Karlsruher Institut für Technologie

Bachelorarbeit

Die Multilevel Monte Carlo Methode und deren Anwendung am Beispiel der linearen Transportgleichung

Tim Buchholz

25.03.20

Betreuung: Prof.Dr. Christian Wieners und M.Sc. Niklas Baumgarten

Fakultät für Mathematik

Karlsruher Institut für Technologie

Inhaltsverzeichnis

1 Einleitung

TODO (wird zu einem späteren Zeitpunkt eingefügt)

2 Grundlagen

2.1 Analytische/numerische Grundlagen

Sei $\mathcal{D} \subseteq \mathbb{R}^d$ offen für $d \in \mathbb{N}$ und $\|\cdot\|$ eine Norm auf $\mathbb{R}^d$. Die folgenden Definitionen und Sätze sollen als Grundlagen für die weiteren Betrachtungen dieser Thesis dienen. Insbesondere wollen wir hierbei meist auf konkrete Beweise verzichten und verweisen dahingehend auf die Literatur. Die analytischen Grundlagen bauen zum Teil auf der Vorlesung Rand- und Eigenwertprobleme aus dem Sommersemester 2019 von Herrn Prof. Dr. Reichel auf, sind aber auch z.B. in [14] oder [16] zu finden.

Definition 2.1. (Einige Operatoren)

(a) Für $F : \mathbb{R}^d \to \mathbb{R}^d$ ist die Divergenz von F definiert durch

$$\operatorname{div} : \begin{cases} \mathbb{R}^d \to \mathbb{R} \\ F \mapsto \operatorname{div} F = \nabla \cdot F \coloneqq \sum_{i=1}^{d} \frac{\partial F_i}{\partial x_i} \end{cases}$$

(b) Für $f : \mathbb{R}^d \to \mathbb{R}$ und $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}^d$ ist die partielle Ableitung von f nach dem sogenannten Multiindex α definiert durch

$$\partial^\alpha f \coloneqq \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}} = \frac{\partial^{\alpha_1 + \cdots + \alpha_d} f}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}$$

Satz 2.2. (Gaußscher Integralsatz für Lipschitz-Gebiete)
Sei $\mathcal{D} \subset \mathbb{R}^d$ ein beschränktes Lipschitz-Gebiet und sei n der äußere Einheitsnormalenvektor an $\partial\mathcal{D}$. Dann gilt:

$$\int_{\mathcal{D}} \frac{\partial f}{\partial x_i} \, \mathrm{d}x = \int_{\partial\mathcal{D}} f n_i \, \mathrm{d}a$$

für jede Funktion $f \in C^1(\overline{\mathcal{D}})$.
Oft erscheint der Gaußsche Integralsatz auch in folgender Form:

$$\int_{\mathcal{D}} \operatorname{div} F \, \mathrm{d}x = \int_{\partial\mathcal{D}} F \cdot n \, \mathrm{d}a$$

wobei $F : \mathcal{D} \to \mathbb{R}^d$ ein Vektorfeld ist. Die Komponentenfunktionen von $F = (f_1, \ldots, f_d)$ sollen dann $f_i \in C^1(\overline{\mathcal{D}})$ für $i = 1, \ldots, n$ erfüllen.

Folgerung 2.3. (mehrdimensionale partielle Integration)
Sei $f \in C^1(\mathbb{R}^d, \mathbb{R})$ und $F : \mathcal{D} \to \mathbb{R}^d$ ein stetig partiell differenzierbares Vektorfeld. Dann gilt:

$$\int_{\mathcal{D}} f \operatorname{div}(F) \, \mathrm{d}x = \int_{\partial\mathcal{D}} fF \cdot n \, \mathrm{d}a - \int_{\mathcal{D}} F \cdot \nabla f \, \mathrm{d}x$$

Definition 2.4. (schwache Ableitung)
Sei $f \in L^1_{\text{loc}}(\mathcal{D})$. Dabei bezeichne $L^1_{\text{loc}}(\mathcal{D})$ den Raum der lokal integrierbaren Funktionen auf $\mathcal{D}$. Wir sagen v besitzt eine schwache Ableitung zum Multiindex α, falls eine Funktion $v \in L^1_{\text{loc}}$ existiert, mit

$$\int_{\mathcal{D}} f\partial^\alpha \phi \,\mathrm{d}x = (-1)^{|\alpha|} \int_{\mathcal{D}} v\phi \,\mathrm{d}x \qquad \forall \phi \in C_c^\infty(\mathcal{D})$$

In diesem Zusammenhang nennen wir Φ auch Testfunktion und wir definieren $D^\alpha f \coloneqq v$ als die schwache Ableitung von f zum Multiindex α.

Bemerkung. Per Konvention ist für $\alpha = (0, \ldots, 0) \quad \partial^\alpha f = f$

Definition 2.5. (Sobolevräume)
Sei $\mathcal{D} \subseteq \mathbb{R}^d$ offen, $k \in \mathbb{N}$ und $1 \leq p \leq \infty$. Weiter sei $L : C^1(\mathcal{D}, \mathbb{R}^m) \to L^\infty(\mathcal{D}, \mathbb{R}^k)$ ein linearer Differentialoperator erster Ordnung und $L^\star : C^1(\mathcal{D}, \mathbb{R}^k) \to L^\infty(\mathcal{D}, \mathbb{R}^m)$ der zugehörige adjungierte Operator. Es gelte also

$$\int_{\mathcal{D}} Lf \cdot \phi \,\mathrm{d}x = \int_{\mathcal{D}} f \cdot L^\star \phi \,\mathrm{d}x \text{ für } f \in C_c^1(\mathcal{D}, \mathbb{R}^m) \text{ , } \phi \in C_c^1(\mathcal{D}, \mathbb{R}^k)$$

Dann sind:

(a) $W^{k,p}(\mathcal{D}) \coloneqq \{f \in L^p(\mathcal{D})$ und die schwachen Ableitungen $\partial^\alpha f$ existieren, mit $\partial^\alpha f \in L^p(\mathcal{D})$ für alle $\alpha \in \mathbb{N}_0^d, |\alpha| \leq k\}$

(b) $\|f\|_{k,p} = \|f\|_{W^{k,p}(\mathcal{D})} \coloneqq \begin{cases} (\sum\limits_{|\alpha|\leq k} \int_{\mathcal{D}} |\partial^\alpha f|^p \,\mathrm{d}x)^{\frac{1}{p}} & , 1 \leq p < \infty \\ \sum\limits_{|\alpha|\leq k} \|\partial^\alpha f\|_\infty & , p = \infty \end{cases}$

(c) $W_0^{k,p}(\mathcal{D}) \coloneqq \overline{C_c^\infty(\mathcal{D})}^{\|\cdot\|_{k,p}}$. Über den sogenannten Spursatz erhält man eine äquivalente Charakterisierung durch: $W_0^{k,p}(\mathcal{D}) = \{f \in W^{k,p}(\mathcal{D}) : f|_{\partial\mathcal{D}} = 0\}$

(d) Im Falle $p = 2$ schreibt man aufgrund der Tatsache, dass es sich dann bei $W^{k,p}(\mathcal{D})$ um einen Hilbertraum handelt, oft auch $H^k(\mathcal{D}) \coloneqq W^{k,p}(\mathcal{D})$

(e) $H(L, \mathcal{D}) \coloneqq \{f \in L^2(\mathcal{D}, \mathbb{R}^m) : \exists v \in L^2(\mathcal{D}, R^k)$ mit $\int_{\mathcal{D}} v \cdot \phi \,\mathrm{d}x = \int_{\mathcal{D}} f \cdot L^\star \phi \,\mathrm{d}x \; \forall \phi \in C_c^1(\mathcal{D}, \mathbb{R}^k)\}$

Satz 2.6. (Multiplikation mit Testfunktionen und Integration)
Sei $\mathcal{D} \subset \mathbb{R}^d$ offen und $f \in L^1_{\text{loc}}(\mathcal{D})$ und $\int_{\mathcal{D}} f\phi \,\mathrm{d}x = 0 \; \forall \phi \in C_c^\infty(\mathcal{D})$,
dann gilt $f \equiv 0$.

2.2 Stochastische Grundlagen

An dieser Stelle wollen wir an einige grundlegende Resultate der Wahrscheinlichkeitstheorie erinnern. Außerdem führen wir dabei auch Teile der Notation ein, die wir an späterer Stelle noch brauchen werden. Als Referenzen sind vor allem [5], die Vorlesung Wahrscheinlichkeitstheorie von Herrn Prof. Dr. Henze (SS18) sowie [22] zu nennen.
Sei $\Omega \neq \emptyset$ eine beliebige nichtleere Teilmenge. Einige grundlegende Begriffe der Maß- und Wahrscheinlichkeitstheorie wollen wir an dieser Stelle voraussetzen, sie sind aber ebenfalls in [22] zu finden. Dazu zählen:

- eine σ-Algebra $\mathcal{A} \subset \mathcal{P}(\Omega)$
- die von einem Mengensystem $\mathcal{M} \subset \mathcal{P}(\Omega)$ erzeugte σ-Algebra $\sigma(\mathcal{M})$
- ein Maß μ auf einer σ-Algebra $\mathcal{A}$
- das Maß-Integral einer messbaren Funktion $f : \Omega \to \overline{\mathbb{R}}$ über einem Maßraum $(\Omega, \mathcal{A}, \mu)$
- die Borel'sche σ-Algebra $\mathcal{B}$, sowie die Begriffe 'Borelmenge' und 'Borel-messbar'.
- ein Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$
- Zufallsvariablen und deren Verteilungen
- stochastische Unabhängigkeit

Sei ab nun $(\Omega, \mathcal{A}, \mathbb{P})$ stets ein Wahrscheinlichkeitsraum und $\mathcal{B}^n$ die Borelsche σ-Algebra über $\mathbb{R}^n$.

Definition 2.7. (stetig verteilte Zufallsvariablen und Zufallsvektoren)
Ein Zufallsvektor $X = (X_1, \ldots, X_n)$ heißt stetig verteilt, wenn eine nichtnegative (Borel-)messbare Funktion $f : \mathbb{R}^n \to \mathbb{R}$ mit $\int_{\mathbb{R}^n} f(x)\,\mathrm{d}x = 1$ existiert, sodass

$$\mathbb{P}^X(B) \coloneqq \mathbb{P}(X \in B) = \int_B f(x)\,\mathrm{d}x, \quad B \in \mathcal{B}^n$$

In diesem Fall heißt f Dichte von X (bzw. von $\mathbb{P}^X$).
Ist $n = 1$ spricht man einfach von einer stetig verteilten Zufallsvariable mit Dichte f.

Definition 2.8. (Erwartungswert)
$X : \Omega \to \overline{\mathbb{R}}$ sei eine Zufallsvariable. Der Erwartungswert von X existiert genau dann, wenn $\int_\Omega |X|\,\mathrm{d}\mathbb{P} < \infty$. In diesem Fall heißt

$$\mathbb{E}[X] \coloneqq \int_\Omega X\,\mathrm{d}\mathbb{P}$$

der Erwartungswert von X. Ist X eine stetig verteilte Zufallsvariable mit Dichte f, so gilt:

$$\int_\Omega X\,\mathrm{d}\mathbb{P} = \int_{\mathbb{R}} x f(x)\,\mathrm{d}x$$

Definition 2.9. (Normalverteilung und multivariate Normalverteilung)

(a) Eine Zufallsvariable X heißt normalverteilt mit Parametern μ und σ^2, falls X die Dichte

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}$$

besitzt. In diesem Fall schreiben wir oft auch $X \sim \mathcal{N}(\mu, \sigma^2)$. Ist spezieller $\mu = 0$ und $\sigma^2 = 1$ so heißt X standardnormalverteilt.

(b) Sei nun $X = (X_1, \ldots, X_n)$ ein Zufallsvektor, $\mu = (\mu_1, \ldots, \mu_n) \in \mathbb{R}^n$ und $C = (\sigma_{ij})_{1 \le i,j \le n} \in \mathbb{R}^{n \times n}$ eine symmetrische positiv-definite Matrix. X bestitzt eine (nicht ausgeartete) multivariate Normalverteilung mit Parametern μ und C, falls die Dichte von X durch

$$f(x) = \frac{1}{\sqrt{(2\pi)^k \det C}} \exp\left(-\frac{1}{2}(x-\mu)^\top C^{-1}(x-\mu)\right), \quad x \in \mathbb{R}^n$$

gegeben ist. Wir schreiben auch $X \sim \mathcal{N}_n(\mu, C)$. Insbesondere kann man mit recht elementaren Methoden einsehen, dass dann auch für jedes $j \in \{1, \ldots, n\}$ $X_j \sim \mathcal{N}(\mu_j, \sigma_{jj})$ gilt und außerdem die Einträge σ_{ij} der Matrix C gerade die Kovarianzen $\mathrm{Cov}(X_i, X_j)$ darstellen. Ein Beweis hierfür findet sich im Appendix. Dieser fasst mehrere Resultate aus [5] zusammen.

Satz 2.10. (starkes Gesetz großer Zahlen)
Es sei $(X_n)_{n\in\mathbb{N}}$ eine u.i.v.-Folge und es gelte $\mathbb{E}[|X_1|] < \infty$. Dann gilt für fast alle $\omega \in \Omega$

$$\mathbb{E}[X_1] = \lim_{n\to\infty} \sum_{i=1}^{n} X_i(\omega)$$

das heißt es existiert eine Menge $\mathfrak{N} \subset \Omega$ mit $\mathbb{P}(\mathfrak{N}) = 0$ und obige Aussage gilt für alle $\omega \notin \mathfrak{N}$. Eine solche Menge $\mathfrak{N}$ heißt auch Nullmenge. In der Literatur findet man diese Art der Konvergenz oft auch unter dem Namen der ($\mathbb{P}$)-fast sicheren Konvergenz.

Satz 2.11. (Chebyshev Ungleichung)
Sei X eine Zufallsvariable mit $\mathbb{E}[X^2] < \infty$, dann gilt für jedes $\epsilon > 0$:

$$\mathbb{P}\left(|X - \mathbb{E}[X]| \ge \epsilon\right) \le \frac{\mathbb{V}[X]}{\epsilon^2}$$

Satz 2.12. (Zentraler Grenzwertsatz)
Sei $(X_n)_{n\in\mathbb{N}}$ eine u.i.v.-Folge und es gelte $\mathbb{E}[X_1^2] < \infty$. Mit $\mathbb{V}[X_1]$ bezeichnen wir die Varianz der Zufallsvariable X_1:

$$\mathbb{V}[X_1] = \mathbb{E}[X_1^2] - \mathbb{E}[X_1]^2 = \mathbb{E}[(X_1 - \mathbb{E}[X_1])^2]$$

Dann gilt für eine standardnormalverteilte Zufallsvariable $\widetilde{\mathcal{N}}$:

$$\hat{S}_n := \frac{\sum_{j=1}^{n} X_j - n\mathbb{E}[X_1]}{\sqrt{n\mathbb{V}[X_1]}} \xrightarrow{\mathcal{D}} \widetilde{\mathcal{N}}$$

Dabei bezeichnet $\xrightarrow{\mathcal{D}}$ die Konvergenz in Verteilung und ist genau dann erfüllt, wenn

$$\lim_{n\to\infty} \mathbb{P}^{\hat{S}_n}((-\infty, x]) = \mathbb{P}^{\widetilde{\mathcal{N}}}((-\infty, x])$$

für alle Stetigkeitsstellen der Verteilungsfunktion $\mathbb{P}^{\widetilde{\mathcal{N}}}((-\infty, \cdot])$ von $\widetilde{\mathcal{N}}$ erfüllt ist.

Definition 2.13. (Zufallsfelder)

(a) Sei $\mathcal{D} \subset \mathbb{R}^d$ eine nichtleere Menge und $d \geq 1$. Wir nennen $X : \Omega \times \mathcal{D} \to \mathbb{R}$ ein Zufallsfeld, wenn für jedes feste $x \in \mathcal{D}$ die Funktion $X(x) := X(\cdot, x)$ eine Zufallsvariable ist.
Ist spezieller $d = 1$ so wird die Parametermenge auch oft mit T bezeichnet und man spricht von einem stochastischen Prozess. In der Literatur findet man Zufallsfelder vor allem unter der englischen Bezeichnung 'random fields'.
Außerdem definieren wir die Erwartung eines Zufallsfeldes durch

$$\mu(x) := \mathbb{E}[X(x)] = \int_\Omega X(w, x)\, \mathrm{d}\mathbb{P}(\omega) \quad \text{für } x \in \mathcal{D}$$

und die zu einem Zufallsfeld gehörige Kovarianzfunktion $C : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ durch

$$C(x, y) := \mathrm{Cov}[X(x), X(y)] = \mathbb{E}\left[(X(x) - \mathbb{E}[X(x)])\,(X(y) - \mathbb{E}[X(y)])\right]$$

(b) Ein Zufallsfeld X heißt weiter Gauß'sches Zufallsfeld, falls für jede Wahl $n \in \mathbb{N}$ und $x = (x_1, \ldots, x_n) \in \mathcal{D}$ der Zufallsvektor $\hat{X} = (\hat{X}_1, \ldots, \hat{X}_n)^\top = (X(x_1), \ldots, X(x_n))$ eine (nicht ausgeartete) multivariate Normalverteilung mit Parametern $\mu(x)$ und $C(x, y)$ besitzt. Wir haben in 2.9 bereits gesehen, dass wir so gerade den Erwartungswertvektor μ und die Kovarianzmatrix C festlegen.

(c) Ein Zufallsfeld $Y : \Omega \times \mathcal{D} \to \mathbb{R}$ heißt hingegen log-normal verteilt oder kurz lognormal-Feld, falls das durch

$$\begin{aligned} X : \Omega \times \mathcal{D} &\to \mathbb{R} \\ (\omega, x) &\mapsto \log(Y(\omega, x)) \end{aligned}$$

definierte Feld ein Gauß'sches Zufallsfeld ist. Umgekehrt lässt sich jedes lognormal-Feld also durch ein Gauß'sches Zufallsfeld X erzeugen. So ist dann $\tilde{Y} : \Omega \times \mathcal{D} \to \mathbb{R}$, $(\omega, x) \mapsto \exp(X(\omega, x))$ auch ein lognormal-Feld.

3 Die Monte Carlo Methode

3.1 Herleitung und Beispiel

Wie in [25] wollen wir uns, um die Monte Carlo Methode von Grund auf einzuführen, zunächst mit der numerischen Integration beschäftigen. Grundsätzlich handelt es sich bei der Monte Carlo Methode um einen sogenannten Erwartungswertschätzer. Bevor wir also ein Problem mithilfe der Monte Carlo Methode lösen können, müssen wir die Größe, welche wir berechnen wollen, zunächst in der Form eines Erwartungswertes ausdrücken. Wir suchen dann also einen Erwartungswert $\mathbb{E}[X]$, wobei X eine Zufallsvariable, einen Zufallsvektor oder gar ein Zufallsfeld beschreiben kann. Mithilfe der Monte-Carlo-Methode können wir dann versuchen eben diesen Erwartungswert zu schätzen. Dazu müssen wir X simulieren können. Damit ist gemeint, dass wir in der Lage sein müssen eine Realisierung $(x_1, \ldots, x_n)$ von $(X_1, \ldots, X_n)$ zu generieren (oft sagt man auch in Anlehnung an das Bernoulli'sche Urnenmodell 'zu ziehen'). Dabei sollen die Zufallsgrößen $X_1, \ldots, X_n$ unabhängig sein und die gleiche Verteilung besitzen wie die Zufallsgröße X. Außerdem sei vorausgesetzt, dass der Erwartungswert $\mathbb{E}[X] < \infty$ existiert. Anschließend wird der gesuchte Erwartungswert durch

$$\mathbb{E}[X] \approx \frac{1}{n}(x_1 + \ldots x_n)$$

approximiert.

Beispiel 3.1. (Integral über $[0,1]^d$ - aus [25])
Angenommen wir wollen für $d \geq 1$ folgendes Integral berechnen:

$$I = \int_{[0,1]^d} f(u_1, \ldots, u_d)\, \mathrm{d}u_1 \ldots\, \mathrm{d}u_d$$

Wir können das Integral dann wie folgt als Erwartungswert ausdrücken: Sei $X = f(U_1, \ldots, U_d)$ ein Zufallsvektor, wobei $U_1, \ldots, U_d$ unabhängig und auf $[0,1]$ gleichverteilt sind, d.h. jedes U_i besitzt als Dichte $f_i(x) = \mathbb{1}_{[0,1]}(x)$. Dann ergibt sich so

$$I = \int_{[0,1]^d} f(u_1, \ldots, u_d)\, \mathrm{d}u_1 \ldots\, \mathrm{d}u_d = \mathbb{E}[f(U_1, \ldots, U_d)] = \mathbb{E}[X]$$

Wir haben also das Integral, welches wir berechnen wollen, als Erwartungswert ausgedrückt. Nun müssen wir die Zufallsvariable $X = f(U_1, \ldots, U_d)$ simulieren. Dazu nehmen wir an, gleichverteilte Zufallsvariablen simulieren zu können. Die Simulation solcher Zufallsvariablen spielt in der numerischen Stochastik eine ganz besondere Rolle, denn oft werden andere Verteilungen durch Transformationen auf den Fall einer Gleichverteilung auf $[0,1]$ reduziert. Sei also $(U_i)_{i\geq 1}$ eine Folge unabhängiger Zufallsvariablen mit Gleichverteilung auf $[0,1]$. Wir können dann mithilfe der simulierten Realisierungen $(u_i)_{i\geq 1}$ von

$(U_i)_{i\geq 1}$ die Zufallsvariable X wie folgt definieren: Wir setzen

$$\begin{aligned} X_1 &= f(U_1, \ldots, U_d), & x_1 &= f(u_1, \ldots, u_d) \\ X_2 &= f(U_{d+1}, \ldots, U_{2d}), & x_2 &= f(u_{d+1}, \ldots, u_{2d}) \\ X_i &= f(U_{(i-1)d+1}, \ldots, U_{id}), & x_2 &= f(u_{(i-1)d+1}, \ldots, u_{id}) \end{aligned}$$

Da $(U_i)_{i\geq 1}$ eine Folge unabhängiger Zufallsvariablen ist, erhalten wir so unter der einzigen echten Voraussetzung, dass f messbar ist, nach Blockungslemma ebenfalls eine Folge unabhängiger Zufallsvariablen $(X_i)_{i\geq 1}$. Außerdem erhalten wir so für ein großes $n \in \mathbb{N}$ eine gute Approximation von I durch:

$$I = \mathbb{E}[X] \approx \frac{1}{n}(x_1 + \cdots + x_n) = \frac{1}{n}(f(u_1, \ldots, u_d) + \cdots + f(u_{(n-1)d+1}, \ldots, u_{nd}))$$

Inbesondere haben wir keinerlei Regularität an f vorausgesetzt, es genügt bereits die bloße Messbarkeit von f .

Oft wollen wir über eine andere Grundmenge als $[0,1]^d$ integrieren. Bei endlichen Mengen, etwa einer beschränkten Borelmenge $B \subset \mathbb{R}^d$ mit $0 < |B| \coloneqq \lambda^d(B)$ (hierbei ist $\lambda^d(\cdot)$ das Borel-Lebesgue-Maß) lässt sich $I = \int_B f(x)\,\mathrm{d}x$ ähnlich wie in obigem Beispiel berechnen. Für einen Zufallsvektor U mit Gleichverteilung $U(B)$ auf B existiert nämlich der Erwartungswert $f(U)$ und es gilt:

$$\mathbb{E}[f(U)] = \int_B f(x)\frac{1}{|B|}\,\mathrm{d}x = \frac{I}{|B|}$$

Wieder simulieren wir $(U_i)_{i\geq 1}$ als Folge unabhängiger Zufallsvariablen mit identischer Verteilung zu U. Dann erhalten wir:

$$I = |B| \cdot \mathbb{E}[f(U)] \approx \frac{|B|}{n}\sum_{j=1}^{n} f(u_j)$$

Wollen wir hingegen ein Integral über $\mathbb{R}^d$ auswerten, muss es uns in der Form

$$I = \int_{\mathbb{R}^d} g(x)f(x)\,\mathrm{d}x = \int_{\mathbb{R}^d} g(x_1, \ldots, x_d)f(x_1, \ldots, x_d)\,\mathrm{d}x$$

vorliegen. Dabei sei $f(x)$ nichtnegativ und $\int_{\mathbb{R}^d} f(x)\,\mathrm{d}x = 1$. Dann lässt sich I schreiben als $I = \mathbb{E}[g(X)]$ für eine Zufallsvariable X mit Werten in $\mathbb{R}^d$ und Verteilung $f(x)\,\mathrm{d}x$. Wir können also I approximieren durch

$$I \approx \frac{1}{n}\sum_{i=1}^{n} g(x_i) \quad ,$$

wobei $(x_i)_{i\geq 1}$ Realisierungen der Zufallsvariablen $(X_i)_{i\geq 1}$ sind, welche unabhängig und identisch zu X verteilt seien.

