Diplomarbeit

Stochastische Gradientenverfahren zur Optimerung unter Echtzeitbedingungen in der adaptiven Optik

Johannes Lüdke

23. Februar 2012

Betreuer: Prof. Dr. Peter Kunkel

Mathematische Fakultät, Universität Leipzig

geschrieben als Diplomand bei Dr. rer. nat. Uwe Völker

Institut für Technische Physik des Deutschen Zentrums für Luft- und Raumfahrt e.V. (DLR)

Inhaltsverzeichnis

Einleitung

Diese Arbeit befasst sich mit den stochastischen Gradientenverfahren als Methoden der numerischen Optimierung, die für eine Anwendung in der adaptiven Optik untersucht werden. Sie wird insbesondere für den Fall der Propagation von Licht durch die turbulente Atmosphäre zur Verbesserung der Abbildungseigenschaften eines optischen Systems verwendet. Durch die zeitliche Dynamik der Turbulenz ergibt sich die Echtzeit-Anforderung, sie wird daher zusammen mit dem physikalischen Hintergrund der Anwendung im ersten Kapitel dargestellt.

Anschließend wird dies mathematisch gefasst und die sich ergebenden Anforderungen werden erläutert. Der Lösungsansatz und die zu betrachtenden Verfahren werden vorgestellt, insbesondere das *Simultaneous Perturbation Stochastic Approximation* Verfahren (SPSA-Verfahren), das aufgrund der geringen Anzahl benötigter Zielfunktionsauswertungen pro Iteration sehr gut zu den Echtzeit-Anforderungen der Anwendung passt.

Bevor zur theoretischen Analyse dieser Verfahren übergegangen wird, folgt ein Kapitel, das die benötigte wahrscheinlichkeitstheoretische Grundlage für die Behandlung der stochastischen Gradientenverfahren schafft.

Zur Definition und Abgrenzung des Verfahrens werden in Kapitel 4 zunächst die Gradientenverfahren und das ableitungsfreie Verfahren des steilsten Abstiegs behandelt. Für das diesem Verfahren entsprechende *Stochastic-Approximation*-Verfahren *Finite Differences Stochastic Approximation* wie auch für das vielversprechende SPSA-Verfahren wird die Konvergenz bis auf ein Theorem für *Stochastic-Approximation*-Verfahren aus [KC78] zurückgeführt. Dabei wird dem Beweis aus [Spa05] zunächst gefolgt, u.a. aufgrund von Messbarkeitsüberlegungen werden dann aber geänderte Voraussetzungen verwendet.

Für das SPSA-Verfahren wird außerdem ein Ausblick auf die asymptotische Theorie gegeben, die jedoch aufgrund der Echtzeitbedingungen in der Anwendung in dieser Arbeit keine entscheidende Bedeutung hat.

Das Optimierungsverfahren ist dann an einem realen adaptiv-optischen System im Labor getestet worden. Kapitel 5 beschäftigt sich mit der Umsetzung des Verfahrens in einen Algorithmus. Dazu werden Vorschläge zur Parameterwahl, welche die Konvergenztheorie berücksichtigen, und Beispiel-Code in C++ angegeben.

In Kapitel 6 wird dann konkret die Nutzung des Verfahrens für die Anwendung am Beispiel des Labor-Testsystems beschrieben. Dabei werden die Bedingungen des Verfahrens bezogen auf die Anwendung gewürdigt. Nach der Darstellung und Bewertung eines Optimierungsdurchgangs wird darauf eingegangen, unter welchen Bedingungen das SPSA-Verfahren sinnvoll für die adaptive Optik eingesetzt werden kann und ein Ausblick auf die mögliche Verwendung in einem adaptiv-optischen System gegeben.

Kapitel 1

Physikalisches Modell

1.1 Einleitung

Im Kapitel 1 wird der physikalische Hintergrund der betrachteten Anwendung kurz beschrieben. Das Modell der atmosphärischen Turbulenz wird erläutert, wichtige Kenngrößen werden dazu eingeführt und das *Frozen-Turbulence*-Modell wird vorgestellt. Daraufhin wird in die adaptive Optik eingeführt und die Methoden der Wellenfrontsensor-losen und Wellenfrontsensor-gestützten adaptiven Optik werden gegenübergestellt. Im Anschluss werden noch technische Aspekte erörtert, um eine Abschätzung der zeitlichen Beschränkung zu erhalten.

1.2 Modell der atmosphärischen Turbulenz

In diesem Abschnitt wird die Ursache der atmosphärischen Störungen kurz erklärt, wichtige Kenngrößen werden dazu eingeführt und das *Frozen-Turbulence*-Modell wird vorgestellt. Anschließend wird noch auf das Strehl-Verhältnis als einer Form der Charakterisierung der Güte einer Abbildung eingegangen.

Atmosphärische Störungen

In der astronomischen Sternenbeobachtung (*Imaging*-Anwendung der Astronomie) hat das einfallende Licht eines Sterns die Form einer ebenen Welle, die Wellenvektoren des Lichts fallen also parallel ein. Aufgrund des typischerweise großen Abstands ist diese Näherung dort gültig, wird aber auch sonst oft als zumindest lokal gültige Annahme aufgefasst. Im monochromatischen Fall gilt für die elektrische Feldstärke des Lichts als elektromagnetische Welle

$$\begin{aligned} \vec{E}(\vec{x},t) &= \vec{E}(\vec{x})\ \mathrm{e}^{i\varphi_0}\ \mathrm{e}^{-i(\omega t-\vec{k}\cdot\vec{x})} \\ &= \vec{E}(\vec{x})\ \mathrm{e}^{i(\vec{k}\cdot\vec{x}+\varphi_0)}\ \mathrm{e}^{-i\omega t} \end{aligned} \tag{1.1}$$

am Ort $\vec{x}$ zum Zeitpunkt t mit (reellem) Amplitudenfaktor $\vec{E}(\vec{x})$, Kreisfrequenz ω und Wellenvektor $\vec{k}$.

Dabei ist $\varphi(\vec{x}) = \vec{k}\cdot\vec{x}+\varphi_0$ die Phase des schwingenden elektrischen Feldes, das die Lichtwelle beschreibt. Die *Wellenfront* ist eine Fläche gleicher Phase,

d.h.

$$\begin{aligned} \mathbb{W} &= \{\vec{x}: \quad \varphi(\vec{x}) = \varphi^*\} \ , \\ \text{Darstellung} \quad \vec{x} &= (x,\ y,\ \Phi(x,y))^T, \end{aligned} \tag{1.2}$$

bei einer ebenen ungestörten Welle also eine ebene Fläche. Man kann Φ wie folgt darstellen:

$$\Phi(x,y) = n\,\lambda\,\Delta\phi(x,y), \tag{1.3}$$

wobei $\Delta\phi(x,y)$ der Phasenfehler zum jeweiligen Punkt ist, λ die Wellenlänge und n eine natürliche Zahl.

Der Brechungsindex eines Materials beeinflusst die Geschwindigkeit, mit der das Licht im Medium propagiert. Vernachlässigt man weitere Effekte, so gilt

$$v = \frac{c_0}{n(\lambda)}, \tag{1.4}$$

wobei n der Brechungsindex des Mediums, λ die Wellenlänge, c_0 die Vakuum-Lichtgeschwindigkeit (Naturkonstante) und v die Lichtgeschwindigkeit im Medium ist.

Neben der Wellenlängenabhängigkeit des Brechungsindex (Dispersion) besitzt der Brechungsindex von Luft auch eine Temperatur- und Druckabhängigkeit, d.h. das einfallende Licht propagiert unterschiedlich schnell durch wärmere und kältere Bereiche der Atmosphäre. Genauer gilt für den Brechungsindex n zum Zeitpunkt t am Punkt $\vec{x}$

$$n(\vec{x},t) = n_0 + n_1(\vec{x},t)$$

mit dem mittleren Brechungsindex der Atmosphäre $n_0 \approx 1$ und einem fluktuierenden Anteil n_1. Die Orts- und Zeitabhängigkeit des Anteils n_1 kann durch die Abhängigkeit dieses Anteils von Druck und Temperatur erklärt werden, die im vorliegenden inhomogenen Medium zeitlichen und örtlichen Änderungen unterworfen sind. Für diese Abhängigkeit von Druck und Temperatur gilt im sichtbaren Spektralbereich nach [RW96, Kapitel 3.2] näherungsweise

$$n_1(p,T) = n(p,T) - 1 = \frac{77.6p}{T}\ 10^{-6} \tag{1.5}$$

mit der Temperatur T in Kelvin und dem Luftdruck p in mbar.

Die Phasenfehler und Wellenfrontstörungen entstehen aus diesen Fluktuationen des Brechungsindex. Die Lichtwellen durchlaufen ein inhomogenes Medium, welches im Fall atmosphärischer Turbulenz auch eine zeitliche Dynamik aufweist. Mithin resultiert eine deformierte Wellenfront, vergleiche Abbildungen 1.1 und 1.2.

Kenngrößen – r_0, θ_0, f_G

In diesem Unterabschnitt wird eine Einführung in die wichtigsten Kenngrößen zur Charakterisierung atmosphärischer Turbulenz gegeben, [Tys00, Kapitel 1-2] folgend.

In den folgenden Formeln ist C_n^2 die *atmosphärische Strukturkonstante*: Sie ist ein Maß für die Stärke der Turbulenz und hängt von der Höhe und der Temperatur ab, ändert sich also im Jahres- und Tagesverlauf, ja sogar minütlich. Bei astronomischen Anwendungen bewegt sich C_n^2 in der Größenordnung von

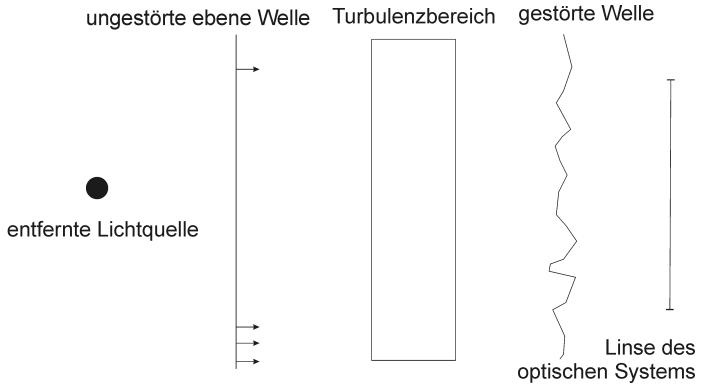


Abbildung 1.1: Atmosphärische Turbulenz in der Astronomie – einfallendes Licht eines weit entfernten Objekts erhält Wellenfront-Störungen bei Propagation durch die Erdatmosphäre auf dem Weg zum optischen System. Abbildung sinngemäß aus ROGGEMANN und WELSH: *Imaging through Turbulence*, S.66 [RW96].

$10^{-15} \dots 10^{-18}\,\mathrm{m}^{-2/3}$. In größerer Höhe nehmen die Dichte der Luft und der Temperaturgradienten ab, weniger dichte Luft bedeutet geringere Turbulenz, und oberhalb der Jetstream-Winde in etwa 10km Höhe erreicht C_n^2 die Größenordnung $10^{-18}\,\mathrm{m}^{-2/3}$, [Tys00, S. 35]. Bei bodennahen Anwendungen steigt die Größenordnung auf $10^{-14}\,\mathrm{m}^{-2/3}$ bis $10^{-12}\,\mathrm{m}^{-2/3}$ (z.B. mittags im Sommer) an, vergleiche die Messung auf der Freistrahlstrecke des *DLR* in Lampoldshausen [Grü10, Abschnitt 3.1].

Zur Formulierung der Kenngrößen der atmosphärischen Turbulenzen wird im Folgenden über den Propagationsweg des einfallenden Lichts, d.h. über die Gerade vom Aussendepunkt $\vec{x}_1$ bis zum Empfangspunkt $\vec{x}_2$, integriert. k ist die (Kreis-)Wellenzahl, $k = |\vec{k}| = \frac{2\pi}{\lambda}$ und β der Zenitwinkel (Winkel unter dem Zenit). Die Gerade durch Aussendepunkt $\vec{x}_1$ und Empfangspunkt $\vec{x}_2$ bezeichnet man als optische Achse.
Die wichtigsten Kenngrößen sind:

1. Der *Friedparameter* r_0, auch *atmosphärische Kohärenzlänge* genannt:

$$r_0 = \left(0.423\ k^2\ \sec\beta \int\limits_{\text{Weg}} C_n^2(h)\ dh \right)^{-\frac{3}{5}}. \tag{1.6}$$

   Dies gilt für ebene Wellen und $\beta < 60°$. In der allgemeinen Definition bezieht man sich auf eine Wellenfrontfehler-Grenze von $\mathbb{V}\mathrm{ar}(\varphi) = 1\,\mathrm{rad}^2$.[1] r_0 beschreibt die räumliche Kohärenz und bewegt sich für sichtbares Licht abhängig von der Turbulenzstärke in der Größenordnung einiger Zentimeter bis zu wenigen Millimetern.

   In erster Näherung ist der Friedparameter so definiert, dass das Auflösungsvermögen von Teleskopen mit einer Apertur, also einer Öffnungsweite, kleiner als r_0 nicht verringert wird.

[1] $\mathbb{V}\mathrm{ar}(X)$ ist eine Bezeichnung für die Varianz einer Zufallsgröße X, siehe Definition 3.36.

2. Der isoplanatische Winkel θ_0:

$$\theta_0 = \left(2.91 \; k^2 \; \sec^{\frac{8}{3}} \beta \int\limits_{\text{Weg}} C_n^2(h) h^{\frac{5}{3}} dh \right)^{-\frac{3}{5}} . \tag{1.7}$$

   Die Wellenfrontstörungen durch die Atmosphäre sind anisoplanatisch, also abhängig von der Richtung, aus der das Licht empfangen wird. Der isoplanatische Winkel ist ein Maß dafür, unter welchem Winkelabstand man noch von nahezu wirkungsgleichen atmosphärischen Störungen ausgehen kann. Typische Werte sind hier $7 \ldots 10\,\mu\text{rad}$ (bei $\lambda = 500\,\text{nm}$). In der allgemeinen Definition bezieht man sich ebenfalls auf die Wellenfrontfehler-Grenze $\mathbb{V}\text{ar}(\varphi) = 1\,\text{rad}^2$.

3. Die *Greenwood-Frequenz* f_G – sie beschreibt die zeitliche Charakterstik:

$$f_G = 2.31 \lambda^{-\frac{6}{5}} \left(\sec \beta \int\limits_{\text{Weg}} C_n^2(h) \; v_{\text{Wind}}^{\frac{5}{3}}(h) dh \right)^{\frac{3}{5}} , \tag{1.8}$$

   $v_{\text{Wind}}(h)$ ist dabei die transversale Windgeschwindigkeit in der Höhe h.

Der Kehrwert der Greenwood-Frequenz wird als *atmosphärische Zeitkonstante* τ_0 bezeichnet, $\tau_0 = \frac{1}{f_G}$.

In den Formeln (1.6) bis (1.8) beschreibt $C_n^2(h)$ ein höhenabhängiges Turbulenzprofil, das die Entstehungsgeschichte der wissenschaftlichen Betrachtung der optischen Eigenschaften von Turbulenz in der Atmosphäre reflektiert. Bei horizontalen Anwendungen wird üblicherweise ein konstanter Wert von C_n^2 über den Propagationspfad angenommen.

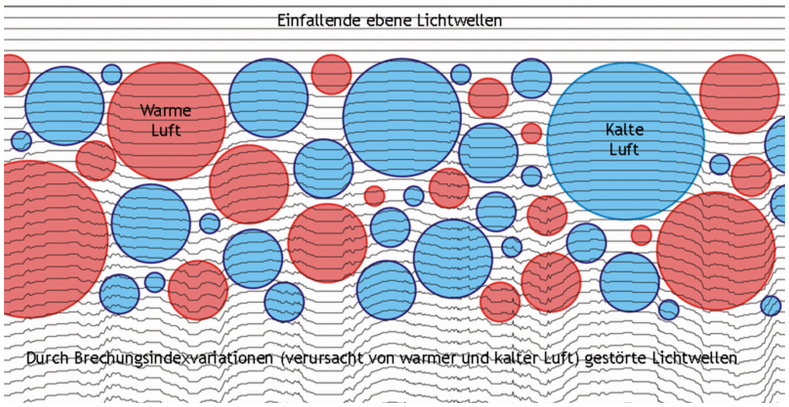


Abbildung 1.2: Modell der Turbulenzzellen und deren Beeinflussung der Wellenfront. Aus: [Wik11]

Frozen-Turbulence-Modell

Um den zeitlichen Verlauf der Turbulenz zu modellieren, geht man in der Regel von TAYLOR's *frozen-flow-Hypothese* aus [Tys00, S. 40]. Ohne die Details auszuführen, liegt dieser die Vorstellung zugrunde, dass die Brechungsindex-Fluktuationen über einen hinreichend kurzen Zeitraum konstant bleiben – bis auf den Einfluss des Windes. Die Turbulenz-Schicht mit festen Brechungsindices bewegt sich also mit dem Wind durch den Propagationspfad (genaugenommen nur mit dem Windanteil in der Ebene orthogonal zur optischen Achse). Der Vorteil dieser Annahme ist, dass man zeitliche Änderungen durch räumliche Verschiebungen ausdrücken kann.

Strehl-Verhältnis

Man verwendet ein Beugungsintegral unter anderem um die Intensität des einfallenden Lichts in der Beobachtungsebene bei der Beugung des Lichts durch eine Blende zu berechnen. Für eine kreisförmige Blende mit Durchmesser D, der Wellenlänge λ des einfallenden Lichts und dem Abstand R vom Beobachtungsort zur Blende nimmt man für $\frac{D^2}{\lambda R} \ll 1$ die sogenannte Fernfeldnäherung des Beugungsintegrals an (FRAUENHOFER-Beugung). Mit dieser gilt für die maximale Intensität I_0 auf der optischen Achse bei ebener Wellenfront mit gleichverteilter Intensität der Ausgangswelle

$$I_0 = P\frac{\pi D^2}{4\lambda^2 R^2}, \tag{1.9}$$

wobei P die Gesamtleistung durch die Apertur ist.

Der *RMS-Wellenfront-Fehler* („wavefront error“) σ ist als Wurzel aus dem Durchschnitt des Integrals der Quadrate aller Wellenfrontfehler über die gesamte Apertur definiert:

$$\sigma^2 = \frac{1}{\text{Oberfläche der Apertur}} \iint\limits_{\text{Apertur}} \big(\Phi(x,y) - \Phi_M(x,y)\big)^2 dxdy$$

mit der Wellenfront $\Phi(x,y)$ (vergleiche (1.2)) und ihrem Mittelwert $\Phi_M(x,y)$, der für eine ebene Welle konstant ist.
Für die Intensität auf der optischen Achse gilt

$$I = I_0\ S,$$

das *Strehl-Verhältnis* S gibt also die Verringerung der Intensität durch die optischen *Aberrationen* (Abbildungsfehler) an und wird bei kleinem Wellenfrontfehler ($\sigma < \frac{\lambda}{10}$) durch

$$S = \exp\left(-\left(\frac{2\pi\sigma}{\lambda}\right)^2\right) \tag{1.10}$$

genähert [Tys00, S. 23].

1.3 Adaptive Optik: Korrektur der Wellenfrontfehler

In diesem Abschnitt erfolgt eine kurze Einführung in die adaptive Optik und die beiden grundlegenden Methoden der Umsetzung adaptiver Optik werden erläutert und einander gegenübergestellt, zunächst die Definition adaptiver Optik:

Def. 1.1 (Adaptive Optik)**.**
Die Adaptive Optik *(AO) ist eine Technik zur Verbesserung der Abbildung optischer Systeme, welche Wellenfrontfehler, die bei der Propagation des Lichts entstehen, reduziert bzw. kompensiert.*[2] *Die Wellenfrontfehler können durch atmosphärische Turbulenz entstehen, es kann sich um statische oder nichtstatische Aberrationen des optischen Systems selbst handeln.*

Phasenkonjugation[3]

Man kann die Wellenfrontfehler korrigieren, in dem man zur richtigen Zeit entgegengesetzte Aberrationen auf die Welle aufprägt – dies nennt man *Phasenkonjugation.* Mathematisch ausgedrückt entsteht durch Multiplikation von $\mathrm{e}^{-i\varphi}$ (komplexe Konjugation der Phase $\varphi(\vec{x})$ im Argument der Exponentialfunktion) mit der elektromagnetischen Welle $E(\vec{x},t) = |E(\vec{x},t)|\ \mathrm{e}^{i\varphi(\vec{x})}\ \mathrm{e}^{i\omega t}$ (mit $\vec{x} = (x,y)$ in der Ebene orthogonal zur optischen Achse) die unverschobene Welle

$$E(\vec{x},t) = |E(\vec{x},t)|\ \mathrm{e}^{i\omega t}.$$

Um das physikalisch umzusetzen, kann man einen deformierbaren Spiegel nutzen. Da es aber auch andere Technologien gibt, wird hier der allgemeinere Begriff *Wellen(front-)korrekturelement* verwendet.

Hypothetisch könnte man die Korrektur der atmosphärischen Phasenstörungen vornehmen, indem man alle bzw. genügend viele Atmosphären-Parameter ermittelt und dann aus dem Modell die genauen Phasenstörungen erhält, die man zur Phasenkonjugation nutzt.

Anwendungsszenarien der adaptiven Optik

Neben der astronomischen Beobachtung ergeben sich auch weitere Anwendungsszenarien, z.B. die Übertragung von Daten mit Laserlicht im Freistrahl-Verfahren. Dafür ist man darauf angewiesen, am Ort des Empfangs eine gute Strahlqualität zu erhalten. Ein weiteres Anwendungsgebiet der adaptiven Optik findet sich bei der Energie-Übertragung mittels Laser-Strahlung oder dem aktiven *Imaging*, bei dem das Objekt beleuchtet wird (typischerweise von Lasern) und die Bildgebung auf den gestreuten Strahlen beruht. Anstatt Wellenfrontfehler des einfallenden Lichts zu kompensieren, kann man auch dem ausgehenden Laserstrahl beim Aussenden geeignete Phasenkorrekturen aufprägen, die zusammen mit den atmosphärischen Phasenfehlern einen „guten“ Strahl am Ziel liefern (Laserstrahl-Steuerung bzw. *preforming*). Neben der Korrektur der atmosphärischen Turbulenz kann man dabei erwarten, dass das adaptiv-optische System die statischen

[2] Für diese Arbeit wird auch die manchmal davon unterschiedene *aktive Optik* als adaptive Optik aufgefasst.

[3] In der nichtlinearen Optik wird dieser Begriff für eine bestimmten Medien inhärente Eigenschaft gebraucht, hier hingegen wirkt ein erst anzusteuerndes Element phasenkonjugierend.

Aberrationen der optischen Bauteile (Teleskopspiegel, Linsen, ...) mitkompensieren kann. Für Weltraum-Teleskope kann ein adaptiv-optisches System auch nur für diesen Zweck ausgelegt werden, da man dort keine kostengünstige Möglichkeit hat nachzujustieren, oder z.B. Leichtbau-Spiegel benutzt, die sich erst vor Ort entfalten und dabei systembedingt Fehler induzieren.

Insgesamt kann man die folgenden Einsatzszenarien nennen. Adaptive Optik wird bereits verwendet

- in der Astronomie,
- in Mikroskopie-Anwendungen und
- für Untersuchungen in der Augenheilkunde.

In der Forschung wird der Einsatz

- in der Freistrahl-Datenübertragung mit Lasern,
- in der Energieübertragung mit Lasern,
- in Laser-Verstärkern und
- in Weltraum-Teleskopen

vorbereitet bzw. erprobt.

Für einige Anwendungsszenarien benutzt man die *Laserleitstern*-Technik: Bei dieser Technik strahlt man in die Nähe des betrachteten Bereichs und korrigiert auf diesen künstlichen Laserleitstern hin. Wegen des Anisoplanatismus sollte sich dieser nah am zu beobachtenden Objekt befinden. Der hierfür verwendete Laser strahlt selbst auch durch die Atmosphäre, außerdem ist die Lasersicherheit zu berücksichtigen.

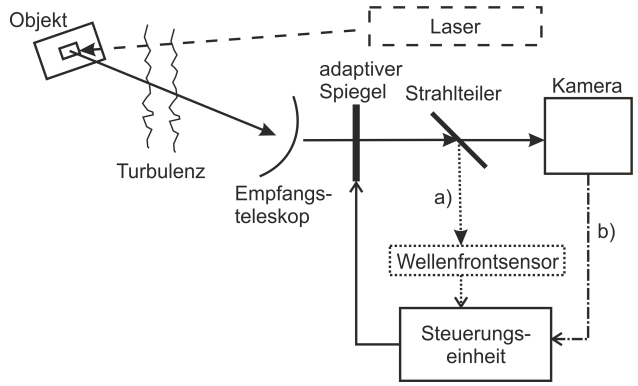


Abbildung 1.3: Aufbau und Funktionsweise eines adaptiv-optischen Systems. Zeichnung angelehnt an [PR07, Abb. 1] und [RW96, Abb. 5.1]. Man beachte die verschiedenen Pfeilformen für elektrische und optische Übertragung.

In Abbildung 1.3 wird grundlegend Aufbau und Funktionsweise eines adaptiv-optischen Systems vorgestellt. Dabei wird Licht von einem eventuell per Laser angestrahlten Objekt von einem Teleskop empfangen. Der Strahl trifft auf

einen adaptiven Spiegel und danach auf eine Kamera. Bei der Laserleitstern-Technik würde ein geeigneter Laser in den Himmel nahe des beobachteten Objekts strahlen, und die adaptive Optik würde auf diesen künstlichen Leitstern hin optimieren, während die Kamera den Bildausschnitt mit dem zu untersuchenden Objekt aufnimmt.

a) Für ein Wellenfrontsensor-basiertes adaptiv-optisches System wird ein Teil des eingehenden Strahles auf den Wellenfrontsensor gelenkt, aus dessen Daten die Steuerungseinheit eine Wellenfront rekonstruiert und dementsprechend neue Steuersignale an den Spiegel sendet. Die Kamera dient in diesem Fall nur der Aufnahme des Objekts, die Bilddaten fließen nicht in das Optimierungsverfahren des Regelprozesses ein.

b) Für ein Wellenfrontsensor-loses adaptiv-optisches System wird kein Wellenfrontsensor verwendet, und man berechnet direkt aus dem Kamerabild C einen Systemleistungswert y. Die Steuerungseinheit bestimmt daraufhin aus den bisherigen und dem aktuellen Systemleistungswert eine neue Spiegelstellung.

Der grundlegende, später noch angepasste Laboraufbau für ein adaptiv-optisches System, an dem im Rahmen dieser Arbeit Optimierungsläufe durchgeführt wurden, ist in den Abbildungen 6.4 und 6.5 dargestellt. Korrigiert man nicht den einfallenden Strahl, sondern sendet einen vorkorrigierten Strahl aus, so spricht man von Preforming:

Beispiel (Laserstrahl-Steuerung). Für die *Laserstrahl-Steuerung* (*preforming*), die in Abbildung 1.4 dargestellt ist, will man z.B. die Laserleistung auf einen kleinen Bereich fokussieren. Dies kann z.B. der Energieübertragung oder Datenkommunikation dienen. Man kann dann den Laser-Strahl *vorformen*, ihm also über einen adaptiven Spiegel eine Wellenfront aufprägen, die dann zusammen mit der Wellenfrontstörung durch Turbulenz am Zielpunkt eine gute Übertragungsqualität ergibt. Der entstehende Laser-Spot wird wieder beobachtet und abgebildet und auf ihn hin optimiert. Bei großen Distanzen kann sich das Problem langer Lichtlaufzeiten. ergeben, weil das Verfahren unter Annahme der „*frozen turbulence*“ arbeitet, siehe Seite 9.

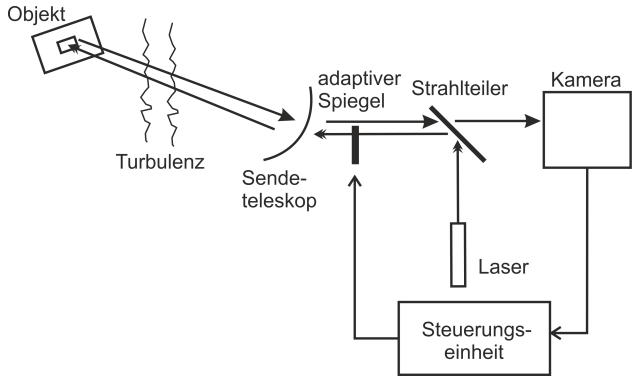


Abbildung 1.4: Adaptive Optik für die Laserstrahl-Steuerung

Wellenfrontsensor-gestützte adaptive Optik

Dieser klassische Ansatz wird auch als *Wellenfront-Konjugation* oder „wavefront reversal“ bezeichnet. Er besteht darin, die Wellenfront zu messen und daraus eine Korrektur-Stellung des deformierbaren Spiegels zu berechnen. Zur Messung der Wellenfrontstörung verwendet man einen *Wellenfrontsensor*, z.B. einen Shack-Hartmann-Sensor. Über die sogenannte *Phasen-Rekonstruktion* wird aus den Wellenfrontsensor-Daten errechnet, welche Wellenfront die einfallende Welle hatte. Bei dieser Herangehensweise sollte man die verwendeten Geometrien von Wellenfrontsensor und adaptivem Spiegel aufeinander abstimmen. Hierbei wird oft nur die Wellenfront korrigiert, die Modulation der Amplitude aber ignoriert. In der astronomischen Anwendung ist diese häufig relativ klein, bedingt durch die vergleichsweise guten Turbulenzbedingungen.

Für bodennahe Anwendungen mit schlechteren Turbulenzbedingungen wie Laser-*Imaging* oder Freistrahl-Datenübertragung kann die Wellenfront-Bestimmung durch einen Wellenfrontsensor problematisch werden, da dieser bei entsprechend großer Szintillation in den intensitätsschwachen Bereichen keine Wellenfrontfehler-Daten liefern kann und 2π-Phasensprünge nicht erkennen kann.

Bei dieser wie auch bei der folgenden Methode für die adaptive Optik hängt die erreichbare Verbesserung auch noch davon ab, wie gut man die gewünschte korrigierende Wellenfront mit dem Wellenfrontkorrekturelement einstellen kann.

Da man letztlich (zumindest in den meisten Anwendungsszenarien) direkt an der als Systemleistungswert dargestellten Leistung des Systems interessiert ist und nicht konkret an der Phasenkorrektur, gibt es einen zweiten Zweig der adaptiven Optik, bei dem man direkt einen Systemleistungswert optimiert („system perfomance metric“ in [Vor+00]).

Wellenfrontsensor-lose adaptive Optik

Sie wird unter anderem auch als *Wellenfrontsteuerung durch modellfreie Optimierung* und als nichtkonjugierte adapative Optik bezeichnet, siehe z.B. [Vor+00] Faktoren, die diesen Ansatz attraktiv erscheinen lassen, sind nach [Vor+00]:

- Die Probleme, die starke Szintillationen (Intensitätsvariation aufgrund der atmosphärischen Störung) der Wellenfrontsensor-basierten adaptiven Optik bereiten können, könnten vermieden werden.

- Klassische adaptive Optik benötigt eine ausreichend starke Referenzwelle. Für bestimmte Anwendungen wird dafür die Laserleitstern-Technik eingesetzt, die Probleme mit sich bringt.

- Effiziente modellfreie Optimierungsalgorithmen sind verfügbar geworden (Vorontsov [Vor+00] bezieht sich hier insbesondere auf SPGD[4] und seine VLSI[5]-Hardware-Implementierung).

- Neue Wellenkorrekturelemente sind verfügbar geworden, was die Erwartung weckt, dass das gesamte adaptiv-optische System günstig, schnell und einfach sein kann. Ein Beispiel dafür sind mikroelektromechanische deformierbare Spiegel (MEMS-Spiegel, MEMS steht dabei für *micro-electromechanical system*, Mikrosystem).

[4] Siehe zum Zusammenhang von SPSA und SPGD die Ausführungen auf Seite 24.
[5] Very-large-scale Integration (ein Integrationsgrad bei integrierten Schaltkreisen)

Wesentlich ist also, dass man die beschriebenen Beschränkungen des Wellenfrontsensors vermeiden will. Dies spielt bei den klassischen relativ turbulenzschwachen Anwendungsszenarien weniger eine Rolle als bei den neueren Szenarien z.B. für die Freistrahl-Datenübertragung.

Für die Wellenfrontsensor-lose adaptive Optik kann man nun eine geeignete Systemleistungsfunktion, ein geeignetes Optimierungsverfahren und passende Parameter für die algorithmische Umsetzung wählen. Dies steht im Gegensatz zur Wellenfrontsensor-gestützten adaptiven Optik, wo die Zielbedingung fest ist (nämlich der Ausgleich der Phasenstörungen). Die gewählte Systemleistungsfunktion wird dann maximiert.

Def. 1.2. Unter der *Qualitätsmetrik*, kurz *Metrik* eines adaptiv optischen Systems versteht man eine Funktion, die einer optischen Aufnahme (z.B. Kammerabild einer CCD-Kamera, aber auch anderes) einen Wert zuordnet, der eine Systemleistungsfunktion des adaptiv-optischen System beschreibt. Es handelt sich dabei nicht um eine Metrik im mathematischen Sinn des Wortes.

Je nach gewünschtem *Optimierungsziel* kann man verschiedene *Systemleistungsfunktionen* wählen. Eine solche Funktion dient dazu, die optische Qualität gut wiederzugeben.

Es gibt zum Beispiel die folgenden Optimierungszielsetzungen:

- Erreichen eines vorher definierten Strahlprofils,
- Höchstmögliche Bildschärfe eines bestimmten Bildbereichs (vergleiche Autofokussierung),
- Höchstmögliche Strehl-Zahl des optischen Systems, und
- Konzentration der Intensität in einem Bildbereich (*Power-in-the-Bucket*-Anwendung).

