ und

$$\|\mathbf{w}^{n+1} - \mathbf{w}^*(\alpha^{n+1})\|_2 \;\leq\; \frac{\tau\sqrt{\alpha^{n+1}}}{2C} \;\leq\; \tilde{\varepsilon}. \tag{4.31}$$

(ii) Mit der Lösung $\mathbf{w}^* = \lim_{\alpha\to 0} \mathbf{w}^*(\alpha)$ gilt

$$\|\mathbf{w}^n - \mathbf{w}^*\|_2 \;\leq\; \|\mathbf{w}^n - \mathbf{w}^*(\alpha^n)\|_2 + \|\mathbf{w}^*(\alpha^n) - \mathbf{w}^*\|_2 \overset{(4.31)}{\leq} \frac{\tau\sqrt{\alpha^n}}{2C} + L\alpha^n,$$

mit der Lipschitzkonstanten $L > 0$, die aufgrund der Stetigkeit des zentralen Pfades $\mathbf{w}^*(\alpha)$ für $\alpha \in (0, \alpha^0]$ existiert.

□

4.3 Active-Set-Methode

In diesem Unterkapitel wird ein weiteres Verfahren vorgestellt, mit dem das Optimierungsproblem (4.2) gelöst werden kann, wenn die Voraussetzung (4.11) erfüllt ist. Bei diesem Verfahren handelt es sich um eine *Active-Set-Methode*, die beim iterativen Lösen des Problems (4.2) lokal quadratische und global lineare Konvergenz aufweist. Diese Methode ist einem primal-dualen Innere-Punkte-Verfahren vorzuziehen, da neben der gleichen Konvergenzordnung die Dimension des zu lösenden Newtonsystems wesentlich kleiner ist.

Eine Active-Set-Methode bietet sich insbesondere wegen der „einfachen“ Form der Ungleichungsnebenbedingungen, die durch

$$a_k \leq x_k^i \leq b_k, \quad 0 \leq p_i, \quad i \in I_\ell, \quad k \in K_d \tag{4.32}$$

gegeben sind, an. Zum Herleiten dieses Verfahrens werden die Begriffe *aktiv* bzw. *inaktiv* benötigt.

Definition 4.3.1. *(**aktive und inaktive Komponenten**) Sei $(\mathbf{x}, \mathbf{p}) \in \mathcal{S}$ mit $\mathbf{x} = ({\mathbf{x}^1}^\top, \ldots, {\mathbf{x}^\ell}^\top)^\top$, $\mathbf{p} = (p_1, \ldots, p_\ell)^\top$ und $\mathcal{S}$ aus (4.5). Mit $x_k^i \in [a_k, b_k]$ und $p_i \geq 0$ $\forall\, i \in I_\ell$, $k \in K_d$ sagt man, dass*

- x_k^i *aktiv ist, wenn* $x_k^i = a_k$ *oder* $x_k^i = b_k$ *gilt,*
- x_k^i *inaktiv ist, wenn* $a_k < x_k^i < b_k$ *gilt,*
- p_i *aktiv ist, wenn* $p_i = 0$ *gilt,*
- p_i *inaktiv ist, wenn* $p_i > 0$ *gilt.*

Bei einer Active-Set-Methode nutzt man die Abhängigkeit zwischen der primalen $\bar{\mathbf{x}}$ und der dualen Lösung $\bar{\boldsymbol{\lambda}}$ des Problems (4.2) aus, die durch die Komplementaritätsbedingung

$$\bar{\lambda}_{a,k}^i \cdot (a_k - \bar{x}_k^i) = 0, \quad \bar{\lambda}_{b,k}^i \cdot (\bar{x}_i - b) = 0, \quad \bar{\lambda}_{p,i} \cdot \bar{p}_i = 0, \tag{4.33}$$

$\forall\, i \in I_\ell$, $k \in K_d$ gegeben ist. Man sieht, dass entweder der Lagrangemultiplikator $\bar{\lambda}_{a,k}^i$ (bzw. $\bar{\lambda}_{b,k}^i$, $\bar{\lambda}_{p,i}$), die Nebenbedingung $a_k - \bar{x}_k^i$ (bzw. $\bar{x}_k^i - b_k$, $\bar{p}_i$) oder beide gleich Null sind. Diese Eigenschaft erlaubt es der Active-Set-Methode in jedem Iterationsschritt auf die Ungleichungsnebenbedingungen (4.32) zu verzichtet, wobei nach jedem Schritt eine Projektion auf den zulässigen Bereich S notwendig ist. Da S lediglich durch die Box Constraints (4.32) beschränkt ist, ist eine solche Projektion für Problem (4.2) einfach umzusetzen.

Durch Weglassen der Ungleichungsnebenbedingungen in (4.2) erhält man das Optimierungsproblem

$$\left.\begin{aligned} &\min_{\mathbf{x},\mathbf{p}} J(\mathbf{x},\mathbf{p}) = -\ln|M(\mathbf{x},\mathbf{p})|, \\ &\text{mit der Fisher-Informationsmatrix} \\ &M(\mathbf{x},\mathbf{p}) \;=\; \sum_{i=1}^{\ell} p_i \int_T \nabla_{\boldsymbol{\theta}} u(t,\mathbf{x}^i,\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} u(t,\mathbf{x}^i,\boldsymbol{\theta})^T dt, \\ &\text{unter den Nebenbedingungen} \\ &\sum_{i=1}^{\ell} p_i \;=\; 1, \quad \mathbf{p} \;=\; (p_1,\ldots,p_\ell)^\top \in \mathbb{R}^\ell, \quad \mathbf{x} \;=\; ({\mathbf{x}^1}^\top,\ldots,{\mathbf{x}^\ell}^\top)^\top \in \mathbb{R}^{d\ell}, \end{aligned}\right\} \tag{4.34}$$

sowie das zugehörige Lagrangefunktional $\tilde{\mathcal{L}} : \mathbb{R}^{d\ell} \times \mathbb{R}^\ell \times \mathbb{R} \to \mathbb{R}$ mit

$$\tilde{\mathcal{L}}(\mathbf{x},\mathbf{p},\mu) \;=\; J(\mathbf{x},\mathbf{p}) + \mu \left(\sum_{i=1}^{\ell} p_i - 1 \right).$$

Regularität in einer lokalen Lösung

Da das Optimierungsproblem (4.34) nur eine Nebenbedingung beinhaltet, die zudem affin ist, ist nach [43] die Regularitätsbedingung von Robinson in jedem Punkt $(\mathbf{x},\mathbf{p}) \in \mathbb{R}^{d\ell} \times \mathbb{R}^\ell$ erfüllt, insbesondere also in jedem lokalen Minimum $(\bar{\mathbf{x}},\bar{\mathbf{p}}) \in \mathbb{R}^{d\ell} \times \mathbb{R}^\ell$.

Optimalitätsbedingungen erster Ordnung

Da die Regularitätsbedingung von Robinson erfüllt ist, kann der Satz von Kuhn-Tucker angewendet werden.

Satz 4.3.1. (Satz von Kuhn-Tucker für das Optimierungsproblem (4.34), [43]) *Unter der Voraussetzung (4.1) sei* $(\bar{\mathbf{x}}, \bar{\mathbf{p}}) \in \mathbb{R}^{d\ell} \times \mathbb{R}^{\ell}$ *eine lokale Optimallösung von (4.34). Da das Problem (4.34) in* $(\bar{\mathbf{x}}, \bar{\mathbf{p}})$ *regulär ist, existiert ein* $\bar{\mu} \in \mathbb{R}$*, so dass*

(i) $\bar{\mu} \cdot \left(\sum_{i=1}^{\ell} \bar{p}_i - 1 \right) = 0$ *(Bedingung vom komplementären Schlupf),*

(ii) $\nabla_{\mathbf{x},\mathbf{p}} \tilde{\mathcal{L}}(\bar{\mathbf{x}}, \bar{\mathbf{p}}, \bar{\mu}) = 0.$

Die Gleichungen $\nabla_{\mathbf{x},\mathbf{p}} \tilde{\mathcal{L}}(\bar{\mathbf{x}}, \bar{\mathbf{p}}, \bar{\mu}) = 0$ und die Bedingung vom komplementären Schlupf aus Satz 4.3.1 stellen mit $(\bar{\mathbf{x}}, \bar{\mathbf{p}}, \bar{\mu}) \in \mathbb{R}^{d\ell} \times \mathbb{R}^{\ell} \times \mathbb{R}$ die *reduzierten KKT-Bedingungen*

$$\left.\begin{aligned} \partial_{x_k^i} \tilde{\mathcal{L}}(\bar{\mathbf{x}}, \bar{\mathbf{p}}, \bar{\mu}) &= \partial_{x_k^i} J(\bar{\mathbf{x}}, \bar{\mathbf{p}}) = 0, \\ \partial_{p_i} \tilde{\mathcal{L}}(\bar{\mathbf{x}}, \bar{\mathbf{p}}, \bar{\mu}) &= \partial_{p_i} J(\bar{\mathbf{x}}, \bar{\mathbf{p}}) + \bar{\mu} = 0, \\ \bar{\mu} \cdot \left(\sum_{i=1}^{\ell} \bar{p}_i - 1 \right) &= 0, \end{aligned}\right\} \tag{4.35}$$

für $i \in I_\ell$ und $k \in K_d$ dar.

Folgender Satz gibt einen Zusammenhang zwischen den KKT-Bedingungen des Ursprungsproblems (4.9) und den Bedingungen (4.35) wieder.

Satz 4.3.2. *Sei* $(\bar{\mathbf{x}}, \bar{\mathbf{p}}) \in S$ *eine lokale Minimallösung des Optimierungsproblems (4.2) und* $(\bar{\mathbf{x}}, \bar{\mathbf{p}}, \bar{\boldsymbol{\lambda}}, \bar{\mu}) \in \mathcal{W}$ *der zugehörige KKT-Punkt, der die Bedingung (4.9) erfüllt. Dann ist*

- $\partial_{x_k^i} \mathcal{L}(\bar{\mathbf{x}}, \bar{\mathbf{p}}, \bar{\boldsymbol{\lambda}}, \bar{\mu}) = \partial_{x_k^i} \tilde{\mathcal{L}}(\bar{\mathbf{x}}, \bar{\mathbf{p}}, \bar{\mu}) = 0$*, wenn* $\bar{x}_k^i$ *inaktiv ist,*
- $\partial_{p_i} \mathcal{L}(\bar{\mathbf{x}}, \bar{\mathbf{p}}, \bar{\boldsymbol{\lambda}}, \bar{\mu}) = \partial_{p_i} \tilde{\mathcal{L}}(\bar{\mathbf{x}}, \bar{\mathbf{p}}, \bar{\mu}) = 0$*, wenn* $\bar{p}_i$ *inaktiv ist,*
- $\partial_{x_k^i} \tilde{\mathcal{L}}(\bar{\mathbf{x}}, \bar{\mathbf{p}}, \bar{\mu}) \in \mathbb{R}$*, wenn* $\bar{x}_k^i$ *aktiv ist,*
- $\partial_{p_i} \tilde{\mathcal{L}}(\bar{\mathbf{x}}, \bar{\mathbf{p}}, \bar{\mu}) \in \mathbb{R}$*, wenn* $\bar{p}_i$ *aktiv ist,*

mit dem Lagrangefunktional $\mathcal{L} : \mathcal{W} \to \mathbb{R}$ *aus (4.7).*

Beweis:
Da $(\bar{\mathbf{x}}, \bar{\mathbf{p}}) \in S$ eine lokale Minimallösung des Optimierungsproblems (4.2) darstellt, existiert nach Satz 4.1.2 ein $(\bar{\boldsymbol{\lambda}}, \bar{\mu}) \in \mathbb{R}_+^{(d+1)\ell} \times \mathbb{R}$, so dass die KKT-Bedingungen (4.9) erfüllt sind. Da die Komplementaritätsbedingungen (4.33) erfüllt sind, ist $\lambda_{a,k}^i = \lambda_{b,k}^i = 0$, wenn $\bar{x}_k^i$ inaktiv ist und $\lambda_{p,i} = 0$, wenn $\bar{p}_i$ inaktiv ist.

□

Mit einer optimalen Lösung $(\bar{\mathbf{x}}, \bar{\mathbf{p}})$ sind somit nur die Gleichungen aus (4.35) von (4.2) erfüllt, für die die zugehörigen Komponenten $\bar{x}_k^i$, bzw. $\bar{p}_i$ inaktiv sind. Da im Vorhinein nicht bekannt ist, welche Komponenten der optimalen Lösung $(\bar{\mathbf{x}}, \bar{\mathbf{p}}) \in S$ aktiv und welche inaktiv sind, wird bei der Active-Set-Methode in jedem Iterationsschritt auf diese Eigenschaft geprüft.

4.3.1 Algorithmus

Das Nullstellenproblem (4.35), dargestellt durch $F(\bar{\mathbf{x}}, \bar{\mathbf{p}}, \bar{\mu}) = 0$, kann mit dem in [26] beschriebenen Newton-artigen Verfahren gelöst werden. Es handelt sich um eine Active-Set-Methode, bei der in jedem Iterationsschritt mit einer reduzierten Hessematrix das zugehörige Newtonsystem gelöst wird. Falls die Lösung eine zulässige Richtung darstellt, kann durch Schrittweitensteuerung die neue Iterierte bestimmt werden. Wenn diese Lösung allerdings keine zulässige Richtung ist, wird stattdessen ein Gradientenprojektionsschritt angewendet um das Zielfunktional zu verkleinern. Der für den Gradientenprojektionsschritt benötigte Projektionsoperator $\mathcal{P}_{\mathcal{S}} : \mathbb{R}^{(d+1)\ell} \times \mathbb{R} \to \mathcal{S} \times \mathbb{R}$ ist definiert durch

$$\mathcal{P}_{\mathcal{S}}(\mathbf{w}) = \arg \min_{\mathbf{y} \in \mathcal{S} \times \mathbb{R}} \|\mathbf{w} - \mathbf{y}\|_2^2. \tag{4.36}$$

Zur übersichtlichen Darstellung sei im folgenden mit $n \in \mathbb{N}_0$

$$\begin{aligned} DF(\mathbf{x}^n, \mathbf{p}^n, \mu^n)\, \mathbf{s}^n &= -F(\mathbf{x}^n, \mathbf{p}^n, \mu^n) \\ (\mathbf{x}^{n+1^\top}, \mathbf{p}^{n+1^\top}, \mu^{n+1})^\top &= (\mathbf{x}^{n^\top}, \mathbf{p}^{n^\top}, \mu^n)^\top + (\mathbf{s}_x^{n^\top}, \mathbf{s}_p^{n^\top}, s_\mu^n)^\top \end{aligned} \tag{4.37}$$

das nichtreduzierte Newtonsystem, wobei $DF(\mathbf{x}, \mathbf{p}, \mu) = \nabla_{\mathbf{x},\mathbf{p},\mu} F(\mathbf{x}, \mathbf{p}, \mu)$ die Jacobimatrix darstellt und $\mathbf{s}^n := (\mathbf{s}_x^{n^\top}, \mathbf{s}_p^{n^\top}, s_\mu^n)^\top$. Für den folgenden Algorithmus sei

$$DF_i(\mathbf{x}^n, \mathbf{p}^n, \mu^n)\, \mathbf{s}^n = -F_i(\mathbf{x}^n, \mathbf{p}^n, \mu^n)$$

die $i-$te Zeile des Newtonsystems (4.37) mit $i = 1, \ldots, (d+1)\ell + 1$ und

$$F_{\mathbf{x},\mathbf{p}}(\mathbf{w}) := \Big(F_1(\mathbf{w}), \ldots, F_{(d+1)\ell}(\mathbf{w})\Big)^\top.$$

Algorithmus 4.3.1. (Active-Set Newton-artige Methode zur Bestimmung eines D-optimalen Designs)

(S.0) *Initialisierung:*
Wähle ein $\mathbf{w}^0 = (\mathbf{x}^0, \mathbf{p}^0, \mu^0) \in \mathcal{S}^0 \times \mathbb{R}$ *und setze* $n := 0$.
Wähle ein $\sigma > 0$, $c > 0$ *und ein* $\varepsilon > 0$ *hinreichend klein.*

(S.1) *Setzen der aktiven und inaktiven Menge:*
Setze

$$\begin{aligned} \mathcal{A}^n &:= \{k + (i-1)d : x_k^{i,n} = a_k \vee x_k^{i,n} = b_k\} \cup \{d\ell + i : p_i^n \le \delta\}, \\ \mathcal{I}^n &:= \{1, \ldots, (d+1)\ell\} \setminus \mathcal{A}^n. \end{aligned}$$

(S.2) *Berechnung einer Suchrichtung:*
Bestimme die Lösung $\mathbf{s}^n$ *des Systems*

$$\begin{aligned} DF_i(\mathbf{w}^n)\mathbf{s}^n &= -F_i(\mathbf{w}^n), && \text{für } i \in \mathcal{I}^n, \\ s_i^n &= 0, && \text{für } i \in \mathcal{A}^n. \end{aligned}$$

(S.3) *Schrittweitenbestimmung:*
Bestimme das größte $\alpha^n \in [0, 1]$, *so dass* $\mathbf{w}^n + \alpha^n \mathbf{s}^n \in \mathcal{S} \times \mathbb{R}$.

(S.4) *Berechnung der neuen Iterierten:*
Ist $J(\mathbf{x}^n + \alpha^n \mathbf{s}_x^n, \mathbf{p}^n + \alpha^n \mathbf{s}_p^n) \leq J(\mathbf{x}^n, \mathbf{p}^n)$*, dann:*
setze $\mathbf{w}^{n+1} := \mathbf{w}^n + \alpha^n \mathbf{s}^n$ *und gehe zu (S.6).*
Sonst gehe zu (S.5).

(S.5) *Gradientenprojektion:*
Setze $\mathbf{w}^n(t) := \mathcal{P}_{\mathcal{S}}\big(\mathbf{w}^n - tF(\mathbf{w}^n)\big)$ *mit* $\mathcal{P}_{\mathcal{S}}$ *aus (4.36).*
Bestimme das größte $\alpha \in [0,1]$*, so dass mit* $t^n := c \cdot \alpha$

$$J(\mathbf{x}^n(t^n), \mathbf{p}^n(t^n)) \;\leq\; J(\mathbf{x}^n, \mathbf{p}^n) - \sigma F_{\mathbf{x},p}(\mathbf{w}^n)^\top \begin{pmatrix} \mathbf{x}^n - \mathbf{x}^n(t^n) \\ \mathbf{p}^n - \mathbf{p}^n(t^n) \end{pmatrix}$$

gilt und setze $\mathbf{w}^{n+1} := \mathbf{w}^n(t^n)$.

(S.6) *Abbruchbedingung:*
Ist $\dfrac{\|\mathbf{w}^{n+1} - \mathbf{w}^n\|_2}{\|\mathbf{w}^0\|_2} < \varepsilon$*: Ein lokales Minimum von (4.2) ist gefunden. Stop.*
Sonst gehe zu (S.1).

Die Aktiv- bzw. Inaktivsetzung in Algorithmus 4.3.1 wird demnach wie folgt durchgeführt: Sei $(\mathbf{x}^n, \mathbf{p}^n, \mu^n) \in S \times \mathbb{R}$ gegeben. Setzt man $(\mathbf{x}^n, \mathbf{p}^n, \mu^n)$ in System (4.37) ein und löst dieses, dann

(1) repräsentiert die Lösung $(\mathbf{s}_x{}^\top, \mathbf{s}_p{}^\top, s_\mu)^\top$ entweder eine Abstiegsrichtung bezüglich des Zielfunktionals

(2) oder $(\mathbf{s}_x{}^\top, \mathbf{s}_p{}^\top, s_\mu)^\top$ ist keine Abstiegsrichtung bezüglich des Zielfunktionals.

Fall (1): Sei $(\mathbf{s}_x{}^\top, \mathbf{s}_p{}^\top, s_\mu)^\top$ eine Abstiegsrichtung, dann existiert ein $\tilde{\alpha} \in [0,1]$, so dass mit

$$(\mathbf{x}^{n+1}(\alpha), \mathbf{p}^{n+1}(\alpha), \mu^{n+1}(\alpha)) := (\mathbf{x}^n + \alpha \mathbf{s}_x, \mathbf{p}^n + \alpha \mathbf{s}_p, \mu^n + \alpha s_\mu)$$

die Ungleichung

$$J(\mathbf{x}^{n+1}(\alpha), \mathbf{p}^{n+1}(\alpha)) \;\leq\; J(\mathbf{x}^n, \mathbf{p}^n)$$

für alle Schrittweiten $\alpha \in [0, \tilde{\alpha}]$ erfüllt ist. Insbesondere existiert ein

$$\alpha_{\max} \;=\; \max\{\alpha \in [0,1] : (\mathbf{x}^{n+1}(\alpha), \mathbf{p}^{n+1}(\alpha), \mu^{n+1}(\alpha)) \in S\}.$$

Ist $\alpha_{\max} < 1$, dann existiert mindestens ein Punkt $x_k^{i,n+1}(\alpha_{\max})$ oder $p_i^{n+1}(\alpha_{\max})$, der auf dem Rand des zulässigen Bereichs S liegt. Diese Punkte werden aktiv gesetzt.

Fall (2): Ist $(\mathbf{s}_x{}^\top, \mathbf{s}_p{}^\top, s_\mu)^\top$ keine Abstiegsrichtung, dann wird als Richtung der Gradient $\mathbf{s} = -F(\mathbf{w}^n)$ verwendet, da dieser immer in Richtung des steilsten Abstiegs bezüglich des Zielfunktionals zeigt. Um die Iterierte $\mathbf{w}^{n+1} \in \mathcal{S} \times \mathbb{R}$ zu erhalten, wird nun das maximale $\alpha \in [0,1]$ derart bestimmt, so dass mit der Schrittweite $t^n = c \cdot \alpha$ und $\mathbf{w}^{n+1} := \mathcal{P}_{\mathcal{S}}\Big(\mathbf{w}^n + t^n \mathbf{s}\Big)$ das Zielfunktional wie in Algorithmus 4.3.1 beschrieben verkleinert wird. Es werden nun alle x_k^i aktiv gesetzt, für die entweder $x_k^i = a_k$ oder $x_k^i = b_k$ gilt und es werden alle p_i aktiv gesetzt mit $p_i = 0, i \in I_\ell$.

4.3.2 Konvergenz

Die folgenden Konvergenzaussagen sind der Veröffentlichung [26] entnommen.

Satz 4.3.3. (Konvergenz, [26])
Unter der Voraussetzung (4.11) sei $(\mathbf{x}^*, \mathbf{p}^*) \in \mathcal{S}$ *eine optimale Lösung des Optimierungsproblems (4.2). Dann generiert der Algorithmus 4.3.1 mit dem Startwert* $\mathbf{w}^0 := (\mathbf{x}^0, \mathbf{p}^0, \mu^0) \in \mathcal{S}^0 \times \mathbb{R}$ *eine Folge* $(\mathbf{w}^n)_{n=0}^{\infty}$ *mit:*

(i) Die Folge $(\mathbf{w}^n)_{n=0}^{\infty}$ *konvergiert gegen* $\mathbf{w}^* := (\mathbf{x}^*, \mathbf{p}^*, \mu^*) \in \mathcal{S} \times \mathbb{R}$*, wobei* $\mu^* \in \mathbb{R}$ *die duale Lösung darstellt.*

(ii) Die Konvergenzgeschwindigkeit ist global q-superlinear, d. h. es gilt global

$$\lim_{n\to\infty} \frac{\|\mathbf{w}^{n+1} - \mathbf{w}^*\|_2}{\|\mathbf{w}^n - \mathbf{w}^*\|_2} \;=\; 0.$$

(iii) Die Konvergenzgeschwindigkeit ist lokal q-quadratisch, d. h. es gilt lokal mit einer Konstanten $C > 0$

$$\|\mathbf{w}^{n+1} - \mathbf{w}^*\|_2 \;\leq\; C\|\mathbf{w}^n - \mathbf{w}^*\|_2^2.$$

4.4 Verfahren bei schwacher Differenzierbarkeit

In den Kapiteln 4.2 und 4.3 wurden zwei Verfahren vorgestellt, mit denen das Optimierungsproblem (4.2) gelöst werden kann, wenn die Lösungen der Sensitivitätsgleichungen erster Ordnung (2.33) bezüglich der Ortsvariablen $\mathbf{x}$ zweimal stetig differenzierbar sind, d. h. es wurde vorausgesetzt, dass für jedes feste $t \in T$ und $\boldsymbol{\theta} \in \Theta$ mit $u(\mathbf{x}) := u(t, \mathbf{x}, \boldsymbol{\theta})$

$$\nabla_{\boldsymbol{\theta}} u(\mathbf{x}) \in \mathcal{C}^2(\Omega; \mathbb{R}^m) \tag{4.38}$$

gilt. Diese Verfahren können allerdings nicht angewendet werden, wenn die Voraussetzung (4.38) nicht erfüllt ist.

Da die Sensitivitäten $\nabla_{\boldsymbol{\theta}} u(\mathbf{x})$ als Finite-Element-Lösung im allgemeinen H^1-Funktionen sind und daraus resultiert, dass das zu minimierende Zielfunktional $J(\mathbf{x}, \mathbf{p})$ im Ort ebenfalls eine H^1-Funktion ist, wird in diesem Abschnitt eine Methode vorgestellt, mit der das Optimierungsproblem (4.2) gelöst werden kann, obwohl die erste schwache Ableitung von $J(\mathbf{x}, \mathbf{p})$ bezüglich $\mathbf{x}$ eine L^2-Funktion darstellt und somit unstetig sein kann.

Ferner wird angenommen, dass die Sensitivitäten im Ort genau einmal schwach differenzierbar sind, da im Fall einer höheren Regularität auf ein Newtonartiges Verfahren (wie in Kapitel 4.2 und 4.3 beschrieben) zurückgegriffen werden kann. Somit ist

$$\nabla_{\boldsymbol{\theta}} u(\mathbf{x}) \in H^1(\Omega, \mathbb{R}^m), \quad \nabla_{\boldsymbol{\theta}} u(\mathbf{x}) \notin H^2(\Omega, \mathbb{R}^m),$$

und folglich gilt für fest gewähltes $\mathbf{p} \in \mathbb{R}^\ell$ mit $p_i \geq 0, i \in I_\ell$

$$J(\mathbf{x}, \mathbf{p}) \in H^1(\Omega, \mathbb{R}^m), \quad J(\mathbf{x}, \mathbf{p}) \notin H^2(\Omega, \mathbb{R}^m). \tag{4.39}$$

Um eine Lösung von (4.2) numerisch berechnen zu können, kann ein Gradientenverfahren genutzt werden. Da allerdings das Zielfunktional bezüglich der Gewichte

$p_i > 0$, $i \in I_\ell$ weiterhin zweimal stetig differenzierbar ist, werden die Variablen $\mathbf{x}$ und $\mathbf{p}$ getrennt voneinander betrachtet um das Optimierungsproblem (4.2) zweistufig lösen zu können. Dann kann iterativ das optimale $\bar{\mathbf{p}}$ mit dem Newtonverfahren und das optimale $\bar{\mathbf{x}}$ mit der Methode des steilsten Abstiegs bestimmt werden. Dadurch erhält man zwar ein Verfahren von der selben Konvergenzordnung wie beim klassischen Gradientenverfahren, dafür löst man zwei wesentlich kleinere Probleme, von denen eines sogar global quadratisch konvergiert.

Um ein zweistufiges Optimierungsproblem zu erhalten, wird der zulässige Bereich S aus (4.5) zunächst in die Teilbereiche $\bar{\Omega}^\ell$ mit

$$\bar{\Omega}^\ell = \{\mathbf{x} = ({\mathbf{x}^1}^\top, \ldots, {\mathbf{x}^\ell}^\top)^\top \in \mathbb{R}^{d\ell} : \mathbf{a} \leq \mathbf{x}^i \leq \mathbf{b} \in \mathbb{R}^d \; \forall\, i \in I_\ell\}$$

und

$$S_p := \{\mathbf{p} = (p_1, \ldots, p_\ell) \in \mathbb{R}^\ell : \; p_i \geq 0 \; \forall\, i \in I_\ell, \; \sum_{i=1}^{\ell} p_i = 1\}$$

zerlegt. So kann das Optimierungsproblem (4.2) umgeschrieben werden zu

$$\min J(\mathbf{x}, \mathbf{p}) = -\ln |M(\mathbf{x}, \mathbf{p})| \quad \text{u.d.N.} \quad \mathbf{x} \in \bar{\Omega}^\ell, \, \mathbf{p} \in S_p. \tag{4.40}$$

Es gilt:

- Nach Satz 4.2.5 ist für jedes fest gewählte $\tilde{\mathbf{x}} \in \bar{\Omega}^\ell$ das Optimierungsproblem

  $$\min J(\tilde{\mathbf{x}}, \mathbf{p}) = -\ln |M(\tilde{\mathbf{x}}, \mathbf{p})| \quad \text{u.d.N.} \quad \mathbf{p} \in S_p, \tag{4.41}$$

  konvex. Da $J(\tilde{\mathbf{x}}, \mathbf{p}) \in \mathcal{C}^2(S_p; \mathbb{R})$ ist, kann das Problem (4.41) mit dem Innere-Punkte-Verfahren aus Kapitel 4.2 oder mit der Active-Set-Methode aus Kapitel 4.3 gelöst werden. Da bei der Active-Set-Methode ein kleineres System als beim Innere-Punkte-Verfahren zu lösen ist und beide Verfahren die selbe Konvergenzordnung aufweisen, ist die Active-Set-Methode dem Innere-Punkte-Verfahren vorzuziehen.

- Für fest gewähltes $\tilde{\mathbf{p}} \in S_p$ kann das Optimierungsproblem

  $$\min J(\mathbf{x}, \tilde{\mathbf{p}}) = -\ln |M(\mathbf{x}, \tilde{\mathbf{p}})| \quad \text{u.d.N.} \quad \mathbf{x} \in \bar{\Omega}^\ell, \tag{4.42}$$

  nicht konvex sein. Da nach (4.39) $J(\mathbf{x}, \tilde{\mathbf{p}}) \in H^1(\Omega^\ell; \mathbb{R})$ ist, kann das Problem (4.42) weder mit dem Innere-Punkte-Verfahren aus Kapitel 4.2, noch mit der Active-Set-Methode aus Kapitel 4.3 gelöst werden. Daher wird nun ein Gradientenverfahren vorgestellt, welches das Problem (4.42) numerisch lösen kann.

4.4.1 Gradientenverfahren

Für fest gewähltes $\tilde{\mathbf{p}} \in S_p$ kann das Optimierungsproblem (4.42) mit der Methode des steilsten Abstiegs gelöst werden. Hierbei handelt es sich um ein Gradientenverfahren, welches iterativ eine optimale Lösung $\bar{\mathbf{x}} \in \bar{\Omega}^\ell$ bestimmen kann. Der Algorithmus lautet:

Algorithmus 4.4.1. (Methode des steilsten Abstiegs)

(S.0) *Setze* $n := 0$. *Wähle ein* $\mathbf{x}^0 \in \Omega^\ell$ *und* $\varepsilon > 0$ *hinreichend klein.*

(S.1) *Solange* $\|\nabla_{\mathbf{x}} J(\mathbf{x}^n, \tilde{\mathbf{p}})\|_2 > \varepsilon$*:*

Bestimme eine geeignete Schrittweite $\alpha^n \in (0,1)$, *so dass*
$\mathbf{x}^{n+1} := \mathbf{x}^n - \alpha^n \nabla_{\mathbf{x}} J(\mathbf{x}^n, \tilde{\mathbf{p}}) \in \bar{\Omega}^\ell$ *und*
$J(\mathbf{x}^{n+1}, \tilde{\mathbf{p}}) < J(\mathbf{x}^n, \tilde{\mathbf{p}})$ *gilt.*
Setze $n := n + 1$.

(S.2) *Ist* $\|\nabla_{\mathbf{x}} J(\mathbf{x}^n, \tilde{\mathbf{p}})\|_2 \leq \varepsilon$*: Eine optimale Lösung ist gefunden. Stop.*

Der Gradient $\nabla_{\mathbf{x}} J(\mathbf{x}, \tilde{\mathbf{p}}) \in L^2(\Omega^\ell; \mathbb{R}^{d\ell})$ ist auf Nullmengen allerdings nicht definiert und kann in einem Punkt $\tilde{\mathbf{x}} \in \bar{\Omega}^\ell$ nicht ohne weiteres ausgewertet werden. Daher werden die Komponenten

$$\partial_{x_k^i} J(\tilde{x}, \tilde{p}) \;=\; -2\tilde{p}_i \int_T \partial_{x_k^i} \left[\nabla_{\boldsymbol{\theta}} u(t, \tilde{\mathbf{x}}^i)\right]^\top M(\tilde{\mathbf{x}}, \tilde{\mathbf{p}})^{-1} \nabla_{\boldsymbol{\theta}} u(t, \tilde{\mathbf{x}}^i) dt, \tag{4.43}$$

$\forall\, i \in I_\ell,\ k \in K_d$ mithilfe der Delta-Distribution $\delta(y)$ [45] berechnet.

Definition 4.4.1. (Delta-Distribution, Dirac-Folge)
Die Delta-Distribution ist eine stetig lineare Abbildung $\delta : \mathbb{R} \to \mathbb{R}_0^+$ *mit der Eigenschaft*

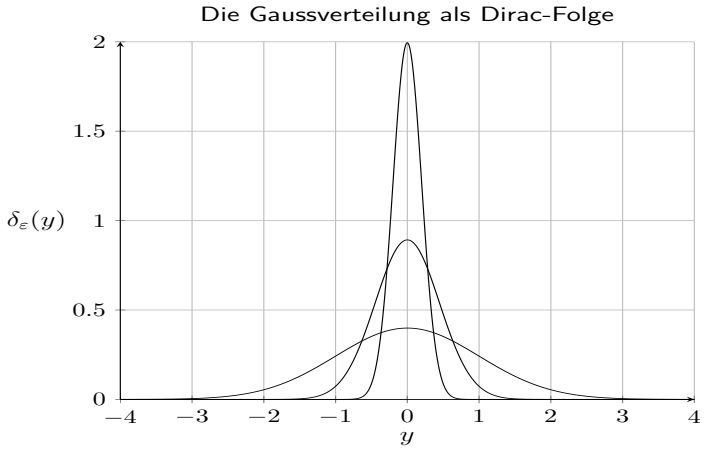
$$\delta(y) = \left\{ \begin{array}{ll} +\infty, & y = 0 \\ 0, & y \neq 0 \end{array} \right\} \quad \textit{und} \quad \int_a^b \delta(y)\, dy = \left\{ \begin{array}{ll} 1, & 0 \in [a,b] \\ 0, & 0 \notin [a,b] \end{array} \right\}.$$

Es gilt stets $\displaystyle\int_a^b f(x)\delta(x - x_0)\, dx = \left\{ \begin{array}{ll} f(x_0), & x_0 \in [a,b] \\ 0, & x_0 \notin [a,b] \end{array} \right\}.$

Die Delta-Distribution kann als Grenzwert einer schwach konvergenten Funktionalfolge über symmetrische Verteilungsfunktionen $\delta_\varepsilon(x)$ *definiert werden, die Dirac-Folge genannt wird. Eine typische Dirac-Folge ist die Gaussverteilung*

$$\delta_\varepsilon(y) := \frac{1}{\sqrt{2\pi\varepsilon}} \exp\left(-\frac{y^2}{2\varepsilon}\right), \tag{4.44}$$

die für $\varepsilon \to 0^+$ *gegen die Delta-Distribution konvergiert. Der Träger der Gaussverteilung kann durch* $\operatorname{supp}(\delta_\varepsilon(y)) \approx [-3\sqrt{\varepsilon}, 3\sqrt{\varepsilon}]$ *beschrieben werden.*



Um den Gradienten $\nabla_{\mathbf{x}} J(\mathbf{x}, \tilde{\mathbf{p}})$ in $\tilde{\mathbf{x}} = ({\tilde{\mathbf{x}}^1}^\top, \ldots, {\tilde{\mathbf{x}}^\ell}^\top)^\top \in \bar{\Omega}^\ell$ bei fest gewähltem $\tilde{\mathbf{p}} \in \mathcal{S}_p$ auszuwerten, wird zunächst die L^2-Funktion $\partial_{x_k^i} \left[\nabla_{\boldsymbol{\theta}} u(t, \tilde{\mathbf{x}}^i)\right] \; \forall\, i \in I_\ell$ in $\tilde{\mathbf{x}}^i \in \bar{\Omega}$ im distributionellen Sinn bestimmt und anschließend in (4.43) eingesetzt. Da die Inverse der Fisher-Informationsmatrix $M(\tilde{\mathbf{x}}, \tilde{\mathbf{p}})^{-1}$ und die Finite-Element Lösung $\nabla_{\boldsymbol{\theta}} u(t, \tilde{\mathbf{x}}^i)$ wie üblich ausgewertet werden können, erhält man dann die gesuchte Größe.

Zur übersichtlichen Darstellung werden im folgenden die Sensitivitäten durch

$$w_j(t, \mathbf{x}) := \partial_{\theta_j} u(t, \mathbf{x}), \quad j = 1, \ldots, m$$

dargestellt. Dann gilt mit der Gauss-Verteilung $\delta_\varepsilon(y)$ aus (4.44) und der Menge

$$B(\mathbf{x}^i, \varepsilon) := \{\mathbf{y} \in \bar{\Omega} : \; \|\mathbf{x}^i - \mathbf{y}\|_2 \leq 3\sqrt{\varepsilon}\}$$

die Approximation

$$\partial_{x_k^i} w_j(t, \mathbf{x}^i) \;\; \approx \;\; \lim_{\varepsilon \to 0} \int_{B(x^n, \varepsilon)} \partial_{x_k^i} w_j(t, \mathbf{y}) \prod_{k=1}^{d} \delta_\varepsilon(y_k - x_k^i)\, dy, \tag{4.45}$$

für $i \in I_\ell$ und $k \in K_d$.

Für die Berechnung des Funktionswertes $\partial_{x_k^i} w_j(t, \mathbf{x}^i)$ werden nun zwei Fälle unterschieden. Zunächst wird angenommen, dass der Punkt $\mathbf{x}^i$ im strikten Inneren einer Zelle $T_m \subset \mathcal{T}_h$[3] liegt. Da in diesem Fall immer ein $\varepsilon_0 > 0$ existiert, so dass $B(\mathbf{x}^i, \varepsilon) \subset T_m \; \forall\, \varepsilon \in (0, \varepsilon_0)$ gilt, hängt der gesuchte Funktionswert ausschließlich von der Sensitivität $w_j(t, \mathbf{x}^i)_{|_{T_m}}$ ab. Um den gesuchten Funktionswert zu erhalten, wird $w_j(t, \mathbf{x}^i)_{|_{T_m}}$ als Linearkombination

$$w_j(t, \mathbf{x}^i)_{|_{T_m}} \;=\; \sum_{r=1}^{N} w_{j,r}(t) \varphi_r(\mathbf{x}^i)_{|_{T_m}} \;=\; \sum_{r=1}^{N} w_{j,r}(t) \varphi_r(\mathbf{x}^i) \mathbb{1}_{S_r \cap T_m}(\mathbf{x}^i),$$

mit der Indikatorfunktion $\mathbb{1}$ beschrieben, wobei $(\varphi_r(\mathbf{x}))_{r=1}^N$ die Ansatzfunktionen sind und $S_r := \mathrm{supp}(\varphi_r)$ ist. Da $\varphi_r(\mathbf{x}) \; \forall r = 1, \ldots, N$ stückweise linear ist, gilt folglich

$$\partial_{x_k^i} w_j(t, \mathbf{x}^i)_{|_{T_m}} \;=\; \sum_{r=1}^{N} w_{j,r}(t)\, \partial_{x_k^i} \varphi_r(\mathbf{x}^i)_{|_{T_m}} \;=\; \sum_{r=1}^{N} w_{j,r}(t) c_{m,r,k} \mathbb{1}_{S_r \cap T_m}(\mathbf{x}^i), \tag{4.46}$$

so dass mit den Konstanten $c_{m,r,k} := \partial_{x_k^i} \varphi_r(\mathbf{x}^i)_{|_{T_m}}$ die Gleichung (4.45) umformuliert werden kann zu

$$\begin{aligned}
\partial_{x_k^i} w_j(t, \mathbf{x}^i) &\overset{(4.46)}{=} \lim_{\varepsilon \to 0} \int_{B(\mathbf{x}^i, \varepsilon)} \left(\sum_{r=1}^{N} w_{j,r}(t) \partial_{x_k^i} \varphi_r(\mathbf{x}^i) \right) \prod_{k=1}^{d} \delta_\varepsilon(y_k - x_k^i)\, dy \\
&\overset{(4.46)}{=} \lim_{\varepsilon \to 0} \int_{B(\mathbf{x}^i, \varepsilon)} \underbrace{\left(\sum_{r=1}^{N} w_{j,r}(t) c_{m,r,k} \mathbb{1}_{S_r \cap T_m}(\mathbf{x}^i) \right)}_{\text{räumlich konstant}} \prod_{k=1}^{d} \delta_\varepsilon(y_k - x_k^i)\, dy \\
&= \left(\sum_{r=1}^{N} w_{j,r}(t) c_{m,r,k} \mathbb{1}_{S_r \cap T_m}(\mathbf{x}^i) \right) \lim_{\varepsilon \to 0} \underbrace{\int_{B(\mathbf{x}^i, \varepsilon)} \prod_{k=1}^{d} \delta_\varepsilon(y_k - x_k^i)\, dy}_{=1 \; \forall \varepsilon > 0}.
\end{aligned}$$

[3] $\mathcal{T}_h$ ist die Menge aller Zellen T_i mit $\bigcup_{i=1}^N T_i = \Omega$ und $T_i \cap T_j$ für $i \neq j$ ist eine Nullmenge.