Betrachten wir nun wieder die Monte Carlo Methode in einem etwas abstrakteren Sinne ganz allgmein. An der Stelle, an der wir letztlich die Realisierungen einer Zufallsvariable eingesetzt haben, also einen Erwartungswert durch $\mathbb{E}[X] \approx \frac{1}{n}(x_1 + \dots x_n)$ approximiert haben, haben wir stets gefordert, dass n groß ist. Es stellt sich nun die Frage, wann n groß genug ist. Wir wollen uns deshalb noch abschließend damit beschäftigen, wann und wie die Methode konvergiert und was wir über die Genauigkeit der Approximation aussagen können.

3.2 Konvergenz und Genauigkeit

Damit die Methode überhaupt in irgendeiner Weise als nützlich zu erachten ist, bedarf es Möglichkeiten, den Fehler

$$\epsilon_n = \frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X]$$

abzuschätzen. Um diesem Problem beizukommen, bedienen wir uns zweier zentraler Aussagen der Wahrscheinlichkeitstheorie. Zum einen sagt uns das starke Gesetz großer Zahlen 2.10, dass unter der Voraussetzung $\mathbb{E}[|X|] < \infty$ der Fehler ϵ_n für $n \to \infty$ für fast alle $\omega \in \Omega$ gegen 0 konvergiert. Wir erhalten also zunächst Konvergenz der Methode in einem sehr grundlegendem Sinn. Aus dem zentralen Grenzwertsatz 2.12 lassen sich zum anderen Aussagen über die Genauigkeit der Methode und letztlich somit auch der Art der Konvergenz ableiten. Nach 2.12 erhalten wir nämlich für eine u.i.v.-Folge $(X_i)_{i\in\mathbb{N}}$ mit gleicher Verteilung wie X und $\mathbb{E}[X^2] < \infty$, dass

$$\frac{\sqrt{n}}{\sqrt{\mathbb{V}[X]}}\epsilon_n = \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i - \sqrt{n}\mathbb{E}[X]}{\sqrt{\mathbb{V}[X]}} = \frac{\sum_{i=1}^{n} X_i - n\mathbb{E}[X]}{\sqrt{n\mathbb{V}[X]}} =: \hat{S}_n \xrightarrow{\mathcal{D}} \widetilde{\mathcal{N}} \text{ für } n \to \infty \ ,$$

wobei $\widetilde{\mathcal{N}}$ eine standardnormalverteilte Zufallsvariable ist. Die Wurzel der Varianz wird im Folgenden noch des Öfteren auftauchen, weswegen wir an dieser Stelle die sogenannte Standardabweichung $\sigma := \sqrt{\mathbb{V}[X]}$ einführen. Da also

$$\lim_{n\to\infty} \mathbb{P}^{\hat{S}_n}((-\infty, x]) = \mathbb{P}^{\widetilde{\mathcal{N}}}((-\infty, x])$$

gilt, ist insbesondere für $a \leq b$

$$\lim_{n\to\infty} \mathbb{P}(\frac{\sigma}{\sqrt{n}}a \leq \epsilon_n \leq \frac{\sigma}{\sqrt{n}}b) = \lim_{n\to\infty} \mathbb{P}(a \leq \hat{S}_n \leq b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \, \mathrm{d}x \ .$$

An dieser Stellen wollen wir kurz innehalten und uns überlegen, was obiges Resultat für den Fehler der Monte Carlo Methode denn praktisch gesehen bedeutet.

- Der zentrale Grenzwertsatz liefert uns kein zu der Folgerung aus dem starken

Gesetz großer Zahlen vergleichbares Resultat, denn es ist $\lim_{n\to\infty} \mathbb{P}(\epsilon_n = 0) = 0$ nach obiger Überlegung.

- Der zentrale Grenzwertsatz erlaubt uns ebenso nicht eine für andere Verfahren typische Fehlerschranke der Form $\epsilon_n \leq M_n$ für eine von n und möglicherweise anderen Faktoren, wie z.B. Ausgangsdaten, abhängigen Schranke M_n aufzustellen. Grund dafür ist, dass der Träger von $\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ ganz $\mathbb{R}$ ist.

- Was der zentrale Grenzwertsatz uns jedoch erlaubt, ist, ein sogenanntes 95% Konfidenzintervall für ϵ_n zu bestimmen. Das bedeutet, dass das tatsächliche Ergebnis mit einer Wahrscheinlichkeit von mindestens 95% im gegebenen Intervall enthalten ist. Denn, da
$$\mathbb{P}(|N| \leq 1.96) \approx 0.95$$
können wir wegen
$$\lim_{n\to\infty} \mathbb{P}(-1.96\frac{\sigma}{\sqrt{n}} \leq \epsilon_n \leq 1.96\frac{\sigma}{\sqrt{n}}) \approx 0.95 \quad (\star)$$
ein Konfidenzintervall für $\mathbb{E}[X]$ der Form
$$[\hat{\mu} - 1.96\frac{\sigma}{\sqrt{n}}, \hat{\mu} + 1.96\frac{\sigma}{\sqrt{n}}] \quad \text{für ein } \hat{\mu} \in \mathbb{R}$$
angeben. In der Praxis nehmen wir näherungsweise an, dass $(\star)$ auch für ein festes $n \in \mathbb{N}$ erfüllt ist, und entledigen uns so des Grenzwertes. Somit wird dann insbesondere die Wahl $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}(x_1, \ldots, x_n)$ gerechtfertigt.

Wir erhalten also (unter den eben erklärten Annahmen) eine Konvergenzrate des (wahrscheinlichen) Fehlers von $\frac{\sigma}{\sqrt{n}}$. Dieses Resultat mag auf den ersten Blick relativ ernüchternd wirken, allerdings existieren Fälle, in denen solch eine langsame Methode die bestmögliche ist. [25] nennt hierzu zum Beispiel Integrale in mehr als 100 Dimensionen oder besonders schwere parabolische Differentialgleichungen. Denn anders als andere Verfahren, besonders deutlich wird dies erneut auf der Ebene der Quadratur (vgl Beispiel 4.1), sind Monte Carlo Methoden nicht vom sogenannten 'Curse of dimensionality' betroffen. Während bei anderen Quadraturformeln die Anzahl der benötigten Quadraturpunkte mit der Dimension im Exponent steigt, gelten obige Resultate unabhängig von der Dimension. Wir werden später Zufallsprobleme mit sehr hohen Dimensionen betrachten, da wir in einer Bodenschicht jede Zelle als einzelne Zufallsvariable betrachten werden. Deswegen ziehen wir die Monte Carlo Methode bzw. später die Multilevel Monte Carlo Methode einem anderen Ansatz zum Lösen stochastischer partieller Differentialgleichungen, wie etwa stochastische Finite Elemente, vor. Außerdem lohnt es sich zu erwähnen, dass wir im Falle der numerischen Integration - bis auf Integrierbarkeit und Messbarkeit - keine Voraussetzungen an die Regularität der zu Funktion f gestellt haben.
Obiges Resultat legt außerdem nahe, dass es entscheidend für eine Aussage über die Konvergenz und Güte der Methode ist, die Standardabweichung σ zu kennen, oder zumindest über einen guten Schätzer für σ zu verfügen. Falls uns σ bzw. $\mathbb{V}$ nämlich sogar

exakt bekannt ist, können wir die sogenannte Chebyshev Ungleichung 2.11 ausnutzen: Da $(X_i)_{i\in\mathbb{N}}$ eine u.i.v.-Folge mit Verteilung wie X ist, gilt nämlich mit den üblichen Rechenregeln für die Varianz (zu finden z.B. in [5] auf den Seiten 778 und 779)

$$\mathbb{V}[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}[X] = \frac{\mathbb{V}[X]}{n}$$

Dann besagt die Chebychev Ungleichung für alle $t \geq 0$:

$$\mathbb{P}\left(|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X]| \geq t\right) \leq \frac{\mathbb{V}[X]}{nt^2}$$

Für uns bedeutet das insbesondere, dass für jedes $\epsilon \in (0, 1]$ die berechnete Monte-Carlo Approximation $\frac{1}{n}\sum_{i=1}^{N}$ mit einer Wahrscheinlichkeit von $1-\epsilon$ weniger als $\left(\frac{\mathbb{V}[X]}{n\epsilon}\right)^{\frac{1}{2}}$ von dem tatsächlichen Erwartungswert $\mathbb{E}[X]$ entfernt ist. In der Literatur (z.B. in [32]) finden sich einige Weiterentwicklungen der Monte Carlo Methode. Abgesehen von der Multilevel Monte Carlo Methode, welche wir in Abschnitt 4 behandeln werden, wollen wir uns hier auf die oben erklärte Standard-Variante beschränken.

4 Die Multilevel Monte Carlo Methode

4.1 Motivation und Beispiel

Nachdem wir im dritten Abschnitt die Monte Carlo Methode betrachtet haben, wollen wir uns nun einer Weiterentwicklung der Monte Carlo Methode, der sogenannten Multi Level Monte Carlo Methode zuwenden. Grundsätzlich liegt dieselbe Situation vor wie bei der Monte Carlo Methode: Wir wollen wieder eine Größe bestimmen, welche sich nach geeigneter Modellierung in der Form eines Erwartungswertes $\mathbb{E}[X]$ einer Zufallsvariablen X schreiben lässt. Besonders, wenn diese Größe mit der Lösung von gewöhnlichen oder partiellen Differentialgleichungen zusammenhängen, wie wir sie später betrachten wollen, hat man nun jedoch die Wahl, mit welcher Genauigkeit die numerische Lösung des zugrunde liegenden Problems, z.B. der Differentialgleichung, erfolgen soll. Beispielsweise können wir im Falle der numerischen Lösung von Differentialgleichung Zeitschrittweiten und/oder Gitterweiten der Ortsdiskretisierung festlegen. Wir werden dann in diesem Zusammenhang auch von verschiedenen (Genauigkeits-)Leveln sprechen. An dieser Stelle tritt stets ein typischer Zwiespalt auf:

- Zum Einen wollen wir möglichst genau rechnen. Dies legt die Wahl von besonders kleinen Zeitschrittweiten bzw. feinen Gittern zur Ortsdiskretisierung nahe.
- Zum Anderen wollen wir die Anzahl der Rechenschritte bzw. die Rechenzeit möglichst gering halten. Dies spricht hingegen für große Zeitschritte bzw. grobe Gitter.

Zusätzlich zu der oft bereits alleine anspruchsvollen Aufgabe, solche Probleme numerisch zu lösen, müssen wir also stets einen für unsere Bedürfnisse passenden Kompromiss aus möglichst genauer numerischer Approximation und geringem (oder zumindest machbarem) Rechenaufwand eingehen. Obwohl dies zunächst wie eine zusätzliche Hürde erscheint und Mehraufwand vermuten lässt, stellt sich heraus, dass eine solche Wahl der Genauigkeit im Kontext von Monte Carlo Methode sich durchaus als nützlich erweisen kann. Die Multi Level Monte Carlo Methode, wir werden im Folgenden auch oft vom sogenannten Multi Level Monte Carlo Schätzer sprechen, ist der Prototyp einer Familie sogenannter Varianz-reduzierender Methoden, welche das Ziel haben, die naive Monte Carlo Methode in Sachen Konvergenzrate und Effizienz zu schlagen. Bevor wir erklären wie genau die Multilevel Monte Carlo Methode im Allgemeinen dabei vorgeht, möchten wir die Funktionsweise wieder anhand eines Beispiels erklären, welches in [21] ausführlich beleuchtet wird.

Beispiel 4.1. (Wieder ein Integral über $[0,1]^d$)
Seien nun $d_1, d_2 \in \mathbb{N}$. Wie bereits im letzten Abschnitt setzen wir uns die Aufgabe das Integral einer Funktion f zunächst über $[0,1]^{d_1}$ zu bestimmen. Damit wir aber überhaupt in oben erklärte Situation kommen und von verschiedenen 'Leveln' sprechen können, sei f nun zusätzlich abhängig von einem Parameter $\lambda \in \Lambda \subseteq \mathbb{R}^{d_2}$, also $f : \Lambda \times [0,1]^{d_1} \to \mathbb{R}$. Um bei den folgenden Überlegungen die Notation so schlank wie möglich zu halten, betrachten wir an dieser Stelle nur einen konkreten Spezialfall:

Sei $d_1 = d_2 = 1$ und $f \in C([0,1] \times [0,1], \mathbb{R})$, d.h. wir wollen das Integral

$$I(\lambda) = \int_0^1 f(\lambda, u)\,\mathrm{d}u$$

für alle $\lambda \in \Lambda = [0,1]$ bestimmen, wir suchen also nach einer Funktion in Abhängigkeit von λ.

Monte Carlo Schätzer für $I(\lambda)$
Wollen wir an dieser Stelle einen normalen Monte Carlo Schätzer nutzen, stellt sich die Frage, wie wir mit dem zusätzlichen Parameter umgehen sollen. Die wohl naheliegendste und einfachste Idee ist, zunächst für ein festes $m \in \mathbb{N}$ ein Gitter $\{\lambda_i = \frac{i}{m}, i = 0, \dots, m\}$ festzulegen und für jedes λ_i wie im letzten Abschnitt vorzugehen und für ein $n \in \mathbb{N}$

$$I(\lambda_i) \approx \hat{I}(\lambda_i) \coloneqq \frac{1}{n} \sum_{k=1}^{n} f(\lambda_i, x_k)$$

zu schätzen. Dabei seien wieder $(x_k)_{k=1,\dots,n}$ Realisierungen von unabhängigen auf $[0,1]$ gleichverteilten Zufallsvariablen $(X_k)_{k=1,\dots,n}$. Anschließend lässt sich aus den so ermittelten Werten durch Interpolation einen Schätzer für die gesamte Funktion $I(\lambda)$ bestimmen. Grundsätzlich sind verschiedene Interpolationsansätze möglich. Für dieses grundlegende Beispiel wählen wir stückweise lineare Interpolation. Wir erhalten so für alle $\lambda \in \Lambda$:

$$I(\lambda) \approx (PI)(\lambda) = \sum_{i=0}^{m} \hat{I}(\lambda_i)\varphi_i(\lambda)$$

mit $\varphi_i \coloneqq \mathbb{1}_{\{|\lambda - \lambda_i| \le m\}}(1 - m|\lambda - \lambda_i|)$. Ein solcher Interpolationsansatz lässt sich insbesondere auf mehrdimensionale Gitter übertragen. Somit erhalten wir für $I(\lambda)$:

$$I(\lambda) \approx \mathcal{I}_{MC}(\lambda) \coloneqq \sum_{i=0}^{m} \left(\frac{1}{n} \sum_{k=1}^{n} f(\lambda_i, x_k) \right) \varphi_i(\lambda) = \frac{1}{n} \sum_{k=1}^{n} (Pf(\cdot, x_k))(\lambda)$$

Als Fehler dieser Methode können wir die mittlere quadratische Abweichung, verbunden mit einer beliebigen Norm, betrachten; wir wählen hierbei die L^2-Norm. Wir erhalten so

$$\epsilon(\mathcal{I}_{MC}) = \left(\mathbb{E}[\|I - \mathcal{I}_{MC}\|^2_{L^2([0,1])}] \right)^{\frac{1}{2}} = \left(\mathbb{E}\left[\int_0^1 |I(\lambda) - \mathcal{I}_{MC}(\lambda)|^2 \,\mathrm{d}\lambda \right] \right)^{\frac{1}{2}}$$

Ist f zusätzlich stetig differenzierbar im Parameter λ, kann gezeigt werden, dass

$$\epsilon(\mathcal{I}_{MC}) = \mathcal{O}(n^{-\frac{1}{2}} + m^{-1}) \ .$$

Gleichzeitig ist die Anzahl der arithmetischen Operationen, Funktionsaufrufe und gene-

rierter Zufallszahlen in $\mathcal{O}(mn)$. Wir sehen also, dass wir an dieser Stelle genau diesen Zwiespalt antreffen, welchen wir zuvor abstrakt beschrieben haben. Aus diesem Grund wollen wir nun einen Multilevel Monte Carlo Schätzer für $I(\lambda)$ einführen.

Multilevel Monte Carlo Schätzer für $I(\lambda)$
Wir betrachten nun eine Familie von Gittern $\{\lambda_{li} = \frac{i}{m_l} : m_l = 2^l \ , i = 0, 1, \dots m_l\}$ für $l = 0, \dots, L$. Analog zu oben führen wir zugehörige Interpolationsoperatoren

$$(P_l I)(\lambda) = \sum_{i=0}^{m_l} \hat{I}(\lambda_{li})\varphi_{li} \quad (l = 0, \dots, m)$$

ein. Wir können nun also insbesondere $P \coloneqq P_L$ also Teleskopsumme darstellen. Es gilt nämlich:

$$P = P_L = P_0 + \sum_{l=1}^{L}(P_l - P_{l-1}) \ .$$

Der Monte Carlo Schätzer von oben lässt sich (mit $P_{-1} \coloneqq 0$)dann durch

$$\mathcal{I}_{MC} = \sum_{l=0}^{L} \frac{1}{n} \sum_{k=1}^{n}(P_l - P_{l-1})f(\cdot, x_k)$$

umschreiben. Um nun tatsächlich einen Nutzen aus der Aufteilung in verschiedene Level zu ziehen und einen guten Kompromiss zwischen Kosten und Fehler herzustellen, erlauben wir nun zusätzlich die Anzahl der Zufallsauswertungen n von Level zu Level zu variieren. Wir wählen also $(n_l)_{l=0,\dots,L} \in \mathbb{N}^{L+1}$. Außerdem seien $\{X_{lj}, l = 0, \dots, L \ , j = 1, \dots, n_l\}$ unabhängige auf $[0, 1]$ gleichverteilte Zufallsvariablen und $(x_{lj})_{l=0,\dots,L \ ,j=1,\dots,n_l}$ zugehörige Realisierungen. Dann erhalten wir den Multilevel Monte Carlo Schätzer

$$I(\lambda) \approx \mathcal{I}_{MLMC}(\lambda) = \sum_{l=0}^{L} \frac{1}{n_l} \sum_{j=1}^{n_l}((P_l - P_{l-1})f(\cdot, x_{lj}))(\lambda) \ .$$

Der bedeutendste Schritt ist an dieser Stelle eine passende Wahl der n_l. Bei diesem Beispiel wollen wir uns darauf beschränken, eine passende Wahl anzugeben und den Nutzen hervorzuheben, welchen wir durch diese Wahl erlangen. So zeigt sich, dass, wie in [21] ausführlicher erklärt, eine passende Wahl beispielsweise durch $n_l = \Theta(2^{-\frac{3l}{2}}n)$ für ein $n \in \mathbb{N}$ groß genug gegeben ist. Dann kann für den analog wie für den MC-Schätzer definierten (RMSE-)Fehler gezeigt werden, dass

$$\epsilon(\mathcal{I}_{MLMC}) = \mathcal{O}(n^{-\frac{1}{2}} + n^{-\frac{1}{2}}) = \mathcal{O}(n^{-\frac{1}{2}}) \ .$$

Zugleich zeigt sich, dass die Anzahl der benötigten Rechenoperationen inklusive Funktions- und Zufallszahlauswertungen diesmal in der Komplexitätsklasse $\mathcal{O}(n)$ enthalten ist. Genaueres dazu findet sich ebenfalls in [21] (Abschnitt 2.3) . Verglichen mit der Standard (Ein-Level) Monte Carlo Methode können wir nun also eine Approximation für die ge-

samte Familie von Integralen $I(\lambda)$ mit einem Fehler von $\mathcal{O}(n^{-\frac{1}{2}})$, aber den Kosten von $\mathcal{O}(n)$ berechnen. Das ist durchaus erstaunlich, denn bereits die Kosten der Auswertung eines einzigen Integrals $I(\lambda)$ für ein festes $\lambda \in \Lambda$ liegen in $\mathcal{O}(n)$. Die MLMC Methode ist so gesehen in gewisser Weise von optimaler Ordnung, denn auch die Berechnung eines einzigen Samples auf höchstem Level benötigt $\mathcal{O}(n)$ Operationen.

Wir sehen also, dass die Multilevel Monte Carlo Methode in Situationen, in denen wir bei der Wahl von Zeitschrittweiten und/oder feinen Gittern zur Ortsdiskretisierung zwischen Anzahl an Rechenoperationen und Genauigkeit einen Kompromiss finden müssen, einen Ein-Level Ansatz, wie die Standard Monte Carlo Methode, durchaus übertreffen kann. Der Kern dieser Methode bildet dabei eine geschickte Wahl der Anzahl n_l der Zufallssamples, welche wir auf je einem Level auswerten. Wie wir in unserem Fall diese Wahl durchführen, soll an anderer Stelle in Abschnitt 6 ausführlich erläutert werden, in welchem wir die bisher zunächst beispielhaft anhand der Integration eingeführte Multilevel Monte Carlo Methode auf das probabilistische Transportproblem, welches wir in Abschnitt 5 bereits näher beleuchtet haben, übertragen werden. Mehr zu Monte Carlo und Multilevel Monte Carlo Methoden für Parameterintegrale findet sich neben [21] auch in [20].