Für die *Power-in-the-Bucket*-Anwendung betrachtet man das Verhältnis der Intensität in einem Bildbereich zur Gesamtintensität, siehe Def. 1.3. Die Zielsetzungen überschneiden sich teilweise: Eine maximale Strehl-Zahl geht z.B. mit einem maximalen *Power-in-the-Bucket*-Wert einher. Zur Ermittlung des *Power-in-the-Bucket*-Werts kann ein kleinerer Bildbereich ausreichen als zur Bildschärfe-Bestimmung, die Strehl-Zahl hingegen gilt nur für einen Punkt auf der optischen Achse, so dass der *Power-in-the-Bucket*-Wert den Vorteil hat, aussagekräftiger zu sein.

1.4 Technische Aspekte

Technische Beschreibung der Funktion

Die Funktion f, welche die Steuersignale $\mathbf{x}$ für das Wellenfrontkorrektur-Element auf den Systemleistungswert y abbildet, kann man in zwei Abschnitte unterteilen,

$$\begin{aligned} f = f_1 \circ f_2 : \quad & \mathbb{R}^m \longrightarrow \mathbb{R} \\ & \mathbf{x} \mapsto y, \end{aligned} \tag{1.11}$$

mit

$$\begin{aligned} f_1 : \quad & \mathbb{R}^m \longrightarrow \mathbb{R}^{n_x \times n_y} \\ & \mathbf{x} \mapsto C \end{aligned} \tag{1.12}$$

und der Hintereinanderausführung $(f \circ g)(x) := f(g(x))$. f_1 bildet dabei die Steuersignale $\mathbf{x}$ auf das Kamerabild C der Größe $n_x \times n_y$ ab. Dieser Schritt beinhaltet die physikalischen Begebenheiten und trägt zum wesentlichen Teil des *Rauschens* bei. Dies speist sich aus verschiedenen Quellen, unter anderem aus

- der Strahllage-Qualität eines eventuell eingesetzten Lasers,
- dem Rauschen der Kamera und
- den Einschränkungen des adaptiven Spiegels (wie Nachschwingen).

Die Anzahl der Steuerkanäle des Spiegels m hat dabei die Größenordnung $25 \ldots 4000$. Das Kamerabild mit beispielsweise 1200×1000 Pixeln oder 256×256 Pixeln (bei 16 bit-Farbtiefe also $2,4$ MB bzw. $0,125$ MB pro Bild) wird danach in einen Computer oder in ein speziell dafür entwickeltes Gerät übertragen und ein Systemleistungswert (Metrikwert) ermittelt:

$$\begin{aligned} f_2 : \quad & \mathbb{R}^{n_x \times n_y} \longrightarrow \mathbb{R} \\ & C \mapsto y. \end{aligned} \tag{1.13}$$

In der Praxis kann es auch sein, dass man nicht einen, sondern mehrere verschiedene Metrikwerte ermittelt und bei einer Annäherung an das Optimum in eine andere Metrik wechselt. Die Behandlung der bei der Verwendung mehrerer Metrikwerte aufkommenden Fragen der Vektoroptimierung wird in dieser Diplomarbeit nicht geleistet und ist für die Praxis der adaptiven Optik von nachgeordnetem Interesse.

Der adaptive Spiegel

Als adaptiver Spiegel wird ein deformierbarer Spiegel bezeichnet, der zur Korrektur von Phasenstörungen verwendet werden kann.

Bei deformierbaren Membranspiegeln hängt es von den Randbedingungen ab, welche Wellenfronten gut reproduzierbar sind. Nach dem Anlegen eines Steuersignals gibt es eine Nachschwingzeit, die man abwarten muss. Außerdem lassen sich die einzelnen Spiegelsegmente nicht komplett unabhängig voneinander verstellen, sondern bewegen benachbarte Segmente mit (Übersprechen, *Cross-coupling*). Die deformierbaren Spiegel werden in dieser Arbeit als idealisiert angenommen. Somit kann man davon ausgehen, dass die Steuersignale die Spiegelstellungen so beeinflussen, als wäre jede gewünschte Wellenfront darstellbar.

Ansonsten wären diese Fragen zu stark vom konkret gewählten Spiegelmodell abhängig. Durch die modellfreie Optimierung ist das „aus der Sicht des Verfahrens“ Teil der Funktion f, von der man den Einfluss der Steuersignaländerungen auf den Systemleistungswert ohnehin nicht kennt.

Power-in-the-Bucket-Wert

Den *Power-in-the-Bucket*-Wert definiert man für ein konkretes System wie folgt:

Def. 1.3 (*Power-in-the-Bucket*-Wert)**.** Sei $\bar{I}_r$ der Mittelwert der Helligkeitswerte der im Kreis mit Radius r liegenden Kamerapixel, und sei A_r der Flächeninhalt des Kreises. Dann ergibt sich der *Power-in-the-Bucket*-Wert zu:

$$P_r(C) = \frac{\bar{I}_r \cdot A_r}{\bar{I}_{\mathrm{ROI}} \cdot A_{\mathrm{ROI}}}, \tag{1.14}$$

wobei der Index ROI die entsprechende Größe für das gesamte betrachtete Kamerabild bezeichnet.

Modale Ansteuerung

Man kann den adaptiven Spiegel entweder zonal ansteuern, d.h., die Spannung für jeden Aktuator einzeln regeln, was vom adaptiven Spiegel dann in die entsprechende Aktuator-Auslenkung umgesetzt wird. Im Gegensatz dazu gibt es auch die modale Ansteuerung über sogenannte Zernike-Moden [Nol76], das sind orthogonale Polynome $Z_n^l(r, \varphi)$, die verwendet werden, um Wellenfrontfehler und Abbildungsfehler zu beschreiben. Bei der modalen Ansteuerungsart fasst man die Koeffizienten von m Zernike-Moden als Steuersignalvektor $\mathbf{x}$ zusammen, der die Gewichte in der Linearkombination aus Polynomen beschreibt.

Zeitliche Beschränkungen

Mithilfe des bisher Dargestellten kann man eine Grenze für die maximale Anzahl von Funktionsauswertungen herleiten, die bei der Wellenfrontsensor-losen adaptiven Optik zur Verfügung stehen.

Die Greenwood-Frequenz f_G (siehe Gleichung (1.8)) beträgt z.B. auf einem der am besten geeigneten Orte für astronomische Beobachtungen, dem Haleakalā auf Hawaii, 20 Hz ($\tau_0 = 50\,\mathrm{ms}$) und im Allgemeinen kann man von etwa 30 Hz ($\tau_0 = 33\,\mathrm{ms}$) ausgehen [Tys00, S.6], also von der Größenordnung her einige 10Hz ($\tau_0 \approx 100\,\mathrm{ms} \ldots 30\,\mathrm{ms}$).

Die Anzahl der Funktionsauswertungen, die man zur Verfügung hat, bevor sich die Beziehung Steuersignale – Bildgüte durch die Variation der Turbulenz geändert hat, ergibt sich mit

$$\#\text{Funktionsauswertungen} = \frac{\tau_0}{\tau_{\mathrm{DM}} + \tau_{\mathrm{Cam}} + \tau_{\mathrm{Contr}}}$$

mit der atmosphärischen Zeitkonstante τ_0 (s. S. 8), der Zeit τ_{DM}, die der adaptive Spiegel benötigt, um die neuen Steuersignale in Spiegelstellungen umzusetzen, der Zeit τ_{Cam}, die die Kamera zur Aufnahme braucht und der Zeit τ_{Contr} , die die Steuerungseinheit benötigt, um einen Systemleistungswert aus dem Kamerabild zu bestimmen und aus dem bisherigen Verlauf ein neues Steuersignal für den Spiegel zu erzeugen.

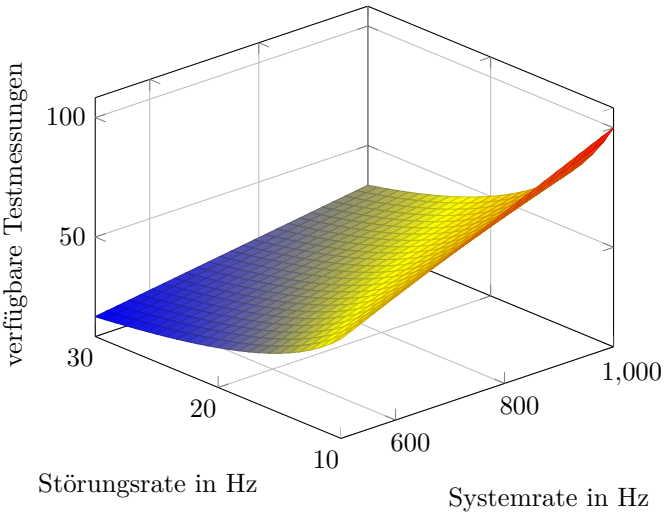


Abbildung 1.5: Die Anzahl der verfügbaren Funktionsauswertungen für verschiedene Geschwindigkeiten des adaptiv-optischen Systems und turbulente Störungen von $10 \ldots 30\,\mathrm{Hz}$ liegt zwischen 20 und 100.

In Tabelle 1.1 wird die Systemrate durch eine Betrachtung der verwendeten Komponenten abgeschätzt. Die Dauer der Spiegeleinstellzeit hängt typischerweise vom Spiegelhub ab. Der Erhalt des Kamerabilds wird derzeit durch die Datenrate der Übertragung begrenzt. Mit dem aktuellen Test-Laboraufbau sind nur Systemraten von weniger als 20 Hz realisierbar. Durch den Einsatz besserer Hardware kann die Systemrate aber in den unteren kHz-Bereich gebracht werden, wenn man davon ausgeht, dass man die Berechnungen (neuer Steuersignale und eines Systemleistungswerts) ausreichend beschleunigen kann, zum Beispiel durch Verkleinerung der *Region of Interest* der Kamera, durch Parallelisierung der Metrik-Berechnung oder durch Hardware-Implementierung.

In Abbildung 1.5 wird die Anzahl der verfügbaren Funktionsauswertungen dargestellt bei Störungsraten von $10 \ldots 30\,\mathrm{Hz}$ und Systemraten von $500 \ldots 1000\,\mathrm{Hz}$. Es zeigt sich, dass $20 \ldots 100$ Funktionsauswertungen zur Verfügung stehen. Für die Untersuchung in der praktischen Anwendung wird man letztlich eher auf die Anzahl der Funktionsauswertungen schauen, die für eine signifikante Verbesserung nötig sind und daraufhin die einzelnen Komponenten (Spiegel, Kamera, Steuerungseinheit) auf ihre Geschwindigkeit hin aussuchen.

Im Anschluss an die folgende mathematische Behandlung wird das Verfahrens in einem adaptiv-optischen Laborsystem in Kapitel 5 und 6 im Hinblick auf die Anwendung untersucht.

Adaptive Optik			
		Frequenz	Dauer
Spiegel-Neueinstellung			
MEMS-Spiegel – Beispiel Boston Micromachines 4K-DM		20 KHz	0.05 ms
Adaptiver Spiegel im Labor-Testaufbau	Nachschwingzeit	20 Hz	50.0 ms
Aufnahme Kamerabild			
Es gibt Kameras, die mehrere Millionen Bilder pro Sekunde aufnehmen können.		$\sim$ MHz	0.001 ms
Übertragung Kamerabild			
Framegrabber mit 600 MB/s	2.4 MB (1600×1200)	250 Hz	4.0 ms
	0.125 MB (256×256)	$\sim$ kHz	0.2 ms
Berechnung Metrikwert			
hängt von verwendeter Bildgüte-Funktion und vom Computer ab			x_1 ms
Berechnung neuer Steuersignale			
hängt vom verwendeten Algorithmus und vom Computer ab			x_2 ms
Zusammen			
im besten Fall:		$\leq$ einige KHz	$0.26 + x$ ms
im langsamsten Fall:		$\leq$ 20 Hz	$54 + x$ ms
Störungs-Rate			
Greenwood-Frequenz bzw. atmosphärische Zeitkonstante		10 . . . 30 Hz	100 . . . 33 ms

Tabelle 1.1: Übersicht der Zeitkonstanten von Komponenten einer Wellenfrontsensor-losen adaptiven Optik.
Über den adaptiven Spiegel gibt auch Tabelle 6.1 Auskunft. Ein *Framegrabber* dient zur Übertragung des Kamerabilds in den Computer.

Kapitel 2

Mathematisches Modell

Dieses Kapitel schließt die Lücke zwischen dem physikalischen Einleitungskapitel und der folgenden mathematischen Behandlung. Insbesondere werden grundlegende Notationen eingeführt, um über stochastische Gradientenverfahren sprechen zu können, und das folgende wahrscheinlichkeitstheoretische Kapitel wird motiviert.

2.1 Formulierung der Optimierungsaufgabe

Das physikalische Problem als Optimierungsaufgabe

Aus dem in der Einleitung formulierten physikalischen Hintergrund ergibt sich die folgende Optimierungsaufgabe. In der Formulierung der Aufgabe wird $\mathbf{x}$ ein m-dimensionaler Vektor sein, der die Steuersignale für das Wellenfrontkorrekturelement modelliert, und $\mathbb{H} = \{\mathbf{x} : a_i \leq \mathbf{x}_i \leq b_i; a_i \leq b_i\}$ ein Hyperquader, der also komponentenweise Beschränkungen enthalten kann. Die Zielfunktion f in der Optimierungsaufgabe

$$f(\mathbf{x}) \rightarrow \max_{\mathbf{x} \in \mathbb{H} \subset \mathbb{R}^m}!$$

setzt sich aus verschiedenen Komponenten zusammen und umfasst den folgenden Prozess, der auch in Abbdilung 2.1 schematisch dargestellt ist. Zuerst gibt man dem Wellenkorrekturelement (etwa einem deformierbaren Spiegel) neue Steuersignale. Der Lichtstrahl bzw. Laserstrahl propagiert dann mit den aufgeprägten Korrekturen durch optische Elemente mit Aberrationen (dies können ständige Aberrationen sein oder sie können zum Beispiel aufgrund thermischer Verformung o.ä. entstehen) und evtl. durch turbulente Atmosphäre. Schließlich trifft der Strahl auf eine Kamera, und aus dem Kamerabild bestimmt man einen Metrikwert. In Abhängigkeit von den Steuersignalen soll also die Metrik maximiert werden.

Für die folgende mathematische Behandlung wird äquivalent von einem Minimierungsproblem

$$f(\mathbf{x}) \rightarrow \min_{\mathbf{x} \in \mathbb{X} \subset \mathbb{R}^m}! \tag{2.1}$$

ausgegangen mit einem zulässigen Bereich $\mathbb{X}$ (siehe Lemma 4.1). Mit $\mathbf{x}^*$ werden Minima von (2.1) bezeichnet und konkrete Funktionswerte von f mit y ($y = f(\mathbf{x})$).

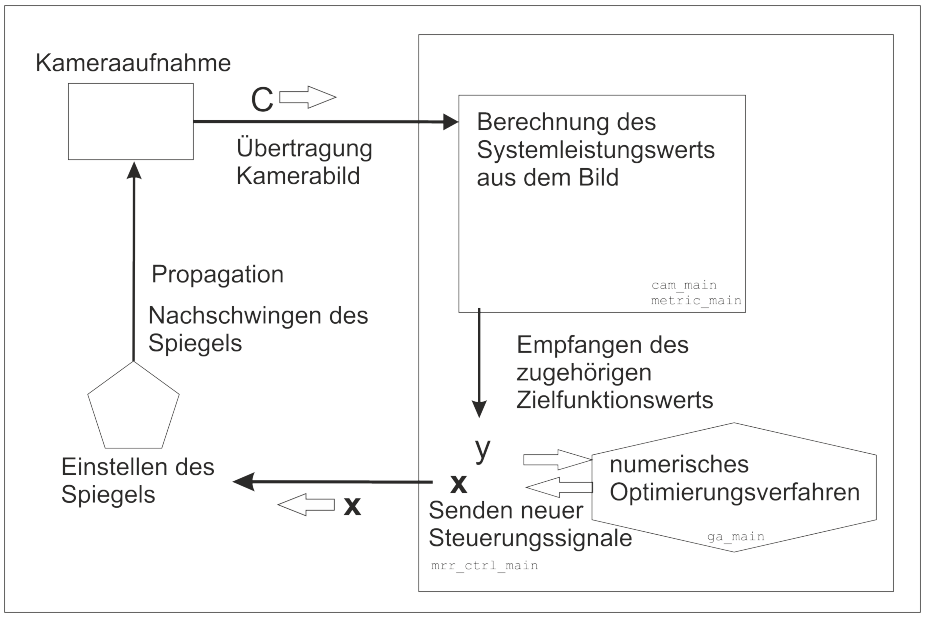


Abbildung 2.1: Schema des Ablaufs der Funktionsauswertung am Beispiel des Laborsystems

Bevor die verwendeten Größen im Falle von mit Störungen verbundenen Funktionsauswertungen, also Messfehlern, eingeführt werden können, sei kurz die verwendete wahrscheinlichkeitstheoretische Notation geklärt. Diese wird im folgenden Kapitel vertieft.
$(\Omega, \mathcal{F}, \mathbb{P})$ sei in den folgenden Kapiteln ein Wahrscheinlichkeitsraum, E ein messbarer Raum, der mit der σ-Algebra $\mathcal{E}$ ausgestattet sei. Ist $E = \mathbb{R}$ oder $E = \mathbb{R}^m$, so sei dieser mit der BORELschen σ-Algebra $\mathcal{B}$ bzw. $\mathcal{B}^n$ und dem Lebesgue-Maß λ ausgestattet. Eine Funktion $X : \Omega \to E$ ist *integrierbar* oder ihr Erwartungswert existiert, wenn $\mathbb{E}(|X|) < \infty$. $\mathbb{V}\mathrm{ar}(X)$ sei die Varianz der Zufallsgröße X, sie ist für quadratintegrierbare Zufallsgrößen endlich.

Problemformulierung mit Rauschen

Im Folgenden wird der Fall betrachtet, in dem statt der Zielfunktion f selbst nur durch Rauschen gestörte Messungen $\tilde{f}(\mathbf{x})$ vorliegen. Das Minimierungsproblem erhält dann die Form

$$\mathbb{E}\big(\tilde{f}(\mathbf{x})\big) \to \min_{\mathbf{x} \in \mathbb{X} \subset \mathbb{R}^m}! \tag{2.2}$$

Die folgende Definition für unabhängiges Rauschen wird hier zu Vergleichszwecken angegeben, die tatsächlichen Bedingungen an Erwartungswert, Varianz und Unabhängigkeit finden sich im Rahmen der Konvergenzanalyse.

Def. 2.1 (Unabhängiges Rauschen)**.**
Eine Familie von Zufallsvariablen $X_s : \Omega \to \mathbb{R}^m$, $s \in \mathbb{R}$ heißt unabhängiges Rauschen*, wenn X_s jeweils ($\mathcal{F}$-$\mathcal{B}$-)messbar sind, $\mathbb{E}(X_s) = 0$, $\mathbb{V}\mathrm{ar}(X_s) = const.$ gilt und alle X_s unabhängig sind.*
Rauschen ist also eine Abbildung $X : \Omega \times \mathbb{R} \to \mathbb{R}^m, \quad (\omega, s) \mapsto X_s(\omega)$.

In dieser Arbeit wird für Rauschen die folgende Notation verwendet:

$$\tilde{f}(\mathbf{x}) = f(\mathbf{x}) + R_{\mathbf{x}}, \tag{2.3}$$

$R_{\mathbf{x}} : \Omega \to \mathbb{R}$ ist eine Zufallsgröße. In diesem Sinne ist auch $\tilde{f}(\mathbf{x})$ eine Zufallsgröße.

2.2 Anforderung an numerische Optimierungsverfahren in der adaptiven Optik

Wegen der Anwendung in der adaptiven Optik ergeben sich folgende Besonderheiten:

- **Echtzeitbedingungen** durch die zeitliche Dynamik der Zielfunktion: Die Lösung der Optimierungsaufgabe (aktuell beste Steuersignale $\mathbf{x}^*$) ist nur sehr kurz nützlich, da sich die verschiedenen Störungsquellen rasch ändern, insbesondere wenn man atmosphärische Turbulenz korrigieren will. Das bedeutet, dass man mit möglichst wenig Funktionsauswertungen bereits Verbesserungen erzielen möchte.
- **Physikalische Funktionsauswertung:** Die Abbildung Steuersignal $\mathbf{x}$ auf Metrikwert y, die die Zielfunktion f darstellt, wird nicht analytisch oder auf bestimmte Art numerisch berechnet, sondern ergibt sich zwischen Steuersignalen und Kamerabild als physikalische Messung.
- **Hohe Dimension des Grundraums**
- **Live-Bedingung:** Es gibt keine echten Test-Zielfunktionsauswertungen– jedes Steuersignal wirkt sich auf das aktuelle Bild aus.

Daraus resultieren die Anforderungen an das numerische Optimierungsverfahren:

- **Verwendung nur weniger Funktionsauswertungen pro Iteration** trotz hoher Dimension, und schnelle Erzielung erster Verbesserungen.
- **Robustheit gegen Rauschen:** Sowohl für die beschriebenen, dem System inhärenten Quellen des Rauschens, als auch um die Robustheit gegen die Veränderung der Zielfunktion mit der Zeit zu erhalten (dies simuliert man als Rauschen und nimmt relativ „gutartige“, stetige Verändung an), muss das Verfahren robust gegenüber Rauschen in der Zielfunktion sein.
- **Lokale Optimierung:** Durch die Echtzeitbedingungen und die zeitliche Änderung der Zielfunktion ist es entscheidender, schnell eine Verbesserung der Bildgüte zu erzielen, als unbedingt das globale Optimum zu finden. Bei geringen atmosphärischen Störungen oder unter Laborbedingungen kann dies aber auch durchaus wünschenswert sein.

Für die numerische Analyse geht man von einem konstanten Modell aus, d.h., die Funktion $\mathbf{x} \mapsto f(\mathbf{x})$, die Steuersignale auf Systemleistungswert abbildet, sei nicht von der Zeit abhängig. In der Anwendung „adaptive Optik zur Korrektur atmosphärischer Turbulenz“ entspricht dies der Betrachtung eines kurzen Zeitraums im „*frozen-turbulence*-Fenster“, in dem man die Turbulenz als konstant

annehmen kann. Weiterhin geht man davon aus, dass man das Verhalten des Verfahrens an den Übergangsstellen zwischen den „*frozen-turbulence*-Fenstern“ durch seine Robustheit gegenüber Rauschen einschätzen kann.

2.3 Charakterisierung des Lösungsansatzes

Man behandelt das Optimierungsproblem als *modellfreie Optimierung*, versucht also nicht ein Modell für den Zusammenhang von Steuersignal und Bildgüte zu konstruieren. Die zugehörige Anwendung in der adaptiven Optik wird daher auch als *auf modellfreier Optimierung basierende Wellenfront-Kontrolle* (oder früher als *aperture tagging* bzw. *image-sharpening techniques* bezeichnet, siehe [Vor+00]).

2.4 Charakterisierung der Verfahren

Nach der Beschreibung des Optimierungsproblems und der Wahl des modellfreien Ansatzes wird nun eine erste Definition der Verfahren gegeben.

2.4.1 Definition der Verfahren

Zur Lösung des Optimierungsproblems werden *Abstiegsverfahren* untersucht.

Def. 2.2. *Abstiegsverfahren* sind Verfahren mit der Iterationsvorschrift

$$\mathbf{x}_{k+1} = \mathbf{x}_k + a_k \mathbf{p}_k(\mathbf{x}_k),\ k = 0, 1, \ldots, \tag{2.4}$$

wobei $\mathbf{p}_k(\mathbf{x}_k)$ eine Abstiegsrichtung in $\mathbf{x}_k$ ist und $a_k > 0$ eine Schrittweitenfolge. Die Definition der Abstiegsrichtung und eine Erklärung der Wortwahl erfolgt in Abschnitt 4.2.1. Speziell für den Fall $\mathbf{p}_k(\mathbf{x}_k) = -\nabla f(\mathbf{x}_k)$ wird es als *Verfahren des steilsten Abstiegs* bezeichnet und erhält die Form

$$\mathbf{x}_{k+1} = \mathbf{x}_k - a_k \nabla f(\mathbf{x}_k). \tag{2.5}$$

Abgekürzt soll es in dieser Arbeit als SD-Verfahren bezeichnet werden, nach *method of steepest descent*. Die deutsche Abkürzung SA-Verfahren nach „Steilster Abstieg“ wäre in dieser Arbeit irreführend, da es sich nicht um eine Variante der später noch einzuführenden *Stochastic-Approximation*-Verfahren handelt, die mit SA abgekürzt werden.

Da im physikalischen System keine direkten Informationen über den Gradienten ∇f zugänglich sind, muss man sich mit einer Gradientennäherung $\hat{\mathbf{g}}_k$ an ∇f behelfen, die man nur aus Funktionsauswertungen der Funktion f bzw. im Falle von Messfehlern aus $\tilde{f}$ errechnet.

Def. 2.3 (Ableitungsfrei)**.** Verfahren heißen *ableitungsfrei*, wenn sie keine Auswertungen des Gradienten ∇f benutzen, sondern nur Funktionsauswertungen der Zielfunktion f selbst verwenden.

Diese Begriffswahl kommt aus [Sch79, Kapitel 9] und findet sich analog im Englischen als *gradient-free* in [Spa05, Kapitel 6].

Eine Möglichkeit ist es, eine Gradientennäherung durch komponentenweise einseitige Differenzenquotienten zu erhalten:

$$\hat{\mathbf{g}}_h(\mathbf{x}) = \left(\frac{f(\mathbf{x} + h\mathbf{e}_i) - f(\mathbf{x})}{h} \right)_{i=1,\ldots,m}, \tag{2.6}$$

mit dem i-ten Einheitsvektor $\mathbf{e}_i$ und einer Konstante $h > 0$. Mit $\mathbf{p}_k(\mathbf{x}_k) = -\hat{\mathbf{g}}_k(\mathbf{x}_k)$ erhält man das ableitungsfreie Gradientenverfahren, kurz GSD-Verfahren. Auf dieses wird im Abschnitt 4.2.4 genauer eingegangen.

Im Falle einer verrauschten Zielfunktion wird die Folge der Iterierten $\mathbf{x}_k$ zu einer Folge von Zufallsvektoren $\mathbf{X}_k$ und man befindet sich auf dem Terrain der Theorie der *Stochastic-Approximation*-Verfahren.

Falls der Gradient direkt verfügbar, aber mit Rauschen behaftet ist, erhält das Verfahren des steilsten Abstiegs (2.5) die Form

$$\mathbf{X}_{k+1} = \mathbf{X}_k - a_k \tilde{\mathbf{g}}(\mathbf{X}_k) \,, \tag{2.7}$$

mit $\tilde{\mathbf{g}}(\mathbf{x}) = \nabla f(\mathbf{x}) + R'_{\mathbf{x}}$. $R'_{\mathbf{x}}$ sind wieder Zufallsgrößen, die die Messfehler bei der direkten Gradientenmessung modellieren. Dies soll zunächst als stochastisches Verfahren des steilsten Abstiegs (SSD nach *stochastic steepest descent*) bezeichnet werden.

Im ableitungsfreien Fall verwendet man einen Gradientenschätzer. In der mathematischen Statistik wird als eine der wesentlichen Eigenschaften eines *Schätzers* die Erwartungstreue definiert.

Def. 2.4 (Erwartungstreue und Bias)**.** Sei $P : \Omega \to E$ eine messbare Funktion. P ist *erwartungstreuer Schätzer* eines Parameters $p \in E$, falls $\mathbb{E}(P) = p$. Die *Verzerrung* (*Bias*) definiert man entsprechend:

$$\mathrm{Bias}(P) \coloneqq \mathbb{E}(P - p). \tag{2.8}$$

Ist $\mathrm{Bias}(P) \neq 0$, so spricht man von einem *verzerrten* Schätzer.

Das damit zum ableitungsfreien Verfahren des steilsten Abstiegs korrespondierende *Stochastic-Approximation*-Verfahren bezeichnet man als SA-Verfahren der Kiefer-Wolfowitz-Form (KW-SA, [Chi97]), seine allgemeine Form ist

$$\mathbf{X}_{k+1} = \mathbf{X}_k - a_k \hat{\mathbf{g}}_k(\mathbf{X}_k) \tag{2.9}$$

mit einer Gradientenschätzung $\hat{\mathbf{g}}_k(\mathbf{X}_k)$.

Diese Bezeichnungen gründen sich auf dem Beginn der *Stochastic-Approximation*-Theorie in den Arbeiten von Robbins und Monro: „A Stochastic Approximation Method“, [RM51] aus dem Jahr 1951 und von Kiefer und Wolfowitz: „Stochastic Estimation of the Maximum of a Regression Function“, [KW52] aus dem Jahr 1952. Der mehrdimensionale Fall der KW-SA-Methode geht laut [Spa05, S. 151] auf [Blu54] zurück.

Verwendet man finite Differenzen zur Bestimmung eines Gradientenschätzers, also Differenzenquotienten-Näherungen in jeder Komponente, so wird das entstehende SA-Verfahren mit *Finite Differences Stochastic Approximation* bezeichnet, kurz FDSA [Spa05, S.151]. Bei diesem Verfahren hat die Gradientenschätzung die Form

$$\hat{\mathbf{g}}_k(\mathbf{X}_k) = \hat{\mathbf{g}}_k^{\mathrm{FD1}}(\mathbf{X}_k) \coloneqq \left(\frac{\tilde{f}(\mathbf{X}_k + h_k\mathbf{e}_i) - \tilde{f}(\mathbf{X}_k)}{h_k} \right)_{i=1,\ldots,m} \tag{2.10}$$

mit $h_k > 0$. FD1 steht dabei für die Approximation 1. Ordnung bei dem verwendeten einseitigen Differenzenquotienten, vergleiche Lemma 4.11. Um zwischen den beiden Schrittweitenfolgen a_k und h_k zu unterscheiden, nennt man a_k die *Iterierten-Schrittweitenfolge* und h_k die *Gradientenschätzungs-Schrittweitenfolge*.

Die Notation (2.3) erweiternd, wird für die Funktionsauswertungen, die der Gradientenschätzung dienen, eine zusätzliche Notation für die Messfehler eingeführt:

$$R_k^{i+} = R_{\mathbf{X}_k + h_k \mathbf{e}_i} \text{ und}$$
$$R_k^0 = R_{\mathbf{X}_k},$$

so dass für die verrauschte Funktion $\tilde{f}$ gilt:

$$\tilde{f}(\mathbf{X}_k + h_k \mathbf{e}_i) =: f(\mathbf{X}_k + h_k \mathbf{e}_i) + R_k^{i+} \text{ und}$$
$$\tilde{f}(\mathbf{X}_k) =: f(\mathbf{X}_k) + R_k^0.$$

Dabei werden R_k^{i+} und R_k^0 im Weiteren als eigenständige Zufallsgrößen ohne Rückgriff auf $R_{\mathbf{x}}$ behandelt. Die Bedingungen an die das Rauschen modellierenden Zufallsgrößen finden sich im Abschnitt 4.3.6.

Das *Simultaneous Perturbation Stochastic Approximation* Verfahren (Stochastisches Approximationsverfahren der gleichzeitigen Störungen[1]), wird in Spall: *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*, [Spa05] aus der Klasse der *Stochastic-Approximation*-Verfahren hergeleitet. In der adaptiven Optik wurde eine Variante davon u.a. von Mikhail Vorontsov, Direktor des *Intelligent Optics Laboratory* sowie Professor und Stiftungsprofessor für das *Ladar and Free Space Optical Communications Institute (LOCI)* an der Universität Dayton [Vor] unter dem Namen *Stochastic Parallel Gradient Descent* (stochastisches paralleles Gradientenabstiegsverfahren) erfolgreich eingesetzt [Vor+00; VS98]. „Parallel“ im Namen bezieht sich dabei darauf, dass für die Test-Funktionsauswertungen zur Gradientenschätzung alle Komponenten von $\mathbf{x}$ gleichzeitig verändert werden. Es wird als SPGD-Verfahren abgekürzt und kann als SPSA-Verfahren aufgefasst werden, siehe Bemerkung 4.56.

Beim *Simultaneous Perturbation Stochastic Approximation* Verfahren verwendet man die Gradientenschätzung

$$\hat{\mathbf{g}}_k(\mathbf{X}_k) = \hat{\mathbf{g}}_k^{\mathrm{SP}}(\mathbf{X}_k) = \left(\frac{\tilde{f}(\mathbf{X}_k + h_k \mathbf{D}_k) - \tilde{f}(\mathbf{X}_k - h_k \mathbf{D}_k)}{2 h_k \mathbf{D}_{ki}} \right)_{i=1,\ldots,m}, \tag{2.11}$$

wobei $h_k > 0$, $\mathbf{D}_{ki} \neq 0$ fast sicher. Die genauen Bedingungen an $\mathbf{D}_k$ werden im Abschnitt 4.3.7 angegeben und erläutert. Die von Spall vorgeschlagene Wahl ist hierbei $\mathbf{D}_{ki}$ unabhängig jeweils $+1$--1-Bernoulli-verteilt zu wählen. In jeder Komponente von $\hat{\mathbf{g}}_k$ wird im Unterschied zu FDSA die gleiche Differenz von Funktionswerten verwendet. Wenn das SPSA-Verfahren sich gut verhält, liegt hier eine wesentliche Chance, die Gradientenschätzung durch nur 2 bzw. eine konstante Anzahl von Funktionsauswertungen zu erhalten. Im Unterschied

[1] Diese werden im Folgenden als *Perturbationen* bezeichnet, um begrifflich klar von Messfehlern/Rauschen getrennt zu sein.

dazu ist bei deterministischen Verfahren die benötigte Anzahl von Funktionsauswertungen in der Größenordnung der Dimension des Suchraums. Man muss natürlich erwarten, dass weniger verfügbare Information auch zu einer schlechteren Schätzung für den Gradienten führen, allein deswegen muss das Verhalten des Verfahrens aber nicht schlechter werden, siehe Abschnitt 4.3.7.