Wenn $\mathbf{x}^i$ im strikten Inneren der Zelle T_m liegt kann der gewünschte Funktionswert demnach wie folgt berechnet werden:

$$\partial_{x_k^i} w_j(t, \mathbf{x}^i) \;=\; \sum_{r=1}^{N} w_{j,r}(t) c_{m,r,k} \mathbb{1}_{S_r \cap T_m}(\mathbf{x}^i). \tag{4.47}$$

Jetzt wird abschließend gezeigt, wie $\partial_{x_k^i} w_j(t, \mathbf{x}^i)$ bestimmt werden kann, wenn der Punkt $\mathbf{x}^i$ auf dem Rand mindestens einer Zelle $T_m \subset \mathcal{T}_h$ liegt. In diesem Fall wird wiederum ausgenutzt, dass die partielle Ableitung der Sensitivitäten in Richtung x_k^i als Linearkombination (4.46) geschrieben werden kann. Da in diesem Fall immer ein $\varepsilon_0 > 0$ existiert, so dass $B(\mathbf{x}^i, \varepsilon) \subset \bigcup\{T_m \in \mathcal{T}_h : \mathbf{x}^i \in T_m\}$ für alle $\varepsilon \in (0, \varepsilon_0]$ ist, kann Gleichung (4.45) umformuliert werden zu

$$\begin{aligned}
\partial_{x_k^i} w_j(t, \mathbf{x}^i) &\overset{(4.46)}{=} \lim_{\varepsilon \to 0} \int_{B(\mathbf{x}^i, \varepsilon)} \left(\sum_{r=1}^{N} w_{j,r}(t) \partial_{x_k^i} \varphi_r(\mathbf{x}^i) \right) \prod_{k=1}^{d} \delta_\varepsilon(y_k - x_k^i)\, dy \\
&\overset{(4.46)}{=} \lim_{\varepsilon \to 0} \sum_{T_m \in \mathcal{T}_h} \int_{T_m \cap B(\mathbf{x}^i, \varepsilon)} \underbrace{\left(\sum_{r=1}^{N} w_{j,r}(t) c_{m,r,k} \mathbb{1}_{S_r \cap T_m}(\mathbf{x}^i) \right)}_{\text{räumlich konstant}} \prod_{k=1}^{d} \delta_\varepsilon(y_k - x_k^i)\, dy \\
&= \lim_{\varepsilon \to 0} \sum_{T_m \in \mathcal{T}_h} \left(\sum_{r=1}^{N} w_{j,r}(t) c_{m,r,k} \mathbb{1}_{S_r \cap T_m}(\mathbf{x}^i) \right) \int_{T_m \cap B(\mathbf{x}^i, \varepsilon)} \prod_{k=1}^{d} \delta_\varepsilon(y_k - x_k^i)\, dy \\
&\overset{(4.44)}{=} \lim_{\varepsilon \to 0} \sum_{T_m \in \mathcal{T}_h} \left(\sum_{r=1}^{N} w_{j,r}(t) c_{m,r,k} \mathbb{1}_{S_r \cap T_m}(\mathbf{x}^i) \right) \frac{\mu(T_m \cap B(\mathbf{x}^i, \varepsilon))}{\mu(B(\mathbf{x}^i, \varepsilon))},
\end{aligned}$$

wobei $\mu(T_m \cap B(\mathbf{x}^i, \varepsilon))$ das Volumen von $T_m \cap B(\mathbf{x}^i, \varepsilon)$ und $\mu(B(\mathbf{x}^i, \varepsilon))$ das Volumen der Menge $B(\mathbf{x}^i, \varepsilon)$ darstellt. Für die Gaussverteilung (4.44) ist das Verhältnis $\mu(T_m \cap B(\mathbf{x}^i, \varepsilon)) : \mu(B(\mathbf{x}^i, \varepsilon))$ für jedes $\varepsilon \in (0, \varepsilon_0]$ gleich und stellt somit eine positive Konstante dar. Somit kann der gesuchte Funktionswert bestimmt werden durch

$$\partial_{x_k^i} w_j(t, \mathbf{x}^i) \;=\; \sum_{T_m \in \mathcal{T}_h} \sum_{r=1}^{N} w_{j,r}(t) c_{m,r,k} \, \frac{\mu(T_m \cap B(\mathbf{x}^i, \varepsilon_0))}{\mu(B(\mathbf{x}^i, \varepsilon_0))}. \tag{4.48}$$

Gleichung (4.47) ist ein Spezialfall von Gleichung (4.48), denn falls $\mathbf{x}^i$ im Inneren einer Zelle $T_m \in \mathcal{T}_h$ liegt, gilt

$$\frac{\mu(T_m \cap B(\mathbf{x}^i, \varepsilon_0))}{\mu(B(\mathbf{x}^i, \varepsilon_0))} \;=\; \begin{cases} 1, & B(\mathbf{x}^i, \varepsilon_0) \subset T_m, \\ 0, & B(\mathbf{x}^i, \varepsilon_0) \notin T_m. \end{cases}$$

4.4.2 Algorithmus

Nachdem beschrieben wurde, wie man den Gradienten $F(\mathbf{x}, \tilde{\mathbf{p}}) := \nabla_{\mathbf{x}} J(\mathbf{x}, \tilde{\mathbf{p}})$ mit

$$F_i(\mathbf{x}, \tilde{\mathbf{p}}) \;\in L^2(\Omega^\ell) \quad \forall i = 1, \ldots, d\ell,$$

in einem Punkt $\mathbf{x} \in \bar{\Omega}^\ell$ auswerten kann, wird nun der Algorithmus der zweistufigen Active-Set-Methode (ZASM) angegeben, mit dem man das Optimierungsproblem (4.2) im H^1-konformen Fall lösen kann.

Algorithmus 4.4.2. (Zweistufige Active-Set-Methode zur Bestimmung eines D-optimalen Designs)

(S.0) *Initialisierung:*
Wähle ein $\mathbf{x}^0 \in \Omega^\ell$ *und ein* $\mathbf{p}^0 \in S_p$ *und setze* $n := 0$.
Wähle ein $\sigma > 0$, $c > 0$ *und ein* $\varepsilon > 0$ *hinreichend klein.*

(S.1) *Lösen des konvexen Teilproblems (4.41):*
Bestimme mit dem Algorithmus 4.3.1 die optimalen Gewichte

$$\mathbf{p}^{n+1} := \min_{\mathbf{p} \in S_p} J(\mathbf{x}^n, \mathbf{p}).$$

(S.2) *Setzen der aktiven und inaktiven Menge für das Teilproblem (4.42):*
Setze

$$\mathcal{A}^n := \{k + (i-1)d : x_k^{i,n} = a_k \vee x_k^{i,n} = b_k\},$$
$$\mathcal{I}^n := \{1, \ldots, d\ell\} \setminus \mathcal{A}^n.$$

(S.3) *Berechnung der Suchrichtung:*
Bestimme die Richtung $\mathbf{s}^n$ *mithilfe von (4.43) und (4.48):*

$$s_i^n = -F_i(\mathbf{x}, \tilde{\mathbf{p}}), \quad \text{für } i \in \mathcal{I}^n,$$
$$s_i^n = 0, \qquad\qquad \text{für } i \in \mathcal{A}^n,$$

Ist $\|\mathbf{s}^n\|_2 < \varepsilon$*: Eine optimale Lösung ist gefunden. Stop.*
Sonst gehe zu (S.4).

(S.4) *Schrittweitenbestimmung und setzen der neuen Iterierten:*
Bestimme das größte $\alpha \in [0, 1]$*, so dass* $\mathbf{x}^n + \alpha \mathbf{s}^n \in \bar{\Omega}^\ell$.
Setze $\mathbf{x}^{n+1} := \mathbf{x}^n + \alpha \mathbf{s}^n$, $n := n + 1$ *und gehe zu (S.1).*

4.4.3 Komplexität

Folgende Tabelle gibt einen Überblick über die Komplexität der einzelnen Schritte von Algorithmus 4.4.2 in Abhängigkeit von

- der Anzahl der unbekannten, zu schätzenden Parameter $m > 0$,
- der Schrittweite in der Zeit $\Delta t > 0$
- und der Schrittweite $\Delta\alpha > 0$.

	Anzahl der erforderlichen Operationen	Speicheraufwand
(S.0)	$\mathcal{O}(1)$	$\mathcal{O}(m^2)$
(S.1)	$\mathcal{O}(m^4(\Delta t)^{-1})$	$\mathcal{O}(m^2)$
(S.2)	$\mathcal{O}(1)$	$\mathcal{O}(m^2)$
(S.3)	$\mathcal{O}(m^4(\Delta t)^{-1})$	$\mathcal{O}(m^2)$
(S.4)	$\mathcal{O}(m^2(\Delta\alpha)^{-1})$	$\mathcal{O}(1)$

Tabelle 4.1: Komplexität der zweistufigen Active-Set-Methode zur Bestimmung eines D-optimalen Designs.

Kapitel 5

Vergleich zweier Verfahren zur numerischen Bestimmung eines D-optimalen Designs am Beispiel eines instationären Problems

Nachdem in den Kapiteln 3 und 4 zwei Möglichkeiten zur Bestimmung eines D-optimalen Designs vorgestellt wurden, werden diese in diesem Kapitel am Beispiel einer zweidimensionalen Wärmeleitungsgleichung angewendet und anschließend miteinander verglichen. Dieses Beispiel ist dem Buch [85] entnommen.

Das in Kapitel 3 beschriebene Lösungsverfahren [85] basiert auf einer *Exchange-type Methode* [82], wohingegen das in Kapitel 4.2 hergeleitete *Innere-Punkte-Verfahren* und die in Kapitel 4.3 beschriebene *Active-Set-Methode* das Optimierungsproblem klassisch lösen. Da - wie bereits in Kapitel 4 erwähnt - die *Active-Set-Methode* dem *Innere-Punkte-Verfahren* vorzuziehen ist, wird in dieser Arbeit ein D-optimales Design am Beispiel einer zweidimensionalen Wärmeleitungsgleichung mit der *Exchange-type Methode* und der *Active-Set-Methode* gelöst.

Dieses Kapitel gliedert sich wie folgt: Nachdem zunächst das Beispiel einer zweidimensionalen Wärmeleitungsgleichung vorgestellt wird, wird anschließend beschrieben, wie diese mit der Methode der finiten Elemente und dem Crank-Nicolson Verfahren gelöst werden kann. Die zur Bestimmung eines optimalen Designs benötigten Sensitivitäten werden durch Lösen der sogenannten Sensitivitätsgleichungen bestimmt. Diese Gleichungen, die selbst parabolische Differentialgleichungen darstellen, werden in diesem Kapitel aufgestellt und numerisch gelöst. Abschließend wird die *Exchange-type Methode* mit der *Active-Set-Methode* verglichen, indem für unterschiedliche Anzahl von Freiheitsgraden sowohl im Raum als auch in der Zeit jeweils ein D-optimales Design bestimmt wird. Dann können durch Bestimmung des relativen Fehlers in der Lösung und dem benötigten Zeitaufwand beide Methoden miteinander verglichen werden.

5.1 Beschreibung des zweidimensionalen Wärmeleitungsexperimentes

Gegeben sei eine dünne, quadratische Metallplatte mit wärmeisolierter Oberfläche, die im folgenden durch die abgeschlossene Menge $\bar{\Omega} = [0,1]^2$ dargestellt wird:

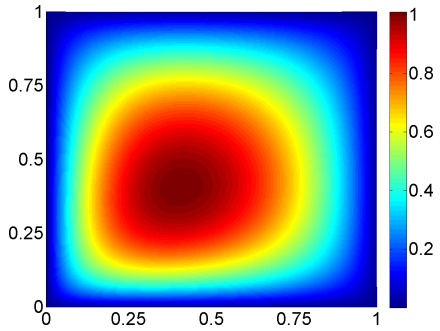


Abbildung 5.1: Temperaturverteilung einer quadratischen Metallplatte.

Es soll herausgefunden werden, um welches Metall es sich handelt. Wie in [85] beschrieben kann zur Realisierung dessen eine Parameteridentifikation genutzt werden, die den Diffusionskoeffizienten

$$\kappa(\mathbf{x}, \boldsymbol{\theta}) \quad = \quad \theta_1 + x_1\theta_2 + x_2\theta_3, \tag{5.1}$$

mit $\mathbf{x} = (x_1, x_2)^\top \in \Omega$ und $\boldsymbol{\theta} := (\theta_1, \theta_2, \theta_3)^\top \geq 0$ ermittelt und auf diese Weise das Material identifizieren kann.

Um den Diffusionskoeffizienten schätzen zu können, werden im vorhinein $M > 0$ Experimente durchgeführt. Diese werden wie folgt bewerkstelligt:

(1) Die Metallplatte wird zunächst auf konstant $5°C$ in ganz Ω heruntergekühlt.

(2) Hat die gesamte Metallplatte die konstante Temperatur von $5°C$ erreicht, beginnt das eigentliche Experiment: Es werden $\ell \in \mathbb{N}$ Sensoren auf der Metallplatte angebracht um die Temperatur an diesen Stellen ermitteln zu können.

(3) Eine Minute lang - dargestellt durch das Intervall $T := (0,1)$ - wird nun der Rand der Metallplatte konstant von $5°C$ zu $0°C$ heruntergekühlt. In dieser Zeit wird mit den Sensoren die Temperatur gemessen.

Notation: Die im k-ten Experiment ermittelten Messdaten werden mit $z_i^k(t), i \in I_\ell$ bezeichnet. Nach M Versuchen erhält man die gemittelten Messreihen

$$\tilde{z}_i(t) = \frac{1}{M}\sum_{k=1}^{M} z_i^k(t).$$

Zweidimensionale Wärmeleitungsgleichung als parabolische Differentialgleichung

Das beschriebene Experiment kann durch die parabolische Differentialgleichung

$$\partial_t u(t, \mathbf{x}, \boldsymbol{\theta}) = \nabla_{\mathbf{x}} \cdot [\kappa(\mathbf{x}, \boldsymbol{\theta}) \nabla_{\mathbf{x}} u(t, \mathbf{x}, \boldsymbol{\theta})], \quad \mathbf{x} \in \Omega, \quad t \in T \tag{5.2}$$

mit der Anfangsbedingung

$$u(0, \mathbf{x}, \boldsymbol{\theta}) = 5, \quad \mathbf{x} \in \Omega \tag{5.3}$$

und der Randbedingung

$$u(t, \mathbf{x}, \boldsymbol{\theta}) = 5(1-t), \quad \mathbf{x} \in \partial\Omega, \quad t \in T \tag{5.4}$$

dargestellt werden [85].

Parameteridentifikation

Wie in Kapitel 2.2 beschrieben, können die unbekannten Parameter $\boldsymbol{\theta}^* := (\theta_1^*, \theta_2^*, \theta_3^*)^\top$ mithilfe eines Maximum-Likelihood-Schätzers $\hat{\boldsymbol{\theta}} := (\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3)^\top$ näherungsweise ermittelt werden, indem das Optimierungsproblem

$$\hat{\boldsymbol{\theta}} \;=\; \arg\min_{\Theta} J(\boldsymbol{\theta}) \;=\; \frac{1}{2}\int_T \|\tilde{\mathbf{z}}(t) - \mathcal{P}u(t, \mathbf{x}, \boldsymbol{\theta})\|_2^2 \, dt$$

gelöst wird, wobei $u(t, \mathbf{x}, \boldsymbol{\theta})$ eine Lösung der parabolischen Differentialgleichung

$$\begin{aligned}
\partial_t u(t, \mathbf{x}, \boldsymbol{\theta}) &= \nabla_{\mathbf{x}} \cdot [\kappa(\mathbf{x}, \boldsymbol{\theta}) \nabla_{\mathbf{x}} u(t, \mathbf{x}, \boldsymbol{\theta})], \quad \mathbf{x} \in \Omega, \quad t \in T, \\
u(0, \mathbf{x}, \boldsymbol{\theta}) &= 5, \quad \mathbf{x} \in \Omega, \\
u(t, \mathbf{x}, \boldsymbol{\theta}) &= 5(1-t), \quad \mathbf{x} \in \partial\Omega, \quad t \in T,
\end{aligned}$$

und

$$\tilde{\mathbf{z}}(t) \;:=\; (\tilde{z}_1(t), \ldots, \tilde{z}_\ell(t))^\top$$

den Vektor der gemittelten Messreihen darstellt und

$$\mathcal{P}u(t, \mathbf{x}, \boldsymbol{\theta}) \;:=\; (u(t, x_1, \boldsymbol{\theta}), \ldots, u(t, x_\ell, \boldsymbol{\theta}))^\top$$

mit dem Projektionsoperator $\mathcal{P}$ aus Definition 2.2.1. Das Ergebnis $\hat{\boldsymbol{\theta}}$ liefert dann eine Schätzung des exakten Wertes $\boldsymbol{\theta}^*$.

Optimales Design

Wie in Kapitel 3 beschrieben kann die Genauigkeit des Schätzers $\hat{\boldsymbol{\theta}}$ maximiert werden, indem die zugehörige Kovarianzmatrix $\mathrm{cov}(\hat{\boldsymbol{\theta}})$ bezüglich des D-Optimalitätskriteriums (3.1) minimiert wird. Hierfür wird vorausgesetzt, dass der Messfehler $\varepsilon(t, \mathbf{x}, \boldsymbol{\theta}^*)$ normalverteilt ist mit

$$E\{\varepsilon(t, \mathbf{x}, \boldsymbol{\theta}^*)\} = 0, \quad E\{\varepsilon(t, \mathbf{x}, \boldsymbol{\theta}^*)\varepsilon(t', \mathbf{x}', \boldsymbol{\theta}^*)\} = \sigma^2 \delta_{ij} \delta(t - t').$$

Dann kann die Kovarianzmatrix $\mathrm{cov}(\hat{\boldsymbol{\theta}})$ durch die Inverse der Fisher-Informationsmatrix approximiert werden. In diesem Kapitel wird die Fisher-Informationsmatrix bezüglich (3.1) minimiert um eine optimale Messstellenkonstellation zu ermitteln. In Vorbereitung dessen wird zunächst beschrieben, wie die parabolische Differentialgleichung (5.2) mit (5.3) - (5.4) mit der Methode der finiten Elemente und dem Crank-Nicolson Verfahren gelöst werden kann.

5.2 Schwache Formulierung und Diskretisierung

Wie in Kapitel 2.2.2 erläutert, wird die parabolische Differentialgleichung (5.2) für jedes fest gewählte $t \in T$ mit der Methode der finiten Elemente [12] gelöst. Um die Existenz der zweiten Ableitung nicht fordern zu müssen, wird hierfür die Zustandsgleichung (5.2) in schwacher Form benötigt. Analog zur Vorgehensweise in Kapitel 2.2.2 wird die Gleichung (5.2) mit einer beliebigen, aber festen Testfunktion $\varphi \in H^1(\Omega)$ multipliziert und anschließend über Ω integriert. Für eine übersichtliche Darstellung ist im folgenden

$$L := L^2(\Omega), \quad V := H^1(\Omega) \quad \text{und} \quad V_0 := H_0^1(\Omega).$$

In diesem Kapitel wird vorausgesetzt, dass

$$u(t, \mathbf{x}, \boldsymbol{\theta}) \in \mathcal{C}^1(T; H^1(\Omega; \mathcal{C}^1(\Theta; \mathbb{R}))) \tag{5.5}$$

und mit $u(t)(\mathbf{x}) := u(t, \mathbf{x}, \boldsymbol{\theta})$ für jedes fest gewählte $\boldsymbol{\theta} \in \Theta$

$$u(t) \in V \tag{5.6}$$

gilt. Dann gilt für jedes $t \in T$ mit $\kappa(\mathbf{x}) := \kappa(\mathbf{x}, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$ fest gewählt:

$$\begin{aligned}\int_\Omega \Big(\partial_t u(t) - \nabla_{\mathbf{x}} \cdot (\kappa(\mathbf{x}) \nabla_{\mathbf{x}} u(t))\Big)\, \varphi \, dx \;&=\; \big(\partial_t u(t) - \nabla_{\mathbf{x}} \cdot (\kappa(\mathbf{x}) \nabla_{\mathbf{x}} u(t)), \varphi\big)_L \qquad (5.7)\\ &\overset{(5.1)}{=} \partial_t \big(u(t), \varphi\big)_L - \big(\kappa(\mathbf{x}) \Delta_{\mathbf{x}} u(t), \varphi\big)_L - \theta_2 \big(\partial_{x_1} u(t), \varphi\big)_L - \theta_3 \big(\partial_{x_2} u(t), \varphi\big)_L \;=\; 0\end{aligned}$$

$\forall\, \varphi \in V$.

Mit der ersten Greenschen Identität folgt

$$\begin{aligned}\big(\kappa(\mathbf{x}) \Delta_{\mathbf{x}} u(t), \varphi\big)_L \;&=\; \big((\theta_1 + x_1\theta_2 + x_2\theta_3) \Delta_{\mathbf{x}} u(t), \varphi\big)_L\\ &=\; \theta_1 \big(\Delta_{\mathbf{x}} u(t), \varphi\big)_L + \theta_2 \big(\Delta_{\mathbf{x}} u(t), x_1\, \varphi\big)_L + \theta_3 \big(\Delta_{\mathbf{x}} u(t), x_2\, \varphi\big)_L\\ &=\; -\theta_1 \big(\nabla_{\mathbf{x}} u(t), \nabla_{\mathbf{x}} \varphi\big)_L + \theta_1 \big(\nabla_{\mathbf{x}} u(t) \cdot \mathbf{n}, \varphi\big)_{L^2(\partial\Omega)}\\ &\quad\; - \theta_2 \big(\nabla_{\mathbf{x}} u(t), \nabla_{\mathbf{x}} (x_1 \varphi)\big)_L + \theta_2 \big(\nabla_{\mathbf{x}} u(t) \cdot \mathbf{n}, x_1\, \varphi\big)_{L^2(\partial\Omega)}\\ &\quad\; - \theta_3 \big(\nabla_{\mathbf{x}} u(t), \nabla_{\mathbf{x}} (x_2 \varphi)\big)_L + \theta_3 \big(\nabla_{\mathbf{x}} u(t) \cdot \mathbf{n}, x_2\, \varphi\big)_{L^2(\partial\Omega)},\\ \big(\partial_{x_1} u(t), \varphi\big)_L \;&=\; -\big(u(t), \partial_{x_1} \varphi(t)\big)_L + \big(u(t), \varphi\big)_{L^2(\partial\Omega)},\\ \big(\partial_{x_2} u(t), \varphi\big)_L \;&=\; -\big(u(t), \partial_{x_2} \varphi(t)\big)_L + \big(u(t), \varphi\big)_{L^2(\partial\Omega)}.\end{aligned}$$

Wie in [21] beschrieben, lautet die schwache Formulierung des inhomogenen Dirichletproblems (5.2) mit der Bilinearform $\alpha : V \times V \to \mathbb{R}$, definiert durch

$$\begin{aligned}a(v(t), \varphi) \;&=\; \theta_1 \big(\nabla_{\mathbf{x}} v(t), \nabla_{\mathbf{x}} \varphi\big)_L + \theta_2 \big(\nabla_{\mathbf{x}} v(t), \nabla_{\mathbf{x}} (x_1 \varphi)\big)_L + \theta_3 \big(\nabla_{\mathbf{x}} v(t), \nabla_{\mathbf{x}} (x_2 \varphi)\big)_L\\ &\quad + \big(v(t), \partial_{x_1} \varphi\big)_L + \big(v(t), \partial_{x_2} \varphi\big)_L \qquad (5.8)\end{aligned}$$

folgendermaßen:

Sei $\boldsymbol{\theta} \in \Theta$ fest gewählt. Gesucht ist für fest gewähltes $t \in T$ die Lösung $u(t) \in V$, so dass

$$\left.\begin{aligned} u(t) &= u_g(t) + v(t), \quad v(t) \in V_0, \\ \partial_t (v(t), \varphi)_L + a(v(t), \varphi) &= f(u_g(t), \varphi) \quad \forall\, \varphi \in V_0, \end{aligned}\right\} \tag{5.9}$$

mit der rechten Seite

$$f(u_g(t), \varphi) = -\partial_t (u_g(t), \varphi)_L - a(u_g(t), \varphi) \tag{5.10}$$

gilt, wobei $u_g(t) \in V$ so definiert ist, dass $u_g(t) = 1 - t$ auf dem Rand $\partial\Omega$ ist.

Im folgenden Abschnitt wird beschrieben, wie eine schwache Lösung von (5.9) mit der Methode der Finiten Elemente ermittelt werden kann.

5.2.1 Finite-Elemente-Diskretisierung im Ort

Um Gleichung (5.9) für jedes fest gewählte $t \in T$ mit der Methode der finiten Elemente numerisch lösen zu können, wird zunächst der Lösungsraum V_0 auf einen endlichdimensionalen Teilraum $V_{0,h}$ eingeschränkt. Man erhält dann das zu lösende, diskrete Variationsproblem:

Sei $\boldsymbol{\theta} \in \Theta$ fest gewählt. Gesucht ist $u_h(t) \in V_h$, so dass

$$\left.\begin{aligned} u_h(t) &= u_{g,h}(t) + v_h(t), \quad v_h(t) \in V_{0,h}, \\ \partial_t (v_h(t), \varphi_h)_L + a(v_h(t), \varphi_h) &= f(u_{g,h}(t), \varphi_h) \quad \forall\, \varphi_h \in V_{0,h}, \end{aligned}\right\} \tag{5.11}$$

wobei $u_{g,h}(t) \in V_h$ so gewählt wird, dass $u_{g,h}(t) = 1 - t$ auf dem Rand $\partial\Omega$ ist.

Im folgenden sei

$$V_h := \{v \in \mathcal{C}^0(\bar{\Omega}) : v_{|_{T_i}} \in \mathcal{P}_1 \ \forall\, T_i \in \mathcal{T}_h\}$$

mit dem Raum der Polynome vom Grad kleiner gleich 1

$$\mathcal{P}_1 := \{v(\mathbf{x}) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 \text{ mit } \alpha_0, \alpha_1, \alpha_2 \in \mathbb{R}\},$$

und

$$V_{0,h} := \{v \in V_h : v(\mathbf{x}) = 0 \ \forall\, \mathbf{x} \in \partial\Omega\}.$$

Um das diskrete Variationsproblem (5.11) zu lösen, wird das abgeschlossene Gebiet $\bar{\Omega} = [0,1]^2$ in kongruente Dreieckselemente $\mathcal{T}_h = \{T_1, T_2, \ldots, T_n\}$ - wie in Abbildung 5.2 dargestellt - zerlegt.

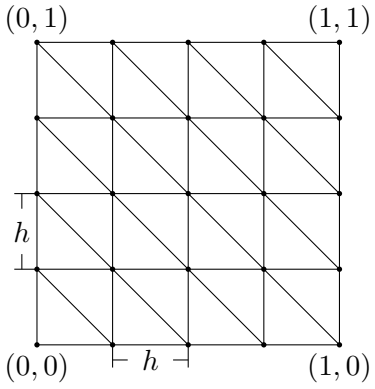


Abbildung 5.2: Triangulierung des Einheitsquadrates

Da $V_{0,h}$ endlich dimensional ist, gilt

$$N(h) := \dim(V_{0,h}) < \infty.$$

Eine Basis von $V_{0,h}$ wird mit $\{\Psi_i\}_{i=1}^{N(h)}$ bezeichnet. Dann kann jedes Element $v_h \in V_h$ dargestellt werden als Linearkombination

$$v_h(\mathbf{x}) = \sum_{i=1}^{N(h)} v_i \Psi_i(\mathbf{x}).$$

Insbesondere kann jede Testfunktion $\varphi \in V_{0,h}$ durch eine solche Linearkombination dargestellt werden. Dadurch vereinfacht sich das Variationsproblem zu:

Sei $\boldsymbol{\theta} \in \Theta$ fest gewählt. Gesucht ist $u_h(t) \in V_h$, so dass

$$\left.\begin{aligned} u_h(t) &= u_{g,h}(t) + v_h(t), \quad v_h(t) \in V_{0,h}, \\ \partial_t (v_h(t), \Psi_i)_L + a(v_h(t), \Psi_i) &= f(u_{g,h}(t), \Psi_i) \quad \forall\, i = 1, \ldots, N(h), \end{aligned}\right\} \tag{5.12}$$

wobei $u_{g,h}(t) \in V_h$ so gewählt wird, dass $u_{g,h}(t) = 1 - t$ auf dem Rand $\partial\Omega$ ist.

Im folgenden wird

- $N(h)$ derart gewählt, dass $\sqrt{N(h)} \in \mathbb{N}$ gilt,
- das diskrete Gebiet Ω_h dargestellt durch

  $$\Omega_h := \{\mathbf{x} = (x_1, x_2)^\top \in \Omega;\, x_1 = kh,\, x_2 = lh \text{ mit } k, l \in \{1, \ldots, \sqrt{N(h)}\}\},$$

  wobei für die Schrittweite $h := (\sqrt{N(h)} - 1)^{-1}$ gilt.
- eine Basis $\{\Psi_i(\mathbf{x})\}_{i=1}^{N(h)}$ des Raumes V_h derart gewählt, dass $\Psi_i(\mathbf{x}^j) = \delta_{ij}\ \ \forall\, \mathbf{x}^j \in \Omega_h$ und $\forall\, i, j \in \{1, \ldots, N(h)\}$ erfüllt ist.

5.2.2 Zeitliche Diskretisierung

Mit fest gewähltem $u_{g,h}(t) \in V_h$, so dass $u_{g,h}(t) = 1 - t$ auf dem Rand $\partial\Omega$ ist, stellt die Gleichung

$$\partial_t (v_h(t), \Psi_i)_L + a(v_h(t), \Psi_i) \quad = \quad f(u_{g,h}(t), \Psi_i) \quad \forall\, i = 1, \ldots, N(h)$$

aus (5.12) mit der Anfangsbedingung

$$(v_h(0), \Psi_i)_L = (5, \Psi_i)_L \quad \forall\, i = 1, \ldots, N(h),$$

eine gewöhnliche Differentialgleichung erster Ordnung dar. Diese Differentialgleichung kann durch Trennung der Variablen und anschließender Integration gelöst werden. Man erhält

$$\begin{aligned}
(v_h(t), \Psi_i)_L \quad &= \quad (v_h(0), \Psi_i)_L + \int_0^t -a(v_h(s), \Psi_i) + f(u_{g,h}(s), \Psi_i)\, ds, \\
&\overset{(5.10)}{=} \quad (v_h(0), \Psi_i)_L - \int_0^t a(v_h(s), \Psi_i) + a(u_{g,h}, \Psi_i)\, ds \\
&\qquad\quad -(u_{g,h}, \Psi_i)_L + (\overline{v}_h(0), \Psi_i)_L \\
&= \quad (v_h(0) - u_{g,h}(t) + u_{g,h}(0), \Psi_i)_L - \int_0^t a(v_h(s) + u_{g,h}(s), \Psi_i)\, ds \\
&\overset{u_h = v_h + u_{g,h}}{=} \quad (u_h(0) - u_{g,h}(t), \Psi_i)_L - \int_0^t a(u_h(s), \Psi_i)\, ds,
\end{aligned}$$

$\forall\, i = 1, \ldots, N(h)$.

Da das Integral

$$\int_0^t a(u_h(s), \Psi_i)\, ds$$

im allgemeinen nicht analytisch berechnet werden kann, wird eine numerische Integrationsmethode verwendet. Zur Realisierung einer solchen Methode wird in dieser Arbeit zunächst das Zeitintervall $\overline{T} = [0, 1]$ in $N(\Delta t) - 1$ äquidistante Intervalle der Form $[t_n, t_{n+1}]$ mit $\Delta t := t_{n+1} - t_n$ zerlegt, wobei $0 < N(\Delta t) \in \mathbb{N}$, $t_1 = 0$ und $t_{N(\Delta t)} = 1$ ist. Dann gilt für $n = 1, \ldots, N(\Delta t) - 1$

$$(v_h(t_{n+1}), \Psi_i)_L \quad = \quad (u_h(t_n) - u_{g,h}(t_{n+1}), \Psi_i)_L - \int_{t_n}^{t_{n+1}} a(u_h(s), \Psi_i)\, ds,$$

$\forall\, i = 1, \ldots, N(h)$.

Die Integrale werden mit der *Trapezregel* numerisch berechnet. Wie in [76] dargestellt gilt mit $\Delta t = t_{n+1} - t_n$

$$\int_{t_n}^{t_{n+1}} a(u_h(s), \Psi_i)\, ds \;=\; \frac{\Delta t}{2} \Big(a(u_h^{n+1}, \Psi_i) + a(u_h^n, \Psi_i) \Big) + O((\Delta t)^3). \tag{5.13}$$

Mithilfe der Trapezregel können dann die Näherungen mit $u_{g,h}^n \approx u_{g,h}(t_n)$

$$v_h^n \approx v_h(t_n) \quad \text{und} \quad u_h^n = v_h^n + u_{g,h}^n \approx u_h(t_n)$$

mit dem *Crank-Nicolson-Verfahren* [76] sukzessive bestimmt werden, indem für $n = 1, \dots, N(\Delta t) - 1$ die Variationsprobleme

$$\begin{aligned}(v_h^{n+1}, \Psi_i)_L + \frac{\Delta t}{2} a(v_h^{n+1}, \Psi_i) \;\; &= \\ &(u_h^n - u_{g,h}^{n+1}, \Psi_i)_L + \frac{\Delta t}{2}\Big(a(u_h^n, \Psi_i) - a(u_{g,h}^{n+1}, \Psi_i)\Big)\end{aligned}$$

$\forall\, i = 1, \dots, N(h)$ gelöst werden.

Bemerkung 5.2.1. *Wie in [76] beschrieben ist der Fehler $\|v_h^n - v_h(t_n)\|_{V_0}$ des Crank-Nicolson-Verfahrens von der Größenordnung $O((\Delta t)^2)$.*

5.3 Sensitivitätsgleichungen

Um die Fisher-Informationsmatrix

$$M(\vec{\mathbf{x}}, \mathbf{p}) \;=\; \sum_{i=1}^{\ell} p_i \int_T \nabla_{\boldsymbol{\theta}} u(t, \mathbf{x}^i, \boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} u(t, \mathbf{x}^i, \boldsymbol{\theta})^\top dt \;\in\; \mathbb{R}^{3\times 3}, \tag{5.14}$$

mit $\vec{\mathbf{x}} = ({\mathbf{x}^1}^\top, \dots, {\mathbf{x}^\ell}^\top)^\top$ und $\mathbf{x}^i \in \Omega\ \forall\, i \in I_\ell$ berechnen zu können, werden die Sensitivitäten

$$\nabla_{\boldsymbol{\theta}} u(t, \mathbf{x}, \boldsymbol{\theta}) \;=\; \Big(\partial_{\theta_1} u(t, \mathbf{x}, \boldsymbol{\theta}), \partial_{\theta_2} u(t, \mathbf{x}, \boldsymbol{\theta}), \partial_{\theta_3} u(t, \mathbf{x}, \boldsymbol{\theta})\Big)^\top,$$

benötigt. Da (5.5) vorausgesetzt wurde, ist für jedes fest gewählte $\boldsymbol{\theta}$

$$\partial_{\theta_i} u(t, \mathbf{x}, \boldsymbol{\theta}) \in \mathcal{C}^1(T; V)$$

für alle $i \in \{1, 2, 3\}$.

In diesem Abschnitt wird beschrieben, wie der Gradient $\nabla_{\boldsymbol{\theta}} u(t, \mathbf{x}, \boldsymbol{\theta})$ durch Lösen der zugehörigen Sensitivitätsgleichungen numerisch ermittelt werden kann: Wie beim Zustand $u(t, \mathbf{x}, \boldsymbol{\theta})$ werden die Sensitivitäten $\partial_{\theta_i} u(t, \mathbf{x}, \boldsymbol{\theta})$ als schwache Lösung eines Variationsproblems mit der Methode der finiten Elemente ermittelt, wobei bezüglich der Zeit auf das Crank-Nicolson-Verfahren zurückgegriffen wird. Die Herleitung der kontinuierlichen sowie der diskreten schwachen Formulierung der Sensitivitätsgleichungen erfolgt analog zur Variationsformulierung in Abschnitt 5.2.