4.2 Konvergenz und Genauigkeit

Da wir in Abschnitt 6 noch einmal ausführlich auf die Eigenschaften des Verfahrens für unsere konkrete Anwendung eingehen werden, soll dieser Unterabschnitt noch einmal etwas allgemeiner auf die stochastischen Hintergründe eingehen. Als Referenz ist hierbei das Kapitel 9.5 über Monte-Carlo Funktionen in [32] zu nennen. Betrachten wir also wieder etwas allgemeiner eine Folge unabhängiger Zufallsvariablen $Y_1, Y_2 \dots$ mit zugehörigen Realisierungen $y_1, y_2, \dots$. Diese sollen dabei alle die identische Verteilung wie eine weitere Zufallsvariable Y mit zugehöriger Dichte g_Y besitzen. Wir wollen diesmal den Erwartungswert einer Zufallsvariablen X berechnen, wobei X mithilfe einer messbaren Funktion f, die alle (unten aufgeführten) Voraussetzungen des Transformationssatzes erfülle, folgendermaßen ausgedrückt werden kann:

$$X = f(Y)$$

Wir fordern nun wieder, dass der Erwartungswert $\mathbb{E}[|X|] < \infty$ existiert, und diese Forderung ist für die Konvergenz der Methode ebenso wichtig wie scharf. Wollen wir uns überlegen, was das maßtheoretisch bedeutet, erhalten wir: Ist für $\{g_Y > 0\} \subseteq O$ die Menge O offen und $f : \mathbb{R} \to \mathbb{R}$ eine Borel-messbare Abbildung, deren Restriktion auf O stetig differenzierbar ist, eine nirgends verschwindende Funktionaldeterminante besitze und O bijektiv auf eine Menge $V \subset \mathbb{R}$ abbilde. Dann muss die Dichte g_X der Zufallsvariable X auf V integrierbar sein. Dabei gilt für die Dichte von X nach dem Transformationssatz:

$$g_X(t) := \begin{cases} \frac{g_y\left(f^{-1}(t)\right)}{|\det f'(f^{-1}(t))|} & , \text{falls } t \in V \\ 0 & , \text{sonst} \end{cases}$$

Wir nehmen nun außerdem an, dass wir über eine Hierarchie $(f_l)_{l\in\{0,\dots,L\}}$ verfügen. Dabei sei $f = f_L$ und wir bezeichnen l als Level-Parameter. Aufgrund der Linearität des Erwartungswertes kann der Erwartungswert von X dann folgendermaßen ausgedrückt werden:

$$\mathbb{E}[X] = \mathbb{E}[f(Y)] = \mathbb{E}[f_0[Y]] + \sum_{l=1}^{L} \mathbb{E}[f_l(Y) - f_{l-1}(Y)]$$

Wir können nun jeden Summanden einzeln durch einen Monte Carlo Ansatz schätzen. Dazu seien $(Y_{l,i})_{l\in\{0,\dots,L\},i\in\{1,\dots,n_l\}}$ unabhängige Zufallsvariablen aus der Folge $(Y_i)_{i\in\mathbb{N}}$. Dann gilt:

$$\mathbb{E}[X] \approx \frac{1}{n_0}\sum_{i=1}^{n_0} f_0(Y_{0,i}) + \sum_{l=1}^{L}\frac{1}{n_l}\sum_{i=1}^{n_l}(f_l(Y_{l,i}) - f_{l-1}(Y_{l,i}))$$

An dieser Stelle scheint obige Darstellung keinen wirklichen Vorteil gegenüber dem Standard Monte Carlo Schätzer zu besitzen, wir müssen aber nun beachten, dass zum Einen die Anzahl der benötigten Rechenoperation zur Berechnung von $f_l(Y_{l,i})$ unter Umständen für niedrige Werte l deutlich geringer ausfällt, als dies für größere l der Fall ist. Zum Anderen gilt für den Fehler der Monte Carlo Schätzung (vgl. Abschnitt 3), dass der Fehler des l-ten Summanden wie $\sqrt{\frac{\mathbb{V}[f_l(Y)-f_{l-1}(Y)]}{n_l}}$ konvergiert. Das heißt insbesondere, dass falls $\mathbb{V}[f_l(Y)-f_{l-1}(Y)]$ klein ausfällt, auch kleinere n_l gewählt werden können, als bei der Standard Monte Carlo Methode, für welche bei einer großen Varianz $\mathbb{V}[f(Y)] = \mathbb{V}[f_L(Y)]$ ein sehr großes $n = n_L$ für eine gute Approximation benötigt werden. Andererseits müssen wir aber auch beachten, dass wir uns damit zusätzlich zum Schätzfehler, durch die Approximation von f, auch einen Approximationsfehler einhandeln. In der Praxis, wie z.B. bei partiellen Differentialgleichungen, verfügen wir aber auch oft nicht über f selbst, sondern nur über verschieden gute Approximationen. Wir müssen also in der späteren Anwendung ganz genau prüfen, wie und ob wir aus der Multilevel Monte Carlo Methode tatsächlich einen Nutzen ziehen können. Diese Grundidee, die Varianz klein zu halten, damit die benötigte Anzahl der auszuwertenden Zufallssamples gering gehalten werden kann, ist namensgebend für die sogenannten 'Varianz reduzierenden Methoden' zur Verbesserung der Monte Carlo Methode. Weitere Details der Konvergenzanalyse finden sich in für unser Problem angepasster Form in Abschnitt 6. Bevor wir dazu kommen können, müssen wir aber im folgenden Abschnitt zunächst das zu betrachtende Problem sowie zugehörige Löser einführen, welche später die Rolle der Approximationen f_l übernehmen werden.

5 Das lineare Transportproblem

5.1 Problemstellung

5.1.1 Deterministisches Problem

Sei $\mathbb{T} = [0, T]$ ein Zeitintervall für $T > 0$ und $\mathcal{D} \subset \mathbb{R}^d, d \in \mathbb{N}$ ein beschränktes, offenes und konvexes Lipschitz-Gebiet mit Rand $\partial\mathcal{D} = \Gamma_{\mathrm{D}} \dot{\cup} \Gamma_{\mathrm{N}}$. Wie bereits in der Einleitung beschrieben, wollen wir den Transport eines Stoffes in einer porösen Bodenschicht auf Grundlage eines vorhandenen Flusses beschreiben. Als modellhaftes Problem soll uns hierfür die Regenwasserversickerung dienen: In einer porösen Bodenschicht befindet sich zum Zeitpunkt $t = 0$ ein Stoff (beispielsweise Öl) in einer gegebenen Anfangskonzentration und -verteilung. Nun sickert Regenwasser in diese poröse Bodenschicht ein. Zusätzlich wollen wir weitere Zuflüsse des Fremdstoffes über den Einflussrand $\Gamma_{\mathrm{in}} \subset \partial\mathcal{D}$ zulassen. Wir sind letztendlich an der Konzentration dieser Substanz an einer Stelle $x \in \overline{\mathcal{D}}$ zu einem Zeitpunkt $t \in \mathbb{T}$ interessiert.
Bevor allerdings die Konzentration als Lösung des Transportproblems bestimmt werden kann, muss zunächst das Flussvektorfeld $q : \overline{\mathcal{D}} \to \mathbb{R}^d$ berechnet werden.
Sei hierfür $p : D \to \mathbb{R}$ der hydrostatische Druck, $\kappa : D \to (\mathbb{R}_{\mathrm{sym}})^{d\times d}$ der Permeabilitätstensor und $G = (0, 0, p_0 g_0)^\top$. Wie bereits in der Einleitung angedeutet, kann der Fluss des Regenwassers durch das Darcy-Gesetz $q = -\kappa(\nabla p + G)$ modelliert werden. Durch $u(x) := p(x) + p_0 g_0 x_3$ vereinfacht sich das Darcy-Gesetz zu $q = -\kappa\nabla u$.
Nehmen wir die physikalische Annahme hinzu, dass der Fluss q 'quellfrei' sein soll, also an keiner Stelle Masse verschwinden oder erscheinen kann, erhalten wir das Potentialströmungsproblem:

$$\text{Bestimme } u : \overline{\mathcal{D}} \to \mathbb{R} \text{ und } q : \overline{\mathcal{D}} \to \mathbb{R}^2 \text{ mit}$$

$$\text{(PS)} \begin{cases} \operatorname{div} q = 0 & ,\text{in } \mathcal{D} \\ q = -\kappa\nabla u & ,\text{in } \mathcal{D} \\ u = u_D & ,\text{auf } \Gamma_D \\ -q \cdot n = g_N & ,\text{auf } \Gamma_N \end{cases}$$

Bemerkung. Wir wollen aus verschiedenen Gründen direkt die sogenannte gemischte Formulierung des Potentialströmungsproblem nutzen. Näheres dazu findet sich im nächsten Abschnitt.

Anschließend suchen wir die Dichteverteilung $\rho : \mathcal{D} \times \mathbb{T} \to \mathbb{R}_{\geq 0}$ einer transportierten Substanz (in unserem Modell das Öl).
Gegeben sei dazu die Anfangsverteilung $\rho_0 : \mathcal{D} \to \mathbb{R}_{\geq 0}$ und der Einfluss der Substanz über die Zeit $\rho_{\mathrm{in}} : \Gamma_{\mathrm{in}} \times \mathbb{T} \to \mathbb{R}_{\geq 0}$ mit $\Gamma_{\mathrm{in}} := \{z \in \partial\mathcal{D} : q(z) \cdot n(z) \leq 0\} \subset \partial\mathcal{D}$. Dabei ist $n(z)$ der äußere Normalenvektor im (Rand-)Punkt z. Wir bedienen uns wieder der Physik und fordern die Erfüllung der Bilanzgleichung

$$\forall K \subseteq \mathcal{D}, t \in \mathbb{T} : \frac{d}{dt} \int_K \rho(x,t)\,\mathrm{d}x + \int_{\partial K} \rho(x,t) q(x) \cdot n(x)\,\mathrm{d}a = 0.$$

Wenden wir für ein zulässiges $K \subseteq \mathcal{D}$ und $\rho, q \in C^1(\mathcal{D})$ den Satz von Gauß, an erhalten wir

$$\int_K \partial_t \rho(x,t) + \operatorname{div}(\rho q)(x,t)\,\mathrm{d}x = 0$$

und können so die lineare Transportgleichung ableiten:

$$\partial_t \rho + \operatorname{div}(\rho q) = 0 \text{ in } \mathcal{D} \times (0,T]$$

Mit den entsprechenden Rand- und Anfangswerten erhalten wir so:

$$\text{Bestimme } \rho : \overline{\mathcal{D}} \times \mathbb{T} \to \mathbb{R}_{\geq 0}, \text{ sodass}$$

$$\text{(TP)} \begin{cases} \partial_t \rho + \operatorname{div}(\rho q) = 0 & \text{,in} \quad \mathcal{D} \times (0,T) \\ \rho(x,t) = \rho_{\text{in}}(x,t) & \text{,auf} \quad \Gamma_{\text{in}} \times (0,T) \\ \rho(x,0) = \rho_0(x) & \text{,auf} \quad \mathcal{D} \end{cases}$$

5.1.2 Probabilistisches Problem

In dem letzten Unterabschnitt sind wir bereits bei der Lösung des Potentialströmungsproblems davon ausgegangen, sämtliche benötigten Randwerte sowie den Permeabilitätstensor κ exakt für das gesamte Gebiet $\mathcal{D}$ zu kennen. Wir wollen uns von dieser durchaus starken Annahme lösen und deshalb zusätzlich die Permeabilität κ mit Mitteln der Stochastik modellieren. Sei dazu $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und ab nun $d = 2$, also $\mathcal{D} \subseteq \mathbb{R}^2$.

Bemerkung. Grundsätzlich funktionieren die vorgestellten Verfahren auch für $d = 3$, wir wollen uns aber der Anschaulichkeit halber auf zwei Dimensionen beschränken. Das so betrachtete Gebiet $\mathcal{D}$ lässt sich so z.B. als Querschnitt einer Bodenschicht interpretieren.

Weiter sei nun $\kappa(\cdot, x) : \Omega \to \mathbb{R}_{\geq 0}$ die (vom Zufall abhängige) Permeabilität. Wie schon an anderer Stelle (z.B. in [24]) wollen wir die Permeabilität als lognormal-Feld modellieren. Unser so entstehendes Problem fällt somit in den Bereich der Uncertainty Quantification und ist gegeben durch:

Für jedes benötigte $\omega \in \Omega$, bestimme $u(\omega, \cdot) : \overline{\mathcal{D}} \to \mathbb{R}$ und $q(\omega, \cdot) : \overline{\mathcal{D}} \to \mathbb{R}^2$ mit

$$\text{(PS)} \begin{cases} \operatorname{div}(q(\omega, x)) &= 0 & \text{, für } x \in \mathbb{D} \\ q(\omega, x) &= -\kappa(\omega)\nabla u(\omega, x) & \text{, für } x \in \mathcal{D} \\ -q(\omega, x) \cdot n &= g_N(x) & \text{, für } x \in \Gamma_N \\ u(\omega, x) &= u_D(x) & \text{, für } x \in \Gamma_D \end{cases}$$

Für $q : \Omega \times \overline{\mathcal{D}} \to \mathbb{R}^2$, bestimme $\rho : \Omega \times \overline{\mathcal{D}} \times \mathbb{T} \to \mathbb{R}_{\geq 0}$ mit

$$\text{(TP)} \begin{cases} \partial_t \rho(\omega, x, t) + \operatorname{div}(\rho(\omega, x, t) q(\omega, x)) &= 0 & \text{, für } (x, t) \in \mathcal{D} \times (0, T] \\ \rho(\omega, x, t) &= \rho_{\text{in}}(x, t) & \text{, für } (x, t) \in \Gamma_{\text{in}} \times \mathbb{T} \\ \rho(\omega, x, 0) &= \rho_0(x) & \text{, für } x \in \mathcal{D} \end{cases}$$

für die Anfangs- und Randwerte:

$g_N : \Gamma_{\text{N}} \to \mathbb{R}$
$u_D : \Gamma_{\text{D}} \to \mathbb{R}$
$\rho_{\text{in}} : \Gamma_{\text{in}} \times \mathbb{T} \to \mathbb{R}_{\geq 0}$
$\rho_0 : \mathcal{D} \to \mathbb{R}_{\geq 0}$
wobei $\partial\mathcal{D} = \Gamma_{\text{D}} \dot{\cup} \Gamma_{\text{N}}$ und $\Gamma_{\text{in}} \coloneqq \{z \in \partial\mathcal{D} : q(z) \cdot n(z) \leq 0\} \subset \partial\mathcal{D}$

Dabei stellen wir uns die Aufgabe, den Erwartungswert eines gegebenen Zielfunktionals $Q(\rho(\omega))$ zu berechnen, etwa dem Ausfluss der transportierten Substanz über den Rand. An dieser Stelle können wir dann, nachdem wir uns in den nächsten zwei Unterabschnitten damit beschäftigt haben, wie wir obige Probleme numerisch lösen, die MLMC Methode nutzen, um diesen Erwartungswert zu berechnen.

5.2 Numerische Lösung des Potentialströmungsproblem

Bemerkung. Die beiden folgenden Abschnitte bauen im Wesentlichen auf den beiden Vorlesungen 'Einführung in das Wissenschaftliche Rechnen' (SS 2019) und 'Finite Elemente Methoden' (WS 2019/2020) von Herrn Prof. Dr. Wieners auf. Dem entsprechend sind als Quellen neben [3], [2] und [18] vor allem die Mitschriebe zu den oben genannten Vorlesungen, sowie die Berichte zum Rechnerpraktikum mit M++ [1] zu nennen.

Wie bereits in obigem Abschnitt erwähnt, sollen sich die nächsten beiden Abschnitte damit beschäftigen, wie wir die oben beschriebenen Probleme für ein festes $\omega \in \Omega$ numerisch lösen können. Wir wollen dabei im Folgenden auf eine Möglichkeit eingehen, diese Berechnung numerisch durchzuführen. Insbesondere werden dabei jene Verfahren beschrieben, welche wir auch später innerhalb der MLMC Methode in M++ nutzen wollen. Da wir in diesen beiden Abschnitten $\omega \in \Omega$ ohnehin festhalten, genügt es zudem das deterministische Problem zu betrachten.
Sowohl das hybride Finite Elemente Verfahren, welches wir zur Lösung des Potentialströmungsproblem nutzen wollen, als auch das Discontinuous Galerkin Vefahren, mit dessen Hilfe wir das Transportproblem lösen wollen, bauen auf der Finite Elemente Theorie auf. Diese ist im Wesentlichen in der zweiten Hälfte des 20. Jahrhunderts entstanden, ist aber bis heute in praktischer wie auch in theoretischer Sicht aktuell. Die Grundidee ist hierbei, die vorliegenden Rand-Anfangswertaufgaben in einem passenden endlichen Unterraum zu lösen. Dabei löst man sich auf analytischer Seite zunächst oft von einzelnen Regularitäts- und Differenzierbarkeitsbedingungen und führt einen sogenannten schwachen Lösungsbegriff ein (vergleiche Abschnitt 2.1). Statt nun aber solch eine schwache Lösung in einem unendlich dimensionalen Funktionenraum, wie beispielsweise in den Sobolevräumen $H^1(\mathcal{D})$ oder $H_0^1(\mathcal{D})$ zu bestimmen, zieht man sich auf endlich dimensionale Unterräume zurück.
Die folgende Definition entstammt [3] und geht ursprünglich (1978) auf Ciarlet zurück.

Definition 5.1. Sei

- $K \subseteq \mathbb{R}^d$ eine beschränkte abgeschlossene Menge mit einem nichtleeren Inneren und stückweise stetig differenzierbarem Rand
- $\mathcal{P}$ ein endlich dimensionaler Funktionenraum auf K
- $\Xi = \{\eta_1, \eta_2, \ldots, \eta_k\}$ eine Basis für $\mathcal{P}'$

Dann heißt $(K, \mathcal{P}, \Xi)$ ein finites Element.

Wir wollen im Folgenden diese theoretische Definition zwar im Hinterkopf behalten, aber wie in [2] meist nur mit den sogenannten Finite-Elemente-Räumen arbeiten. Dabei wird eine geeignete Zerlegung $\mathcal{T} = \{K_1, K_2, \ldots, K_M\}$ von $\mathcal{D}$ in endlich viele Teilgebiete gewählt. Anschließend betrachten wir einen endlichen Raum von Funktionen, die eingeschränkt auf diese Teilgebiete von einfacher Gestalt sind, beispielsweise bieten sich oft polynomielle Darstellungen niedrigen Grades an. Ein solches Teilgebiet $K \in \mathcal{T}$ nennen

wir Finites Element oder auch Zelle und fordern implizit, verbunden mit dem betrachteten Funktionenraum, die Erfüllung der obigen Definition.
Im Falle $\mathcal{D} \subseteq \mathbb{R}^2$ kommen so z.B. Dreiecke oder Vierecke in Frage, in $\mathcal{D} \subseteq \mathbb{R}^3$ können Tetraeder, Würfel, Quader und andere genutzt werden.
Sei nun $\mathcal{D} \subseteq \mathbb{R}^2$ zudem ein polygonales Gebiet, um eine einfache Zerlegung in Dreiecke oder Vierecke zu gewährleisten.

Definition 5.2. 1. Eine Zerlegung $\mathcal{T} = \{K_1, K_2, \dots, K_M\}$ von $\mathcal{D}$ in Dreiecks- oder Viereckselemente heißt zulässig, wenn folgende Eigenschaften erfüllt sind:

- $\overline{\mathcal{D}} = \bigcup_{i=1}^{M} K_i$
- Für $i \neq j$ ist $K_i \cap K_j$
  a) ein gemeinsamer Eckpunkt von K_i und K_j
  b) eine gemeinsame Kante von K_i als auch von K_j
  c) oder $K_i \cap K_j = \emptyset$

2. Wir schreiben oft $\mathcal{T}_h$ anstatt $\mathcal{T}$, wenn jedes Element einen Durchmesser von höchstens h besitzt .

3. Eine Familie von Zerlegungen $\{\mathcal{T}_h\}$ heißt uniform, wenn ein $\delta > 0$ existiert, sodass jedes $K \in \mathcal{T}_h$ einen Kreis mit Radius r_K enthält mit $r_K \geq \frac{h}{\delta}$.

Abbildung 1: Zulässige Zerlegung und unzulässige Zerlegung mit hängendem Knoten

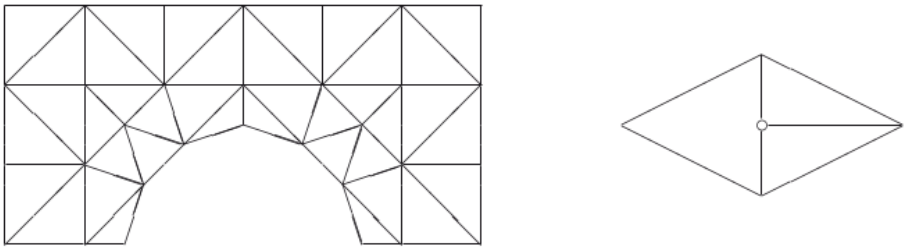


Abbildung aus [2] Seite 58

Wir werden außerdem im Laufe der Thesis dazu übergehen, ähnlich wie bereits im Abschnitt über die Multilevel Monte Carlo Methode auch bei Zerlegungen von 'Leveln' zu sprechen. Dabei betrachten wir stets eine uniforme Familie zulässiger Zerlegungen $\{\mathcal{T}_h\}_{h \in \mathcal{H}}$ und fordern dabei, dass die Indexmenge $\mathcal{H}$ eine ganz bestimmte Form hat. Genauer soll

$$\mathcal{H} = \{h_0, h_1 \coloneqq \frac{h_0}{2}, h_2 \coloneqq \frac{h_1}{2} = \frac{h_0}{4}, \dots\} \text{ für ein } h_0 > 0$$

gelten. Insbesondere gelte also $\overline{\mathcal{H}} \ni 0$. Sprechen wir dann von Level i meinen wir damit die Zerlegung $\mathcal{T}_{h_i} \in \{\mathcal{T}_h\}$. Zudem führen wir für alle Zerlegungen folgende Bezeichnungen ein:

- ein $K \in \mathcal{T}$ nennen wir Zelle
- ein $z \in \mathcal{V}_K \coloneqq \{z_{K,0}, z_{K,1}, z_{K,2}, z_{K,3}\} \subset \mathbb{R}^2$ nennen wir Knoten und $\mathcal{V}_K$ die Menge der Knoten von K
- $\mathcal{V}_\mathcal{T} \coloneqq \bigcup_{K\in\mathcal{T}} \mathcal{V}_K$ sei die Menge aller Knoten
- $\mathcal{F} \coloneqq (\{\partial K_1 \cap \partial K_2 : K_1, K_2 \in \mathcal{T}\} \cup \{\partial K_1 \cap \partial\mathcal{D} : K_1 \in \mathcal{T}\}) \setminus \{\emptyset\}$ sei die Menge aller Seiten
- $\mathcal{F}_K \coloneqq (\{\partial K \cap \partial K' : K' \in \mathcal{T}\} \cup \{\partial K \cap \partial\mathcal{D}\}) \setminus \{\emptyset\}$ sei die Menge aller Seiten von K
- $\partial\mathcal{D}_h \coloneqq \bigcup_{F\in\mathcal{F}} F$ sei der Rand von $\mathcal{D}_h$.

5.2.1 Schwache Formulierung

Betrachten wir also die deterministische Version des Potentialströmungsproblem:

$$\text{Bestimme } u : \overline{\mathcal{D}} \to \mathbb{R} \text{ und } q : \overline{\mathcal{D}} \to \mathbb{R}^2 \text{ mit}$$

$$\text{(PS)} \begin{cases} \operatorname{div} q = 0 & \text{,in } \mathcal{D} \quad (1) \\ q = -\kappa\nabla u & \text{,in } \mathcal{D} \quad (2) \\ u = u_D & \text{,auf } \Gamma_D \\ -q\cdot n = g_N & \text{,auf } \Gamma_N \end{cases}$$

Satz 2.6 sagt uns , dass wir in obiger Formulierung Gleichung (1) mit Testfunktionen $\phi \in H^1(\mathcal{D})$ und Gleichung (2) mit Testfunktionen $\psi \in H^1(\operatorname{div}, \mathcal{D})$ multiplizieren und anschließend über $\mathcal{D}$ integrieren können und so eine äquivalente schwache Formulierung herleiten:

$$\int_\mathcal{D} \operatorname{div}(q)\phi \,\mathrm{d}x = 0 \text{ für alle Testfunktionen } \phi : \mathcal{D} \to \mathbb{R}$$

$$\int_\mathcal{D} (q + \kappa\nabla u)\cdot\psi \,\mathrm{d}x = 0 \text{ für alle Testfunktionen } \psi : \mathcal{D} \to \mathbb{R}^2$$

Da κ weiter symmetrisch positiv definit ist, lässt sich letztere Gleichung zu

$$\begin{aligned} & \int_\mathcal{D} \kappa^{-1}(q + \kappa\nabla u)\cdot\psi \,\mathrm{d}x = 0 \\ \Leftrightarrow \qquad & \int_\mathcal{D} \nabla u\cdot\psi \,\mathrm{d}x = -\int_\mathcal{D} (\kappa^{-1}q)\cdot\psi \,\mathrm{d}x \qquad (\star) \end{aligned}$$

umformen. Außerdem wollen wir nun noch die Dirichlet-Randbedingungen $u = u_D$ auf Γ_D einfließen lassen. Dazu verwenden wir den Satz von Gauß:

$$\int_{\partial\Omega} (u\psi)\cdot n \,\mathrm{d}a \overset{\text{Gauß}}{=} \int_\Omega \operatorname{div}(u\psi)\,\mathrm{d}x = \int_\Omega \nabla u\cdot\psi \,\mathrm{d}x + \int_\Omega u\operatorname{div}(\psi)\,\mathrm{d}x \quad (\psi : \Omega \to \mathbb{R}^2)$$

Wählen wir nun unseren Ansatzraum so, dass für die Funktion ψ gilt $\psi \cdot n = 0$ auf Γ_N. Damit folgt

$$\int_{\Gamma_D} (u_D\psi)\cdot n \,\mathrm{d}a \overset{\substack{\psi\cdot n|_{\Gamma_N}=0 \\ u|_{\Gamma_D}=u_D}}{=} \int_{\partial\Omega} (u\psi)\cdot n \,\mathrm{d}a = \underbrace{\int_\Omega \nabla u\cdot\psi \,\mathrm{d}x}_{\overset{(\star)}{=} -\int_\Omega (\kappa^{-1}q)\cdot\psi \,\mathrm{d}x} + \int_\Omega u \,\mathrm{div}(\psi)\,\mathrm{d}x.$$

Die Neumann-Randbedingung $(\kappa\nabla u)\cdot n = g_N$ auf Γ_N wird durch die Wahl des Lösungsraumes erfüllt.