Auch hierfür wird eine angepasste Notation für die Messfehler, die bei den Funktionsauswertungen auftreten, die der Gradientenschätzung dienen, eingeführt:

$$R_k^+ = R_{\mathbf{X}_k + h_k \mathbf{D}_k} \text{ und}$$
$$R_k^- = R_{\mathbf{X}_k - h_k \mathbf{D}_k},$$

so dass für die verrauschte Funktion gilt:

$$\tilde{f}(\mathbf{X}_k + h_k \mathbf{D}_k) =: f(\mathbf{X}_k + h_k \mathbf{D}_k) + R_k^+ \text{ und}$$
$$\tilde{f}(\mathbf{X}_k - h_k \mathbf{D}_k) =: f(\mathbf{X}_k - h_k \mathbf{D}_k) + R_k^-.$$

Dabei werden $R_k^\pm$ im Weiteren als eigenständige Zufallsgrößen ohne Rückgriff auf $R_{\mathbf{x}}$ behandelt. Die Bedingungen an die das Rauschen modellierenden Zufallsgrößen finden sich im Abschnitt 4.3.7.

Da es für die Behandlung der Gradientenschätzung reicht, die bei der Schätzung auftretenden Messfehlerdifferenzen zu betrachten, definiert man spezieller:

$$R_k^{\mathrm{FD1}} := R_k^{i+} - R_k^0, \tag{2.12}$$
$$R_k^{\mathrm{FD2}} := R_k^{i+} - R_k^{i-} \text{ und} \tag{2.13}$$
$$R_k^{\mathrm{SP}} := R_k^+ - R_k^-. \tag{2.14}$$

Es schließt nun eine erste Gegenüberstellung der Verfahren an (vergleiche auch Tabelle 2.1). In Abschnitt 4.3.3 wird dies mit dem entsprechenden wahrscheinlichkeitstheoretischen Hintergrundwissen vertieft. Dort wird dann unter anderem auch die Anzahl der benötigten Zielfunktionsauswertungen pro Iteration verglichen.

Tabelle 2.1: Gegenüberstellung von SD-, FDSA- und SPSA-Verfahren

SD	FDSA	SPSA
sind Gradienten- und lokale Optimierungsverfahren		
deterministisch	stochastisch	
direkte Gradientenmessung	ableitungsfrei	
keine Messfehler	verrauschte Zielfunktion	
kein Zufall	1 Zufallsquelle (Rauschen)	2 Zufallsquellen (Rauschen und Perturbation)

2.4.2 Parameter und Varianten

Für das **GSD-Verfahren** kann man die *Art der Gradientennäherung* wählen. Grundlegend geht es dabei um die Wahl der Approximation 1. oder 2. Ordnung, also GN1 oder GN2, gleichwohl wären auch Mischformen denkbar.

Außerdem muss man die *Art der Liniensuche* wählen (siehe Abschnitt 4.2.3). Einstellbare Parameter sind die Startschrittweite a, der Parameter h in der Gradientennäherung und eventuelle Liniensuchparameter.

Beim **FDSA-Verfahren**, das im Prinzip gleichgelagert ist, ist es üblich, keine Liniensuche durchzuführen und die Abstände der Gradientenschätzung h und die Iterations-Schrittweite a als Schrittweitenfolgen $a_k = \frac{a}{k^\alpha}$, $h_k = \frac{h}{k^\gamma}$ zu wählen, so dass die Parameter a, h, α, γ festzulegen sind.

Beim **SPSA-Verfahren** wählt man ebenfalls diese Schrittweitenfolgen. Zusätzlich ist die *Art der Perturbationen* $\mathbf{D}_k$ festzulegen, wobei es dort aber eine verbreitete Standardwahl gibt. Außerdem kann man bestimmte Erweiterungen vornehmen, z.B. die Gradientenschätzer-Mittelung (5.2), siehe dazu Abschnitt 6.8.1.

Somit ist das Optimierungsproblem charakterisiert und es wurden die relevanten Schreibweisen für die Verfahren eingeführt.

Kapitel 3

Grundlagen Wahrscheinlichkeitstheorie

3.1 Einleitung

Im SPSA-Verfahren hat man es mit zwei Zufallsquellen zu tun: den zufälligen Perturbationen $\mathbf{D}_k$ für die Gradientenschätzung $\hat{\mathbf{g}}_k(\mathbf{X}_k)$ und den Messfehlern $R_k^{\pm}$. Beide gehen in den stochastischen Prozess $\mathbf{X}_k$ ein (für eine formale Definition eines stochastischen Prozesses siehe Def. 3.51):

$$\mathbf{X}_{k+1} = \mathbf{X}_k - a_k u(\mathbf{X}_k, Z_k), \quad Z_k = (\overline{\mathbf{D}}_k, R_k^{\mathrm{SP}}), \overline{\mathbf{D}}_k = h_k \mathbf{D}_k, \tag{3.1}$$

wobei u eine Notation für die Gradientenschätzung ist, die die Abhängigkeit von den eingehenden Zufallsgrößen berücksichtigt.

Dadurch ergibt sich die Notwendigkeit, mit bedingten Erwartungswerten und anderen Ausdrücken mit mehreren Zufallsgrößen bei verschiedenen Unabhängigkeitsbeziehungen umgehen zu können. Zur Kontrolle der Abweichungen der Gradientenschätzung $\hat{\mathbf{g}}_k(\mathbf{X}_k)$ von $\nabla f(\mathbf{X}_k)$ im SPSA-Verfahren verwendet man außerdem eine Martingalungleichung von Doob. Daher werden im folgenden Kapitel die benötigte wahrscheinlichkeitstheoretische Notation eingeführt, Aussagen der Theorie der stochastischen Prozesse und der allgemeinen Wahrscheinlichkeitstheorie für die Anwendung im Analysekapitel vorbereitet und Rechenregeln hergeleitet, die sich in der benötigten Form nicht in der Standardliteratur finden.

Dieses Kapitel baut wesentlich auf den Skripten von Prof. König zu den Vorlesungen Wahrscheinlichkeitstheorie I+II und Stochastische Prozesse I+II auf (gehalten in Leipzig zwischen 2005 und 2009), sowie für die Martingaltheorie auf dem Buch [Wil91]. Außerdem wurden die Bücher [GS06], [LR79] und [Rao84] verwendet.

Um die Aussage von Spall im SPSA-Konvergenzbeweis, dass für einen bestimmten Ausdruck ein Martingal vorliegt und man die Martingalungleichung von Doob anwenden kann, auf solide wahrscheinlichkeitstheoretische Basis zu stellen, wurde auf `math.stackexchange.com` zurückgegriffen. Insbesondere hat sich ergeben, dass der fragliche Ausdruck ein Submartingal ist, und man die Doobsche Ungleichung, die in dieser Arbeit in Korollar 3.64 und Bemerkung

3.65 in die benötigte Form gebracht wird, auch auf Submartingale anwenden kann.

Einige wahrscheinlichkeitstheoretische Fakten, die man zur Behandlung der bedingten Erwartungswerte mit mehr als 2 Zufallsgrößen für die Analyse des SPSA-Verfahrens benötigt, sind durch Diskussionen auf
Math.StackExchange deutlich geworden. Insbesondere baut die Idee für Hilfssatz 3.13 auf [Msee] auf, beim Beweis wird in dieser Arbeit dann ein etwas anderer Weg gegangen. Ebenso baut Regel (iii) und (iv) in Korollar 3.22 auf Überlegungen in [Msei] und [Mseg] auf. Die Idee zum Einfügen von Hilfssatz 3.25 ist nach [Mseg]. Lemma 3.33 wurde mit [Msed] entwickelt. Überall, wo auf Ideen daraus referenziert wird, befinden sich in dieser Arbeit dennoch die ausgearbeiteten Beweise.

3.2 Zufallsgrößen

In diesem Kapitel seien weiterhin E, E_n messbare Räume mit σ-Algebren $\mathcal{E}$ bzw. $\mathcal{E}_n$. $\mathbb{R}$ sei mit der BORELschen σ-Algebra $\mathcal{B}$ und $\mathbb{R}^m$ entsprechend mit der m-dimensionalen BORELschen σ-Algebra $\mathcal{B}^m$ ausgestattet, jeweils mit dem zugehörigen Lebesgue-Maß λ. $(\Omega, \mathcal{F}, \mathbb{P})$ sei ein Wahrscheinlichkeitsraum.

Def. 3.1. Eine messbare Abbildung $f : \mathbb{R}^{n_1} \to \mathbb{R}^{n_2}$ heißt BOREL-Funktion. Eine messbare Abbildung $X : \Omega \to E$ heißt *Zufallsgröße* (*Zufallsvektor* für $E = \mathbb{R}^n$).

Def. 3.2. Eine Zufallsgröße $X : \Omega \to \mathbb{R}$ heißt *integrierbar*, falls $\int_\Omega |X(\omega)| \, \mathbb{P}(d\omega)$ konvergiert, d.h., falls der Erwartungswert

$$\mathbb{E}(|X|) \coloneqq \int_\Omega |X(\omega)| \, \mathbb{P}(d\omega)$$

existiert ($\Leftrightarrow \mathbb{E}(|X|) < \infty$).

Die Menge aller solcher Zufallsgrößen bezeichnet man mit $\mathcal{L}^1 = \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ (Die $\mathcal{L}^p$-Räume für höhere Integrierbarkeit werden in Def. 3.36 definiert). Eine Zufallsvariable X heißt *zentriert*, falls $\mathbb{E}(X) = 0$.

Def. 3.3. Für die fast sichere Gleichheit zweier Zufallsgrößen X und Y (d.h. $\mathbb{P}(X = Y) = 1$) wird die Notation $X \eqsim Y$ eingeführt.

Die Zufallsgröße $\widetilde{X} : \Omega \to E$ heißt *Version* von $X : \Omega \to E$, falls $\widetilde{X} \eqsim X$.

Def. 3.4. Für Zufallsvektoren X sei $\|X\|$ die euklidische Norm im Bildraum, also $\|X\| \coloneqq \sqrt{\sum_{i=1}^n X_i^2}$. $\|X\|$ ist demnach eine Zufallsvariable.

Def. 3.5. Der *σ-Operator* $\sigma(\cdot)$ bildet Mengensysteme $\mathcal{C}$ (auf Ω) auf die kleinste σ-Algebra $\sigma(\mathcal{C})$ ab, die $\mathcal{C}$ enthält. Er ist monoton, d.h., aus $\mathcal{C}_1 \subseteq \mathcal{C}_2$ folgt $\sigma(\mathcal{C}_1) \subseteq \sigma(\mathcal{C}_2)$ und hat die Eigenschaft $\sigma(C) = C$, falls C eine σ-Algebra ist (vergleiche [Kön08, S. 41]).

Die folgende Definition ist aus [Kön09, Abschnitt 7.1] entnommen:

Def. 3.6. Für eine Zufallsgröße $X : \Omega \to E$ wird die von X erzeugte σ-Algebra gemäß $X^{-1}(\mathcal{E}) \equiv \sigma(X) \coloneqq \{X^{-1}(A) \,|\, A \in \mathcal{E}\}$ definiert. Allgemeiner definiert man: Für Zufallsgrößen $X_i : \Omega \to E_i$ ist

$$\sigma\Big(\bigcup_{i \in I} X_i^{-1}(\mathcal{E}_i)\Big) \equiv \sigma\Big(\bigcup_{i \in I} \sigma(X_i)\Big) \tag{3.2}$$

die von $(X_i)_{i\in I}$ *erzeugte σ-Algebra* (für eine beliebige Indexmenge I).

Insbesondere gilt:

$$\sigma(X_1,\dots,X_n) := \sigma(\sigma(X_1)\cup\dots\cup\sigma(X_n)). \tag{3.3}$$

Dabei verwendet man die Urbildmengen-Notation:

$$X^{-1}(A) := \{\omega\in\Omega \mid X(\omega)\in A\},$$

d.h. konkret z.B. $X^{-1}(\{x\}) = \{\omega\in\Omega \mid X(\omega)=x\}$. Außerdem führt man die Notation $\bigvee_{i\in I}\mathcal{F}_i := \sigma(\bigcup_{i\in I}\mathcal{F}_i)$ ein.

Def. 3.7. Die *Verteilung* einer Zufallsgröße $X:\Omega\to E$, $E=\mathbb{R},\mathbb{R}^m$ definiert man als das Bildmaß $\mathbb{P}\circ X^{-1}$ auf $(E,\mathcal{E})$ ([Kön09, S.94] oder [Wil91, Def. 3.9])

Zufallsgrößen mit symmetrischer Verteilung sind zentriert: Sei X eine Zufallsgröße mit symmetrischer Verteilung f, das heißt $f(-t)=f(t), t\in\mathbb{R}$.
Nach [Kön09, 4.2.5(a)]:
$\mathbb{E}(X) = \int_{\mathbb{R}} tf(t)dt = \int_{-\infty}^{0} tf(t)dt + \int_0^\infty tf(t)dt = \int_0^\infty -tf(-t)dt + \int_0^\infty tf(t)dt =$
$-\int_0^\infty tf(t)dt + \int_0^\infty tf(t)dt = 0.$

Lemma 3.8. *Die Indikatorfunktion*

$$\mathbb{1}_A(x) = \begin{cases} 1 & \text{falls } x\in A \\ 0 & \text{falls } x\notin A \end{cases} \tag{3.4}$$

hat die Eigenschaft: Ist $A\subseteq B$, dann ist $\mathbb{1}_A \le \mathbb{1}_B$.

Beweis. Die Ungleichung $\mathbb{1}_A\le\mathbb{1}_B$ ist erfüllt, falls $\mathbb{1}_A(x)=0$. Ist $\mathbb{1}_A(x)=1$, so ist $x\in A\subseteq B$, d.h. $\mathbb{1}_B(x)=1$, da $x\in B$. □

Lemma 3.9. *Sei $(A_n)_n$ eine Mengenfolge, $A_n\subseteq A_{n+1}$ und sei $A := \bigcup_{n=1}^{\infty} A_n$, dann gilt:*

$$\mathbb{1}_{A_n}\to\mathbb{1}_A \text{ (punktweise)}, \ \mathbb{1}_{A_n}\le\mathbb{1}_A, \ n\to\infty. \tag{3.5}$$

Beweis. Da $A_n\subseteq A$ folgt $\mathbb{1}_{A_n}\le\mathbb{1}_A$ nach Lemma 3.8. Zu zeigen bleibt:

$$\lim_{n\to\infty}\left|\mathbb{1}_{A_n}(x)-\mathbb{1}_A(x)\right| = 0$$

In den Fällen $x\notin A_n, x\in A$ und $x\in A_n, x\notin A$ ist $\mathbb{1}_{A_n}(x)-\mathbb{1}_A(x)\neq 0$, wobei $x\in A_n, x\notin A$ wegen $A_n\subseteq A$ nicht möglich ist, also gilt

$$\left|\mathbb{1}_{A_n}(x)-\mathbb{1}_A(x)\right| = \mathbb{1}_{A\setminus A_n}(x).$$

Betrachte ein festes x. Es gilt $A\setminus A_{n+1}\subseteq A\setminus A_n$, d.h. $\mathbb{1}_{A\setminus A_{n+1}}\le\mathbb{1}_{A\setminus A_n}$. $0\le\mathbb{1}_{A\setminus A_n}\le 1$ ist eine beschränkte und monoton fallende, also konvergente Folge. Sei $c := \lim_{n\to\infty}\mathbb{1}_{A\setminus A_n}(x)$. Ist $c\neq 0$, so müsste $c=1$ gelten und es gäbe ein n_0, so dass für alle $n\ge n_0$
$\mathbb{1}_{A\setminus A_n}(x)=1$, d.h. $x\in A\setminus A_n\ \forall n$, im Widerspruch zu $A=\bigcup_{n=1}^{\infty}A_n$. □

Das Doob-Dynkin-Lemma findet sich z.B. in [Rao84, Satz 3, S. 7].

Lemma 3.10 (Doob-Dynkin-Lemma)**.**
$X : \Omega \to \mathbb{R}$ ist $\sigma(Y_1, \dots, Y_n)$-messbar genau dann, wenn es eine Borel*-messbare Funktion $f : \mathbb{R}^n \to \mathbb{R}$ gibt mit $X = f(Y_1, \dots, Y_n)$.*

Beweis. Setze in [Rao84]: $g \coloneqq X$, $f \coloneqq (Y_1, \dots, Y_n)$, $\Sigma \coloneqq \mathcal{F}$, $\mathcal{A} \coloneqq \mathcal{B}^n$. □

Lemma 3.11. *Ist X $\mathcal{F}_1$-messbar und $\mathcal{F}_1 \subseteq \mathcal{F}_2$ eine σ-Subalgebra, dann ist X auch $\mathcal{F}_2$-messbar.*

Beweis. $\forall A \in \mathcal{E} : X^{-1}(A) \in \mathcal{F}_1 \subseteq \mathcal{F}_2$. □

Bemerkung 3.12. Hat $\mathcal{F}$ die Form $\sigma(Y_i \,|\, i \in I)$ für gewisse Zufallsvariablen Y_i, dann ist $\sigma(Z) \subseteq \mathcal{F}$ eine σ-Subalgebra, falls $Z \in \{Y_i \,|\, i \in I\}$.
Es gilt nämlich: $\sigma(Z) \subseteq \bigcup_{i\in I} \sigma(Y_i) \subseteq \sigma\big(\bigcup_{i\in I} \sigma(Y_i)\big) = \mathcal{F}$ nach der Monotonie des σ-Operators und (3.3).

Der folgende Satz ist eingefügt mit Informationen aus [Msee] und macht eine Aussage über die Dividierbarkeit von $\mathcal{L}^1$-Zufallsgrößen:

Hilfssatz 3.13. *Wenn $E(\left|\frac{1}{X}\right|) < \infty$, dann $X \neq 0$.*

Beweis. $\mathbb{E}\left(\left|\frac{1}{X}\right|\right) = \int_{\{\omega\in\Omega:X(\omega)=0\}} \left|\frac{1}{X(\omega)}\right| \mathbb{P}(d\omega) + \int_{\{\omega\in\Omega:X(\omega)\neq 0\}} \left|\frac{1}{X(\omega)}\right| \mathbb{P}(d\omega)$ und das erste Integral wäre unendlich, wenn $\mathbb{P}(X = 0) \neq 0$.

Genauer folgt mit [Kön09, Def. 6.5.1(ii)] für

$$\int_{\{\omega\in\Omega:X(\omega)=0\}} \left|\frac{1}{X(\omega)}\right| \mathbb{P}(d\omega) = \int \mathbb{1}_{\{X=0\}} \left|\frac{1}{X}\right| \mathbb{P}(d\omega)$$

mit der Folge $f_n = n\mathbb{1}_{\{X=0\}}$ nichtnegativer einfacher Funktionen

$$\int \mathbb{1}_{\{X=0\}} \left|\tfrac{1}{X}\right| \mathbb{P}(d\omega) = \lim_{n\to\infty} \int n\mathbb{1}_{\{X=0\}} \mathbb{P}(d\omega) = \lim_{n\to\infty} n \underbrace{\int \mathbb{1}_{\{X=0\}} \mathbb{P}(d\omega)}_{=\mathbb{P}(X=0)=:c}.$$

Das ist der Grenzwert der Folge $x_n = cn$, der ausschließlich für $c = 0$ endlich ist. Also ist $P(X = 0) = 0$. □

Bemerkung 3.14. Ist eine Zufallsgröße X fast sicher ungleich 0, so kann man sie auf der Nullmenge $\{X = 0\}$ auf einen beliebigen endlichen Wert abändern und erhält eine Zufallsgröße $\widetilde{X}$ mit den Eigenschaften $\widetilde{X} \neq 0$ und $\widetilde{X} \eqsim X$. So erhält man eine überall invertierbare Version der Zufallsgröße $X : \Omega \to \mathbb{R}$. Man bekommt solch eine Version z.B., indem man setzt $\widetilde{X} := X\mathbb{1}_{\{X\neq 0\}}$ (denn für diese gilt $\mathbb{P}(X \neq \widetilde{X}) = \mathbb{P}(X = 0) = 0$, da $\{X = 0\}$ eine Nullmenge ist).

Sind die Komponenten des Zufallsvektors $X_i \neq 0$, so gibt es eine Version $\widetilde{X}$ des Zufallsvektors X mit $\widetilde{X}_i \neq 0$ und $\widetilde{X} \eqsim X$. Seien $\widetilde{X}_i$ die Versionen von X_i mit $\widetilde{X}_i \neq 0$. Für diese gilt $\mathbb{P}(\widetilde{X}_i \neq X_i) = 0$. Für $\widetilde{X} \coloneqq (\widetilde{X}_1, \dots, \widetilde{X}_n)$ gilt $X \eqsim \widetilde{X}$, denn sogar $\mathbb{P}(X \neq \widetilde{X}) = \mathbb{P}(\{X_1 \neq \widetilde{X}_1\} \cup \dots \cup \{X_n \neq \widetilde{X}_n\}) \leq \mathbb{P}(X_1 \neq \widetilde{X}_1) + \dots + \mathbb{P}(X_n \neq \widetilde{X}_n) = 0$ nach der (σ-)Subadditivität von Maßen.

In dieser Situation erklärt man den Ausdruck $\frac{c}{\mathbf{X}}$ für einen m-dimensionalen Zufallsvektor $\mathbf{X}$ wie folgt:

$$\frac{c}{\mathbf{X}} \coloneqq \left(\frac{c}{\widetilde{\mathbf{X}}_1}, \dots, \frac{c}{\widetilde{\mathbf{X}}_m}\right). \tag{3.6}$$

Def. 3.15. Die Definition 2.4 der Erwartungstreue eines Schätzers P erweiternd heißt eine Folge von Schätzern P_k *asymptotisch erwartungstreu*, wenn

$$\lim_{k\to\infty} \text{Bias}(P_k) = 0. \tag{3.7}$$

Der folgende Satz ist nach [Kön09, Satz 6.8.1].

Satz 3.16 (Monotoner Konvergenzsatz)**.** *Sei $0 \leq (X_n)_n \leq X$ f.s., $X_n \to X$ f.s. Dann gilt:*

$$\lim_{n\to\infty} \mathbb{E}(X_n) = \mathbb{E}(X). \tag{3.8}$$

Die rechte Seite $\mathbb{E}(X)$ ist endlich, falls $X \in \mathcal{L}^1$.

Es folgt die Cauchy-Schwarzsche Ungleichung aus [Kön09, Satz 3.5.6]:

Satz 3.17 (Cauchy-Schwarzsche Ungleichung)**.**
Für zwei Zufallsgrößen X und Y gilt:

$$\left(\mathbb{E}(XY)\right)^2 \leq \mathbb{E}(X^2)\ \mathbb{E}(Y^2). \tag{3.9}$$

3.3 Unabhängigkeit

Def. 3.18 (Unabhängigkeit von Zufallsgrößen)**.**
Die Zufallsgrößen $\{X_n\}_{n\in I}$ ($X_n : \Omega \to E_n$) heißen unabhängig, wenn die σ-Algebren $X_n^{-1}(\mathcal{E}_n)$ unabhängig sind. Insbesondere:

(a) Ein Vektor $(X_1, \ldots, X_m)$ von Zufallsgrößen $X_i : \Omega \to \mathbb{E}$ heißt unabhängig, wenn die σ-Algebren $X_1^{-1}(\mathcal{E}), \ldots, X_m^{-1}(\mathcal{E})$ unabhängig sind.

(b) Zwei Vektoren $X : \Omega \to \mathbb{R}^{n_1}$, $Y : \Omega \to \mathbb{R}^{n_2}$ heißen unabhängig, wenn die σ-Algebren $X^{-1}(\mathcal{B}^{n_1})$ und $Y^{-1}(\mathcal{B}^{n_2})$ unabhängig sind.

Betrachtet man den Ausdruck $\sigma(X, Y)$, so kann dies als erzeugte σ-Algebra von zwei Zufallsvariablen X, Y verstanden werden (nach (3.3)), oder (X, Y) kann wie in Def. 3.18(a) als Zufallsvektor aufgefasst werden und die davon erzeugte σ-Algebra gemeint sein. Beide Varianten beschreiben aber die gleiche σ-Algebra. Das folgende Lemma ist nach [Mseg]:

Lemma 3.19. *Seien $X, Y : \Omega \to \mathbb{R}$ Zufallsgrößen. Die vom Zufallsvektor $T = (X, Y)$ erzeugte σ-Algebra ist gleich der von den Zufallsgrößen X, Y erzeugte σ-Algebra $\mathcal{G} \coloneqq \sigma(X, Y)$, d.h., es gilt $\sigma(T) = \mathcal{G}$.*

Beweis. Der Beweis ist nach [Piaa].

„$\subseteq$“ Zu zeigen ist, dass $T^{-1}(B) \in \mathcal{G} \quad \forall B \in \mathcal{B}^2$. Betrachtet man die Menge $\mathcal{D} = \{B \in \mathcal{B}^2 \,|\, T^{-1}(B) \in \mathcal{G}\}$, so ist dies äquivalent zu $\mathcal{B}^2 \subseteq \mathcal{D}$. Es gilt $\mathcal{B} \times \mathcal{B} \subseteq \mathcal{B}^2$, sogar $\mathcal{B}^2 = \sigma(\mathcal{B} \times \mathcal{B})$, siehe z.B. [Wil91, S. 80]. Ist B von der Form $B_1 \times B_2$, $B_1, B_2 \in \mathcal{B}$. $T^{-1}(B_1 \times B_2) = X^{-1}(B_1) \cap Y^{-1}(B_2)$ $\forall B_1, B_2 \in \mathcal{B}$ und es folgt $\mathcal{B} \times \mathcal{B} \subseteq \mathcal{D}$. Wende auf die Inklusionen $\mathcal{B} \times \mathcal{B} \subseteq \mathcal{D} \subseteq \mathcal{B}^2$ den $\sigma(\cdot)$-Operator an. Es folgt:

$$\sigma(\mathcal{B} \times \mathcal{B}) \subseteq \sigma(\mathcal{D}) \subseteq \sigma(\mathcal{B}^2), \text{ d.h.}$$
$$\mathcal{B}^2 \subseteq \mathcal{D} \subseteq \mathcal{B}^2,$$

da $\mathcal{D}$ eine σ-Alegra ist. D.h. $\mathcal{D} = \mathcal{B}^2$ und es folgt $\sigma(T) \subseteq \mathcal{G}$.

„$\supseteq$“ Es gilt: $X^{-1}(B) = T^{-1}(B \times \mathbb{R})$ sowie $Y^{-1}(B) = T^{-1}(\mathbb{R} \times B) \quad \forall B \in \mathcal{B}$, daher folgt: $\sigma(X) \cup \sigma(Y) \subseteq \sigma(T) = T^{-1}(\mathcal{B}^2)$. Anwenden des σ-Operators liefert: $\mathcal{G} \equiv \sigma(\sigma(X) \cup \sigma(Y)) \subseteq \sigma(T)$.

Der Zwischenschritt, warum $\mathcal{D}$ eine σ-Algebra ist, wurde hier nicht ausformuliert. In [Wil91, S. 207] findet sich auch eine Herleitung, warum $\sigma(Y) = \sigma(Y_1, Y_2)$, $Y = (Y_1, Y_2)$. □

Ein Lemma zur Unabhängigkeit nach [Kön09, Lemma 3.2.11] wird angegeben:

Lemma 3.20. *Sei $(X_i)_{i \in I}$ eine Familie unabhängiger Zufallsgrößen und I eine beliebige Indexmenge. I_1 und I_2 seien disjunkte Teilmengen von I. Definiere die Zufallsvariablen $Y_1 = f_1\big((X_i)_{i \in I_1}\big)$ und $Y_2 = f_2\big((X_i)_{i \in I_2}\big)$ mit geeigneten[1] Funktionen f_1 und f_2. Dann sind Y_1 und Y_2 unabhängig.*

Das folgende Lemma entspricht Lemma 7.2.13(b) in [Kön09]:

Lemma 3.21. *Seien $X_n : \Omega \to E_n$ unabhängige Zufallsgrößen, $\varphi_n : E_n \to E'_n$ $\mathcal{E}_n - \mathcal{E}'_n$-messbare Abbildungen, dann sind $\varphi_n \circ X_n$ ebenfalls unabhängig.*

Korollar 3.22. *Daraus folgt insbesondere:*

(i) Seien $X : \Omega \to \mathbb{R}^m$ und $Y : \Omega \to E$ unabhängig, dann sind auch X_i und Y unabhängig.

(ii) Seien $X : \Omega \to \mathbb{R}^m$ und $Y : \Omega \to E$ unabhängig, alle Komponenten $X_i \neq 0$, dann sind auch Y und X_i / X_l unabhängig für alle $i \neq l$.

(iii) Seien $X : \Omega \to \mathbb{R}$; $Y : \Omega \to \mathbb{R}^m$ unabhängige Zufallsgrößen und $\frac{1}{X} \in \mathcal{L}^1$. Dann sind auch Y und $\frac{1}{X} : \Omega \to \mathbb{R}$ unabhängig.

(iv) Seien $X : \Omega \to \mathbb{R}^{n_1}$, $Y : \Omega \to \mathbb{R}^{n_2}$ unabhängige Zufallsgröße und $\frac{1}{X_i} \in \mathcal{L}^1$, $i = 1, \ldots, n_1$, dann sind für festes i auch Y und $\frac{1}{X_i} : \Omega \to \mathbb{R}$ unabhängig [Msei; Mseg].

Beweis.

(i) Wähle $X_1 = Y$, $\varphi_1 \coloneqq \mathrm{id} : E \to E$ und
$X_2 = X$, $\varphi_2 \coloneqq \Phi_i : \mathbb{R}^m \to \mathbb{R}$ die Abbildung auf die i-te Koordinate, sie ist stetig also messbar, in Lemma 3.21.

(ii) Wähle $X_1 = Y$, $\varphi_1 \coloneqq \mathrm{id} : E \to E$ und
$X_2 = X$, $\varphi_2 : \mathbb{R}^m \to \mathbb{R}, X \mapsto \frac{X_i}{X_l} = \frac{\Phi_i(X)}{\Phi_l(X)}$ in Lemma 3.21. φ_2 ist messbar, da $X_l(\omega) \neq 0 \ \forall \omega \in \Omega$, [Kön09, Lemma 6.4.6c].

(iii) Da $X \neq 0$ nach Hilfsatz 3.13, kann man O.B.d.A. X auf einer $\mathbb{P}$-Nullmenge abändern, so dass $X : \Omega \to \mathbb{R} \setminus \{0\}$. Wende Lemma 3.21 an für $n = 2$,
$X_1 = Y$, $\varphi_1 = \mathrm{id} : \mathbb{R}^m \to \mathbb{R}^m$,
$X_2 = X$, $\varphi_2 : \mathbb{R} \setminus \{0\} \to \mathbb{R}$, $x \mapsto \frac{1}{x}$.

[1] [Kön09] ist an dieser Stelle nicht genauer. Das Lemma wird aber für den Fall verwendet, dass f_1 und f_2 die identische Funktion sind, so dass die Erfülltheit der Voraussezung in diesem Fall als sinnvolle Annahme erscheint.

(iv) Nach (i) sind auch X_i, Y unabhängig und dann nach (iii) auch $\frac{1}{X_i}$ und Y.

□

Hilfssatz 3.23. *Seien die σ-Algebren $\mathcal{E}_1, \mathcal{E}_2$ unabhängig und sei $\mathcal{D}_1$ eine σ-Subalgebra von $\mathcal{E}_1$, dann sind $\mathcal{D}_1$ und $\mathcal{E}_2$ unabhängig.*

Beweis. Spezialfall von [Kön09, Lemma 7.2.3(a)] mit $I = \{1, 2\}, \mathcal{E}_2 = \mathcal{D}_2$. □

Die Idee für den folgenden Satz beruht auf [Piaa].

Lemma 3.24. *Seien $X, Y, Z : \Omega \to E$ Zufallsgrößen. X und (Y, Z) seien unabhängig. Dann ist X auch von Y unabhängig.*

Beweis. Zu zeigen ist, dass die Unabhängigkeit von $\sigma(X), \sigma(Y, Z)$ die Unabhängigkeit von $\sigma(X)$, $\sigma(Y)$ impliziert. Nach Bemerkung 3.12 ist $\sigma(Y) \subseteq \sigma(Y, Z)$ und aus Hilfssatz 3.23 folgt die Behauptung mit $\mathcal{E}_1 = \sigma(Y, Z)$, $\mathcal{D}_1 = \sigma(Y)$, $\mathcal{E}_2 = \sigma(X)$. □

Das folgende Lemma ist nach [Mseg].

Hilfssatz 3.25. *Für eine Zufallsvariable $T : \Omega \to E$ und eine messbare Funktion $u : E \to E'$ gilt: $\sigma(u(T)) \subseteq \sigma(T)$.*

Beweis. Sei $A \in \sigma(u(T))$, d.h. $\exists B \in \mathcal{E}'$ so, dass $A = \{u(T) \in B\}$.
Setze $C \coloneqq u^{-1}(B) \overset{n.\,Def.}{=} \{s \,|\, u(s) \in B\}$. Dann gilt $A = \{T \in C\}$, d.h. $A \in \sigma(T)$. □

Hilfssatz 3.26. *Wenn X $\mathcal{A}$-messbar ist, dann ist $\sigma(X)$ eine Teil-σ-Algebra von $\mathcal{A}$.*

Beweis. Es ist $X^{-1}(B) \in \mathcal{A}$ für alle $B \in \mathcal{E}$. Zu zeigen ist: Für alle $C \in \sigma(X)$ gilt $C \in \mathcal{A}$. Da $\sigma(X) = \{X^{-1}(B) \,|\, B \in \mathcal{E}\}$ (Def. 3.6), existiert für jedes $C \in \sigma(X)$ ein $B \in \mathcal{E}$ mit $C = X^{-1}(B)$. □

Hilfssatz 3.27. *Seien $X, Z : \Omega \to \mathbb{R}^m$ und $Y : \Omega \to \mathbb{R}$ Zufallsgrößen. Sind X und (Y, Z) unabhängig, $Y \neq 0$, dann sind auch X und $(\frac{1}{Y}, Z)$ unabhängig.*

Beweis. Wende Hilfssatz 3.25 auf $u : (y, z) \mapsto (\frac{1}{y}, z)$ an.
Dann ist $\sigma(\frac{1}{Y}, Z) \subseteq \sigma(Y, Z)$ und die Behauptung folgt mit Hilfssatz 3.23 (mit $\mathcal{E}_2 = \sigma(X)$ und $\mathcal{E}_1 = \sigma(Y, Z)$, $\mathcal{D}_1 = \sigma(\frac{1}{Y}, Z)$). □

Das folgende Lemma ist nach [Kön09, Satz 7.2.14].