Zunächst werden die Sensitivitätsgleichungen aufgestellt. Diese erhält man, indem die Zustandsgleichung (5.2) jeweils nach θ_1, θ_2 und θ_3 partiell abgeleitet wird. Es ergibt sich für $i \in \{1, 2, 3\}$ und dem euklidischen Skalarprodukt $\langle \cdot, \cdot \rangle$

$$\begin{aligned}&\partial_{\theta_i}\Big(\partial_t u(t, \mathbf{x}, \boldsymbol{\theta}) - \nabla_{\mathbf{x}} \cdot [\kappa(x, \boldsymbol{\theta}) \nabla_{\mathbf{x}} u(t, \mathbf{x}, \boldsymbol{\theta})]\Big) \\ &= \partial_t [\partial_{\theta_i} u(t, \mathbf{x}, \boldsymbol{\theta})] - \partial_{\theta_i}\left(\kappa(x, \boldsymbol{\theta}) \Delta_{\mathbf{x}} u(t, \mathbf{x}, \boldsymbol{\theta}) + \left\langle \begin{pmatrix}\theta_2 \\ \theta_3\end{pmatrix}, \nabla_{\mathbf{x}} u(t, \mathbf{x}, \boldsymbol{\theta}) \right\rangle\right) \\ &= \partial_t [\partial_{\theta_i} u(t, \mathbf{x}, \boldsymbol{\theta})] - \partial_{\theta_i} \kappa(x, \boldsymbol{\theta}) \Delta_{\mathbf{x}} u(t, \mathbf{x}, \boldsymbol{\theta}) - \kappa(x, \boldsymbol{\theta}) \Delta_{\mathbf{x}} [\partial_{\theta_i} u(t, \mathbf{x}, \boldsymbol{\theta})] \\ &\quad - [\partial_{\theta_i} \theta_2] \partial_{x_1} u(t, \mathbf{x}, \boldsymbol{\theta}) - [\partial_{\theta_i} \theta_3] \partial_{x_2} u(t, \mathbf{x}, \boldsymbol{\theta}) - \left\langle \begin{pmatrix}\theta_2 \\ \theta_3\end{pmatrix}, \nabla_{\mathbf{x}} [\partial_{\theta_i} u(t, \mathbf{x}, \boldsymbol{\theta})] \right\rangle \;=\; 0.\end{aligned}$$

Somit lauten die drei Sensitivitätsgleichungen mit $u := u(t, \mathbf{x}, \boldsymbol{\theta})$

$$\partial_t \left[\partial_{\theta_1} u\right] = \Delta_{\mathbf{x}} u + (\theta_1 + x_1\theta_2 + x_2\theta_3)\, \Delta_{\mathbf{x}} \left[\partial_{\theta_1} u\right] + \left\langle \begin{pmatrix} \theta_2 \\ \theta_3 \end{pmatrix}, \nabla_{\mathbf{x}} \left[\partial_{\theta_1} u\right] \right\rangle, \tag{5.15}$$

$$\partial_t \left[\partial_{\theta_2} u\right] = \Delta_{\mathbf{x}} u + (\theta_1 + x_1\theta_2 + x_2\theta_3)\, \Delta_{\mathbf{x}} \left[\partial_{\theta_2} u\right] + \left\langle \begin{pmatrix} \theta_2 \\ \theta_3 \end{pmatrix}, \nabla_{\mathbf{x}} \left[\partial_{\theta_2} u\right] \right\rangle + \partial_{x_1} u, \tag{5.16}$$

$$\partial_t \left[\partial_{\theta_3} u\right] = \Delta_{\mathbf{x}} u + (\theta_1 + x_1\theta_2 + x_2\theta_3)\, \Delta_{\mathbf{x}} \left[\partial_{\theta_3} u\right] + \left\langle \begin{pmatrix} \theta_2 \\ \theta_3 \end{pmatrix}, \nabla_{\mathbf{x}} \left[\partial_{\theta_3} u\right] \right\rangle + \partial_{x_2} u. \tag{5.17}$$

Die Lösungen der Sensitivitätsgleichungen besitzen für $i \in \{1, 2, 3\}$ die Anfangswerte

$$\partial_{\theta_i} u(0, \mathbf{x}, \boldsymbol{\theta}) \;=\; 0 \quad \forall\, \mathbf{x} \in \Omega, \quad \boldsymbol{\theta} \in \Theta,$$

sowie die Randwerte

$$\partial_{\theta_i} u(t, \mathbf{x}, \boldsymbol{\theta}) \;=\; 0 \quad \forall\, t \in T, \quad \mathbf{x} \in \partial\Omega, \quad \boldsymbol{\theta} \in \Theta.$$

Sei $w_i(t)(\mathbf{x}) := \partial_{\theta_i} u(t, \mathbf{x}, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$ fest. Dann ist $w_i(t) \in V_0$, $t \in T$. Mit der Bilinearform $\alpha : V \times V \to \mathbb{R}$ aus Gleichung (5.8) erhält man dann die schwache Formulierung der Sensitivitätsgleichungen:

Sei $u(t) \in V$ eine schwache Lösung des Variationsproblems (5.9). Gesucht sind $w_1(t), w_2(t), w_3(t) \in V_0$, so dass

$$\begin{aligned}
\partial_t (w_1(t), \varphi)_L + a(w_1(t), \varphi) &= -(\nabla_{\mathbf{x}} u(t), \nabla_{\mathbf{x}} \varphi)_L, \\
&=: f_1(u(t), \varphi), \\
\partial_t (w_2(t), \varphi)_L + a(w_2(t), \varphi) &= -(\nabla_{\mathbf{x}} u(t), \nabla_{\mathbf{x}} \varphi)_L - (u(t), \partial_{x_1} \varphi)_L, \\
&=: f_2(u(t), \varphi), \\
\partial_t (w_3(t), \varphi)_L + a(w_3(t), \varphi) &= -(\nabla_{\mathbf{x}} u(t), \nabla_{\mathbf{x}} \varphi)_L - (u(t), \partial_{x_2} \varphi)_L, \\
&=: f_3(u(t), \varphi),
\end{aligned}$$

$\forall\, \varphi \in V_0$.

Analog zu Abschnitt 5.2 wird der Sobolevraum V_0 auf den Teilraum $V_{0,h}$ eingeschränkt. Mit der Triangulierung $\mathcal{T}_h$ und der Basis $\{\Psi_i\}_{i=1}^{N(h)}$ aus Abschnitt 5.2 kann das diskrete Variationsproblem formuliert werden:

Sei $u_h(t) \in V_h$ eine schwache Lösung des diskreten Variationsproblems (5.9). Gesucht sind $w_{h,1}(t), w_{h,2}(t), w_{h,3}(t) \in V_{0,h}$, so dass für $j = 1, 2, 3$

$$\partial_t (w_{h,j}(t), \Psi_i)_L + a(w_{h,j}(t), \Psi_i) \;=\; f_j(u_h(t), \Psi_i),$$

$\forall\, i = 1, \ldots, N(h)$ erfüllt ist.

Mit dem Crank-Nicolson-Verfahren gilt dann für $j = 1, 2, 3$

$$(w_{h,j}^{n+1}, \Psi_i)_L + \frac{\Delta t}{2} a(w_{h,j}^{n+1}, \Psi_i) = (w_{h,j}^{n}, \Psi_i)_L + \frac{\Delta t}{2} \Big(a(w_{h,j}^{n}, \Psi_i) + f_j(u_h^{n+1} - u_h^{n}, \Psi_i) \Big),$$

$\forall\, i = 1, \ldots, N(h)$.

5.4 Numerische Ergebnisse

In diesem Abschnitt wird das in Kapitel 3.3.3 beschriebene *Standardverfahren* und die in Kapitel 4.4 beschriebene zweistufige Active-Set Newton-artige Methode am Beispiel der Wärmeleitungsgleichung (5.2) angewendet um ein D-optimales Design $(\vec{\mathbf{x}}^*, \mathbf{p}^*)$ numerisch zu ermitteln. Anschließend werden beide Verfahren anhand der numerischen Ergebnisse miteinander verglichen. Das Ziel ist die Bestimmung einer optimalen Messstellenkonstellation $(\vec{\mathbf{x}}^*, \mathbf{p}^*)$, so dass bei einer Parameteridentifikation die Parameter $(\theta_1, \theta_2, \theta_3)^\top$ aus (5.1) bestmöglich geschätzt werden können.

Wie in Kapitel 3.2 erläutert, wird für die Bestimmung eines D-optimalen Designs eine erste gute Schätzung $\hat{\boldsymbol{\theta}}^1 = (\hat{\theta}_1^1, \hat{\theta}_2^1, \hat{\theta}_3^1)^\top$ des exakten Modellparameters $\boldsymbol{\theta}^* = (\theta_1^*, \theta_2^*, \theta_3^*)^\top$ benötigt. In [85] wurde beschrieben, dass hierfür der Schätzer

$$\hat{\boldsymbol{\theta}}^1 \quad := \quad (0.1, 0.3, 0.3)^\top \tag{5.18}$$

gewählt werden kann.

Da insgesamt $m = 3$ Parameter geschätzt werden, existiert nach Satz 3.3.2 eine optimale Messstellenkonfiguration mit maximal $\frac{3\cdot(3+1)}{2} = 6$ Designpunkten. Für die numerischen Experimente wird als Startdesign

$$\begin{Bmatrix} \vec{\mathbf{x}}^{0\top} \\ \mathbf{p}^{0\top} \end{Bmatrix} = \begin{Bmatrix} (0.1, 0.1) & (0.1, 0.5) & (0.1, 0.9) & (0.9, 0.1) & (0.9, 0.5) & (0.9, 0.9) \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{Bmatrix} \tag{5.19}$$

mit $(\vec{\mathbf{x}}^0, \mathbf{p}^0) \in \bar{\Omega}^6 \times [0,1]^6$ gewählt.

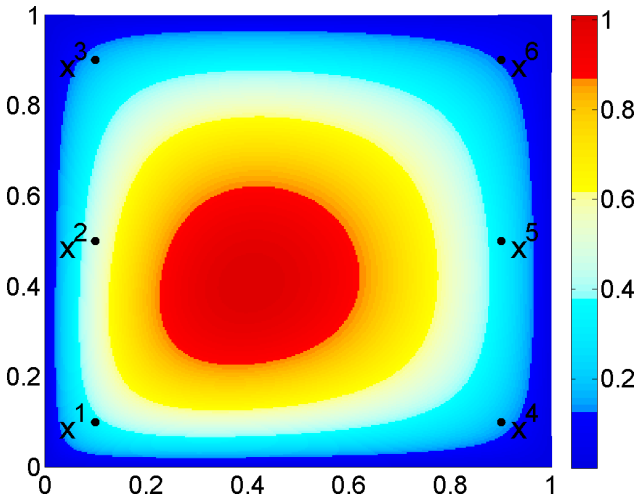


Abbildung 5.3: Lösung $u(t, \mathbf{x}, \hat{\boldsymbol{\theta}}^1)$ der Wärmeleitungsgleichung (5.2) zum Zeitpunkt $t = 1$ mit den Designpunkten $\vec{\mathbf{x}}^0$. Diese Punkte erhalten die Startgewichte $p_i^0 = \frac{1}{6}$ für $i = 1, \ldots, 6$.

5.4.1 Darstellung der numerischen Lösung der Zustands- und Sensitivitätsgleichungen

Bevor im folgenden Abschnitt ein D-optimales Design $(\vec{\mathbf{x}}^*, \mathbf{p}^*)$ ermittelt wird, werden zunächst die hierfür erforderlichen numerischen Lösungen der Wärmeleitungsgleichung (5.2) und der Sensitivitätsgleichungen (5.15), (5.16) und (5.17) abgebildet.

Folgende Abbildungen zeigen die numerische Lösung der Wärmeleitungsgleichung (5.2) zu den Zeitpunkten $t = 0.00$ sec, $t = 0.24$ sec und $t = 1.00$ sec.

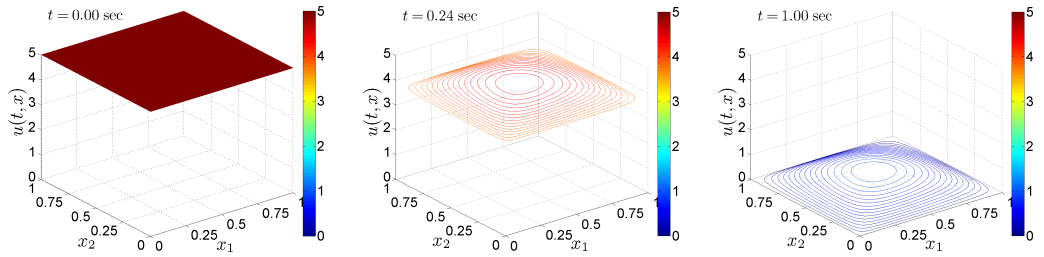


Abbildung 5.4: Die Lösung $u(t, \mathbf{x}, \hat{\boldsymbol{\theta}}^1)$ zu den Zeitpunkten $t = 0.00$ sec, $t = 0.24$ sec und $t = 1.00$ sec.

Nachdem eine Lösung $u(t, \mathbf{x}, \hat{\boldsymbol{\theta}}^1)$ der Wärmeleitungsgleichung (5.2) ermittelt wurde, wird diese in die Gleichungen (5.15), (5.16) und (5.17) eingesetzt um die Sensitivitäten $\partial_{\theta_1} u(t, \mathbf{x}, \hat{\boldsymbol{\theta}}^1)$, $\partial_{\theta_2} u(t, \mathbf{x}, \hat{\boldsymbol{\theta}}^1)$ und $\partial_{\theta_3} u(t, \mathbf{x}, \hat{\boldsymbol{\theta}}^1)$ ermitteln zu können.

Folgende Abbildungen zeigen die numerischen Lösungen der Sensitivitätsgleichungen zu den Zeitpunkten $t = 0.00$ sec, $t = 0.24$ sec und $t = 1.00$ sec:

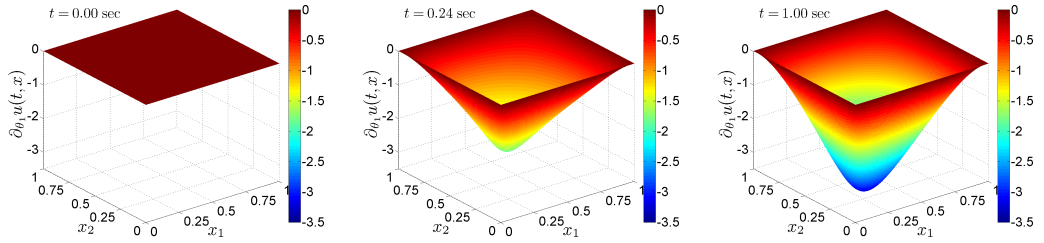


Abbildung 5.5: Die Sensitivität $\partial_{\theta_1} u(t, \mathbf{x}, \hat{\boldsymbol{\theta}}^1)$ zu den Zeitpunkten $t = 0.00$ sec, $t = 0.24$ sec und $t = 1.00$ sec.

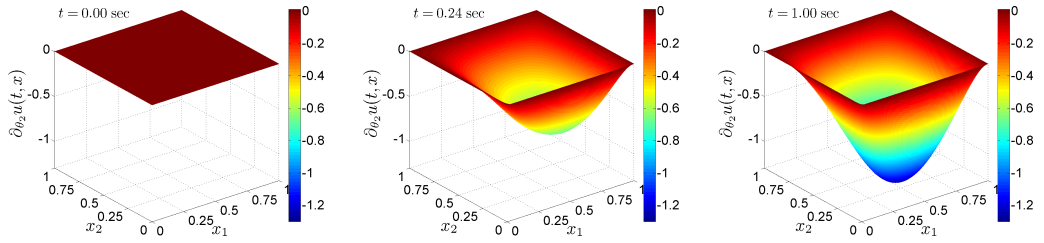


Abbildung 5.6: Die Sensitivität $\partial_{\theta_2} u(t, \mathbf{x}, \hat{\boldsymbol{\theta}}^1)$ zu den Zeitpunkten $t = 0.00$ sec, $t = 0.24$ sec und $t = 1.00$ sec.

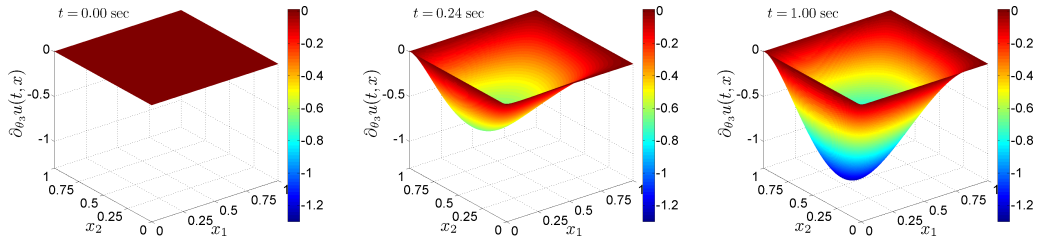


Abbildung 5.7: Die Sensitivität $\partial_{\theta_3} u(t, \mathbf{x}, \hat{\boldsymbol{\theta}}^1)$ zu den Zeitpunkten $t = 0.00$ sec, $t = 0.24$ sec und $t = 1.00$ sec.

Da die Lösung $u(t, \mathbf{x}, \hat{\boldsymbol{\theta}}^1)$ zum Zeitpunkt $t = 0$ und für jedes $t \in T$ am Rand des Gebietes Ω unabhängig von den Modellparametern θ_i für $i = 1, 2, 3$ ist, sind in diesen Bereichen die Sensitivitäten Null.

Im folgenden Kapitel 5.4.2 werden in Abhängigkeit von den gewählten Schrittweiten $h > 0$ und $\Delta t > 0$ die zugehörigen D-optimalen Designs ermittelt. Hierfür wird sowohl das in Kapitel 3 beschriebene Lösungsverfahren als auch die in Kapitel 4 erläuterte Active-Set-Methode verwendet. Auf Basis dieser numerischen Ergebnisse werden dann diese beiden Verfahren in Abhängigkeit von h und Δt miteinander verglichen.

5.4.2 Numerisch ermittelte D-optimale Designs

In Kapitel 5.2 wurde beschrieben, dass in dieser Arbeit sowohl die Wärmeleitungsgleichung (5.2) als auch die Sensitivitätsgleichungen (5.15), (5.16) und (5.17) mit der Methode der finiten Elemente und dem Crank-Nicolson-Verfahren gelöst werden. Die diesbezüglich gewählte räumliche Schrittweite wird mit $h > 0$ und die zeitliche Schrittweite mit $\Delta t > 0$ bezeichnet.

In Abhängigkeit von der Anzahl der Freiheitsgrade

$$N(h) \;:=\; h^{-2} \in \mathbb{N}$$

und

$$N(\Delta t) \;:=\; (\Delta t)^{-1} \in \mathbb{N}$$

werden in diesem Abschnitt Algorithmus 3.3.1 (*exchange-type*) und Algorithmus 4.4.2 (*active-set*) zur Bestimmung eines D-optimalen Designs $\mathbf{w}^* = (\vec{\mathbf{x}}^*, \mathbf{p}^*)$ verwendet. Anschließend werden diese numerischen Ergebnisse anhand des relativen Fehlers im D-optimalen Design und im Zielfunktional miteinander verglichen. Die räumliche Schrittweite $h > 0$ und die zeitliche Schrittweite $\Delta t > 0$ werden dabei so gewählt, dass stets $N(h) \in \mathbb{N}$ und $N(\Delta t) \in \mathbb{N}$ gilt.

Um den relativen Fehler in der h- und Δt-abhängigen Lösung

$$\mathbf{w}_{h,\Delta t} \;:=\; \mathbf{w}(h, \Delta t) \;=\; \Big(\vec{\mathbf{x}}(h, \Delta t), \mathbf{p}(h, \Delta t)\Big) \;\in\; \bar{\Omega}^\ell \times [0,1]^\ell$$

zu bestimmen, wird

1. zunächst bei fest gewählter Schrittweite $\Delta t = 10^{-3}$ im Raum verfeinert („h-Verfeinerung“)

2. und anschließend bei fest gewählter Schrittweite $h = 640^{-1}$ in der Zeit verfeinert („Δt-Verfeinerung“)

um jeweils ein D-optimales Design $\mathbf{w}_{h,\Delta t}$ mit den Algorithmen 3.3.1 und 4.4.2 numerisch zu ermitteln.

Bemerkung 5.4.1. *Aus [85] ist bekannt, dass für das Designproblem (5.2) mit (5.1)*

$$\ell \;=\; 3$$

gilt und somit das zugehörige D-optimale Design genau 3 *Messstellen beinhaltet.*

Notation

- Im folgenden wird ein h-abhängiges Design bei fest gewählter, zeitlicher Schrittweite $\Delta t^* := 10^{-3}$ mit

$$\mathbf{w}_{h,\Delta t^*} \;:=\; \mathbf{w}(h, \Delta t^*) \tag{5.20}$$

bezeichnet. Für den Grenzwert gilt

$$\mathbf{w}^*_{h,\Delta t^*} \;:=\; \lim_{h\to 0} \mathbf{w}(h, \Delta t^*). \tag{5.21}$$

- Ein Δt-abhängiges Design wird bei fest gewählter, räumlicher Schrittweite $h^* := 640^{-1}$ mit

$$\mathbf{w}_{h^*,\Delta t} \;:=\; \mathbf{w}(h^*, \Delta t) \tag{5.22}$$

bezeichnet. Für den Grenzwert gilt

$$\mathbf{w}^*_{h^*,\Delta t} \;:=\; \lim_{\Delta t\to 0} \mathbf{w}(h^*, \Delta t). \tag{5.23}$$

- Für fest gewähltes $\Delta t^* := 10^{-3}$ und groß gewähltem $N(h)$ wurde der Grenzwert $\mathbf{w}^*_{h,\Delta t^*} := (\mathbf{x}^*_{h,\Delta t^*}, \mathbf{p}^*_{h,\Delta t^*})$ durch

$$\mathbf{x}^*_{h,\Delta t^*} := \begin{pmatrix} (0.63792, 0.27316)^\top \\ (0.27316, 0.63792)^\top \\ (0.15364, 0.15364)^\top \end{pmatrix} \quad \text{und} \quad \mathbf{p}^*_{h,\Delta t^*} := \begin{pmatrix} 0.33 \\ 0.33 \\ 0.34 \end{pmatrix}, \tag{5.24}$$

approximiert. Für den Zielfunktionswert gilt

$$J(\mathbf{w}^*_{h,\Delta t^*}) \;=\; 3.31284.$$

- Für fest gewähltes $h^* := 640^{-1}$ und groß gewähltem $N(\Delta t)$ wurde der Grenzwert $\mathbf{w}^*_{h^*,\Delta t} := (\mathbf{x}^*_{h^*,\Delta t}, \mathbf{p}^*_{h^*,\Delta t})$ durch

$$\mathbf{x}^*_{h^*,\Delta t} := \begin{pmatrix} (0.63752, 0.27331)^\top \\ (0.27331, 0.63752)^\top \\ (0.15314, 0.15314)^\top \end{pmatrix} \quad \text{und} \quad \mathbf{p}^*_{h^*,\Delta t} := \begin{pmatrix} 0.33 \\ 0.33 \\ 0.34 \end{pmatrix}, \tag{5.25}$$

approximiert. Für den Zielfunktionswert gilt

$$J(\mathbf{w}^*_{h^*,\Delta t}) \;=\; 3.30951.$$

D-optimale Designpunkte bei $\Delta t^* := 10^{-3}$ fest gewählt

Zunächst werden die ermittelten Designpunkte $\vec{\mathbf{x}}_{h,\Delta t^*}$ eines D-optimalen Designs $\mathbf{w}_{h,\Delta t^*}$ abgebildet, die bei fest gewähltem $\Delta t^* := 10^{-3}$ und unterschiedlicher Anzahl an Freiheitsgraden $N(h)$ mit den Algorithmen 3.3.1 und 4.4.2 berechnet wurden. Die Messstellen aus (5.24), die für $\Delta t^* := 10^{-3}$ als optimal angenommen werden, sind jeweils rot dargestellt.

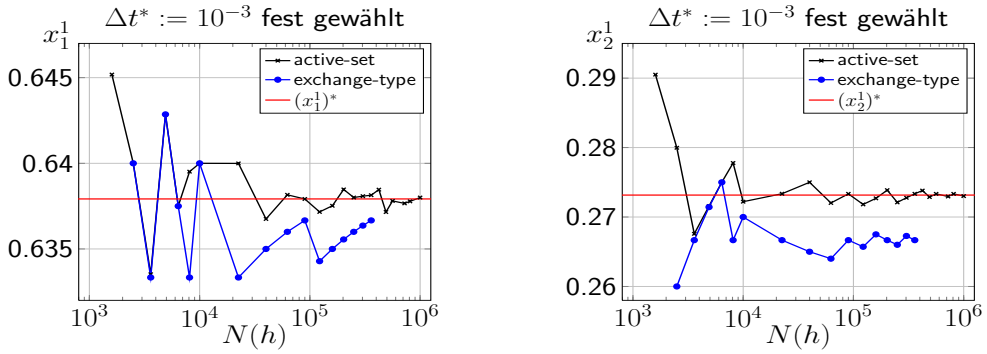


Abbildung 5.8: Die mit Algorithmus 3.3.1 (*exchange-type*) und Algorithmus 4.4.2 (*active-set*) berechnete Messstelle $\mathbf{x}^1 = (x_1^1, x_2^1)^\top$ in Abhängigkeit von $N(h)$.

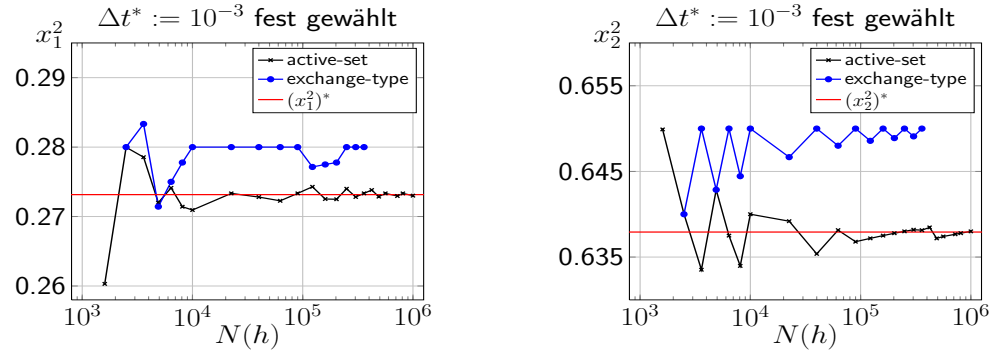


Abbildung 5.9: Die mit Algorithmus 3.3.1 (*exchange-type*) und Algorithmus 4.4.2 (*active-set*) berechnete Messstelle $\mathbf{x}^2 = (x_1^2, x_2^2)^\top$ in Abhängigkeit von $N(h)$.

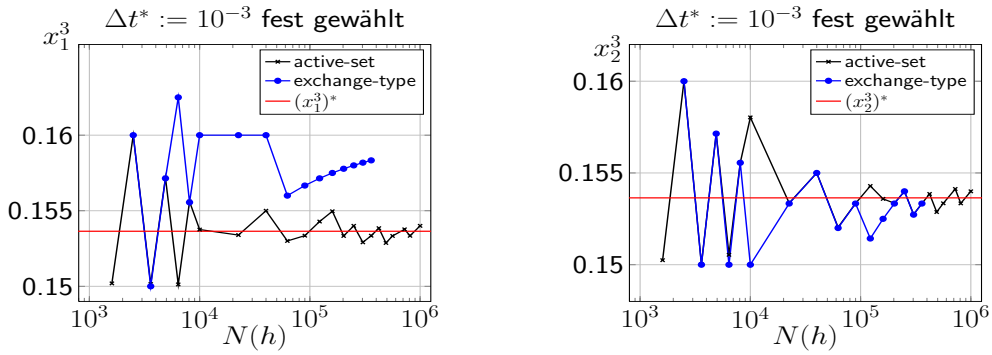


Abbildung 5.10: Die mit Algorithmus 3.3.1 (*exchange-type*) und Algorithmus 4.4.2 (*active-set*) berechnete Messstelle $\mathbf{x}^3 = (x_1^3, x_2^3)^\top$ in Abhängigkeit von $N(h)$.

D-optimale Desigpunkte bei $h^* := 640^{-1}$ fest gewählt

Nun werden die ermittelten Designpunkte $\vec{\mathbf{x}}_{h^*,\Delta t}$ eines D-optimalen Designs $\mathbf{w}_{h^*,\Delta t}$ in Abhängigkeit von $N(\Delta t)$ bei fest gewähltem $h^* := 640^{-1}$ abgebildet. Die Messstellen aus (5.25), die für $h^* := 640^{-1}$ als optimal angenommen werden, sind rot dargestellt.

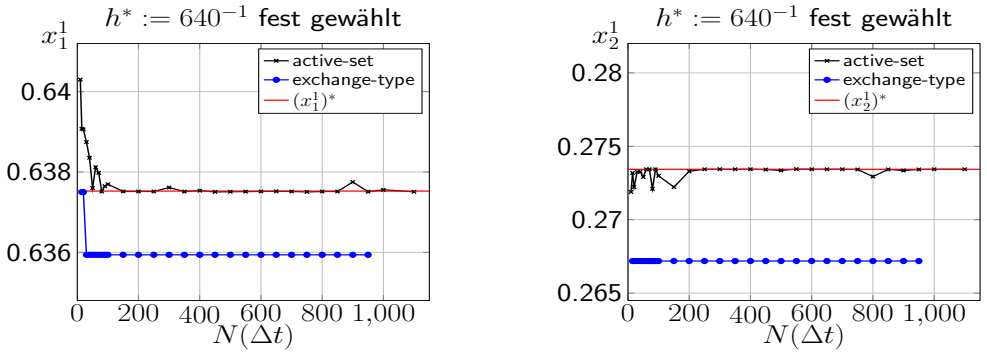


Abbildung 5.11: Die mit Algorithmus 3.3.1 (*exchange-type*) und Algorithmus 4.4.2 (*active-set*) berechnete Messstelle $\mathbf{x}^1 = (x_1^1, x_2^1)^\top$ in Abhängigkeit von $N(\Delta t)$.

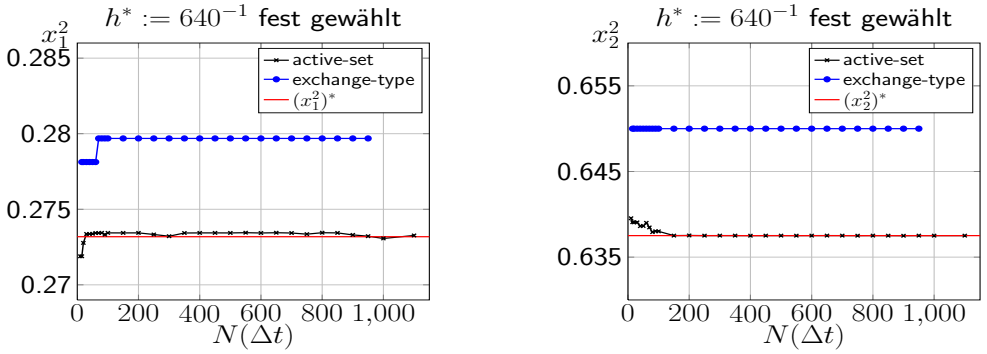


Abbildung 5.12: Die mit Algorithmus 3.3.1 (*exchange-type*) und Algorithmus 4.4.2 (*active-set*) berechnete Messstelle $\mathbf{x}^2 = (x_1^2, x_2^2)^\top$ in Abhängigkeit von $N(\Delta t)$.

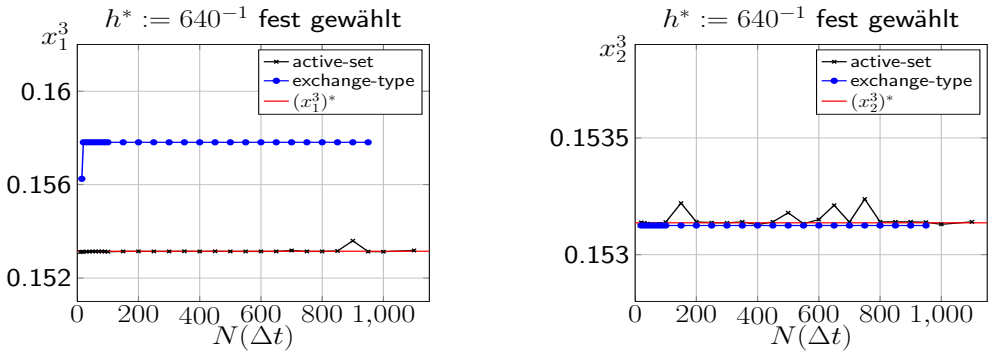


Abbildung 5.13: Die mit Algorithmus 3.3.1 (*exchange-type*) und Algorithmus 4.4.2 (*active-set*) berechnete Messstelle $\mathbf{x}^3 = (x_1^3, x_2^3)^\top$ in Abhängigkeit von $N(\Delta t)$.

Darstellung der Iterierten der zweistufigen Active-Set-Methode

Ausgehend vom Startdesign $(\vec{\mathbf{x}}^0, \mathbf{p}^0) \in \bar{\Omega}^6 \times [0,1]^6$ aus (5.19) werden in folgender Abbildung 5.14 die Iterierten der zweistufigen Active-Set-Methode dargestellt. Für die numerische Berechnung der Lösung der Wärmeleitungsgleichung (5.2) und der Lösungen der Sensitivitätsgleichungen (5.15) - (5.17) wurden die Schrittweiten $h = 640^{-1}$ und $\Delta t = 10^{-3}$ gewählt.

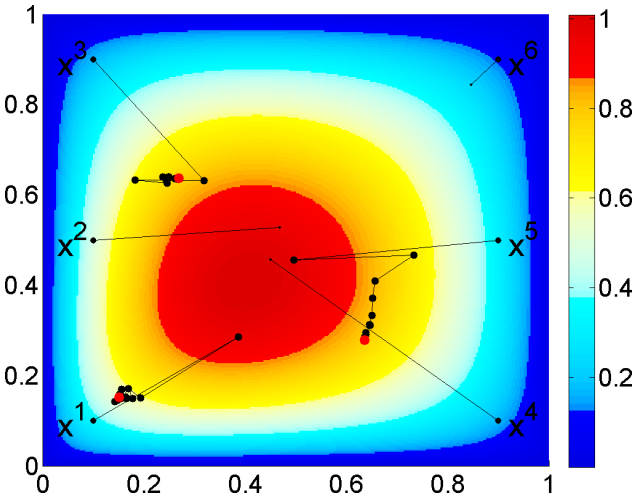


Abbildung 5.14: Lösung der Wärmeleitungsgleichung (5.2) zum Zeitpunkt $t = 1$. Zu sehen sind die Iterierten der zweistufigen Active-Set-Methode bei der Bestimmung eines D-optimalen Designs.

Wie in Kapitel 4.4 beschrieben, verläuft die Umsetzung der Active-Set-Methode zur Bestimmung eines D-optimalen Designs zweistufig:

(1) In jedem Iterationsschritt werden zunächst die Gewichte $\mathbf{p}^n \in [0,1]^6$ als fest gewählt angenommen um mithilfe eines Gradientenverfahrens eine Messstellenkonstellation $\vec{\mathbf{x}}^{n+1} \in \bar{\Omega}^6$ zu ermitteln, so dass der Zielfunktionswert reduziert wird.

(2) Anschließend werden für die Messstellen $\vec{\mathbf{x}}^{n+1} \in \bar{\Omega}^6$ mit dem Newtonverfahren die zugehörigen, optimalen Gewichte $\mathbf{p}^{n+1} \in [0,1]^6$ ermittelt.

Man erkennt in Abbildung 5.14, dass die Gewichte der Designpunkte $\mathbf{x}^i$ für $i \in \{2,4,6\}$ nach einem Iterationsschritt aktiv gesetzt wurden und auch aktiv bleiben, d.h. $p_i^n = 0$ für $i \in \{2,4,6\}$ und $n \geq 1$. Es wurden folgende D-optimale Designs mit der zweistufigen Active-Set-Methode berechnet:

$$\begin{Bmatrix} \vec{\mathbf{x}}^{1\top} \\ \mathbf{p}^{1\top} \end{Bmatrix} = \begin{Bmatrix} (0.39, 0.29) & (0.47, 0.53) & (0.32, 0.63) & (0.45, 0.46) & (0.50, 0.46) & (0.85, 0.85) \\ 0.33 & 0 & 0.33 & 0 & 0.33 & 0 \end{Bmatrix},$$

$$\vdots$$

$$\begin{Bmatrix} \vec{\mathbf{x}}^{59\top} \\ \mathbf{p}^{59\top} \end{Bmatrix} = \begin{Bmatrix} (0.15, 0.15) & (0.47, 0.53) & (0.27, 0.64) & (0.45, 0.46) & (0.64, 0.27) & (0.85, 0.85) \\ 0.33 & 0 & 0.33 & 0 & 0.33 & 0 \end{Bmatrix}.$$

5.4.3 Relativer Fehler in der Lösung $\mathbf{w}_{h,\Delta t}$ und im Zielfunktional $J(\mathbf{w}_{h,\Delta t})$

Relativer Fehler in den Lösungen $\mathbf{w}_{\mathbf{h},\mathbf{\Delta t}^*}$ und $\mathbf{w}_{\mathbf{h}^*,\mathbf{\Delta t}}$

Nachdem mit Algorithmus 3.3.1 (*exchange-type*) und Algorithmus 4.4.2 (*active-set*) für unterschiedliche Anzahl an Freiheitsgraden $N(h)$ und $N(\Delta t)$ jeweils ein D-optimales Design ermittelt wurde, wird in diesem Abschnitt sowohl der relative Fehler in der Lösung $\mathbf{w}_{h,\Delta t^*}$

$$R_1(\mathbf{w}_{h,\Delta t^*}) \quad := \quad \frac{\|\mathbf{w}^*_{h,\Delta t^*} - \mathbf{w}_{h,\Delta t^*}\|_2}{\|\mathbf{w}^*_{h,\Delta t^*}\|_2}, \tag{5.26}$$

als auch der relative Fehler in der Lösung $\mathbf{w}_{h^*,\Delta t}$

$$R_2(\mathbf{w}_{h^*,\Delta t}) \quad := \quad \frac{\|\mathbf{w}^*_{h^*,\Delta t} - \mathbf{w}_{h^*,\Delta t}\|_2}{\|\mathbf{w}^*_{h^*,\Delta t}\|_2}, \tag{5.27}$$

in Abhängigkeit von $N(h)$ und $N(\Delta t)$ dargestellt. Für die Berechnung des relativen Fehlers werden die Grenzwerte $\mathbf{w}^*_{h,\Delta t^*}$ und $\mathbf{w}^*_{h^*,\Delta t}$ aus (5.24) und (5.25) gewählt.

Für $R_1(\mathbf{w}_{h,\Delta t^*})$ aus (5.26) und dem in (5.27) definierten $R_2(\mathbf{w}_{h^*,\Delta t})$ gilt in Abhängigkeit von der Anzahl der Freiheitsgrade $N(h)$ und $N(\Delta t)$:

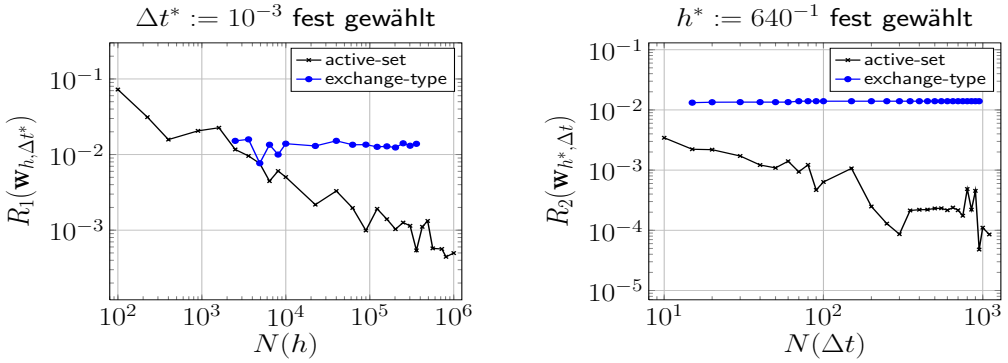


Abbildung 5.15: Darstellung des relativen Fehlers $R_1(\mathbf{w}_{h,\Delta t^*})$ und des relativen Fehlers $R_2(\mathbf{w}_{h^*,\Delta t})$ in Abhängigkeit von der Anzahl der Freiheitsgrade im Ort $N(h) \in \mathbb{N}$ und in der Zeit $N(\Delta t) \in \mathbb{N}$.

Der relative Fehler $R_1(\mathbf{w}_{h,\Delta t^*})$

Man sieht in Abbildung 5.15, dass bei der Bestimmung eines D-optimalen Designs mit dem Algorithmus 4.4.2 (*zweistufigen Active-Set-Methode*) der relative Fehler bei einer h-Verfeinerung eine Konvergenz gegen Null mit der Ordnung $\frac{1}{2}$ aufweist.

Bei der Bestimmung eines D-optimalen Designs mit dem Algorithmus 3.3.1 (*exchange-type*) ist hingegen der relative Fehler $R_1(\mathbf{w}_{h,\Delta t^*})$ nahezu konstant $\approx 10^{-2}$. Der Grund hierfür liegt im Problems des *„Clusterings"*, welches in Kapitel 3.3.3 und in [85] erläutert wurde: Befindet sich ein Designpunkt $\mathbf{x}^i$ in einem aktuellen Design

$$\boldsymbol{\xi}^n \quad = \quad \left\{ \begin{array}{cccc} \mathbf{x}^1, & \mathbf{x}^2, & \ldots, & \mathbf{x}^\ell \\ p_1, & p_2, & \ldots, & p_\ell \end{array} \right\},$$

dann werden alle potentiell neuen Designpunkte $\tilde{\mathbf{x}}$, die in einer gewissen Umgebung von $\mathbf{x}^i$ liegen, nicht in das neue Design $\boldsymbol{\xi}^{n+1}$ aufgenommen um das beschriebene *Cluster-Problem* zu vermeiden. Dadurch kann die einmal gesetzte Messstelle $\mathbf{x}^i$ in einer Umgebung von $\mathbf{x}^i$ nicht korrigiert werden. Somit kann generell nicht erwartet werden, dass bei einer h-Verfeinerung der relative Fehler gegen Null konvergiert.