Wir erhalten so folgende schwache Formulierung:

Bestimme (q,u) mit $q\cdot n = -g_N$ auf Γ_N und

$$\text{(sPS)} \begin{cases} \int_{\mathcal{D}} \kappa^{-1}q\cdot\psi \,\mathrm{d}x - \int_{\mathcal{D}} u \,\mathrm{div}(\psi)\,\mathrm{d}x &= -\int_{\Gamma_D}(u_D\psi)\cdot n\,\mathrm{d}a \\ \int_{\mathcal{D}} \mathrm{div}(q)\,\phi\,\mathrm{d}x &= 0 \end{cases}$$

für alle (ψ,ϕ) in einem geeigneten Testraum mit $\psi\cdot n = 0$ auf Γ_N

5.2.2 Diskretisierung

Sei $\widetilde{\mathcal{T}}$ eine zulässige Zerlegung von $\mathcal{D}$ und alle Bezeichnungen wie oben. Wir nummerieren zunächst die Zellen und die Seiten durch:

$$\begin{aligned} \mathcal{F} &= \{F_1,\dots,F_{|\mathcal{F}|}\} && \text{globale Seitennummerierung} \\ \mathcal{T} &= \{K_1,\dots,K_{|\widetilde{\mathcal{T}}|}\} && \text{globale Zellennummerierung} \end{aligned}$$

Dabei sei im Weiteren $N \coloneqq |\mathcal{F}|$ die Anzahl der Seiten und $|\mathcal{T}|$ die Anzahl der Zellen. Als Nächstes soll es nun Ziel sein, eine Lösung der im letzten Abschnitt erklärten schwachen Formulierung in einem endlich dimensionalen Finite Elemente Ansatzraum zu bestimmen. Um aber hierfür genau diese Räume definieren zu können, benötigen wir zuerst sogenannte Basisfunktionen, genauer die Seiten- und die Zellenbasis.

Definition 5.3. (Seiten- und Zellenbasis)

(a) $\{\psi_i\}_{i=1}^N$ heißt Seitenbasis und ist definiert durch

$$\forall i,j\in\{1,\dots,N\} : \int_{F_j} \psi_i\cdot n^K \,\mathrm{d}a = \pm\delta_{i,j} \text{ und } \psi_i|_K \in \mathbb{P}_1(K,\mathbb{R}^2)\cap C(\overline{\mathcal{D}}) \ (K\in\mathcal{T})$$

(b) $\{\mu_i\}_{i=1}^{|\mathcal{T}|}$ heißt Zellenbasis und ist gegeben durch

$$\forall i\in\{1,\dots,|\mathcal{T}|\} : \mu_i \coloneqq \mathbb{1}_{K_i}.$$

Anschließend können wir mithilfe dieser Basisfunktionen die Testräume bzw. Finite Elemente Räume definieren:

Definition 5.4. (Ansatzräume)

(a) $W_h \coloneqq \operatorname{span}\{\psi_1, \dots, \psi_N\}$ (Seitenansatzraum/ Raum für ψ und q_h)

(b) $W_h(g) \coloneqq \{\psi_h \in W_h : \int_F \psi_h \cdot n \,\mathrm{d}a = \int_F g \,\mathrm{d}a \;\text{ für alle } F \subseteq \Gamma_{\mathrm{N}})\}$

(c) $\mathcal{Q}_h \coloneqq \operatorname{span}\{\mu_1, \dots, \mu_{|\mathcal{T}|}\}$ (Zellenansatzraum/ Raum für ϕ und u_h)

Zusammen mit der schwachen Formulierung (5.2.1) erhalten wir so das nun diskretisierte Problem:

$$\begin{aligned}&\text{Bestimme } (q_h, u_h) \in W_h(-g_N) \times \mathcal{Q}_h \text{ mit}\\ &\begin{cases} \int_\Omega \kappa^{-1} q_h \cdot \psi_h \,\mathrm{d}x - \int_\Omega u_h \operatorname{div}(\psi_h) \,\mathrm{d}x &= -\int_{\Gamma_D} (u_D \psi_h) \cdot n \,\mathrm{d}a \\ \int_\Omega \operatorname{div}(q_h)\, \phi_h \,\mathrm{d}x &= 0 \end{cases}\\ &\text{für alle } (\psi_h, \phi_h) \in W_h(0) \times \mathcal{Q}_h\end{aligned}$$

5.3 Formulierung als LGS

Wir können nun damit beginnen, das so entstandene endlich dimensionale Problem in ein Lineares Gleichungs System umzuformulieren. Dazu definieren wir:

$$\begin{aligned}\underline{A} \in \mathbb{R}^{N\times N} \text{ mit } \underline{A}[n,k] &\coloneqq \int_\Omega \kappa^{-1} \psi_n \cdot \psi_k \,\mathrm{d}x\\ \underline{B} \in \mathbb{R}^{|\mathcal{T}|\times N} \text{ mit } \underline{B}[m,k] &\coloneqq -\int_\Omega \mu_m \operatorname{div}(\psi_k) \,\mathrm{d}x\\ \underline{b} \in \mathbb{R}^{N} \text{ mit } \underline{b}[k] &\coloneqq -\int_{\Gamma_D} u_D \psi_k \cdot n \,\mathrm{d}a\end{aligned}$$

und (für die Randbedingungen)

$$\underline{W}(g) \coloneqq \left\{ \underline{q} \in \mathbb{R}^N : \underline{q}[k] = \int_{F_k} g \,\mathrm{d}a \;(\text{für } k \text{ mit } F_k \subseteq \Gamma_N) \right\}$$

Unser zu lösendes Problem lässt sich so mit $q_h = \sum_{n=1}^{N} \underline{q}[n]\psi_n$ und $u_h = \sum_{m=1}^{M} \underline{u}[m]\mu_m$ umformen zu

$$\begin{aligned}&\text{Bestimme}(\underline{q}, \underline{u}) \in \underline{W}(-g_N) \times \mathbb{R}^{|\mathcal{T}|} \text{ mit}\\ &\begin{cases} \underline{A}\underline{q} + \underline{B}^T \underline{u} &= \underline{b} \\ \underline{B}\underline{q} &= 0 \end{cases}\end{aligned}$$

oder anders geschrieben

$$\text{Bestimme}(\underline{q}, \underline{u}) \in \underline{W}(-g_N) \times \mathbb{R}^{|\mathcal{T}|} \text{ mit}$$
$$\left\{ \begin{pmatrix} \underline{A} & \underline{B}^T \\ \underline{B} & 0 \end{pmatrix} \begin{pmatrix} \underline{q} \\ \underline{u} \end{pmatrix} = \begin{pmatrix} \underline{b} \\ 0 \end{pmatrix}. \right.$$

Wir haben so eine diskrete gemischte Formulierung des Potentialströmungsproblems hergeleitet und können mit dieser aus gegebenen Rand- und Anfangswerten ein Flussvektorfeld q erzeugen, welches der obigen Differentialgleichung genügt. Es handelt sich hierbei um das gemischte Finite Elemente Verfahren. In M++ selbst lösen wir das Potentialströmungsproblem durch eine Abwandlung dieses Verfahrens. Wir diskretisieren dazu eine äquivalente Formulierung von (sPS) und erhalten so mit dem hybriden Finite Elemente Verfahren die gleichen Ergebnisse, die auch der vorgestellte gemischte Ansatz liefern würde, bei besserer Effizienz und guter Parallelisierbarkeit. Da das Potentialströumgsproblem in dieser Thesis primär dazu genutzt werden soll, das Vektorfeld q zu bestimmen, soll uns aus theoretischer Sicht aber obige Formulierung genügen und wir verweisen hinsichtlich der Lösung mit hybriden gemischten Finiten Elementen, neben einem kleinen, Überblick verschaffendem Abschnitt im Appendix 10.2, auf die Literatur, wie etwa [4] oder [30].

5.4 Numerische Lösung des Transportproblems

In diesem Abchnitt soll nun, nachdem wir $q(\omega, \cdot)$ als Finite-Elemente-Lösung des Potentialströmungsproblems erhalten haben, die numerische Lösung des linearen Transportproblems behandelt werden:

Für $\omega \in \Omega$ und $q(\omega, \cdot) : \overline{\mathcal{D}} \to \mathbb{R}^2$, bestimme $\rho(\omega, \cdot) : \overline{\mathcal{D}} \times \mathbb{T} \to \mathbb{R}_{\geq 0}$ mit

$$\text{(pTP)} \begin{cases} \partial_t \rho(\omega, x, t) + \operatorname{div}(\rho(\omega, x, t) q(\omega, x)) = 0 & \text{, für } (x,t) \in \mathcal{D} \times (0, T] \\ \rho(\omega, x, t) = \rho_{\text{in}}(x, t) & \text{, für } (x,t) \in \Gamma_{\text{in}} \times \mathbb{T} \\ \rho(\omega, x, 0) = \rho_0(x) & \text{, für } x \in \mathcal{D} \end{cases}$$

Insbesondere wollen wir an dieser Stelle wieder $\omega \in \Omega$ festhalten und betrachten deshalb zunächst nur das deterministische Problem wie in 5.1.1:

Bestimme $\rho : \overline{\mathcal{D}} \times \mathbb{T} \to \mathbb{R}_{\geq 0}$, sodass

$$\text{(dTP)} \begin{cases} \partial_t \rho(x, t) + \operatorname{div}(\rho(x, t) q(x)) = 0 & \text{,in } \mathcal{D} \times (0, T) \\ \rho(x, t) = \rho_{\text{in}}(x, t) & \text{,auf } \Gamma_{\text{in}} \times (0, T) \\ \rho(x, 0) = \rho_0(x) & \text{,auf } \mathcal{D} \end{cases}$$

Wir greifen dabei auf ein sogenanntes discontinuous Galerkin Verfahren zurück, welches für diese Problemklasse bereits an anderen Stellen (z.B. in [11]) erprobt wurde. Ursprünglich geht das Discontinuous Galerkin Verfahren auf Reed und Hill [28] zurück. Einen guten (wenn auch mittlerweile etwas in die Jahre gekommenen) Überblick über die Anwendung von discontinuous Galerkin Verfahren bietet [9]. Grundätzlich handelt es sich beim discontinuous Galerkin Verfahren ebenfalls um einen FEM Ansatz, der zwar Ähnlichkeiten zum Finite Elemente Verfahren aufweist, welches wir im letzten Abschnitt gesehen hatten, aber auch einige bedeutende Unterschiede besitzt, auf welche wir im Folgenden besonders eingehen wollen. Anders als zuvor das Potentialströmungsproblem ist die lineare Transportgleichung nämlich sowohl orts- als auch zeitabhängig. Daher werden wir die lineare Transportgleichung zunächst im Ort diskretisieren. Wir erhalten so eine Semidiskretisierung, welche wir anschließend mit einem Zeitintegrator, wie beispielsweise der impliziten Mittelpunktsregel, in eine Volldiskretisierung überführen.

5.4.1 Diskretisierung

Wie bereits weiter oben beschrieben, werden wir im Folgenden zunächst den Raum diskretisieren und anschließend die so entstandene Semidiskretisierung in eine Volldiskretisierung auflösen. Insgesamt wollen wir das discontinuous Galerkin Verfahren mit einem Zeitintegrator, wie der impliziten Mittelpunktsregel oder einem klassischen Runge-Kutta-Verfahren nutzen. Zunächst führen wir die analytische Flussfunktion ein.

Definition 5.5. (Flussfunkion)
Zu einem gegebenen Flussvektorfeld $q : \mathcal{D} \to \mathbb{R}^2$ ist die Flussfunktion Υ definiert als:

$$\Upsilon : \mathrm{Abb}(\mathcal{D} \times \mathbb{T}, \mathbb{R}) \to \mathrm{Abb}(\mathcal{D} \times \mathbb{T}, \mathbb{R}^2)$$
$$\rho \mapsto \rho q$$

Für eine klassische Lösung ρ von (dTP) gilt dann insbesondere $\partial_t \rho = -\operatorname{div}(\Upsilon(\rho))$ auf $\mathcal{D} \times (0, T]$.
Halten wir also zunächst $t \in \mathbb{T}$ und leiten so die Semidiskretisierung her.
Sei nun $\mathcal{T}$ eine zulässige Triangulierung von $\mathcal{D}$ aus Dreiecken wie in 5.2 und $(\cdot, \cdot)_A$ das $L^2(A)-$Skalarprodukt. Wir wählen als Lösungs-/Testraum $\mathcal{Q}_h = \prod_{K \in \mathcal{T}} \mathbb{P}_p(K, \mathbb{R})$ für ein festes $p \geq 0$. Anders als zuvor fordern wir für unsere Lösungs- und Testfunktionen diesmal aber nicht die Stetigkeit auf $\mathcal{D}$. Da so $\mathcal{Q}_h$ nicht im betrachteten analytischen Lösungs- und Testraum liegt, etwa $\mathcal{Q}_h \nsubseteq H^1(\mathcal{D})$, nennt man $\mathcal{Q}_h$ auch einen nichtkonformen Ansatzraum. Außerdem lässt sich im Allgemeinen auch die später bestimmte Lösung $\rho_h \in \mathcal{Q}_h$ (definiert auf $\mathcal{D}_h = \bigcup_{K \in \mathcal{T}} K$) nicht stetig auf $\mathcal{D}$ fortsetzen, denn für eine beliebige innere Kante F kann der Grenzwert von ρ_h auf den anliegenden Zellen K, K' ($\overline{F} = \partial K \cap \partial K'$) unterschiedlich sein.
Trotzdem müssen wir auch auf den inneren Kanten $\mathcal{F}^0 \subset \mathcal{F}$ festlegen, welcher Grenzwert in einem solchen Falle gewählt wird.
Dazu führen wir als Pendant zur analytischen Flussfunktion (vgl. 5.5) auch eine numerische Flussfunktion ein. Grundsätzlich kommen mehrere solche Flussfunktionen in Frage, welche direkten Einfluss auf Eigenschaften des entstehenden Verfahrens besitzen. Wir entscheiden uns an dieser Stelle für den weit verbreiteten sogenannten upwind flux:

Definition 5.6. (upwind flux)
Sei $K \in \mathcal{T}$ eine beliebige Zelle und $F \in \mathcal{F}_K$ eine Kante von K. Dann ist

$$\Upsilon^\star : \mathrm{Abb}(\mathcal{D} \times \mathbb{T}, \mathbb{R}) \to \mathrm{Abb}(\mathcal{D} \times \mathbb{T}, \mathbb{R}^2)$$
$$\rho_h \mapsto \begin{cases} \Upsilon(\rho_h|_K), & \text{für } q \cdot n_F^K \geq 0 \\ \Upsilon(\rho_h|_{K'}), & \text{für } q \cdot n_F^K < 0 \text{ und } \overline{F} = \partial K \cap \partial K' \end{cases}$$

Sei nun also ρ klassische Lösung von (dTP) mit $\partial_t \rho = -\operatorname{div}(\Upsilon(\rho))$ auf $\mathcal{D}$. Dann gilt nach Satz von Gauß:

$$\int_{\partial \mathcal{D}} \rho q \cdot n \phi \, \mathrm{d}a = \int_{\partial \mathcal{D}} \Upsilon(\rho) \cdot n \phi \, \mathrm{d}a = \int_{\mathcal{D}} \operatorname{div}(\Upsilon(\rho)\phi) \, \mathrm{d}x \tag{5.1}$$

Das Integral über den Rand von $\mathcal{D}$ können wir nach der folgenden kleinen Vorüberlegung auch als Integral über alle Kanten der gewählten Zerlegung $\mathcal{T}$ ausdrücken:
Es gilt nämlich für alle inneren Kanten, also solche Kanten F, für die zwei Zellen K und K' existieren, sodass $\overline{F} = K \cap K'$ ist, dass $\int_F \Upsilon^\star(\rho) \cdot n^K \phi \, \mathrm{d}a = -\int_F \Upsilon^\star(\rho) \cdot n^{K'} \phi \, \mathrm{d}a$ stets erhalten ist.
Summieren wir also zunächst über alle Zellen, summieren anschließend die Integrale über alle Kanten und ersetzen dabei den analytischen durch den numerischen Fluss, erhalten

wir gerade wieder obiges Randintegral. Es gilt also:

$$\sum_{K\in\mathcal{T}}\sum_{F\in\mathcal{F}_K}\int_F \Upsilon^\star(\rho)\cdot n^K\phi\,\mathrm{d}a = \int_{\partial\mathcal{D}}\Upsilon(\rho)\cdot n\phi\,\mathrm{d}a \overset{5.1}{=} \int_{\mathcal{D}}\operatorname{div}(\Upsilon(\rho)\phi)\,\mathrm{d}x$$

Nach der Produktregel der Divergenz lässt sich das letzte Integral auswerten zu:

$$\int_{\mathcal{D}}\operatorname{div}(\Upsilon(\rho)\phi)\,\mathrm{d}x = \int_{\mathcal{D}}\phi\operatorname{div}(\Upsilon(\rho)) + \Upsilon(\rho)\cdot\nabla\phi\,\mathrm{d}x \overset{\text{Vor. an }\rho}{=} -\int_{\mathcal{D}}\partial_t\rho\phi\,\mathrm{d}x + \int_{\mathcal{D}}\Upsilon(\rho)\cdot\nabla\phi\,\mathrm{d}x$$

Durch Umstellen und das Zusammenfassen der obigen Resultate erhalten wir so:

$$\sum_{K\in\mathcal{T}}\int_K\partial_t\rho\phi\,\mathrm{d}x = \sum_{K\in\mathcal{T}}\int_K\Upsilon(\rho)\cdot\nabla\phi\,\mathrm{d}x - \sum_{K\in\mathcal{T}}\sum_{F\in\mathcal{F}_K}\int_F\Upsilon^\star(\rho)\cdot n^K\phi\,\mathrm{d}a$$

Dabei wurde zusätzlich ausgenutzt, dass es sich bei den Kanten um Nullmengen handelt und wir so das Integral über $\mathcal{D}$ als Summe der Integrale über alle Zellen auffassen können. Nutzen wir nun noch aus, dass für den Fluss $\rho(x,t) = \rho_{\text{in}}(x,t)$ für $x\in\Gamma_{\text{in}}$ gilt, kommen wir so auf

$$\sum_{K\in\mathcal{T}}\int_K\partial_t\rho\phi\,\mathrm{d}x = \sum_{K\in\mathcal{T}}\int_K\Upsilon(\rho)\cdot\nabla\phi\,\mathrm{d}x - \sum_{K\in\mathcal{T}}\left(\sum_{\substack{F\in\mathcal{F}_K\\F\not\subseteq\Gamma_{\text{in}}}}\int_F\Upsilon^\star(\rho)\cdot n^K\phi\,\mathrm{d}a + \sum_{\substack{F\in\mathcal{F}_K\\F\subseteq\Gamma_{\text{in}}}}\rho_{\text{in}}q\cdot n^K\phi\,\mathrm{d}a\right) \tag{5.2}$$

Die Semidiskretisierung ist motiviert durch (5.2) und lautet: Bestimme $\rho_h\in\mathcal{Q}_h$, sodass für alle $\phi_h\in\mathcal{Q}_h$ gilt:

$$\sum_{K\in\mathcal{T}}(\partial_t\rho_h,\phi_h)_K = \sum_{K\in\mathcal{T}}\left((\Upsilon(\rho_h),\nabla\phi_h)_K - \sum_{\substack{F\in\mathcal{F}_K\\F\not\subseteq\Gamma_{\text{in}}}}(\Upsilon^\star(\rho_h)\cdot n^K,\phi_h)_F - (\rho_{\text{in}}q\cdot n^K,\phi_h)_{\partial K\cap\Gamma_{\text{in}}}\right) \tag{5.3}$$

Sei nun G die Anzahl der Zellenfreiheitsgrade. Diese hängen direkt von der Wahl des Lösungs-/Testraumes $\mathcal{Q}_h$, bzw. der Wahl von p ab. Die Wahl $p=0$ liefert gerade ein Finite Volumen Verfahren, G ist dann gerade 1 und die Zellenbasis $\{\mu_K\}_{K\in\mathcal{T}}$ ist durch $\mu_K = \mathbb{1}_K$ gegeben. Für $p\geq 1$ erhalten wir hingegen ein Discontinuous Galerkin Verfahren. Die Anzahl der Zellfreiheitsgrade einer Zelle K ist dann gerade durch $G = p\cdot|\mathcal{V}_K|$ (also p-mal die Anzahl der Ecken von K) gegeben. Die Zellenbasis hat in diesem Fall $G\cdot|\mathcal{T}|$ Elemente und für je G Basisfunktionen μ gilt $\operatorname{supp}(\mu)\subseteq\overline{K}$.
Sei $N = |\mathcal{T}|$ und $\mathcal{T} = \{K_1,\dots,K_{|\mathcal{T}|}\}$ geordnet. Dann lässt sich die Zellenbasis $\{\mu_i\}_{i=1}^{G|\mathcal{T}|}\subset\mathcal{Q}_h$ gemäß dem Träger der Basisfunktion zu $\{\mu_{K_j,i}\}_{j=1,\dots,N;i=1,\dots,G}$ ordnen. Für jede fes-

te Zelle K ergibt sich durch Einsetzen der zugehörigen Basisfunktionen $\{\mu_i\}_{i=1,\dots,G} = \{\mu_{K,i}\}_{i=1,\dots,G}$:

$$(\partial_t \rho_h, \mu_i)_K = \left((\Upsilon(\rho_h), \nabla \mu_i)_K - \sum_{\substack{F \in \mathcal{F}_K \\ F \not\subseteq \Gamma_{\text{in}}}} \left(\Upsilon^*(\rho_h) \cdot n^K, \mu_i\right)_F - \left(\rho_{\text{in}} q \cdot n^K, \mu_i\right)_{\partial K \cap \Gamma_{\text{in}}} \right)$$

TODO Wir erhalten so folgende Darstellung: So ergibt sich die Differentialgleichung

$$\begin{cases} \underline{M} \partial_t \underline{\rho}(t) = \underline{A}\underline{\rho}(t) + \underline{b}(t) \\ \underline{\rho}(0) = \underline{\rho_0} \end{cases}$$

Da dies nun eine gewöhnliche Differentialgleichung ist, können wir die Lösung

$$\underline{\rho}(t) = \exp(t\underline{M}^{-1}\underline{A}) \left(\underline{\rho_0} + \int_0^t \exp(-s\underline{M}^{-1}\underline{A})\underline{b}(s)\,\mathrm{d}s \right) \tag{5.4}$$

explizit angeben. Es handelt sich hierbei aber immer noch um eine semidiskrete Formulierung. Wir wollen deshalb zuletzt noch auf die Herleitung der Zeitintegratoren eingehen. Diese nutzen wir, um unter Verwendung der oben hergeleiteten Semidiskretisierung die numerische Lösung $\underline{\rho}$ sowohl orts- al auch zeitdiskret zu berechnen. Der Ansatz leitet sich hierbei direkt aus dem Resultat (5.4) ab und besteht aus der Integration der Differentialgleichung $\underline{M}\partial_t\underline{\rho} = \underline{A}\underline{\rho} + \underline{b}$ über die Zeit t im Intervall $[t_i, t_{i+1}]$. Dabei ist $t_i = i\delta t$. Hiermit folgt:

$$\underline{M}\underline{\rho}(t_{i+1}) - \underline{M}\underline{\rho}(t_i) = \int_{t_i}^{t_{i+1}} \underline{M}\partial_t\underline{\rho}(t)dt = \int_{t_i}^{t_{i+1}} \underline{A}\underline{\rho}(t) + \underline{b}(t)dt.$$

Mithilfe der Anwendung verschiedener Quadraturformeln lässt sich daraus ein Runge-Kutta Verfahren herleiten. Über die Rechteckformel

$$\int_{t_i}^{t_{i+1}} \underline{A}\underline{\rho}(t) + \underline{b}dt \approx (t_{i+1} - t_i)(\underline{A}\underline{\rho}(t_{i+1}) + \underline{b}(t_{i+1})) = \Delta t(\underline{A}\underline{\rho}(t_{i+1}) + \underline{b}(t_{i+1}))$$

ergibt sich z.B. das implizite Euler Verfahren

$$\underline{\rho}(t_{i+1}) = \underline{\rho}(t_i) + \Delta t\underline{M}^{-1}(\underline{A}\underline{\rho}(t_{i+1}) + \underline{b}(t_{i+1})).$$

Ein weiteres Verfahren dieser Art, welches wir an dieser Stelle verwenden werden, ist die implizite Mittelpunktsregel (der Übersicht wegen für $\underline{b} \equiv 0$):

$$\underline{\rho}(t_{i+1}) = \underline{\rho}(t_i) + \delta t\underline{M}^{-1}(\underline{A}\frac{1}{2}(\underline{\rho}(t_i) + \underline{\rho}(t_{i+1})) + \frac{1}{2}(\underline{\rho}(t_i) + \underline{\rho}(t_{i+1}))).$$

Das so entstehenden Gesamtverfahren ist aufgrund der Kombination von Discontinuous Galerkin Verfahren und Runge-Kutta-Zeitintegratoren in der Literatur oft auch unter dem Namen 'Runge–Kutta discontinuous Galerkin Methods' zu finden. Einen schönen Überblick über diese Verfahrensklasse bietet der Artikel [12]. Nachdem wir nun das Discontinuous Galerkin Verfahren für die lineare Transportgleichung eingeführt und erklärt haben, sollen nun noch auf einige Eigenschaften des Verfahrens verwiesen werden. Dabei wollen wir uns aber beschränken, einige grundlegende Resultate zu nennen und so eher einen groben Überblick mit Referenzen zur Literatur zu geben. Mehr zur numerischen Analyse des Discontinuous Galerkin Verfahren findet sich zum einen in Standardwerken, wie [15], eine schöne Zusammenstellung bietet aber auch [19].
Ebenfalls findet sich in [19] eine grundlegende numerische Analyse des Discontinuous Galerkin Verfahrens angewandt auf die stationäre lineare Transportgleichung. Dabei werden unter anderem die Konsistenz, die sogenannte Galerkin-Orthogonalität, sowie die Stabilität und Konvergenz des Verfahrens behandelt. Mit der numerischen Analyse des Discontinuous Galerkin Verfahrens an sich befassten sich unter anderem LeSaint und Raviart [26], Peterson [27] und Richter [29]. Runge-Kutta DG Verfahren für der linearen Transportgleichung ähnliche Problemstellungen betrachteten Cockburn und Shu in einer 5-teiligen Serie von Arbeiten. Besonders zu nennen sind dabei in unserem Kontext [10] und [8].