Lemma 3.28. *Seien X,Y unabhängige reellwertige Zufallsgrößen, X, Y, $XY \in \mathcal{L}^1$. Dann gilt:*

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y). \tag{3.10}$$

3.4 Bedingter Erwartungswert

Es folgen zwei Definitionen aus [Kön09, Def. 7.3.2 und 7.3.7]:

Def. 3.29. Für eine integrierbare Zufallsgröße $Y \in \mathcal{L}^1$ und eine σ-Subalgebra $\mathcal{A}$ von $\mathcal{F}$ definiert man den bedingten Erwartungswert $\mathbb{E}(Y \,|\, \mathcal{A})$ als die fast sicher eindeutige Zufallsgröße $Z := \mathbb{E}(X \,|\, \mathcal{A})$ mit den Eigenschaften

(i) Z ist $\mathcal{A}$-messbar und

(ii) $\mathbb{E}(X\mathbb{1}_A) = \mathbb{E}(Z\mathbb{1}_A) \quad \forall A \in \mathcal{A}$.

Ersetzt man $\mathcal{A}$ durch $\sigma(Y)$ erhält man die Definition des bedingten Erwartungswerts auf eine Zufallsgröße:

Def. 3.30 (Bedingter Erwartungswert auf eine Zufallsgröße)**.**
Für eine $\mathcal{L}^1$- Zufallsgröße X definiert man den bedingten Erwartungswert

$$\mathbb{E}(X \,|\, Y) := \mathbb{E}(X \,|\, \sigma(Y))$$

bezüglich einer Zufallsgröße $Y : \Omega \to E$, $E = \mathbb{R}, \mathbb{R}^m$, *als die fast sicher eindeutige Zufallsgröße $Z := \mathbb{E}(X \,|\, Y)$ mit den Eigenschaften:*

(i) Z ist $\sigma(Y)$-messbar und

(ii) $\mathbb{E}(X\mathbb{1}_A) = \mathbb{E}(Z\mathbb{1}_A) \quad \forall A \in \sigma(Y)$.

Bemerkung 3.31. Im Hinblick auf die Definition bis auf fast sichere Gleichheit des bedingten Erwartungswerts wurde das Symbol $\doteq$ in Def. 3.3 eingeführt. Auch wenn es nicht gesondert erwähnt wird, sind die Aussagen über den Erwartungswert bis auf fast sichere Gleichheit zu verstehen.

Der bedingte Erwartungswert $\mathbb{E}(Y \,|\, X)$ ist insbesondere eine f.s. endliche, integrierbare Zufallsgröße: Betrachte dazu (ii) für $A = \Omega$ und beachte, dass Y integrierbar. Mit $\mathbb{E}(Z) < \infty$ folgt Z fast sicher endlich.

Lemma 3.32. *Eigenschaften des bedingten Erwartungswerts [Kön09, Satz 7.3.8].*

(i) Er ist linear in X: $\mathbb{E}(X_1 + cX_2 \,|\, \mathcal{A}) = \mathbb{E}(X_1 \,|\, \mathcal{A}) + c\mathbb{E}(X_2 \,|\, \mathcal{A})$.

(ii) Ist Y $\mathcal{A}$-messbar, so gilt $\mathbb{E}(Y \,|\, \mathcal{A}) = Y$.

(iii) Ist $X \leq Y$ f.s., dann gilt $\mathbb{E}(X \,|\, \mathcal{A}) \leq \mathbb{E}(Y \,|\, \mathcal{A})$ *f.s.*

(iv) Der bedingte Erwartungswert $\mathbb{E}(\cdot \,|\, \mathcal{A})$ ist monoton.

(v) Die Turmeigenschaft gilt: Ist $\mathcal{A}_1 \subseteq \mathcal{A}$ eine Teil-σ-Algebra, dann ist $\mathbb{E}\big[\mathbb{E}[X \,|\, \mathcal{A}_1] \,|\, \mathcal{A}\big] \doteq \mathbb{E}\big[\mathbb{E}[X \,|\, \mathcal{A}] \,|\, \mathcal{A}_1\big] \doteq \mathbb{E}[X \,|\, \mathcal{A}_1]$, *[Kön09, Satz 7.3.8.(iv)].*

Lemma 3.33. *Sei $X \in \mathcal{L}^1(\mathbb{P})$.*
Dann gilt für jede σ-Algebra $\mathcal{A}$ auf Ω: Aus $\mathbb{E}(X \,|\, \mathcal{A}) \doteq 0$ folgt $\mathbb{E}(X) = 0$. Insbesondere: Aus $\mathbb{E}[X \,|\, Y] \doteq 0$ folgt $\mathbb{E}(X) = 0$ für alle Zufallsgrößen $Y : \Omega \to \mathbb{R}$.

Beweis. Nach der Definition von $\mathbb{E}[X \,|\, \mathcal{A}]$ gilt $\mathbb{E}\big(\mathbb{E}[X \,|\, \mathcal{A}]\mathbb{1}_A\big) = \mathbb{E}(X\mathbb{1}_A)$.
Für $A = \Omega$ gilt daher : $\mathbb{E}[\underbrace{\mathbb{E}(X \,|\, \mathcal{A})}_{\doteq 0}] = \mathbb{E}(X)$.

□

Lemma 3.34 (Regeln)**.**
Es gelten die folgenden Regeln für Zufallsgrößen X, Y, Z, die nach $\mathbb{R}$ bzw. $\mathbb{R}^m$ abbilden:

(i) Sei X unabhängig von (Y, Z). Dann gilt: $\mathbb{E}[XY \mid Z] = \mathbb{E}(X)\mathbb{E}[Y \mid Z]$.

(ii) Seien $X, Z : \Omega \to \mathbb{R}^m$ unabhängig, $X_i \neq 0$ für alle Komponenten von X und $f : \mathbb{R}^m \to \mathbb{R}$ messbar. Dann gilt:

$$\mathbb{E}\Big[\tfrac{X_i}{X_j} f(Z) \,\Big|\, Z\Big] = \mathbb{E}[f(Z) \mid Z]\; \mathbb{E}\Big(\tfrac{X_i}{X_j}\Big)$$

für $i \neq j$.

(iii) Wenn X, Y unabhängig und X, Z unabhängig sind, dann sind auch X, $\sigma(Y) \cup \sigma(Z)$ unabhängig.

(iv) Es gilt:

$$\begin{aligned} \mathbb{E}(|X|) \leq c \quad &\Rightarrow \quad \mathbb{E}(X) \leq c \\ X \leq c \quad &\Rightarrow \quad \mathbb{E}(X) \leq c. \end{aligned}$$

(v) $\mathbb{E}(0 \mid \mathcal{A}) \doteq 0$.

(vi) Falls $f(X) \leq \alpha$ und $Y \geq 0$ f.s., so folgt $\mathbb{E}[f(X)Y \mid X] \leq \alpha \mathbb{E}[Y \mid X]$. Insbesondere: $X \leq c$ f.s. $\Rightarrow \mathbb{E}[X \mid \mathcal{A}] \leq c$ f.s.

(vii) Es gilt: $|X| \geq X$ und daher nach Monotonie des bedingten Erwartungswerts $|\mathbb{E}[X \mid \mathcal{A}]| \leq |\mathbb{E}[|X| \mid \mathcal{A}]| = \mathbb{E}(|X| \mid \mathcal{A})$.

Beweis.
[**Regel (i)**] Der Beweis für diese Regel ist nach Piau, [Piaa]. Nach Lemma 3.24 ist X auch unabhängig von Z und von Y. Man testet, ob $\mathbb{E}(X)\mathbb{E}[Y \mid Z]$ die Bedingungen der Definition des bedingten Erwartungswerts $\mathbb{E}(XY \mid Z)$ erfüllt:

(1) $\mathbb{E}(X)\mathbb{E}[Y \mid Z]$ ist $\sigma(Z)$-messbar, da $\mathbb{E}(X)$ eine Konstante ist und die Zufallsgröße $\mathbb{E}[Y \mid Z]$ nach ihrer Definition $\sigma(Z)$-messbar ist.

(2) Zu zeigen ist $\mathbb{E}[\mathbb{E}(X)\ \mathbb{E}(Y \mid Z)\mathbb{1}_A] = \mathbb{E}(XY\mathbb{1}_A) \quad \forall A \in \sigma(Z)$.
Da $E(X)$ ein konstanter Faktor ist, gilt:

$$\begin{aligned} \mathbb{E}\big(\mathbb{E}(X)\ \mathbb{E}[Y \mid Z]\mathbb{1}_A\big) &= \mathbb{E}(X)\ \mathbb{E}\big(\mathbb{E}[Y \mid Z]\mathbb{1}_A\big) \\ &= \mathbb{E}(X)\ \mathbb{E}(Y\mathbb{1}_A) \text{ nach (ii) von Def. für } \mathbb{E}[Y \mid Z] \\ &= \mathbb{E}(XY\mathbb{1}_A), \text{ da } X \text{ unabhängig von } (Z, Y). \end{aligned}$$

[**Regel (ii)**] Um Regel (i) zu verwenden, müssen $\frac{X_i}{X_j}$ und $(Z, f(Z))$ unabhängig sein. Wegen $(Z, f(Z)) = \sigma\Big(\sigma(Z) \cup \sigma(f(Z))\Big)$, $\sigma(f(Z)) \subseteq \sigma(Z)$ nach Lemma 3.25, also $(Z, f(Z)) = \sigma(\sigma(Z)) = \sigma(Z)$, reicht es zu zeigen, dass $\frac{X_i}{X_j}$ unabhängig von Z sind. Benutze dafür Lemma 3.22 (ii) mit $Y = Z$.
[**Regel (iii)**] Diese geht auf einen Spezialfall von [Kön09, Lemma 7.2.3] für σ-Algebren $\mathcal{D}, \mathcal{E}, \mathcal{F}$ zurück: Wenn $\mathcal{D}$ unabhängig von $\mathcal{E}$ und von $\mathcal{F}$, dann ist

$\mathcal{D}$ auch unabhängig von $\mathcal{E} \cup \mathcal{F}$ und entsprechend für Zufallsgrößen: Wenn X unabhängig von Y und von Z ist, dann auch von $\sigma(Y) \cup \sigma(Z)$.
[Regel (iv), 2. Teil] $X \leq c \Rightarrow \mathbb{E}(X) \leq \mathbb{E}(c) = c\mathbb{P}(\Omega) = c$.
[Regel (v)] Nach Definition gilt für $Z = \mathbb{E}(0 \,|\, \mathcal{A})$: $\mathbb{E}(0\, \mathbb{1}_A) = \mathbb{E}(Z \mathbb{1}_A)$. Für $A = \Omega$: $0 = 0\,\mathbb{P}(\Omega) = \mathbb{E}(0) = \mathbb{E}(Z)$, d.h. $Z \doteq 0$.
[Regel (vi)] Man schließt, dass $f(X)Y \leq \alpha Y$ f.s., woraus folgt:
$\mathbb{E}[f(X)Y \,|\, X] \leq \mathbb{E}[\alpha Y \,|\, X] = \alpha \mathbb{E}[Y \,|\, X]$ nach Linearität und Monotonie. □

Lemma 3.35. *Sei f messbar und X eine Zufallsgröße.*
Mit der σ-Algebra $\mathcal{G} = \sigma(\{X, Y_i, i = 1, \ldots, n\})$ gilt $\mathbb{E}[f(X) \,|\, \mathcal{G}] = f(X)$.

Beweis. Nach dem Lemma 3.10 (Doob-Dynkin-Lemma) ist $f(X)$ $\mathcal{G}$-messbar und man verwendet Eigenschaft (ii) des bedingten Erwartungswerts. □

3.5 $\mathcal{L}^1$ und $\mathcal{L}^2$

Eine an [Kön09, S. 81] angelehnte Definition des $\mathcal{L}^p$-Raums wird angeführt:

Def. 3.36. Für ein Zufallsgröße $X : \Omega \to \mathbb{R}$ und $p \in [1, \infty]$ definiert man:

$$\|X\|_{\mathcal{L}^p} = \begin{cases} \big(\mathbb{E}(|X|^p)\big)^{1/p} & \text{falls } p < \infty \\ \sup\{c \geq 0 : \mathbb{P}(|X| > c) \geq 0\} & \text{falls } p = \infty \end{cases} \tag{3.11}$$

und

$$\mathcal{L}^p(\Omega, \mathcal{F}, \mathbb{P}) = \mathcal{L}^p \coloneqq \{X : X \text{ ist eine Zufallsgröße und } \|X\|_{\mathcal{L}^p} < \infty\}. \tag{3.12}$$

Man verwendet insbesondere auch die Bezeichnung $\|X\|_{\mathcal{L}^\infty} = \operatorname{ess\,sup}_{\omega \in \Omega} |X(\omega)|$.

Für $\mathcal{L}^2$-Zufallsgrößen definiert man die Varianz $\mathbb{V}\mathrm{ar}(X)$ durch

$$\mathbb{V}\mathrm{ar}(X) = \mathbb{E}\Big(\big(X - \mathbb{E}(X)\big)^2\Big).$$

Der folgende Satz ist aus [Kön09, Satz 6.7.5]:

Satz 3.37 (Hölder'sche Ungleichung)**.** *Seien $p, q \in [1, \infty]$ und $\frac{1}{p} + \frac{1}{q} = 1$ und seien $X \in \mathcal{L}^p$ und $Y \in \mathcal{L}^q$. Dann ist $XY \in \mathcal{L}^1$ und*

$$\mathbb{E}(XY) \leq \mathbb{E}(X^p)^{1/p} \mathbb{E}(Y^q)^{1/q}. \tag{3.13}$$

Die Minkowski-Ungleichung aus [Kön09, Satz 6.7.9] wird für Zufallsgrößen angegeben:

Satz 3.38 (Minkowski-Ungleichung)**.** *Sei $p \in [1, \infty]$.*
Dann gilt für alle $X, Y \in \mathcal{L}^2$:

$$\Big(\mathbb{E}\big((X+Y)^p\big)\Big)^{1/p} \leq \Big(\mathbb{E}(X^p)\Big)^{1/p} + \Big(\mathbb{E}(Y^p)\Big)^{1/p}. \tag{3.14}$$

Def. 3.39. Man definiert den Erwartungswert eines Zufallsvektors $\mathbf{X} : \Omega \to \mathbb{R}^m$ als $\mathbb{E}(\mathbf{X}) \coloneqq \big(\mathbb{E}(\mathbf{X}_1), \ldots, \mathbb{E}(\mathbf{X}_m)\big)$ (siehe etwa [Kön09, S. 50]). Für einen Zufallsvektor $\mathbf{X}$ wird $\mathbf{X} \in \mathcal{L}^p$ dementsprechend so definiert, dass die Komponenten $\mathbf{X}_i$ $\mathcal{L}^p$-Zufallsgrößen sind.

Die folgende Definition geht auf [Wil91, Abschnitt 12.0] zurück.

Def. 3.40. Sei X_k eine Folge von Zufallsgrößen. X_k heißt *in $\mathcal{L}^p$ beschränkt* (im Fall $p = 2$ auch gleichmäßig varianzbeschränkt),wenn

$$\sup_{k\geq 0} \mathbb{E}(|X_k|^p) < \infty. \tag{3.15}$$

Das heißt, es gibt ein $c \geq 0$ unabhängig von k, so dass $\mathbb{E}(|X_k|^p) \leq c \ \forall k$. Die Bezeichnung varianzbeschränkt geht auf das folgende Lemma zurück:

Lemma 3.41. *Sei $X \in \mathcal{L}^1$. $X \in \mathcal{L}^2$ gilt genau dann, wenn $\mathbb{V}\mathrm{ar}(X) < \infty$.*

Beweis. Nach der STEINERschen Formel [Kön09, S. 94] gilt $\mathbb{V}\mathrm{ar}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$, sodass $\mathbb{V}\mathrm{ar}(X) < \infty$ genau dann, wenn $\mathbb{E}(|X|^2) < \infty$. □

Der folgende Satz ist aus [Kön09, 6.7.2 und 7.3.15] entnommen:

Satz 3.42 (JENSENsche Ungleichung)**.**
Sei $\mathbb{I} \subset \mathbb{R}$ ein Intervall, $X \in \mathcal{L}^1(\mathbb{P})$ eine $\mathbb{I}$-wertige Zufallsgröße und $\varphi : \mathbb{I} \to \mathbb{R}$ konvex. Dann gilt:

$$\mathbb{E}(\varphi(X)) \geq \varphi\left(\mathbb{E}(X)\right), \tag{3.16}$$

$$\mathbb{E}(\varphi(X) \,|\, \mathcal{A}) \geq \varphi\left(\mathbb{E}(X \,|\, \mathcal{A})\right) \ \textit{f.s.} \tag{3.17}$$

Für $\varphi(x) = x^2$ folgt insbesondere:

$$c \geq \mathbb{E}(X^2) \quad \Rightarrow c \geq \mathbb{E}(X)^2. \tag{3.18}$$

Lemma 3.43. *Sei $(X_k)_k$ $\mathcal{L}^2$-beschränkt. Dann ist auch $\mathbb{E}[X_k \,|\, \mathcal{A}]$ (mit derselben Konstanten) $\mathcal{L}^2$-beschränkt. Ist $X \in \mathcal{L}^2$, dann ist $\mathbb{E}[X \,|\, \mathcal{A}] \in \mathcal{L}^2$.*

Beweis. Da X_k $\mathcal{L}^2$-beschränkt gibt es ein $c \geq 0$, so dass $\mathbb{E}(X_k^2) \leq c$. Nach der Definition des bedingten Erwartungswerts gilt für die Zufallsgröße $Y_k \coloneqq X_k^2$: $\mathbb{E}\big(\mathbb{E}[Y_k \,|\, \mathcal{A}]\big) = \mathbb{E}(Y_k \mathbb{1}_\Omega) \leq c$ und damit nach der JENSENschen Ungleichung (3.17) für $\varphi(x) = x^2$: $(\mathbb{E}[X_k \,|\, \mathcal{A}])^2 \leq \mathbb{E}[X_k^2 \,|\, \mathcal{A}]$. Mit der Monotonie des Erwartungswerts:

$$\mathbb{E}\big((\mathbb{E}[X_k \,|\, \mathcal{A}])^2\big) \leq \mathbb{E}\big(\mathbb{E}[X_k^2 \,|\, \mathcal{A}]\big) = \mathbb{E}(X_k^2 \mathbb{1}_\Omega) \leq c, \tag{3.19}$$

d.h., $Z_k \coloneqq \mathbb{E}[X_k \,|\, \mathcal{A}]$ ist $\mathcal{L}^2$-beschränkt. Setzt man $X_k = X \ \forall k$ für ein $X \in \mathcal{L}^2$, folgt die zweite Aussage des Lemmas. □

Korollar 3.44. *Es gilt: $\mathcal{L}^2 \subset \mathcal{L}^1$.*

Beweis. Sei $Y \in \mathcal{L}^2$, d.h., es gibt ein $c \in \mathbb{R}$, so dass $E(Y^2) \leq c$. Nach der JENSENschen Ungleichung (3.18) für $X = |Y|$ gilt $E\big(|Y|\big)^2 \leq c$, insbesondere $E\big(|Y|\big) < \infty$ oder $Y \in \mathcal{L}^1$. □

Dieses Resultat lässt sich auf Folgen von Zufallsgrößen erweitern:

Lemma 3.45. *Jede $\mathcal{L}^2$-beschränkte Folge von Zufallsgrößen ist auch $\mathcal{L}^1$-beschränkt.*

Beweis. Sei X_k $\mathcal{L}^2$-beschränkt, d.h. $\sup_{k\geq 0}\mathbb{E}(|X_k|^2) < \infty$, es existiert also ein c, dass nicht von k abhängt, so dass

$$\mathbb{E}(|X_k|^2) \leq c \tag{3.20}$$

Nach der JENSENschen Ungleichung (3.16) für $X = |X_k|$ folgt $\mathbb{E}(|X_k|)^2 \leq c'$, d.h. $\sup_{k\geq 0}\mathbb{E}(|X_k|) \leq c$, wie gezeigt werden sollte. □

Lemma 3.46. *Ist der Zufallsvektor $\mathbf{X}$ eine $\mathcal{L}^2$-Zufallsgröße, dann ist auch seine Norm $\|X\|$ eine $\mathcal{L}^2$-Zufallsgröße.*

Beweis. Es gilt:

$$\mathbb{E}(\|\mathbf{X}\|^2) = \mathbb{E}(X_1^2 + \cdots + X_m^2) = \mathbb{E}(X_1^2) + \cdots \mathbb{E}(X_m^2) < \infty \tag{3.21}$$

nach der Linearität des Erwartungswerts und da nach Definition 3.39 die Komponenten $\mathbf{X}_i$ von $\mathbf{X}$ $\mathcal{L}^2$-Zufallsgrößen sind. □

Def. 3.47. Eine Folge $(X_k)_k$ von Zufallsvariablen $X_k : \Omega \to E$, $E = \mathbb{R}, \mathbb{R}^m$ heißt *orthogonal*, wenn gilt

$$\mathbb{E}(X_k^T X_l) = 0 \text{ für } \; k \neq l. \tag{3.22}$$

Lemma 3.48. *Die Summe zweier $\mathcal{L}^2$-beschränkter Folgen von Zufallsgrößen ist wieder $\mathcal{L}^2$-beschränkt.*

Beweis. Für alle k gelte $\mathbb{E}(X_k^2) \leq c_1$, $\mathbb{E}(Y_k^2) \leq c_2$. Verwende Satz 3.38 für $p = 2$:

$$\Big(\mathbb{E}\big((X_k + Y_k)^2\big)\Big)^{1/2} \leq \Big(\underbrace{\mathbb{E}(X_k^2)}_{\leq c_1}\Big)^{1/2} + \Big(\underbrace{\mathbb{E}(Y_k^2)}_{\leq c_2}\Big)^{1/2}, \tag{3.23}$$

d.h., die Wurzel auf der linken Seite von (3.23) ist für alle k beschränkt und damit der Ausdruck $\mathbb{E}\big((X_k + Y_k)^2\big)$ selbst auch. □

Lemma 3.49. $\mathbf{E}_k$ *sei eine orthogonale Folge von $\mathcal{L}^2$-Zufallsgrößen, setze* $\mathbf{M}_{k_0;n} \coloneqq \sum_{k=k_0}^{n} \mathbf{E}_k$. *Dann gilt:*

$$\mathbb{E}\Big(\Big\|\mathbf{M}_{k_0;n}\Big\|^2\Big) = \sum_{k=k_0}^{n} \mathbb{E}\big(\,\|\mathbf{E}_k\|^2\,\big). \tag{3.24}$$

Beweis. (1) in der folgenden Rechnung wird durch die Orthogonalität der $\mathbf{E}_k$

gerechtfertigt. O.B.d.A. sei $k_0 = 1$.

$$\begin{aligned}
\mathbb{E}(\|\mathbf{M}_{1;n}\|^2) &= \mathbb{E}\left(\left\|\sum_{k=1}^{n}\mathbf{E}_i\right\|^2\right) = \mathbb{E}\left(\left\|\mathbf{E}_1 + \cdots \mathbf{E}_m\right\|\right) \\
&= \mathbb{E}\Big((\mathbf{E}_{11} + \cdots + \mathbf{E}_{n1})^2 + \cdots + (\mathbf{E}_{1m} + \cdots \mathbf{E}_{nm})^2\Big) \\
&= \mathbb{E}\Big(\mathbf{E}_{11}^2 + \cdots + \mathbf{E}_{n1}^2 + 2\sum_{k>l}\mathbf{E}_{k1}\,\mathbf{E}_{l1} \\
&\qquad \vdots \\
&\qquad + \mathbf{E}_{1m}^2 + \cdots + \mathbf{E}_{nm}^2 + 2\sum_{k>l}\mathbf{E}_{km}\,\mathbf{E}_{lm}\Big) \overset{(1)}{=} \mathbb{E}\Big(\sum_{k=1}^{n}\underbrace{\sum_{i=1}^{m}\mathbf{E}_{ki}^2}_{=\|\mathbf{E}_k\|^2}\Big).
\end{aligned}$$

Mit der Linearität des Erwartungswerts (3.24) folgt die Behauptung. □

Eindimensionale Orthogonalität lässt sich mehrdimensional erweitern:

Lemma 3.50. *Ist für jedes $i = 1, \ldots, m$ $\mathbf{Y}_{k\,i}$ eine orthogonale Folge, dann ist $\mathbf{Y}_k$ eine orthogonale Folge.*

Beweis. Verwende, dass für alle i gilt: $\mathbb{E}(\mathbf{Y}_{k\,i}\mathbf{Y}_{n\,i}) = 0$ $(k \neq n)$:

$$\mathbb{E}({\mathbf{Y}_k}^T\mathbf{Y}_n) = \mathbb{E}(\mathbf{Y}_{k\,1}\mathbf{Y}_{n\,1} + \mathbf{Y}_{k\,2}\mathbf{Y}_{n\,2} + \cdots + \mathbf{Y}_{k\,m}\mathbf{Y}_{n\,m}) = \sum_{i=1}^{m}\underbrace{\mathbb{E}(\mathbf{Y}_{k\,i}\mathbf{Y}_{n\,i})}_{=0} = 0.$$

□

3.6 Martingale

Nun folgt der letzte wahrscheinlichkeitstheoretische Abschnitt, der die Martingaltheorie vorstellt, insofern sie für diese Arbeit relevant ist. Insbesondere wird die Anwendung der Doobschen Martingalungleichung vorbereitet. Eine Definition für stochastische Prozesse aus [Kön09, Definition 10.1.1] wird angeführt:

Def. 3.51. Eine Folge $(X_n)_{n\geq 0}$ von $(E, \mathcal{E})$-wertigen Zufallsgrößen $X_n : \Omega \to E$ heißt E-wertiger *stochastischer Prozess* oder im Falle von $E = \mathbb{R}$ einfach stochastischer Prozess.

Die folgenden beiden Definitionen sind [Kön06, Definition 1.1.2 und Bemerkung 1.1.3(ii)] entnommen.

Def. 3.52 (Filtrierung)**.** Eine *Filtrierung* von $\mathcal{F}$ ist eine aufsteigende Folge $(\mathcal{F}_n)_n$ von Teil-σ-Algebren von $\mathcal{F}$, d.h. $\mathcal{F}_{n_1} \subseteq \mathcal{F}_{n_2}$ für alle $n_1 \leq n_2$.

Def. 3.53 ($\mathcal{F}_n$-Martingal)**.** Eine Folge $(M_n)_n$ integrierbarer Zufallsgrößen auf Ω, bei der jedes M_n $\mathcal{F}_n$-messbar ist, heißt *$\mathcal{F}_n$-Martingal* oder *Martingal bezüglich der Filtrierung $\mathcal{F}_n$*, falls

$$\mathbb{E}(M_{n+1} \,|\, \mathcal{F}_n) = M_n. \tag{3.25}$$

Ist $\mathcal{F}_n$ die kanonische Filtrierung $\mathcal{F}_n = \sigma(M_1, \ldots, M_n)$, so nennt man M_n einfach Martingal. Für ein vektorwertiges $\mathcal{F}_n$-Martingal ist jede Komponente ein $\mathcal{F}_n$-Martingal.

Bemerkung. Statt der hier geforderten Bedingung (3.25) für die Martingaleigenschaft wird auch $\mathbb{E}(M_{n_2} \,|\, \mathcal{F}_{n_1}) \doteq M_{n_1}$ für $n_2 \geq n_1$ in der Definition eines Martingals gefordert, was äquivalent ist, siehe [LR79, Satz 6.2.1] oder [Kön06, Bemerkung 1.1.3(ii)].

Lemma 3.54. *Ist M_n $\mathcal{F}_n$-messbar und von der Form $M_n = Y_k + \cdots + Y_n$ $(n = k, k+1, \dots)$, und gelte $\mathbb{E}[Y_{n+1} \,|\, \mathcal{F}_n] \doteq 0$, dann erfüllt M_n die Martingalbedingung* (3.25)*. Speziell lautet die Bedingung für die natürliche Filtrierung $\mathcal{F}_n^M$: $\mathbb{E}[Y_{n+1} \,|\, M_l, l \leq n] \doteq 0$.*

Beweis.

$$\mathbb{E}[M_{n+1} \,|\, \mathcal{F}_n] \doteq \mathbb{E}[M_n + Y_{n+1} \,|\, \mathcal{F}_n] \overset{\text{lin.}}{\doteq} \mathbb{E}[M_n \,|\, \mathcal{F}_n] + 0 \doteq \mathbb{E}[M_n \,|\, \mathcal{F}_n] \doteq M_n,$$

da M_n $\mathcal{F}_n$-messbar. □

Def. 3.55. Eine Folge von Zufallsgrößen X_n heißt eine *Martingaldifferenz*, falls es ein Martingal gibt, so dass $M_{n+1} = M_n + X_n$, $M_0 = 0$.

Das folgende Resultat ist aus [Rao84, Abschnitt 3.5, Satz 2] entnommen:

Lemma 3.56 (Orthogonalität von $\mathcal{L}^2$-Martingalen)**.**
Sei $S_n \in \mathcal{L}^1$, $n \geq 1$ und $\mathcal{F}_n$ eine Filtrierung.

(i) S_n ist genau dann ein Martingal, wenn es sich ausdrücken lässt als $S_n = \sum_{k=1}^n Y_k$ mit $\mathbb{E}(Y_{k+1} \,|\, \mathcal{F}_k) \doteq 0$, $k \geq 1$.

(ii) Wenn S_n ein $\mathcal{F}_n$-Martingal ist und $S_n \in \mathcal{L}^2$, dann bilden die Inkremente Y_k eine orthogonale Folge, $\mathbb{E}(Y_k Y_l) = 0$ $(k < l)$, dabei sei $Y_1 = S_1$.

Das folgende Resultat ist aus [Kön06, Lemma 1.1.5(iii)] entnommen:

Lemma 3.57. *Sei φ konvex, $\mathbb{E}(|\varphi(M_n)|) < \infty, n \in \mathbb{N}$ und M_n ein $\mathcal{F}_n$-Martingal. Dann ist $(\varphi(M_n))_n$ ein $\mathcal{F}_n$-Submartingal.*

Die Idee des folgenden Beweises geht dabei auf die Aussagen aus [Wil91] zurück, zunächst die Definition der Dichtefunktion nach [Wil91, Def. 6.12. und Def. 8.3].

Def. 3.58. Man sagt, dass X eine *Dichtefunktion* (Englisch: *probability density function*, pdf) f_X besitzt, wenn es eine Borel-Funktion $f_X : \mathbb{R} \to [0, \infty]$ gibt, so dass

$$\mathbb{P}(X \in B) = \int_B f_X(x)dx, \, \forall B \in \mathcal{B}. \tag{3.26}$$

f_X ist dabei bis auf Gleichheit fast überall definiert. Man sagt, dass X und Y eine gemeinsame Verteilung besitzen, falls es eine Borel-Funktion $f_{X,Y} : \mathbb{R}^2 \to [0, \infty] \times [0, \infty]$ gibt, so dass

$$\mathbb{P}\big((X, Y) \in B\big) = \iint \mathbb{1}_B(x, y) f_{X,Y}(x, y)dxdy \quad \forall B \in \mathcal{B} \times \mathcal{B}. \tag{3.27}$$

Das folgende Lemma ist nach [Wil91, Def. 9.6]:

Lemma 3.59. *Seien X, Z Zufallsgrößen mit einer* gemeinsamen Dichtefunktion *(Englisch: joint pdf) $f_{X,Z}(x, z)$. Dann ist*

- $f_Z(z) = \int_{\mathbb{R}} f_{X,Z}(x,z)dx$ *eine Dichtefunktion für* Z *und*
- $f_X(x) = \int_{\mathbb{R}} f_{X,Z}(x,z)dz$ *eine Dichtefunktion für* X.

Sei h *eine* Borel*-Funktion auf* $\mathbb{R}$ *und* $h(X) \in \mathcal{L}^1$. *Definiert man die elementare bedingte Dichtefunktion (*elementary conditonal pdf*) gemäß*

$$f_{X \mid Z}(x \mid z) \coloneqq \mathbb{1}_{f_Z(z)\neq 0} \frac{f_{X,Z}(x,z)}{f_Z(z)}, \tag{3.28}$$

sowie

$$g(z) \coloneqq \int_{\mathbb{R}} h(x) f_{X \mid Z}(x \mid z)dx, \tag{3.29}$$

dann ist $g(Z)$ *eine Version der bedingten Erwartung von* $h(X)$ *gegeben* Z, *d.h.*

$$g(Z) = \mathbb{E}[h(X) \mid Z]. \tag{3.30}$$

Das folgende Lemma ist nach [Wil91, Satz 9.10]:

Lemma 3.60. *Seien* Y, Z *unabhängige Zufallsgrößen. Wenn* h *eine beschränkte,* Borel*-messbare Funktion ist und man*

$$\gamma^h(y) \coloneqq \mathbb{E}\big(h(y,Z)\big) \tag{3.31}$$

setzt, dann ist $\gamma^h(Y)$ *eine Version der bedingten Erwartung* $\mathbb{E}[h(Y,Z) \mid Y]$.