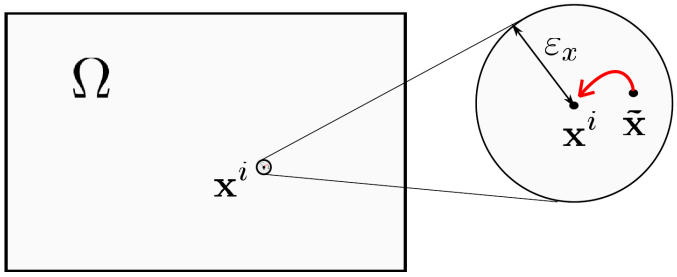


Abbildung 5.16: Befindet sich ein mit Algorithmus 3.3.1 ermittelter Designpunkt $\tilde{\mathbf{x}}$ in einer kleinen Umgebung um einen bereits existierenden Designpunkt $\mathbf{x}^i$, dann muss $\tilde{\mathbf{x}} := \mathbf{x}^i$ gesetzt werden (siehe Kapitel 3.3.3).

Bemerkung 5.4.2. *Für $h < 50^{-1}$ erhält man mit dem Algorithmus 3.3.1 ein Design mit mehr als drei Designpunkten (siehe Anhang B.1) und liegt daher nicht im Lösungsraum $\bar{\Omega}^3 \times [0,1]^3$. Hier kann kein relativer Fehler $R_1(\mathbf{w}_{h,\Delta t^*})$ berechnet werden.*

Der relative Fehler $R_2(\mathbf{w}_{h^*,\Delta t})$

Bei der Bestimmung eines D-optimalen Designs mit dem Algorithmus 4.4.2 (*zweistufige Active-Set-Methode*) konvergiert der relative Fehler $R_2(\mathbf{w}_{h^*,\Delta t})$ bei einer Δt-Verfeinerung mit der Ordnung 1 gegen Null.

Ermittelt man hingegen ein D-optimales Design mit dem Algorithmus 3.3.1 (*exchange-type*), so bleibt der relative Fehler $R_2(\mathbf{w}_{h^*,\Delta t})$ auch hier nahezu konstant $\approx 10^{-2}$. Das der relative Fehler trotz Δt-Verfeinerung nicht kleiner wird liegt daran, dass die mit dem Algorithmus 3.3.1 ermittelten Designpunkte stets auf einem Gitterpunkt der diskreten Menge Ω_h liegen und Ω_h bei Δt-Verfeinerung stets gleich bleibt. Der Algorithmus 3.3.1 liefert bei gleichem Startdesign $\mathbf{w}^0_{h^*,\Delta t}$ hier immer das gleiche D-optimale Design (siehe Anhang B.2).

Bemerkung 5.4.3. *Für $\Delta t < 15^{-1}$ erhält man mit dem Algorithmus 3.3.1 ein Design mit mehr als drei Designpunkten (siehe Anhang B.1) und liegt daher nicht im Lösungsraum $\bar{\Omega}^3 \times [0,1]^3$. Hier kann kein relativer Fehler $R_2(\mathbf{w}_{h^*,\Delta t})$ berechnet werden.*

Nachdem der relative Fehler im D-optimalen Design $\mathbf{w}_{h^*,\Delta t}$ und $\mathbf{w}_{h,\Delta t^*}$ dargestellt wurde, wird nun der relative Fehler im Zielfunktional abgebildet. Mit $J(\mathbf{w}^*_{h,\Delta t^*}) = 3.31284$ und $J(\mathbf{w}^*_{h^*,\Delta t}) = 3.30951$ kann dann der relative Fehler im Zielfunktional $J(\mathbf{w}_{h,\Delta t^*})$ mittels

$$R_3(\mathbf{w}_{h,\Delta t^*}) \quad := \quad \frac{|J(\mathbf{w}^*_{h,\Delta t^*}) - J(\mathbf{w}_{h,\Delta t^*})|}{|J(\mathbf{w}^*_{h,\Delta t^*})|}, \tag{5.28}$$

und der relative Fehler im Zielfunktional $J(\mathbf{w}_{h^*,\Delta t})$ durch

$$R_4(\mathbf{w}_{h^*,\Delta t}) \quad := \quad \frac{|J(\mathbf{w}^*_{h^*,\Delta t}) - J(\mathbf{w}_{h^*,\Delta t})|}{|J(\mathbf{w}^*_{h^*,\Delta t})|}, \tag{5.29}$$

berechnet werden.

Relativer Fehler im Zielfunktional $J(\mathbf{w}_{h,\Delta t^*})$ bei fest gewähltem $\Delta t^* := 10^{-3}$

Die folgende Abbildung zeigt auf der linken Seite den relativen Fehler im Zielfunktional $J(\mathbf{w}_{h,\Delta t^*})$ bei fest gewählter, zeitlicher Schrittweite $\Delta t^* := 10^{-3}$ in Abhängigkeit von $N(h)$, wobei die numerische Lösung $\mathbf{w}_{h,\Delta t^*}$ jeweils mit dem Algorithmus 4.4.2 (*zweistufigen Active-Set-Methode*) und dem Algorithmus 3.3.1 (*exchange-type*) berechnet wurde. Die rechte Abbildung zeigt den jeweiligen Zielfunktionswert $J(\mathbf{w}_{h,\Delta t^*})$ in Abhängigkeit von der Anzahl der Freiheitsgrade $N(h)$.

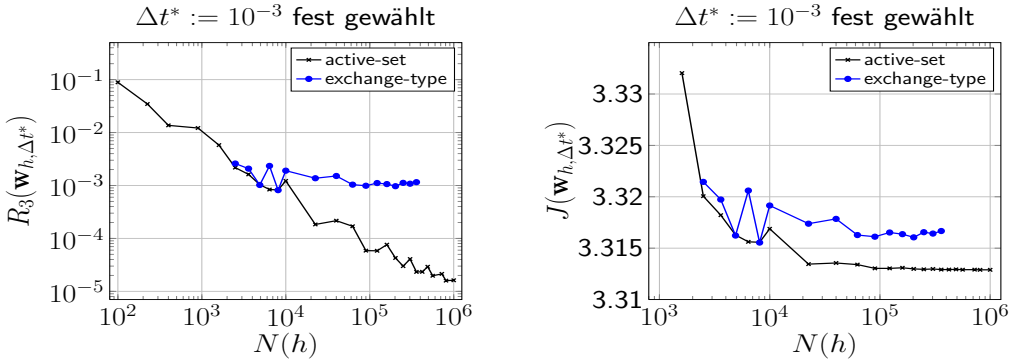


Abbildung 5.17: links: Der relative Fehler $R_3(\mathbf{w}_{h,\Delta t^*})$ in Abhängigkeit von der Anzahl der Freiheitsgrade im Ort $N(h)$. rechts: Der Zielfunktionswert $J(\mathbf{w}_{h,\Delta t^*})$.

Bei der Bestimmung eines D-optimalen Designs mit dem Algorithmus 4.4.2 (*zweistufige Active-Set-Methode*) konvergiert der relative Fehler bei einer h-Verfeinerung mit der Ordnung 1 gegen Null.

Nutzt man hingegen den Algorithmus 3.3.1 (*exchange-type*), dann sieht man, dass der relative Fehler $R_3(\mathbf{w}_{h,\Delta t^*})$ in etwa konstant $\approx 10^{-3}$ ist. Wie beim relativen Fehler $R_1(\mathbf{w}_{h,\Delta t^*})$, der ebenfalls fast konstant ist, kann dieses mit der Strategie zur Vermeidung von *„Clusterings“* begründet werden.

Was den Zielfunktionswert $J(\mathbf{w}_{h,\Delta t^*})$ anbelangt, erhält man mit Algorithmus 4.4.2 stets ein D-optimales Design mit kleinerem, zugehörigem Zielfunktionswert, als mit den mit Algorithmus 3.3.1 ermittelten Lösungen $\mathbf{w}_{h,\Delta t^*}$.

Relativer Fehler im Zielfunktional $J(\mathbf{w}^*_{h^*,\Delta t})$ bei fest gewähltem $h^* := 640^{-1}$

Folgende Abbildung 5.18 zeigt den relativen Fehler $R_4(\mathbf{w}_{h^*,\Delta t})$ bei einer Δt-Verfeinerung, wobei $h^* := 640^{-1}$ fest gewählt wurde. Man sieht, dass der relative Fehler sowohl beim Standardverfahren (*exchange-type*), als auch bei der zweistufigen Active-Set-Methode (*active-set*) linear gegen Null konvergiert, wobei die Active-Set-Methode stets einen kleineren Zielfunktionswert liefert.

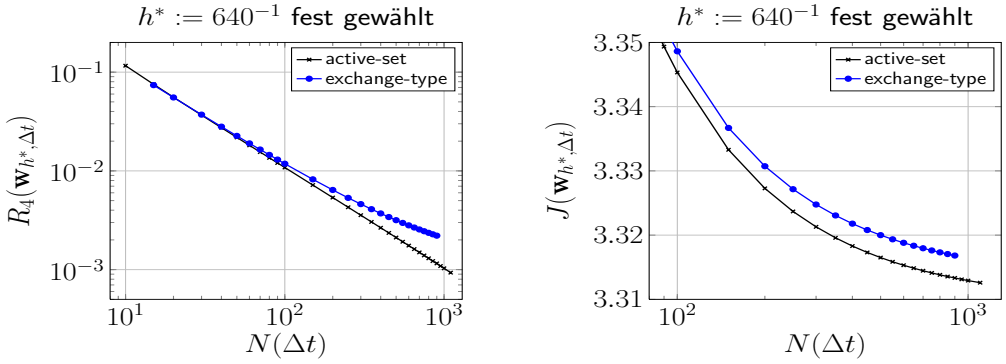


Abbildung 5.18: links: Der relative Fehler $R_4(\mathbf{w}_{h^*,\Delta t})$ in Abhängigkeit von der Anzahl der Freiheitsgrade $N(\Delta t)$. rechts: Das Zielfunktionswert $J(\mathbf{w}_{h^*,\Delta t})$.

Wie bereits beschrieben erhält man mit Algorithmus 3.3.1 (*exchange-type*) bei fest vorgegebener Schrittweite $h^* := 640^{-1}$ für jedes $\Delta t < 15^{-1}$ stets das gleiche D-optimale Design (siehe Anhang B.2). Da sich demnach die Lösung $\mathbf{w}_{h^*,\Delta t}$ beim *Standardverfahren* nicht verändert, wird in Abbildung (5.18) die Fehlerordnung dargestellt, die aus dem Crank-Nicolson Verfahren resultiert: Es gilt mit den Fisher-Informationsmatrizen

$$M(\mathbf{x}^*_{h^*,\Delta t}, \mathbf{p}^*_{h^*,\Delta t}) =: M^*_{h^*,\Delta t} \quad \text{und} \quad M(\mathbf{x}_{h^*,\Delta t}, \mathbf{p}_{h^*,\Delta t}) =: M_{h^*,\Delta t}$$

die Abschätzung

$$\begin{aligned}
|J(\mathbf{w}^*_{h^*,\Delta t}) - J(\mathbf{w}_{h^*,\Delta t})| &= \Big|\ln\det(M_{h^*,\Delta t}) - \ln\det(M^*_{h^*,\Delta t})\Big| \\
&= \left|\ln\frac{\det(M_{h^*,\Delta t})}{\det(M^*_{h^*,\Delta t})}\right| = \left|\ln\frac{\det(M_{h^*,\Delta t}) - \det(M^*_{h^*,\Delta t}) + \det(M^*_{h^*,\Delta t})}{\det(M^*_{h^*,\Delta t})}\right| \\
&\le \left|\ln\frac{|\det(M_{h^*,\Delta t}) - \det(M^*_{h^*,\Delta t})| + \det(M^*_{h^*,\Delta t})}{\det(M^*_{h^*,\Delta t})}\right| \\
&\overset{\text{Bem. 5.2.1}}{=} \left|\ln\frac{O((\Delta t)^2) + \det(M^*_{h^*,\Delta t})}{\det(M^*_{h^*,\Delta t})}\right| \\
&= C\Big|\ln O((\Delta t)^2)\Big| = C'\Big|\ln O(\Delta t)\Big| = O(\Delta t),
\end{aligned}$$

mit Konstanten $C > 0$ und $C' = 2C > 0$.

Nachdem die relativen Fehler im D-optimalen Design und im Zielfunktional abgebildet wurden, wird nun ein Überblick über die benötigte Rechenzeit und die Anzahl an Iterationsschritten in Abhängigkeit von $N(h)$ und $N(\Delta t)$ gegeben.

5.4.4 Zeit- und Iterationsaufwand in Abhängigkeit von $N(h)$ und $N(\Delta t)$

In diesem Abschnitt wird der jeweilige Zeit- und Iterationsaufwand angegeben, der zur Berechnung eines D-optimalen Designs mit der zweistufigen Active-Set-Methode aus Kapitel 4.4 und dem Standardverfahren aus Kapitel 3.3.3 bei unterschiedlicher Wahl von Freiheitsgraden $N(h)$ und $N(\Delta t)$ ermittelt wurden.

Übersicht Zeitaufwand

Die folgende Abbildung 5.19 gibt einen Überblick über den benötigten Zeitaufwand (in Sekunden) in Abhängigkeit von $N(h)$ und $N(\Delta t)$ an:

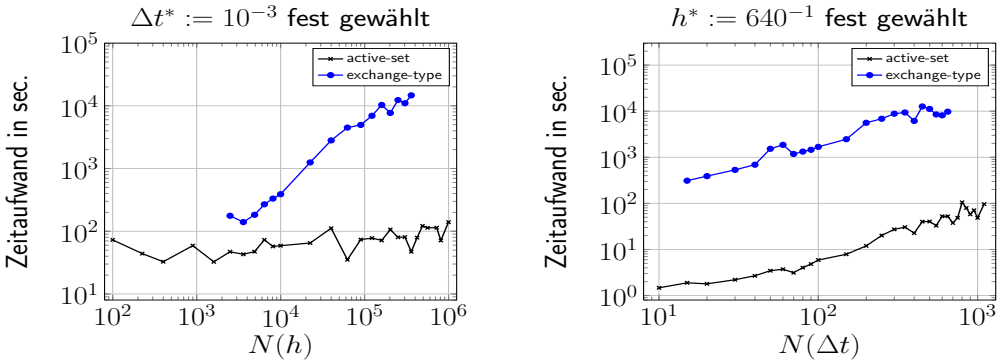


Abbildung 5.19: Der gemessene Zeitaufwand in Abhängigkeit von der Anzahl der Freiheitsgrade $N(h)$ und $N(\Delta t)$.

Zeitaufwand in Abhängigkeit von $N(h)$

Da der Algorithmus 4.4.2 (*active-set*) unabhängig von der Anzahl der Freiheitsgrade im Raum $N(h) \in \mathbb{N}$ ist und die Lösung iterativ mittels einem Gradientenverfahren im Raum $\bar{\Omega}^\ell$ „frei“ gesucht wird (wobei $\ell \in \mathbb{N}$ fest vorgegeben ist), benötigt man beim Bestimmen der Lösung $\mathbf{w}_{h,\Delta t^*}$ für jedes $h > 0$ in etwa die selbe Rechenzeit.

Im Unterschied dazu werden beim Algorithmus 3.3.1 (*exchange-type*) in jedem Iterationsschritt alle Gitterpunkte „abgesucht“, so dass der Rechenaufwand mit der Anzahl an Freiheitsgraden $N(h)$ wächst. Insgesamt steigt der Rechen- und folglich der Zeitaufwand beim Standardverfahren linear mit der Erhöhung der Anzahl der Freiheitsgrade $N(h)$.

Zeitaufwand in Abhängigkeit von $N(\Delta t)$

Sowohl beim Algorithmus 4.4.2 (*active-set*) als auch beim Algorithmus 3.3.1 (*exchange-type*) müssen in jedem Iterationsschritt Zeitintegrale berechnet werden. Bei einer Δt-Verfeinerung, wo folglich die Anzahl an Freiheitsgraden in der Zeit $N(\Delta t)$ erhöht wird, steigt dabei der Rechenaufwand linear an. Dieser lineare Anstieg ist in Abbildung 5.19 deutlich zu erkennen.

Übersicht Iterationsaufwand

Folgende Tabellen zeigen die Anzahl der benötigten Iterationsschritte bei der Bestimmung eines D-optimalen Designs mit dem Algorithmus 4.4.2 (*active-set*) und dem Algorithmus 3.3.1 (*exchange-type*) in Abhängigkeit von der Anzahl an Freiheitsgraden im Raum $N(h)$ und in der Zeit $N(\Delta t)$. Man sieht, dass bei beiden Verfahren in etwa die selbe Anzahl an Iterationsschritten benötigt werden:

$N(h)$	1e2	4e2	9e2	1.6e3	2.5e3	3.6e3	4.9e3	6.4e3	8.1e3	1e4
active-set	71	44	53	41	65	55	62	68	72	56
exchange-type	59	52	56	87	62	59	61	62	62	61

$N(h)$	2.25e4	4e4	6.25e4	9e4	1.225e5	1.6e5	2.025e5	2.5e5
active-set	53	91	47	58	62	57	83	67
exchange-type	62	62	62	62	62	62	62	62

$N(h)$	3.025e5	3.6e5	4.225e5	4.9e5	5.625e5	7.225e5	8.1e5	1e6
active-set	60	64	65	77	72	64	63	81
exchange-type	60	62						

Tabelle 5.1: Anzahl der Iterationen in Abhängigkeit von den Freiheitsgraden im Raum $N(h)$, wobei $\Delta t := 10^{-3}$ fest gewählt wurde.

$N(\Delta t)$	50	100	150	200	250	300	350	400	450	500	550
active-set	54	62	63	70	62	68	67	56	67	63	48
exchange-type	62	62	62	62	62	62	61	61	61	61	

$N(\Delta t)$	600	650	700	750	800	850	900	950	1000	1100
active-set	67	61	73	61	56	73	44	57	59	68
exchange-type	61	61	61	61	61	61	61			

Tabelle 5.2: Anzahl der Iterationen in Abhängigkeit von den Freiheitsgraden in der Zeit $N(\Delta t)$, wobei $h := 640^{-1}$ fest gewählt wurde.

5.5 Fazit

In diesem Kapitel wurden zwei Verfahren zur Bestimmung eines D-optimalen Designs am Beispiel einer zweidimensionalen Wärmeleitungsgleichung umgesetzt:

- ein Standardverfahren aus dem Jahre 2005 (*exchange-type*), welches in [85] hergeleitet wurde,
- eine zweistufige Active-Set-Methode (*active-set*) , mit der die in Kapitel 4 hergeleitete, neue Vorgehensweise zur Bestimmung eines D-optimalen Designs numerisch umgesetzt werden kann.

Um diese Verfahren miteinander vergleichen zu können, wurde die Wärmeleitungsgleichung (5.2) mit den zugehörigen Sensitivitätsgleichungen erster Ordnung (5.15) - (5.17) für unterschiedliche Anzahl an Freiheitsgraden sowohl im Raum $(N(h) \in \mathbb{N})$, als auch in der Zeit $(N(\Delta t) \in \mathbb{N})$ numerisch berechnet. Diese Lösungen werden zur numerischen Bestimmung eines D-optimalen Designs benötigt. Auf diese Weise konnte untersucht werden, wie sich der relative Fehler im D-optimalen Design und der relative Fehler im Zielfunktional in Abhängigkeit von $N(h)$ und $N(\Delta t)$ bei Anwendung der beiden Lösungsverfahren verhält.

Es stellte sich heraus, dass die *Active-Set-Methode* zwar in etwa die selbe Anzahl an Iterationsschritten benötigte, wie das *Standardverfahren*, allerdings viel weniger Rechenzeit in Anspruch nimmt. Zudem ermittelte die *Active-Set-Methode* stets ein D-optimales Design mit kleinerem Zielfunktionswert, als das Standardverfahren.

Was die relativen Fehler im D-optimalen Design und im Zielfunktional anbelangte, sah man sehr deutlich, dass die relativen Fehler bei der *Active-Set-Methode* in allen Fällen gegen Null konvergieren. Bei der *Standardmethode* sah man allerdings, dass sich die relativen Fehler in Abbildung 5.15 und Abbildung 5.17 in Abhängigkeit von $N(h)$ und $N(\Delta t)$ nahezu konstant verhielten. Einzig der relative Fehler im Zielfunktional bei fest gewähltem h^* in Abbildung 5.18 zeigte lineare Konvergenz gegen Null. Wie aber beschrieben wurde, zeigte sich in dieser Abbildung das Verhalten des Diskretisierungsfehlers in der Fisher-Informationsmatrix, der bei kleiner werdender Schrittweite Δt linear gegen Null konvergiert.

Aufgrund der guten Konvergenzeigenschaften der *Active-Set-Methode* bietet sich dieses Verfahren an, wenn ein D-optimales Design so exakt wie möglich ermittelt werden soll. Da - wie in Kapitel 3.3.3 beschrieben - bei der *Standardmethode* das sogenannte *Cluster-Problems* auftritt, kann bei diesem Verfahren nämlich im allgemeinen keine Konvergenz bei h- und Δt-Verfeinerung erwartet werden.

Ist man hingegen lediglich daran interessiert herauszufinden, wo sich in etwa die D-optimalen Messstellen befinden, kann die *Standardmethode* angewendet werden. Da bei dieser Methode keine Ableitungen erforderlich sind, sondern lediglich die Fisher-Informationsmatrix berechnet werden muss, kann die Implementation dieses Verfahrens schnell und einfach umgesetzt werden.

Kapitel 6

Parameteridentifikation und optimale Versuchsplanung in der präparativen Säulenchromatographie

Die Chromatographie ist ein Separationsverfahren, bei dem Stoffgemische zur anschließenden Untersuchung in ihre Bestandteile aufgereinigt werden. Man unterscheidet zwischen der *analytischen* und der *präparativen Chromatographie*. Bei der analytischen Chromatographie ist das Ziel, das Vorhandensein von Stoffen nachzuweisen und ihre Konzentration zu bestimmen. Dagegen werden bei der präparativen Chromatographie die einzelne Bestandteile eines Stoffgemisches zur weiteren Verwendung separiert. Gegenstand dieses Kapitels ist die präparative Säulenchromatographie zur Aufreinigung von Proteingemischen.

Eine chromatographische Aufreinigung erfolgt immer mithilfe zweier nicht miteinander mischbaren Phasen: die *mobilen Phase* (einem beweglichen Medium) und *stationäre Phase* (einem unbeweglichen Trennmedium). Je nach Beschaffenheit der mobilen und stationären Phase unterscheidet man zwischen verschiedenen Chromatographiearten. Die wichtigste Vertreter sind:

Trennmechanismus	stationäre Phase	mobile Phase
Papierchromatographie	Cellulose	Wasser und/oder organisches Lösungsmittel
Gaschromatographie (Säule)	poröses Säulenmaterial, Film auf Säulenwand	Gas (z.B. Wasserstoff, Stickstoff)
Säulenchromatographie	Granulat ($\varnothing 74$ - $140 \mu m$) aus z.B. Cellulose, Kieselgel oder Aluminiumoxid	Flüssigkeit
Hochdruckflüssigkeits-chromatographie	Granulat $\varnothing \leq 10 \mu m$ aus z.B. Cellulose, Kieselgel oder Aluminiumoxid	Flüssigkeit (wird durch Hochdruckpumpe in Bewegung versetzt)

Tabelle 6.1: Die stationäre und mobile Phase verschiedener Chromatographiearten.

Die unterschiedlichen Chromatographiearten haben gemeinsam, dass stets die mobile Phase an einer unbeweglichen stationären Phase vorbeiströmt. Dabei werden die Komponenten, die schwächer an die stationäre Phase gebunden werden, von der mobilen Phase stärker mitgerissen und verlassen somit die stationäre Phase früher. Auf der anderen Seite verlassen die Bestandteile, die stärker an die stationäre Phase gebunden werden, später die Säule. Durch diese Eigenschaft können unterschiedliche Komponenten voneinander getrennt werden.

Eine bewährte Methode zum Aufreinigen von Proteingemischen ist die präparative Säulenchromatographie [75]. Wie in Abbildung 6.1 dargestellt werden hier Proteine voneinander getrennt, indem sie durch eine Säule gedrückt werden, die mit einer stationären Phase gepackt ist.

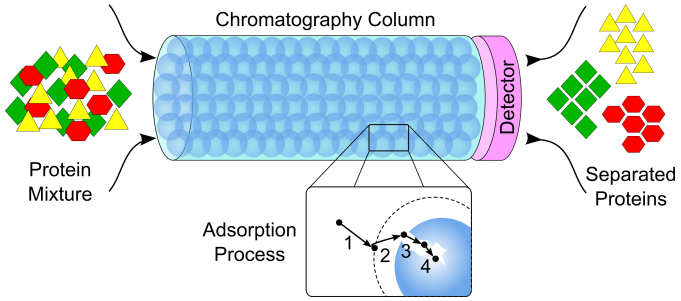


Abbildung 6.1: Präparative Säulenchromatographie zum Aufreinigen von Proteingemischen.[4]

Die einzelnen Proteine weisen abhängig vom pH-Wert der Lösung unterschiedliche Adsorbtions- und Desorbtionseigenschaften auf. Dieses wird bei der Säulenchromatographie ausgenutzt indem die stationäre Phase als ein Adsorbent gewählt wird. Da Salz generell stärker adsorbiert als Proteine kann ein Chromatographieprozess wie folgt beschrieben werden: Zunächst wird ein Proteingemisch in die Säule gedrückt, so dass eine Bindung zwischen der stationären Phase und den Proteinen entsteht. Jetzt kann durch kontrolliertes Hinzufügen einer Salzkonzentration erreicht werden, dass die Proteine wiederholt adsorbieren und desorbieren. Bei geeigneter Steuerung der Salzkonzentration können die Proteine auf diese Weise voneinander getrennt werden.

Da es von großer Relevanz ist, die Proteine nach einem Trennprozess identifizieren zu können, wird in diesem Kapitel gezeigt, wie dieses mit der in Kapitel 2.2 beschriebenen Parameteridentifikation realisiert werden kann. Eine bewährte Vorgehensweise zur Identifizierung von Proteinen ist die Schätzung der sogenannten *steric-mass action* (SMA) Parameter [13]. Bei den SMA-Parametern handelt es sich um insgesamt vier Modellparameter, die ein Protein eindeutig bestimmen. Dazu zählt man

(1) die *Charakteristische Ladung* $\nu > 0$,

(2) den *Adsorbtionskoeffizienten* $k_a > 0$,

(3) den *Desorbtionskoeffizienten* $k_d > 0$ und

(4) den *Schirmungskoeffizienten* $\gamma > 0$

[4] Bild: Tobias Hahn, EMCL, Karlsruher Institut für Technologie

eines Proteins. Clayton A. Brooks und Steven M. Cramer haben untersucht und gezeigt [13], wie Proteine nichtlinear von ihnen abhängen.

In diesem Kapitel wird beschrieben, wie die SMA-Parameter eines Proteins durch die in Kapitel 2.2 beschriebene Parameteridentifikationsmethode geschätzt werden können. In Vorbereitung dessen wird im folgenden Abschnitt zunächst das Konvektions-Diffusions-Gleichungssystem zur Beschreibung eines präparativen Säulenchromatographieprozesses vorgestellt, in dem die SMA-Parameter als Modellparameter vorkommen. Da zur Schätzung dieser Parameter eine numerische Lösung dieses Modells erforderlich ist, wird zudem gezeigt, wie diese mithilfe der Methode der finiten Elemente und dem Crank-Nicolson-Verfahren ermittelt werden kann.

In Kapitel 2.3 wurden drei Verfahren beschrieben, mit denen die SMA-Parameter aus ermittelten Messdaten geschätzt werden können. Diese Verfahren werden in diesem Kapitel am Beispiel des Proteins Lysozym bei pH 7 angewendet und miteinander verglichen. Für die Umsetzung dieser Verfahren werden die Sensitivitäten erster Ordnung und für ein Verfahren sogar die Sensitivitäten zweiter Ordnung benötigt. Wie diese als Finite-Elemente-Lösung ermittelt werden können, wird in dieser Arbeit detailliert beschrieben.

Da im allgemeinen die zur Parameterschätzung benötigten Messdaten lediglich am Ausgang einer Chromatographiesäule erhoben werden, wird anschließend anhand des Proteins Lysozym eine optimale Versuchsplanung durchgeführt um zu überprüfen, inwieweit diese Messposition sinnvoll gewählt ist und wo die D-optimalen Messstellen liegen.

6.1 Herleitung des Systems zur Beschreibung eines präparativen Chromatographieprozesses

Ein präparativer Chromatographieprozess zum Aufreinigen von Proteinen kann durch ein gekoppeltes System von parabolischen und gewöhnlichen Differentialgleichungen dargestellt werden. Dieses System besteht aus einem Haupt- und einem SMA-Modell, wobei beide Modelle zusammen die zeitliche Änderung der Protein- und der Salzkonzentration in einer Chromatographiesäule beschreiben. Hierfür unterteilt man die Säule in die Bereiche *stationäre Phase*, *Poren* und *mobile Phase*.

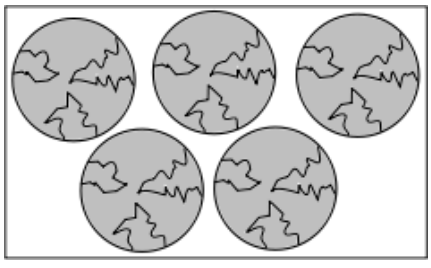


Abbildung 6.2: Das Gesamtvolumen V_{stat} der stationären Phase mit den Poren.

Die Konzentration einer Komponente (Protein oder Salz) in der stationären Phase wird mit $q_i(t,x)$, in den Poren mit $c_{p,i}(t,x)$ und in der mobilen Phase mit $c_i(t,x)$ bezeichnet. Als Index für die Salzkonzentration wird im folgenden $i := 0$ gewählt. Da vorausgesetzt

wird, dass das Proteingemisch aus insgesamt $0 < K \in \mathbb{N}$ Proteinen besteht, ist mit

$$I_K := \{1, \ldots, K\}$$

die Menge

$$I_{K,0} := I_K \cup \{0\}$$

die Indexmenge aller Komponenten.

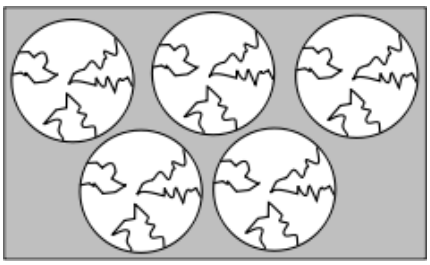


Abbildung 6.3: Das Volumen der gesamten Chromatographiesäule abzüglich V_{stat} ist das *Innere der Säule* und wird mit V_{int} bezeichnet. Durch diesen Bereich strömt die mobile Phase.

Das Hauptmodell beschreibt die zeitliche Änderung der Protein- und der Salzkonzentration in der mobilen Phase und in den Poren. Zusammen mit dem Diffusionskoeffizienten $D_{ax} > 0$ und der als konstant gewählten Geschwindigkeit $u_{ax} > 0$ besagt das Hauptmodell

$$\partial_t c_i(t,x) = D_{ax}\Delta_x c_i(t,x) - u_{ax}\,\nabla_x c_i(t,x) - \kappa_i\,[c_i(t,x) - c_{p,i}(t,x)], \quad (6.1)$$

$$\partial_t c_{p,i}(t,x) = \eta_i\,[c_i(t,x) - c_{p,0}(t,x)] - \frac{1-\varepsilon_p}{\varepsilon_p}\partial_t q_i(t,x), \quad (6.2)$$

für $i \in I_{K,0}$. Die Konstanten $\kappa_i > 0$ und $\eta_i > 0$ werden im folgenden noch beschrieben. Die Packungsdichte

$$\varepsilon = \frac{V_{stat}}{(V_{stat} + V_{int})} > 0 \quad (6.3)$$

beschreibt das Verhältnis zwischen der stationären Phase mit Poren und dem Gesamtvolumen der Säule. Die Porengröße

$$\varepsilon_p = \frac{V_{pore}}{V_{stat}} \quad (6.4)$$

beschreibt das Verhältnis vom Volumen der Poren V_{pore} zu V_{stat}.

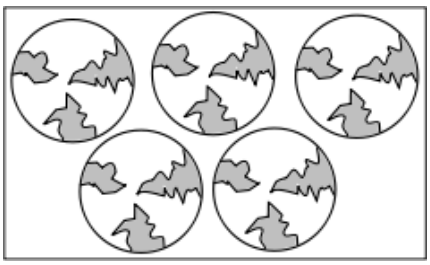


Abbildung 6.4: Die Poren haben ein Gesamtvolumen von V_{pore}.

Das Hauptmodell wird nun beschrieben. Die hierfür benötigte Theorie ist weitestgehend dem Buch [75], der Arbeit [13] und dem Skript [53] entnommen.

6.1.1 Hauptmodell

Für die Herleitung werden folgende Voraussetzungen gewählt:

Voraussetzung 6.1.1.

(V1) Die stationäre Phase besteht aus kugelförmigen Partikeln mit gleichem Radius, welche homogen angeordnet sind.

(V2) Sowohl die Dichte, als auch die Geschwindigkeit der Fluide sind konstant.

(V3) Der Prozess ist isothermisch.

(V4) Keine Konvektion in den Partikeln. Der flüssige Zustand in den Poren wird als stationär angenommen und wird nicht durch die Bewegung der mobilen Phase beeinträchtigt.

(V5) Es werden keine Gelpermeationseffekte (Größenausschlusseffekte) in Betracht gezogen. Somit wird vorausgesetzt, dass alle gelösten Stoffe den gesamten Porenraum durchdringen.

Sei K die Anzahl der Proteinkomponenten, die voneinander getrennt werden sollen. Dann sind zusammen mit der Salzkonzentration insgesamt $K+1$ Komponenten am Chromatographieprozess beteiligt. Für jede Komponente $i \in I_{K,0}$ gilt genau eine Konvektions-Diffusionsgleichung. Die Herleitung dieser Gleichung verläuft für jede Komponente analog.

Als Grundlage der Konvektions-Diffusions-Gleichung gilt das *Prinzip der Massenerhaltung*. Dieses besagt, dass für jedes feste Kontrollvolumen $V \subset \Omega$

$$\frac{d}{dt}\int_V \varrho(t,x)\,dx \;=\; -\int_{\partial V} j(t,x)\cdot n\,d\sigma + \int_V f(t,x)\,dx \quad \forall\, V \subset \Omega, \tag{6.5}$$

gilt, wobei $\varrho(t,x)$ die Dichte des Fluids, $j(t,x)$ die Teilchenstromdichte und $f(t,x)$ die Quelldichte darstellt. Zusammen mit der Geschwindigkeit $v(t,x)$ gilt der Zusammenhang

$$j(t,x) \;=\; \varrho(t,x)v(t,x).$$

Gleichung (6.5) besagt, dass die zeitliche Massenänderung sich aus dem Massenabfluss am Rand und dem Massenzufluss der Quelle ergibt. Nach dem Divergenztheorem [53] gilt ferner

$$\int_V \nabla_x \cdot j(t,x)\,dx \;=\; \int_{\partial V} j(t,x)\cdot n\,d\sigma,$$

so dass (6.5) umformuliert werden kann zu

$$\frac{d}{dt}\int_V \varrho(t,x)\,dx \;=\; -\int_V \nabla_x \cdot j(t,x)\,dx + \int_V f(t,x)\,dx \quad \forall\, V \subset \Omega. \tag{6.6}$$

Wie in [53] beschrieben, lässt sich die Teilchenstromdichte $j(t,x)$ zerlegen in einen Diffusionsstromdichtenanteil $j_D(t,x)$ und einen Transportstromdichtenanteil $j_T(t,x)$, d. h.

$$j(t,x) \;=\; j_D(t,x) + j_T(t,x).$$

Was den Diffusionsanteil anbelangt besagt das erste Fick'sche Gesetz [53], dass sich der Diffusionsstrom in Richtung des größten Konzentrationsgefälles bewegt. Daher ist

$$j_D(t,x) \;=\; -D\,\nabla_x c(t,x)$$

mit dem Diffusionskoeffizienten $D > 0$ und der Konzentration $c(t,x)$. Es wird nun vorausgesetzt, dass die Dichte $\varrho(t,x)$ proportional zur Konzentration $c(t,x)$ ist, d. h.

$$\varrho(t,x) \;=\; \varrho_0\, c(t,x),$$

mit $\varrho_0 > 0$ konstant.

Für den Transportanteil gilt mit der Geschwindigkeit $v_T(t,x)$ nach [53]

$$j_T(t,x) \;=\; v_T(t,x)\,\varrho(t,x) \;=\; \varrho_0\, v_T(t,x)\, c(t,x),$$

so dass für Gleichung (6.6) gilt (Argumente werden weggelassen)

$$\begin{aligned}
\varrho_0 \frac{d}{dt}\int_V c\,dx &= -\int_V \nabla_x\cdot(j_D + j_T)\,dx + \int_V f\,dx \\
&= -\int_V \nabla_x\cdot j_D\,dx - \int_V \nabla_x\cdot j_T\,dx + \int_V f\,dx \qquad (6.7)\\
&= D\int_V \Delta_x c\,dx - \varrho_0\int_V \nabla_x\cdot(v_T\,c)\,dx + \int_V f\,dx \\
&= D\int_V \Delta_x c\,dx - \varrho_0\int_V (\nabla_x\cdot v_T)\,c\,dx - \varrho_0\int_V v_T\,\nabla_x c\,dx + \int_V f\,dx,
\end{aligned}$$

$\forall\, V \subset \Omega$, wobei zur besseren Übersicht

$$c \;:=\; c(t,x), \quad j_D := j_D(t,x), \quad j_T := j_T(t,x) \quad \text{und} \quad f := f(t,x)$$

gesetzt wurde. Folglich ist

$$\frac{d}{dt}\int_V c\,dx \;=\; D_{ax}\int_V \Delta_x c\,dx - \int_V (\nabla_x\cdot v_T)\,c\,dx - \int_V v_T\,\nabla_x c\,dx + \frac{1}{\varrho_0}\int_V f\,dx,$$

$\forall\, V \subset \Omega$ mit $D_{ax} := \dfrac{D}{\varrho_0}$.

Da die Voraussetzung 6.1.1 erfüllt ist, ist die Transportgeschwindigkeit des Fluids $v_T(t,x)$ in jedem Kontrollvolumen $V \subset \Omega$ konstant und wird im folgenden durch

$$u_{ax} \;:=\; v_T(t,x) > 0$$

dargestellt. Dann ist $\nabla_x \cdot v_T = 0$ und es gilt

$$\frac{d}{dt}\int_V c\,dx \;=\; D_{ax}\int_V \Delta_x c\,dx - u_{ax}\int_V \nabla_x c\,dx + \frac{1}{\varrho_0}\int_V f\,dx, \qquad (6.8)$$

$\forall\, V \subset \Omega$.