5.5 Eigenschaften des Discontinuous Galerkin Verfahren

5.5.1 Lösungsbegriffe

Wie zuvor bereits beim Potentialströmungsproblem können wir auch für das Transportproblem eine sogenannte schwache Formulierung bestimmen. Diese hängt im Fall des Transportproblems eng mit der Semidiskretisierung zusammen und lautet mit $\partial\mathcal{D} = \Gamma_{\text{in}} \dot\cup \Gamma_{\text{out}}$ und $\rho(x,t) = \rho_{\text{in}}(x,t)$ für $(x,t) \in \Gamma_{\text{in}} \times (0,T)$:

Definition 5.7. $\rho \in L_1(\mathcal{D} \times (0,T))$ heißt schwache Lösung des linearen Transportproblems, falls es für ein gegebenes $q : \overline{\mathcal{D}} \to \mathbb{R}^2$ folgende Bedingungen erfüllt:

$$
\begin{array}{lll}
\text{(swTP)} & B(\rho,\phi) & = \langle b,\phi\rangle \quad \forall \phi \in H^1(\mathcal{D}\times\mathbb{T}) \text{ mit } \phi(\cdot,T) = 0 \text{ und } \phi|_{\Gamma_{\text{out}}} = 0 \\
\text{Dabei sind :} & & \\
& B(\rho,\phi) & \coloneqq \int_0^T \int_{\mathcal{D}} \rho(\partial_t\phi + q\nabla\phi)\,\mathrm{d}x\,\mathrm{d}t - \int_{\Gamma_{\text{out}}} \rho q\cdot n\phi\,\mathrm{d}a\,\mathrm{d}t \\
& \langle b,\phi\rangle & \coloneqq \int_{\Gamma_{\text{in}}} \rho_{\text{in}} q\cdot n\phi\,\mathrm{d}a\,\mathrm{d}t - \int_{\mathcal{D}} \rho_0\phi(0)\,\mathrm{d}x \\
& \Gamma_{\text{out}} & \coloneqq \{z\in\partial\mathcal{D} : q(z)\cdot n(z) > 0\} \\
& \Gamma_{\text{in}} & \coloneqq \{z\in\partial\mathcal{D} : q(z)\cdot n(z) \le 0\}
\end{array}
$$

Es gilt an dieser Stelle außerdem:

Lemma 5.8. (Zusammenhang der Lösungsbegriffe)

1. Ist ρ eine klassische Lösung, so ist ρ auch eine schwache Lösung.

2. Ist $\rho \in C^2(\mathcal{D} \times \mathbb{T}, \mathbb{R})$ und eine schwache Lösung, so ist ρ auch eine klassische Lösung.

Beweis. Sei $\phi : \mathcal{D} \times \mathbb{T} \to \mathbb{R}$ eine beliebige Testfunktion aus $H^1(\mathcal{D} \times \mathbb{T})$, für die $\phi(\cdot, T) = 0$ und $\phi|_{\Gamma_{\text{out}}} = 0$ gelte. Wir halten zunächst fest, dass der Raum $H_0^1(\mathcal{D} \times \mathbb{T})$ vollständig in dem so betrachteten Testraum enthalten ist. Wir beginnen nun mit der Differentialgleichung $\partial_t \rho(x,t) + \operatorname{div}(\rho(x,t)q(x)) = 0$, multiplizieren zunächst mit einer Testfunktion ϕ aus dem Testraum und integrieren anschließend über den Raum-Zeitzylinder $\mathcal{D} \times \mathbb{T}$:

$$\int_{\mathcal{D}\times\mathbb{T}} \big(\partial_t \rho + \operatorname{div}(\rho q)\big)\phi \,\mathrm{d}(x,t) = \underbrace{\int_{\mathcal{D}\times\mathbb{T}} \partial_t \rho \phi \,\mathrm{d}(x,t)}_{(1)} + \underbrace{\int_{\mathcal{D}\times\mathbb{T}} \operatorname{div}(\rho q)\phi \,\mathrm{d}(x,t)}_{(2)}$$

Betrachten wir nun zunächst Integral (1), so folgt mit partieller Integration:

$$\begin{aligned}
\int_{\mathcal{D}\times\mathbb{T}} \partial_t \rho \phi \,\mathrm{d}(x,t) &= \int_{\mathcal{D}} \int_{\mathbb{T}} \partial_t \rho \phi \,\mathrm{d}t \,\mathrm{d}x \\
&= \int_{\mathcal{D}} \Big(- \int_{\mathbb{T}} \rho \partial_t \phi \,\mathrm{d}t + \big[\rho\phi\big]_0^T \Big) \,\mathrm{d}x \\
&= - \int_{\mathcal{D}\times\mathbb{T}} \rho \partial_t \phi \,\mathrm{d}(x,t) + \int_{\mathcal{D}} \underbrace{\rho(x,T)\phi(x,T)}_{=0} - \underbrace{\rho(x,0)}_{=\rho_0(x) \text{ auf } \mathcal{D}} \phi(x,0) \,\mathrm{d}x \\
&= - \int_{\mathcal{D}\times\mathbb{T}} \rho \partial_t \phi \,\mathrm{d}(x,t) - \int_{\mathcal{D}} \rho_0 \phi(x,0) \,\mathrm{d}x
\end{aligned}$$

Außerdem können wir Integral (2) mit 2.3 wie folgt ausdrücken:

$$\begin{aligned}
\int_{\mathcal{D}\times\mathbb{T}} \operatorname{div}(\rho q)\phi \,\mathrm{d}(x,t) &= \int_{\mathbb{T}} \int_{\mathcal{D}} \operatorname{div}(\rho q)\phi \,\mathrm{d}x \,\mathrm{d}t \\
&= \int_{\mathbb{T}} \Big(- \int_{\mathcal{D}} \rho q \nabla \phi \,\mathrm{d}x + \int_{\partial\mathcal{D}} \rho q \cdot n \ \phi \,\mathrm{d}a \Big) \,\mathrm{d}t \\
&= - \int_{\mathcal{D}\times\mathbb{T}} \rho q \nabla \phi) \,\mathrm{d}(x,t) + \int_{\mathbb{T}} \int_{\partial\mathcal{D}} \rho q \cdot n \phi \,\mathrm{d}a \,\mathrm{d}t
\end{aligned}$$

Wenn ρ eine klassische Lösung ist, dann existieren insbesondere $\partial_t \rho$ und $\operatorname{div}(\rho q)$ und obige Umformungen sind zulässig, d.h. ρ erfüllt auch die schwache Formulierung.
Ist ρ hingegen eine schwache Lösung die zusätzlich in $C^2(\mathcal{D} \times \mathbb{T}, \mathbb{R})$ liegt, so lassen sich alle oben durchgeführten Umformungen auch in die andere Richtung durchführen und wir erhalten:

$$\int_{\mathcal{D}\times\mathbb{T}} (\partial_t \rho + \operatorname{div}(\rho q))\phi \,\mathrm{d}(x,t) = 0 \quad \forall \phi \in H_0^1(\mathcal{D} \times \mathbb{T})$$

Dann folgt mit (2.6), dass ρ auch die ursprüngliche Differentialgleichung

$\partial_t \rho(x,t) + \operatorname{div}(\rho(x,t)q(x)) = 0$ erfüllt. Somit ist ρ also auch klassische Lösung.

□

Auch die Semidiskretisierung können wir in ähnlicher Form, wie eben noch die schwache Formulierung, ausdrücken. Sei dazu

$$B_h(\rho_h, \phi) \coloneqq \sum_{i=1}^{N} (\partial_t \rho_h, \phi_h)_{K_i} - \left(\sum_{i=1}^{N} (\Upsilon(\rho_h), \nabla \phi_h)_{K_i} - \sum_{\substack{F \in \mathcal{F}_{K_i} \\ F \not\subseteq \Gamma_{\text{in}}}} (\Upsilon^\star(\rho_h) \cdot n^K, \phi_h)_F \right)$$

$$\langle b_h, \phi_h \rangle \coloneqq (\rho_{\text{in}} q \cdot n^K, \phi_h)_{\partial K_i \cap \Gamma_{\text{in}}}$$

Dann löst $\rho_h \in Q_h$ die Semidiskretisierung (5.3), wenn $B_h(\rho_h, \phi) = \langle l_h, \phi_h \rangle$ für alle $\phi_h \in Q_h$ erhalten ist.

5.5.2 Konsistenz

Abschnitt 5.4.1 hat sich damit beschäftigt, aus dem ursprünglich unendlich dimensionalen Problem letztendlich eine volldiskretisierte Verfahrensvorschrift in einem endlichen Ansatzraum herzuleiten. Fragen wir nun nach der Konsistenz des Verfahrens, stellen wir damit zugleich die Frage, ob wir immer noch die richtige Gleichung lösen. Genauer heißt das Verfahren genau dann konsistent, wenn eine analytische Lösung ρ des ursprünglichen Problems (dTP) auch die hergeleitete Verfahrensvorschrift erfüllt. Wir betrachten zunächst etwas abstrakter den formalen Prozess der Diskretisierung einer Gleichung

$$B\rho = 0$$

. Dabei werde das abstrakte Problem $B\rho = 0$ mit einem Operator $B : E \to F$ mithilfe der Abbildungen L_1, L_2, Φ_h diskretisiert. Zu einer Lösung der abstrakten Gleichung $z \in E$ existiere genau ein $z_h \in E_h$ und der lokale Diskretisierungsfehler sei gegeben durch $l_h \coloneqq B_h L_1(z) = \phi_h(B) L_1(z) \in F_h$. Diese Situation kann durch folgendes Diagramm verdeutlicht werden:

Abbildung 2: Diskretisierungsprozess

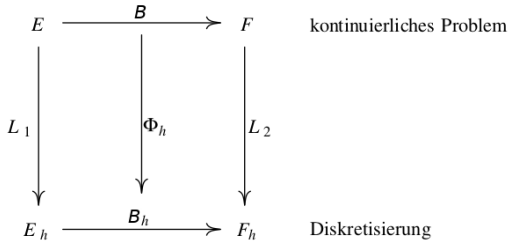


Abbildung leicht abgeändert aus [5] Seite 390

Das diskretisierte Problem $B_h\rho_h = 0$ heißt genau dann konsistent, wenn für eine analytische Lösung $\rho^\star \in E$ gilt, dass $\lim_{h\to 0}\|\Phi_h(B)L_1\rho^\star - L_2B\rho^\star\|_{F_h} = 0$. Betrachten wir zunächst in einem Zwischenschritt die Semidiskretisierung (5.3). Die gewählte Herleitung dieser Formulierung soll dabei nahelegen, dass für eine exakte (und glatte) Lösung ρ die Konsistenz für die Semidiskretisierung erfüllt ist. In oben eingeführter Notation gilt also gerade $B_h(\rho^\star, \phi)$ für alle Testfunktionen ϕ. Auch wenn es sich bei $\mathcal{Q}_h$ um einen nicht konformen Ansatzraum handelt, ist der Herleitung der Semidiskretisierung dennoch zu entnehmen, dass auch $B_h(\rho^\star, \phi_h)$ für alle $\phi_h \in \mathcal{Q}_h$ gilt $(\star)$. Wichtig ist dabei unter anderem auch die Wahl der numerischen Flussfunktion, der upwind flux erhält aber gerade gewünschten Eigenschaften. Mehr dazu findet sich in [19]. Für die Zeitdiskretisierung in Form der impliziten Mittelpunktsregel gilt als einstufige Gauß-Quadratur $\|\Phi_h(B)L_1\rho^\star - L_2B\rho^\star\|_{F_h} = O(h^2)$, womit sie insbesondere konsistent ist.
Daher gilt so für das kombinierte Verfahren, dass eine klassische Lösung $\rho^\star$ somit auch die Volldiskretisierung in obigem Sinne erfüllt.

5.5.3 Galerkin Orthogonalität

Eine direkte Folgerung aus der Konsistenz und $(\star)$ stellt die Galerkin Orthogonalität dar. Für eine Lösung der Semidiskretisierung ρ_h und eine analytische Lösung im klassischen Sinne $\rho^\star$ gilt dann nämlich:

$$B_h(\rho^\star - \rho_h, \phi_h = 0) \quad \forall \phi_h \in \mathcal{Q}_h$$

5.5.4 Stabilität und Konvergenz

Die Stabilität ist bei der numerischen Lösung der Transportgleichung einer der wesentlichen Gründe, weswegen wir das Discontinuous Galerkin Verfahren einem Standard-Finite-Elemente-Ansatz vorziehen. Es zeigt sich nämlich, dass ein normales Finite Elemente Verfahren, wie wir es zuvor bei der Lösung des Potentialströmungsproblems genutzt haben, beim Transportproblem instabil ist. Grund dafür ist, dass $\|q\cdot\nabla\rho_h\|$ beliebig groß werden kann. Für die DG-Diskretisierung mit upwind flux kann hingegen gezeigt werden, dass die Lösung des Transportproblems stabil ist. Auf eine theoretische Stabilitätsanalyse möchten wir aber an dieser Stelle verzichten und verweisen z.B. auf [19] oder [15]. Wie bereits an früherer Stelle erwähnt findet sich in [19] auch eine grundlegende numerische Analyse des Discontinuous Galerkin Verfahren angewandt auf die stationäre lineare Transportgleichung. Ebenso wollen wir bezüglich der Konvergenz des Verfahrens auf entsprechende Literatur verweisen. Wir werden später Konvergenzannahmen stellen und diese in entsprechenden Experimenten verifizieren. In der Literatur finden sich aber auch theoretische Konvergenzbeweise, meist unter Verwendung vielfältiger funktionalanalytischer Grundlagen und unter gewissen Regularitätsvoraussetzungen an die bestimmte Lösung. Die theoretische Betrachtung von Discontinuous Galerkinverfahren ist bereits Gegenstand vielfältiger wissenschaftlicher Arbeiten, aber zugleich ein breites Feld, welches noch lange nicht vollständig erforscht ist.

6 Anwendung der Multilevel Monte Carlo Methode auf das Transportproblem

Nachdem wir in Abschnitt 2 an einige Grundlagen erinnert, in Abschnitt 3 und 4 sowohl die Monte Carlo Methode, als auch die Multilevel Monte Carlo Methode eingeführt und in Abschnitt 5 das Transportproblem, sowie das Potentialströmungsproblem inklusive numerischer Verfahren, welche zur Lösung dergleichen genutzt werden können, erklärt haben, soll nun dieser Abschnitt dazu dienen, die bisherigen Ergebnisse zu bündeln und die Anwendung der Multilevel Monte Carlo Methode auf partielle Differentialgleichungen am Beispiel des Transportproblem nahe zu legen. Dabei nimmt dieser Abschnitt einen zentralen Platz in dieser Thesis ein, weswegen wir noch einmal darlegen wollen, was genau unser Ziel ist und wie wir die uns an dieser Stelle zu Verfügung stehenden Mittel einsetzen, um das Gewünschte zu erreichen. Sei hierzu $\mathcal{D} \subset \mathbb{R}^2$ beschränktes Polygongebiet, $\mathbb{T} = (0, T]$ für ein $T > 0$ und $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum.

Für ein stochastisches Flussvektorfeld $q : \Omega \times \overline{\mathcal{D}} \to \mathbb{R}^2$, bestimme $\rho : \Omega \times \overline{\mathcal{D}} \times \mathbb{T} \to \mathbb{R}_{\geq 0}$ mit

$$\text{(pTP)} \begin{cases} \partial_t \rho(\omega, x, t) + \operatorname{div}(\rho(\omega, x, t) q(\omega, x)) &= 0 & \text{, für } (x,t) \in \mathcal{D} \times \mathbb{T} \\ \rho(\omega, x, t) &= \rho_{\text{in}}(x, t) & \text{, für } (x,t) \in \Gamma_{\text{in}} \times \mathbb{T} \\ \rho(\omega, x, 0) &= \rho_0(x) & \text{, für } x \in \mathcal{D} \end{cases}$$

für die Anfangs- und Randwerte:

$g_N : \Gamma_{\text{N}} \to \mathbb{R}$
$u_D : \Gamma_{\text{D}} \to \mathbb{R}$
$\rho_{\text{in}} : \Gamma_{\text{in}} \times \mathbb{T} \to \mathbb{R}_{\geq 0}$
$\rho_0 : \mathcal{D} \to \mathbb{R}_{\geq 0}$
wobei $\partial\mathcal{D} = \Gamma_{\text{D}} \dot\cup \Gamma_{\text{N}}$ und $\Gamma_{\text{in}} \coloneqq \{z \in \partial\mathcal{D} : q(z) \cdot n(z) \leq 0\} \subset \partial\mathcal{D}$

Genauer sei $Q(\omega) = J(\rho(\omega))$ ein gegebenes Zielfunktional, dann ist es unser Ziel den Erwartungswert $\mathbb{E}[Q(\omega)]$ möglichst genau zu bestimmen. Beispielsweise kann eine beliebige Norm von $\rho(\omega)$ als Zielfunktional betrachtet werden. Unser Modellproblem lautet also:

$$\text{(MP)} \begin{cases} \text{Für ein stochastisches Flussvektorfeld } q : \Omega \times \overline{\mathcal{D}} \to \mathbb{R}^2 \\ \text{und ein Zielfunktional } J \text{ , bestimme } \mathbb{E}[J(\rho)] \\ \text{mit } \rho \text{ als Lösung von (pTP) inklusive Anfangs- und Randwerten} \end{cases} \tag{6.1}$$

Dabei erhalten wir das stochastische Flussvektorfeld q, wie bereits in 5.1.2 erklärt selbst als Lösung des Potentialströmungsproblem. Da wir an der numerischen Lösung partieller Differentialgleichungen interessiert sind und wir die im Allgemeinen unendlich dimensionale Lösung ρ durch eine endlich dimensionale Lösung $\rho_{h,\Delta t}$ approximieren, betrachten wir $Q_{h,\Delta t}(\omega) \coloneqq J(\rho_{h,\Delta t}(\omega))$. Wir nutzen dabei das in Abschnitt 5.4 behandelte Disconti-

nuous Galerkin Verfahren zur Lösung des linearen Transportproblems. Außerdem werden wir im Folgenden eine uniforme Familie $\{\mathcal{T}_h\}$ von Zerlegungen von $\mathcal{D}$ (vgl. Abschnitt 5.2 Definition 5.2) als Diskretisierungsgitter für die Ortsdiskretisierung, sowie $\mathbb{T}_{\Delta t}$ als Zerlegung von $\mathbb{T}$ betrachten. Insbesondere wählen wir später Δt in Abhängigkeit von h, etwa $\Delta t = ch$ für ein $c > 0$ und betrachten dann $Q_h(\omega) \coloneqq Q_{h,ch}(\omega) = J(\rho_h(\omega)) \coloneqq J(\rho_{h,ch}(\omega))$. An dieser Stelle sei betont, dass es sich in diesem Abschnitt bei ρ_h um eine volldiskrtisierte Approximation an ρ handelt und nicht die Semidiskretisierung aus Abschnitt 5.4 gemeint ist. Um die Notation im Folgenden zu erleichtern schreiben wir für $a, b \in \mathbb{R}_{>0}$ $a \lesssim b$, falls $\frac{a}{b}$ gleichmäßig beschränkt und insbesondere unabhängig von den Parametern h und n ist.

Annahme 6.1. (Konvergenz im Erwartungswert des Zielfunktionals)
In obiger Situation gelte

$$\mathbb{E}[Q_h] \to \mathbb{E}[Q] \text{ für } h \to 0$$

Genauer existiere ein $\alpha > 0$, sodass

$$|\mathbb{E}[Q_h - Q]| \lesssim h^\alpha$$

Wir nennen α dann auch die Konvergenzrate von Q_h.