Es bietet sich die folgende Erweiterung dieses Lemmas an:

Lemma 3.61. *Sei* X *eine Zufallsgröße und* $\mathcal{A}$ *eine* σ*-Algebra, so dass* $\sigma(X)$ *und* $\mathcal{A}$ *unabhängig sind.* Y *sei eine* $\mathcal{A}$*-messbare Zufallsgröße und* h *eine beschränkte,* Borel*-messbare Funktion. Definiert man* $\gamma^h(y) \coloneqq \mathbb{E}\big(h(y,Z)\big)$, *dann ist* $\gamma^h(Y)$ *eine Version des bedingten Erwartungswerts* $E\big[h(Y,Z) \mid \mathcal{A}\big]$.

Für $\mathcal{A} = \sigma(Y)$ enthält dies das vorangegange Lemma.

Beweis. Die Idee des Beweises ist nach [Msec] und baut die Beweisskizze aus [Wil91] aus, den Satz von Fubini zu verwenden. Die beiden Eigenschaften der Definition des bedingten Erwartungswerts $\mathbb{E}\big[h(Y,Z) \mid \mathcal{A}\big]$ sind für $\gamma^h(Y)$ zu prüfen.

- Zunächst ist nach dem Doob-Dynkin-Lemma $\gamma^h(Y)$ $\mathcal{A}$-messbar, da Y $\mathcal{A}$-messbar ist.
- Außerdem muss gezeigt werden, dass für alle $A \in \mathcal{A}$ gilt

  $$\mathbb{E}\big(h(Y,Z)\mathbb{1}_A\big) = \mathbb{E}\big(\gamma^h(Y)\mathbb{1}_A\big). \tag{3.32}$$

  Y und $\mathbb{1}_A$ sind $\mathcal{A}$-messbar und Z ist unabhängig von A. Sei μ die Verteilung von $(Y, \mathbb{1}_A)$ und ν die Verteilung von Z. Es gilt:

  $$\mathbb{E}\big(h(Y,Z)\mathbb{1}_A\big) = \iint h(y,z)\mathbb{1}_A(x)d\mu(y,x)d\nu(z). \tag{3.33}$$

  Wegen

  $$\gamma^h(y) = \int h(y,z)d\nu(z)$$

folgt für die rechte Seite von (3.32):

$$\begin{aligned}\mathbb{E}\big(\gamma^h(Y)\mathbb{1}_A\big) &= \int \gamma^h(y)\mathbb{1}_A(x)d\mu(y,x)\\ &= \int\left(\int h(y,z)d\nu(z)\right)\mathbb{1}_A(x)d\mu(y,x) = \iint h(y,z)\mathbb{1}_A(x)d\nu(z)d\mu(y,x).\end{aligned}$$

Nach dem Satz von Fubini ist dies wegen Gleichung (3.33) gleich der linken Seite von (3.32).

□

Bemerkung. Lemma 3.61 soll für den Fall verwendet werden, dass $h(Y,Z) \in \mathcal{L}^1$ und h Borel-messbar. In dieser Arbeit wird angenommen, dass es sich auf diesen Fall erweitern lässt.[2] Alternativ müsste man im folgenden Satz die Bedingung $u(x,y)$ sei beschränkt aufnehmen.
Da u die Form $(x,y) \mapsto \frac{f(x+y_2)-f(x-y_2)}{2h_k y_2} + \frac{y_1}{y_2}$ hat, gilt das höchstens auf einer Teilmenge.

Die Argumentation zur Behandlung eines Fehlerterms in Spalls Konvergenzsatz (Theorem 4.48) beruht darauf, dass man in der folgenden Situation ein Martingal vorliegen hat:

Lemma 3.62 (Martingalsatz)**.** *Seien X_0 und Z_k, $k \geq 0$ unabhängige Zufallsgrößen. Mit glatten Funktionen $u, v_k : E \times E \to E$ gelte*

$$\begin{aligned}X_{k+1} &= v_k(X_k, Z_k),\\ U_k &= u(X_k, Z_k) \text{ und}\\ Y_k &= U_k - \mathbb{E}[U_k \,|\, X_k].\end{aligned}$$

u sei Borel-messbar und U_k eine integrierbare Zufallsgröße. Für jedes $k \geq 0$ sei $\mathcal{G}_k$ eine σ-Algebra, sodass (X_k, Z_k) $\mathcal{G}_k$-messbar sind und Z_{k+1} unabhängig von $\mathcal{G}_k$ ist. Dann gilt

$$\mathbb{E}[Y_{n+1} \,|\, \mathcal{G}_n] \eqsim 0 \tag{3.34}$$

und damit ist nach Lemma 3.54 $M_n = Y_1 + \cdots + Y_n$ für jede Filtrierung $\mathcal{F}_n$, die die an $\mathcal{G}_n$ gestellten Bedingungen erfüllt, ein $\mathcal{F}_n$-Martingal.

Bemerkung. Eine mögliche Wahl für $\mathcal{F}_n$ wäre die Filtrierung
$\sigma(X_0) \vee \sigma(Z_0, \ldots, Z_n)$. Dabei ist X_n nach dem Doob-Dynkin-Lemma $\mathcal{F}_n$-messbar vermöge

$$X_{n+1} = v_n(X_n, Z_n) = v_n(v_{n-1}(X_{n-1}, Z_{n-1}), Z_n) = \ldots = \breve{v}(X_0; Z_0, \ldots, Z_n). \tag{3.35}$$

Da X_0 und $Z_k, k \geq 0$ unabhängig sind, sind also auch Z_{n+1} und $X_0, Z_0, Z_1, \ldots, Z_n$ unabhängig.

Beweis der Bemerkung. Es wird gezeigt, dass Z_{n+1} unabhängig von $\mathcal{G}_n$ für die Wahl $\mathcal{G}_n = \mathcal{F}_n$ erfüllt ist. Dazu verwendet man Lemma 3.20 und setzt

$$I_1 = \{0, \ldots, n+1\},\ I_2 = \{n+2\},$$

[2] Im Beweis in [Wil91] wird die Aussage im Wesentlichen auf den Satz von Fubini zurückgeführt und es scheint so, dass die Beschränktheitsforderung die Regularität sichern soll, und diese auch mit $h(Y,Z) \in \mathcal{L}^1$ hinreichend gegeben ist.

sowie

$$Y_1 = (X_0, Z_0, Z_1, \ldots, Z_n) \text{ und}$$
$$Y_2 = Z_{n+1}.$$

Es wird angenommen, dass die Voraussetzung, dass die im Lemma erwähnten Funktionen f_1 und f_2 „geeignet“ seien, für den hier benutzten Fall identischer Funktionen erfüllt ist.

Man beachte noch, dass $\sigma(Y_1) = \mathcal{F}_n$ nach Lemma 3.19. □

Im Rahmen von Spalls SPSA-Konvergenzanalyse stellt dies den Punkt dar, bei dem Spall behauptet, dass $\sum_{k=k_0}^{n} a_k \mathbf{E}_k$ ein Martingal ist. Mit den von Spall in [Spa92; Spa05] beschriebenen Filtrierungen konnte ich die Messbarkeitsvoraussetzung in Definition 3.53 nicht nachvollziehen wie in Bemerkung 4.20 beschrieben wird. Die Idee für die hier verwendete Filtrierung und der hier angegebene Martingalsatz geht auf [Mseh] und [Msef] zurück.

Beweis von Lemma 3.62. $\mathbb{E}[U_k \,|\, X_k]$ ist wohldefiniert, da $U_k = u(X_k, Z_k)$ eine $\mathcal{L}^1$-Zufallsgröße ist. Definiere die Funktion $w_n : E \to E, x \mapsto \mathbb{E}\big(u(x, Z_n)\big)$.
$X_{n+1} = \breve{v}(X_0; Z_0, \ldots, Z_n)$ ist $\mathcal{G}_n$-messbar nach Lemma 3.10 (Doob-Dynkin-Lemma). Z_{n+1} ist unabhängig von $\mathcal{G}_n$. X_{n+1} ist $\mathcal{G}_n$-messbar und daher eine σ-Subalgebra von $\mathcal{G}_n$ nach Hilfssatz 3.26, also sind Z_{n+1} und X_{n+1} unabhängig nach Hilfssatz 3.23.

Daher kann man nun Lemma 3.61 nach einer Idee von [Msec] für $h = u$, d.h. $\gamma^h = w_{n+1}$, verwenden. Das erste Mal für $\mathcal{A} = \sigma(X_{n+1})$, ein zweites Mal für $\mathcal{A} = \mathcal{G}_n$. Daraus folgt:

$$\mathbb{E}\big(u(X_{n+1}, Z_{n+1}) \,|\, X_{n+1}\big) = w_{n+1}(X_{n+1}) = \mathbb{E}\big(u(X_{n+1}, Z_{n+1}) \,|\, \mathcal{G}_n\big). \tag{3.36}$$

Nun kann man schließen:

$$\begin{aligned}
\mathbb{E}[Y_{n+1} \,|\, \mathcal{G}_n] &\eqsim \mathbb{E}[U_{n+1} - \mathbb{E}(U_{n+1} \,|\, X_{n+1}) \,|\, \mathcal{G}_n] \\
&\eqsim \mathbb{E}[U_{n+1} \,|\, \mathcal{G}_n] - \mathbb{E}[\mathbb{E}(U_{n+1} \,|\, X_{n+1}) \,|\, \mathcal{G}_n] \\
&\eqsim \mathbb{E}[U_{n+1} \,|\, \mathcal{G}_n] - \mathbb{E}[U_{n+1} \,|\, X_{n+1}] \eqsim 0.
\end{aligned}$$

Für die Zusatzaussage ist noch zu zeigen, dass M_n eine $\mathcal{G}_n$-messbare $\mathcal{L}^1$-Zufallsgröße ist. Da aus $U_n \in \mathcal{L}^1$ folgt, dass $Y_n \in \mathcal{L}^1$, gilt also $M_n \in \mathcal{L}^1$. Die geforderte Messbarkeitseigenschaft von M_n gilt, wenn sie für Y_n gilt. $\mathbb{E}[U_n \,|\, X_n]$ ist als bedingter Erwartungswert auf X_n X_n-messbar, also insbesondere $\mathcal{G}_n$-messbar (Lemma 3.11). $U_n = u(X_n, Z_n)$ ist $\mathcal{G}_n$-messbar, da (X_n, Z_n) $\mathcal{G}_n$-messbar ist (Doob-Dynkin-Lemma). □

Es gilt die folgende Variante der Doobschen Martingal-Ungleichung nach [Kön06, Satz 1.7.2(i)] (in [GS06, (7.8.2)] auch als Doob-Kolmogorov-Ungleichung bezeichnet):

Lemma 3.63 (Doobsche Martingal-Ungleichung nach [Kön06])**.**
Sei S_n ein positives Submartingal. Dann gilt für alle $\eta > 0$

$$\mathbb{P}(|S_n|^* \geq \eta) \equiv \mathbb{P}(\max_{k=1,\ldots,n} |S_k| \geq \eta) \leq \frac{1}{\eta^2} \mathbb{E}(|S_n|^2) \tag{3.37}$$

mit $|S_n|^* = \max_{k=1,\ldots,n} |S_k|$.

Es gilt die folgende Variante der Doobschen Martingalungleichung:

Korollar 3.64. *Sei S_n ein positives Submartingal. Dann gilt für alle $\eta > 0$*

$$\mathbb{P}(\sup_{n\geq 0} |S_n| \geq \eta) \leq \frac{1}{\eta^2} \lim_{n\to\infty} \mathbb{E}(|S_n|^2), \tag{3.38}$$

wobei der Grenzwert auf der rechten Seite eventuell ∞ ist.

Bemerkung. In dieser Arbeit wird angenommen, dass sich dieses Korollar auf den Fall nichtnegativer Submartingale erweitern lässt.

Beweis. In der Form (3.37) der Doobschen Martingalungleichung geht man beidseitig zum $\lim_{n\to\infty}$ über:

$$\lim_{n\to\infty} \mathbb{P}(\max_{k=1,\dots,n} |S_n| \geq \eta) \leq \frac{1}{\eta^2} \lim_{n\to\infty} \mathbb{E}(|S_n|^2).$$

Zu zeigen bleibt, dass

$$\lim_{n\to\infty} \mathbb{P}(\max_{k=1,\dots,n} |S_k| \geq \eta) = \mathbb{P}(\sup_{n\geq 0} |S_n| \geq \eta), \text{ also dass}$$
$$\lim_{n\to\infty} \mathbb{E}(\mathbb{1}_{A_n}) = \mathbb{E}(\mathbb{1}_A)$$

mit

$$A_n = \{\omega \in \Omega : \max_{k=0,\dots,n} |S_k| \geq \eta\} \text{ und}$$
$$A = \{\omega \in \Omega : \sup_{n\geq 0} |S_k| \geq \eta\}.$$

Das Konvergenzresultat für Indikatorfunktionen von Mengenfolgen nach Lemma 3.9 soll verwendet werden. $A_n \subseteq A_{n+1}$, denn $\max_{k=0,\dots,n} |s_k| \leq \max_{k=0,\dots,n+1} |s_k|$. Es gilt $A = \bigcup A_k$, denn:

„$\subseteq$“ Sei $\omega \in A$, d.h. mit $s_k := S_k(\omega)$, $\sup_{n\geq 0} |s_k| \geq \eta$, d.h., dass ein $k_0 \geq 0$ existiert, so dass $|s_{k_0}| \geq \eta$, also $\omega \in A_{k_0} \subseteq \bigcup_{k=0}^{\infty} A_n$.

„$\supseteq$“ Sei $\omega \in \bigcup_{k=0}^{\infty} A_k$, d.h. mit $s_k := S_k(\omega)$, dass ein $k_0 \geq 0$ existiert, so dass $\max_{k\leq k_0} |s_k| \geq \eta$. Wegen $\sup_{k\geq 0} |s_k| \geq \max_{k\leq k_0} |s_k| \geq \eta$ ist $\omega \in A$.

Also ist $0 \leq \mathbb{1}_{A_n} \to \mathbb{1}_A, \mathbb{1}_{A_n} \leq \mathbb{1}_A$, woraus nach dem monotonen Konvergenzsatz $\lim_{n\to\infty} \mathbb{E}(\mathbb{1}_{A_n}) = \mathbb{E}(\mathbb{1}_A)$ folgt, was den Beweis abschließt. □

Bemerkung. Diese Doobsche Ungleichung entspricht der, auf die sich auch in [KC78, S. 27] berufen wird und auf die Spall in [Spa92, S. 335] verweist.

Bemerkung 3.65. Setzt man $\widetilde{S}_n = S_{n+k_0}$, so gilt

$$\mathbb{P}\left(\sup_{n\geq 0} \left|\widetilde{S}_n\right| \geq \eta\right) \leq \frac{1}{\eta^2} \lim_{n\to\infty} \mathbb{E}\left(\left|\widetilde{S}_n\right|^2\right),$$

d.h. wegen $\sup_{n\geq 0} \left|\widetilde{S}_n\right| = \sup_{n\geq 0} |S_{n+k_0}| = \sup_{n\geq k_0} |S_n|$ und $\lim_{n\to\infty} x_{n+k_0} = \lim_{n\to\infty} x_n$:

$$\mathbb{P}(\sup_{n\geq k_0} |S_n| \geq \eta) \leq \frac{1}{\eta^2} \lim_{n\to\infty} \mathbb{E}(|S_n|^2). \tag{3.39}$$

Kapitel 4

Analyse der numerischen Verfahren

4.1 Optimierungsverfahren

Der Ausdruck $\mathcal{C}(\mathbb{R}^{n_1}, \mathbb{R}^{n_2})$ bezeichne den Raum der stetigen Funktionen $f : \mathbb{R}^{n_1} \to \mathbb{R}^{n_2}$ und $\mathcal{C}^l(\mathbb{R}^{n_1}, \mathbb{R}^{n_2})$ den der l-mal stetig differenzierbaren Funktionen. Als Vektornorm wird die 2-Norm $\|\mathbf{x}\| = \sqrt{\mathbf{x}^T\mathbf{x}}$ verwendet.

Die Behandlung eines Minimierungsproblems, obwohl man in der Anwendung ein Maximierungsproblem betrachtet, wird dadurch gerechtfertigt, dass $\min_{\mathbf{x}\in\mathbb{X}} f(\mathbf{x})$ und $\max_{\mathbf{x}\in\mathbb{X}} -f(\mathbf{x})$ dieselbe Lösungsmenge besitzen:

Lemma 4.1. *Sei $f \in \mathcal{C}(\mathbb{R}^m, \mathbb{R})$, $\mathbb{X} \subseteq \mathbb{R}^m$ nichtleer. Dann gilt:*

$$\arg\min_{\mathbf{x}\in\mathbb{X}} f(\mathbf{x}) = \arg\max_{\mathbf{x}\in\mathbb{X}} \big(-f(\mathbf{x})\big),$$

d.h., $\min_{\mathbf{x}\in\mathbb{X}} f(\mathbf{x})$ besitzt dieselbe Lösungsmenge wie $\max_{\mathbf{x}\in\mathbb{X}} -f(\mathbf{x})$.

Beweis. Sei $\mathbf{x}^*$ eine Lösung von $\min_{\mathbf{x}\in\mathbb{X}} f(\mathbf{x})$, d.h., $f(\mathbf{x}^*) \leq f(\mathbf{x})$ für alle $\mathbf{x} \in \mathbb{X}$
$\overset{\cdot(-1)}{\Longleftrightarrow}$ $-f(\mathbf{x}^*) \geq -f(\mathbf{x})$ für alle $\mathbf{x} \in \mathbb{X}$, d.h. $\mathbf{x}^*$ löst $\max_{\mathbf{x}\in\mathbb{X}} -f(\mathbf{x})$. □

Def. 4.2 (Stationäre Punkte)**.** Ist $\mathbf{x}^*$ eine lokale Minimalstelle und $f : \mathbb{R}^m \to \mathbb{R}$ stetig differenzierbar in einer offenen Umgebung von $\mathbf{x}^*$, dann ist $\nabla f(\mathbf{x}^*) = 0$. Punkte mit $\nabla f(\mathbf{x}) = 0$ heißen *stationäre Punkte* von f.

Eine Strategie zur Entwicklung numerischer Optimierungsverfahren basiert darauf, solche stationären Punkte zu finden.

Im nächsten Abschnitt werden (deterministische) Abstiegsverfahren behandelt, insbesondere das Verfahren des steilsten Abstiegs, bevor im darauffolgenden Abschnitt stochastische Gradientenverfahren untersucht werden.

4.2 Abstiegsverfahren und das Verfahren des steilsten Abstiegs

Als Ausgangspunkt für die Behandlung der stochastischen Gradientenverfahren werden die deterministischen Gradientenverfahren vorgestellt.

4.2.1 Definition der Verfahren

Der Abschnitt zu den Abstiegsverfahren ist im Wesentlichen NOCEDAL und WRIGHT: *Numerical Optimization*, Kapitel 2 [NW06] entnommen.

Def. 4.3 (Abstiegsrichtung)**.** Sei $\mathbf{p} \in \mathbb{R}^m \setminus \{0\}$ und θ der Winkel zwischen $\mathbf{p}$ und $-\nabla f(\mathbf{x})$. Dann heißt $\mathbf{p}$ *Abstiegsrichtung* von f im Punkt $\mathbf{x}$, wenn θ kleiner als $\frac{\pi}{2}$ ist. Für den Winkel γ zwischen $\mathbf{p}$ und $\nabla f(\mathbf{x})$ gilt: $\cos\gamma = \frac{\nabla f(\mathbf{x})^T \mathbf{p}}{\|\nabla f(\mathbf{x})\|\|\mathbf{p}\|}$, $\theta = \pi - \gamma$.

Die Wortwahl erklärt sich aus dem Folgenden: Sei $\nabla f(\mathbf{x}) \neq 0$. Dann gilt nach dem TAYLORschem Lehrsatz

$$f(\mathbf{x} + \varepsilon\mathbf{p}) = f(\mathbf{x}) + \varepsilon \cdot \nabla f(\mathbf{x})^T \mathbf{p} + O(\varepsilon^2)$$

und weil $\mathbf{p}$ eine Abstiegsrichtung ist, gilt $\nabla f(\mathbf{x})^T \mathbf{p} = \|\mathbf{p}\| \, \|\nabla f(\mathbf{x})\| \cos\gamma < 0$, da $\gamma > \frac{\pi}{2}$. Also folgt: $f(\mathbf{x}) > f(\mathbf{x} + \varepsilon\mathbf{p}) \quad \forall \varepsilon > 0$ hinreichend klein.

Bemerkung 4.4. Die auf 1 normierte Richtung des *steilsten Abstiegs* ist die Richtung, in der f in Richtung $\mathbf{p}$ am stärksten abnimmt, also wo der Koeffizient von ε am kleinsten ist, d.h. $\mathbf{p}$ ist die Lösung des Minimierungsproblems

$$\nabla f(\mathbf{x})^T \mathbf{p} \to \min_{\mathbf{p} \in \mathbb{R}^m}! \quad \text{s.t. } \|\mathbf{p}\| = 1. \tag{4.1}$$

$\nabla f(\mathbf{x})^T \mathbf{p} = \|\mathbf{p}\| \, \|\nabla f(\mathbf{x})\| \cos\gamma = \|\nabla f(\mathbf{x})\| \cos\gamma$ ist minimal für $\cos\gamma = -1$ und das Minimum wird von $\mathbf{p} = -\frac{\nabla f(\mathbf{x})}{\|\nabla f(\mathbf{x})\|}$ angenommen. Diese Richtung ist orthogonal zu den Höhenlinien der Funktion f [NW06, S.21].

Damit ist die Definition 2.2 (Abstiegsverfahren) aus Kapitel 2 vollständig untermauert.

Bemerkung 4.5. Die Richtung des steilsten Abstiegs ist nicht immer am effektivsten zur Lösung eines gegebenen Problems. Für schlecht skalierte Probleme (gleiche Änderungen in den Komponenten von $\mathbf{x}$ verursachen sehr unterschiedliche Änderungen der Zielfunktion) zum Beispiel sollte man Verfahren 2. Ordnung, also NEWTON- und Quasi-NEWTON-Verfahren verwenden (vergleiche [NW06, ab S.44]). Auch für diese könnte man stochastische Alternativen entwickeln. In Hinblick auf die Echtzeitbedingungen beschränkt sich diese Arbeit aber auf stochastische Gradientenverfahren, die auf dem Steilsten-Abstieg-Verfahren (2.5) beruhen, da ableitungsfreie Verfahren 2. Ordnung mehr Funktionsauswertungen pro Iteration benötigen.

4.2.2 Zusammenhang mit der Differentialgleichung

Der folgende Abschnitt basiert auf [Spa05, S. 109]. Das Verfahren des steilsten Abstiegs kann man mit der Differentialgleichung

$$\dot{\mathbf{x}} = -\mathbf{g}(\mathbf{x}) \tag{4.2}$$

mit Anfangswert $\mathbf{x}(0) = \mathbf{x}_0$ in Verbindung bringen:

Die Iterationsvorschrift (2.5) lässt sich in der folgenden Weise schreiben:

$$\frac{\mathbf{x}_{k+1} - \mathbf{x}_k}{a_k} = -\mathbf{g}(\mathbf{x}_k). \tag{4.3}$$

Fasst man dann a_k als Zeitinkrement auf, in dem man $a_k = t_{k+1} - t_k$, d.h. $t_{k+1} = \sum_{k'=0}^{k} a_{k'}$, setzt und $\mathbf{x}$ eine Funktion sei, für die $\mathbf{x}(t_k) = \mathbf{x}_k$ gilt, so erhält (4.3) die Form

$$\frac{\mathbf{x}(t_{k+1}) - \mathbf{x}(t_k)}{t_{k+1} - t_k} = -\mathbf{g}\big(x(t_k)\big). \tag{4.4}$$

Gilt nun $a_k \to 0$, so kann man diese Differenzengleichung als Näherung an die DGL (4.2) auffassen.

Ein Optimum $\mathbf{x}^*$ von f ist (bei unrestringierter Optimierung, d.h. $\mathbb{X} = \mathbb{R}^m$) insbesondere ein stationärer Punkt, d.h., es gilt $\nabla f(\mathbf{x}^*) = \mathbf{0}$, und damit ist $\mathbf{x}^*$ ein Gleichgewichtspunkt[1] der DGL, es gilt:

$$\dot{\mathbf{x}} = -\nabla f\big(\underbrace{\mathbf{x}(t)}_{=\mathbf{x}^*}\big) = \mathbf{0}.$$

Dies ist dann auch ein Fixpunkt der Iteration ($\mathbf{x}_{k+1} = \mathbf{x}_k - a_k \cdot \mathbf{0}$).

4.2.3 Liniensuche und Schrittweitenwahl

Dieser Abschnitt beruht auf [NW06, Kapitel 3].
Nach der Festlegung einer Richtung stellt sich die Frage nach der zu verwendenden Schrittweite a_k. Diese bestimmt man durch eine Liniensuche, optimal wäre ein Minimum von

$$\Phi(a_k) \coloneqq f(\mathbf{x}_k + a_k \mathbf{p}_k), \quad a_k > 0. \tag{4.5}$$

Da eine exakte Lösung dieses 1-dimensionalen Minimierungsproblems aber oft zu viel Aufwand erfordert, hilft man sich durch Verwendung bestimmter Liniensuchbedingungen, z.B. den Wolfe-Bedingungen.

Im Folgenden werden die Bezeichnungen $f_k \coloneqq f(\mathbf{x}_k)$ und $\nabla f_k \coloneqq \nabla f(\mathbf{x}_k)$ verwendet.

Def. 4.6. Man fasst die folgenden beiden Bedingungen unter dem Namen Wolfe-*Bedingungen* zusammen [NW06, S. 34].

- Armijo-Bedingung:

$$\Phi(a_k) \leq f(\mathbf{x}_k) + c_1 a_k \, \nabla {f_k}^T \mathbf{p}_k, \quad c_1 \in (0,1), \tag{4.6a}$$

  d.h., der Funktionswert sollte hinreichend sinken.

[1] Eine Definition folgt in Abschnitt 4.3.4.

- Krümmungsbedingung (englisch *curvature condition*):

$$\Phi'(a_k) \equiv \nabla f(\mathbf{x}_k + a_k \mathbf{p}_k)^T \mathbf{p}_k \geq c_2 \ \nabla f(\mathbf{x}_k)^T \mathbf{p}_k \tag{4.6b}$$

  mit $c_2 \in (c_1, 1)$. Damit wird verhindert, dass zu kleine Werte für a_k akzeptiert werden.

Der folgende Satz zeigt, dass für Abstiegsrichtungen Schrittweiten existieren, für die die WOLFE-Bedingungen erfüllt sind. Er ist aus [NW06] entnommen. Hinreichend ist etwa, dass f auf $\mathbb{X}$ nach unten beschränkt ist.

Satz 4.7. *Sei $f \in \mathcal{C}^1(\mathbb{R}^m, \mathbb{R})$ und $\mathbf{p}_k$ eine Abstiegsrichtung in $\mathbf{x}_k$. f sei nach unten beschränkt entlang der Geraden $\{\mathbf{x}_k + a\mathbf{p}_k \,|\, a > 0\}$ und sei $0 < c_1 < c_2 < 1$.*
Dann existiert ein nichtleeres Intervall von Schrittlängen, die die WOLFE*-Bedingungen* (4.6) *erfüllen.*

Beweis. Der Beweis folgt [NW06]. $l(a) = f(\mathbf{x}_k) + ac_1 \nabla {f_k}^T \mathbf{p}_k$, $c_1 \in [0, 1]$ ist nach unten unbeschränkt, $\Phi(a) = f(\mathbf{x}_k + a\mathbf{p}_k)$ hingegen nach unten beschränkt. Es gibt also einen Schnittpunkt. Sei a' der kleinste Schnittpunkt, d.h.

$$f(\mathbf{x}_k + a'\mathbf{p}_k) = f(\mathbf{x}_k) + a'c_1 \nabla {f_k}^T \mathbf{p}_k. \tag{4.7}$$

Wegen $f(\mathbf{x}_k + a\mathbf{p}_k) = f(\mathbf{x}_k) + a\nabla {f_k}^T \mathbf{p}_k + O(a^2)$, $a \to 0$ ist für hinreichend kleine $a \leq a'$: $\Phi(a) \leq l(a)$ nach der Definition von a', also für alle $a \leq a'$. Das heißt, die Armijo-Bedingung 4.6a ist erfüllt für alle $a \leq a'$. Man wendet nun den Mittelwertsatz in der Form (A.1) an: Es existiert ein $a'' \in (0, a')$, so dass

$$f(\mathbf{x}_k + a'\mathbf{p}_k) = f(\mathbf{x}_k) + a'\nabla f(\mathbf{x}_k + a''\mathbf{p}_k)^T \mathbf{p}_k. \tag{4.8}$$

Mit (4.7) und (4.8) zusammen folgt dann:

$$\nabla f(\mathbf{x}_k + a''\mathbf{p}_k)\mathbf{p}_k = c_1 \nabla {f_k}^T \mathbf{p}_k > c_2 \nabla {f_k}^T \mathbf{p}_k,$$

d.h. a'' erfüllt beide WOLFE-Bedingungen. Da beide Ungleichungen sogar strikt gelten, gibt es wegen der Stetigkeit von f das behauptete Intervall. □

Andere übliche Voraussetzungen sind die GOLDSTEIN-Bedingungen, siehe dazu [NW06, S. 36].

Der folgende Satz ist aus [NW06, Abschnitt 3.2] entnommen.

Satz 4.8 (ZOUTENDIJK)**.** *Betrachte ein Abstiegsverfahren* (2.4)*, bei dem die Schrittweitenfolge a_k den* WOLFE*-Bedingungen* (4.6) *genügt. Nehme an, dass f nach unten beschränkt ist und stetig differenzierbar auf einer offenen Teilmenge $\mathbb{X}^0$, die die Niveaumenge $\mathcal{L} \coloneqq \{\mathbf{x} \in \mathbb{R}^m : f(\mathbf{x}) \leq f(\mathbf{x}_0)\}$ enthält ($\mathbf{x}_0$ sei der Startpunkt der Iteration). Nimm an, dass ∇f* LIPSCHITZ*-stetig ist mit* LIPSCHITZ*-Konstante L. θ_k sei der Winkel zwischen $\mathbf{p}_k$ und der Richtung des steilsten Abstiegs $-\nabla f_k$, d.h. $\cos\theta_k = \frac{-\nabla f_k^T \mathbf{p}_k}{\|\nabla f_k\| \|\mathbf{p}_k\|}$.*[2]
Dann gilt ZOUTENDIJK*'s Bedingung:*

$$\sum_{k \geq 0} \cos^2(\theta_k) \left\|\nabla f_k\right\|^2 < \infty. \tag{4.9}$$

[2] In [NW06] ist θ_k als der Winkel zwischen $\mathbf{p}_k$ und $-\nabla f_k$ definiert [NW06, S. 37], aber θ im vorhergehenden Abschnitt als Winkel zwischen $\mathbf{p}$ und ∇f [NW06, S. 21]. Letzterer wird in dieser Arbeit mit γ bezeichnet.

Beweis. Es wird dem Beweis in [NW06] gefolgt. Wegen (4.6b) und der Iterationsvorschrift (2.4) gilt:

$$(\nabla f_{k+1} - \nabla f_k)^T \mathbf{p}_k \geq (c_2 - 1)\nabla {f_k}^T \mathbf{p}_k.$$

Wegen der LIPSCHITZ-Stetigkeit gilt:

$$(\nabla f_{k+1} - \nabla f_k)^T \mathbf{p}_k \leq L(a_k \mathbf{p}_k)^T \mathbf{p}_k = a_k L \left\|\mathbf{p}_k\right\|^2.$$

Zusammmen ergibt sich:

$$a_k \geq \frac{(c_2 - 1)\nabla {f_k}^T \mathbf{p}_k}{L \left\|\mathbf{p}_k\right\|^2}.$$

Das Einsetzen in die ARMIJO-Bedingung (4.6a) ergibt:

$$\begin{aligned} f(\mathbf{x}_k + a_k \mathbf{p}_k) &\leq f(\mathbf{x}_k) + c_1 a_k \nabla {f_k}^T \mathbf{p}_k, \text{ d.h.} \\ f_{k+1} &\leq f_k + \frac{c_2 - 1}{L} c_1 \frac{(\nabla {f_k}^T \mathbf{p}_k)^2}{\left\|\mathbf{p}_k\right\|^2} \qquad \leq f_k - c \cos^2 \theta_k \left\|\nabla f_k\right\|^2 \end{aligned}$$

mit $c = c_1 \frac{1-c_2}{L}$. Durch Aufsummieren von $k' = 0$ bis k erhält man:

$$f_1 + \cdots + f_{k+1} \leq f_0 + \cdots + f_k - c \sum_{k'=0}^{k} \cos^2 \theta_{k'} \left\|\nabla f_{k'}\right\|^2, \text{ also}$$

$$f_{k+1} \leq f_0 - c \sum_{k'=0}^{k} \cos^2 \theta_{k'} \left\|\nabla f_{k'}\right\|^2.$$

Da f nach unten beschränkt ist, gibt es eine positive Konstante c_3 mit

$$c \sum_{k'=0}^{k} \cos^2 \theta_{k'} \left\|\nabla f_{k'}\right\|^2 \leq f_0 - f_{k+1} \leq c_3 \ \forall k,$$

so dass nach Grenzwertbildung $k \to \infty$ folgt

$$\sum_{k=0}^{\infty} \cos^2 \theta_k \left\|\nabla f_k\right\|^2 \leq \infty.$$

□

Korollar 4.9. *Die Methode des steilsten Abstiegs* (2.5) *konvergiert unter den Voraussetzungen von Satz 4.8 gegen einen stationären Punkt von f.*

Beweis. Aus (4.9) folgt: $\cos^2(\theta_k) \left\|\nabla f_k\right\|^2 \to 0$. Es gilt $\cos(\theta_k) = 1$ und daher $\lim_{k\to\infty} \left\|\nabla f_k\right\| = 0$. □

Es muss aber nicht unbedingt die Richtung des steilsten Abstiegs sein, es reicht, wenn die Abstiegsrichtung $\mathbf{p}_k$ „nah genug“ an der Richtung des steilsten Abstiegs $-\nabla f$ ist, also einen Winkel von weniger als 90° einnimmt, genauer, dass $|\theta_k| \leq \theta_{\max} < \frac{\pi}{2}$. Dies entspricht im folgenden Satz der Forderung $\cos(\theta_k) \geq \delta > 0$.