Nun wird der Quellterm $\displaystyle\int_V f(t,x)\,dx$ betrachtet. Wie in [75] beschrieben ist die Kraft $f(t,x)$ bei der präparativen Säulenchromatographie proportional zur Differenz $c(t,x) - c_p(t,x)$, wobei $c(t,x)$ die Konzentration einer Komponente in der mobilen Phase und

$c_p(t,x)$ die Konzentration der selben Komponente in den Poren darstellt. Somit existiert eine Konstante $\kappa > 0$, so dass

$$f(t,x) \;=\; -\varrho_0\,\kappa\,(c(t,x) - c_p(t,x))$$

ist. Somit gilt

$$\frac{d}{dt}\int_V c\,dx \;=\; D_{ax}\int_V \Delta_x c\,dx - u_{ax}\int_V \nabla_x c\,dx - \kappa\int_V c - c_p\,dx. \tag{6.9}$$

Es wird nun vorausgesetzt, dass $\frac{d}{dt}c(t,x)$ auf jedem Kontrollvolumen $V \subset \Omega$ gleichmäßig stetig ist, so dass die Vertauschung von Differentiation und Integration erlaubt ist [53] und

$$\int_V \partial_t c - D_{ax}\,\Delta_x c + u_{ax}\,\nabla_x c + \kappa\,(c - c_p)\,dx = 0$$

$\forall\ V \subset \Omega$ gilt. Wie in [53] beschrieben, folgt

$$\partial_t c \;=\; D_{ax}\,\Delta_x c - u_{ax}\,\nabla_x c - \kappa\,(c - c_p).$$

Für die Salz- und Proteinkonzentrationen in der mobilen Phase gelten somit die Konvektions-Diffusionsgleichungen

$$\partial_t c_i(t,x) \;=\; D_{ax}\Delta_x c_i(t,x) - u_{ax}\nabla_x c_i(t,x) - \kappa_i(c_i(t,x) - c_{p,i}(t,x)) \tag{6.10}$$

für $i \in I_{K,0}$. Die gewöhnlichen Differentialgleichungen

$$\partial_t c_{p,i}(t,x) \;=\; \eta_i(c_i(t,x) - c_{p,i}(t,x)) - \frac{1-\varepsilon_p}{\varepsilon_p}\partial_t q_i(t,x) \tag{6.11}$$

beschreiben die zeitliche Änderung der Konzentrationen in den Poren für $i \in I_{K,0}$.
Anfangsbedingung:

Die Funktionen $c_i(0,x), c_{p,i}(0,x)$ und $q_i(0,x)$ werden als konstant angenommen. Üblicherweise ist $c_i(0,x) = 0$ für $i \in I_K$.
Randbedingung:

Für die Stelle $x = 0$ wird für die Konzentration der $n+1$ Komponenten eine Robinsche Randbedingung angenommen

$$c_i(t,0) = c_{in,i}(t) + \frac{D_{ax}}{u_{ax}}\nabla_x c_i(t,0),\ c_{in,i}(t) = g(t) \text{ für } i \in I_{K,0}$$

und für $x = L_c$ gilt Neumann-Null-Randbedingung

$$\nabla_x c_i(t,L_c) = 0 \text{ für } i \in I_{K,0}.$$

Es ist $\kappa_i = 3\,\frac{k_{eff,i}}{r_p}\,\frac{(1-\varepsilon)}{\varepsilon}$ und $\eta_i = 3\,\frac{k_{eff,i}}{r_p}\,\frac{1}{\varepsilon_p}$

6.1.2 SMA-Modell

Das SMA-Modell für einen präparativen Chromatographieprozess beschreibt das Massengleichgewicht beim Ionenaustausch in der stationären Phase und wurde unter anderem in den Arbeiten [13] und [65] hergeleitet. Hierbei handelt es sich um ein gekoppeltes System von $K+1$ gewöhnlichen Differentialgleichungen, die jeweils die zeitliche Änderung einer Komponenten in der stationären Phase $q_i(t,x), i \in I_K$ beschreibt. Es gilt für jedes $i \in I_K$

$$\partial_t q_i(t,x) \;=\; k_{a,i} c_{p,i}(t,x) \left(\Lambda - \sum_{r=1}^{K} (\nu_r + \gamma_r) q_r(t,x) \right) - k_{d,i} q_i(t,x) c_{p,0}(t,x)^{\nu_i}, \quad (6.12)$$

wobei

- $k_{a,i} > 0$ den Adsorbtionskoeffizienten,
- $k_{d,i} > 0$ den Desorbtionskoeffizienten,
- $\nu_i > 0$ die Charakteristische Ladung und
- $\gamma_i > 0$ den Schirmungskoeffizienten

des i-ten Proteins darstellt. Diese Modellparameter werden SMA-Parameter genannt.

Zwischen der Salzkonzentration $q_0(t,x)$ und den Proteinen $q_i(t,x)$ in der stationären Phase gilt zudem der Zusammenhang

$$\partial_t q_0(t,x) \;=\; -\sum_{i=1}^{K} \nu_i \, \partial_t q_i(t,x). \quad (6.13)$$

6.2 Schwache Formulierung der Modellgleichungen

Das Konvektions-Diffusionsgleichungssystem zur Beschreibung eines präparativen Chromatographieprozesses wurde in Kapitel 6.1 vorgestellt und besteht aus den untereinander gekoppelten Gleichungen

$$\left.\begin{aligned}
\partial_t c_i(t,x) &= D_{ax} \Delta_x c_i(t,x) - u_{ax} \nabla_x c_i(t,x) - \kappa_i \big(c_i(t,x) - c_{p,i}(t,x)\big), \\
\partial_t c_{p,i}(t,x) &= \eta_i \big(c_i(t,x) - c_{p,i}(t,x)\big) - \frac{1-\varepsilon_p}{\varepsilon_p} \partial_t q_i(t,x), \\
\partial_t q_0(t,x) &= -\sum_{r=1}^{K} \nu_r \, \partial_t q_r(t,x), \\
\partial_t q_j(t,x) &= k_{a,j} c_{p,j}(t,x) \Big(\Lambda - \sum_{r=1}^{K} (\nu_r + \gamma_r) q_r(t,x) \Big)^{\nu_j} - k_{d,j} q_j(t,x) c_{p,0}(t,x)^{\nu_j},
\end{aligned}\right\} \quad (6.14)$$

mit der Robinschen-Randbedingungen

$$c_i(0,t) \;=\; c_{in,i}(t) + \frac{D_{ax}}{u_{ax}} \nabla_x c_i(0,t) \quad (6.15)$$

und der homogenen Neumann-Randbedingung

$$\nabla_x c_i(t, L_c) \;=\; 0, \tag{6.16}$$

für $i \in I_{K,0}$ und $j \in I_K$, wobei $c_{in,i}(t)$ die Konzentration der i-ten Komponente darstellt, die von außen in die Chromatographiesäule gedrückt wird.

Die Anfangsbedingungen

$$c_i(0,x) = g_i, \quad c_{p,i}(0,x) = g_{p,i} \quad \text{und} \quad q_i(0,x) = f_i \tag{6.17}$$

werden für $i \in I_{K,0}$ in Ω als konstant angenommen.

Da dieses gekoppelte System im allgemeinen nicht analytisch gelöst werden kann, wird dieses numerisch gelöst. Die Ortsdiskretisierung erfolgt - wie in Kapitel 5.2.1 am Beispiel der zweidimensionalen Wärmeleitungsgleichung beschrieben - mithilfe der Methode der finiten Elemente. Man erhält ein System von gewöhnlichen Differentialgleichungen erster Ordnung, welches numerisch mit dem Crank-Nicolson-Verfahren analog zur Vorgehensweise in Kapitel 5.2.2 gelöst wird.

Eine numerische Lösung des Systems (6.14) mit (6.15) - (6.17) wird benötigt um

(1) die unbekannten, proteinspezifischen SMA-Parameter

$$\boldsymbol{\theta}^* := (\boldsymbol{\theta}^{1,*\,\top}, \ldots, \boldsymbol{\theta}^{K,*\,\top})^\top \quad \text{mit} \quad \boldsymbol{\theta}^{i,*} := (k_{a,i}^*, k_{d,i}^*, \nu_i^*, \gamma_i^*)^\top \in \mathbb{R}^4 \tag{6.18}$$

aus Gleichung (6.14) mithilfe eines Parameterschätzers $\hat{\boldsymbol{\theta}}$ näherungsweise ermitteln zu können. Der Parameterschätzers $\hat{\boldsymbol{\theta}} := \hat{\boldsymbol{\theta}}(\mathbf{z}(t))$ hängt von Messdaten

$$\mathbf{z}(t) \;=\; (z_1(t,x_1), \ldots, z_\ell(t,x_\ell))^\top$$

ab, die an den fest gewählten Messpositionen $\vec{\mathbf{x}} = (x_1, \ldots, x_\ell)^\top \in \Omega^\ell$ erhoben wurden.

(2) ein D-optimales Design numerisch ermitteln zu können. Ein D-optimales Design gibt an, wie die Messstellen $\vec{\mathbf{x}} = (x_1, \ldots, x_\ell)^\top$ zu wählen sind, damit die Kovarianzmatrix cov$(\hat{\boldsymbol{\theta}})$ des Schätzers

$$\hat{\boldsymbol{\theta}} \;=\; (\hat{\boldsymbol{\theta}}^{1\,\top}, \ldots, \hat{\boldsymbol{\theta}}^{K\,\top})^\top$$

bezüglich des D-Optimalitätskriteriums (2.7) minimal ist. Dadurch kann der unbekannte Modellparameter $\boldsymbol{\theta}^*$ durch $\hat{\boldsymbol{\theta}}$ bestmöglich geschätzt werden.

Das Konvektions-Diffusionsgleichungssystem (6.14) mit (6.15) - (6.17) wird räumlich mit der Methode der finiten Elemente gelöst. In Vorbereitung dessen wird zunächst die zugehörige schwache Formulierung analog zu Kapitel 5.2 aufgestellt.

Zur übersichtlichen Darstellung ist im folgenden

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

wobei

$$\Omega := (0, L_c)$$

gilt und $0 < L_c \in \mathbb{R}$ die Länge der Chromatographiesäule darstellt.

In diesem Kapitel wird vorausgesetzt, dass für die Konzentrationen der Komponenten $i \in I_{K,0}$ für jedes fest gewählte $\boldsymbol{\theta} \in \Theta$

$$\begin{aligned} c_i(t,x,\boldsymbol{\theta}) &\in \mathcal{C}^1(T;V), && (6.19)\\ c_{p,i}(t,x,\boldsymbol{\theta}) &\in \mathcal{C}^1(T;V), && (6.20)\\ q_i(t,x,\boldsymbol{\theta}) &\in \mathcal{C}^1(T;V) && (6.21) \end{aligned}$$

ist und mit $c_i(t)(x) := c_i(t,x,\boldsymbol{\theta}), c_{p,i}(t)(x) := c_{p,i}(t,x,\boldsymbol{\theta})$ und $q_i(t)(x) := q_i(t,x,\boldsymbol{\theta})$ für jedes $t \in T$

$$c_i(t) \in V, \quad c_{p,i}(t) \in V \quad \text{und} \quad q_i(t) \in V \tag{6.22}$$

gilt.

Um die schwache Formulierung zu erhalten, wird wie in Kapitel 5.2 die Differentialgleichungen (6.14) mit einer beliebigen, aber festen Testfunktion $\varphi \in V$ multipliziert und anschließend über Ω integriert. Für die erste Gleichung aus (6.14) gilt dann für $i \in I_{K,0}$, $\boldsymbol{\theta} \in \Theta$ fest gewählt

$$\partial_t(c_i(t),\varphi)_L - D_{ax}(\Delta_x c_i(t),\varphi)_L + u_{ax}(\nabla_x c_i(t),\varphi)_L + \kappa_i(c_i(t) - c_{p,i}(t),\varphi)_L = 0,$$

$\forall\, \varphi \in V$. Nach der *ersten Greenschen Identität* gilt zudem

$$\begin{aligned} (\Delta_x c_i(t),\varphi)_L &= \left[\nabla_x c_i(t)(x)\varphi(x)\right]_{x=0}^{x=L_c} - (\nabla_x c_i(t), \nabla_x\varphi)_L \\ &= \nabla_x c_i(t)(L_c)\varphi(L_c) - \nabla_x c_i(t)(0)\varphi(0) - (\nabla_x c_i(t), \nabla_x\varphi)_L \\ &\overset{(6.16)}{=} -\nabla_x c_i(t)(0)\varphi(0) - (\nabla_x c_i(t), \nabla_x\varphi)_L \\ &\overset{(6.15)}{=} -\frac{u_{ax}}{D_{ax}}(c_i(t)(0) - c_{in,i}(t))\varphi(0) - (\nabla_x c_i(t), \nabla_x\varphi)_L. \end{aligned}$$

Somit lautet die erste Gleichung aus (6.14) in schwacher Form mit der Abbildung $a_i : V^3 \to \mathbb{R}$ für $i \in I_{K,0}$

$$\begin{aligned} \partial_t(c_i(t),\varphi)_L &= -D_{ax}(\nabla_x c_i(t), \nabla_x\varphi)_L - u_{ax}(\nabla_x c_i(t),\varphi)_L \\ &\quad -\kappa_i(c_i(t) - c_{p,i}(t),\varphi)_L - u_{ax}\, c_i(t)(0)\,\varphi(0) + u_{ax}\, c_{in,i}(t)\,\varphi(0) \\ &=: a_i(c_i(t), c_{p,i}(t), \varphi) + u_{ax}\, c_{in,i}(t)\,\varphi(0), \end{aligned} \tag{6.23}$$

$\forall \varphi \in V$.

Für die übrigen Gleichungen aus (6.14) gilt mit den Abbildungen $a_{p,i} : V^4 \to \mathbb{R}$, $b_0 : V^{K+1} \to \mathbb{R}$ und $b_j : V^{K+3} \to \mathbb{R}$ für $i \in I_{K,0}$ und $j \in I_K$

$$\begin{aligned} \partial_t(c_{p,i}(t),\varphi)_L &= \eta_i(c_i(t) - c_{p,i}(t),\varphi)_L - \frac{1-\varepsilon_p}{\varepsilon_p}\partial_t(q_i(t),\varphi)_L \\ &=: a_{p,i}(c_i(t), c_{p,i}(t), q_i(t), \varphi), \end{aligned} \tag{6.24}$$

$$\begin{aligned} \partial_t(q_0(t),\varphi)_L &= -\partial_t \sum_{i=1}^{K} \nu_i(q_i(t),\varphi)_L \\ &=: b_0(q_1(t),\ldots,q_K(t),\varphi), \end{aligned} \tag{6.25}$$

$$\begin{aligned}\partial_t(q_j(t),\varphi)_L &= k_{a,j}(c_{p,j}(t)\left(\Lambda-\sum_{r=1}^{K}(\nu_r+\gamma_r)q_r(t)\right)^{\nu_j},\varphi)_L \\ &\quad -k_{d,j}(q_j(t)\,c_{p,0}(t)^{\nu_j},\varphi)_L \\ &=: b_j(c_{p,0}(t),c_{p,j}(t),q_1(t),\ldots,q_K(t),\varphi),\end{aligned} \tag{6.26}$$

$\forall\varphi\in V$. Somit lautet das zu lösende Problem in der variationellen Formulierung:

Sei $\boldsymbol{\theta}$ fest gewählt. Gesucht sind die Konzentrationen $c_i(t), c_{p,i}(t), q_i(t)\in V$ für alle $i\in I_{K,0}$, so dass

$$\left.\begin{aligned}\partial_t(c_i(t),\varphi)_L &= a_i(c_i(t),c_{p,i}(t),\varphi)+u_{ax}\,c_{in,i}(t)\,\varphi(0),\\ \partial_t(c_{p,i}(t),\varphi)_L &= a_{p,i}(c_i(t),c_{p,i}(t),q_i(t),\varphi),\\ \partial_t(q_i(t),\varphi)_L &= \begin{cases}b_0(q_1(t),\ldots,q_K(t),\varphi), & \text{für } i=0,\\ b_i(c_{p,0}(t),c_{p,i}(t),q_1(t),\ldots,q_K(t),\varphi), & \text{sonst,}\end{cases}\end{aligned}\right\} \tag{6.27}$$

mit

$$(c_i(0),\varphi)_L=(g_i,\varphi)_L,\quad (c_{p,i}(0),\varphi)_L=(g_{p,i},\varphi)_L,\quad (q_i(0),\varphi)_L=(f_i,\varphi)_L, \tag{6.28}$$

$\forall\varphi\in V$ erfüllt sind.

Das System (6.27) besteht aus $K+1$ parabolischen und $2(K+1)$ gewöhnlichen Differentialgleichungen. Wie dieses gekoppelte System räumlich mit der Methode der finiten Elemente und zeitlich mit dem Crank-Nicolson-Verfahren gelöst werden kann, wird in den Kapiteln 6.4.1 und 6.4.2 beschrieben.

Zunächst werden die Sensitivitätsgleichungen erster Ordnung hergeleitet. Diese werden für die Schätzung der SMA-Parameter $\boldsymbol{\theta}^*$ und zur Bestimmung eines D-optimalen Designs benötigt. Da in dieser Arbeit neben einem Gradientenverfahren und dem Gauss-Newtonverfahren auch das klassische Newtonverfahren zur Schätzung der SMA-Parameter verwendet wird, werden zudem die hierfür benötigten Sensitivitätsgleichungen zweiter Ordnung aufgestellt.

6.3 Schätzung der SMA-Parameter

Wie in Kapitel 2.3 erläutert, können die unbekannten SMA-Parameter

$$\boldsymbol{\theta}^*=({\boldsymbol{\theta}^{1,*}}^\top,\ldots,{\boldsymbol{\theta}^{K,*}}^\top)^\top,\quad \text{mit}\quad \boldsymbol{\theta}^{i,*}=(k_{a,i}^*,k_{d,i}^*,\nu_i^*,\gamma_i^*)^\top, \tag{6.29}$$

die in dem System (6.14) enthalten sind, durch Lösen eines Optimierungsproblems

$$\hat{\boldsymbol{\theta}} \;=\; \arg\min_{\Theta} J(\boldsymbol{\theta}) \;=\; \arg\min_{\Theta}\frac{1}{2}\int_T\left\|\tilde{\mathbf{z}}(t)-\mathcal{P}\left(\sum_{i=1}^{K}c_i(t,x,\boldsymbol{\theta})\right)\right\|_2^2 dt, \tag{6.30}$$

geschätzt werden, wobei $c_i(t,x,\boldsymbol{\theta})$ für $i\in I_K$ eine Lösung von (6.27) - (6.28) ist,

$$\tilde{\mathbf{z}}(t):=\left(\tilde{z}_1(t),\ldots,\tilde{z}_\ell(t)\right)^\top$$

den Vektor der gemittelten Messreihen darstellt und

$$\mathcal{P}\left(\sum_{i=1}^{K} c_i(t,x,\boldsymbol{\theta})\right) := \sum_{i=1}^{K}(c_i(t,x_1,\boldsymbol{\theta}),\ldots,c_i(t,x_\ell,\boldsymbol{\theta}))^\top$$

den Projektionsoperator $\mathcal{P}$ aus Definition 2.2.1. Die gemittelte Messreihe $\tilde{z}_i(t)$ an einer Messposition $x_i \in \overline{\Omega}$ für $i \in I_\ell$ erhält man nach $M > 0$ Experimenten durch

$$\tilde{z}_i(t) := \frac{1}{M}\sum_{k=1}^{M} z_i^k(t),$$

wobei $z_i^k(t)$ die Messreihe an der Stelle x_i im k-ten Experiment darstellt. Das Ergebnis $\hat{\boldsymbol{\theta}}$ liefert dann eine Schätzung des exakten Wertes $\boldsymbol{\theta}^*$ aus (6.29).

6.3.1 Sensitivitätsgleichungen erster Ordnung

In diesem Abschnitt werden die Sensitivitätsgleichungen erster Ordnung aufgestellt, die durch partielle Ableitung der Zustandsgleichungen (6.14) nach den SMA-Parametern $\boldsymbol{\theta}$ gebildet werden.

Sensitivitäten erster Ordnung in der mobilen Phase

In der mobilen Phase gilt mit der Salzkonzentration $c_0(t) := c_0(t,x,\boldsymbol{\theta})$ und den Proteinkonzentrationen $c_i(t) := c_i(t,x,\boldsymbol{\theta})$ für jedes $i \in I_K$ die Konvektions-Diffusionsgleichung

$$\partial_t c_i(t) \;=\; D_{ax}\Delta_x c_i(t) - u_{ax}\nabla_x c_i(t) - \kappa_i(c_i(t) - c_{p,i}(t)) \tag{6.31}$$

mit den Randbedingungen

$$\left.\begin{aligned} \frac{D_{ax}}{u_{ax}}\nabla_x c_i(t,0,\boldsymbol{\theta}) - c_i(t,0,\boldsymbol{\theta}) + c_{in,i}(t) &= 0,\\ \nabla_x c_i(t,L_c,\boldsymbol{\theta}) &= 0, \end{aligned}\right\} \tag{6.32}$$

und der Anfangsbedingung

$$c_i(0) \;=\; g_i. \tag{6.33}$$

Leitet man die Differentialgleichungen (6.31) partiell nach den SMA-Parametern ab, erhält man für $i \in I_{K,0}$ und $j \in I_K$

$$\begin{aligned}
\partial_t[\partial_{k_{a,j}} c_i(t)] &= D_{ax}\Delta_x[\partial_{k_{a,j}} c_i(t)] - u_{ax}\nabla_x[\partial_{k_{a,j}} c_i(t)] - \kappa_i\Big(\partial_{k_{a,j}} c_i(t) - \partial_{k_{a,j}} c_{p,i}(t)\Big),\\
\partial_t[\partial_{k_{d,j}} c_i(t)] &= D_{ax}\Delta_x[\partial_{k_{d,j}} c_i(t)] - u_{ax}\nabla_x[\partial_{k_{d,j}} c_i(t)] - \kappa_i\Big(\partial_{k_{d,j}} c_i(t) - \partial_{k_{d,j}} c_{p,i}(t)\Big),\\
\partial_t[\partial_{\nu_j} c_i(t)] &= D_{ax}\Delta_x[\partial_{\nu_j} c_i(t)] - u_{ax}\nabla_x[\partial_{\nu_j} c_i(t)] - \kappa_i\Big(\partial_{\nu_j} c_i(t) - \partial_{\nu_j} c_{p,i}(t)\Big),\\
\partial_t[\partial_{\gamma_j} c_i(t)] &= D_{ax}\Delta_x[\partial_{\gamma_j} c_i(t)] - u_{ax}\nabla_x[\partial_{\gamma_j} c_i(t)] - \kappa_i\Big(\partial_{\gamma_j} c_i(t) - \partial_{\gamma_j} c_{p,i}(t)\Big).
\end{aligned}$$

Dieses System besteht aus insgesamt $4K(K+1)$ Konvektions-Diffusions-Gleichungen und wird mit

$$\boldsymbol{\theta}^j \;=\; (\theta_1^j,\theta_2^j,\theta_3^j,\theta_4^j)^\top \;:=\; (k_{a,j},k_{d,j},\nu_j,\gamma_j)^\top \tag{6.34}$$

im folgenden durch

$$\partial_t[\partial_{\theta_k^j} c_i(t)] = D_{ax}\Delta_x[\partial_{\theta_k^j} c_i(t)] - u_{ax}\nabla_x[\partial_{\theta_k^j} c_i(t)] - \kappa_i\Big(\partial_{\theta_k^j} c_i(t) - \partial_{\theta_k^j} c_{p,i}(t)\Big) \quad (6.35)$$

für $i \in I_{K,0}$, $j \in I_K$ und $k = 1, \ldots, 4$ dargestellt.

Die Eingangskonzentration $c_{in,i}(t)$ der i-ten Komponente ist unabhängig von $\boldsymbol{\theta} \in \Theta$, so dass für $i \in I_{K,0}$, $j \in I_K$ und $k = 1, \ldots, 4$

$$\partial_{\theta_k^j} c_{in,i}(t) \;=\; 0$$

gilt. Somit erhält man für die Sensitivitätsgleichungen (6.35) die Robinschen-Randbedingungen

$$\partial_{\theta_k^j} c_i(t, 0, \boldsymbol{\theta}) \;=\; \frac{D_{ax}}{u_{ax}}\nabla_x[\partial_{\theta_k^j} c_i(t, 0, \boldsymbol{\theta})] \quad (6.36)$$

und die homogenen Neumann-Randbedingungen

$$\nabla_x[\partial_{\theta_k^j} c_i(t, L_c, \boldsymbol{\theta})] \;=\; \mathbf{0} \quad (6.37)$$

für alle $i \in I_{K,0}$, $j \in I_K$ und $k = 1, \ldots, 4$. Da (6.33) ebenfalls unabhängig von den SMA-Parametern $\boldsymbol{\theta}$ ist, erhält man für die Sensitivitäten in der mobilen Phase die homogenen Anfangsbedingungen

$$\partial_{\theta_k^j} c_i(0, x, \boldsymbol{\theta}) \;=\; 0, \quad (6.38)$$

für alle $i = 0, \ldots, K$, $j = 1, \ldots, K$ und $k = 1, \ldots, 4$.

Um die Sensitivitätsgleichungen in der mobilen Phase in schwacher Form zu erhalten, werden die Gleichungen (6.35) mit einer beliebigen, aber festen Testfunktion $\varphi \in V$ multipliziert und über Ω integriert. Es gilt dann für $i \in I_{K,0}$, $j \in I_K$ und $k = 1, \ldots, 4$

$$\begin{aligned}\partial_t(\partial_{\theta_k^j} c_i(t), \varphi)_L - D_{ax}(\Delta[\partial_{\theta_k^j} c_i(t)], \varphi)_L + u_{ax}(\nabla_x[\partial_{\theta_k^j} c_i(t)], \varphi)_L \\ +\kappa_i(\partial_{\theta_k^j} c_i(t) - \partial_{\theta_k^j} c_{p,i}(t), \varphi)_L \;=\; 0,\end{aligned}$$

$\forall\, \varphi \in V$.

Nach der *ersten Greenschen Identität* gilt

$$\begin{aligned}(\Delta_x[\partial_{\theta_k^j} c_i(t)], \varphi)_L \;&=\; \Big[\nabla_x[\partial_{\theta_k^j} c_i(t)(x)]\varphi(x)\Big]_{x=0}^{x=L_c} - (\nabla_x[\partial_{\theta_k^j} c_i(t)], \nabla_x\varphi)_L \\ &=\; \nabla_x[\partial_{\theta_k^j} c_i(t)(L_c)]\varphi(L_c) - \nabla_x[\partial_{\theta_k^j} c_i(t)(0)]\varphi(0) \\ &\quad -(\nabla_x[\partial_{\theta_k^j} c_i(t)], \nabla_x\varphi)_L \\ &\overset{(6.36)}{=}\; -\nabla_x[\partial_{\theta_k^j} c_i(t)(0)]\varphi(0) - (\nabla_x[\partial_{\theta_k^j} c_i(t)], \nabla_x\varphi)_L \\ &\overset{(6.37)}{=}\; -\frac{u_{ax}}{D_{ax}}[\partial_{\theta_k^j} c_i(t)(0)]\varphi(0) - (\nabla_x[\partial_{\theta_k^j} c_i(t)], \nabla_x\varphi)_L.\end{aligned}$$

Somit lauten die ersten $4K(K+1)$ Sensitivitätsgleichungen erster Ordnung in schwacher Form mit der Abbildung $a_i : V^3 \to \mathbb{R}$ aus (6.23) für $i \in I_{K,0}$, $j \in I_K$ und $k = 1, \ldots, 4$

$$\partial_t(\partial_{\theta_k^j} c_i(t), \varphi)_L \;=\; a_i(\partial_{\theta_k^j} c_i(t), \partial_{\theta_k^j} c_{p,i}(t), \varphi), \quad (6.39)$$

mit der Anfangsbedingung

$$\left(\partial_{\theta_k^j} c_i(0), \varphi\right)_L = 0 \tag{6.40}$$

$\forall \varphi \in V$.

Sensitivitäten erster Ordnung in den Poren

In den Poren gilt mit der Salzkonzentration $c_{p,0}(t) := c_{p,0}(t, x, \boldsymbol{\theta})$ und den Proteinkonzentrationen $c_{p,j}(t) := c_{p,j}(t, x, \boldsymbol{\theta})$ für $j \in I_K$ die gewöhnliche Differentialgleichung

$$\partial_t c_{p,i}(t) \;=\; \eta_i\Big(c_i(t) - c_{p,i}(t)\Big) - \frac{1-\varepsilon_p}{\varepsilon_p}\partial_t q_i(t), \tag{6.41}$$

mit der Anfangsbedingung

$$c_{p,i}(0) \;=\; g_{p,i} \tag{6.42}$$

für $i \in I_{K,0}$. Leitet man die Gleichungen (6.41) partiell nach $\boldsymbol{\theta}$ ab, erhält man für $i \in I_{K,0}$ und $j \in I_K$ die $4K(K+1)$ Sensitivitätsgleichungen

$$\begin{aligned}
\partial_t[\partial_{k_{a,j}} c_{p,i}(t)] &= \eta_i\Big(\partial_{k_{a,j}} c_i(t) - \partial_{k_{a,j}} c_{p,i}(t)\Big) - \frac{1-\varepsilon_p}{\varepsilon_p}\partial_t[\partial_{k_{a,j}} q_i(t)],\\
\partial_t[\partial_{k_{d,j}} c_{p,i}(t)] &= \eta_i\Big(\partial_{k_{d,j}} c_i(t) - \partial_{k_{d,j}} c_{p,i}(t)\Big) - \frac{1-\varepsilon_p}{\varepsilon_p}\partial_t[\partial_{k_{d,j}} q_i(t)],\\
\partial_t[\partial_{\nu_j} c_{p,i}(t)] &= \eta_i\Big(\partial_{\nu_j} c_i(t) - \partial_{\nu_j} c_{p,i}(t)\Big) - \frac{1-\varepsilon_p}{\varepsilon_p}\partial_t[\partial_{\nu_j} q_i(t)],\\
\partial_t[\partial_{\gamma_j} c_{p,i}(t)] &= \eta_i\Big(\partial_{\gamma_j} c_i(t) - \partial_{\gamma_j} c_{p,i}(t)\Big) - \frac{1-\varepsilon_p}{\varepsilon_p}\partial_t[\partial_{\gamma_j} q_i(t)],
\end{aligned}$$

wobei diese Gleichungen mit $\boldsymbol{\theta}^j$ aus (6.34) für $k = 1, \ldots, 4$ im folgenden durch

$$\partial_t[\partial_{\theta_k^j} c_{p,i}(t)] \;=\; \eta_i\Big(\partial_{\theta_k^j} c_i(t) - \partial_{\theta_k^j} c_{p,i}(t)\Big) - \frac{1-\varepsilon_p}{\varepsilon_p}\partial_t[\partial_{\theta_k^j} q_i(t)] \tag{6.43}$$

dargestellt wird.

Wie in der mobilen Phase sind auch die Anfangsbedingungen der Konzentrationen in den Poren aus (6.42) unabhängig von den SMA-Parametern $\boldsymbol{\theta}$. Somit gilt zum Zeitpunkt $t = 0$

$$\partial_{\theta_k^j} c_{p,i}(0) \;=\; 0, \tag{6.44}$$

für $i \in I_{K,0}$, $j \in I_K$ und $k = 1, \ldots, 4$.

Analog zu den Sensitivitätsgleichungen in der mobilen Phase werden die Gleichungen (6.43) + (6.44) mit einer beliebigen, aber festen Testfunktion $\varphi \in V$ multipliziert und anschließend über Ω integriert. Auf diese Weise erhält man die Sensitivitätsgleichungen erster Ordnung in den Poren in schwacher Form. Es gilt für $i \in I_{K,0}$, $j \in I_K$ und $k = 1, \ldots, 4$

$$\partial_t(\partial_{\theta_k^j} c_{p,i}(t), \varphi)_L - \eta_i(\partial_{\theta_k^j} c_i(t) - \partial_{\theta_k^j} c_{p,i}(t), \varphi)_L + \frac{1-\varepsilon_p}{\varepsilon_p}\partial_t(\partial_{\theta_k^j} q_i(t), \varphi)_L \;=\; 0,$$

$\forall\, \varphi \in V$.

Somit lauten die $4K(K+1)$ Sensitivitätsgleichungen erster Ordnung in den Poren in schwacher Form mit der Abbildung $a_{p,i} : V^4 \to \mathbb{R}$ aus (6.24) für $i \in I_{K,0}$, $j \in I_K$ und $k = 1, \ldots, 4$

$$\partial_t (\partial_{\theta_k^j} c_{p,i}(t), \varphi)_L \;=\; a_{p,i}(\partial_{\theta_k^j} c_i(t), \partial_{\theta_k^j} c_{p,i}(t), \partial_{\theta_k^j} q_i(t), \varphi), \tag{6.45}$$

mit der Anfangsbedingung

$$(\partial_{\theta_k^j} c_{p,i}(0), \varphi)_L = 0, \tag{6.46}$$

$\forall \varphi \in V$.

Sensitivitäten erster Ordnung der Salzkonzentration in der stationären Phase

In der stationären Phase gilt für die Salzkonzentration $q_0(t)$ die gewöhnliche Differentialgleichung

$$\partial_t q_0(t) \;=\; -\sum_{r=1}^{K} \nu_r\, \partial_t q_r(t),$$

mit der Anfangsbedingung

$$q_0(0) \;=\; f_0.$$

Leitet man diese Gleichung nach den SMA-Parametern $\boldsymbol{\theta}$ ab, gilt mit $\boldsymbol{\theta}^j$ aus (6.34) für $j \in I_K$

$$\left.\begin{aligned}
\partial_t [\partial_{\theta_1^j} q_0(t)] &= -\textstyle\sum_{r=1}^{K} \theta_3^r\, \partial_t [\partial_{\theta_1^j} q_r(t)], \\
\partial_t [\partial_{\theta_2^j} q_0(t)] &= -\textstyle\sum_{r=1}^{K} \theta_3^r\, \partial_t [\partial_{\theta_2^j} q_r(t)], \\
\partial_t [\partial_{\theta_3^j} q_0(t)] &= -\partial_t q_j(t) - \textstyle\sum_{r=1}^{K} \theta_3^r\, \partial_t [\partial_{\theta_3^j} q_r(t)], \\
\partial_t [\partial_{\theta_4^j} q_0(t)] &= -\textstyle\sum_{r=1}^{K} \theta_3^r\, \partial_t [\partial_{\theta_4^j} q_r(t)],
\end{aligned}\right\} \tag{6.47}$$

mit der Anfangsbedingung für $k = 1, \ldots, 4$

$$\partial_{\theta_k^j} q_0(0) \;=\; 0. \tag{6.48}$$

Multipliziert man (6.47) mit einer beliebigen, aber festen Testfunktion $\varphi \in V$ und integriert über Ω, gilt mit der Abbildung $Q_k : V^2 \to \mathbb{R}$ für $k = 1, \ldots, 4$ definiert durch

$$Q_k(v(t), \varphi) \;=\; \begin{cases} -\partial_t (v(t), \varphi)_L, & \text{für } k = 3, \\ 0, & \text{sonst,} \end{cases}$$

und mit der Abbildung $b_0 : V^{K+1} \to \mathbb{R}$ aus (6.25)

$$\partial_t (\partial_{\theta_k^j} q_0(t), \varphi)_L \;=\; b_0(\partial_{\theta_k^j} q_1(t), \ldots, \partial_{\theta_k^j} q_K(t), \varphi) \;+\; Q_k(q_j(t), \varphi), \tag{6.49}$$

mit der Anfangsbedingung für $j \in I_K$ und $k = 1, \ldots, 4$

$$(\partial_{\theta_k^j} q_0(0), \varphi)_L = 0, \tag{6.50}$$

$\forall \varphi \in V$.

Sensitivitäten erster Ordnung der Proteine in der stationären Phase

Für die i-te Proteinkonzentration $q_i(t)$ gilt die Gleichung

$$\partial_t q_i(t) \;=\; k_{a,i} c_{p,i} \Big(\Lambda - \sum_{r=1}^{K} (\nu_r + \gamma_r) q_r(t) \Big)^{\nu_j} - k_{d,i} q_i(t) c_{p,1}(t)^{\nu_i} \tag{6.51}$$

mit der Anfangsbedingung

$$q_i(0) \;=\; f_i.$$

Leitet man diese Gleichungen partiell nach den SMA-Parametern $\boldsymbol{\theta}$ ab, erhält man die $4K^2$ Sensitivitätsgleichungen in der stationären Phase. Um diese Gleichungen in schwacher Form zu erhalten, werden diese wiederum mit einer beliebigen, aber festen Testfunktion $\varphi \in V$ multipliziert und über Ω integriert. Zusammen mit

- der Abbildung $A_i : V^{2K+5} \to \mathbb{R}$ für $i \in I_K$ definiert durch

$$\begin{aligned} A_i(c_{p,0}(t), c_{p,i}(t), q_1(t), \ldots, q_K(t), v_1(t), v_2(t), w_1(t), \ldots, w_K(t), \varphi) \;=\; & \\ & k_{a,i} \big(v_1(t) \Big(\Lambda - \sum_{r=1}^{K} (\nu_r + \gamma_r) q_r(t) \Big)^{\nu_i}, \varphi \big)_L \\ & - k_{a,i} \nu_i \big(c_{p,i}(t) \Big(\Lambda - \sum_{r=1}^{K} (\nu_r + \gamma_r) q_r(t) \Big)^{\nu_i - 1} \sum_{r=1}^{K} (\nu_r + \gamma_r) w_r(t), \varphi \big)_L \\ & - k_{d,i} \big(w_i(t) c_{p,0}(t)^{\nu_i}, \varphi \big)_L - k_{d,i} \nu_i \big(q_i(t) c_{p,0}(t)^{\nu_i - 1} v_2(t), \varphi \big)_L, \end{aligned}$$

- der Abbildung $F^i_{j,1} : V^{2K+3} \to \mathbb{R}$ definiert durch

$$F^i_{j,1}(c_{p,0}(t), \ldots, c_{p,K}(t), q_0(t), \ldots, q_K(t), \varphi) \;=\; \begin{cases} \big(c_{p,i}(t) \Big(\Lambda - \sum_{r=1}^{K} (\nu_r + \gamma_r) q_r(t) \Big)^{\nu_i}, \varphi \big)_L & \text{falls } i = j, \\ 0, & \text{sonst,} \end{cases}$$

- der Abbildung $F^i_{j,2} : V^{2K+3} \to \mathbb{R}$ definiert durch

$$F^i_{j,2}(c_{p,0}(t), \ldots, c_{p,K}(t), q_0(t), \ldots, q_K(t), \varphi) \;=\; \begin{cases} -\big(q_i(t) c_{p,0}(t)^{\nu_i}, \varphi \big)_L & \text{falls } i = j, \\ 0, & \text{sonst,} \end{cases}$$

- der Abbildung $F^i_{j,3} : V^{2K+3} \to \mathbb{R}$ definiert durch

$$F^i_{j,3}(c_{p,0}(t),\dots,c_{p,K}(t),q_0(t),\dots,q_K(t),\varphi) \;=\; \begin{cases} k_{a,i}(c_{p,i}(t)\Big(\Lambda-\sum_{r=1}^{K}(\nu_r+\gamma_r)q_r(t)\Big)^{\nu_i}\ln\Big(\Lambda-\sum_{r=1}^{K}(\nu_r+\gamma_r)q_r(t)\Big),\varphi)_L \\ -k_{a,i}\nu_i(c_{p,i}(t)\Big(\Lambda-\sum_{r=1}^{K}(\nu_r+\gamma_r)q_r(t)\Big)^{\nu_i-1}q_j(t),\varphi)_L \\ -k_{d,i}(q_i(t)c_{p,0}(t)^{\nu_i}\ln(c_{p,0}(t)),\varphi)_L & \text{falls } i=j, \\ -k_{a,i}\nu_i(c_{p,i}(t)\Big(\Lambda-\sum_{r=1}^{K}(\nu_r+\gamma_r)q_r(t)\Big)^{\nu_i-1}q_j(t),\varphi)_L, & \text{sonst,} \end{cases}$$

- der Abbildung $F^i_{j,4} : V^{2K+3} \to \mathbb{R}$ definiert durch

$$F^i_{j,4}(c_{p,0}(t),\dots,c_{p,K}(t),q_0(t),\dots,q_K(t),\varphi) \;=\; -k_{a,i}\nu_i(c_{p,i}(t)\Big(\Lambda-\sum_{r=1}^{K}(\nu_r+\gamma_r)q_r(t)\Big)^{\nu_i-1}q_j(t),\varphi)_L,$$

erhält man die Sensitivitätsgleichungen in der stationären Phase in schwacher Form:

Es gilt mit $\boldsymbol{\theta}^j$ aus (6.34) für $i \in I_K$, $j \in I_K$ und $k = 1,\dots,4$

$$\begin{aligned}\partial_t(\partial_{\theta^j_k}q_i(t),\varphi)_L \;&=\; F^i_{j,k}(c_{p,0}(t),\dots,c_{p,K}(t),q_0(t),\dots,q_K(t),\varphi) \\ &\quad + A_i(c_{p,0}(t),c_{p,i}(t),q_1(t),\dots,q_K(t),\partial_{\theta^j_k}c_{p,0}(t),\partial_{\theta^j_k}q_1(t),\dots,\partial_{\theta^j_k}q_K(t),\varphi),\end{aligned}$$

$\forall\varphi \in V$. Da die Anfangsbedingungen der Proteinkonzentrationen in der stationären Phase unabhängig von den SMA-Parametern $\boldsymbol{\theta}$ sind, gilt zum Zeitpunkt $t = 0$ für $i \in I_K$, $j \in I_K$ und $k = 1,\dots,4$

$$\partial_t(\partial_{\theta^j_k}q_i(0),\varphi) = 0,$$

$\forall\varphi \in V$.