6.1 Die Monte Carlo Methode

Wie bereits im Kontext der numerischen Integration wollen wir die Monte Carlo Methode als Ausgangspunkt nutzen und anschließend bei der Betrachtung der Multilevel Monte Carlo Methode auch auf entscheidende Unterschiede zu und Vorteile gegenüber der Monte Carlo Methode eingehen. Sowohl bei der Monte Carlo Methode, als auch bei der Multilevel Variante approximieren wir den Erwartungswert $\mathbb{E}[Q_h]$ durch einen Schätzwert $\widehat{Q}_h$. Um die Genauigkeit und die Kosten zu bemessen betrachten wir zum einen den sogenannten 'root mean square error' (RMSE)

$$e(\widehat{Q}_h) \coloneqq \left(\mathbb{E}\left[(\widehat{Q}_h - \mathbb{E}[Q])^2\right]\right)^{\frac{1}{2}} \tag{6.2}$$

zum anderen die Anzahl an floating-point-Rechenoperationen $C_\epsilon(\widehat{Q}_h)$, die benötigt werden um einen RMSE mit $e(\widehat{Q}_h) \leq \epsilon$ zu erhalten. Zu beachten ist hierbei, dass in dem RMSE einige Fehlerquellen gemeinsam betrachtet werden. So gehen sowohl der Discontinuous Galerkin Fehler des Transportproblems (Approximation von ρ durch ρ_h), der Approximationsfehler der Approximation von Q durch Q_h, als auch der statistische Fehler des Schätzers $\widehat{Q}_h$ in den RMSE mit ein. Insbesondere bedeutet also $e(\widehat{Q}_h) \leq \epsilon$, dass alle oben genannten Fehlerquellen kleiner als ϵ ausfallen. Betrachten wir also, wie bei der Monte Carlo Methode, welche wir gleich noch einmal kurz behandeln wollen, nur eine einzige Zerlegung $\mathcal{T}$ der Familie $\mathcal{T}_h$, so kann es sein, dass es gar nicht möglich ist den RMSE durch ein bestimmtes ϵ zu beschränken, da einer der Approximationsfehler für das gewählte feste Level bereits alleine größer als ϵ ist. Bei der Standard Monte

Carlo Methode schätzen wir $\mathbb{E}[Q]$ durch den Mittelwert n unabhängiger gleichverteilter Zufallssamples und erhalten so

$$\widehat{Q}_{h,n}^{\mathrm{MC}} \coloneqq \frac{1}{n}\sum_{i=1}^{n} Q_h(\omega_i) = \sum_{i=1}^{n} J(\rho_h(\omega_i)) \tag{6.3}$$

Dabei modellieren wir das zufällige Flussvektorfeld $q(\omega_i) : \mathcal{D} \to \mathbb{R}^2$, indem wir zunächst für $\kappa(\omega_i) : \mathcal{D} \to (\mathbb{R}_{\mathrm{sym}})^{d\times d}$ ein lognormal-verteiltes unabhängiges Zufallsfeld erzeugen und anschließend das Potentialströmungsproblem

Für $\kappa(\omega_i) : \mathcal{D} \to (\mathbb{R}_{\mathrm{sym}})^{d\times d}$, bestimme $u(\omega_i,\cdot) : \overline{\mathcal{D}} \to \mathbb{R}$ und $q(\omega_i,\cdot) : \overline{\mathcal{D}} \to \mathbb{R}^2$ mit

$$\text{(PS)} \begin{cases} \operatorname{div}(q(\omega_i,x)) &= 0 & \text{, für } x \in \mathcal{D} \\ q(\omega_i,x) &= -\kappa(\omega_i)\nabla u(\omega_i,x) & \text{, für } x \in \mathcal{D} \\ -q(\omega_i,x)\cdot n &= g_N(x) & \text{, für } x \in \Gamma_N \\ u(\omega_i,x) &= u_D(x) & \text{, für } x \in \Gamma_D \end{cases}$$

lösen. Zur Erzeugung des lognormal-verteilten Zufallsfeldes können wir auf entsprechende Algorithmen zurückgreifen, in unserem Fall etwa dem sogenannten Circulant Embedding. Der Algorithmus wurde 1997 erstmals in [13] vorgestellt und erzeugt Gauß'sche Zufallsfelder auf regulären Gittern und basiert auf der Fast Fourier Transformation, welche in der Literatur auch oft unter der Abkürzung FFT zu finden ist. Dabei werden spezielle Strukturen der Kovarianzmatrix ausgenutzt. Anschließend kann das Gauß'sche Zufallsfeld über eine einfache Transformation in ein lognormal Feld überführt werden. Mehr zu Circulant Embedding findet sich z.B. in [31] Abschnitt 12. Auch die grundsätzliche Idee Circulant Embedding für die Modellierung der Ausgangsdaten in stochastischen partiellen Differentialgleichungen zu nutzen ist keineswegs neu und findet sich z.B. in [6] oder [7]. Wir haben nun alle Mittel in der Hand um die Monte Carlo Methode angewandt auf das Transportproblem als Algorithmus zu formulieren. Dabei fassen wir die eben erklärte Erzeugung des zufälligen Vektorfeldes in der Funktion 'RndVecField' zusammen. Außerdem gehen wir davon aus, dass bereits alle übrigen freien Parameter, sowie die Rand- und Anfangswerte fest gewählt sind.

Algorithm 1: Monte Carlo Methode angewandt auf das Transportproblem

Input : h, n
Output: $\widehat{Q}_{h,n}^{\mathrm{MC}}$

Initialisiere: $\Sigma = 0, i = 0$
while $i < n$ **do**
  Erzeuge ein Zufallssample: $q(\omega_i, x) \leftarrow$ RndVecField
  Löse das Transportproblem: $\rho_h(\omega_i, x) \leftarrow$ Löse mit Discontinuous Galerkin
  Berechne das zugehörige Zielfuinktional: $Q_h(\omega_i) \leftarrow$ Berechne Zielfunktional
  Setze: $\Sigma = \Sigma + Q_h(\omega_i)$, i = i+1
end

Result: $\widehat{Q}_{h,n}^{\mathrm{MC}} = \Sigma/n$

Wir nehmen an dieser Stelle an, dass sich die Gesamtkosten an Rechenoperationen, welche für die Berechnung eines $Q_h(\omega_i)$ benötigt werden, durch

$$C(Q_h(\omega_i)) \lesssim h^{-\gamma}$$

beschränken lassen. Im Folgenden wollen uns überlegen, wie wir unter dieser Annahme die Kosten für C_ϵ abschätzen können. Diese Überlegungen finden sich auch in [7]. Der 'mean square error' $e(\widehat{Q}_{h,n}^{\mathrm{MC}})^2$, also das Quadrat des RMSE, lässt sich nämlich auch folgendermaßen betrachten:

$$\begin{aligned} e(\widehat{Q}_{h,n}^{\mathrm{MC}})^2 &= \mathbb{E}\left[\left(\widehat{Q}_{h,n}^{\mathrm{MC}} - \mathbb{E}[\widehat{Q}_{h,n}^{\mathrm{MC}}] + \mathbb{E}[\widehat{Q}_{h,n}^{\mathrm{MC}}] - \mathbb{E}[Q]\right)^2\right] \\ &= \mathbb{E}\left[(\widehat{Q}_{h,n}^{\mathrm{MC}} - \mathbb{E}[\widehat{Q}_{h,n}^{\mathrm{MC}}])^2\right] + \left(\mathbb{E}[\widehat{Q}_{h,n}^{\mathrm{MC}}] - \mathbb{E}[Q]\right)^2 \\ &= \mathbb{V}[\widehat{Q}_{h,n}^{\mathrm{MC}}] + \left(\mathbb{E}[\widehat{Q}_{h,n}^{\mathrm{MC}}] - \mathbb{E}[Q]\right)^2 \end{aligned} \tag{6.4}$$

Dabei nutzen wir, dass für zwei unabhängige Zufallsvariablen a und b mit Erwartungswert $\mathbb{E}[a] = 0 = \mathbb{E}[b]$ stets $\mathbb{E}[(a+b)^2] = \mathbb{E}[a^2] + \mathbb{E}[b^2]$ gilt. Da weiter $\mathbb{E}[\widehat{Q}_{h,n}^{\mathrm{MC}}] = \mathbb{E}[Q_h]$ und $\mathbb{V}[\widehat{Q}_{h,n}^{\mathrm{MC}}] = \frac{1}{n^2} n \mathbb{V}[Q_h] = \frac{1}{n}\mathbb{V}[Q_h]$ gilt, erhalten wir so

$$e(\widehat{Q}_{h,n}^{\mathrm{MC}})^2 = \underbrace{\frac{1}{n}\mathbb{V}[Q_h]}_{(1)} + \underbrace{(\mathbb{E}[Q_h - Q])^2}_{(2)} \ . \tag{6.5}$$

Dabei ist (1), wie wir eben bereits gesehen haben, die Varianz des Monte Carlo Schätzers und spiegelt daher den Schätzfehler wider. Unter der zusätzlichen Annahme, dass Erwartungswert und Varianz des Schätzers existieren, konvergiert dieser Fehler nach dem starken Gesetz großer Zahlen für $n \to \infty$ gegen 0. (2) hingegen hängt als Quadrat über den Gesamtapproximationsfehler zwischen Q_h und Q nur von der Diskretisierungsschrittweite h ab. Ist das verwendete Lösungsverfahren konvergent, so geht (2) gegen 0

für $h \to 0$. Wollen wir den RMSE also durch ein $\epsilon > 0$ beschränken, ist es hinreichend dafür zu sorgen, dass sowohl (1) also auch (2) kleiner als $\frac{\epsilon^2}{2}$ ausfallen. Wir müssen also h klein genug und zugleich n groß genug wählen. Genauer kann dies, unter der Annahme, dass $\mathbb{V}[Q_h]$ annäherend konstant ist und somit nicht von h abhängt, erreicht werden, indem wir $n \gtrsim \epsilon^{-2}$ und $h \lesssim \epsilon^{\frac{1}{\alpha}}$ (vgl. Annahme 6.1). Da wir weiter angenommen hatten, dass $C(Q_h(\omega_i)) \lesssim h^{-\gamma}$ gilt, erhalten wir

$$C(\widehat{Q}^{\mathrm{MC}}_{h,n}) \lesssim nh^{-\gamma}$$

und somit Gesamtkosten für einen erwarteten Schätzfehler kleiner als ϵ von

$$C_\epsilon(\widehat{Q}^{\mathrm{MC}}_{h,n}) \lesssim \epsilon^{-2-\frac{\gamma}{\alpha}} \tag{6.6}$$

6.2 Die Multilevel Monte Carlo Methode

Wie wir bereits in Abschnitt 4 dargelegt haben, ist die entscheidende Idee der Multilevel Monte Carlo Methode Zufallssamples auf mehreren verschiedenen Leveln $\{h_l : l = 0, \dots, L\}$ mit $h_0 > h_1 > \cdots > h_L$ zu betrachten. Wir erinnern an dieser Stelle daran, dass wir für die Menge aller betrachtbaren Level $\mathcal{H} = \{h_0, h_1 \coloneqq \frac{h_0}{2}, h_2 \coloneqq \frac{h_1}{2} = \frac{h_0}{4}, \dots\}$ gefordert hatten, insbesondere war hier $0 \in \overline{\mathcal{H}}$. Theoretisch gesehen berechnet der Algorithmus, welchen wir an späterer Stelle noch formulieren wollen, für ein gegebenes ϵ nur Samples auf den für diese Genauigkeit benötigten Leveln, einer endlichen Untermenge von $\mathcal{H}$. Die Menge der tatsächlich betrachteten Samples $\tilde{\mathcal{H}} = \{h_l : l = 0, \dots, L\} \subset \mathcal{H}$ ist also stets endlich. Wie bereits in 4 erhalten wir

$$\mathbb{E}[Q_{h_L}] = \mathbb{E}[Q_{h_0}] + \sum_{l=0}^{L} \mathbb{E}[Q_{h_l} - Q_{h_{l-1}}]$$

Um uns an dieser Stelle die Notation zu vereinfachen führen wir an dieser Stelle für $l = 0, \dots, L$ die Zufallsvariable Y_l ein mit $Y_0(\omega) \coloneqq Q_{h_0}(\omega) = J(\rho_{h_0}(\omega))$ und $Y_l \coloneqq Q_{h_l}(\omega) - Q_{h_{l-1}}(\omega) = J(\rho_{h_l}(\omega)) - J(\rho_{h_{l-1}}(\omega))$ für $l = 1, \dots, L$. Man beachte hierbei, dass wie bereits an früher Stelle erwähnt, beim Vergleich zweier verschiedener Level das gleiche Sample zugrunde gelegt wird. So folgt:

$$\mathbb{E}[Q_{h_L}] = \sum_{l=0}^{L} \mathbb{E}[Y_l]$$

Dann ist der Multilevel Monte Carlo Schätzer gegeben durch die Summe der Monte Carlo Schätzer $\widehat{Y}^{\mathrm{MC}}_{h,n_l}$ für die einzelnen Erwartungswerte $\mathbb{E}[Y_l]$:

$$\widehat{Q}^{\mathrm{MLMC}}_{\tilde{\mathcal{H}},\{n_l\}_{l=0}^L} = \sum_{l=0}^{L} \widehat{Y}^{\mathrm{MC}}_{l,n_l} = \frac{1}{n_0} \sum_{i_0=0}^{n_0} Q_{h_0}(\omega_{i_0}) + \sum_{l=1}^{L} \frac{1}{n_l} \sum_{i_l=1}^{n_l} \left(Q_{h_l}(\omega_{i_l}) - Q_{h_{l-1}}(\omega_{i_l}) \right) \tag{6.7}$$

Da für jedes l $\widehat{Y}_{l,n_l}^{\mathrm{MC}}$ getrennt berechnet und somit jeder Erwartungswert $\mathbb{E}[Y_l]$ unabhängig geschätzt wird, erhalten wir die Varianz des MLMC-Schätzers durch $\mathbb{V}[\widehat{Q}^{\mathrm{MLMC}}_{\tilde{\mathcal{H}},\{n_l\}_{l=0}^L}] = \sum_{l=0}^L \frac{1}{n_l}\mathbb{V}[Y_l]$. Wie in 6.4 lässt sich dann der mean square error dann ausdrücken durch:

$$\begin{aligned} e(\widehat{Q}^{\mathrm{MLMC}}_{\tilde{\mathcal{H}},\{n_l\}_{l=0}^L})^2 &= \mathbb{E}\left[(\widehat{Q}^{\mathrm{MLMC}}_{\tilde{\mathcal{H}},\{n_l\}_{l=0}^L} - \mathbb{E}[Q])^2\right] \\ &= \mathbb{E}\left[(\widehat{Q}^{\mathrm{MLMC}}_{\tilde{\mathcal{H}},\{n_l\}_{l=0}^L} - \mathbb{E}[\widehat{Q}^{\mathrm{MLMC}}_{\tilde{\mathcal{H}},\{n_l\}_{l=0}^L}] + \mathbb{E}[\widehat{Q}^{\mathrm{MLMC}}_{\tilde{\mathcal{H}},\{n_l\}_{l=0}^L}] - \mathbb{E}[Q])^2\right] \\ &= \mathbb{E}\left[(\widehat{Q}^{\mathrm{MLMC}}_{\tilde{\mathcal{H}},\{n_l\}_{l=0}^L} - \mathbb{E}[\widehat{Q}^{\mathrm{MLMC}}_{\tilde{\mathcal{H}},\{n_l\}_{l=0}^L}])^2\right] + \left(\mathbb{E}[\widehat{Q}^{\mathrm{MLMC}}_{\tilde{\mathcal{H}},\{n_l\}_{l=0}^L}] - \mathbb{E}[Q]\right)^2 \\ &= \underbrace{\sum_{l=0}^{L} \frac{1}{n_l}\mathbb{V}[Y_l]}_{(1)} + \underbrace{(\mathbb{E}[Q_{h_L} - Q])^2}_{(2)} \end{aligned} \tag{6.8}$$

Bemerkung. Was lässt sich hieraus ablesen?

- Unter Annahme 6.1 gilt $\mathbb{V}[Y_l] = \mathbb{V}[Q_{h_l} - Q_{h_{l-1}}] \overset{l\to\infty}{\to} 0$, das bedeutet für uns: Je größer l und damit je feiner die Gitterweite h_l ist, desto weniger Samples werden benötigt um $\mathbb{E}[Y_l]$ zu schätzen.
- Das niedrigste betrachtete Level $l = 0$ kann unabhängig von ϵ fest gewählt werden. So bleibt insbesondere die benötigte Anzahl an Rechenoperationen pro Sample auf dem niedrigsten Level konstant, auch wenn $\epsilon \to 0$ geht. Bei der tatsächlichen Anwendung muss allerdings h_0 so gewählt werden, dass zumindest ein Mindestmaß an Auflösung des Problems gegeben ist. Deswegen werden wir später beim Algorithmus die Level beginnend bei l_0 indizieren. Um in der Theorie aber die Notation so schlank wie möglich zu halten, haben wir uns bewusst dafür entschieden bei der Indizierung in der Theorie mit Level $l = 0$ zu beginnen.

Wir sind nun in der Lage folgendes zentrale Resultat aus [7] nachzuvollziehen:

Satz 6.2.
Sei in obiger Situation $\widehat{Y}_l \coloneqq \widehat{Y}_{l,n_l}^{\mathrm{MC}}$ der Monte Carlo Schätzer für Y_l und $C_l \coloneqq C(Y_l^{(i)})$ die Anzahl der Rechenoperationen, welche für die Berechnung eines Samples von Y_l benötigt werden. Es seien $\alpha, \beta, \gamma, c_1, c_2, c_3 > 0$ Konstanten mit $\alpha \geq \frac{1}{2}\min(\beta, \gamma)$ und

(a) $|\mathbb{E}[Q_{h_l} - Q]| \leq c_1 h_l^{\alpha}$ (vgl. Annahme 6.1)

(b) $\mathbb{V}[Y_l] = \mathbb{V}[Q_{h_l} - Q_{h_{l-1}}] \leq c_2 h_l^{\beta}$

(c) $C_l \leq c_3 h_l^{-\gamma}$.

Dann existiert für jedes $0 < \epsilon < \frac{1}{e}$ ein $L \in \mathbb{N}$ und ein zugehöriges $h_L \in \mathcal{H}$, sodass für $\tilde{\mathcal{H}} = \{h_l\}_{l=0}^L \subset \mathcal{H}$ gelten:

$$e(\widehat{Q}^{\mathrm{MLMC}}_{\tilde{\mathcal{H}},\{n_l\}_{l=0}^L})^2 = \mathbb{E}\left[\left(\widehat{Q}^{\mathrm{MLMC}}_{\tilde{\mathcal{H}},\{n_l\}_{l=0}^L} - \mathbb{E}[Q]\right)^2\right] < \epsilon^2 ,$$

$$C_\epsilon(\widehat{Q}^{\mathrm{MLMC}}_{\tilde{\mathcal{H}},\{n_l\}_{l=0}^L}) \leq \tilde{c} \begin{cases} \epsilon^{-2}, & \text{falls } \beta > \gamma \\ \epsilon^{-2}\log(\epsilon)^2, & \text{falls } \beta = \gamma \\ \epsilon^{-2-(\gamma-\beta)/\alpha}, & \text{falls } \beta < \gamma \end{cases}$$

Dabei darf $\tilde{c}$ von c_1, c_2 und c_3 abhängen.

Beweis. Betrachten wir zunächst den Monte Carlo Schätzer $\widehat{Y}_l$, so gilt nach den Rechenregeln für Erwartungswerte

$$\mathbb{E}[\widehat{Y}_l] = \begin{cases} \mathbb{E}[Q_{h_l}], & l = 0 \\ \mathbb{E}[Q_{h_l} - Q_{h_{l-1}}], & l > 0 \end{cases} \qquad (\star)$$

Wir nehmen o.B.d.A an, dass $h_0 = 1$. Ist dies nicht der Fall, lassen sich die Konstanten c_1, c_2, c_3 und $\tilde{c}$ entsprechend skalieren. Wir wählen nun $L \coloneqq \lceil \alpha^{-1}\log_2(\sqrt{2}c_1\epsilon^{-1})\rceil < \alpha^{-1}\log_2(\sqrt{2}c_1\epsilon^{-1}) + 1$. Dann gilt:

$$2^{-\alpha}\frac{\epsilon}{\sqrt{2}} < c_1 2^{-\alpha L} \leq \frac{\epsilon}{\sqrt{2}} \qquad (\star\star)$$

Mit $(\star)$ und (a) gilt dann mit $\tilde{\mathcal{H}} = \{0, \dots, L\}$

$$\left(\mathbb{E}[\widehat{Q}^{\mathrm{MLMC}}_{\tilde{\mathcal{H}},\{n_l\}_{l=0}^L} - \mathbb{E}[Q]]\right)^2 = \left(\mathbb{E}[Q_{h_L}] - \mathbb{E}[Q]\right)^2 \leq c_1 h_L^\alpha = c_1 2^{-\alpha L} \leq \frac{1}{2}\epsilon^2 \ .$$

Nach (6.8) ist

$$e(\widehat{Q}^{\mathrm{MLMC}}_{\tilde{\mathcal{H}},\{n_l\}_{l=0}^L})^2 = \underbrace{\mathbb{V}[\widehat{Q}^{\mathrm{MLMC}}_{\tilde{\mathcal{H}},\{n_l\}_{l=0}^L}]}_{(1)} + \underbrace{\left(\mathbb{E}[Q_{h_L} - Q]\right)^2}_{(2)}$$

. Für das gewählte L ist also bereits $(2) \leq \frac{1}{2}\epsilon^2$. Um also $e(\widehat{Q}^{\mathrm{MLMC}}_{\tilde{\mathcal{H}},\{n_l\}_{l=0}^L})^2 < \epsilon^2$ zu gewährleisten, müssen wir nachweisen, dass für entsprechend gewählte $\{n_l\}_{l=0}^L$ auch (1) kleiner als $\frac{1}{2}\epsilon^2$ ausfällt. Wir nutzen dazu die linke Ungleichung aus $(\star\star)$ und erhalten

$$\sum_{l=0}^{L} 2^{\gamma l} < \frac{2^{\gamma L}}{1 - 2^{-\gamma}} < \frac{2^\gamma(\sqrt{2}c_1)^{\frac{\gamma}{\alpha}}}{1 - 2^{-\gamma}}\epsilon^{-\frac{\gamma}{\alpha}} \qquad (6.9)$$

Wir führen nun eine Fallunterscheidung für das Verhältnis zwischen β und γ durch.

(i) $\beta = \gamma$:
Wir setzen $n_l = \lceil 2\epsilon^{-2}(L+1)c_2 2^{-\beta l}\rceil$ für $l = 0, \dots, L$, dann gilt mit (b):

$$\mathbb{V}[\widehat{Q}^{\mathrm{MLMC}}_{\tilde{\mathcal{H}},\{n_l\}_{l=0}^L}] = \sum_{l=0}^{L}\mathbb{V}[\widehat{Y}_l] \leq \sum_{l=0}^{L} c_2 n_l^{-1} 2^{-\beta l} \leq \frac{1}{2}\epsilon^2$$

Ṡomit gilt also $e(\widehat{Q}^{\mathrm{MLMC}}_{\tilde{\mathcal{H}},\{n_l\}_{l=0}^L}) < \epsilon$. Für die Anzahl an insgesamt benötigten Rechen-

operationen gilt dann mit (c):

$$C_\epsilon(\widehat{Q}^{\mathrm{MLMC}}_{\tilde{\mathcal{H}},\{n_l\}_{l=0}^L}) \leq c_3 \sum_{l=0}^{L} n_l 2^{\gamma l} \leq c_3 \left(2\epsilon^{-2}(L+1)^2 c_2 + \sum_{l=0}^{L} 2^{\gamma l} \right)$$

Für $\epsilon < e^{-1} < 1$ ist $1 < \log \epsilon^{-1}$ und $\epsilon^{-\frac{\gamma}{\alpha}} \leq \epsilon^{-2} \leq \epsilon^{-2}(\log \epsilon)^2$, da $\alpha \geq \frac{1}{2}\gamma$. Nutzen wir nun $L = \lceil \alpha^{-1} \log_2(\sqrt{2} c_1 \epsilon^{-1}) \rceil < \alpha^{-1} \log_2(\sqrt{2} c_1 \epsilon^{-1}) + 1$ erhalten wir

$$C_\epsilon(\widehat{Q}^{\mathrm{MLMC}}_{\tilde{\mathcal{H}},\{n_l\}_{l=0}^L}) \leq \tilde{c}_1 \epsilon^{-2} (\log \epsilon)^2, \quad \text{für ein } \tilde{c}_1 > 0$$

(ii) $\beta > \gamma$:
Wir setzen $n_l = \lceil 2\epsilon^{-2} c_2 (1 - 2^{-(\beta-\gamma)/2})^{-1} 2^{-(\beta-\gamma)l/2} \rceil$, dann ist

$$\sum_{l=0}^{L} \mathbb{V}[\widehat{Y}_l] \leq \frac{1}{2}\epsilon^2 \left(1 - 2^{-(\beta-\gamma)/2}\right) \sum_{l=0}^{L} 2^{-(\beta-\gamma)l/2} < \frac{1}{2}\epsilon^2$$

. Mit $n_l < 2\epsilon^{-2} c_2 (1 - 2^{-(\beta-\gamma)/2})^{-1} 2^{-(\beta-\gamma)l/2} + 1$ ist so

$$C_\epsilon(\widehat{Q}^{\mathrm{MLMC}}_{\tilde{\mathcal{H}},\{n_l\}_{l=0}^L}) \leq c_3 \left(2\epsilon^{-2} c_2 \left(1 - 2^{-(\beta-\gamma)/2}\right)^{-2} + \sum_{l=0}^{L} 2^{\gamma l} \right) .$$

Wiederum folgt mit (6.9), $\epsilon < e^{-1} < 1$ und $\epsilon^{-\frac{\gamma}{\alpha}} \leq \epsilon^{-2}$, dass

$$C_\epsilon(\widehat{Q}^{\mathrm{MLMC}}_{\tilde{\mathcal{H}},\{n_l\}_{l=0}^L}) \leq \tilde{c}_2 \epsilon^{-2}, \quad \text{für ein } \tilde{c}_2 > 0$$

(iii) $\beta < \gamma$
Wir setzen $n_l = \lceil \epsilon^{-2} c_2 2^{(\gamma-\beta)L/2+1} \left(1 - 2^{-(\gamma-\beta)/2}\right)^{-1} 2^{-(\beta+\gamma)l/2} \rceil$. Dann ist

$$\sum_{l=0}^{L} \mathbb{V}[\widehat{Y}_l] < \epsilon^2 2^{-(\gamma-\beta)L/2-1} \left(1 - 2^{-(\gamma-\beta)/2}\right) \sum_{l=0}^{L} 2^{(\gamma-\beta)l/2} < \frac{1}{2}\epsilon^2.$$

Durch obige Wahl von n_l erhalten wir dann

$$\begin{aligned} C_\epsilon(\widehat{Q}^{\mathrm{MLMC}}_{\tilde{\mathcal{H}},\{n_l\}_{l=0}^L}) &\leq c_3 \left(2\epsilon^{-2} c_2 2^{(\gamma-\beta)L/2} (1 - 2^{-(\gamma-\beta)/2})^{-1} \sum_{l=0}^{L} 2^{(\gamma-\beta)l/2} + \sum_{l=0}^{L} 2^{\gamma l} \right) \\ &\leq c_3 \left(2\epsilon^{-2} c_2 2^{(\gamma-\beta)L} (1 - 2^{-(\gamma-\beta)/2})^{-2} + \sum_{l=0}^{L} 2^{\gamma l} \right) . \end{aligned}$$

Nutzen wir nun ein letztes Mal ($\star\star$), so ist $2^{(\gamma-\beta)L} < \left(\sqrt{2} c_1\right)^{\frac{\gamma-\beta}{\alpha}} 2^{\gamma-\beta} \epsilon^{-(\gamma-\beta)/\alpha}$ und

mit $\epsilon < e^{-1} < 1$ gilt wegen $\alpha \geq \frac{1}{2}\beta$ auch $\epsilon^{-\frac{\gamma}{\alpha}} \leq \epsilon^{-2-(\gamma-\beta)/\alpha}$. Mit (6.9) folgt dann

$$C_\epsilon(\widehat{Q}^{\mathrm{MLMC}}_{\tilde{\mathcal{H}},\{n_l\}_{l=0}^L}) \leq \tilde{c}_3 \epsilon^{-2-(\gamma-\beta)\alpha}, \quad \text{für ein } \tilde{c}_3 > 0$$

Mit $\tilde{c} \coloneqq \max\{\tilde{c}_1, \tilde{c}_2, \tilde{c}_3\}$ folgt also die Behauptung.