Korollar 4.10. *Sei ein Abstiegsverfahren der Form* (2.4) *gegeben, für das* $\cos(\theta_k) \geq \delta > 0$ *gesichert ist. Dann konvergiert dieses unter den Voraussetzungen von Satz 4.8 gegen einen stationären Punkt von* f*.*

Beweis. Analog zum vorherigen Beweis folgt aus der ZOUTENDIJK-Bedingung (4.9), dass $\lim_{k\to\infty} \cos^2(\theta_k) \left\|\nabla f_k\right\|^2 = 0$. Da $\cos^2(\theta_k) \geq \delta^2 > 0 \ \forall k$, folgt $\lim_{k\to\infty} \left\|\nabla f_k\right\|^2 = 0$. □

Das Verfahren ist also robust gegenüber Abweichungen von der Richtung des steilsten Abstiegs.

4.2.4 Ableitungsfreie Gradientenverfahren

Bis hierhin wurden Gradientenverfahren nur unter der Prämisse behandelt, dass der Gradient ∇f direkt verfügbar ist. Ist er dies nicht, wie in dem beschriebenen Anwendungsfall der adaptiven Optik, so muss man $\nabla f(\mathbf{x})$ geeignet nähern und dabei nur Funktionsauswertungen $f(\mathbf{x})$ nutzen, also ableitungsfrei arbeiten. Dazu lassen sich komponentenweise Differenzenquotienten nutzen.

Die Finite-Differenzen-Näherung durch einseitige Differenzenquotienten

$$\hat{\mathbf{g}}_h^{\mathrm{GN1}}(\mathbf{x}) = \Big(\frac{f(\mathbf{x}+h\mathbf{e}_i) - f(\mathbf{x})}{h}\Big)_{i=1,\dots,m}$$

wurde schon in Gleichung (2.6) angegeben. Eine weitere Möglichkeit besteht in der Nutzung von symmetrischen, zweiseitigen Differenzenquotienten

$$\hat{\mathbf{g}}_h^{\mathrm{GN2}}(\mathbf{x}) = \Big(\frac{f(\mathbf{x}+h\mathbf{e}_i) - f(\mathbf{x}-h\mathbf{e}_i)}{2h}\Big)_{i=1,\dots,m}. \tag{4.10}$$

Natürlich wären auch Mischformen denkbar sowie die Variante $\hat{\mathbf{g}}_h(\mathbf{x}) = \Big(\frac{f(\mathbf{x}) - f(\mathbf{x}-h\mathbf{e}_i)}{h}\Big)$. In dieser Arbeit wird die Untersuchung auf die Gradientennäherungen $\hat{\mathbf{g}}_h^{\mathrm{GN1}}$ und $\hat{\mathbf{g}}_h^{\mathrm{GN2}}$ bezogen.

Das folgende Lemma basiert auf [NW06, S. 196] und begründet die Bezeichnung GN1 und GN2.

Lemma 4.11 (Approximationsordnung der Gradientennäherungen)**.**
Sei f *eine dreimal stetig differenzierbare Funktion. Die Gradientennäherungen* $\hat{\mathbf{g}}^{GN1}$ *und* $\hat{\mathbf{g}}^{GN2}$ *des GSD-Verfahrens sind im folgenden Sinn Approximationen erster und zweiter Ordnung an den Gradienten* $\nabla f(\mathbf{x})$*:*

$$\hat{\mathbf{g}}_h^{GN1}(\mathbf{x}) = \nabla f(\mathbf{x}) + O(h), \ h \to 0 \tag{4.11}$$

$$\hat{\mathbf{g}}_h^{GN2}(\mathbf{x}) = \nabla f(\mathbf{x}) + O(h^2), \ h \to 0. \tag{4.12}$$

Für die Aussage über GN1 seien die zweiten Ableitungen f_i'' *beschränkt, für die Aussage über GN2 die dritten Ableitungen* f_i'''*.*

Beweis. Mit den Bezeichnungen $f_i'' = \partial_i^2 f$ und $f_i''' = \partial_i^3 f$, $\partial_i = \frac{\partial}{\partial \mathbf{x}_i}$ und $\mathbf{g}_i$ der i-ten Komponente des Gradienten $\mathbf{g} = \nabla f$ gilt nach dem TAYLORschen Lehrsatz (Satz A.5):

$$\begin{aligned} f(\mathbf{x}+h\mathbf{e}_i) &= f(\mathbf{x}) + h\mathbf{g}_i(\mathbf{x}) + \tfrac{1}{2}h^2 f_i''(\mathbf{x}) + \tfrac{1}{6}h^3 f_i'''(\mathbf{x}^{i+}) \text{ und} \\ f(\mathbf{x}-h\mathbf{e}_i) &= f(\mathbf{x}) - h\mathbf{g}_i(\mathbf{x}) + \tfrac{1}{2}h^2 f_i''(\mathbf{x}) - \tfrac{1}{6}h^3 f_i'''(\mathbf{x}^{i-}), \end{aligned} \tag{4.13}$$

wobei $\mathbf{x}^{i\pm}$ auf der Verbindungsstrecke von $\mathbf{x}$ und $\mathbf{x} \pm h\mathbf{e}_i$ liegen. Dabei wurde $\nabla f(\mathbf{x})^T(h\mathbf{e}_i) = h\mathbf{g}_i(\mathbf{x})$ und $(h\mathbf{e}_i)^T\nabla^2 f(\mathbf{x})(h\mathbf{e}_i) = h^2 f_i''(\mathbf{x})$ benutzt. Unter Verwendung der Bezeichnungen aus Satz A.5 gibt es in der Multiindex-Summe nur einen Beitrag, von dem $\mathbf{j}$ mit $\mathbf{j}_i = 3$ und $\mathbf{j}_l = 0$ $(l \neq i)$, d.h.

$$\sum_{|\mathbf{j}|=3} \frac{D^{\mathbf{j}} f(\mathbf{x}^{i\pm})}{\mathbf{j}!}(\pm h\mathbf{e}_i)^{\mathbf{j}} = \frac{D^i D^i D^i f(\mathbf{x}^{i\pm})}{0!\cdots 3!\cdots 0!}(\pm h)^3 = \pm\tfrac{1}{6}h^3 f_i'''(\mathbf{x}^{i\pm}).$$

Nach dieser TAYLOR-Entwicklung ergibt sich für die in $\hat{\mathbf{g}}_h^{\mathrm{GN2}}$ verwendete Differenz von Funktionswerten:

$$f(\mathbf{x} + h\mathbf{e}_i) - f(\mathbf{x} - h\mathbf{e}_i) = 2h\mathbf{g}_i(\mathbf{x}) + \frac{1}{6}\Big(h^3 f_i'''(\mathbf{x}^{i+}) + h^3 f_i'''(\mathbf{x}^{i-})\Big),$$

d.h.

$$\hat{\mathbf{g}}_{hi}^{\mathrm{GN2}} = \mathbf{g}_i(\mathbf{x}) + \tfrac{1}{12}h^2\Big(f_i'''(\mathbf{x}^{i+}) - f_i'''(\mathbf{x}^{i-})\Big).$$

Sei $\mathbb{X}$ eine kompakte Menge, so dass $\mathbf{x}, \mathbf{x}^{i+}, \mathbf{x}^{i-} \in \mathbb{X}^0$ ($\mathbb{X}^0$ sei das Innere von $\mathbb{X}$) und $L^{(3)} \coloneqq \sup_{x\in\mathbb{X}} f_i'''(\mathbf{x})$ (endlich nach Voraussetzung).
Mit $\left|f_i'''(\mathbf{x}^{i+}) - f_i'''(\mathbf{x}^{i-})\right| \leq 2L^{(3)}$ gilt

$$\left|\hat{\mathbf{g}}_{h\,i}^{\mathrm{GN2}}(\mathbf{x}) - \mathbf{g}_i(\mathbf{x})\right| \leq \frac{1}{6}h^2 L^{(3)} \text{ und}$$
$$\hat{\mathbf{g}}_h^{\mathrm{GN2}}(\mathbf{x}) = \nabla f(\mathbf{x}) + O(h^2),\ h \to 0.$$

Für die Gradientennäherung 1. Ordnung betrachtet man die TAYLOR-Entwicklung

$$f(\mathbf{x} + h\mathbf{e}_i) = f(\mathbf{x}) + h\mathbf{g}_i(\mathbf{x}) + \frac{1}{2}h^2 f_i''(\mathbf{x}^{i+})$$

mit einem Punkt $\mathbf{x}^{i+}$ auf der Verbindungsstrecke von $\mathbf{x}$ und $\mathbf{x} + h\mathbf{e}_i$. Es folgt wegen $h > 0$ und mit $L^{(2)} \coloneqq \sup_{\mathbf{x}\in\mathbb{X}} \left|f_i''(\mathbf{x})\right|$:

$$\left|\frac{f(\mathbf{x} + h\mathbf{e}_i) - f(\mathbf{x})}{h} - \mathbf{g}_i(\mathbf{x})\right| = \frac{1}{2}h\underbrace{\left|f_i''(\mathbf{x}^{i+})\right|}_{\leq L^{(2)}}.$$

Also gilt:

$$\left|\hat{\mathbf{g}}_{h\,i}^{\mathrm{GN1}}(\mathbf{x}) - \mathbf{g}_i(\mathbf{x})\right| \leq \frac{1}{2}h\,L^{(2)} \text{ und}$$
$$\hat{\mathbf{g}}_h^{\mathrm{GN1}}(\mathbf{x}) = \nabla f(\mathbf{x}) + O(h),\ h \to 0.$$

□

Für das GSD-Verfahren mit Gradientennäherung GN1 werden $m+1$ Funktionsauswertungen für eine Gradientennäherung benötigt. Durch die Wiederverwendung der Funktionsauswertung $f(\mathbf{x}_k)$ benötigt das Verfahren aber nur $m+1$ Funktionsauswertungen pro Iteration. Im Falle der Verwendung der Gradientennäherung GN2 werden $2m$ Funktionsauswertungen pro Iteration benötigt. Dazu kommt jedoch noch die Funktionsauswertung an der neuen Iterierten $f(\mathbf{x}_{k+1})$, so dass hier $2m + 1$ Funktionsauswertungen pro Iteration benötigt werden.

Man hat Konvergenzaussagen der folgenden Form:

Bemerkung 4.12 (Konvergenz des GSD-Verfahrens). Sind die Voraussetzungen des Satzes 4.8 (Zoutendijk) erfüllt, und wählt man h so, dass für

$$\mathbf{p}_k = -\hat{\mathbf{g}}_h^{\mathrm{GN1/2}}(\mathbf{x}_k) \tag{4.14}$$

die Bedingung $\cos(\theta_k) \geq \delta > 0$ an den Winkel θ_k zwischen $\mathbf{p}_k$ und $-\nabla f(\mathbf{x}_k)$ des Satzes 4.10 erfüllt ist, so konvergiert das GSD-Verfahren gegen einen stationären Punkt von f.

Nun wird auf die numerische Differentiation eingegangen. Der folgende Abschnitt ist nach [Kun07].

Def. 4.13 (Kondition eines Problems)**.**
Die Konditionszahl *eines durch $u : \mathbb{X} \to \mathbb{Y}$ beschriebenen Problems wird als*

$$\kappa = \sup_{x_1, x_2 \in \mathbb{X}, x_1 \neq x_2} \frac{\|u(x_2) - u(x_1)\|}{\|x_2 - x_1\|} \tag{4.15}$$

definiert. Ist $\kappa \approx 1$, so spricht man von einem gut konditionierten Problem. Ist $\kappa = \infty$, spricht man von einem schlecht gestellten Problem.

Die Größe von κ hängt von der Wahl der Normen in (4.15) ab.
Die $\mathcal{C}^0$-Norm stetiger Funktinonen $f : [a, b] \to \mathbb{R}$ definiert man wie folgt: $\|f\|_{\mathcal{C}^0} \coloneqq \max_{t \in [a,b]} |f(t)|$.

Man muss beachten, dass bei der numerischen Differentiation ein schlecht gestelltes Problem vorliegt: Bei der Zuordnung $\mathcal{C}^1([a, b], \mathbb{R}) \ni f \mapsto f'(\bar{t})$ kann man Beispiele finden, bei denen $|f(t)| \leq 1 \forall t \in \mathbb{R}$ gilt, aber $f'(t_0)$ beliebig groß wird (etwa $f(t) = \tanh(ct)$, $t_0 = 0$), so dass für die Konditionszahl κ,

$$\kappa = \sup_{\substack{f_1, f_2 \in \mathcal{C}^1([a,b],\mathbb{R}), \\ f_1 \neq f_2}} \frac{\left|f_2'(\bar{t}_2) - f_1'(\bar{t}_1)\right|}{\|f_2 - f_1\|_{\mathcal{C}^0}}, \tag{4.16}$$

des Problems gilt $\kappa = \infty$. Man muss also sehr genau auf die Auswirkungen von Rundungsfehlern achten.

Bei der Wahl $\|f\| = \|f\|_{C^0} + \left\|\dot{f}\right\|_{C^0}$ läge eine gut konditioniertes Problem vor – dies würde aber voraussetzen, dass man auch die Daten von $\dot{f}$ zur Verfügung hätte, wovon genau nicht ausgegangen wird.

Für das folgende Lemma und die anschließende Bemerkung, deren Aussagen auf [NW06, S. 196] beruhen, kann man für die relative Genauigkeit u bei der Auswertung von f an die Maschinengenauigkeit u denken, die für Maschinenzahlen mit gerader Basis b und Mantissenlänge l in [Kun07] als $u \coloneqq \frac{1}{2} b^{-l+1}$ definiert worden ist. $\mathrm{comp}(x)$ sei der auf dem Computer berechnete Wert von x.

Lemma 4.14. *$\mathbb{X}$ sei eine kompakte Menge, so dass $\mathbf{x}, \mathbf{x} \pm h\mathbf{e}_i \in \mathbb{X}^0$. Die Rechnerarithmetik sei so, dass mit $L \coloneqq \sup_{\mathbf{x} \in \mathbb{X}} |f(\mathbf{x})|$ und u der relativen Genauigkeit bei der Auswertung von f gilt*

$$\Big|\mathrm{comp}\big(f(\mathbf{x})\big) - f(\mathbf{x})\Big| \leq uL \text{ für alle } \mathbf{x}. \tag{4.17}$$

Für die Aussage über $\hat{\mathbf{g}}^{GN1}$ seien die ungemischten zweiten Ableitungen von f beschränkt: $L^{(2)} = \sup_{\mathbf{x}\in\mathbb{X}} \|f_i''(\mathbf{x})\|$. Für die Aussage über $\hat{\mathbf{g}}^{GN2}$ seien die ungemischten dritten Ableitungen beschränkt: $L^{(3)} = \sup_{\mathbf{x}\in\mathbb{X}} \|f_i'''(\mathbf{x})\|$. Die Fehler der Gradientenschätzungen $\hat{\mathbf{g}}^{GN1}$ und $\hat{\mathbf{g}}^{GN2}$ sind dann in der folgenden Weise komponentenweise beschränkt ($\mathbf{g}_i(\mathbf{x})$ sei die i-te Komponente von $\nabla f(\mathbf{x})$):

$$\left| \operatorname{comp}\left(\hat{\mathbf{g}}_h^{GN1}(\mathbf{x})\right)_i - \mathbf{g}_i(\mathbf{x}) \right| \leq \tfrac{L^{(2)}}{2} h + 2\tfrac{uL}{h} \text{ und} \tag{4.18}$$

$$\left| \operatorname{comp}\left(\hat{\mathbf{g}}_h^{GN2}(\mathbf{x})\right)_i - \mathbf{g}_i(\mathbf{x}) \right| \leq \tfrac{L^{(3)}}{6} h^2 + 2\tfrac{uL}{h}. \tag{4.19}$$

Dabei wird angenommen, dass die Berechnung des Differenzenquotienten aus den Funktionswerten keine zusätzlichen Rundungsfehler involviert.

Beweis. Es werden die im Beweis von Lemma 4.11 hergeleiteten Schranken benutzt.

$$\begin{aligned}
&\left| \operatorname{comp}\left(\hat{\mathbf{g}}_h^{\mathrm{GN1}}(\mathbf{x})\right)_i - \mathbf{g}_i(\mathbf{x}) \right| = \left| \frac{\operatorname{comp}\left(f(\mathbf{x}+h\mathbf{e}_i)\right) - \operatorname{comp}\left(f(\mathbf{x})\right)}{h} - \mathbf{g}_i(x) \right| \\
&\leq \left| \frac{\operatorname{comp}\left(f(\mathbf{x}+h\mathbf{e}_i)\right) - \operatorname{comp}\left(f(\mathbf{x})\right)}{h} - \frac{f(\mathbf{x}+h\mathbf{e}_i) - f(\mathbf{x})}{h} \right| \\
&\quad + \left| \frac{f(\mathbf{x}+h\mathbf{e}_i) - f(\mathbf{x})}{h} - \mathbf{g}_i(x) \right| \\
&\leq 2\frac{uL}{h} + \frac{L^{(2)}}{2} h,
\end{aligned}$$

analog für $\hat{\mathbf{g}}_h^{\mathrm{GN2}}$. □

Bemerkung 4.15. Die minimierenden Werte h_1^* und h_2^* für die obere Fehlerschranke bei Näherung 1. bzw. 2. Ordnung sind

$$h_1^* = \sqrt{4u\frac{L}{L^{(2)}}} \text{ und} \tag{4.20}$$

$$h_2^* = \sqrt[3]{6u\frac{L}{L^{(3)}}}. \tag{4.21}$$

In der Praxis geht man davon aus, dass $\left|\frac{L}{L^{(2)}}\right|$ bzw. $\left|\frac{L}{L^{(3)}}\right|$ in der Größenordnung 1 ist, wenn das Problem gut skaliert ist, und wählt dann

$$h^{\mathrm{GN1}} = \sqrt{u} \text{ und } h^{\mathrm{GN2}} = \sqrt[3]{u}, \tag{4.22}$$

um einen guten Ausgleich zwischen Diskretisierungs- und Rundungsfehler zu erreichen. In Abbildung 4.1 ist der Verlauf der oberen Fehlerschranke dargestellt und die übliche Wahl (4.22) markiert (bei einem Verhältnis $\left|\frac{L}{L^{(2)}}\right| = \left|\frac{L}{L^{(3)}}\right| = 2$).

Eine Möglichkeit, um das verrauschte Problem (2.2) zu lösen, wäre $\tilde{f}(\mathbf{x})$ über viele Funktionsauswertungen zu mitteln, dies als $f(\mathbf{x})$ aufzufassen und die genannten deterministischen Optimierungsverfahren anzuwenden. In der *Stochastic-Approximation*-Theorie sieht man das Rauschen in $\tilde{f}(\mathbf{x})$ dagegen direkt als Zufallsgröße an und will die Mittelungen vermeiden.

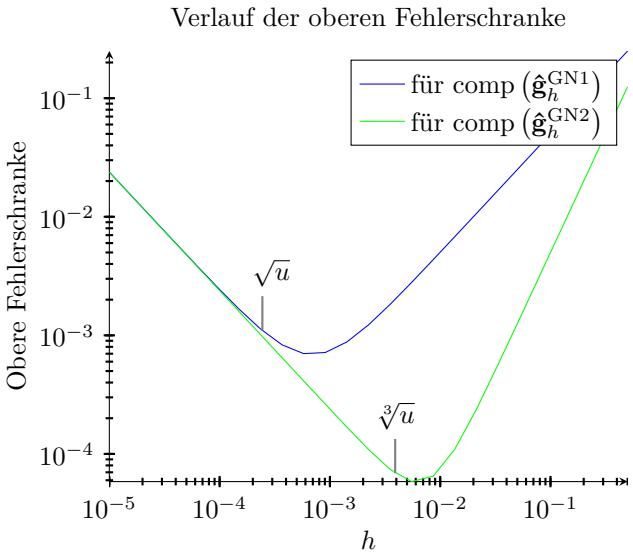


Abbildung 4.1: Rundungs- und Diskretisierungsfehler bei berechneten Gradientennäherungen. Der Wert für die Maschinengenauigkeit wurde mit $u = 2^{-24}$ gewählt. Dies entspricht dem Arbeiten nach IEEE 754 bei einfacher Genauigkeit, für doppelte Genauigkeit (`double`) wäre $u = 2^{-53}$. In (4.18) und (4.19) wurde beispielhaft $L^{(2)} = L^{(3)} = 1$ und $L = 2$ gewählt.

4.3 Stochastische Gradientenverfahren

4.3.1 Einführung

Die ursprüngliche Idee der *Stochastic-Approximation*-Theorie, aus deren Blickwinkel in dieser Arbeit ableitungsfreie stochastische Gradientenverfahren betrachtet werden (genauer des ROBBINS-MONRO-SA-Verfahrens [RM51]), ist die folgende: Die eindeutige Nullstelle x_0 einer Funktion g wird gesucht mit $\tilde{g}(x) = g(x) + R_x$. Es gelte $\mathbb{E}(\tilde{g}(x)) = g(x)$. Das durch die Iterationsvorschrift

$$X_{k+1} = X_k - a_k \tilde{g}(X_k), a_k > 0 \tag{4.23}$$

definierte Verfahren konvergiert dann gegen die Nullstelle x^*, in $\mathcal{L}^2$ falls $\tilde{g}$ gleichmäßig beschränkt ist, g nicht wächst, $g(x_0)$ existiert und positiv ist und $(a_k)_k$ die Bedingungen $\sum_{k=0}^{\infty} a_k = \infty$, $\sum_{k=0}^{\infty} a_k^2 < \infty$ erfüllt.

Bemerkung 4.16. In diesem Sinne kann das stochastische Verfahren des steilsten Abstiegs (2.7) auch als ROBBINS-MONRO-*Stochastic-Approximation* (RM-SA) aufgefasst werden.

Das RM-SA-Verfahren bezieht sich eigentlich darauf, die Nullstellen einer Funktion g zu finden, und wird hier im Rahmen der Gradientenverfahren darauf bezogen, einen stationären Punkt für f zu finden, in dem man die Nullstellensuche für $\mathbf{g} = \nabla f$ durchführt. Daher wird in dieser Arbeit die Bezeichnung $\mathbf{g}$ verwendet, wenn es eigenständig im Nullstellenproblem vorkommt, und $\mathbf{g} = \nabla f$, wenn es um das Minimierungsproblem $f(\mathbf{x}) \to \min!$ geht.

Im Folgenden werden die beiden KIEFER-WOLFOWITZ-artigen *Stochastic-Approximation*-Formen FDSA und SPSA betrachtet, die in der Einleitung bereits definiert worden sind. FDSA zeichnet sich gegenüber SPSA dadurch aus, dass man finite Differenzen, also komponentenweise Differenzenquotienten verwendet, um einen Schätzer für den Gradienten zu erhalten. Neben der in der Einleitung eingeführten FDSA-Form (2.10) mit einseitigen finiten Differenzen für die Gradientenschätzung $\hat{\mathbf{g}}_k(\mathbf{X}_k)$ ist auch die Form

$$\hat{\mathbf{g}}_k(\mathbf{X}_k) = \hat{\mathbf{g}}_k^{\mathrm{FD2}}(\mathbf{X}_k) = \left(\frac{\tilde{f}(\mathbf{X}_k + h_k \mathbf{e}_i) - \tilde{f}(\mathbf{X}_k - h_k \mathbf{e}_i)}{2h_k} \right)_{i=1,\dots,m} \tag{4.24}$$

mit symmetrischen finiten Differenzen gebräuchlich (Approximation 2. Ordnung).

Die folgende Definition wird sowohl in der Analyse des FDSA- als auch des SPSA-Verfahrens genutzt.

Def. 4.17 (Bias, Fehlerterm und Partionierung)**.**
Das KW-SA-Verfahren der Form (2.9) *wird wie folgt unterteilt:*

$$\mathbf{X}_{k+1} = \mathbf{X}_k + a_k\big(-\mathbf{g}(\mathbf{X}_k) - \mathbf{B}_k(\mathbf{X}_k) - \mathbf{E}_k(\mathbf{X}_k)\big) \tag{4.25}$$

(Prototyp für ROBBINS-MONRO/KIEFER-WOLFOWITZ*-artige Verfahren gemäß [KC78, Abschnitt 2.3.1]). Dabei ist* $\mathbf{B}_k$ *der* Bias *der Gradientenschätzung des KW-SA-Verfahrens am k-ten Iterationsschritt*

$$\mathbf{B}_k(\mathbf{X}_k) \coloneqq \mathbb{E}\big[\hat{\mathbf{g}}_k(\mathbf{X}_k) - \mathbf{g}(\mathbf{X}_k) \,|\, \mathbf{X}_k\big] \tag{4.26}$$

und $\mathbf{E}_k$ *der* Fehlerterm

$$\mathbf{E}_k(\mathbf{X}_k) \coloneqq \hat{\mathbf{g}}_k(\mathbf{X}_k) - \mathbb{E}\big[\hat{\mathbf{g}}_k(\mathbf{X}_k) \,|\, \mathbf{X}_k\big]. \tag{4.27}$$

Damit die letzten beiden Ausdrücke wohldefiniert sind, sei $\hat{\mathbf{g}}_k(\mathbf{X}_k)$ *eine* $\mathcal{L}^1$*-Zufallsgröße. Dies geht auch aus den Forderungen des wahrscheinlichkeitstheoretischen Modells hervor, siehe S. 59.*

Beweis. Es zeigt sich, dass die so definierte Aufteilung (4.25) wieder die allgemeine Kiefer-Wolfowitz-Form (2.9) ergibt:

$$\begin{aligned}
&- \mathbf{g}(\mathbf{X}_k) - \mathbf{B}_k(\mathbf{X}_k) - \mathbf{E}_k(\mathbf{X}_k) \\
&= -\mathbf{g}(\mathbf{X}_k) - \mathbb{E}\big[\hat{\mathbf{g}}_k(\mathbf{X}_k) \,|\, \mathbf{X}_k\big] + \underbrace{\mathbb{E}\big[\mathbf{g}(\mathbf{X}_k) \,|\, \mathbf{X}_k\big]}_{=\mathbf{g}(\mathbf{X}_k)} - \hat{\mathbf{g}}_k(\mathbf{X}_k) + \mathbb{E}\big[\hat{\mathbf{g}}_k(\mathbf{X}_k) \,|\, \mathbf{X}_k\big] \\
&= -\hat{\mathbf{g}}_k(\mathbf{X}_k),
\end{aligned}$$

dabei ist $\mathbf{g}(\mathbf{X}_k)$ $\mathbf{X}_k$-messbar (Doob-Dynkin-Lemma) und man verwendet die Eigenschaft (ii) der bedingten Erwartung. □

Bemerkung 4.18. Die Definition der Verzerrung $\mathbf{B}_k(\mathbf{X}_k)$ ist an die Definition 2.4 des Bias(P) eines Schätzers P angelehnt.

Es wird dann für die Gradientenschätzung $\hat{\mathbf{g}}_k^{\mathrm{SP}}$ gelten, dass

$$\mathbb{E}\big[\hat{\mathbf{g}}_k^{\mathrm{SP}}(\mathbf{X}_k) - \nabla f(\mathbf{X}_k) \,|\, \mathbf{X}_k\big] \to 0 \text{ f.s., } k \to \infty. \tag{4.28}$$

Nun wird näher auf die wahrscheinlichkeitstheoretische Modellierung eingegangen.

4.3.2 Wahrscheinlichkeitstheoretisches Modell

Im stochastischen Fall wird die Folge der Iterierten mit $(\mathbf{X}_1, \mathbf{X}_2, \dots)$ bezeichnet, jedes $\mathbf{X}_k$ ist dabei eine Zufallsgröße, $\mathbf{X}_k : \Omega \to \mathbb{R}^m$, d.h., jedes $\omega \in \Omega$ erzeugt eine Folge $\mathbf{X}_1(\omega), \mathbf{X}_2(\omega), \dots$. Der gesamte Zufall, der zum aktuellen Steuersignal $\mathbf{X}_k$ geführt hat, die zufälligen Perturbationen $\mathbf{D}_k$ und das Rauschen in den Messungen $R_{\mathbf{x}}$ werden also von einem ω bestimmt.

Die Struktur ist dabei die folgende: Zunächst gilt (3.1) für die Beziehung von $\mathbf{X}_k, R_k$ und $\mathbf{D}_k$, wie schon in der Einleitung erwähnt:

$$\mathbf{X}_{k+1} = \mathbf{X}_k - a_k u_k(\mathbf{X}_k, Z_k), Z_k = (\overline{\mathbf{D}}_k, R_k^{\mathrm{SP}}) \tag{3.1}$$

mit

$$u_k(\mathbf{X}_k, Z_k) = \hat{\mathbf{g}}_k^{\mathrm{SP}}(\mathbf{X}_k), \tag{4.29}$$

also einer Bezeichnung für die Gradientenschätzung, die die eingehenden Zufallsgrößen berücksichtigt, zufällig generierten Perturbationen $\overline{\mathbf{D}}_k = h_k \mathbf{D}_k$ und der Differenz $R_k^{\mathrm{SP}} = R_k^+ - R_k^-$ der Messfehler an den zur Bestimmung der Gradientenschätzung beteiligten Stellen $\mathbf{X}_k \pm \overline{\mathbf{D}}_k$, also

$$u(\mathbf{X}_k, Z_k) = \frac{f(\mathbf{X}_k + \overline{\mathbf{D}}_k) - f(\mathbf{X}_k - \overline{\mathbf{D}}_k)}{2\overline{\mathbf{D}}_k} + \frac{R_k}{2\overline{\mathbf{D}}_k}. \tag{4.30}$$

Z_k beschreibt das in jeder Iteration neu hinzukommende Zufällige. Den Zusammenhang der verschiedenen Zufallsanteile modelliert man wie folgt:

- $\{R_k^-\}_{k=1,2,\ldots}$, $\{R_k^+\}_{k=1,2,\ldots}$ seien unabhängige Folgen integrierbarer Zufallsgrößen. Sie seien unabhängig von $(\mathbf{X}_k, \mathbf{D}_k)$ (d.h. $\sigma(R_k^-), \sigma(\mathbf{X}_k, \mathbf{D}_k)$ unabhängig und $\sigma(R_k^+), \sigma(\mathbf{X}_k, \mathbf{D}_k)$ unabhängig, siehe Bemerkung 4.19).
- $\mathbf{D}_k$ sei ein Vektor unabhängiger Zufallsgrößen und unabhängig als Folge. Außerdem sei $\mathbf{D}_k$ unabhängig von $\mathbf{X}_k$.
- $\mathbf{X}_k$ sei eine Folge von Zufallsvariablen (m-dimensional).

Bemerkung 4.19. Die Sprechweise „X unabhängig von Y“ ist definiert als Unabhängigkeit der beiden Zufallsvariablen X, Y gemäß Def. 3.18(b). Eine Folge von Zufallsgrößen X_k heißt unabhängig, wenn die Menge $\{X_k\}_{k\in\mathbb{N}}$ gemäß Def. 3.18 unabhängig ist. In der Sprechweise „X unabhängig von (Y, Z)“ ist (Y, Z) die in (3.3) definierte, von Y und Z erzeugte σ-Algebra. Das bedeutet also, $\sigma(X), \sigma(Y, Z)$ sind unabhängig, was nicht mit der Unabhängigkeit von (X, Y, Z) als Zufallsvektor übereinstimmt, bei der $\sigma(X), \sigma(Y), \sigma(Z)$ unabhängig sein sollen. Die Unabhängigkeit R_k^- bzw. R_k^+ von $(\mathbf{X}_k, \mathbf{D}_k)$ schließt die Unabhängigkeit von $\mathbf{X}_k$ und $\mathbf{D}_k$ nach Lemma 3.24 mit ein.

Im Falle des FDSA-Verfahrens gilt analog, aber ohne die Perturbationen $\mathbf{D}_k$:

$$\mathbf{X}_{k+1} = \mathbf{X}_k - a_k u_k(\mathbf{X}_k, Z_k), Z_k = R_k^{\mathrm{FD}} \tag{4.31}$$

mit

$$u_k(\mathbf{X}_k, Z_k) = \hat{\mathbf{g}}_k^{\mathrm{FD1}}(\mathbf{X}_k), \tag{4.32}$$

mit der entsprechenden Differenz R_k^{FD}, also

$$u_k(\mathbf{X}_k, Z_k) = \frac{f(\mathbf{X}_k + h_k \mathbf{e}_i) - f(\mathbf{X}_k)}{h_k} + \frac{R_k}{h_k}. \tag{4.33}$$

Z_k beschreibt wiederum das in jeder Iteration neu hinzukommende Zufällige.

Damit ist $\mathbf{X}_{k+1}$ also für FDSA- und SPSA-Verfahren eine Funktion von $(\mathbf{X}_k, Z_k)$,

$$\mathbf{X}_{k+1} = v(\mathbf{X}_k, Z_k) \tag{4.34}$$

mit $v(\mathbf{X}_k, Z_n) = X_k - a_n u(\mathbf{X}_k, Z_k)$.

$\mathcal{G}_k$ sei die Filtrierung des sukzessive hinzukommenden Zufälligen,

$$\mathcal{G}_k \coloneqq \sigma(\mathbf{X}_0) \vee \sigma(Z_1, \ldots, Z_k). \tag{4.35}$$

Wegen $\mathbf{X}_{k+1} = v(\mathbf{X}_k, Z_k) = v(v(\mathbf{X}_{n-1}, Z_{n-1}), Z_n) = \cdots$ sieht man, dass $\mathbf{X}_{k+1}$ letztlich $\mathcal{G}_k$-messbar ist nach dem Doob-Dynkin-Lemma (Lemma 3.10).