Berechnung der Sensitivitäten erster Ordnung

Sei $\boldsymbol{\theta} \in \Theta$ fest gewählt. Um die Sensitivitäten erster Ordnung für $i \in I_{K,0}$

$$\nabla_{\boldsymbol{\theta}}c_i(t) \in V^{4K}, \quad \nabla_{\boldsymbol{\theta}}c_{p,i}(t) \in V^{4K} \quad \text{und} \quad \nabla_{\boldsymbol{\theta}}q_i(t) \in V^{4K}$$

in $\boldsymbol{\theta} \in \Theta$ zu bestimmen, werden zunächst die Zustände

$$c_i(t) \in V, \quad c_{p,i}(t) \in V \quad \text{und} \quad q_i(t) \in V$$

als Lösung des Zustandssystems (6.27) für $i \in I_{K,0}$ numerisch ermittelt. Mithilfe dieser Lösungen und den Abbildungen

- $Q_k(\varphi) \;:=\; Q_k(q_j(t),\varphi)$,

- $A_i(\partial_{\theta_k^j} c_{p,0}(t), \partial_{\theta_k^j} q_1(t), \ldots, \partial_{\theta_k^j} q_K(t), \varphi) :=$
$$A_i(c_{p,0}(t), c_{p,i}(t), q_1(t), \ldots, q_K(t), \partial_{\theta_k^j} c_{p,0}(t), \partial_{\theta_k^j} q_1(t), \ldots, \partial_{\theta_k^j} q_K(t), \varphi),$$
- $F_{j,k}^i(\varphi) := F_{j,k}^i(c_{p,0}(t), \ldots, c_{p,K}(t), q_0(t), \ldots, q_K(t), \varphi),$

für $i \in I_{K,0}$, $j \in I_K$ und $k = 1, \ldots, 4$ können dann die Sensitivitäten erster Ordnung sukzessive durch Lösen des Systems

$$\left.\begin{aligned}
\partial_t(\partial_{\theta_k^j} c_i(t), \varphi)_L &= a_i(\partial_{\theta_k^j} c_i(t), \partial_{\theta_k^j} c_{p,i}(t), \varphi), \\
\partial_t(\partial_{\theta_k^j} c_{p,i}(t), \varphi)_L &= a_{p,i}(\partial_{\theta_k^j} c_i(t), \partial_{\theta_k^j} c_{p,i}(t), \partial_{\theta_k^j} q_i(t), \varphi), \\
\partial_t(\partial_{\theta_k^j} q_0(t), \varphi)_L &= b_0(\partial_{\theta_k^j} q_1(t), \ldots, \partial_{\theta_k^j} q_K(t), \varphi) + Q_k(\varphi), \\
\partial_t(\partial_{\theta_k^j} q_i(t), \varphi)_L &= A_i(\partial_{\theta_k^j} c_{p,0}(t), \partial_{\theta_k^j} q_1(t), \ldots, \partial_{\theta_k^j} q_K(t), \varphi) + F_{j,k}^i(\varphi),
\end{aligned}\right\} \quad (6.52)$$

mit den Anfangsbedingungen

$$(\partial_{\theta_k^j} c_i(0), \varphi)_L = 0, \quad (\partial_{\theta_k^j} c_{p,i}(0), \varphi)_L = 0 \quad \text{und} \quad (\partial_{\theta_k^j} q_i(0), \varphi)_L = 0, \quad (6.53)$$

$\forall \varphi \in V$ numerisch ermittelt werden.

In dieser Arbeit werden das Zustandssystem (6.27) + (6.28) und die Sensitivitätsgleichungen erster Ordnung (6.52) + (6.53) räumlich mit der Methode der finiten Elemente und in der Zeit mit dem Crank-Nicolson-Verfahren gelöst. In Kapitel 6.4.1 und 6.4.2 wird diese Vorgehensweise beschrieben.

Zunächst werden die Sensitivitätsgleichungen zweiter Ordnung in schwacher Form aufgestellt. Eine Lösung dieser Gleichungen wird benötigt, falls die Schätzung der SMA-Parameter mit dem Newtonverfahren erfolgt.

6.3.2 Sensitivitätsgleichungen zweiter Ordnung

In diesem Abschnitt werden die Sensitivitätsgleichungen zweiter Ordnung in schwacher Form aufgestellt. Diese Gleichungen erhält man, indem die Sensitivitätsgleichungen erster Ordnung wiederum partiell nach den SMA-Parametern abgeleitet werden. Um die schwache Form zu erhalten, werden anschließend diese Gleichungen mit eine beliebigen, aber festen Testfunktion $\varphi \in V$ multipliziert und über Ω integriert. Da die Vorgehensweise analog zu Kapitel 6.3.1 ist, werden hier die Sensitivitätsgleichungen zweiter Ordnung in schwacher Form lediglich aufgelistet.

Sei $\boldsymbol{\theta} \in \Theta$ fest gewählt. Mit einer Lösung

$$c_i(t) \in V, \quad c_{p,i}(t) \in V \quad \text{und} \quad q_i(t) \in V$$

des Zustandssystems (6.27) + (6.28) und eine Lösung

$$\nabla_{\boldsymbol{\theta}} c_i(t) \in V^{4K}, \quad \nabla_{\boldsymbol{\theta}} c_{p,i}(t) \in V^{4K} \quad \text{und} \quad \nabla_{\boldsymbol{\theta}} q_i(t) \in V^{4K}$$

der Sensitivitätsgleichungen erster Ordnung (6.52) + (6.53), erhält man die Sensitivitäten zweiter Ordnung durch Lösen des Systems für $i \in I_{K,0}$, $j, j_2 \in I_K$ und

$k, k_2 = 1, \ldots, 4$

$$\left.\begin{aligned}
\partial_t(\partial_{\theta_{k_2}^{j_2}}[\partial_{\theta_k^j} c_i(t)], \varphi)_L &= a_i(\partial_{\theta_{k_2}^{j_2}}[\partial_{\theta_k^j} c_i(t)], \partial_{\theta_{k_2}^{j_2}}[\partial_{\theta_k^j} c_{p,i}(t)], \varphi), \\
\partial_t(\partial_{\theta_{k_2}^{j_2}}[\partial_{\theta_k^j} c_{p,i}(t)], \varphi)_L &= a_{p,i}(\partial_{\theta_{k_2}^{j_2}}[\partial_{\theta_k^j} c_i(t)], \partial_{\theta_{k_2}^{j_2}}[\partial_{\theta_k^j} c_{p,i}(t)], \partial_{\theta_{k_2}^{j_2}}[\partial_{\theta_k^j} q_i(t)], \varphi) \\
\partial_t(\partial_{\theta_{k_2}^{j_2}}[\partial_{\theta_k^j} q_0(t)], \varphi)_L &= b_0(\partial_{\theta_{k_2}^{j_2}}[\partial_{\theta_k^j} c_{p,0}(t)], \partial_{\theta_{k_2}^{j_2}}[\partial_{\theta_k^j} q_i(t)], \varphi) \; + \; \tilde{Q}_k(\varphi) \\
\partial_t(\partial_{\theta_{k_2}^{j_2}}[\partial_{\theta_k^j} q_i(t)], \varphi)_L &= \\
& A_i(\partial_{\theta_{k_2}^{j_2}}[\partial_{\theta_k^j} c_{p,0}(t)], \partial_{\theta_{k_2}^{j_2}}[\partial_{\theta_k^j} q_1(t)], \ldots, \partial_{\theta_{k_2}^{j_2}}[\partial_{\theta_k^j} q_K(t)], \varphi) \; + \; \tilde{F}_{j,k}^i(\varphi)
\end{aligned}\right\} \quad (6.54)$$

mit den Anfangsbedingungen

$$\begin{aligned}
(\partial_{\theta_{k_2}^{j_2}}[\partial_{\theta_k^j} c_i(0)], \varphi)_L &= 0, \\
(\partial_{\theta_{k_2}^{j_2}}[\partial_{\theta_k^j} c_{p,i}(0)], \varphi)_L &= 0, \\
(\partial_{\theta_{k_2}^{j_2}}[\partial_{\theta_k^j} q_i(0)], \varphi)_L &= 0,
\end{aligned} \quad (6.55)$$

$\forall \varphi \in V$.

6.4 Diskretisierung

6.4.1 Finite-Elemente-Diskretisierung im Ort

In diesem Abschnitt wird gezeigt, wie für jedes feste $t \in T$ eine schwache Lösung des Systems (6.27) + (6.28) mit der Methode der finiten Elemente bestimmt werden kann. Analog zu Kapitel 5.2.1 wird hierfür zunächst der Lösungsraum V auf einen endlichdimensionalen Teilraum V_h eingeschränkt. Man erhält dann das zu lösende, diskrete Variationsproblem:

Sei $\boldsymbol{\theta}$ fest gewählt. Gesucht sind $c_{i,h}(t), c_{p,i,h}(t), q_{i,h}(t) \in V_h$ für $i \in I_{K,0}$, so dass

$$\left.\begin{aligned}
\partial_t(c_{i,h}(t), \varphi)_L &= a_i(c_{i,h}(t), c_{p,i,h}(t), \varphi), \\
\partial_t(c_{p,i,h}(t), \varphi)_L &= a_{p,i}(c_{i,h}(t), c_{p,i,h}(t), q_{i,h}(t), \varphi), \\
\partial_t(q_{i,h}(t), \varphi)_L &= \begin{cases} b_0(q_{1,h}(t), \ldots, q_{K,h}(t), \varphi), & \text{für } i = 0, \\ b_i(c_{p,0,h}(t), c_{p,i,h}(t), q_{1,h}(t), \ldots, q_{K,h}(t), \varphi), & \text{sonst,} \end{cases}
\end{aligned}\right\} \quad (6.56)$$

mit

$$(c_{i,h}(0), \varphi)_L = (g_i, \varphi)_L, \quad (c_{p,i,h}(0), \varphi)_L = (g_{i,p}, \varphi)_L, \quad (q_{i,h}(0), \varphi)_L = (f_i, \varphi)_L, \quad (6.57)$$

$\forall \varphi \in V_h$ erfüllt sind.

Im folgenden sei

$$V_h := \{v \in \mathcal{C}(\bar{\Omega}) : \, v_{|_{T_i}} \in \mathcal{P}_1 \; \forall \, T_i \in \mathcal{T}_n\}$$

mit

$$\mathcal{P}_1 := \{v(x) = \alpha_0 + \alpha_1 x, \, \alpha_0, \alpha_1 \in \mathbb{R}\}$$

den Raum der Polynome vom Grad kleiner gleich 1.

Um das diskrete Variationsproblem (6.56) zu lösen, wird das abgeschlossene Gebiet $\overline{\Omega} = [0, L_c]$ in die äquidistanten Intervalle $T_i := [\chi_i, \chi_{i+1}]$ unterteilt, so dass

$$0 \;=\; \chi_1 \;\leq\; \chi_2 \;\leq\; \ldots \;\leq\; \chi_{N(h)-1} \;\leq\; \chi_{N(h)} \;=\; L_c$$

und $\chi_{i+1} - \chi_i = h > 0$ mit $h := \dfrac{L_c}{N(h)-1}$ ist.

Da V_h endlich dimensional ist, gilt

$$N(h) := \dim(V_h) < \infty.$$

Eine Basis von V_h wird mit $\{\Psi_i\}_{i=1}^{N(h)}$ bezeichnet. Dann können für jedes feste $t \in T$ die Funktionen $c_{i,h}(t), c_{p,i,h}(t), q_{i,h}(t) \in V_h$ dargestellt werden als Linearkombinationen

$$\left.\begin{array}{rcl} c_{i,h}(t)(x) & = & \sum_{j=1}^{N(h)} c_j^i(t)\Psi_j(x), \\ c_{p,i,h}(t)(x) & = & \sum_{j=1}^{N(h)} c_{p,j}^i(t)\Psi_j(x), \\ q_{i,h}(t)(x) & = & \sum_{j=1}^{N(h)} q_j^i(t)\Psi_j(x), \end{array}\right\} \tag{6.58}$$

$\forall\, i \in I_{K,0}$.

Insbesondere kann jede Testfunktion $\varphi \in V_h$ durch eine solche Linearkombination dargestellt werden. Dadurch vereinfacht sich das Variationsproblem zu:

Sei $\boldsymbol{\theta}$ fest gewählt. Gesucht sind $c_{i,h}(t), c_{p,i,h}(t), q_{i,h}(t) \in V_h$ für alle $i \in I_{K,0}$, so dass

$$\left.\begin{array}{l} \partial_t(c_{i,h}(t), \Psi_k)_L = a_i(c_{i,h}(t), c_{p,i,h}(t), \Psi_k), \\ \partial_t(c_{p,i,h}(t), \Psi_k)_L = a_{p,i}(c_{i,h}(t), c_{p,i,h}(t), q_{i,h}(t), \Psi_k), \\ \partial_t(q_{i,h}(t), \Psi_k)_L = \begin{cases} b_0(q_{1,h}(t), \ldots, q_{K,h}(t), \Psi_k), & \text{für } i = 0, \\ b_i(c_{p,0,h}(t), c_{p,i,h}(t), q_{1,h}(t), \ldots, q_{K,h}(t), \Psi_k), & \text{sonst,} \end{cases} \end{array}\right\} \tag{6.59}$$

mit

$$\left.\begin{array}{rcl} (c_{i,h}(0), \Psi_k)_L & = & (g_i, \Psi_k)_L, \\ (c_{p,i,h}(0), \Psi_k)_L & = & (g_{p,i}, \Psi_k)_L, \\ (q_{i,h}(0), \Psi_k)_L & = & (f_i, \Psi_k)_L, \end{array}\right\} \tag{6.60}$$

$\forall\, k = 1, \ldots, N(h)$ erfüllt sind.

Bemerkung 6.4.1. *Mithilfe der Linearkombinationen aus (6.58) lauten die Gleichungen (6.59) ausformuliert:*

$$\begin{aligned} \partial_t \sum_{j=1}^{N(h)} c_j^i(t)\,(\Psi_j, \Psi_k)_L \;=\; & \\ & -D_{ax} \sum_{j=1}^{N(h)} c_i^j(t)(\nabla_x \Psi_j, \nabla_x \Psi_k)_L - u_{ax} \sum_{j=1}^{N(h)} c_j^i(t)(\nabla_x \Psi_j, \Psi_k)_L \\ & -\kappa_i \sum_{j=1}^{N(h)} c_j^i(t)(\Psi_j, \Psi_k)_L + \kappa_i \sum_{j=1}^{N(h)} c_{p,j}^i(t)(\Psi_j, \Psi_k)_L \\ & -u_{ax}\, c_0^i(t)\Psi_k(0) + u_{ax}\, c_{in}^i(t)\Psi_k(0), \end{aligned}$$

$$
\begin{aligned}
\partial_t \sum_{j=1}^{N(h)} c_{p,j}^i(t)\,(\Psi_j,\Psi_k)_L &= \eta_i \sum_{j=1}^{N(h)} c_j^i(t)(\Psi_j,\Psi_k)_L - \eta_i \sum_{j=1}^{N(h)} c_{p,j}^i(t)(\Psi_j,\Psi_k)_L \\
&\quad + \frac{1-\varepsilon_p}{\varepsilon_p}\partial_t \sum_{j=1}^{N(h)} q_j^i(t)(\Psi_j,\Psi_k)_L, \\
\partial_t \sum_{j=1}^{N(h)} q_j^0(t)\,(\Psi_j,\Psi_k)_L &= -\sum_{i=1}^{K} \nu_i \left(\sum_{j=1}^{N(h)} q_j^i(t)(\Psi_j,\Psi_k)_L \right), \\
\partial_t \sum_{j=1}^{N(h)} q_j^i(t)\,(\Psi_j,\Psi_k)_L &= \\
&\quad k_{a,i} \sum_{j=1}^{N(h)} c_{p,j}^i(t)(\Psi_j \left(\Lambda - \sum_{r=1}^{K} (\nu_r+\gamma_r)(\sum_{p=1}^{N(h)} q_p^r(t)\Psi_p) \right)^{\nu_i}, \Psi_k)_L \\
&\quad -k_d \sum_{j=1}^{N(h)} q_j^i(t)(\Psi_j \left(\sum_{r=0}^{n} c_{p,r}^i(t)\Psi_r \right)^{\nu_i}, \Psi_k)_L,
\end{aligned}
$$

$\forall k = 1,\ldots,N(h)$.

Im folgenden wird eine Basis $\{\Psi_i(\mathbf{x})\}_{i=1}^{N(h)}$ des Raumes V_h derart gewählt, dass $\Psi_i(x_j) = \delta_{ij}\ \ \forall\, x_j \in \Omega_h$ und $\forall\, i,j \in \{1,\ldots,N(h)\}$ erfüllt ist, wobei

$$\Omega_h \ := \ \{x \in \Omega : x = kh,\, k \in \{0,\ldots,N(h)-1\}\}$$

mit $h := \dfrac{L_c}{N(h)-1}$ ist.

6.4.2 Zeitliche Diskretisierung

In Kapitel 5.2.2 wurde beschrieben, wie man eine gewöhnliche Differentialgleichung erster Ordnung mit dem Crank-Nicolson-Verfahren lösen kann. Da das Variationsproblem (6.59) für alle $i \in I_{K,0}$ und $k = 1,\ldots,N(h)$ mit den Anfangsbedingungen (6.60) ein System von gewöhnlichen Differentialgleichungen erster Ordnung darstellt, kann dieses ebenfalls mit dem Crank-Nicolson-Verfahren gelöst werden.

Analog zu Kapitel 5.2.2 wird zunächst das abgeschlossene Zeitintervall $\overline{T} = [0,t_f]$ in $N(\Delta t)-1$ äquidistante Intervalle der Form $[t_n,t_{n+1}]$ mit $\Delta t := t_{n+1} - t_n$ zerlegt, wobei $0 < N(\Delta t) \in \mathbb{N}$ und

$$0 \ = \ t_1 \ \leq \ t_2 \ \leq \ \ldots \ \leq \ t_{N(\Delta t)-1} \ \leq \ t_{N(\Delta t)} \ = \ t_f$$

ist. Dann gilt für das diskrete Variationsproblem (6.59) nach Trennung der Variablen

und anschließender Integration für $i \in I_{K,0}$, $j \in I_K$ und $n = 1, \ldots, N(\Delta t) - 1$

$$
\begin{aligned}
(c_{i,h}(t_{n+1}), \Psi_k)_L &= (c_{i,h}(t_n), \Psi_k)_L - \int_{t_n}^{t_{n+1}} a_i(c_{i,h}(s), c_{p,i,h}(s), \Psi_k)\, ds, \\
(c_{p,i,h}(t_{n+1}), \Psi_k)_L &= (c_{p,i,h}(t_n), \Psi_k)_L - \int_{t_n}^{t_{n+1}} a_{p,i}(c_{i,h}(s), c_{p,i,h}(s), q_{i,h}(s), \Psi_k)\, ds, \\
(q_{0,h}(t_{n+1}), \Psi_k)_L &= (q_{0,h}(t_n), \Psi_k)_L - \int_{t_n}^{t_{n+1}} b_0(q_{1,h}(t), \ldots, q_{K,h}(t), \Psi_k)\, ds, \\
(q_{j,h}(t_{n+1}), \Psi_k)_L &= (q_{j,h}(t_n), \Psi_k)_L \\
&\quad - \int_{t_n}^{t_{n+1}} b_j(c_{p,0,h}(t), c_{p,j,h}(t), q_{1,h}(t), \ldots, q_{K,h}(t), \Psi_k)\, ds,
\end{aligned}
$$

$\forall\, k = 1, \ldots, N(h)$.

Die Integrale werden (wie in Kapitel (6.59)) mit der Trapezregel numerisch berechnet, so dass man mit den Näherungen

$$c_{i,h}^n \approx c_{i,h}(t_n) \quad , c_{p,i,h}^n \approx c_{p,i,h}(t_n) \quad \text{und} \quad q_{i,h}^n \approx q_{i,h}(t_n)$$

die Variationsprobleme für $i \in I_{K,0}$ und $j \in I_K$

$$(c_{i,h}^{n+1}, \Psi_k)_L - \frac{\Delta t}{2} a_i(c_{i,h}^{n+1}, c_{p,i,h}^{n+1}, \Psi_k) = (c_{i,h}^n, \Psi_k)_L + \frac{\Delta t}{2} a_i(c_{i,h}^n, c_{p,i,h}^n, \Psi_k),$$

$$
\begin{aligned}
(c_{p,i,h}^{n+1}, \Psi_k)_L - \frac{\Delta t}{2} a_{p,i}(c_{i,h}^{n+1}, c_{p,i,h}^{n+1}, q_{i,h}^{n+1}, \Psi_k) &= \\
(c_{p,i,h}^n, \Psi_k)_L + \frac{\Delta t}{2} a_{p,i}(c_{i,h}^n, c_{p,i,h}^n, q_{i,h}^n, \Psi_k),&
\end{aligned}
$$

$$
\begin{aligned}
(q_{0,h}^{n+1}, \Psi_k)_L - \frac{\Delta t}{2} b_0(q_{1,h}^{n+1}(t), \ldots, q_{K,h}^{n+1}(t), \Psi_k) &= \\
(q_{0,h}^n, \Psi_k)_L + \frac{\Delta t}{2} b_0(q_{1,h}^n(t), \ldots, q_{K,h}^n(t), \Psi_k),&
\end{aligned}
$$

$$
\begin{aligned}
(q_{j,h}^{n+1}, \Psi_k)_L - \frac{\Delta t}{2} b_j(c_{p,0,h}^{n+1}(t), c_{p,j,h}^{n+1}(t), q_{1,h}^{n+1}(t), \ldots, q_{K,h}^{n+1}(t), \Psi_k) &= \\
(q_{j,h}^n, \Psi_k)_L + \frac{\Delta t}{2} b_j(c_{p,0,h}^n(t), c_{p,j,h}^n(t), q_{1,h}^n(t), \ldots, q_{K,h}^n(t), \Psi_k),&
\end{aligned}
$$

erhält. Beginnend mit $n = 1$ erhält man durch sukzessives Lösen dieser Variationsprobleme eine numerische Lösung eines präparativen Chromatographieprozesses, welcher durch die Gleichungen (6.14) mit (6.15) - (6.17) dargestellt wird.

6.5 Numerische Identifizierung des Proteins Lysozym bei pH 7

Das Lysozym ist ein Protein, welches in Sekreten des Atmungs- und Verdauungstraktes sowie in Tränen und Gewebsflüssigkeiten des Menschen vorkommt. Seine Aufgabe ist die Abwehr bakterieller Infektionen. Lysozym verursacht die Auflösung von Bakterienzellen indem es wichtige strukturelle Polysaccharid-Komponenten der Zellwände von

Bakterien abbaut. Es wird daher im medizinischen Bereich eingesetzt - zum Beispiel bei der Behandlung entzündlicher Erkrankungen im Mund und Rachenraum in Form von Lutschtabletten. Üblicherweise wird Lysozym für den kommerziellen Gebrauch aus Eiklar gewonnen, da es hier in besonders hoher Konzentration vorkommt und das am bestuntersuchte Lysozym darstellt [41].

Es ist bekannt, dass das aus Hühnereiweiß gewonnene Lysozym bei einem pH-Wert von 7 die SMA-Parameter

$$\boldsymbol{\theta}_{Lys} \;:=\; (k_a, k_d, \nu, \gamma)^\top \;=\; (5.0, 25.0, 3.29, 44.7)^\top \tag{6.61}$$

besitzt. Da diese Parameter bekannt sind, bietet sich das Beispiel eines Lysozyms bei pH 7 gut an um numerisch die in Kapitel 2.3 beschriebenen Verfahren zur Parameterschätzung zu untersuchen.

Wie in 2.2 beschrieben, werden bei einer Parameteridentifikation Messdaten benötigt. Diese erhält man, indem mehrfach experimentell ein Chromatographieprozess mit einem Lysozym bei pH 7 durchgeführt wird und am Ausgang der Säule die heraustretende Proteinkonzentration gemessen wird. Das Experiment zur Erhebung der Messdaten für die in dieser Arbeit geplanten numerischen Experimente wird in folgendem Unterkapitel beschrieben.

Anschließend wird untersucht, inwieweit man mit den in Kapitel 2.3 beschriebenen Verfahren die gesuchten vier SMA-Parameter aus (6.61) schätzen kann. In Vorbereitung dessen werden diese Verfahren zunächst am vereinfachten, zweidimensionalen Problem getestet, bei dem die Parameter $k_a = 5.0$ und $\gamma = 44.7$ als bekannt vorausgesetzt werden und lediglich k_d und ν geschätzt werden sollen.

6.5.1 Beschreibung des Experiments zur Bestimmung von Messdaten

Um die in Kapitel 2.2 beschriebene Parameteridentifikation zur Bestimmung der SMA-Parameter (6.61) durchführen zu können, werden Messdaten benötigt. Diese erhält man experimentell, indem am Beispiel einer Lysozymlösung die Säulenchromatographie durchgeführt und die Konzentration am Ausgang der Säule gemessen wird.

Für diese Arbeit wurde eine Säule der Länge $L_c = 25$ mm gewählt und mit einer stationären Phase bestehend aus kugelförmiger Cellulose mit einer Porengröße (6.4) von $\varepsilon_p = 0.9$ gefüllt. Hierbei entstand eine Packungsdichte (6.3) von $\varepsilon = 0.35$. Bevor das Experiment startet, wird die Chromatographiesäule zunächst mit einer Salzkonzentration so gefüllt, dass zum Zeitpunkt $t = 0$ in der mobilen Phase $c_0(0, x) = 0.02$ mol $\forall\, x \in [0, 25]$ gilt.
Das Experiment wird 1157 s lang wie folgt durchgeführt:

(1) In den ersten 2.4 s werden konstant 0.0002 mol Lysozym pro Sekunde in die Säule gedrückt, so dass $c_{in,1}(t) = \begin{cases} 0.0002, & t \in [0, 2.4], \\ 0, & t \in (2.4, 1157], \end{cases}$ gilt.

(2) In den ersten 233.3 s werden konstant 0.02 mol Salzkonzentration pro Sekunde hinzugefügt. Anschließend wird die Konzentration dieser Salzlösung linear bis auf

0.5 mol erhöht, wie in Abbildung 6.5 dargestellt. Es gilt

$$c_{in,0}(t) = \begin{cases} 0.02, & t \in [0, 233.3], \\ \dfrac{48}{11545}t - 0.95, & t \in (233.3, 348.75], \\ 0.5, & t \in (348.75, 1557]. \end{cases}$$

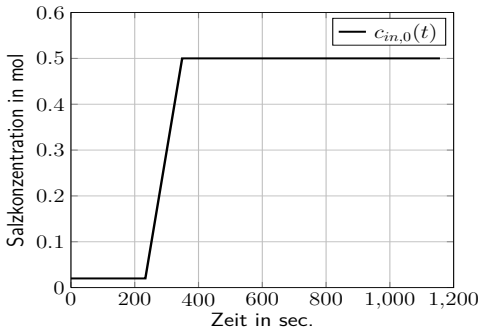


Abbildung 6.5: Die für das Experiment gewählte Salzkonzentration $c_{in,0}(t)$.

Für die in dieser Arbeit durchzuführenden numerischen Berechnungen werden synthetische Messdaten verwendet. In Vorbereitung dessen wurden zwei Messreihen numerisch generiert, indem zunächst für die exakten SMA-Parameter das zugehörige Chromatogramm bestimmt wurde und anschließend die Werte mit einem Fehler gestört wurden, für den eine Normalverteilung $\mathcal{N}(0, \sigma^2)$ mit der Standardabweichung $\sigma = 10^{-6}$ angenommen wurde.

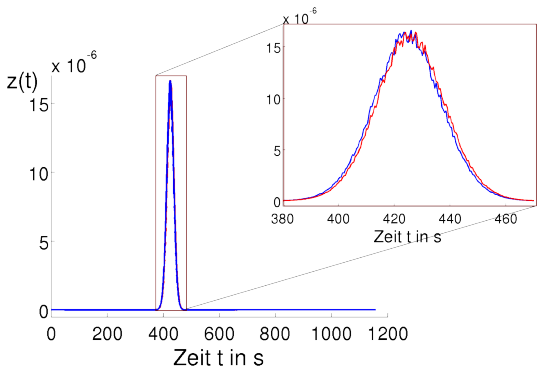


Abbildung 6.6: Die für die Parameteridentifikation verwendeten Messreihen, wobei der Messfehler als normalverteilt mit $\mathcal{N}(0, \sigma^2)$, $\sigma = 10^{-6}$ gewählt wurde.

6.5.2 Darstellung des Parametergebietes

Für die Parameterschätzung wird das Parametergebiet $\Theta := (1, 10) \times (20, 30) \times (2, 5) \times (30, 60)$ gewählt. In diesem Abschnitt wird zunächst untersucht, wie die einzelnen Parameter von Zielfunktional abhängen. Hierfür werden nacheinander jeweils zwei der vier gesuchten Parameter als bekannt vorausgesetzt.

Parametergebiet 1: $\nu^* := 3.29$ und $\gamma^* := 44.7$ fest gewählt

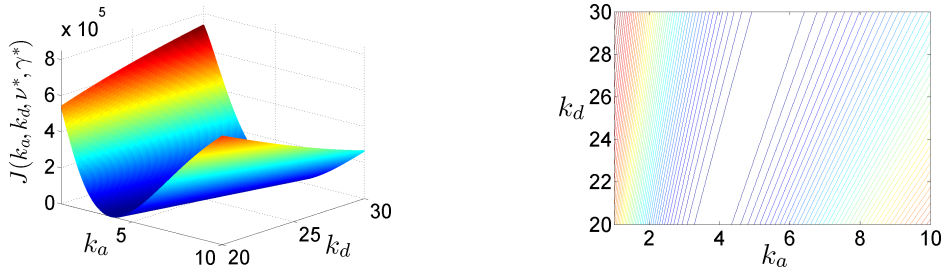


Abbildung 6.7: Das Zielfunktional $J(k_a, k_d, \nu^*, \gamma^*)$ bei fest gewähltem $\nu^* := 3.29$ und $\gamma^* := 44.7$ mit zugehörigem Konturdiagramm.

Parametergebiet 2: $\mathbf{k_d^*} := 25.0$ und $\gamma^* := 44.7$ fest gewählt

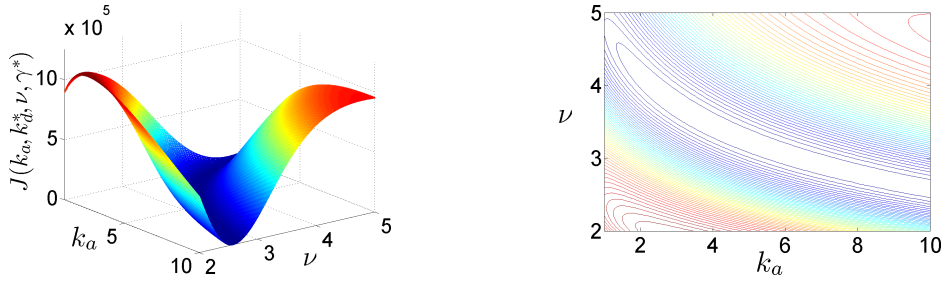


Abbildung 6.8: Das Zielfunktional $J(k_a, k_d^*, \nu, \gamma^*)$ bei fest gewähltem $k_d^* := 25.0$ und $\gamma^* := 44.7$ mit zugehörigem Konturdiagramm.

Parametergebiet 3: $\mathbf{k_a^*} := 5.0$ und $\gamma^* := 44.7$ fest gewählt

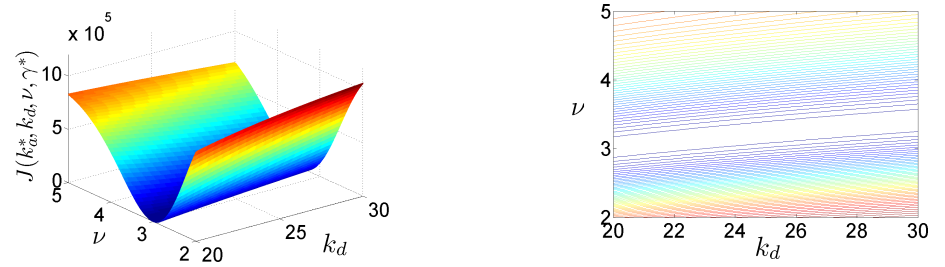


Abbildung 6.9: Das Zielfunktional $J(k_a^*, k_d, \nu, \gamma^*)$ bei fest gewähltem $k_a^* := 5.0$ und $\gamma^* := 44.7$ mit zugehörigem Konturdiagramm.

Parametergebiet 4: $\mathbf{k_a^* := 5.0}$ und $\mathbf{k_d^* := 25.0}$ fest gewählt

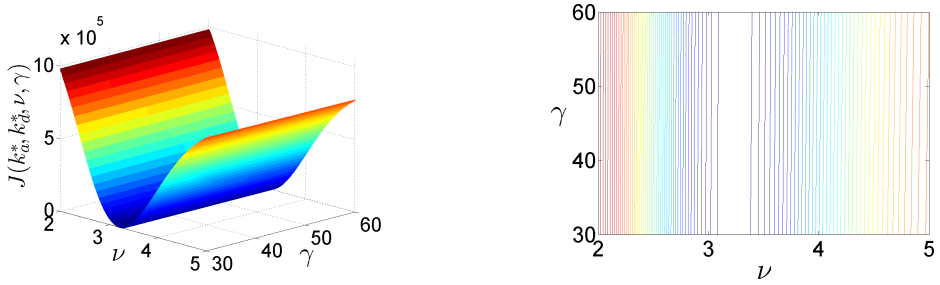


Abbildung 6.10: Das Zielfunktional $J(k_a^*, k_d^*, \nu, \gamma)$ bei fest gewähltem $k_a^* := 5.0$ und $k_d^* := 25.0$ mit zugehörigem Konturdiagramm.

Parametergebiet 5: $\mathbf{k_a^* := 5.0}$ und $\boldsymbol{\nu^* := 3.29}$ fest gewählt

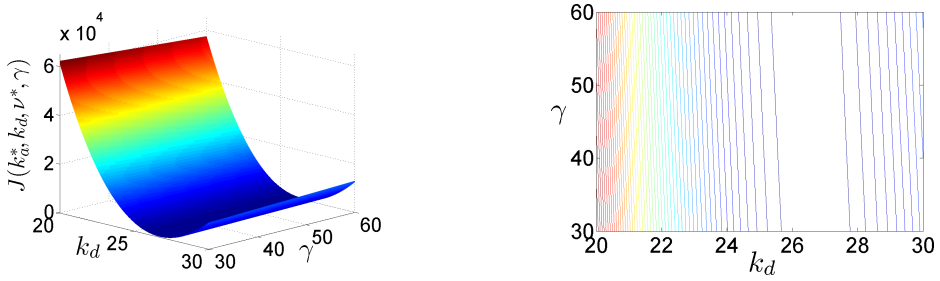


Abbildung 6.11: Das Zielfunktional $J(k_a^*, k_d, \nu^*, \gamma)$ bei fest gewähltem $k_a^* := 5.0$ und $\nu^* := 3.29$ mit zugehörigem Konturdiagramm.

Parametergebiet 6: $\mathbf{k_d^* := 25.0}$ und $\boldsymbol{\nu^* := 3.29}$ fest gewählt

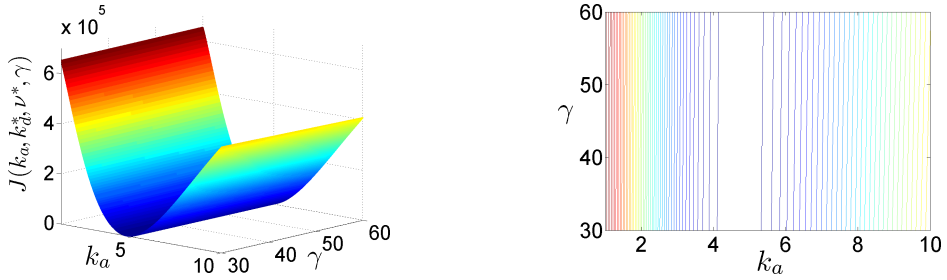


Abbildung 6.12: Das Zielfunktional $J(k_a, k_d^*, \nu^*, \gamma)$ bei fest gewähltem $k_d^* := 25.0$ und $\nu^* := 3.29$ mit zugehörigem Konturdiagramm.

6.5.3 Numerische Bestimmung des Desorbtionskoeffizienten und der Charakteristischen Ladung

In diesem Abschnitt wird an einem vereinfachten, zweidimensionalen Schätzproblem untersucht, inwieweit man

- den Desorbtionskoeffizienten $k_d \in [20, 30]$ und
- den Schirmungskoeffizienten $\nu \in [2, 5]$

aus (6.61) mit den in Kapitel 2.3 beschriebenen Verfahren schätzen kann, wenn der Adsorbtionskoeffizient $k_a^* := 5.0$ und die Charakteristische Ladung $\nu^* := 3.29$ bereits bekannt sind. Es werden folgende Verfahren umgesetzt:

(V1) Methode des steilsten Abstiegs mit Armijo-Schrittweitensteuerung aus Kapitel 2.3.1.

(V2) Gedämpftes Newton-Verfahren mit Armijo-Schrittweitensteuerung (Kapitel 2.3.2).

(V3) Ein hybrides Verfahren aus (V1) und (V2). Bei diesem Verfahren wird solange das Verfahren (V1) verwendet, bis die Hessematrix $\nabla^2_{\boldsymbol{\theta}} J(\boldsymbol{\theta}^n)$ positiv definit ist. Ausgehend von $\boldsymbol{\theta}^n$ wird dann Verfahren (V2) angewendet.

(V4) Gedämpftes Gauss-Newton-Verfahren mit Armijo-Schrittweitensteuerung aus Kapitel 2.3.3.

(V5) Ein hybrides Verfahren aus (V1) und (V4). Da bei Parameteridentifikationen das Verfahren (V1) lokal sehr viele Iterationsschritte benötigt, wird das Verfahren (V1) solange verwendet, bis $|\theta_i^n - \theta_i^{n-1}| < \varepsilon$, $\forall\, i$. Ausgehend von $\boldsymbol{\theta}^n$ wird dann Verfahren (V4) angewendet.

Als Startwert wird

$$(k_d^0, \nu^0) \;:=\; (27.0, 2.5)$$

mit dem Zielfunktionswert

$$J(k_d^0, \nu^0) \;=\; 525133.71$$

gewählt.

Die Verfahren (V1) - (V5) benötigen pro Iterationsschritt jeweils eine Lösung des Zustandssystems (6.27) - (6.28) und eine Lösung der Sensitivitätsgleichungen erster Ordnung (6.52) - (6.53). Die Verfahren (V2) - (V3) benötigen zudem pro Iteration eine Lösung der Sensitivitätsgleichungen zweiter Ordnung (6.54) - (6.55). Wie in Kapitel 6.4.1 und 6.4.2 beschrieben, werden diese Gleichungen mit der Methode der finiten Elemente [12] und dem Crank-Nicolson-Verfahren [76] gelöst. Hierfür werden die Schrittweiten $h := 0.05$ und $\Delta t := 1$ gewählt, so dass im Raum $N(h) = 501$ und in der Zeit $N(\Delta t) = 1157$ Freiheitsgrade vorhanden sind.