□

Wir haben also gezeigt, dass für ein gegebenes $0 < \epsilon < e^{-1}$ stets ein maximales Level L und zugehörige Anzahlen an Zufallssamples $\{n_l\}_{l=0}^L$ existieren, sodass der Multilevel Monte Carlo Schätzer $\widehat{Q}^{\mathrm{MLMC}}_{\tilde{\mathcal{H}},\{n_l\}_{l=0}^L}$ im RMSE höchstens ϵ vom tatsächlichen Erwartungswert $\mathbb{E}[Q]$ entfernt ist. Außerdem haben wir unter entsprechenden Voraussetzungen an $\{n_l\}_{l=0}^L$ obere Schranken für die Anzahl der benötigten Rechenoperationen bewiesen. Verglichen mit den Kosten für die Monte Carlo Methode erhalten wir so für alle zulässigen α, β, γ einen theoretischen Kostenvorteil. Durch eine geschickte, konkrete Wahl von $\{n_l\}_{l=0}^L$ lässt sich zudem für feste Kosten $C(\widehat{Q}^{\mathrm{MLMC}}_{\tilde{\mathcal{H}},\{n_l\}_{l=0}^L}) = \sum_{l=0}^L n_l C_l$ die Varianz des Schätzers minimieren. Wie in [17] näher erklärt kann hierfür

$$n_l = \left\lceil 2\epsilon^{-2} \sqrt{\frac{\mathbb{V}[Y_l]}{C_l}} \left(\sum_{l=0}^{L} \sqrt{\mathbb{V}[Y_l] C_l} \right) \right\rceil \tag{6.10}$$

gewählt werden. Dabei ist $\mathbb{V}[Y_l]$ die geschätzte Varianz und C_l die Anzahl der benötigten Rechenoperationen für ein einzelnes Sample auf Level l. Davon ausgehend können wir nun den Algorithmus formulieren:

Algorithm 2: MLMC angewandt auf das Transportproblem

Input : $\epsilon > 0$, $l_0, L_0 \in \mathbb{N}$ und $N_0 = \{n_{l_0}, \dots, n_{L_0}\}$
Output: $\widehat{Q}^{\text{MLMC}}_{\tilde{\mathcal{H}},\{n_l\}_{l=l_0}^{L}}$

Setze $\tilde{\mathcal{H}} = \{l_0, \dots, L_0\}$ und die Anzahl der benötigten Samples $\Delta N = \{\Delta n_l\}_{l=l_0}^{L_0} = N_0$ und $i = 0$
while $\Delta n_l > 0$ *für mindestens ein* $l \in \tilde{\mathcal{H}}$ **do**
  for *alle* $l \in \tilde{\mathcal{H}}$ *mit* $\Delta n_l > 0$ **do**
    Berechne Zielfunktional und benötigte Kosten: $Y_l, C_l \leftarrow$ MonteCarloEstimator$(h_l, \Delta n_l)$
    Update $C_l, |\mathbb{E}[Y_l]|, \mathbb{V}[Y_l]$ und setze $n_l = \Delta n_l, \Delta n_l = 0$
  end
  Schätze die Exponenten α, β, γ mit den Annahmen von Satz 6.2
  Schätze optimales $N_i = \{n_{l_i}, \dots, n_{L_i}\}$ mit (6.10) und berechne $\Delta N = N_i - N_{i-1}$
  Teste auf schwache Konvergenz $|\mathbb{E}[Q_{h_{L_i}} - Q_{h_{L_i-1}}]| < (2^\alpha - 1)\frac{\epsilon}{\sqrt{2}}$
  if *nicht konvergiert* **then**
    Setze $\tilde{\mathcal{H}} = \{l_0, \dots, L_i, L_{i+1} \coloneqq L_i + 1\}$
    initialisiere $\Delta n_{L_{i+1}}$ und setze $\Delta N = \{\Delta n_l\}_{l \in \tilde{\mathcal{H}}}$
    i = i+1
  else
    $\Delta N = \{0, \dots, 0\}$
  end
end

Result: $\widehat{Q}^{\text{MLMC}}_{\tilde{\mathcal{H}},\{n_l\}_{l=l_0}^{L}}$

Dabei überprüfen wir beim Test auf schwache Konvergenz ob $|\mathbb{E}[Q - Q_{h_{L_i}}]| < \frac{\epsilon}{\sqrt{2}}$ (vgl. (6.8)). Unter der Annahme 6.1 testen wir also ob

$$|\mathbb{E}[Q_{h_{L_i}} - Q_{h_{L_i-1}}]| < (2^\alpha - 1)\frac{\epsilon}{\sqrt{2}}$$

Zu beachten ist, dass diese Version des Algorithmus, ebenso wie die in [17] etwas allgemeinere Variante von heuristischer Natur ist. Es kann je nach Problemstellung nicht garantiert werden, dass mit diesem Verfahren stets ein RMSE kleiner ϵ erreicht werden kann. Satz 6.2 liefert zwar gerade die Garantie, dass es einen Multilevel Monte Carlo Schätzer gibt, welcher einen kleineren RMSE als ϵ besitzt, für die genaue Konstruktion brauchen wir allerdings a-priori Kenntnisse über die Konstanten c_1 und c_2. Obiger Algorithmus verzichtet hingegen auf diese a-priori Kenntnisse und schätzt diese Konstanten on-the-fly. Bei der tatsächlichen Implementierung, welche wir später verwenden werden, geben wir außerdem ein $L_{\max}$ vor. Der Algorithmus bricht ab, wenn $L_{\max}$ überschritten werden würde, und liefert dann kein Ergebnis zurück. Dies liegt aber weniger

an obigem Algorithmus, sondern vielmehr an der Erzeugung des zufälligen Vektorfeldes $q(\omega_i, x) \leftarrow$ RndVecField. Hierbei werden für die Permeabilität $\kappa(\omega_i)$ log-normal verteilte Zufallsfelder mithilfe des Circulant-Embedding Algorithmus erzeugt. Um diese Erzeugung möglichst effizient zu gestalten, werden dafür nötige komplexe Eigenwertberechnungen zu Beginn vor der eigentlichen Multilevel Monte Carlo Methode für alle möglichen Level $\widehat{\mathcal{H}} = \{l_0, \ldots, L_{\max}\}$ durchgeführt. In zukünftigen Arbeiten könnte dies unter Umständen aber auch insofern verbessert werden, dass diese Berechnung ebenfalls in den MLMC Algorithmus integriert wird, aber ohne den Effizienzgewinn durch die einmalige Eigenwertberechnung für jedes Level aufzugeben.

7 Experiment

7.1 Modellproblem

In diesem Abschnitt wollen wir anhand des linearen Transportproblems auch tatsächliche praktische Resultate der Multilevel Monte Carlo Methode beleuchten. Wir setzen dazu

- $\mathcal{D} = (0,1)^2 \subset \mathbb{R}^2$ (also insbesondere wie breits zuvor in der Theorie $n = 2$)
- $\mathbb{T} = [0,T] = [0,1]$ (vgl. Bemerkung 7.1)
- $\rho_{\text{in}} \equiv 0$
- $\rho_0(x) = \begin{cases} c & , \text{ für } x \in B := [0.2, 0.8] \times [0.7, 0.8] \\ c \exp\left(\frac{|\frac{\text{dist}(x,B)}{0.15}|^2}{|\frac{\text{dist}(x,B)}{0.15}|^2 - 1} \right) & , \text{ falls } \text{dist}(x,B) > 0.15 \\ 0 & , \text{ sonst} \end{cases}$

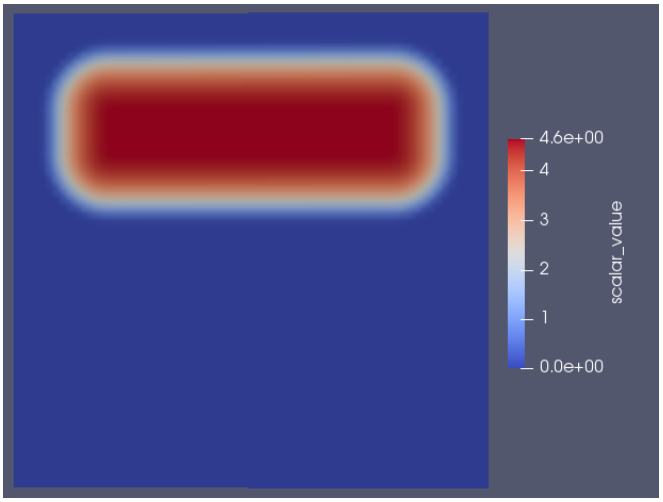
  Dabei sei c so gewählt, dass

$$\int_{\mathcal{D}} \rho_0(x) \, \mathrm{d}x = 1 \ ,$$

  wir skalieren also die zur Beginn im Rechengebiet enthaltene Masse auf 1.

Die Anfangskonzentration ρ_0 entspricht also gerade einer rechteckigen Ansammlung mit Länge 0.6 und Breite 0.1 und Mittelpunkt $(0.5, 0.75)$, welcher in beide Richtungen auf einem 0.15 breiten Streifen mit einer Exponentialfunktion in die Ebene geglättet wird. Wir stellen so für die Anfangsbedingungen unseres Modellproblems einen sehr hohen Grad an Regularität sicher. Visuell kann diese Anfangskonzentration folgendermaßen dargestellt werden:

Abbildung 3: Anfangsbedingung ρ_0



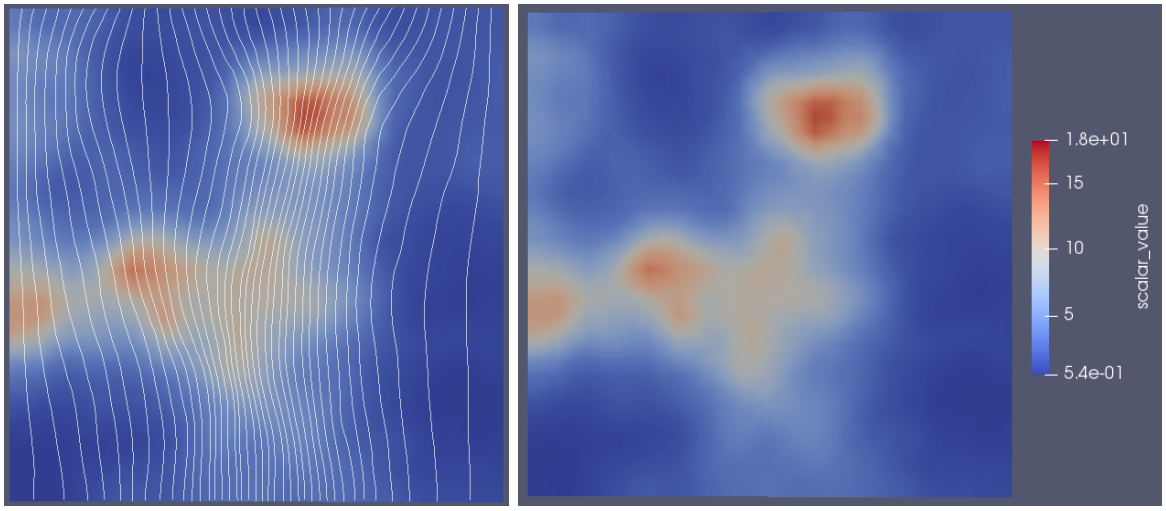
Der Maximalwert ist dabei so gewählt, dass die Gesamtmasse der Anfangsbedingung gerade den Wert 1 ergibt.
Uns interessiert nun folgende Fragestellung:
Wie groß ist der erwartete Anteil der Masse, welcher nach Ablauf des betrachteten Zeitintervalls $\mathbb{T}$ im Rechengebiet $\mathcal{D}$ verbleibt?

Wie bereits an früherer Stelle erklärt, berechnen wir hierzu ein stochastische Flussvektorfeld $q : \Omega \times \overline{\mathcal{D}} \to \mathbb{R}^2$ als Lösung des zugehörigen Potentialströmungsproblems, wobei wir dabei den Permeabilitätstensor $\kappa : \Omega \times \mathcal{D} \to \mathbb{R}_{\geq 0}$ als lognormal verteiltes Zufallsfeld modellieren. Wir identifizieren dabei die Verteilung von κ mit der zugehörigen Kovarianzfunktion:

$$C(x, y) = \sigma^2 \exp(-\frac{\|x - y\|_2^s}{\lambda^s}).$$

Dabei ist $0 < \sigma^2 < \infty$ die Varianz des zugrundeliegenden Gauß'schen Zufallsfeldes, durch $\lambda = (\lambda_1, \lambda_2) \in \mathbb{R}^2$ werden die Korrelationslängen in die verschiedenen Koordinatenrichtungen gegeben und $s \in (1, 2)$ ist ein Glättungsparameter.
Wir können auch den das lognormal verteilte Zufallsfeld und das daraus resultierende stochastische Flussvektorfeld q exemplarisch für ein $\omega \in \Omega$ grafisch darstellen. Bei der Visualisierung des Flussvektorfeldes nutzen wir dabei sogenannte 'Streamlines'. Im Wesentlichen nutzt man dabei einen Zeitintegrator, wie etwa ein Runge Kutta Verfahren, und integriert mit deren Hilfe an verschiedenen Ausgangspunkten entlang des Vektorfeldes.

Abbildung 4: Visualisierung der stochastischen Modellierung



(a) Streamlines des entstehenden Flussvektorfeldes

(b) lognormal verteiltes Zufallsfeld mit obigen Parametern

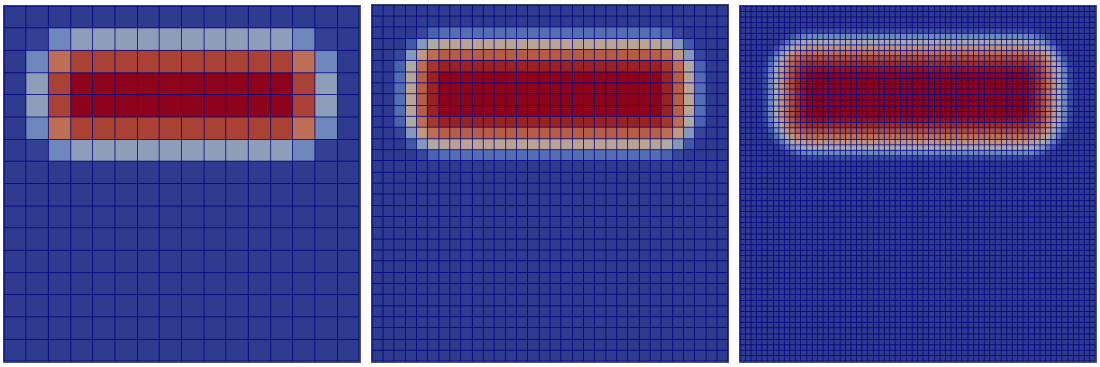
Weiter setzen wir

- $\Gamma_{\mathrm{D}} = \{x = (x_1, x_2) \in \overline{\mathcal{D}} : x_2 = 0\}$

- $\Gamma_{\mathrm{N}} = \partial\mathcal{D} \setminus \Gamma_{\mathrm{D}}$
- $g_N = \begin{cases} 0 & , \text{ falls } x \in \{x \in \Gamma_{\mathrm{N}} : x_1 \in \{0,1\}\} \\ 1 & , \text{ sonst} \end{cases}$
- $u_D \equiv 0$ auf Γ_{D}

Als Zerlegung von $\mathcal{D}$ wählen wir gleichartige Quadrate. Um auf dem geringsten Level ein Mindestmaß an Auflösung zu gewährleisten wählen wir auf Level l_0 eine Zerlegung in $256 = 16^2$ Quadrate. Dies entspricht einer Ortsdiskretisierungsschrittweite von $h_0 = \frac{1}{16} = 0.0625$. Wie bereits im Theorieteil angemerkt, wählen wir von h_0 ausgehend die uniforme Familie von Zerlegungen $\{\mathcal{T}_h\}_{h\in\mathcal{H}}$ mit $\mathcal{H} = \{h_0, h_1 := \frac{h_0}{2}, h_2 := \frac{h_1}{2} = \frac{h_0}{4}, \dots\}$. Auf Level l_1 betrachten wir also $1024 = 32^2$ und auf Level l_k dementsprechend $2^{2(k+4)}$ Quadrate. In M++ entspricht Level l_0 bei der gewählten Diskretisierung 'UnitSquare' gerade Level 4. Die Zerlegungen auf $l_0 = 4, l_1 = 5$ und $l_2 = 6$ lassen sich folgendermaßen darstellen:

Abbildung 5: Zerlegung des Gebietes $\mathcal{D}$ in Finite Elemente



(a) l_0 (256 Zellen) (b) l_1 (1024 Zellen) (c) l_2 (4096 Zellen)

Die Schrittweite für die Diskretisierung in der Zeit setzen wir auf $\Delta t = \frac{h}{8}$. Diese Wahl ist besonders hinsichtlich der Stabilität des Verfahrens wichtig. Bei zu kleinen Zeitschrittweiten treten Oszillationen in der Lösung auf. Obige Wahl hat sich für unser Problem als hinreichend erwiesen. Entsprechend unserer Fragestellung können wir nun das betrachtete Zielfunktional formulieren:

$$Q(\omega) = J(\rho(\omega)) := \int_{\mathcal{D}} \rho(\omega, x, T)\,\mathrm{d}x = \int_{\mathcal{D}} \rho(\omega, x, 1)\,\mathrm{d}x$$

Wir suchen gemäß unserer Fragestellung also gerade nach $\mathbb{E}[Q]$.

Bemerkung 7.1. Die Wahl $T = 1$ ist an dieser Stelle gerade so getroffen, dass das der Fragestellung entsprechende Zielfunktional in gewisser Weise interessant ist. Genauer ist T so gewählt, dass die im Algorithmus auftretende Varianz $\mathbb{V}[Y_l]$ 'groß' ausfällt. Ist T nämlich zu groß gewählt befindet sich für fast alle $\omega \in \Omega$ kaum noch Masse im Gebiet und

die erwartete Endmasse ist $\mathbb{E}[Q] = 0$, während für sehr kleine T Masse zum Zeitpunkt T für fast alle $\omega \in \Omega$ mit der Anfangsmasse übereinstimmt und somit $\mathbb{E}[Q] = 1$. Für $T = 1$ erhalten wir für verschiedene $\omega \in \Omega$ recht unterschiedliche Ergebnisse, da die Masse je nach Beschaffenheit des Flussvektorfeldes schneller oder langsamer durch das Gebiet transportiert wird.

7.2 Ergebnisse

8 Ausblick und Fazit

9 Notation

Folgende Tabelle soll weder stichhaltige Definitionen festlegen, noch die gesamte Notation der Thesis bis auf den letzten Index ausarbeiten. Sie soll dem Leser eher als Orientierung dienen, welche Symbole für die verschiedenen Teilbereiche genutzt werden.