Der folgende Abschnitt benutzt Überlegungen aus [Mseh; Msef].

Bemerkung 4.20 (Filtrierung). Spall verwendet die folgenden beiden Filtrierungen:

- in [Spa05, S. 183] die Filtrierung

$$\mathfrak{J}_k^{\mathrm{Spa05}} = \sigma(\mathbf{X}_0, \ldots, \mathbf{X}_k; \mathbf{D}_0, \ldots, \mathbf{D}_{k-1}), \text{ und} \tag{4.36}$$

- in [Spa92, S. 333] die Filtrierung

$$\mathfrak{J}_k^{\mathrm{Spa92}} = \sigma(\mathbf{X}_0, \ldots, \mathbf{X}_k). \tag{4.37}$$

Durch Überlegungen zu den Messbarkeitsbedingungen wurde hier stattdessen die Filtrierung

$$\mathcal{F}_k = \sigma(\mathbf{X}_0; Z_0, \ldots, Z_k) \tag{4.38}$$

verwendet, vergleiche auch die in [KY03, S. 122] verwendete Filtrierung für einen allgemeineren Fall. $\mathbf{X}_k$ ist $\mathcal{F}_k$-messbar und aus anschaulicher Sicht sollen die Filtrierungen den bis zur k-ten Iteration hinzugekommenen Zufall verkörpern.

Soll M_n,

$$\begin{aligned} &M_n = Y_1 + \cdots + Y_n, \\ &Y_k = u(\mathbf{X}_k, Z_k) - E[u(\mathbf{X}_k, Z_k) \,|\, \mathbf{X}_k] \\ &\text{mit } Z_k = (\overline{\mathbf{D}}_k, R_k), \end{aligned}$$

aber ein $\mathcal{G}_n$-Martingal sein, wobei $\mathcal{G}_n = \sigma(T_1, \ldots, T_n)$ für gewisse Zufallsgrößen T_k, dann muss es nach dem Doob-Dynkin-Lemma eine Borel-messbare Funktion $\breve{w}$ geben, so dass

$$M_n = \breve{w}(T_1, \ldots, T_n). \tag{4.39}$$

Für die beiden Filtrierungen $\mathfrak{J}_k^{\text{Spa05}}$ und $\mathfrak{J}_k^{\text{Spa92}}$ ist das ohne weitere Annahmen nicht ersichtlich: Hinreichend wäre die $\mathcal{G}_n$-Messbarkeit von $U_k = u(\mathbf{X}_k, Z_k)$. Für diese müsste U_k aber eine Borel-Funktion von $(\mathbf{X}_0, \ldots, \mathbf{X}_k; \mathbf{D}_0, \ldots, \mathbf{D}_{k-1})$ bzw. $(\mathbf{X}_0, \ldots, \mathbf{X}_k)$ sein. Eine Zufallsgröße, in die Z_k eingeht, würde also als eine Borel-Funktion dargestellt werden mit nur von Z_k unabhängigen Zufallsgrößen. Dass sich dies für M_n anders verhält als für Y_k ist wegen der Art, mit der Y_k zu M_n beitragen, nicht ersichtlich.

Für die Filtrierung $\mathcal{F}_n = \sigma(\mathbf{X}_0, Z_0, \ldots, Z_n)$ erhält man wegen $\mathbf{X}_{n+1} = \breve{v}(\mathbf{X}_0, Z_0, \ldots, Z_n)$ nach (3.35) eine messbare Abbildung u, die $(\mathbf{X}_0, Z_0, \ldots, Z_n)$ auf $(\mathbf{X}_0, \ldots, \mathbf{X}_{k+1})$ abbildet, ebenso für $(\mathbf{X}_0, \ldots, \mathbf{X}_{k+1}; \mathbf{D}_0, \ldots, \mathbf{D}_{k-1})$ wegen der Wahl von Z_k. Nach Hilfssatz 3.25 gilt also $\mathfrak{J}_{n+1} \subseteq \mathcal{F}_n$.

Die Theorie der *Stochastic-Approximation*-Verfahren bietet den grundlegenden Rahmen für die Behandlung des FDSA- und SPSA-Algorithmus.

Zusammenfassung des Modells

- Die Zufallsgrößen, die das Rauschen an den für die Gradientenschätzung verwendeten Stellen beschreiben, seien unabhängige Folgen integrierbarer Zufallsgrößen und voneinander unabhängig.
  Dies bezieht sich auf R_k^+, R_k^- im Falle des symmetrischen SPSA-Verfahrens, auf $R_k^\pm$, R_k^0 für die asymmetrischen SPSA-Verfahren 1. Ordnung, auf R_k^{i+}, R_k^{i-} für das symmetrische FDSA-Verfahren und auf $R_k^{i\pm}$, R_k^0 für die asymmetrischen FDSA-Verfahren 1. Ordnung.
- Die Zufallsgrößen $\mathbf{X}_k$ hängen in der Form

  $$\mathbf{X}_{k+1} = \mathbf{X}_k - a_k u(\mathbf{X}_k, Z_k)$$

  zusammen mit $Z_k = (\overline{\mathbf{D}}_k, R_k^{\mathrm{SP}})$ (SPSA) oder $Z_k = R_k^{\mathrm{FD}}$ (FDSA). Z_k bezeichnet also das „neu hinzukommende Zufällige“, siehe auch Abschnitt 4.3.4.
- Es gelte $u(\mathbf{X}_k, Z_k) \in \mathcal{L}^2$.
- Die Filtrierung sei $\mathcal{G}_k = \sigma(X_0) \vee \sigma(Z_1, \ldots, Z_n)$.

4.3.3 Gegenüberstellung der Verfahren

Bevor im folgenden Abschnitt auf die *Stochastic-Approximation*-Theorie eingegangen wird, soll an dieser Stelle noch ein Überblick der jetzt vollständig eingeführten Verfahren erfolgen. In Tabelle 4.1 wird der Auswertungsaufwand pro Iteration von GSD-/FDSA- und SPSA-Verfahren für allgemeine Dimension m und speziell für $m = 50$ verglichen. In Tabelle 4.2 werden die Formen aller eingeführten stochastischen und deterministischen Gradientenverfahren noch einmal angegeben. Dies wird durch die Übersicht der entsprechenden Gradientennäherungs- und -schätzungsterme in Tabelle 4.3 vervollständigt.

Verfahren	n für Näherung	
	1. Ordnung	2. Ordnung
Für allgemeines m		
GSD und FDSA	m	$2m+1$
SPSA	%	3
Beispiel für $m = 50$		
GSD und FDSA	50	101
SPSA	%	3

Tabelle 4.1: Vergleich des Auswertungsaufwands.
n ist die vom jeweiligen Algorithmus benötigte Anzahl von Funktionsauswertungen je Iteration.

Name	Form	Erläuterung	Ableitungsfrei?
	Deterministische Gradientenverfahren		
	Es stehen ungestörte Messungen der Zielfunktion f bzw. des Gradienten ∇f zur Verfügung.		
Gradientenabstiegsverfahren (SD)	$\mathbf{x}_{k+1} = \mathbf{x}_k + a_k \mathbf{p}_k(\mathbf{x}_k)$ (2.4)	$\mathbf{p}_k$ Abstiegsrichtung	nein
Verfahren des steilsten Abstiegs	$\mathbf{x}_{k+1} = \mathbf{x}_k - a_k \nabla f(\mathbf{x}_k)$ (2.5)		nein
ableitungsfreies Verfahren des steilsten Abstiegs (GSD)	$\mathbf{x}_{k+1} = \mathbf{x}_k - a_k \hat{\mathbf{g}}_k^{\mathrm{GN}}(\mathbf{x}_k)$	$\hat{\mathbf{g}}_k^{\mathrm{GN}}(\mathbf{x}_k)$ Gradientennäherung	ja
	Stochastische Gradientenverfahren		
	Es stehen nur verrauschte Messungen $\tilde{f}(\mathbf{x})$ der Zielfunktion f bzw. $\tilde{\mathbf{g}}(\mathbf{x})$ des Gradienten ∇f zur Verfügung. Die Folge der Iterierten $\mathbf{x}_k$ wird zu einer Folge von Zufallsgrößen $\mathbf{X}_k$.		
Robbins-Monro-*Stochastic-Approximation* (RM-SA)	$\mathbf{X}_{k+1} = \mathbf{X}_k - a_k \tilde{\mathbf{g}}(\mathbf{X}_k)$ (4.23)	$\tilde{\mathbf{g}}(\mathbf{X}_k)$ gestörter Gradient	nein
Kiefer-Wolfowitz-*Stochastic-Approximation* (KW-SA)	$\mathbf{X}_{k+1} = \mathbf{X}_k - a_k \hat{\mathbf{g}}_k(\mathbf{X}_k)$ (2.9)	$\hat{\mathbf{g}}_k$ Gradientenschätzung	
- *Finite Differences Stochastic Approximation* (FDSA)	$\mathbf{X}_{k+1} = \mathbf{X}_k - a_k \hat{\mathbf{g}}_k^{\mathrm{FD}}(\mathbf{X}_k)$	$\hat{\mathbf{g}}_k^{\mathrm{FD}}$ Finite-Differenzen-Gradientenschätzung (2.10) oder (4.24)	ja
- *Simultaneous Perturbation Stochastic Approximation* (SPSA)	$\mathbf{X}_{k+1} = \mathbf{X}_k - a_k \hat{\mathbf{g}}_k^{\mathrm{SP}}(\mathbf{X}_k)$	$\hat{\mathbf{g}}_k^{\mathrm{SP}}$ *Simultaneous-Perturbation*-Gradientenschätzung (2.11)	ja

Tabelle 4.2: Übersicht der genannten Verfahren

Name	Approximationsordnung 1. Ordnung	2. Ordnung	siehe...
	$\hat{\mathbf{g}}_h$ ist eine Gradientennäherung, d.h. $\hat{\mathbf{g}}_h(\mathbf{x}_k) \xrightarrow{h\to 0} \nabla f(\mathbf{x}_k)$.		
GSD	$\hat{\mathbf{g}}_{h\,i}^{\mathrm{GN1}}(\mathbf{x}_k) = \dfrac{f(\mathbf{x}_k + h\mathbf{e}_i) - f(\mathbf{x}_k)}{h}$	$\hat{\mathbf{g}}_{h\,i}^{\mathrm{GN2}}(\mathbf{x}_k) = \dfrac{f(\mathbf{x}_k + h\mathbf{e}_i) - f(\mathbf{x}_k - h\mathbf{e}_i)}{2h}$	(2.6) und (4.10)
	$\hat{\mathbf{g}}_{ki}$ ist ein Gradientenschätzer, d.h. $\mathbb{E}(\hat{\mathbf{g}}_k(\mathbf{X}_k) \,\vert\, \mathbf{X}_k) \to \nabla f(\mathbf{X}_k) + \mathbf{B}_k(\mathbf{X}_k)$.		
FDSA	$\hat{\mathbf{g}}_{ki}^{\mathrm{FD1}}(\mathbf{X}_k) = \dfrac{\tilde{f}(\mathbf{X}_k + h_k\mathbf{e}_i) - \tilde{f}(\mathbf{X}_k)}{h_k}$	$\hat{\mathbf{g}}_{ki}^{\mathrm{FD2}}(\mathbf{X}_k) = \dfrac{\tilde{f}(\mathbf{X}_k + h_k\mathbf{e}_i) - \tilde{f}(\mathbf{X}_k - h_k\mathbf{e}_i)}{2h_k}$	(2.10) und (4.24)
SPSA	%	$\hat{\mathbf{g}}_{ki}^{\mathrm{SP}}(\mathbf{X}_k) = \dfrac{\tilde{f}(\mathbf{X}_k + h_k\mathbf{D}_k) - \tilde{f}(\mathbf{X}_k - h_k\mathbf{D}_k)}{2h_k\mathbf{D}_{ki}}$	(2.11), Bedingungen an $\mathbf{D}_k$ in Abschnitt 4.3.7

Tabelle 4.3: Formen von $\hat{\mathbf{g}}_k$. Angegeben ist jeweils die i-te Komponente $\hat{\mathbf{g}}_{ki}$.

4.3.4 Aus der *Stochastic-Approximation*-Theorie

Für die folgenden Bedingungen an SA-Verfahren werden weitere Begriffe benötigt. Die Definitionen der erweiterten gleichgradigen Stetigkeit und der Abschnitt über die Interpolation des SA-Prozesses sind Kapitel 4 in KUSHNER und YIN: *Stochastic approximation and recursive algorithms and applications*, [KY03] entnommen.

Der Beweis des Konvergenzresultats für das SPSA-Verfahren nach SPALL [Spa05; Spa92] stützt sich auf einen grundlegenden Konvergenzsatz für Verfahren des ROBBINS-MONRO-SA-Typs aus [KC78], der in diesem Abschnitt angegeben wird.

Die *Stochastic-Approximation*-Theorie nach KUSHNER [KC78; KY03] wird in [KY03] unter anderem für den Fall von Rauschen mit Martingaldifferenz-Eigenschaft entwickelt. Die Idee für diese Annahme an das Rauschen stammt aus dem folgendem Beispiel [KY03, S. 97]:

Beispiel.
Man partioniert eine Folge von Zufallsgrößen oft in einen von der Vergangenheit abhängigen Teil und einen unvorhersagbaren Teil, etwa wie folgt:

$$Y_k \doteq \underbrace{\big(Y_k - \mathbb{E}[Y_k \,|\, Y_l, l < k]\big)}_{=:\delta M_k} + \mathbb{E}[Y_k \,|\, Y_l, l < k], \tag{4.40}$$

wobei δM_k eine Martingaldifferenz (siehe Def. 3.55) ist, da
$M_n \doteq \sum_{k=0}^{n} \big(Y_k - \mathbb{E}[Y_k \,|\, Y_l,\, l < k]\big)$ ein Martingal ist.

Beweis. Dies ist ein Martingal nach Lemma 3.54 wegen

$$\begin{aligned}\mathbb{E}[\delta M_{n+1} \,|\, Y_l, l \le n] &\doteq \mathbb{E}\big[Y_{n+1} - \mathbb{E}[Y_{n+1} \,|\, Y_l, l \le n] \,|\, Y_l, l \le n\big] \\ &\doteq \mathbb{E}[Y_{n+1} \,|\, Y_l, l \le n] - \mathbb{E}[Y_{n+1} \,|\, Y_l, l \le n] \doteq 0\end{aligned}$$

unter Nutzung der Turmeigenschaft nach Lemma 3.32(v). □

Im Beweis für die Konvergenz von SA-Verfahren nach der DGL-Methode wie in [KC78] und [KY03, Abschnitt 5.2] zeigt man zunächst, dass eine stückweise lineare bzw. stückweise konstante Interpolation $\widehat{\mathbf{X}}_k$ des Prozesses $\mathbf{X}_k$ (im erweiterten Sinn) gleichgradig stetig ist. Nach dem Theorem 4.24 (ARZELÀ-ASCOLI) erhält man dann eine konvergente Teilfolge, die einer sogenannten gemittelten DGL genügt. Man benutzt das Verhalten dieser DGL, um Aussagen über das asymptotische Verhalten von $\mathbf{X}_k$ abzuleiten [KY03, S. 117ff.].

Def. 4.21 (Gleichgradige Stetigkeit)**.**
Sei $(f_n)_n$ eine Folge von Funktionen mit $f_n \in \mathcal{C}(\mathbb{R}, \mathbb{R}^m)$, $n \in \mathbb{N}_0$,

- *$\{f_n(0)\}$ sei beschränkt und*
- *es gebe für jedes $\bar{t}$ und $\epsilon > 0$ ein $\delta > 0$, so dass für alle $n \in \mathbb{N}$*

$$\sup_{0 \le t-s \le \delta \ ,\ |t| \le \bar{t}} |f_n(t) - f_n(s)| \le \epsilon. \tag{4.41}$$

Dann nennt man die Folge $(f_n)_n$ gleichgradig stetig *in $\mathcal{C}(\mathbb{R}, \mathbb{R}^m)$.*

Theorem 4.22 (Arzelà-Ascoli)**.**
Sei $(f_n)_n$ eine gleichgradig stetige Folge von $\mathcal{C}(\mathbb{R}, \mathbb{R}^m)$-Funktionen.
Dann enthält sie eine konvergente Teilfolge, die auf jedem beschränkten Intervall gleichmäßig gegen eine stetige Grenzfunktion konvergiert.

Def. 4.23 (Gleichgradige Stetigkeit im erweiterten Sinn)**.**
Sei $(f_n)_n$ eine Folge von Funktionen, bei der jedes f_n eine messbare $\mathbb{R}^m$-wertige Funktion auf $\mathbb{R}$ ist,

- *$\{f_n(0)\}$ beschränkt ist und*
- *es für jedes $\bar{t}$ und $\epsilon > 0$ ein $\delta > 0$ gibt, so dass*

$$\limsup_n \sup_{0 \le t-s \le \delta \ , \ |t| \le \bar{t}} |f_n(t) - f_n(s)| \le \epsilon. \tag{4.42}$$

Dann nennt man die Folge $(f_n)_n$ gleichgradig stetig im erweiterten Sinn*.*

Aus dieser gleichgradigen Stetigkeit im erweiterten Sinn folgt im Allgemeinen nicht, dass die f_n stetig sind, die Aussage von Arzelà-Ascoli gilt aber dennoch, nach [KY03, Theorem 2.2.]:

Theorem 4.24 (Arzelà-Ascoli)**.**
Sei $(f_n)_n$ eine im erweiterten Sinn gleichgradig stetige Folge von $\mathcal{C}(\mathbb{R}, \mathbb{R}^m)$-Funktionen. Dann enthält sie eine gegen eine stetige Grenzfunktion konvergierende Teilfolge, die auf jedem beschränkten Intervall gleichmäßig konvergiert.

Dies stützt die Argumentation von Kushner und Yin [KY03, S. 124], dass die Nutzung der stückweise konstanten Interpolation eine einfachere Notation ermöglicht als etwa lineare Interpolation und dabei durch die fast sichere gleichgradige Stetigkeit im erweiterten Sinn keine Nachteile entstehen.

Def. 4.25 (asymptotisch stabiler Gleichgewichtspunkt einer DGL)**.**
Sei $v : \mathbb{R}^m \to \mathbb{R}^m$ stetig differenzierbar. Für das Differentialgleichungssystem

$$\dot{\mathbf{x}} = v(\mathbf{x}) \tag{4.43}$$

mit $\mathbf{x} = \mathbf{x}(t) \in \mathbb{R}^m$ heißt eine Lösung $\overline{\mathbf{x}} \in \mathcal{C}^1([0, \infty], \mathbb{R}^m)$ aymptotisch stabil*, wenn*

- *$\forall \varepsilon > 0 \quad \exists \delta(\varepsilon) > 0$, so dass für alle $\mathbf{x}^0 \in \mathbb{R}^m$ mit $\left\|\mathbf{x}^0 - \overline{\mathbf{x}}(0)\right\| \le \delta$ gilt: Die Lösung $\mathbf{x} = \mathbf{x}(t)$ mit Anfangswert $\mathbf{x}(0) = \mathbf{x}^0$ existiert $\forall t \ge 0$ und $\|\mathbf{x}(t) - \overline{\mathbf{x}}(t)\| \le \epsilon \quad \forall t \ge 0$.*
- *Außerdem gelte: $\lim_{t\to\infty} \mathbf{x}(t) = \lim_{t\to\infty} \overline{\mathbf{x}}(t)$.*

Eine asymptotisch stabile Lösung $\overline{\mathbf{x}}(t) = \mathbf{x}^* \in \mathbb{R}^n$ heißt asymptotisch stabiler Gleichgewichtspunkt *der DGL* (4.43)*.*

Die Lösungen $x = x(t)$ der Differentialgleichung $\dot{x} = v(x)$ heißen auch *Trajektorien*. Mit $N_\delta(A)$ werde eine δ-Umgebung von A bezeichnet, mit Def. A.8 gilt:

$$N_\delta(A) = \{x \in \mathbb{R}^m \,|\, d(x, A) < \delta\}.$$

Die in KUSHNERs Buch [KY03, S. 104] verwendete Definition der lokalen asymptotischen Stabilität wird angegeben:

Def. 4.26. Eine Menge $A \subseteq H$ heißt *lokal stabil* im Sinne von Ljapunov[3], wenn es zu jedem $\varepsilon > 0$ ein $\delta > 0$ gibt, so dass alle Trajektorien von $\dot{x} = v(x)$, die in $N_\delta(A)$ beginnen, $N_\varepsilon(A)$ nie verlassen. Falls die Trajektorien in A enden, dann nennt man A *asymptotisch stabil* im Sinne von Ljapunov. Wenn dies für alle Anfangsbedingungen gilt, so spricht man von *globaler asymptotischer Stabilität* (nach Ljapunov).

Def. 4.27 (Einzugsbereich einer DGL)**.** Den *Einzugsbereich einer Differentialgleichung* (4.43) mit asymptotisch stabilem Gleichgewichtspunkt $\mathbf{x}^*$ definiert man gemäß

$$\mathbb{D}(\mathbf{x}^*) \coloneqq \{\mathbf{x}_0 \in \mathbb{R}^m \mid \lim_{t\to\infty} \mathbf{x}(t;\mathbf{x}_0) = \mathbf{x}^*\}, \tag{4.44}$$

wobei $\mathbf{x}(t;\mathbf{x}_0)$ die Lösung von (4.43) mit Anfangswert $\mathbf{x}(0) = \mathbf{x}_0$ ist.

Für die Interpolation $\widehat{\mathbf{X}}_0(t) : \mathbb{R} \to \mathbb{R}^m$ der Iterierten $\mathbf{X}_k$ definiert man in [KY03, S. 122] eine Zeitskala in Termen der Schrittweitenfolge a_k (vergleiche Abschnitt 4.2.2):

Def. 4.28 (Zeitskala)**.**
Man definiert

$$t_0 = 0, \quad t_k = \sum_{i=0}^{k-1} a_i. \tag{4.45}$$

Es folgt die Definition der Interpolation aus [KC78, S. 26]:

Def. 4.29. Seien $\mathbf{X}_k$ die Iterierten eines *Stochastic-Approximation*-Prozesses mit Schrittweiten a_k. Man definiert die Interpolation $\widehat{\mathbf{X}}_k$ der Iteriertenfolge $\mathbf{X}_k$ vermöge

$$\widehat{\mathbf{X}}_0(t_k) = \mathbf{X}_k, \tag{4.46}$$

$\widehat{\mathbf{X}}_0(t)$, $t \geq 0$ stückweise linear interpoliert mit Interpolationsintervallen a_k und

$$\widehat{\mathbf{X}}_k(t) = \begin{cases} \widehat{\mathbf{X}}_0(t_k + t) & t \geq -t_k, \\ \mathbf{X}_0 & t \leq -t_k. \end{cases} \tag{4.47}$$

Der folgende Satz ist im Wesentlichen Theorem 2.3.1 aus [KC78] entnommen.

Kushner-Clark Theorem 4.30 (Konvergenz RM-artiger Verfahren)**.**
Sei $\mathbf{X}_k$ eine Folge von Zufallsvariablen, die aus der Iterationsvorschrift (4.25)*,*

$$\mathbf{X}_{k+1} = \mathbf{X}_k + a_k\big(-\mathbf{g}(\mathbf{X}_k) - \mathbf{B}_k(\mathbf{X}_k) - \mathbf{E}_k(\mathbf{X}_k)\big)$$

hervorgehen. Es gelten die folgenden Bedingungen:

(i) $\mathbf{g} \in \mathcal{C}(\mathbb{R}^m, \mathbb{R}^m)$.

[3] Gebräuchlich sind auch die Schreibweisen Ljapunow, Ljapunoff, sowie im Englischen Liapunov.

(ii) $(\mathbf{B}_k)_k$ *sei eine (f.s.) beschränkte Folge* $\mathbb{R}^m$*-wertiger Zufallsvektoren und* $\mathbf{B}_k \xrightarrow{k\to\infty} 0$ *f.s.*

(iii) a_k *sei eine Folge positiver reeller Zahlen, so dass* $a_k \to 0, \sum_k a_k = \infty$.

(iv) $\mathbf{E}_k$ *sei eine Folge* $\mathbb{R}^m$*-wertiger Zufallsvektoren und* $\forall \eta > 0$

$$\lim_{n\to\infty} \mathbb{P}\Big(\sup_{n\geq k_0} \Big\| \sum_{k=k_0}^{n} a_k \mathbf{E}_k \Big\| \geq \eta \Big) = 0. \tag{4.48}$$

(v) $\mathbf{X}_k$ *sei fast sicher beschränkt.*

Dann gibt es eine Nullmenge $\Omega_0 \subseteq \Omega$*, und für* $\omega \notin \Omega_0$ *gilt:*

- $\widehat{\mathbf{X}}_k$ *ist gleichgradig stetig und*
- *der Limes* $\widehat{\mathbf{X}}$ *jeder konvergenten Teilfolge von* $\widehat{\mathbf{X}}_k$ *ist beschränkt und genügt der DGL*

$$\dot{\mathbf{x}} = -\mathbf{g}(\mathbf{x}) \tag{4.49}$$

auf dem Zeitintervall $\mathbb{I} = (-\infty, \infty)$.

Ist $\mathbf{x}^*$ *ein lokal asymptotisch stabiler Gleichgewichtspunkt der DGL (im Sinne von Ljapunov), dann gilt für solche* ω*:*
Wenn es eine kompakte Teilmenge $\mathbb{S}$ *des Einzugsbereichs* $\mathbb{D}(\mathbf{x}^*)$ *gibt, so dass* $\mathbf{X}_k \in \mathbb{S}$ *unendlich oft, dann gilt:*

$$\mathbf{X}_k \xrightarrow{k\to\infty} \mathbf{x}^* \quad \textit{f.s.} \tag{4.50}$$

Die Bedingung (iv) entspricht A.2.2.4″ in [KC78, S.29] und kann abgeschwächt werden zu Bedingung A.2.2.4, [KC78, S. 28], die eigentlich im Theorem 2.3.1 verwendet wird, aber von der hier angegebenen Bedingung impliziert wird. Äquivalente Bedingungen werden auch in [WCK96, Def. 1] behandelt.

Bemerkung 4.31 (zur Beschränktheit der Iterierten). Die Bedingung der Beschränktheit der Iterierten kann Anlass zur Kritik am Konvergenzsatz bieten, wenn man davon ausgehen muss, dass das Erfülltsein einer solche Bedingung für die Anwendung des Satzes schlecht gezeigt werden kann. In [KC78] sind Gründe angegeben, warum die Bedingung nicht einschränkend ist: Will man $\{\mathbf{X}_k\}_{k=1,2,\ldots}$ auf einen Hyperwürfel $\mathbb{H}$ einschränken oder glaubt man, dass alle relevanten Gleichgewichtspunkte der DGL (4.49) darin liegen, so modifiziert man das Verfahren (2.9) mit einer Projektion $\pi_{\mathbb{H}} : \mathbb{R}^m \to \mathbb{H}$, die $\mathbf{x}$ auf $\mathbb{H}$ projiziert und die DGL (4.49) wird zu

$$\dot{\mathbf{x}} = \pi_{\mathbb{H}}\big(\mathbf{g}(\mathbf{x})\big). \tag{4.51}$$

Nach [KC78, Kapitel 5.3.] bleibt dann das Konvergenz-Theorem 4.30 mit (4.51) anstatt von (4.49) gültig.

Dies wird in [KY03] ausgebaut, in dem für das beschränkte Optimierungsproblem

$$f(\mathbf{x}) \to \min_{\mathbf{x}\in\mathbb{H}}!$$

ein zum oben genannten Theorem analoges Resultat angegeben wird: [KY03, S.127, Theorem 2.1].

4.3.5 Vorbereitungen für die Anwendung der *Stochastic-Approximation*-Theorie

Den Beweis des Konvergenzresultates für das FDSA- und SPSA-Verfahren wird man letztlich auf ähnlichem Wege auf das oben angegebene KUSHNER-CLARK-Theorem zurückführen. Die Argumentation wird daher gebündelt und die verwendeten Zwischenresultate in diesem Abschnitt dargestellt.

Das folgende Lemma wird für die Fälle
$T_{k\,i} = \frac{\tilde{f}(\mathbf{X}_k+\overline{\mathbf{D}}_k)-\tilde{f}(\mathbf{X}_k-\overline{\mathbf{D}}_k)}{2\mathbf{D}_{ki}}$ und $T_{k\,i} = \frac{\tilde{f}(\mathbf{X}_k+h_k\mathbf{e}_i)-\tilde{f}(\mathbf{X}_k-h_k\mathbf{e}_i)}{2}$ verwendet.

Lemma 4.32. *Sei* $\mathbf{T_k}$ $\mathcal{L}^2$*-beschränkt. Definiere*

$$\mathbf{E}_k = \mathbf{U}_k - \mathbb{E}[\mathbf{U}_k \,|\, \mathcal{A}] \text{ und}$$
$$\mathbf{U}_k = \frac{1}{h_k}\mathbf{T_k},$$

dann ist

$$\mathbb{E}(\|\mathbf{E}_k\|^2) \leq c h_k^{-2} \tag{4.52}$$

für ein von k *unabhängiges* c*.*

Beweis. Es gilt $\mathbb{E}(\|\mathbf{E}_k\|^2) = \mathbb{E}(\mathbf{E}_{k1}^2 + \cdots + \mathbf{E}_{km}^2) = \mathbb{E}(\mathbf{E}_{k1}^2) + \cdots + \mathbb{E}(\mathbf{E}_{km}^2)$, es reicht also $\mathbf{E}_{ki}$ zu betrachten. Für dieses gilt:

$$\begin{aligned}
\mathbb{E}(\mathbf{E}_{ki}^2) &= \mathbb{E}\Big(\Big|\mathbf{U}_{ki} - \mathbb{E}[\mathbf{U}_{ki} \,|\, \mathcal{A}]\Big|^2\Big) \\
&\leq \mathbb{E}\Big(\Big(\,|\mathbf{U}_{ki}| + |\mathbb{E}[\mathbf{U}_{ki} \,|\, \mathcal{A}]|\Big)^2\Big) \\
&= \mathbb{E}({\mathbf{U}_{ki}}^2) + 2\mathbb{E}\Big(\,|\mathbf{U}_{ki}\mathbb{E}[\mathbf{U}_{ki} \,|\, \mathcal{A}]|\,\Big) + \mathbb{E}\Big(\big(\mathbb{E}[\mathbf{U}_{ki} \,|\, \mathcal{A}]\big)^2\Big).
\end{aligned}$$

Nutze nun die $\mathcal{L}^2$-Beschränktheit von $\mathbf{T}_k$, also die Existenz eines $c \geq 0$, das nicht von k abhängt, mit $\mathbb{E}(|\mathbf{T}_k|^2) \equiv \mathbb{E}(\mathbf{T}_k^2) \leq c$ aus. Es gilt für $\mathbf{U}_{ki}$:

$$\mathbb{E}({\mathbf{U}_{ki}}^2) = \mathbb{E}\big(\frac{1}{h_k^2}\mathbf{T}_k^2\big) = \frac{1}{h_k^2}\mathbb{E}\big(\mathbf{T}_k^2\big) \leq c\frac{1}{h_k^2}.$$

Für $\mathbb{E}\big((\mathbb{E}[\mathbf{U}_{ki} \,|\, \mathcal{A}])^2\big)$ gilt:

$$\mathbb{E}\Big(\big(\mathbb{E}[\mathbf{U}_{ki} \,|\, \mathcal{A}]\big)^2\Big) = \mathbb{E}\Big(\frac{1}{h_k^2}\big(\mathbb{E}[\mathbf{T}_{ki} \,|\, \mathcal{A}]\big)^2\Big) = \frac{1}{h_k^2}\mathbb{E}\Big(\big(\mathbb{E}[\mathbf{T}_{ki} \,|\, \mathcal{A}]\big)^2\Big) \leq \frac{1}{h_k^2}c$$

nach Lemma 3.43. Nach der CAUCHY-SCHWARZschen Ungleichung in der Form (3.9) gilt mit Lemma 3.43

$$\big(\mathbb{E}(|\mathbf{T}_{ki}|\,\mathbb{E}[|\mathbf{T}_{ki}|\,|\,\mathcal{A}])\big)^2 \leq \underbrace{\mathbb{E}(|\mathbf{T}_{ki}|^2)}_{\leq c}\,\underbrace{\mathbb{E}\big((\mathbb{E}[|\mathbf{T}_{ki}|\,|\,\mathcal{A}])^2\big)}_{\leq c} \leq c^2. \tag{4.53}$$

Das heißt für den mittleren Summanden:

$$\mathbb{E}\Big(\Big|\mathbf{U}_{ki}\mathbb{E}[\mathbf{U}_{ki} \,|\, \mathcal{A}]\Big|\Big) = \tfrac{1}{h_k^2}\mathbb{E}\Big(\,|\mathbf{T}_{ki}|\,\underbrace{|\mathbb{E}[\mathbf{T}_{ki} \,|\, \mathcal{A}]|}_{\leq \mathbb{E}[|\mathbf{T}_{ki}|\,|\,\mathcal{A}]}\,\Big) \leq \tfrac{1}{h_k^2}\mathbb{E}\Big(\,|\mathbf{T}_{ki}|\,\mathbb{E}[|\mathbf{T}_{ki}|\,|\,\mathcal{A}]\Big) \leq \tfrac{1}{h_k^2}c.$$

In der ersten Ungleichung wurde Regel (vii) aus Lemma 3.34 verwendet, die zweite verwendet die oben genannte Anwendung der CAUCHY-SCHWARZschen-Ungleichung. Zusammen ergibt sich daraus:

$$\mathbb{E}(\mathbf{E}_{ki}^2) \leq \frac{1}{h_k^2}\widetilde{c},$$

woraus die Behauptung folgt. □

Das folgende Lemma wird für die Fälle

- $Z_k = R_k^{i+} - R_k^{i-}$, $u = \hat{\mathbf{g}}_k^{\mathrm{FD2}}$ und
- $Z_k = (R_k^+ - R_k^-, \overline{\mathbf{D}}_k)$, $u = \hat{\mathbf{g}}_k^{\mathrm{SP}}$

verwendet.