Das globale Minimum liegt bei

$$(k_d^*, \nu^*) \;=\; (21.1762, 3.0733)$$

mit dem Zielfunktionswert

$$J(k_d^*, \nu^*) \;=\; 298.57.$$

Verfahren	# Iterationen	durchn. Zeit pro Iterationsschritt	Zeit gesamt
(V1)	> 9000	275 s	> 1 Monat
(V2)	–	–	–
(V3)	$10 + 11$	238 s bzw. 213 s	39.71 min + 39.11 min
(V4)	–	–	–
(V5)	$10 + 11$	180 s bzw. 160 s	29.96 min + 26.68 min

Tabelle 6.2: Anzahl der Iterationen und durchschnittlicher Zeitaufwand pro Iterationsschritt bei Verwendung der Verfahren (V1) - (V5).

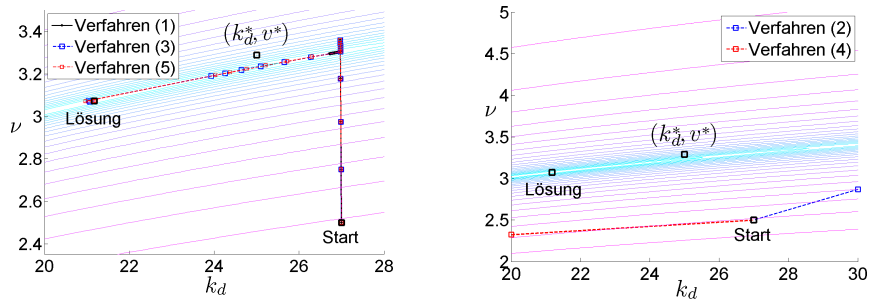


Abbildung 6.13: Lösen eines Parameteridentifikationsproblems mit Startwert $(k_d, \nu) = (21.0, 2.5)$ mit den Verfahren (1) - (5).

6.5.4 Numerische Schätzung der vier SMA-Parameter

Nachdem im vorherigen Abschnitt die Verfahren (V1) - (V5) zur numerischen Schätzung des Desorbtionskoeffizienten $k_d \in [20, 30]$ und der Charakteristischen Ladung $\nu \in [2, 5]$ eines Proteins Lysozym bei pH 7 verwendet wurden, wird in diesem Abschnitt überprüft, inwieweit alle vier SMA-Parameter gleichzeitig geschätzt werden können. Für die Parameteridentifikation werden zunächst die Messreihen verwendet, die für die Parameterschätzung in Kapitel 6.5.3 genutzt wurden. Diese Daten wurden am Ausgang der Chromatographiesäule erhoben.

Als Startwert wird

$$(k_a^0, k_d^0, \nu^0, \gamma^0) \;:=\; (4.5, 23.0, 3.7, 41)$$

gewählt, wobei für den zugehörigen Zielfunktionswert

$$J(k_a^0, k_d^0, \nu^0, \gamma^0) \;=\; 129230.02$$

gilt. Wie in Kapitel 6.5.3 beschrieben, wird für die Realisierung der Verfahren (V1) - (V5) pro Iterationsschritt jeweils eine Lösung des Zustandssystems (6.27) - (6.28) und für jeden Parameter jeweils eine Lösung der Sensitivitätsgleichungen erster Ordnung (6.52) - (6.53) benötigt. Für die Verfahren (V2) - (V3) sind zudem pro Iteration die Lösungen der Sensitivitätsgleichungen zweiter Ordnung (6.54) - (6.55) erforderlich. Die Anwendung der Verfahren (V1) - (V5) ergab, dass sich das globale Minimum bei

$$(k_a^*, k_d^*, \nu^*, \gamma^*) \;=\; (4.6185, 23.0473, 3.2506, 100.0)$$

befindet und den Zielfunktionswert

$$J(k_a^*, k_d^*, \nu^*, \gamma^*) \;=\; 290.90$$

besitzt.

Verfahren	# Iterationen	durchn. Zeit pro Iterationsschritt	Zeit gesamt
(V1)	> 9000	300 s	> 1 Monat
(V2)	–	–	–
(V3)	$7 + 7$	265 s bzw. 278 s	30.94 min + 32.48 min
(V4)	–	–	–
(V5)	$7 + 17$	114 s bzw. 128 s	13.33 min + 36.18 min

Tabelle 6.3: Anzahl der Iterationen und durchschnittlicher Zeitaufwand pro Iterationsschritt bei Verwendung der Verfahren (V1) - (V5).

Wie in Abschnitt 6.5.3 erkennt man hier sehr deutlich, dass die Verfahren (V2) und (V4) für die Schätzung der SMA-Parameter ungeeignet sind. Bei der Verwendung des gedämpften Newton-Verfahrens (V2) muss beachtet werden, dass dieses Verfahren gegen ein lokales Minimum, lokales Maximum oder einen Sattelpunkt konvergieren kann. Nur wenn sich die Startiterierte in einer lokalen Umgebung des gesuchten Minimums befindet kann die Konvergenz von (V2) erwartet werden. Dass das gedämpfte Gauss-Newton-Verfahren (V4) das gesuchte Minimum nicht ermittelt hat, liegt ebenfalls daran, dass sich die Startiterierte nicht nah genug an der optimalen Lösung befindet. Da beim Verfahren (V4) quasi das Newtonverfahren (V2) umgesetzt wird, bei dem die Hessematrix durch die Fisher-Informationsmatrix $M(\vec{\mathbf{x}}, \mathbf{p})$ approximiert wird, kann auch keine Konvergenz gegen die optimale Lösung erwartet werden, wenn (V2) dieses ebenfalls nicht ermitteln kann.

Die Methode des steilsten Abstiegs (V1) konvergiert zwar gegen die gesuchte optimale Lösung, dieses allerdings sehr langsam: Befinden sich die Iterierten in einer lokalen Umgebung der gesuchten optimalen Lösung, dann beträgt der Winkel zwischen dem negativen Gradienten des Zielfunktionals und einer optimalen Abstiegsrichtung nahezu 90°, so dass in jedem Iterationsschritt die Schrittweitensteuerung sehr klein gewählt werden muss um eine Minimierung im Zielfunktional zu bewirken. Daher dauert es sehr lange, bis mit dem Verfahren (V1) eine Lösung ermittelt wird. Da allerdings (V1) sich sehr gut dafür eignet in wenigen Schritten in einer lokalen Umgebung der gesuchten Minimallösung zu gelangen, haben sich (V3) und (V5) als geeignete Verfahren erwiesen um das Parameteridentifikationsproblem zu lösen. Da (V5) im Gegensatz zu (V3) keine Lösung der Sensitivitätsgleichungen zweiter Ordnung (6.54) - (6.55) benötigt, ist das Verfahren (V5) dem Verfahren (V3) vorzuziehen.

Die optimale Lösung

$$\boldsymbol{\theta}^* \;=\; (k_a^*, k_d^*, \nu^*, \gamma^*)^\top \;=\; (4.6185, 23.0473, 3.2506, 100.0)^\top$$

ist allerdings keine gute Schätzung des exakten Modellparameters

$$\boldsymbol{\theta}_{Lys} \;:=\; (5.0, 25.0, 3.29, 44.7)^\top$$

aus (6.61). Für den relativen Fehler gilt

$$R(\boldsymbol{\theta}^*) \;:=\; \frac{\|\boldsymbol{\theta}_{Lys} - \boldsymbol{\theta}^*\|_2}{\|\boldsymbol{\theta}_{Lys}\|_2} \;=\; 1.0731.$$

Insbesondere konnte der Schirmungskoeffizient $\gamma \in [10, 100]$ eines Lysozyms nicht geschätzt werden. Im folgenden Unterkapitel wird daher durch Bestimmung eines D-optimalen Designs überprüft, ob die schlechte Schätzung aus einer falschen Wahl der Messposition resultiert. Bei der numerischen Ermittlung eines D-optimalen Designs wird eine Messstellenkonstellation ermittelt, für die die Determinante der Kovarianzmatrix eines Schätzers minimal ist. Auf diese Weise kann ein unbekannter Parameter bestmöglich geschätzt werden.

Nachdem im folgenden Abschnitt die D-optimalen Messpositionen $\vec{\mathbf{x}} \in \bar{\Omega}^\ell$ ermittelt werden, wird im Anschluss erneut eine Parameterschätzung mithilfe der an $\vec{\mathbf{x}}$ erhobenen Messdaten durchgeführt.

6.6 Bestimmung eines D-optimalen Designs

Da im allgemeinen am Ausgang einer Chromatographiesäule gemessen wird um die für die Parameteridentifikation erforderlichen Messdaten zu erheben, wird in diesem Abschnitt am Beispiel des Proteins Lysozym bei pH 7 untersucht, inwieweit diese Messposition für die Schätzung der vier SMA-Parameter (6.61) geeignet ist. Wie in Kapitel 3 beschrieben, liefert die Fisher-Informationsmatrix

$$M(\vec{\mathbf{x}}, \mathbf{p}) \;=\; \frac{1}{\sigma^2} \sum_{i=1}^{\ell} p_i \int_T \nabla_{\boldsymbol{\theta}} u(t, x_i, \hat{\boldsymbol{\theta}}^1) \nabla_{\boldsymbol{\theta}} u(t, x_i, \hat{\boldsymbol{\theta}}^1)^\top dt, \tag{6.62}$$

mit den Messpositionen $\vec{\mathbf{x}} = (x_1, \ldots, x_\ell) \in \bar{\Omega}^\ell$ und den zugehörigen Gewichten $\mathbf{p} \in [0,1]^\ell$ ein Maß für die Genauigkeit eines Schätzers, da $M(\vec{\mathbf{x}}, \mathbf{p})$ die Inverse der Kovarianzmatrix des Schätzers $\hat{\boldsymbol{\theta}}$ approximiert, falls $\hat{\boldsymbol{\theta}}^1 \approx \boldsymbol{\theta}_{Lys}$ mit $\boldsymbol{\theta}_{Lys}$ aus (6.61) gilt und somit die Voraussetzung 3.1.1 erfüllt ist. Da die exakten Modellparameter $\boldsymbol{\theta}_{Lys}$ bekannt sind, wird in diesem Kapitel $\hat{\boldsymbol{\theta}}^1 := \boldsymbol{\theta}_{Lys}$ gewählt.

Im folgenden Abschnitt werden zunächst die für die Berechnung der Fisher-Informationsmatrix $M(\vec{\mathbf{x}}, \mathbf{p})$ erforderlichen Sensitivitäten erster Ordnung

$$\nabla_\theta u(t, x, \boldsymbol{\theta}_{Lys}) \;\in\; L^2(T; H_0^1(\Omega))$$

numerisch ermittelt und abgebildet. Auf Basis dieser Sensitivitäten wird anschließend die Fisher-Information $\Psi_i(M(\vec{\mathbf{x}}, \mathbf{p}))$ bezüglich der drei Optimalitätskriterien

(1) D-Optimalitätskriterium (Determinante) mit (6.63)
$$\Psi_1(M(\vec{\mathbf{x}}, \mathbf{p})) \;=\; -\ln\det(M(\vec{\mathbf{x}}, \mathbf{p})) \;:=\; -\ln|M(\vec{\mathbf{x}}, \mathbf{p})|,$$

(2) E-Optimalitätskriterium (größter Eigenwert von $\mathrm{cov}\{\hat{\boldsymbol{\theta}}\}$) mit (6.64)
$$\Psi_2(M(\vec{\mathbf{x}}, \mathbf{p})) \;=\; \lambda_{\max}(M(\vec{\mathbf{x}}, \mathbf{p})^{-1}),$$

(3) A-Optimalitätskriterium (Spur) mit (6.65)
$$\Psi_3(M(\vec{\mathbf{x}}, \mathbf{p})) \;=\; \mathrm{tr}(M(\vec{\mathbf{x}}, \mathbf{p})^{-1}),$$

berechnet, falls lediglich am Ausgang der Säule gemessen wird, d. h. $(\vec{\mathbf{x}}, \mathbf{p}) = (L_c, 1)$. Danach wird mit der in Kapitel (4.4) beschriebenen zweistufigen Active-Set-Methode ein D-optimales Design $(\vec{\mathbf{x}}^*, \mathbf{p}^*)$ numerisch ermittelt, welches (6.63) minimiert um eine D-optimale Messstellenkonstellation zu erhalten. An diesen Messstellen werden erneut Messdaten generiert um im Anschluss wiederum eine Parameteridentifikation durchzuführen.

6.6.1 Die Sensitivitäten erster Ordnung beim Lysozym bei pH 7

Um die Fisher-Informationsmatrix $M(\vec{\mathbf{x}}, \mathbf{p})$ berechnen zu können, werden die Sensitivitäten erster Ordnung $\nabla_\theta u(t, x_i, \boldsymbol{\theta}_{Lys})$ benötigt. Diese erhält man durch Lösen der Sensitivitätsgleichungen erster Ordnung (6.52) mit den Anfangsbedingungen (6.53).

Die folgenden Abbildungen zeigen die Sensitivitäten erster Ordnung:

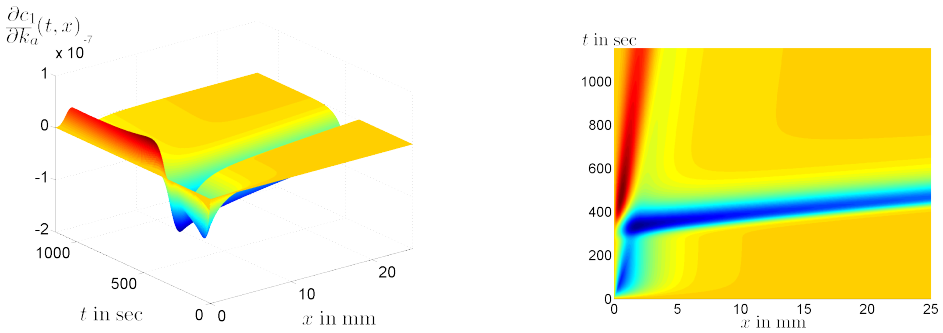


Abbildung 6.14: Die Sensitivität $\partial_{k_a} c_1(t, x, \boldsymbol{\theta}_{Lys})$.

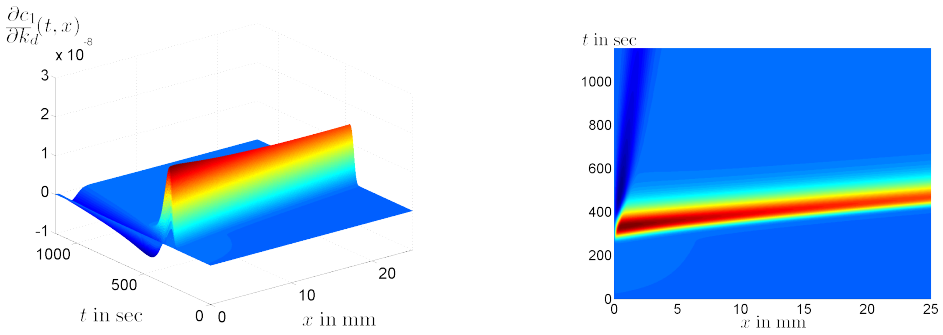


Abbildung 6.15: Die Sensitivität $\partial_{k_d} c_1(t, x, \boldsymbol{\theta}_{Lys})$.

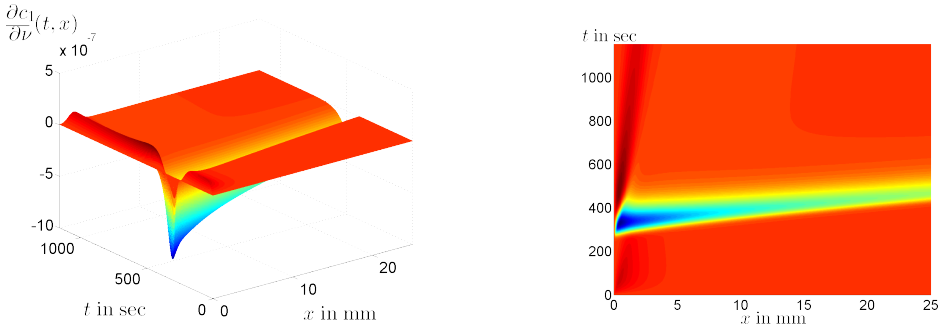


Abbildung 6.16: Die Sensitivität $\partial_{\nu} c_1(t, x, \boldsymbol{\theta}_{Lys})$.

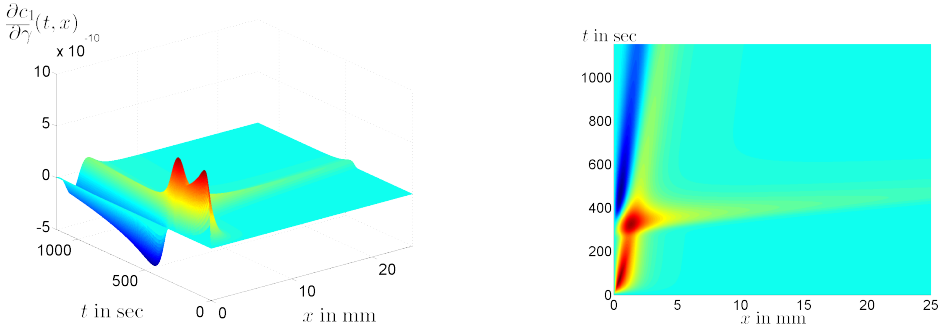


Abbildung 6.17: Die Sensitivität $\partial_{\gamma} c_1(t, x, \boldsymbol{\theta}_{Lys})$.

Mithilfe dieser Sensitivitäten können nun die Fisher-Informationen bezüglich (6.63) - (6.65) berechnet werden, wenn

(1) lediglich am Ausgang einer Chromatographiesäule gemessen wird oder

(2) an den Designpunkten eines D-optimalen Designs die zur Parameterschätzung erforderlichen Messdaten erhoben werden.

Auf diese Weise kann die Qualität eines Maximum-Likelihood-Schätzers $\hat{\boldsymbol{\theta}}$ quantifiziert und für die Fälle (1) und (2) miteinander verglichen werden.

6.6.2 Die Fisher-Information beim Messen am Säulenausgang

Misst man, wie in Kapitel 6.5.1 beschrieben, am Beispiel eines Lysozyms bei pH 7 lediglich am Ausgang einer Chromatographiesäule, dann hat das zugehörige Design mit $x := L_c = 25$ und $p := 1$ die Darstellung

$$\boldsymbol{\xi}_{L_c} := \left\{ \begin{matrix} x \\ p \end{matrix} \right\} = \left\{ \begin{matrix} 25 \\ 1 \end{matrix} \right\}.$$

Mit den in Kapitel 6.6.1 ermittelten Sensitivitäten erster Ordnung wird zunächst die Fisher-Informationsmatrix $M(x,p) \in \mathbb{R}^{4\times 4}$ berechnet: Es ist

$$\begin{aligned} M(x,p) &= \int_T \nabla_{\boldsymbol{\theta}} u(t, L_c, \boldsymbol{\theta}_{Lys}) \nabla_{\boldsymbol{\theta}} u(t, L_c, \boldsymbol{\theta}_{Lys})^\top dt \\ &= \begin{pmatrix} 3.83 \cdot 10^{-1} & -7.58 \cdot 10^{-2} & 1.51 & -2.53 \cdot 10^{-4} \\ -7.58 \cdot 10^{-2} & 1.50 \cdot 10^{-2} & -2.98 \cdot 10^{-1} & 5.00 \cdot 10^{-5} \\ 1.51 & -2.98 \cdot 10^{-1} & 5.92 & -9.95 \cdot 10^{-4} \\ -2.53 \cdot 10^{-4} & 5.00 \cdot 10^{-5} & -9.95 \cdot 10^{-4} & 1.67 \cdot 10^{-7} \end{pmatrix}, \end{aligned}$$

mit den Eigenwerten

$$\lambda_1 = 4.73 \cdot 10^{-5}, \quad \lambda_2 = 5.16 \cdot 10^{-6}, \quad \lambda_3 = 6.32, \quad \lambda_4 = 2.51 \cdot 10^{-11}.$$

Da die Voraussetzung 3.1.1 erfüllt ist, kann die Kovarianzmatrix des Maximum-Likelihood-Schätzers $\hat{\boldsymbol{\theta}}$ durch die Inverse der Fisher-Informationsmatrix approximiert werden. Somit gilt

$$\mathrm{cov}(\hat{\boldsymbol{\theta}}) \approx M(x,p)^{-1} = \begin{pmatrix} 1.96 \cdot 10^{5} & 2.75 \cdot 10^{4} & -6.13 \cdot 10^{4} & -7.68 \cdot 10^{7} \\ 2.75 \cdot 10^{4} & 1.74 \cdot 10^{5} & 4.89 \cdot 10^{3} & 1.86 \cdot 10^{7} \\ -6.13 \cdot 10^{4} & 4.89 \cdot 10^{3} & 2.04 \cdot 10^{4} & 2.72 \cdot 10^{7} \\ -7.68 \cdot 10^{7} & 1.86 \cdot 10^{7} & 2.72 \cdot 10^{7} & 3.98 \cdot 10^{10} \end{pmatrix},$$

mit den Eigenwerten

$$\kappa_1 = 2.12 \cdot 10^{4}, \quad \kappa_2 = 1.94 \cdot 10^{5}, \quad \kappa_3 = 1.58 \cdot 10^{-1}, \quad \kappa_4 = 3.98 \cdot 10^{10}.$$

Wie in Kapitel 2.1.2 beschrieben, liefert die Kovarianzmatrix $\mathrm{cov}(\hat{\boldsymbol{\theta}})$ ein Maß für die Genauigkeit eines Schätzers $\hat{\boldsymbol{\theta}}$:

- Je kleiner die Determinante von cov($\hat{\boldsymbol{\theta}}$), desto kleiner ist das Volumen des zugehörigen Konfidenzellipsoids $\mathcal{K}$.
- Je kleiner der größte Eigenwert von cov($\hat{\boldsymbol{\theta}}$), desto kleiner ist die längste Hauptachse von $\mathcal{K}$.
- Je kleiner die Spur von cov($\hat{\boldsymbol{\theta}}$), desto kleiner ist der Durchschnitt aller Hauptachsen von $\mathcal{K}$

Für das D-, E- und A-Optimalitätskriterium aus (6.63) - (6.65) ergeben sich folgende Fisher-Informationen:

$\Psi_1(M(x,p))$	$\Psi_2(M(x,p))$	$\Psi_3(M(x,p))$
44.69805	$3.98091 \cdot 10^{10}$	$3.98093 \cdot 10^{10}$

Tabelle 6.4: Die Fisher-Information für das D-, E- und A-Optimalitätskriterium aus (6.63) - (6.65) des Einpunkt-Designs $\boldsymbol{\xi}_{L_c}$.

Da der zum Schirmungskoeffizienten $\gamma \in [10, 100]$ zugehörige Eigenwert $\kappa_4 = 3.98 \cdot 10^{10}$ sehr groß ist und der zu κ_4 gehörende Eigenvektor

$$\mathbf{v}^4 = \begin{pmatrix} -1.92 \cdot 10^{-3} \\ 4.68 \cdot 10^{-4} \\ 6.83 \cdot 10^{-4} \\ 0.999998 \end{pmatrix}$$

nahezu den Einsvektor darstellt, ist folglich das Konfidenzintervall bezüglich γ ebenfalls sehr groß, so dass es nicht weiter verwunderlich ist, dass der SMA-Parameter γ mit der im vorherigen Unterkapitel durchgeführten Parameteridentifikation nicht geschätzt werden konnte.

Bemerkung 6.6.1. *Wie in Kapitel 2.1.2 erläutert, definiert der Eigenvektor $\mathbf{v}^i$ die Richtung der i-ten Hauptachse des Konfidenzellipsoids. Im Anhang B.4.1 sind die Eigenvektoren $\mathbf{v}^i$ für $i = 1, \ldots, 4$ aufgelistet.*

In Abbildung 6.17 wurde die Sensitivität $\partial_\gamma c_1(t, x, \boldsymbol{\theta}_{Lys})$ dargestellt. Man erkennt anhand dieser Darstellung sehr deutlich, dass die zur Schätzung des SMA-Parameters γ benötigten Messdaten nicht am Ausgang der Chromatographiesäule erhoben werden sollten, da hier die Sensitivität nahezu Null ist. Es wäre sinnvoller an den Positionen zu messen, an denen die Sensitivität maximal ist. Solche Positionen können durch Bestimmung eines D-optimalen Designs ermittelt werden.

Im folgenden Abschnitt wird mit der in Kapitel 4.4 beschriebenen Active-Set-Methode ein D-optimales Design $(\vec{\mathbf{x}}^*, \mathbf{p}^*)$ numerisch ermittelt, für welches $\Psi_1(M(\mathbf{x}, \mathbf{p}))$ minimal ist. Auf Basis dieses Designs werden dann die zugehörigen Fisher-Informationen berechnet und mit den Werten aus Tabelle 6.4 verglichen.

6.6.3 Die Fisher-Information eines D-optimalen Designs

In Kapitel 4.4 wurde eine zweistufige Active-Set-Methode beschrieben, mit der ein D-optimales Design ermittelt werden kann. Mit dieser Methode wird in diesem Abschnitt

ein solches Design $(\vec{\mathbf{x}}^*, \mathbf{p}^*) \in \bar{\Omega}^\ell \times [0,1]^\ell$ ermittelt, für das die zugehörige Fisher-Informationsmatrix minimal bezüglich des D-Optimalitätskriteriums (6.63) ist. Für die numerische Berechnung der Sensitivitäten erster Ordnung werden die Schrittweiten

$$h \;:=\; 1.25 \cdot 10^{-3} \quad \text{und} \quad \Delta t \;:=\; 1$$

gewählt. Nach Umsetzung der Active-Set-Methode mit Algorithmus 4.4.2 erhält man auf diese Weise das D-optimale Design

$$\boldsymbol{\xi}^* \;:=\; \left\{ \begin{array}{c} \vec{\mathbf{x}}^{*\top} \\ \mathbf{p}^{*\top} \end{array} \right\} \;=\; \left\{ \begin{array}{ccc} 0.65 & 2.15 & 25 \\ 0.4284 & 0.3476 & 0.2239 \end{array} \right\} \tag{6.66}$$

mit der Fisher-Informationsmatrix

$$M(\vec{\mathbf{x}}^*, \mathbf{p}^*) \;=\; \begin{pmatrix} 1.497 & -0.140 & 3.926 & -9.29 \cdot 10^{-3} \\ -0.140 & 0.029 & -1.014 & 7.94 \cdot 10^{-4} \\ 3.926 & -1.014 & 37.436 & -0.023 \\ -9.29 \cdot 10^{-3} & 7.94 \cdot 10^{-4} & -0.023 & 6.52 \cdot 10^{-5} \end{pmatrix}.$$

Die Eigenwerte von $M(\vec{\mathbf{x}}^*, \mathbf{p}^*)$ sind

$$\lambda_1 = 1.07, \quad \lambda_2 = 5.68 \cdot 10^{-4}, \quad \lambda_3 = 37.89, \quad \lambda_4 = 3.37 \cdot 10^{-6}.$$

Da die Fisher-Informationsmatrix die Inverse der Kovarianzmatrix approximiert gilt für die Kovarianzmatrix des Schätzers $\hat{\boldsymbol{\theta}}$, dass

$$\mathrm{cov}(\hat{\boldsymbol{\theta}}) \;\approx\; M(x,p)^{-1} \;=\; \begin{pmatrix} 26.35 & 280.85 & 6.48 & 2641.26 \\ 280.85 & 3918.18 & 92.22 & 25140.56 \\ 6.48 & 92.22 & 2.21 & 586.17 \\ 2641.26 & 25140.56 & 586.17 & 294256.87 \end{pmatrix},$$

mit den Eigenwerten

$$\kappa_1 = 9.31 \cdot 10^{-1}, \quad \kappa_2 = 1.76 \cdot 10^{3}, \quad \kappa_3 = 2.64 \cdot 10^{-2}, \quad \kappa_4 = 2.96 \cdot 10^{5}.$$

Mit κ_i für $i = 1, \ldots, 4$ können nun wiederum die Fisher-Informationen bezüglich des D-, E- und A-Optimalitätskriteriums berechnet werden. Bei Verwendung des D-optimalen Designs aus (6.66) gilt:

$\Psi_1(M(\vec{\mathbf{x}}^*, \mathbf{p}^*))$	$\Psi_2(M(\vec{\mathbf{x}}^*, \mathbf{p}^*))$	$\Psi_3(M(\vec{\mathbf{x}}^*, \mathbf{p}^*))$
16.36638	$2.96443 \cdot 10^5$	$2.98204 \cdot 10^5$

Tabelle 6.5: Die Fisher-Information für das D-, E- und A-Optimalitätskriterium aus (6.63) - (6.65) bei Verwendung des D-optimalen Designs $\boldsymbol{\xi}^*$.

Man sieht sehr deutlich, dass bei der Verwendung des D-optimalen Designs $\boldsymbol{\xi}^*$ alle vier Eigenwerte um mindestens eine Größenordnung kleiner sind als die Eigenwerte der Kovarianzmatrix aus Kapitel 6.6.2. Was das E- und das A-Optimalitätskriterium anbelangt verringert sich die Fisher-Information in beiden Fällen sogar in etwa um den Faktor

10^5, wenn man nicht nur am Säulenausgang misst, sondern auch an den Designpunkten x_1^* und x_2^*. Die Fisher-Information beim D-Optimalitätskriterium (6.63) reduziert sich von $\Psi_1(M(x,p)) = 44.69805$ in Kapitel 6.6.2 auf $\Psi_1(M(\vec{\mathbf{x}}^*, \mathbf{p}^*)) = 16.36638$ in diesem Kapitel. Eine solche Minimierung sieht im ersten Moment nicht sonderlich spektakulär aus. An dieser Stelle muss allerdings beachtet werden, dass die Anwendung des Logarithmus in Ψ_1 das eigentliche Resultat verfälscht. Betrachtet man nämlich die Determinanten der jeweiligen Kovarianzmatrizen erhält man folgende Werte. Es ist

$$|\mathrm{cov}(\hat{\boldsymbol{\theta}})| \;=\; 2.58295 \cdot 10^{19},$$

falls das Einpunkt-Design $\boldsymbol{\xi}_{L_c}$ und

$$|\mathrm{cov}(\hat{\boldsymbol{\theta}})| \;=\; 1.28182 \cdot 10^{7},$$

falls das D-optimale Design $\boldsymbol{\xi}^*$ verwendet wird. Somit wurde die Determinante der Kovarianzmatrix in etwa um den Faktor 10^{12} minimiert.

Da in Kapitel 6.5.4 die SMA-Parameter k_a, k_d und γ nicht gut und der Parameter γ gar nicht geschätzt werden konnte, wird im folgenden Abschnitt erneut eine Parameteridentifikation mithilfe eines Maximum-Likelihood-Schätzers durchgeführt, wobei nun zu den bereits genutzten Messdaten weitere Messreihen hinzugenommen werden, die an den Positionen x_1^* und x_2^* des D-optimalen Designs aus (6.66) erhoben wurden. Da für diese Arbeit keine realen Messdaten zur Verfügung stehen, werden, wie in Kapitel 6.5.1 beschrieben, synthetische Daten verwendet.

6.7 Parameterschätzung auf Basis des D-optimalen Designs

Nachdem in Kapitel 6.6.3 das D-optimale Design

$$\boldsymbol{\xi}^* \;:=\; \left\{ \begin{array}{c} \vec{\mathbf{x}}^{*\top} \\ \mathbf{p}^{*\top} \end{array} \right\} \;=\; \left\{ \begin{array}{ccc} 0.65 & 2.15 & 25 \\ 0.4284 & 0.3476 & 0.2239 \end{array} \right\}$$

numerisch ermittelt wurde, welches die Fisher-Informationsmatrix bezüglich des D-Optimalitätskriteriums (6.63) minimiert, wird in diesem Abschnitt auf Basis dieses Designs erneut eine Schätzung der vier SMA-Parameter des Proteins Lysozym bei pH 7 durchgeführt. Für die Messposition $x_3^* = 25$ werden die selben beiden Messreihen verwendet, wie bei der Parameterschätzung in Kapitel 6.5.4. Für die Designpunkte $x_1^* = 0.65$ und $x_2^* = 2.15$ werden, wie in Kapitel 6.5.1 beschrieben, jeweils zwei Messreihen generiert. Diese werden gemittelt und für die i-te Messposition x_i^* durch $\tilde{z}_i(t)$ dargestellt.

Die SMA-Parameter können nun durch Lösen des Optimierungsproblems

$$\boldsymbol{\theta}^* \;=\; \arg\min_{\Theta} J(\boldsymbol{\theta}) \;=\; \frac{1}{2\sigma^2} \sum_{i=1}^{3} p_i^* \int_T \|\tilde{z}_i(t) - c_1(t, x_i^*, \boldsymbol{\theta})\|_2^2 dt \qquad (6.67)$$

geschätzt werden, wobei $c_1(t, x_i^*, \boldsymbol{\theta})$ die Proteinkonzentration des Lysozyms an der Messposition x_i^* darstellt. Die Konzentration $c_1(t, x_i^*, \boldsymbol{\theta})$ kann für jedes fest gewählte $\boldsymbol{\theta}$ durch numerisches Lösen des Konvektions-Diffusions-Gleichungssystems (6.27) mit den Anfangsbedingungen (6.27) in schwacher Form ermittelt werden.

Das Optimierungsproblem (6.67) wird mit dem Verfahren (V5) aus Kapitel 6.5.4 iterativ gelöst. Als Startwert wird erneut der Vektor

$$\boldsymbol{\theta}^0 \;:=\; (k_a^0, k_d^0, \nu^0, \gamma^0)^\top \;:=\; (4.5, 23.0, 3.7, 41)^\top$$

gewählt, wobei für den zugehörigen Zielfunktionswert

$$J(\boldsymbol{\theta}^0) \;=\; 55462.31$$

gilt. Das nach 43 Iterationsschritten mit dem Verfahren (V5) ermittelte globale Minimum lautet

$$\boldsymbol{\theta}^* \;=\; (k_a^*, k_d^*, \nu^*, \gamma^*)^\top \;=\; (4.8278, 25.6012, 3.2840, 20.4461)^\top \tag{6.68}$$

mit dem Zielfunktionswert

$$J(k_a^*, k_d^*, \nu^*, \gamma^*) \;=\; 130.49.$$

Die folgende Tabelle 6.6 gibt einen Überblick über die Anzahl der Iterationen und dem Zeitaufwand bei der Schätzung der SMA-Parameter wieder.

Verfahren	# Iterationen	durchn. Zeit pro Iterationsschritt	Zeit gesamt
(V5)	$38 + 5$	219 s bzw. 176 s	138.76 min + 14.63 min

Tabelle 6.6: Anzahl der Iterationen und durchschnittlicher Zeitaufwand pro Iterationsschritt bei Verwendung des Verfahrens (V5).

Wie in Kapitel 6.5.3 erläutert, wird beim Verfahren (V5) solange die Methode des steilsten Abstiegs umgesetzt, bis sich die Iterierte in einer lokalen Umgebung des gesuchten Minimums befindet. Anschließend wird das Gauss-Newton-Verfahren verwendet. Für die Schätzung der SMA-Parameter wurde in diesem Abschnitt mit der Methode des steilsten Abstiegs eine lokale Umgebung des Minimums nach 38 Schritten erreicht. Nach weiteren 5 Iterationsschritten ermittelte das Gauss-Newton-Verfahren die Näherung $\boldsymbol{\theta}^*$ aus (6.68) des globalen Minimums.

Anhand der Tabelle 6.7 sieht man deutlich, dass die Schätzwerte $\boldsymbol{\theta}^*$ aus (6.68) die exakten SMA-Parameter besser approximieren, als die geschätzten Parameter aus Kapitel 6.5.4, wo lediglich am Ausgang der Chromatographiesäule gemessen wurde.

SMA-Parameter	erste Schätzung	verbesserte Schätzung	exakter Wert
k_a	4.6185	4.8278	5.0
k_d	23.0473	25.6012	25.0
ν	3.2506	3.2840	3.29
γ	–	20.4461	44.7

Tabelle 6.7: Übersicht der geschätzten SMA-Parameter, wenn nur am Säulenausgang (erste Schätzung) und wenn an den Designpunkten des D-optimalen Designs (verbesserte Schätzung) gemessen wurde.

Der geschätzte Schirmungskoeffizient $\gamma* = 20.4461$ approximiert zwar den exakten Parameter $\gamma_{Lys} = 44.7$ noch nicht gut, dennoch konnte überhaupt einmal der Parameter geschätzt werden, was vorher nicht möglich war. Für den relativen Fehler gilt

$$R(\boldsymbol{\theta}^*) \ := \ \frac{\|\boldsymbol{\theta}_{Lys} - \boldsymbol{\theta}^*\|_2}{\|\boldsymbol{\theta}_{Lys}\|_2} \ = \ 0.4705.$$

Kapitel 7

Zusammenfassung und Ausblick

Zusammenfassung

In dieser Arbeit wurde beschrieben, wie ein unbekannter Modellparameter einer instationären partiellen Differentialgleichung mit der Maximum-Likelihood-Methode geschätzt werden kann, wenn die zur Schätzung verwendeten Messdaten einen normalverteilten Messfehler aufweisen mit Erwartungswert Null. In diesem Fall kann die Kovarianzmatrix des Maximum-Likelihood-Schätzers durch die Inverse der Fisher-Informationsmatrix approximiert werden, welche als Maß für die Zuverlässigkeit eines Schätzers verwendet werden kann.

Der Fokus dieser Arbeit lag neben der numerischen Schätzung von Modellparametern auf der Erhöhung der Zuverlässigkeit eines Schätzers durch Bestimmung eines D-optimalen Designs. Ein D-optimales Design ist die Menge der Messstellen mit zugehörigen Gewichten, für die die Determinante der Fisher-Informationsmatrix maximal ist. Es wurde beschrieben, dass die Zuverlässigkeit einer Parameterschätzung erhöht werden kann, wenn die für die Schätzung erforderlichen Messdaten an den Positionen eines solchen Designs erhoben werden.

In dieser Arbeit wurde eine neue Vorgehensweise hergeleitet, mit der ein D-optimales Design durch Maximierung der Determinante der Fisher-Informationsmatrix mithilfe der klassischen Optimierungstheorie bestimmt werden kann. In diesem Zusammenhang wurden die Lösungsverfahren

- primal-duales Innere-Punkte-Verfahren und
- Active-Set-Methode

beschrieben, mit denen eine D-optimale Lösung numerisch ermittelt werden kann. Diesbezüglich wurde in dieser Arbeit bewiesen, dass das Innere-Punkte-Verfahren global lineare Konvergenz aufweist und es wurde beschrieben, dass die Iterierten der Active-Set-Methode global linear und lokal quadratisch gegen eine D-optimale Lösung konvergieren. Da bei der Active-Set-Methode in jedem Iterationsschritt ein wesentlich kleineres System zu lösen ist als beim Innere-Punkte-Verfahren, wird die Active-Set-Methode zur Bestimmung eines D-optimalen Designs empfohlen. Die in dieser Arbeit beschriebene neue Vorgehensweise zur Bestimmung eines D-optimalen Designs mit der Active-Set-Methode besitzt im Vergleich zu bisherigen Standardmethoden die Vorteile, dass

- die Komplexität der Active-Set-Methode unabhängig von einer räumlichen Diskre-

tisierung ist,

- kein komplettes Abspeichern der numerischen Lösung des Zustandes und der Sensitivitäten erforderlich ist.