Symbol	Beschreibung
div	Divergenz
∇	Gradient
$\mathcal{D} \subset \mathbb{R}^d$	Rechengebiet
$d \in \mathbb{N}$	Dimension des Rechengebietes
$d_1, d_2 \in \mathbb{N}$	Dimensionen des Integrations- und Testraumes in Beispiel 4.1
$(\Omega, \mathcal{A}, \mathbb{P})$	Wahrscheinlichkeitsraum Ω über der σ-Algebra $\mathcal{A}$ mit Wahrscheinlichkeitsmaß $\mathbb{P}$
$\mathbb{T} := (0, T]$	Zeitintervall für ein $T > 0$
f	skalarwertige Funktion
F	vektorwertige Funktion
ϕ	Testfunktion
$\omega, \omega_i, \dots \in \Omega$	Ereignis
$X, X_i, \dots$	Zufallsvariable bzw. Zufallsvektor
$x, x_i, \dots$	Realisierung der zugehörigen Zufallsvariable bzw. des Zufallsvektors
$\mathfrak{N}$	Nullmenge bzgl. $\mathbb{P}$
$\mathcal{N}(\mu, \sigma^2)$	Normalverteilung mit Parametern μ und σ^2
$\mathcal{N}_n(\mu, \sigma^2)$	multivariate Normalverteilung eines n-dim Zufallsvektors
$\widetilde{\mathcal{N}}$	standardnormalverteilte Zufallsvariable
$\mathcal{U}([a, b])$	Gleichverteilung auf dem Intervall $[a, b]$ mit $a < b$
$U, U_i, \dots$	auf $[0, 1]$ gleichverteilte Zufallsvariablen
$\mathbb{E}[X]$	Erwartungswert der Zufallsvariable X
$\mathbb{V}[X]$	Varianz der Zufallsvariablen X
n	Anzahl der betrachteten Zufallsvariablen, später auch Anzahl der betrachteten Samples
$l, \dots, L \in \mathbb{N}_0$	Level, beginnend mit $l = 0$ bis L
$l_0, \dots, L_{\max} \in \mathbb{N}_0$	Level, beginnend mit $l_0 \overset{i.Allg.}{\neq} 0$ bis $L_{\max}$
P	Interpolationsoperator aus Beispiel 4.1
Q, J	Zielfunktionale
κ	Permeabilitätstensor
q	Flussvektorfeld
Γ_{D}	Dirichletrand
Γ_{N}	Neumannrand
Γ_{in}	Einflussrand
h	Schrittweite Ortsdiskretisierung
Δt	Schrittweite Zeitdiskretisierung
ρ	Konzentration eines Stoffes im Rechengebiet

α, β, γ	Konvergenzparameter aus Satz 6.2
$e(\cdot)$	Fehlerfunktion, z.B. RMSE
Υ	Flussfunktion
$\Upsilon^\star$	numerische Flussfunktion
$\mathcal{T}$	Zerlegung von $\mathcal{D}$
$\mathcal{F}$	Menge aller Seiten der Zerlegung $\mathcal{T}$
$\mathcal{V}_\mathcal{T}$	Menge aller Knoten der Zerlegung $\mathcal{T}$
ψ_i	Seitenbasis
μ_i	Zellenbasis
$C(\cdot), C_l, \ldots$	Kosten oder Kostenfunktion
$W^{k,p}, W_0^{k,p}$	Sobolevräume (vgl. 2.1)
H^1, H_0^1	Sobolevräume (vgl. 2.1)
$L^1_{loc}(\mathcal{D})$	Raum der lokal integrierbaren Funktionen auf $\mathcal{D}$
C_c^1	Raum stetig differenzierbarer Funktionen mit kompaktem Träger
C_c^∞	Raum unendlich oft stetig differenzierbarer Funktionen mit kompaktem Träger
$W, W_h, \mathcal{Q}, \mathcal{Q}_h$	Test- und Ansatzräume der Finite Elemente Verfahren, jeweils im entsprechenden Abschnitt definiert
$(K, \mathcal{P}, \mathcal{N})$	finites Element, oft auch nur $K \in \mathcal{T}$ als Zelle
G	Anzahl der Zellenfreiheitsgrade

10 Appendix

10.1 Zusammenhang zwischen multivariater Normalverteilung und Normalverteilung

Satz 10.1. Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $X = (X_1, \ldots, X_n)$ ein Zufallsvektor mit (nicht entarteter) multivariater Normalverteilung mit Parametern $\mu = (\mu_1, \ldots, \mu_n) \in \mathbb{R}^n$ und $C = (\sigma_{ij})_{1 \leq i,j \leq n} \in \mathbb{R}^{n \times n}$.
Dann ist $\mathbb{E}[X] = \mu$ und für alle $i, j \in 1, \ldots, n$ gelten:

$$X_j \sim \mathcal{N}(\mu_j, \sigma_{jj}) \text{ und } \sigma_{ij} = \mathrm{Cov}(X_i, X_j)$$

Beweis. (fasst mehrere Resultate aus [5] zusammen)

Da C symmetrisch positiv definit ist, existiert ein invertierbares $A \in \mathbb{R}^{n \times n}$ mit $C = AA^\top$ (Cholesky-Zerlegung). Weiter sei $Y = (Y_1, \ldots, Y_n)^\top$ ein Zufallsvektor, wobei die einzelnen $Y_1, \ldots, Y_n$ unabhängige und je $\mathcal{N}(0,1)$-verteilte Zufallsvariablen sind. Durch $T(x) \coloneqq Ax + \mu$ erhalten wir somit für $x \in \mathbb{R}^k$ eine stetig differenzierbare Abbildung die den $\mathbb{R}^k$ auf sich selbst abbildet und die Funktionaldeterminante $\det A$ besitzt. Ist Y nun ein n-dimensionaler Zufallsvektor mit Dichte f, so besitzt der Zufallsvektor $Z \coloneqq AY + \mu$ nach dem Transformationssatz die Dichte

$$g(y) = \frac{f(A^{-1}(y - \mu))}{|\det A|}, \quad y \in \mathbb{R}^k.$$

Wir erhalten also mit

$$f(x) = \prod_{j=1}^{n} \left(\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x_j^2}{2} \right) \right) = \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\left(-\frac{x^\top x}{2} \right), \text{ für } x \in \mathbb{R}^n$$

$$g(y) = \frac{1}{(2\pi)^{\frac{n}{2}} |\det A|} \exp\left(-\frac{1}{2} \left(A^{-1}(y - \mu) \right)^\top \left(A^{-1}(y - \mu) \right) \right), \text{ für } y \in \mathbb{R}^n.$$

Wegen $C = AA^\top$, $(A^{-1})^\top = (A^\top)^{-1}$ und $|\det A| = \sqrt{\det C}$ ist somit

$$g(y) = \frac{1}{(2\pi)^{\frac{n}{2}} \sqrt{\det C}} \exp\left(-\frac{1}{2} (y - \mu)^\top C^{-1} (y - \mu) \right), \text{ für } y \in \mathbb{R}^n.$$

Insbesondere ist also $Z \sim X \sim \mathcal{N}_n(\mu, C)$.
Seien nun $A = (a_{ij})_{1 \leq i,j \leq n}$, dann folgt

$$X_j \sim \sum_{l=1}^{n} a_{jl} Y_l + \mu_j \ .$$

Wegen $K_l \coloneqq a_{ij} Y_l \sim \mathcal{N}(0, a_{jl}^2)$ und der Unabhängigkeit der Y_l (mit dem sogennanten Blockungslemma folgt somit Unabhängigkeit der K_l) gilt nach dem Additionsgesetz für

die Normalverteilung

$$X_j \sim \mathcal{N}\left(\mu_j, \sum_{l=1}^{n} a_{jl}^2\right).$$

Aus $C = AA^\top$ folgt schließlich $\sigma_{jj} = \sum_{l=1}^{n} a_{jl}^2$. Es bleibt nun also noch zu zeigen, dass $\mathbb{E}[X] = \mu$ und $\sigma_{ij} = \mathrm{Cov}(X_i, X_j)$. Wir bezeichnen mit $\mathrm{Cov}(X) \coloneqq (\mathrm{Cov}(X_i, X_j))_{1 \le i,j \le n}$ die Kovarianzmatrix. Es ist $\mathbb{E}[Y] = 0$ und $\mathrm{Cov}(Y) = I_n$. Es gilt also:

$$\mathbb{E}[X] = \mathbb{E}[AY + \mu] = (\mathbb{E}[\sum_{l=1}^{n} K_l + \mu_j]) = (\mathbb{E}[\sum_{l=1}^{n} a_{jl} Y_l + \mu_j]) = A\mathbb{E}[Y] + \mu = \mu$$
$$\mathrm{Cov}(X) = \mathrm{Cov}(AY + \mu) = \mathrm{Cov}(AY) = A\mathrm{Cov}(Y)A^\top = AA^\top = C$$

□

10.2 Referenzzelle und Hybridisierung

10.2.1 Referenzzelle

Bevor wie uns an dieser Stelle der Hybridisierung widmen können, wollen wir noch kurz auf einen wichtigen Aspekt der Implementierung finiter Elemente eingehen. An dieser Stelle hat es sich nämlich bereits oft als nützlich erwiesen, eine sogenannte Referenzzelle einzuführen. Statt sich die Daten jeder Zelle statisch zu speichern und dann darauf zuzugreifen, gehen wir dabei stets von der Referenzzelle aus und können über je eine linear affine Abbildung in den tatsächlichen Zellen operieren. Da wir ausschließlich Vierecke verwenden wollen wir uns dementsprechend an dieser Stelle auf Vierecke beschränken.

Definition 10.2. Das Referenzviereck $\square$ ist definiert als

$$\hat{K} \coloneqq \mathrm{conv}\{\hat{\mathcal{V}}\}, \text{ wobei } \hat{\mathcal{V}} \coloneqq \left\{\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}0\\1\end{pmatrix}, \begin{pmatrix}1\\1\end{pmatrix}\right\}$$

Die Seiten von $\square$ sind

$$\hat{F}_0 \coloneqq \mathrm{conv}\left\{\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1\\0\end{pmatrix}\right\}$$
$$\hat{F}_1 \coloneqq \mathrm{conv}\left\{\begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}1\\1\end{pmatrix}\right\}$$
$$\hat{F}_2 \coloneqq \mathrm{conv}\left\{\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}0\\1\end{pmatrix}\right\}$$
$$\hat{F}_3 \coloneqq \mathrm{conv}\left\{\begin{pmatrix}0\\1\end{pmatrix}, \begin{pmatrix}0\\0\end{pmatrix}\right\}$$

Weiter sei $\{\hat{\psi}_i\}_{i=0}^3$ die Seitenbasis aus Hütchenfunktionen und $\hat{n}$ sei der äußere Norma-

lenvektor von $\hat{K}$.

Abbildung 6: Referenzzelle

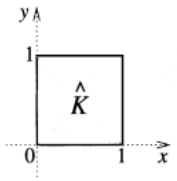


Abbildung modifiziert aus [23] Seite 51

Bemerkung. $\forall i,j \in \{0,1,2,3\} : \int_{F_j} \hat{\psi}_i \cdot \hat{n}\,\mathrm{d}a = \delta_{i,j}$ und $\hat{\psi}_i \in \mathbb{P}_1(\hat{K}, \mathbb{R}^2)$.

Weiter setzen wir noch

$$\text{die Menge der Seiten} \qquad \hat{\mathcal{F}} := \left\{\hat{F}_0, \hat{F}_1, \hat{F}_2, \hat{F}_3\right\}$$
$$\text{und den Seitenansatzraum} \qquad \hat{W} := \operatorname{span}\left\{\psi_0, \psi_1, \psi_2, \psi_3\right\}.$$

Transformation von $\hat{K}$ zu K: Für ein beliebiges $K \in \mathcal{K}$ wollen wir jetzt eine Seitenbasis $\{\psi_0^K, \psi_1^K, \psi_2^K, \psi_3^K\}$ berechnen (Wie bisher gegeben durch $\forall i \in \{0,1,2,3\} : \psi_i^K \in \mathbb{P}_1(K, \mathbb{R}^2)$ und $\int_{F_j^K} \psi_i^K \cdot n^K\,\mathrm{d}a = \delta_{i,j}$, wobei n^K äußere Normale von K und F_j^K beliebige Seite von K). Dazu betrachten wir die affine Transformationsabbildung φ_K von $\hat{K}$ zu K:

$$\varphi_K : \hat{K} \to K, \varphi_K(\xi) = z_{0,K} + B_K\xi \text{ mit passenden } B_K \in \mathbb{R}^{2\times 2} \text{ und}$$
$$J_K := \det(B_K) > 0.$$

Abbildung 7: Affine lineare Transformation von $\hat{K}$ nach K

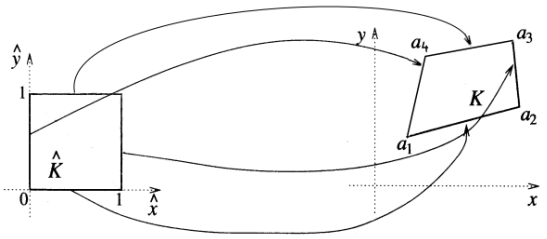


Abbildung modifiziert aus [23] Seite 53

Lemma 10.3. Es gilt: $\tilde{n}^K = \frac{1}{|B_K^{-T}\hat{n}|} B_K^{-T}\hat{n}$ ist Normale zu ∂K.

Die Seitenbasis auf K ist dann gegeben durch

$$\psi_i^K = J_K^{-1} B_K \hat{\psi}_i \circ \varphi_K \ (i \in \{0,1,2,3\})$$

Die globale Seitenbasis $\{\psi_j\}_{j=1}^{|\mathcal{F}|}$ auf $\mathcal{D}$ erhalten wir dann mithilfe einer weiteren Abbildung l, die zwischen der Seitennummerierung in einer Zelle K und der globalen Seitennummerierung vermittelt. Es ist dabei

$$l : \mathcal{K} \times \{0,1,2,3\} \to \{1,\dots,|\mathcal{F}|\}, (K,i) \mapsto l(K,i).$$

Wir setzen nun also $\psi_j (j \in \{1,\dots,|\mathcal{F}|\})$ durch

$$\psi_j(x) = \begin{cases} \psi_i^K(x), & \text{falls } j = l(K,i) \\ 0, & \text{sonst.} \end{cases}$$

Bemerkung. Für alle Zellen $K \in \mathcal{K}$ von denen F_j eine anliegende Seite ($\overline{K} \cap F_j \neq \emptyset$) und F_j lokal mit $i \in \{0,1,2,3\}$ nummeriert ist, gilt:

$$\psi_j|_K = \psi_i^K.$$

10.2.2 Hybridisierung

Wir betrachten die Räume

$$W_K \coloneqq \left\{ \psi_K : K \to \mathbb{R}^2 : \psi_K = J_K^{-1} B_K \hat{\psi} \circ \varphi_K^{-1}, \hat{\psi} \in \hat{W} \right\}$$

$$W_{\mathcal{K}} \coloneqq \prod_{K \in \mathcal{K}} W_K, \qquad M_h \coloneqq \prod_{F \in \mathcal{F}} \mathbb{P}_0(F)$$

$$M_h(u_D) \coloneqq \left\{ \mu_h \in M_h : \forall F \subset \Gamma_D \int_F \mu_h \,\mathrm{d}a = \int_F u_D \,\mathrm{d}a \right\}$$

An dieser Stelle nutzen wir die folgende Äquivalenz:

$$\psi_h \in W_h \iff \left[\psi_h \in W_{\mathcal{K}} \text{ und } (\psi_{K_1} - \psi_{K_2}) \cdot n^F = 0 \ (F = \partial K_1 \cap \partial K_2 \in \mathcal{F}^\circ) \right]$$

Und untersuchen folgendes Problem:

$$\text{Bestimmme } (q_h, u_h, \lambda_h) \in W_{\mathcal{K}} \times \mathcal{Q}_h \times M_h(u_D) \text{ mit}$$

$$\begin{cases} (1) \displaystyle\int_K \kappa^{-1} q_h \psi_K \,\mathrm{d}x - \int_K u_h \operatorname{div}(\psi_K) \,\mathrm{d}x = -\int_{\partial K} \lambda_h \psi_K \cdot n^K \,\mathrm{d}a \\ (2) \displaystyle\int_K \operatorname{div}(q_h) \phi_K \,\mathrm{d}x = 0 \\ (3) \displaystyle\sum_{K \in \mathcal{K}} \int_{\partial K} q_h \cdot n \mu_h \,\mathrm{d}a = -\int_{\Gamma_N} g_N \mu_h \,\mathrm{d}a \end{cases}$$

$$\text{für alle } K \in \mathcal{K}, \psi_K \in W_K, \phi_K \in \mathcal{Q}_h \text{ und } \mu_h \in M_h(0)$$

Dieses Problem ist äquivalent zu dem diskreten gemischten FE-Problem, welches wir

zuvor betrachtet haben:

$$\text{Bestimme } (q_h, u_h) \in W_h(-g_N) \times \mathcal{Q}_h \text{ mit}$$
$$\begin{cases} \int_\Omega \kappa^{-1} q_h \cdot \psi_h \,\mathrm{d}x - \int_\Omega u_h \operatorname{div}(\psi_h)\,\mathrm{d}x = -\int_{\Gamma_D} u_D \psi_h \cdot n \,\mathrm{d}a \\ -\int_\Omega \operatorname{div}(q_h)\phi_h \,\mathrm{d}x = 0 \end{cases}$$
$$\text{für alle } (\psi_h, \phi_h) \in W_h(0) \times \mathcal{Q}_h$$

Für ein festes $K \in \mathcal{K}$ ergibt sich mit der Wahl einer Basis von W_K, $\mathcal{Q}_h$ und M_h eine Formulierung als LGS mit Nebenbedingung, wobei $\underline{q}_K \coloneqq \underline{R}_K \underline{q}$, $\underline{u}_K \coloneqq \underline{R}_K \underline{u}$.

$$\text{Bestimme } \underline{q}, \underline{u} \text{ und } \underline{\lambda} \text{ mit}$$
$$\begin{cases} (1) \begin{pmatrix} \underline{A}_K & \underline{B}_K \\ \underline{B}_K^T & 0 \end{pmatrix} \begin{pmatrix} \underline{q}_K \\ \underline{u}_K \end{pmatrix} = \begin{pmatrix} -\underline{C}_K \underline{R}_K \underline{\lambda} \\ 0 \end{pmatrix} \\ (2) \sum_{K \in \mathcal{K}} (\underline{R}_K \underline{\mu})^T \underline{C}_K \underline{q}_K \;\; = \underline{\mu}^T \underline{b} \end{cases}$$
$$\text{für alle } \underline{\mu} \text{ mit } \underline{\mu}[F] = 0 \text{ für } F \in \Gamma_D \cap \mathcal{F}$$

$$\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \\ \underline{\mu} \end{pmatrix}^T \underbrace{\left(\begin{array}{cccccc|c} \underline{A}_{K_1} & \underline{B}_{K_1} & & & & & \underline{C}_{K_1}\underline{R}_{K_1} \\ \underline{B}_{K_1} & 0 & & & & & 0 \\ & & \underline{A}_{K_2} & \underline{B}_{K_2} & & & \underline{C}_{K_2}\underline{R}_{K_2} \\ & & \underline{B}_{K_2} & 0 & & & 0 \\ & & & & & \ddots & \\ \hline \underline{R}_{K_1}^T \underline{C}_{K_1}^T & 0 & \underline{R}_{K_2}^T \underline{C}_{K_2}^T & 0 & & & 0 \end{array}\right)}_{=: \left(\begin{array}{c|c} \underline{D} & \underline{E} \\ \hline \underline{E}^T & 0 \end{array}\right)} \underbrace{\begin{pmatrix} \underline{q}_{K_1} \\ \underline{u}_{K_1} \\ \underline{q}_{K_2} \\ \underline{u}_{K_2} \\ \vdots \\ \underline{\lambda} \end{pmatrix}}_{=: \begin{pmatrix} \left(\begin{smallmatrix} \underline{q}_{K_i} \\ \underline{u}_{K_i} \end{smallmatrix}\right)_{K_i \in \mathcal{K}} \\ \underline{\lambda} \end{pmatrix}} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ \underline{\mu}^T \underline{b} \end{pmatrix}$$

Mit dem Schurkomplement $\underline{S} \coloneqq \underline{E}^T \underline{D}^{-1} \underline{E}$ folgt

$$\underline{\mu}^T \underline{S}\, \underline{\lambda} = \underline{\mu}^T \underline{b} \text{ für alle } \underline{\mu} \text{ mit } \underline{\mu}[F] = 0 \text{ für } F \in \Gamma_D \cap \mathcal{F}$$

Sobald wir $\underline{\lambda}_k \coloneqq \underline{R}_K \underline{\lambda}$ bestimmt haben, können wir auch das obere LGS (1) lösen, um $\underline{q}_K$ und $\underline{u}_K$ zu erhalten.

Literatur

[1] M++ (meshes, multigrid and more). `http://www.math.kit.edu/ianm3/page/mplusplus/de`. Accessed: 2019-10-17.

[2] D. Braess. *Finite Elemente: Theorie, schnelle Löser und Anwendungen in der Elastizitätstheorie*. Springer-Verlag, 2013.

[3] S. Brenner and R. Scott. *The mathematical theory of finite element methods*, volume 15. Springer Science & Business Media, 2007.

[4] F. Brezzi and M. Fortin. *Mixed and hybrid finite element methods*, volume 15. Springer Science & Business Media, 2012.

[5] M. Brokate, N. Henze, F. Hettlich, A. Meister, G. Schranz-Kirlinger, and T. Sonar. *Grundwissen Mathematikstudium*. Springer, 2016.

[6] J. Charrier. Strong and weak error estimates for elliptic partial differential equations with random coefficients. *SIAM Journal on numerical analysis*, 50(1):216–246, 2012.

[7] K. A. Cliffe, M. B. Giles, R. Scheichl, and A. L. Teckentrup. Multilevel monte carlo methods and applications to elliptic pdes with random coefficients. *Computing and Visualization in Science*, 14(1):3, 2011.

[8] B. Cockburn, S. Hou, and C.-W. Shu. The runge-kutta local projection discontinuous galerkin finite element method for conservation laws. iv. the multidimensional case. *Mathematics of Computation*, 54(190):545–581, 1990.

[9] B. Cockburn, G. E. Karniadakis, and C.-W. Shu. The development of discontinuous galerkin methods. In *Discontinuous Galerkin Methods*, pages 3–50. Springer, 2000.

[10] B. Cockburn and C.-W. Shu. Tvb runge-kutta local projection discontinuous galerkin finite element method for conservation laws. ii. general framework. *Mathematics of computation*, 52(186):411–435, 1989.

[11] B. Cockburn and C.-W. Shu. The runge–kutta discontinuous galerkin method for conservation laws v: multidimensional systems. *Journal of Computational Physics*, 141(2):199–224, 1998.

[12] B. Cockburn and C.-W. Shu. Runge–kutta discontinuous galerkin methods for convection-dominated problems. *Journal of scientific computing*, 16(3):173–261, 2001.

[13] C. R. Dietrich and G. N. Newsam. Fast and exact simulation of stationary gaussian processes through circulant embedding of the covariance matrix. *SIAM Journal on Scientific Computing*, 18(4):1088–1107, 1997.

[14] M. Dobrowolski. *Angewandte Funktionalanalysis: Funktionalanalysis, Sobolev-Räume und elliptische Differentialgleichungen*. Springer-Verlag, 2010.

[15] A. Ern and J.-L. Guermond. Theory and practice of finite elements. 2004. *Applied Mathematical Sciences*, 2004.

[16] L. C. Evans. *Partial differential equations*. American Mathematical Society, Providence, R.I., 2010.

[17] M. B. Giles. Multilevel monte carlo methods. *Acta Numerica*, 24:259–328, 2015.

[18] M. Hanke-Bourgeois. *Grundlagen der numerischen Mathematik und des wissenschaftlichen Rechnens*, volume 1. Springer, 2002.

[19] R. Hartmann. Numerical analysis of higher order discontinuous Galerkin finite element methods. In H. Deconinck, editor, *VKI LS 2008-08: CFD - ADIGMA course on very high order discretization methods, Oct. 13-17, 2008*. Von Karman Institute for Fluid Dynamics, Rhode Saint Genèse, Belgium, 2008.

[20] S. Heinrich. *Random approximation in numerical analysis*. Universität Kaiserslautern. Fachbereich Informatik, 1992.

[21] S. Heinrich. Multilevel monte carlo methods. In *International Conference on Large-Scale Scientific Computing*, pages 58–67. Springer, 2001.

[22] A. Klenke. *Wahrscheinlichkeitstheorie*, volume 1. Springer, 2006.

[23] P. Knabner and L. Angermann. *Numerik partieller Differentialgleichungen: eine anwendungsorientierte Einführung*. Springer-Verlag, 2013.

[24] P. Kumar, P. Luo, F. J. Gaspar, and C. W. Oosterlee. A multigrid multilevel monte carlo method for transport in the darcy–stokes system. *Journal of Computational Physics*, 371:382–408, 2018.

[25] B. Lapeyre, E. Pardoux, and R. Sentis. *Introduction to Monte Carlo methods for transport and diffusion equations*, volume 6. Oxford University Press on Demand, 2003.

[26] P. Lesaint and P.-A. Raviart. On a finite element method for solving the neutron transport equation. *Publications mathématiques et informatique de Rennes*, (S4):1–40, 1974.

[27] T. E. Peterson. A note on the convergence of the discontinuous galerkin method for a scalar hyperbolic equation. *SIAM Journal on Numerical Analysis*, 28(1):133–140, 1991.

[28] W. H. Reed and T. Hill. Triangular mesh methods for the neutron transport equation. Technical report, Los Alamos Scientific Lab., N. Mex.(USA), 1973.

[29] G. R. Richter. An optimal-order error estimate for the discontinuous galerkin method. *Mathematics of Computation*, 50(181):75–88, 1988.

[30] J. E. Roberts and J.-M. Thomas. Mixed and hybrid methods. 1991.

[31] V. Schmidt. *Stochastic geometry, spatial statistics and random fields*. Springer, 2014.

[32] T. J. Sullivan. *Introduction to uncertainty quantification*, volume 63. Springer, 2em015.

Erklärung

Ich versichere wahrheitsgemäß, die Arbeit selbstständig verfasst, alle benutzten Hilfsmittel vollständig und genau angegeben und alles kenntlich gemacht zu haben, was aus Arbeiten anderer unverändert oder mit Abänderungen entnommen wurde, sowie die Satzung des KIT zur Sicherung guter wissenschaftlicher Praxis in der jeweils gültigen Fassung beachtet zu haben.

Ort, den Datum