Die Active-Set-Methode zur Bestimmung eines D-optimalen Designs wurde in dieser Arbeit am Beispiel eines präparativen Säulenchromatographieprozesses zur Aufreinigung von Proteingemischen umgesetzt. Das Ziel war die Verbesserung der Zuverlässigkeit des Maximum-Likelihood-Schätzers, der die SMA-Parameter von Proteinen schätzen kann. Da üblicherweise am Ausgang einer Chromatographiesäule die erforderlichen Messdaten erhoben werden, wurde durch Bestimmung eines D-optimalen Designs am Beispiel des Proteins Lysozym bei pH 7 untersucht, wie sinnvoll diese klassische Messdatenerhebung für die Parameterschätzung ist. Durch numerische Berechnung der jeweiligen Fisher-Informationsmatrizen stellte sich heraus, dass die Fisher-Information bei Verwendung eines D-optimalen Designs in etwa 10^{12} mal kleiner ist, als wenn die Daten lediglich am Ausgang einer Säule erhoben werden würden, was eine enorme Verbesserung der Zuverlässigkeit des Schätzers bedeutet. Da die exakten SMA-Parameter eines Lysozyms bei pH 7 bekannt sind, konnte die klassische Schätzung mit der Schätzung bei Verwendung eines D-optimalen Designs verglichen werden: Es stellte sich heraus, dass einer der vier SMA-Parameter gar nicht geschätzt werden konnte und die anderen drei Parameter nicht besonders gut, wenn lediglich am Ausgang der Chromatographiesäule gemessen wurde. Bei Verwendung eines D-optimalen Designs konnten hingegen alle vier SMA-Parameter geschätzt werden, wobei die auf diese Weise ermittelten Schätzwerte die exakten SMA-Parameter genauer approximierten, als die bei der klassischen Parameterschätzung.

Ausblick

Die numerischen Ergebnisse am Beispiel des Proteins Lysozym bei pH 7 haben gezeigt, dass eine Parameterschätzung verbessert werden kann, wenn die für die Schätzung benötigten Messdaten an den Positionen eines D-optimalen Designs erhoben werden. Die Zuverlässigkeit des Maximum-Likelihood-Schätzers zur Approximation der SMA-Parameter kann allerdings noch weiter erhöht werden: Zusätzlich zur Bestimmung einer D-optimalen Messstellenkonstellation kann die Determinante der Fisher-Informationsmatrix weiter vergrößert werden, indem zum Beispiel die

- optimale Packungsdichte $\varepsilon > 0$ aus (6.3),
- optimale Porengröße $\varepsilon_p > 0$ aus (6.4),
- optimale Säulenlänge $L_c > 0$ aus Kapitel 6.2,
- optimale Salzkonzentration am Eingang der Säule $c_{in,0}(t)$ aus (6.15)

durch Lösen eines Optimierungsproblems ermittelt wird. Die Bestimmung dieser Größen wurde in dieser Arbeit nicht betrachtet, sollte aber zukünftig für eine optimale Schätzung der SMA-Parameter in Betracht gezogen werden.

In dieser Arbeit wurde beschrieben, wie im allgemeinen eine optimale Messstellenkonstellation bezüglich des D-Optimalitätskriteriums (2.7) mithilfe der klassischen Optimierungstheorie ermittelt werden kann. Je nach Anwendung kann allerdings die Bestimmung einer optimalen Messstellenkonstellation bezüglich

- des E-Optimalitätskriteriums aus (2.8) oder
- des A-Optimalitätskriteriums aus (2.9)

sinnvoller sein. So bewirkt zum Beispiel die Maximierung der Fisher-Informationsmatrix bezüglich (2.8) eine Minimierung der längsten Hauptachse des Konfidenzellipsoids. Eine auf diese Weise verbesserte Zuverlässigkeit eines Schätzers ist zum Beispiel dann zu wählen, wenn genau ein Parameter von mehreren schlecht geschätzt werden kann. Für anschließende Arbeiten wäre die Untersuchung der Bestimmung eines optimalen Designs bezüglich (2.8) und (2.9) analog zur Vorgehensweise in Kapitel 4 mithilfe der klassichen Optimierungstheorie von Interesse. Wie in Kapitel 4 könnten dann Lösungsverfahren zur Bestimmung eines optimalen Designs beschrieben und Konvergenzaussagen getätigt werden.

Interessant wäre außerdem eine Erweiterung dieser Arbeit auf den Bereich der sogenannten *Multiphysik*, wenn mehrere untereinander gekoppelte, mathematische Modelle für die Schätzung von Modellparametern verwendet werden. Aufgrund der Kopplung dieser Modelle kann eine Parameterschätzung und die Bestimmung eines D-optimalen Designs zur Erhöhung der Zuverlässigkeit einer solchen Schätzung sehr komplex sein. Daher wäre es interessant, wenn anhand von gekoppelten Modellen untersucht wird, unter welchen Voraussetzungen ein D-optimales Design zur Verbesserung der Schäzung von Modellparametern existiert und mit welchen Lösungsmethoden ein D-optimales Design numerisch ermittelt werden kann.

Anhang A

Hessematrix

Mit $J(\mathbf{x},\mathbf{p}) = -\ln|M(\mathbf{x},\mathbf{p})|$, $\mathbf{x}\in\Omega^\ell$, $\mathbf{p}\in[0,1]^\ell$ gilt für die Hessematrix $\nabla^2_{\mathbf{x}} J(\mathbf{x},\mathbf{p})$:

- Berechnung der Diagonalelemente der Hessematrix $\nabla^2_{\mathbf{x}} J(\mathbf{x},\mathbf{p})$:

$$\begin{aligned}
\frac{\partial^2 J(x,p)}{\partial (x_i^r)^2} &= -2\,p_i \frac{\partial}{\partial x_i^r}\left[\int_T y(t,x_i)^\top M^{-1}\frac{\partial y(t,x_i)}{\partial x_i^r}\,dt\right] \\
&= -2\,p_i\int_T y(t,x_i)^\top M^{-1}\,\frac{\partial^2 y(t,x_i)}{\partial (x_i^r)^2}\,dt \;-\; 2\,p_i\int_T y(t,x_i)^\top \frac{\partial M^{-1}}{\partial x_i^r}\frac{\partial y(t,x_i)}{\partial x_i^r}\,dt \\
&\quad -2p_i\int_T \frac{\partial y(t,x_i)}{\partial x_i^r}^\top M^{-1}\,\frac{\partial y(t,x_i)}{\partial x_i^r}\,dt \\
&\overset{(4.10)}{=} -2\,p_i\int_T y(t,x_i)^\top M^{-1}\,\frac{\partial^2 y(t,x_i)}{\partial (x_i^r)^2}\,dt \\
&\quad +2\,p_i^2\int_T y(t,x_i)^\top M^{-1}\,\frac{\partial M}{\partial x_i^r}\,M^{-1}\frac{\partial y(t,x_i)}{\partial x_i^r}\,dt \\
&\quad -2p_i\int_T \frac{\partial y(t,x_i)}{\partial x_i^r}^\top M^{-1}\,\frac{\partial y(t,x_i)}{\partial x_i^r}\,dt \\
&= -2\,p_i\int_T y(t,x_i)^\top M^{-1}\,\frac{\partial^2 y(t,x_i)}{\partial (x_i^r)^2}\,dt \\
&\quad +2\,p_i^2\int_T y(t,x_i)^\top M^{-1}\int_T y(t,x_i)\frac{\partial y(t,x_i)}{\partial x_i^r}^\top dt\,M^{-1}\frac{\partial y(t,x_i)}{\partial x_i^r}\,dt \\
&\quad +2\,p_i^2\int_T y(t,x_i)^\top M^{-1}\int_T \frac{\partial y(t,x_i)}{\partial x_i^r}y(t,x_i)^\top dt\,M^{-1}\frac{\partial y(t,x_i)}{\partial x_i^r}\,dt \\
&\quad -2p_i\int_T \frac{\partial y(t,x_i)}{\partial x_i^r}^\top M^{-1}\,\frac{\partial y(t,x_i)}{\partial x_i^r}\,dt
\end{aligned}$$

- Berechnung der Nichtdiagonalelemente der Hessematrix $\Delta_x J(x,p)$.
  Sei $i\neq j$:

$$\frac{\partial^2 J(x,p)}{\partial x_i^r\,\partial x_j^s} = -2\,p_i\frac{\partial}{\partial x_j^s}\left[\int_T y(t,x_i)^\top M^{-1}\,\frac{\partial y(t,x_i)}{\partial x_i^r}\,dt\right]$$

$$
\begin{aligned}
&= -2\,p_i \left[\int_T y(t,x_i)^\top \frac{\partial M^{-1}}{\partial x_j^s} \frac{\partial y(t,x_i)}{\partial x_i^r}\, dt \right] \\
&\overset{(4.10)}{=} +2\,p_i^2 \int_T y(t,x_i)^\top M^{-1} \int_T \frac{\partial y(t,x_j)}{\partial x_j^s} y(t,x_j)^\top dt\, M^{-1} \frac{\partial y(t,x_i)}{\partial x_i^r}\, dt \\
&\quad +2\,p_i^2 \int_T y(t,x_i)^\top M^{-1} \int_T y(t,x_j) \frac{\partial y(t,x_j)}{\partial x_j^s}^\top dt\, M^{-1} \frac{\partial y(t,x_i)}{\partial x_i^r}\, dt \\
&= +2\,p_i^2 \int_T y(t,x_j)^\top M^{-1} \int_T \frac{\partial y(t,x_i)}{\partial x_i^r} y(t,x_i)^\top dt M^{-1} \frac{\partial y(t,x_j)}{\partial x_j^s}\, dt \\
&\quad +2\,p_i^2 \int_T y(t,x_j)^\top M^{-1} \int_T y(t,x_i) \frac{\partial y(t,x_i)}{\partial x_i^r}^\top dt M^{-1} \frac{\partial y(t,x_j)}{\partial x_j^s}\, dt \\
&= \frac{\partial^2 J(x,p)}{\partial x_j^r\, \partial x_i^s}
\end{aligned}
$$

Sei $i = j$ und $r \neq s$:

$$
\begin{aligned}
\frac{\partial^2 J(x,p)}{\partial x_i^r\, \partial x_i^s} &= -2\,p_i \frac{\partial}{\partial x_i^r} \left[\int_T y(t,x_i)^\top M^{-1} \frac{\partial y(t,x_i)}{\partial x_i^s}\, dt \right] \\
&= -2\,p_i \int_T y(t,x_i)^\top M^{-1} \frac{\partial^2 y(t,x_i)}{\partial x_i^r\, \partial x_i^s}\, dt \\
&\quad +2\,p_i^2 \int_T y(t,x_i)^\top M^{-1} \int_T y(t,x_i) \frac{\partial y(t,x_i)}{\partial x_i^s}^\top dt\, M^{-1} \frac{\partial y(t,x_i)}{\partial x_i^r}\, dt \\
&\quad +2\,p_i^2 \int_T y(t,x_i)^\top M^{-1} \int_T \frac{\partial y(t,x_i)}{\partial x_i^s} y(t,x_i)^\top dt\, M^{-1} \frac{\partial y(t,x_i)}{\partial x_i^r}\, dt \\
&\quad -2p_i \int_T \frac{\partial y(t,x_i)}{\partial x_i^s}^\top M^{-1} \frac{\partial y(t,x_i)}{\partial x_i^r}\, dt \\
&= \frac{\partial^2 J(x,p)}{\partial x_i^s\, \partial x_i^r}
\end{aligned}
$$

Die Hessematrix $\Delta_x J(x,p)$ ist demnach symmetrisch. □

Anhang B

Tabellen und Abbildungen

B.1 Unzulässige Lösungen des Problems (5.2)

Für $h > 50^{-1}$ mit $\Delta t^* := 10^{-3}$ fest und $\Delta t > 15^{-1}$ mit $h^* := 640^{-1}$ fest, liegen die mit dem Algorithmus 3.3.1 ermittelten Lösungen $(\mathbf{x}^*(h, \Delta t), \mathbf{p}^*(h, \Delta t))$ des Problems (5.2) nicht im Lösungsraum $\bar{\Omega}^3 \times [0,1]^3$. Mit

$$\boldsymbol{\xi}^*(h, \Delta t) \ := \ \begin{Bmatrix} \mathbf{x}^*(h, \Delta t)^\top \\ \mathbf{p}^*(h, \Delta t)^\top \end{Bmatrix}$$

wurden folgende D-optimale Designs mit dem Algorithmus 3.3.1 ermittelt:

$$\boldsymbol{\xi}^*(10^{-1}, \Delta t^*) \ = \ \begin{Bmatrix} (0.6, 0.3) & (0.3, 0.7) & (0.2, 0.2) & (0.7, 0.3) & (0.3, 0.6) \\ 0.08 & 0.25 & 0.33 & 0.25 & 0.08 \end{Bmatrix}$$

$$\boldsymbol{\xi}^*(15^{-1}, \Delta t^*) \ = \ \begin{Bmatrix} (0.27, 0.6) & (0.13, 0.13) & (0.67, 0.27) & (0.27, 0.67) & (0.6, 0.27) \\ 0.21 & 0.33 & 0.12 & 0.12 & 0.21 \end{Bmatrix}$$

$$\boldsymbol{\xi}^*(20^{-1}, \Delta t^*) \ = \ \begin{Bmatrix} (0.3, 0.65) & (0.15, 0.15) & (0.65, 0.25) & (0.25, 0.65) & (0.65, 0.3) \\ 0.21 & 0.33 & 0.12 & 0.12 & 0.21 \end{Bmatrix}$$

$$\boldsymbol{\xi}^*(30^{-1}, \Delta t^*) \ = \ \begin{Bmatrix} (0.27, 0.63) & (0.17, 0.17) & (0.63, 0.27) & (0.3, 0.63) \\ 0.29 & 0.33 & 0.33 & 0.04 \end{Bmatrix}$$

$$\boldsymbol{\xi}^*(40^{-1}, \Delta t^*) \ = \ \begin{Bmatrix} (0.27, 0.65) & (0.15, 0.15) & (0.63, 0.25) & (0.65, 0.3) \\ 0.33 & 0.33 & 0.19 & 0.15 \end{Bmatrix}$$

$$\boldsymbol{\xi}^*(h^*, 10^{-1}) \ = \ \begin{Bmatrix} (0.53, 0.3) & (0.28, 0.65) & (0.16, 0.15) & (0.64, 0.27) \\ 0.02 & 0.33 & 0.34 & 0.32 \end{Bmatrix}$$

B.2 Numerische Lösung von (5.2) mit Algorithmus 3.3.1 bei fest gewähltem $\mathbf{h^* > 0}$

Bei fest gewählter Schrittweite $h^* := (\sqrt{N(h^*)} - 1)^{-1} > 0$, wobei $N(h^*) \in \mathbb{N}$ derart gewählt wird, dass $\sqrt{N(h^*)} \in \mathbb{N}$ gilt, erhält man das Gitter Ω_{h^*} durch

$$\Omega_{h^*} \ := \ \{\mathbf{x} = (x_1, x_2)^\top \in \bar{\Omega} : \ x_1 = kh^*, \ x_2 = lh^* \text{ mit } k, l \in \{1, \ldots, \sqrt{N(h^*)}\}\}.$$

Wendet man den Algorithmus 3.3.1 auf das Designproblem (5.2) mit (5.1) an und nutzt dabei das Gitter Ω_{h^*} mit der Schrittweite $h^* := 640^{-1}$, dann erhält man mit jeder Wahl $\Delta t \leq 15^{-1}$ das D-optimale Design

$$\boldsymbol{\xi}(h^*, \Delta t) = \begin{Bmatrix} (0.2781, 0.6500) & (0.1578, 0.1531) & (0.6359, 0.2672) \\ 0.333 & 0.334 & 0.333 \end{Bmatrix}.$$

B.3 Numerische Lösungen eines präparativen Chromatographieprozesses

In Kapitel 6.5 wurde beschrieben, wie die SMA-Parameter eines Lysozyms bei pH 7 mithilfe der präparativen Säulenchromatographie geschätzt werden können. Zur Realisierung dessen werden die numerischen Lösungen der Gleichungen (6.59) mit den Anfangsbedingungen (6.60) benötigt, die einen solchen Prozess beschreiben.

Folgende Abbildungen zeigen die mit der Methode der finiten Elemente ermittelten numerischen Lösungen dieser Gleichungen.

Die numerische Lösung $\mathbf{c_0(t, x)}$ (Salzkonzentration in der mobilen Phase)

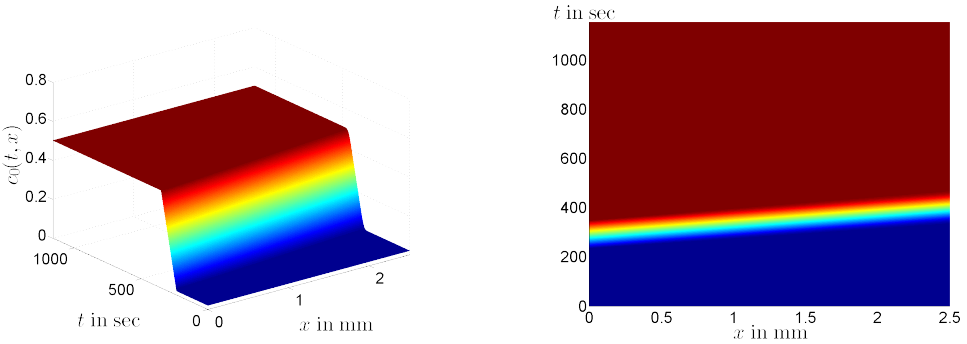


Abbildung B.1: Die Salzkonzentration in der mobilen Phase $c_0(t, x)$.

Die numerische Lösung $\mathbf{c_{p,0}(t, x)}$ (Salzkonzentration in den Poren)

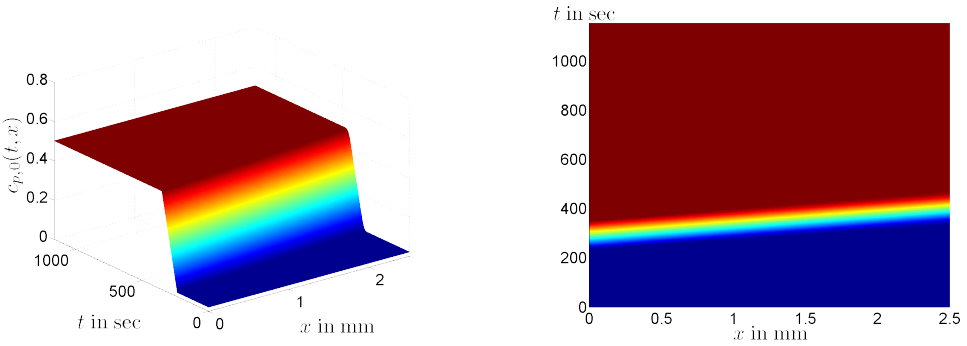


Abbildung B.2: Die Salzkonzentration in den Poren $c_{p,0}(t, x)$.

Die numerische Lösung $\mathbf{c_1(t,x)}$ (Lysozymkonzentration in der mobilen Phase)

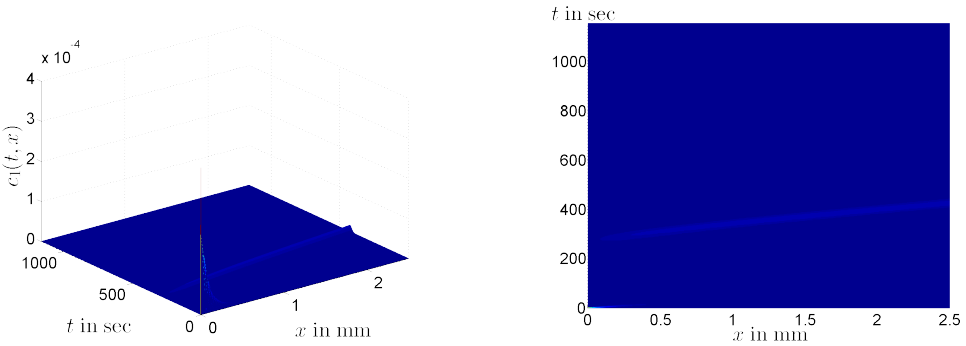


Abbildung B.3: Die Lysozymkonzentration in der mobilen Phase $c_1(t,x)$.

Die numerische Lösung $\mathbf{c_{p,1}(t,x)}$ (Lysozymkonzentration in den Poren)

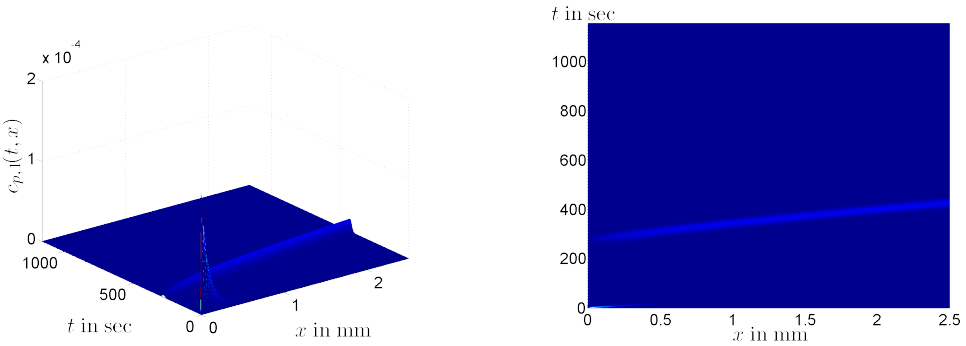


Abbildung B.4: Die Lysozymkonzentration in den Poren $c_{p,1}(t,x)$.

Die numerische Lösung $\mathbf{q_1(t,x)}$ (Lysozymkonzentration in der stationären Phase)

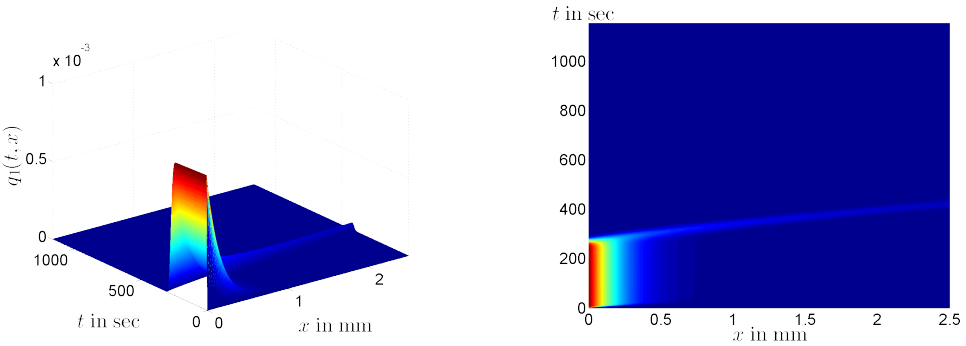


Abbildung B.5: Die Lysozymkonzentration in der stationären Phase $q_1(t,x)$.

B.4 Eigenvektoren zur Darstellung eines Konfidenzellipsoids

B.4.1 Messdatenerhebung am Säulenausgang

Die zu den Eigenwerten

$$\kappa_1 = 2.12 \cdot 10^4, \quad \kappa_2 = 1.94 \cdot 10^5, \quad \kappa_3 = 1.58 \cdot 10^{-1}, \quad \kappa_4 = 3.98 \cdot 10^{10}.$$

der Kovarianzmatrix cov($\hat{\boldsymbol{\theta}}$) aus Kapitel 6.6.2 zugehörigen Eigenvektoren sind

$$\mathbf{v}^1 = \begin{pmatrix} 0.882 \\ -0.402 \\ -0.245 \\ 2.1 \cdot 10^{-3} \end{pmatrix}, \mathbf{v}^2 = \begin{pmatrix} -0.401 \\ -0.914 \\ 0.056 \\ -3.8 \cdot 10^{-4} \end{pmatrix}, \mathbf{v}^3 = \begin{pmatrix} 0.246 \\ -0.049 \\ 0.968 \\ 1.6 \cdot 10^{-4} \end{pmatrix}, \mathbf{v}^4 = \begin{pmatrix} -1.9 \cdot 10^{-3} \\ 4.7 \cdot 10^{-4} \\ 6.8 \cdot 10^{-4} \\ 0.999998 \end{pmatrix}.$$

B.4.2 Messdatenerhebung an den Designpunkten des D-optimalen Designs

Die zu den Eigenwerten

$$\kappa_1 = 9.31 \cdot 10^{-1}, \quad \kappa_2 = 1.76 \cdot 10^3, \quad \kappa_3 = 2.64 \cdot 10^{-2}, \quad \kappa_4 = 2.96 \cdot 10^5.$$

der Kovarianzmatrix cov($\hat{\boldsymbol{\theta}}$) aus Kapitel 6.6.3 zugehörigen Eigenvektoren sind

$$\mathbf{v}^1 = \begin{pmatrix} -0.994 \\ 0.028 \\ 0.108 \\ 6.3 \cdot 10^{-3} \end{pmatrix}, \mathbf{v}^2 = \begin{pmatrix} 0.031 \\ 0.996 \\ 0.024 \\ -0.086 \end{pmatrix}, \mathbf{v}^3 = \begin{pmatrix} 0.107 \\ -0.027 \\ 0.994 \\ 6.4 \cdot 10^{-4} \end{pmatrix}, \mathbf{v}^4 = \begin{pmatrix} 9.0 \cdot 10^{-3} \\ 0.086 \\ 2.0 \cdot 10^{-3} \\ 0.996 \end{pmatrix}.$$

Tabellenverzeichnis

Abbildungsverzeichnis

Literaturverzeichnis

[1] A. Adams, J. Fournier. *Sobolev Spaces*. Academic Press. 2003

[2] John Aldrich. *R. A. Fisher and the Making of Maximum Likelihood 1912-1922*. Statistical Science, Volume 12, Issue 3, pp. 162-176. 1997

[3] U. Altenhöner, M. Meurer, J. Strube, H. Schmidt-Traub. *Parameter estimation for the simulation of liquid chromatography*. Journal of Chromatography A, Volume 769, Issue 1, pp. 59-69. 1997

[4] A. Andrzejewska, K. Kaczmarski, G. Guiochon. *Theoretical study of the accuracy of the pulse method, frontal analysis, and frontal analysis by characteristic points for the determination of single component adsorption isotherms*. Journal of Chromatography A, Volume 1216, Issue 7, pp. 1067-1083. 2009

[5] H. Arellano-Garcia, J. Schöneberger, S. Körkel. *Optimale Versuchsplanung in der chemischen Verfahrenstechnik*. Chemie Ingenieur Technik, Volume 79, Issue 10. 2007

[6] A. Atkinson, R. Bailey. *One hundred years of the design of experiments on and off the pages of Biometrika*. Biometrika, Volume 88, Issue 1, pp. 53-97. 2001

[7] S. Bandara, J. Schlöder, R. Eils, H. Bock, T. Meyer. *Optimal Experimental Design for Parameter Estimation of a Cell Signaling Model*. PLoS Computational Biology, Volume 5, Issue 11. 2009

[8] J. Banga, E. Balsa-Canto. *Parameter estimation and optimal experimental design*. Essays Biochem., Volume 45, pp. 195-209. 2008

[9] T. Barza, V Löffler, H. Arellano-Garciaa, G. Woznya. *Optimal determination of steric mass action model parameters for β-lactoglobulin using static batch experiments*. Journal of Chromatography A, Volume 1217, Issue 26, pp. 4267-4277. 2010

[10] R. Becker, B. Vexler. *A posteriori error estimation for finite element discretization of parameter identification problems*. Numerische Mathematik, Volume 96, Number 3, 435-459. 2003

[11] M. Bernakiewicz, M. Viceconti. *The role of parameter identification in finite element contact analyses with reference to orthopaedic biomechanics applications*. Journal of Biomechanics, Volume 35, Issue 1, pp. 61-67. 2002

[12] D. Braess. *Finite Elemente: Theorie, schnelle Löser und Anwendungen in der Elastizitätstheorie*. Springer-Verlag. 2003

[13] C. A. Brooks, S.M. Cramer. *Steric Mass-Action Ion Exchange: Displacement Profiles and Induced Salt Gradients*. AIChE Journal, Vol. 38, No.12. 1992

[14] G. Carta, A. R. Ubiera, T. M. Pabst. *Protein Mass Transfer Kinetics in Ion Exchange Media: Measurements and Interpretations*. Chemical Engineering & Technology, Volume 28, Issue 11, pp. 1252-1264. 2005

[15] S. Chan, N. Titchener-Hooker, D. Bracewell. *A systematic approach for modeling chromatographic processes - Application to protein purification*. AIChE Journal, Volume 54, Issue 4, pp. 965 - 977. 2008

[16] D. Connifee. *R.A. Fisher and the development of statistics - a view in his centenary year*. Journal of the Statistical and Social Inquiry Society of Ireland, Volume 26, Issue 3. 1990

[17] C. Czado, T. Schmidt. *Mathematische Statistik*. Springer. 2011

[18] A. Dale. *Optimal Experimental Design for Event-Related fMRI*. Human Brain Mapping 8, pp. 109-114. 1999

[19] C. Eck, H. Garcke, P. Knabner. *Mathematische Modellierung*. Springer. 2010

[20] E. Emmrich. *Gewöhnliche und Operator-Differentialgleichungen*. Vieweg. 2004

[21] A. Ern, J. Guermond. *Theory and Practice of Finite Elements*. Springer. 2004

[22] L. Fahrmeir, T. Kneib, S. Lang. *Regression - Modelle, Methoden und Anwendungen*. Springer. 2009

[23] W. T. Federer. *Experimental Design, Theory and Application*. The macmillan Company. 1955

[24] V. V. Federov. *Theory of optimal experiments*. Academic Press. 1992

[25] V. V. Federov, Peter Hackl. *Model-Oriented Design of Experiments*. Springer-Verlag New York. 1997

[26] M. Ferris, O. Mangasarian, J. Pang. *Complementarity: applications, algorithms, and extensions*. Kluwer Academic Publisher, 179-200. 2001

[27] G. Fischer. *Lineare Algebra*. Vieweg. 2002

[28] S. Gallant, A. Kundu, S. Cramer. *Optimization of step gradient separations: Consideration of nonlinear adsorption*. Biotechnology and Bioengineering, Volume 47, Issue 3, pp. 355-372. 1995

[29] S. Gallant, A. Kundu, S. Cramer. *Modeling non-linear elution of proteins in ion-exchange chromatography*. Journal of Chromatography A, Volume 702, Issue 1-2, pp. 125-142. 1995

[30] S. Gallant, A. Kundu, S. Cramer. *Optimization of preparative ion-exchange chromatography of proteins: linear gradient separations*. Journal of Chromatography A, Volume 725, Issue 2, pp. 295-314. 1996

[31] Carl Geiger, Christian Kanzow. *Theorie und Numerik restringierter Optimierungsaufgaben*. Springer. 2002

[32] S. Ghosh. *Multivariate Analysis, Design of Experiments and Survey Sampling*. 1999

[33] G. C. Goodwin, R. L. Payne. *Dynamic System Identification: Experimental Design and Data Analysis*. Academic Press. 1977

[34] G. Guiochon. *Preparative liquid chromatography*. Journal of Chromatography A, Volume 24, Issue 1, pp. 165-189. 2002

[35] G. Guiochon, Attila Felinger, Dean G. Shirazi, Anita M. Katti. *Fundamentals of Preparative and Nonlinear Chromatography*. 2006

[36] W. W. Hager, H. Zhang. *A New Active Set Algorithm for Box Constrained Optimization*. SIAM Journal on Optimization, Volume 17, Issue 2. 2006

[37] L. Held. *Methoden der statistischen Interferenz. Likelihood and Bayes*. Spektrum Akademischer Verlag Heidelberg. 2008

[38] M. Hill, R. Cooley, D. Pollock. *A Controlled Experiment in Ground Water Flow Model Calibration*. European Journal of Operational Research, Volume 36, Issue 3, pp. 520-535. 1998

[39] P. Hoel. *Efficiency Problems in Polynomial Estimation*. Annals of Mathematical Statistics, Volume 29, Issue 4, pp. 1134-1145. 1958

[40] P. Hoel. *minimax design in two-dimensional regression*. Annals of Mathematical Statistics, Volume 36, Issue 4, pp. 1097-1106. 1965

[41] H. Horton, L. Moran, K. Scrimgeour, M.Perry, J. Rawn. *Biochemie*. Pearson Studium. 2009

[42] A. F. Izmailovy, M. V. Solodov. *An Active-Set Newton Method for Mathematical Programs with Complementarity Constraints*. SIAM Journal on Optimization, Volume 19, Issue 3. 2006

[43] F. Jarre, Josef Stoer. *Optimierung*. Springer. 2004

[44] R. Jategaonkar. *Bounded-Variable Gauss-Newton Algorithm for Aircraft Parameter Estimation*. Journal of Aircraft, Volume 37, Issue 4, pp. 742-744. 200

[45] M. Kallenrode. *Rechenmethoden der Physik*. Springer. 2005

[46] K. Kaltenböck. *Chromatographie für Einsteiger*. WILEY-VCH Verlag. 1997

[47] C. T. Kelley. *Iterative Methods for Optimization*. SIAM Frontiers in Applied Mathematics, no 18. 1999

[48] J. Kiefer, J. Wolfowitz. *Optimum Designs in Regression Problems*. Annals of Mathematical Statistics, Volume 30, Issue 2, pp. 271-294. 1959.

[49] B. Klar, F. Lindner. *Skriptum zur Vorlesung Statistik im WS 2011/12*. Institut für Stochastik, KIT

[50] S. Körkel, I. Bauer, H. Bock, J. Schlöder. *A sequential approach for nonlinear optimum experimental design in DAE systems*. Scientific Computing in Chemical Engineering II, Volume 2, Springer Verlag. 1999

[51] Stefan Körkel. *Numerische Methoden für Optimale Versuchsplanungsprobleme bei nichtlinearen DAE-Modellen*. Doktorarbeit, Universität Heidelberg. 2002

[52] S. Körkel, I. Bauer, H. Bock, J. Schlöder. *Statistische Versuchsplanung*. GIT Labor-Fachzeitschrift 10, pp. 820-823. 2007

[53] R. Kornhuber, C. Schütte. *Numerik von partiellen Differentialgleichungen*. Skript, FU Berlin. 2011

[54] O. Krafft. *Lineare statistische Modelle und optimale Versuchsplanung*. Vandenhoeck & Ruprecht Göttingen. 1978

[55] M. Krapp, J. Nebel. *Methoden der Statistik*. Vieweg+Teubner. 2011

[56] Z. Li , V. Milenkovic. *Constructing Strongly Convex Hulls Using Exact or Rounded Arithmetic*. Algorithmica, Volume 8, pp. 345-364. 1992

[57] K. Liebscher, D. Liebscher. *Zu der von Gauss gegebenen Begründung der Methode der kleinsten Quadrate*. Potsdam-Babelsberg Die Sterne, Volume 53, Issue 1, pp. 15-21. 1977

[58] P. Lindner, B. Hitzmann. *Experimental design for optimal parameter estimation of an enzyme kinetic process based on the analysis of the Fisher information matrix*. Journal of Theoretical Biology, Volume 238, Issue 1, pp. 111-123. 2006

[59] A. J. Miller, N. Nguyen *A Fedorov Exchange Algorithm for D-Optimal Design*. Journal of the Royal Statistical Society. Series C (Applied Statistics), Volume 43, Issue 4, pp. 669-677. 1993

[60] T. Mitchell. *An Algorithm for the Construction of „D-Optimal“ Experimental Designs*. Technometrics, Volume 16, Issue 2, pp. 203 - 210. 1974

[61] D. Montgomery. *Design and analysis of experiments*. Wiley. 2005

[62] S. Nickel, O. Stein, K.-H. Waldmann. *Operations Research*. Springer. 2011

[63] J. Nocedal, St. J. Wright. *Numerical Optimization*. Springer. 2006

[64] R. Oliva. *Model calibration as a testing strategy for system dynamics models*. European Journal of Operational Research, Volume 151, Issue 3, pp. 552-568. 2003

[65] A. Osberghaus, M. Haindl, E. von Lieres, J. Hubbuch. *Optimizing a chromatographic three component separation: A comparison of mechanistic and empiric modeling approaches*. Journal of Chromatography A. 2011

[66] A. Pázman. *Foundations of Optimum Experimental Design*. Mathematics and Its Applications. 1986

[67] L. Pronzato, E. Walter. *Experimental Design for Estimating the Optimum Point in a Response Surface*. Acta Applicandae Mathematicae, Vol. 33, No.1. 1993

[68] F. Pukelsheim. *Optimal design of experiments*. Society for Industrial and Applied Mathematics. 2006

[69] A. Quarteroni, A. Valli. *Numerical Approximation of Partial Differential Equations*. Springer. 1994

[70] Z. H. Quereshi, T. S. Ng, G. C. Goodwin. *Optimum experimental design for identification of distributed parameter systems*. International Journal of Control, Volume 31, Number 1, 21-29. 1980

[71] J. Raol, G. Girija, J. Singh. *Modelling And Parameter Estimation Of Dynamic Systems*. IEE Control Engineering. 2004

[72] R. Schlittgen. *Einführung in die Statistik - Analyse und Modellierung von Daten*. Oldenbourg Wissenschaftsverlag GmbH. 2003

[73] K. Schmidt, G. Trenkler. *Mathematische Statistik I*. Einführung in die Moderne Matrix-Algebra: Mit Anwendungen in der Statistik, Springer. 2006

[74] T. Schmidt. *Mathematische Statistik I*. Vorlesungsmitschrieb, Universität Leipzig. 2006

[75] H. Schmidt-Traub. *Preparative Chromatography*. WILEY-VCH Verlag. 2005

[76] H. Schwarz, N. Köckler. *Numerische Mathematik*. Vieweg+Teubner Verlag. 2011

[77] G. Schwedt. *Chromatographische Trenntechniken*. G. Thieme Verlag. 1997

[78] B. Schweizer. *Partielle Differentialgleichungen*. Skript zur Vorlesung, TU Dortmund. 2011

[79] S. Senn. *Francis Galton and regression to the mean*. Wiley, Volume 8, Issue 3, pp. 124-126. 2011

[80] K. R. Shah, B. K. Sinha. *Theory of Optimal Designs*. Springer-Verlag Berlin Heidelberg. 1989

[81] K. Siebertz, D. Bebber, T. Hochkirchen. *Varianten der statistischen Versuchsplanung*. Springer. 2010

[82] M. Stehlík. *Some Properties of Exchange Design Algorithms Under Correlation*. Research Report Series, Nummer 28. 2006

[83] M. Stehlík, W. Müller. *Fisher information in the design of computer simulation experiments*. Journal of Physics: Conference Series, Volume 135, Issue 1. 2008

[84] V. Thomée. *Galerkin Finite Element Methods for Parabolic Problems*. Springer. 2006

[85] D. Ucinski. *Optimal Measurement Methods for Distributed Parameter System Identification*. CRC Press. 2005

[86] D. Ucinski. *Measurement Optimization for Parameter Estimation in Distributed Systems*. Technical University Press Zielona Góra. 1999

[87] D. Ucinski and A. Atkinson. *Experimental Design for Time Dependent Models with Correlated Observations*. Studies in Nonlinear Dynamics & Econometrics, Volume 8, Issue 2. 2004

[88] M. Ulbrich, S. Ulbrich. *Primal-dual interior point methods for pde-constrained optimization*. Mathematical Programming, Springer, 2007

[89] A. Wakolbinger. *Statistik*. Skript zur Vorlesung, Universität Frankfurt, 2002

[90] St. J. Wright. *Primal-dual interior point methods*. SIAM, Philadelphia, 1997

[91] Henry P. Wynn *The Sequential Generation of D-Optimum Experimental Designs*. Annals of Mathematical Statistics, Volume 41, Issue 5, pp. 1655-1664. 1970.

[92] E. Zeidler. *Applied Functional Analysis: Main Principles and Their Applications*. Springer-Verlag. 1995

[93] Z. Zhang. *Parameter Estimation Techniques: A Tutorial with Application to Conic Fitting*. Image and Vision Computing Journal, Volume 15, Issue 1, pp. 59-76. 1997

[94] E. Zivot, J. Wang. *Generalized Method of Moments*. Springer. 2006