Lemma 4.33 (Martingalbeschränkung des SA-Fehlerterms)**.**
Seien $\mathbf{X}_0$ und Z_k, $k \geq 0$ unabhängige Zufallsgrößen. Mit einer glatten Funktion $u : E \times E \to E$ gelte

$$\begin{aligned}\mathbf{X}_{k+1} &= \mathbf{X}_k - a_k u(\mathbf{X}_k, Z_k),\\ \mathbf{U}_k &= u(\mathbf{X}_k, Z_k) \in \mathcal{L}^2 \text{ und}\\ \mathbf{E}_k &= \mathbf{U}_k - \mathbb{E}[\mathbf{U}_k \,|\, \mathbf{X}_k].\end{aligned}$$

Es sei $\mathcal{F}_n = \sigma(\mathbf{X}_0) \vee \sigma(Z_1, \ldots, Z_n)$, $\mathbb{E}(\|\mathbf{E}_k\|^2) = O(h_k^{-2})$ und $\sum_{k\in\mathbb{N}_0} a_k^2/h_k^2 < \infty$. Dann gilt:

$$\lim_{k_0\to\infty} \mathbb{P}\left(\sup_{n\geq k_0} \left\| \sum_{k=k_0}^{n} a_k \mathbf{E}_k \right\| \geq \eta\right) \doteq 0 \ \forall \eta > 0. \tag{4.54}$$

Die Idee, dass $\left\| \sum_{k=k_0}^{n} a_k \mathbf{E}_k \right\|$ ein Submartingal ist, und man die DOOBsche Martingalungleichung für Submartingale statt für Martingale verwendet (wie es in [Spa92] scheint), geht auf [Piab] zurück.

Beweis. Der Beweis wird die DOOBsche Martingalungleichung, die auch für nichtnegative Submartingale gilt, und die Orthogonalität von $\mathcal{L}^2$-Zufallsgrößen ausnutzen, um (4.54) zu erreichen.

Mit $v_k(x, z) = x - a_k u_k(x, z)$ ist

$$\mathbf{M}_{k_0;n} = \sum_{k=k_0}^{n} a_k \mathbf{E}_k$$

nach Lemma 3.62 (Martingalsatz) ein $\mathcal{F}_n$-Martingal.[4] Aus den Voraussetzungen folgt nach Lemma 3.44, dass U_k eine $\mathcal{L}^1$-Zufallsgröße ist.

Die Normabbildung $x \mapsto \|x\|$ ist konvex auf $\mathbb{R}^m$, denn

$$\|c_1\mathbf{x} + c_2\mathbf{y}\| \leq |c_1|\,\|\mathbf{x}\| + |c_2|\,\|\mathbf{y}\| = c_1\,\|\mathbf{x}\| + c_2\,\|\mathbf{y}\|$$

für $c_1, c_2 \geq 0$, $c_1 + c_2 = 1$. Wegen Lemma 3.43 ist mit U_k auch $\mathbb{E}(U_k \,|\, \mathbf{X}_k)$ eine $\mathcal{L}^2$-Zufallsgröße, also auch $\mathbf{E}_k = U_k - \mathbb{E}(U_k \,|\, \mathbf{X}_k) \in \mathcal{L}^2$ und auch $M_n \in \mathcal{L}^2$.

[4] Als Teil des Beweises zur Konvergenz des SPSA-Verfahrens entspricht dies dem Schritt in [Spa92, S. 335], an dem $\sum_{k=k_0}^{n} a_k \mathbf{E}_k$ als Martingal bezeichnet wird, ohne die Filtrierung anzugeben, siehe auch die Bemerkung zur Wahl der Filtrierung (S. 57).

Nach Lemma 3.46 ist $\|\mathbf{M}_{k_0;n}\| \in \mathcal{L}^2$ und da $\mathcal{L}^2 \subset \mathcal{L}^1$ nach Korollar 3.44 ist $\|\mathbf{M}_{k_0;n}\| \in \mathcal{L}^1$, und man erhält mit Lemma 3.57 für $\varphi = \|\cdot\|$, dass $\|M_n\|$ ein $\mathcal{F}_n$-Submartingal ist. Beachte in diesem Zusammenhang Def. 3.39.

Auf dieses nichtnegative Submartingal wird nun die Variante (3.39) der Doobschen Martingalungleichung nach Lemma 3.64 angewendet und man erhält:

$$\mathbb{P}(\sup_{n\geq k_0} \|\mathbf{M}_{k_0;n}\| \geq \eta) \leq \frac{1}{\eta^2} \lim_{n\to\infty} \mathbb{E}(\|\mathbf{M}_{k_0;n}\|^2) =: r. \tag{4.55}$$

Da nach Lemma 3.56 die Komponenten $\mathbf{E}_{ki}$ von $\mathbf{E}_k$ orthogonal sind, und damit nach Lemma 3.50 $\mathbf{E}_k$ eine orthogonale Folge ist, folgt nach (3.24):

$$\mathbb{E}(\|\mathbf{M}_{k_0;n}\|^2) = \sum_{k=k_0}^{n} \mathbb{E}(\|a_k\mathbf{E}_k\|^2) = \sum_{k=k_0}^{n} a_k^2 \mathbb{E}(\|\mathbf{E}_k\|^2). \tag{4.56}$$

Nach (4.56) ist dann $r = \frac{1}{\eta^2}\sum_{k=k_0}^{\infty} a_k^2\mathbb{E}(\|\mathbf{E}_k\|^2)$. Die Summe konvergiert, da nach Voraussetzung $\mathbb{E}(\|\mathbf{E}_k\|^2) \leq ch_k^2$ und $\sum_{n\in\mathbb{N}_0} a_k^2/h_k^2 < \infty$. Aus diesem Grund gilt $\lim_{k_0\to\infty}\sum_{k=k_0}^{\infty} a_k^2\mathbb{E}(\|\mathbf{E}_k\|^2) = 0$. Daraus folgt für den Grenzübergang $k_0 \to \infty$ in (4.55)

$$\lim_{k_0\to\infty} \mathbb{P}(\sup_{n\geq k_0} \|\mathbf{M}_{k_0;n}\| \geq \eta) \leq 0.$$

□

4.3.6 Konvergenzanalyse des *Finite Differences Stochastic Approximation* Verfahrens

In diesem Abschnitt gilt die folgende Bezeichnung: $R_k \coloneqq R_k^{\mathrm{FD2}} \equiv R_k^{i+} - R_k^{i-}$.

Die folgenden Bedingungen sind nach [Spa05], wurden aber in einigen Punkten geändert. Unter diesen gilt der nachfolgende Konvergenzsatz für das FDSA-Verfahren. Über die Änderungen gibt eine darauf folgende Bemerkung Auskunft.

Bedingungen an das FDSA-Verfahren (bei Verwendung symmetrischer finiter Differenzen)

B.1′ (Schrittweitenfolgen)
$a_k > 0, h_k > 0, a_k \to 0, h_k \to 0,$

$$\sum_{k=0}^{\infty} a_k = \infty \text{ und } \sum_{k=0}^{\infty} \frac{a_k^2}{h_k^2} < \infty. \tag{4.57}$$

B.2′ (Beziehung zu gewöhnlicher DGL)
Sei $\mathbf{g}$ stetig auf $\mathbb{R}^m$. Nehme an, dass die Differentialgleichung

$$\dot{\mathbf{x}} = -\mathbf{g}(\mathbf{x}) \tag{4.58}$$

einen asymptotisch stabilen Gleichgewichtspunkt in $\mathbf{x}(t) = \mathbf{x}^*$ hat.

B.3′ (Beschränktheit der Iterierten)
Es gelte $\sup_{k\geq 0} \|\mathbf{X}_k\| < \infty$ f.s., außerdem liege $\mathbf{X}_k$ unendlich oft in einer kompakten Teilmenge des Einzugsbereichs der DGL (4.58) aus **B.2′**.

B.4′ (Bedingungen an das Rauschen)
Es gelte

$$\mathbb{E}[R_k \,|\, \mathbf{X}_k] \eqsim 0 \quad \forall k \tag{4.59}$$

und R_k sei gleichmäßig varianzbeschränkt, d.h.

$$\mathbb{E}(R_k^2) \leq c \quad \forall k. \tag{4.60}$$

R_k^{i+} und R_k^{i-} seien unabhängig. $R_k^{i\pm}$ seien integrierbare Zufallsgrößen. $\mathbf{X}_0, R_k$, $k \geq 0$ seien unabhängig.

B.5′ (Glattheit von f)
Es sei $f \in \mathcal{C}^2$ und die partiellen dritten Ableitungen $f_i'''(\mathbf{x})$ existieren, seien stetig und beschränkt.

Bemerkung 4.34 (Veränderungen zu SPALL). Die erste Bedingung in **B.4′** lautet in [Spa05]

$$\mathbb{E}[R_k^{i+} - R_k^{i-} \,|\, \mathfrak{J}_k] \eqsim 0. \tag{4.61}$$

Da in dieser Arbeit eine andere Filtrierung verwendet wurde, ist die Bedingung auf das hier in den Beweisen Benötigte abgeändert. Die zweite Bedingung in **B.4′** lautet in [Spa05]

$$\mathbb{E}\big[(R_k^{i\pm})^2 \,|\, \mathfrak{J}_k\big] \leq c \text{ f.s.} \tag{4.62}$$

für ein c unabhängig von k. Damit dies wohldefiniert ist, muss $R_k^{i\pm} \in \mathcal{L}^2$ gelten. Für den in dieser Arbeit dargelegten Beweisgang hat sich gezeigt, dass zusätzlich statt (4.62) R_k gleichmäßig varianzbeschränkt zu fordern ist.

Außerdem wurde, um die Bedingungen des Modells zu erfüllen, die Forderung der Unabhängigkeit und Integrierbarkeit der $R_k^{i\pm}$ hinzugefügt. Auch $\mathbf{X}_0, R_k$, $k \geq 0$ unabhängig wurde hinzugefügt.

Bemerkung 4.35. Dass $u(\mathbf{X}_k, Z_k)$ $\mathcal{L}^1$-Zufallsgrößen sein sollen, damit die Bias-Definition 4.17 wohldefiniert ist, wird im Beweis des FDSA-Konvergenztheorems (Theorem 4.37) gezeigt. Alternativ gilt das auch direkt. Da $\mathbf{X}_k$ f.s. beschränkt und $h_k \to 0$, gibt es einen kompakten Hyperwürfel $\mathbb{H}$, so dass $\mathbf{X}_k \pm h_k \mathbf{e}_i \in \mathbb{H}$ f.s. $\forall k$, denn

$$\sup_k \|\mathbf{X}_k \pm h_k \mathbf{e}_i\| \leq \sup_k \|\mathbf{X}_k\| + |h_k| \cdot 1 = \sup_k \|\mathbf{X}_k\| + \sup_k h_k < \infty.$$

Die stetige Funktion f ist auf $\mathbb{H}$ beschränkt, d.h.

$$f(\mathbf{x}) \leq c \text{ für } \mathbf{x} \in \mathbb{H}.$$

Daher gilt:

$$\mathbb{E}\big(\,|u(\mathbf{X}_k, Z_k)|\,\big) = \frac{1}{2h_k}\mathbb{E}\big(\underbrace{|f(\mathbf{X}_k + h_k\mathbf{e}_i) - f(\mathbf{X}_k - h_k\mathbf{e}_i)|}_{\leq 2c \text{ f.s.}}\big) + \frac{1}{2h_k}\mathbb{E}\big(\,|R_k|\,\big).$$

Dieser Ausdruck endlich, da aus Lemma 3.45 folgt, dass $R_k \in \mathcal{L}^1$.

Konvergenz- und Bias-Lemma

Vor dem eigentlichen Konvergenzresultat folgt ein Lemma, das die Eigenschaft von $\hat{\mathbf{g}}_k^{\mathrm{FD}}$, ein Schätzer für den Gradienten zu sein, charakterisiert.

Lemma 4.36 (FDSA-Bias-Lemma)**.**
*Die dritten Ableitungen von f seien beschränkt. Nehme **B.4′** für die Messfehler $R_k^{i\pm}$ an. Laut Modell gelte $u(\mathbf{X}_k, Z_k) \in \mathcal{L}^1$. Mit $h_k \to 0$, $k \to \infty$ gilt dann*

$$\mathbf{B}_k(\mathbf{X}_k) = O(h_k^2),\ k \to \infty, \tag{4.63}$$

und $(\mathbf{B}_k(\mathbf{X}_k))_k$ ist f.s. beschränkt.

Wegen $h_k \to 0$, $k \to \infty$ ist dann $\hat{\mathbf{g}}_k^{\mathrm{FD}}$ ein asymptotisch unverzerrter Schätzer des Gradienten ∇f_k gemäß der angepassten Verzerrungsdefinition (Definition 4.17).

Beweis.

$$\begin{aligned}\mathbb{E}[\hat{\mathbf{g}}_{ki}(\mathbf{X}_k)\,|\,\mathbf{X}_k] &\doteq \mathbb{E}\left[\frac{f(\mathbf{X}_k + h_k\mathbf{e}_i) - f(\mathbf{X}_k - h_k\mathbf{e}_i)}{2h_k} + \frac{R_k^{i+} - R_k^{i-}}{2h_k}\,\middle|\,\mathbf{X}_k\right] \\ &\doteq \frac{f(\mathbf{X}_k + h_k\mathbf{e}_i) - f(\mathbf{X}_k - h_k\mathbf{e}_i)}{2h_k} + \underbrace{\mathbb{E}\left[\frac{R_k^{i+} - R_k^{i-}}{2h_k}\,|\,\mathbf{X}_k\right]}_{\doteq 0},\end{aligned}$$

da $f(\mathbf{X}_k \pm h_k\mathbf{e}_i)$ $\mathbf{X}_k$-messbar und nach $\mathbf{B.4'}$. Also gilt:

$$\mathbb{E}[\hat{\mathbf{g}}_k(\mathbf{X}_k)\,|\,\mathbf{X}_k] \doteq \hat{\mathbf{g}}_{h_k}^{\mathrm{GN2}} \circ \mathbf{X}_k, \tag{4.64}$$

d.h., aus Lemma 4.11 folgt

$$\mathbb{E}[\hat{\mathbf{g}}_k(\mathbf{X}_k) \,|\, \mathbf{X}_k] \doteq \nabla f(\mathbf{X}_k) + O(h_k^2),\ k \to \infty \tag{4.65}$$

und nach Lemma 3.35 für $\mathbf{g}(\mathbf{X}_k) = \nabla f(\mathbf{X}_k)$ bedeutet dies:

$$\mathbf{B}_k(\mathbf{X}_k) \doteq \mathbb{E}[\hat{\mathbf{g}}_k(\mathbf{X}_k) - \mathbf{g}(\mathbf{X}_k) \,|\, \mathbf{X}_k] \doteq O(h_k^2),\ k \to \infty. \tag{4.66}$$

Zur Beschränktheit: Der bedingte Erwartungswert $\mathbb{E}\big[u(\mathbf{X}_k, Z_k) \,|\, \mathbf{X}_k\big]$ ist wohldefiniert und damit endlich, da $u(\mathbf{X}_k, Z_k)$ integrierbar ist.
Wegen $\mathbf{B}_k \doteq \mathbb{E}\big[u(\mathbf{X}_k, Z_k) \,|\, \mathbf{X}_k\big] - \mathbf{g}(\mathbf{X}_k)$ ist daher $\mathbf{B}_k$ für jedes k fast sicher endlich. Mit $h_k \to 0$ gilt gleichzeitig $\mathbf{B}_k \to 0$ f.s., das heißt, $(\mathbf{B}_k)_k$ ist fast sicher beschränkt. □

Bemerkung. Dieses Resultat lässt sich entsprechend auch auf den Fall einseitiger finiter Differenzen erweitern, dann mit $\mathbf{B}_k(\mathbf{X}_k) = O(h_k)$.

Es folgt ein Konvergenzresultat für das FDSA-Verfahren. Nach der Idee von [Spa05, Theorem 6.1] wird dies auf das Kushner-Clark-Theorem zurückgeführt.[5]

Theorem 4.37 (Konvergenz des FDSA-Verfahrens)**.**
Das Problem (2.2) *habe eine eindeutige Lösung* $\mathbf{x}^*$*. Dann konvergiert das Verfahren* (2.9) *mit Gradientenschätzung* (4.24) *unter den Bedingungen* ***B.1′*** *–* ***B.5′*** *fast sicher, d.h.*

$$\mathbf{X}_k \xrightarrow{k\to\infty} \mathbf{x}^* \quad \textit{f.s.} \tag{4.67}$$

Beweis. Die Voraussetzungen des Kushner und Clark-Theorems 4.30 werden überprüft:

- $\mathbf{g}$ ist stetig, da $f \in \mathcal{C}^2$.
- $\mathbf{B}_k$ ist f.s. beschränkt und $\mathbf{B}_k \to 0, k \to \infty$ folgt aus dem Bias-Lemma für das FDSA-Verfahren mit $h_k \to 0, k \to \infty$ nach $\mathbf{B.1}'$.
- Die Bedingungen an a_k entsprechen denen in $\mathbf{B.1}'$.
- $\mathbf{X}_k$ ist f.s. beschränkt nach $\mathbf{B.3}'$.

Zu zeigen bleibt ((iv)). Dies folgt aus der Anwendung des Lemmas 4.32 (dies sichert $\mathbb{E}(\|\mathbf{E}_k\|^2) \leq ch_k^{-2}$) und des Lemmas 4.33 (Martingalbeschränkung des SA-Fehlerterms).

Zur Anwendung von Lemma 4.32 setzt man $U_k = u(\mathbf{X}_k, Z_k)$. T_k ist entsprechend $T_{ki} = \frac{f(\mathbf{X}_k + h_k\mathbf{e}_i) - f(\mathbf{X}_k - h_k\mathbf{e}_k)}{2} + \frac{R_k}{2}$ und ist $\mathcal{L}^2$-beschränkt, da

$$\begin{aligned}\mathbb{E}(|T_k|^2) = \mathbb{E}\left(\left|\frac{f(\mathbf{X}_k + h_k\mathbf{e}_i) - f(\mathbf{X}_k - h_k\mathbf{e}_i)}{2}\right|^2\right) + \mathbb{E}\left(\left|\frac{R_k}{2}\right|^2\right) \\ + \frac{1}{2}\mathbb{E}\Big(\,|f(\mathbf{X}_k + h_k\mathbf{e}_i) - f(\mathbf{X}_k - h_k\mathbf{e}_i)|\,|R_k|\,\Big).\end{aligned} \tag{4.68}$$

[5] In [Spa05, Theorem 6.1] wird auf Bedingungen in Kapitel 4 zurückgeführt, unter denen [Spa05, Theorem 4.1] gilt, welches wiederum auf das Kushner-Clark-Theorem verweist. Letztlich nimmt der in dieser Arbeit vorgestellte Beweisweg eine etwas andere Form an, da dieser Zwischenschritt weggelassen wird und die Beweiskette mit dem Beweis des SPSA-Verfahrens zusammengelegt wurde.

Wegen **B.3′** und $h_k \to 0$ (**B.1′**) gibt es einen abgeschlossenen Hyperquader $\mathbb{H}$, so dass $\mathbf{X}_k \pm h_k \mathbf{e}_i \in \mathbb{H}$ f.s. $\forall k$. Auf dem Kompaktum $\mathbb{H}$ nimmt die stetige Funktion f Minimum und Maximum an, also gibt es ein von k unabhängiges $c \in \mathbb{R}$ mit

$$|f(\mathbf{X}_k + h_k\mathbf{e}_i) - f(\mathbf{X}_k - h_k\mathbf{e}_i)| \leq |f(\mathbf{X}_k + h_k\mathbf{e}_i)| + |f(\mathbf{X}_k - h_k\mathbf{e}_i)| \overset{\text{f.s}}{\leq} 2c.$$

Also

$$\mathbb{E}(|f(\mathbf{X}_k + h_k\mathbf{e}_i) - f(\mathbf{X}_k - h_k\mathbf{e}_i)|^2) \leq 4c^2,$$

so dass der 1. Summand von (4.68) beschränkt ist, für den 2. Summanden von (4.68) gilt dies, da R_k $\mathcal{L}^2$-beschränkt ist und für den 3. Summanden beachte man, dass

$$\mathbb{E}\Big(|f(\mathbf{X}_k + h_k\mathbf{e}_i) - f(\mathbf{X}_k - h_k\mathbf{e}_i)|\, |R_k| \Big) \leq 2c\mathbb{E}(|R_k|)$$

und R_k nach Lemma 3.45 $\mathcal{L}^1$-beschränkt ist.

Lemma 4.33 wendet man an mit $Z_k = R_k^{\mathrm{FD2}}$ und $u(\mathbf{X}_k, Z_k) = \hat{\mathbf{g}}_k^{\mathrm{FD2}}(\mathbf{X}_k)$. $U_k \in \mathcal{L}^2$ gilt dabei wegen $U_k = \frac{1}{h_k} T_k$ und T_k $\mathcal{L}^2$-beschränkt. □

Bemerkung 4.38 (Deterministische Betrachtung des FDSA-Verfahrens). In der Bedingung **B.1′**, wie auch in der entsprechenden SPSA-Bedingung **B.1″**, wird $h_k \to 0$ gefordert. Im Abschnitt zu den ableitungsfreien Gradientenverfahren hat man gesehen, dass man, um einen bestmöglichen Ausgleich zwischen Rundungs- und Diskretisierungsfehlern zu erreichen, h^{GN} nicht zu klein wählen darf (vergleiche Bemerkung 4.15). Als relative Genauigkeit u kann man nun statt der Maschinengenauigkeit aus Lemma 4.14 eine obere Schranke r für das Rauschen in der Zielfunktion heranziehen. Hat man $|R_{\mathbf{x}}| \leq r$ für alle $\mathbf{x}$, d.h.

$$\left|\tilde{f}(\mathbf{x}) - f(x)\right| = |R_{\mathbf{x}}| \leq r \ \forall \mathbf{x}$$

für die verrauschte Zielfunktion, so ist der Fehler $\left\|\hat{\mathbf{g}}_k^{FD}(\mathbf{x}_k) - \nabla f(\mathbf{x}_k)\right\|$ in der folgenden Weise komponentenweise beschränkt ($\mathbf{g}_i$ sei die i-te Komponente des Gradienten $\nabla f(\mathbf{x})$):

$$\left|\hat{\mathbf{g}}_{k\,i}^{\mathrm{FD1}}(\mathbf{x}_k) - \mathbf{g}_i(\mathbf{x}_k)\right| \leq \tfrac{L^{(2)}}{2} h_k + 2\tfrac{r}{h_k} \text{ und} \tag{4.69}$$

$$\left|\hat{\mathbf{g}}_{k\,i}^{\mathrm{FD2}}(\mathbf{x}_k) - \mathbf{g}_i(\mathbf{x}_k)\right| \leq \tfrac{L^{(3)}}{6} h_k^2 + 2\tfrac{r}{h_k}. \tag{4.70}$$

Dabei setzt man wie in Lemma 4.11 voraus, dass f_i'' bzw. f_i''' beschränkt sind. Nimmt man an, dass durch die vorausgesetzte Wohlskaliertheit des Problems $L^{(2)}$ und $L^{(3)}$ in der Größenordnung 1 sind und ist r in der Größenordnung 1, so ergeben sich gute Wahlen für h analog zu Bemerkung 4.15 als

$$h^{\mathrm{FD1}} = \sqrt{r} \text{ und } h^{\mathrm{FD2}} = \sqrt[3]{r}. \tag{4.71}$$

Das Konvergenzresultat gemäß der *Stochastic-Approximation*-Theorie wird dabei nicht entkräftet, da es sich nur um eine obere Schranke handelt. Gleichwohl legt das hier Gesagte nahe, diese Probleme der numerischen Differentiation bei der Umsetzung des Verfahrens in einen Algorithmus durch eine entsprechende Wahl von h_k zu berücksichtigen.

Beweis. Die Aussage von (4.69) wird gezeigt.

$$\begin{aligned}
&\left|\hat{\mathbf{g}}_{ki}^{\mathrm{FD1}}(\mathbf{x}_k) - \mathbf{g}_i(\mathbf{x}_k)\right| \doteq \left|\frac{f(\mathbf{x}_k + h_k\mathbf{e}_i) - f(\mathbf{x}_k)}{h_k} - \mathbf{g}_i(\mathbf{x}) + \frac{R_{\mathbf{x}_k+h_k\mathbf{e}_i} - R_{\mathbf{x}_k}}{h_k}\right| \\
&\overset{\text{f.s}}{\leq} \left|\hat{\mathbf{g}}_{h_k\, i}^{\mathrm{GN1}}(\mathbf{x}_k) - \mathbf{g}_i(\mathbf{x}_k)\right| + \frac{1}{h_k}\big(\,|R_{\mathbf{x}_k+h_k\mathbf{e}_i}| + |R_{\mathbf{x}_k}|\,\big) \\
&\overset{\text{f.s}}{\leq} \frac{L^{(2)}}{2}h_k + 2\frac{r}{h_k},
\end{aligned}$$

wobei man die im Beweis von Lemma 4.11 angegebenen Schranken für $\hat{\mathbf{g}}_{h_k\, i}^{\mathrm{GN1}}(\mathbf{x}_k)$ ausnutzt. □

Bemerkung 4.39. Wählt man, wie in Bemerkung 4.38 erwähnt, $h = \text{const.}$, so ergibt sich aus diesem Bias-Lemma, dass $\hat{\mathbf{g}}_k^{\mathrm{FD2}}(\mathbf{X}_k)$ ein verzerrter Schätzer des Gradienten $\nabla f(\mathbf{X}_k)$ ist (im Sinne der angepassten Bias-Definition). Mit Blick auf Korollar 4.10 stellt dies an sich aber zunächst kein Hindernis für ein Gradientenverfahren dar. Das Konvergenzresultat gilt in der angegeben Form aber zunächst nicht mehr, da insbesondere die Bedingung (ii) des für den Beweis verwendeten KUSHNER-CLARK-Theorems (Theorem 4.30) nicht mehr erfüllt ist.

Genauso wie beim GSD-Verfahren werden hier für FD1 $m + 1$ Funktionsauswertungen für die Gradientenschätzung und pro Iteration benötigt. Für FD2 sind analog $2m$ Funktionsauswertungen für die Gradientenschätzung nötig und $2m + 1$ Funktionsauswertung pro Iteration.

Einordnung der Bedingungen für die Konvergenz

Bemerkung 4.40. An dieser Stelle werden die Voraussetzungen gewürdigt. Zu den Bedingungen soll vorausgreifend auf Bemerkung 4.50 verwiesen werden, die die Bedingungen des SPSA-Verfahrens einordnet. Die FDSA-Bedingungen sind – neben dem offensichtlichen Wegfallen der Bedingungen an die Perturbationen $\mathbf{D}_k$ – in zwei Punkten unwesentlich einfacher als die des SPSA-Verfahrens:

- Statt der $\mathcal{L}^2$-Beschränktheit von $\frac{\tilde{f}(\mathbf{X}_k \pm \overline{\mathbf{D}}_k)}{\mathbf{D}_k}$, wobei dafür bei BERNOULLI-verteilten Perturbationen $\mathbf{D}_k$ gemäß Bemerkung 4.45 hinreichend ist, dass $R_k^{\pm}$ und $f(\mathbf{X}_k \pm \overline{\mathbf{D}}_k)$ in $\mathcal{L}^2$-beschränkt sind, muss dies nur für $R_k^{\pm}$ gelten, was sehr plausibel erscheint. In der strikten Definition des Rauschens (Definition 2.1) wird die Varianz der das Rauschen modellierenden Zufallsgrößen sogar als konstant festgelegt.
- Hier werden die Bedingungen der Regularität nur an die ungemischten dritten Ableitungen f_i''' gestellt.

Das FDSA-Verfahren im rauschfreien Fall

Im deterministischen Fall ist $f = \tilde{f}$ und das FDSA-Verfahren wird zum GSD-Verfahren. Es stellt sich also die Frage, wie sich das FDSA- und GSD-Konvergenzresulat in diesem Fall zueinander verhalten.

Bemerkung 4.41. Das Konvergenzresultat des GSD-Verfahrens nach Bemerkung 4.12 ergibt sich nicht aus dem des FDSA-Verfahrens. Die Bedingung **B.4′** ist wegen der Rauschfreiheit erfüllt. Die Forderungen **B.3′** und **B.5′** des FDSA-Konvergenzresultats gehen über die des GSD-Verfahrens hinaus. Ein großer

Unterschied liegt auch darin, dass hier mit fixen Schrittweitenfolgen gearbeitet wird. Die WOLFE-Bedingungen werden also im Allgemeinen nicht erfüllt. Keines der beiden Resultate impliziert also das andere.

4.3.7 Konvergenzanalyse des *Simultaneous Perturbation Stochastic Approximation* Verfahrens

In diesem Abschnitt gilt die folgende Bezeichnung: $R_k \coloneqq R_k^{\mathrm{SP}} \equiv R_k^+ - R_k^-$.

In diesem Teil der Arbeit wird das SPALLsche Konvergenztheorem für das *Simultaneous Perturbation Stochastic Approximation* Verfahren hergeleitet. Dieses wurde in der Einleitung bereits als KIEFER-WOLFOWITZ-SA-Verfahren (2.9) mit Gradientenschätzung (2.11) eingeführt. Für den Beweis wird [Spa05] und [Spa92] gefolgt, der das Konvergenzresultat auf das KUSHNER-CLARK-Theorem zurückgeführt.

Die Idee, vom FDSA- bzw. KW-SA-Verfahren aus ableitungsfreie stochastische Gradientenverfahren zu entwicklen, die mit wesentlich weniger Funktionsauswertungen für die Gradientenschätzung auskommen, gibt es bereits seit einiger Zeit. Die erste Idee dazu scheint auf ERMOLIEV 1969 zurückzugehen, der ein SA-Verfahren mit *Random Directions* vorschlägt. Dies wird auch in [KC78] auf Konvergenz untersucht und die Effizienz der des FDSA-Verfahrens gegenübergestellt. Siehe dazu und zu den Unterschieden zum in dieser Arbeit vorgestellten SPSA-Verfahren [Spa05, Kapitel 6.8].

Auf der nächsten Seite werden die Konvergenzbedingungen angegeben.

Bedingungen an SPSA-Verfahren (bei Verwendung symmetrischer finiter Differenzen)

B.1″ (Schrittweitenfolgen) (wie **B.1′**)
$a_k > 0, h_k > 0, a_k \to 0, h_k \to 0, \sum_{k=0}^{\infty} a_k = \infty$ und $\sum_{k=0}^{\infty} \frac{a_k^2}{h_k^2} < \infty$.

B.2″ (Beziehung zu gewöhnlicher DGL) (wie **B.2′**)
Sei ∇f stetig auf $\mathbb{R}^m$. Nehme an, dass die Differentialgleichung

$$\dot{\mathbf{x}} = -\nabla f(\mathbf{x}) \tag{4.72}$$

einen asymptotisch stabilen Gleichgewichtspunkt in $\mathbf{x}^*$ hat.

B.3″ (Beschränktheit der Iterierten) (wie **B.3′**)
Es gelte $\sup_{k\geq 0} \|\mathbf{X}_k\| < \infty$ f.s., außerdem liege $\mathbf{X}_k$ unendlich oft in einer kompakten Teilmenge des Einzugsbereichs der DGL (4.72) aus **B.2″**.

B.4″ (Rauschen $R_k^{\pm}$ und sein Verhältnis zu den Perturbationen $\mathbf{D}_k$)
Für alle k sei

$$\mathbb{E}\big[R_k^+ - R_k^- \,|\, \mathbf{X}_k, \mathbf{D}_k\big] \doteq 0 \tag{4.73}$$

und Messung und Störung verhalten sich so, dass

$$\frac{\tilde{f}(\mathbf{X}_k \pm h_k \mathbf{D}_k)}{\mathbf{D}_k} \text{ in } \mathcal{L}^2 \text{ beschränkt} \tag{4.74}$$

ist.

B.5″ (Glattheit von f)
f sei dreimal stetig differenzierbar mit beschränkten dritten Ableitungen.

B.6″ (Statistische Eigenschaft der Perturbationen)
$\mathbf{D}_{ki}$ seien unabhängig für alle k, i und unabhängig von $\mathbf{X}_k$.
$\mathbf{D}_{ki}$, $i = 1, \ldots, m$ haben die gleiche, um 0 symmetrische Verteilung,
$|\mathbf{D}_k|$ sei beschränkt und $\frac{1}{\mathbf{D}_k}$ sei $\mathcal{L}^2$-beschränkt.

B.7″ (Zusammenhang der Zufallsgrößen)
$(R_k^-)_k, (R_k^+)_k$ seien unabhängige Folgen integrierbarer Zufallsgrößen, voneinander unabhängig und unabhängig von $(\mathbf{X}_k, \mathbf{D}_k)$.
$\mathbf{X}_0$ und $(R_k, \mathbf{D}_k)$, $k \geq 0$ seien unabhängige Zufallsgrößen.

Bemerkung 4.42. Im Vergleich zu den Bedingungen von Spall wurden die folgenden Änderungen vorgenommen:

- Die Filtrierung wurde geändert. In der hier vorgestellten Konvergenzanalyse reicht es z.B. aus, dass $\mathbf{D}_{ki}$ von $\mathbf{X}_k$ unabhängig sind, anstatt von $\mathfrak{J}_k$ wie in [Spa05] gefordert. Spall gibt auch an, dass $\mathbb{E}[\hat{\mathbf{g}}_k(\mathbf{X}_k) \,|\, \mathbf{X}_k] \doteq \mathbb{E}[\hat{\mathbf{g}}_k(\mathbf{X}_k) \,|\, \mathfrak{J}_k]$ [Spa05, S. 334], was diese Änderung als nicht besonders problematisch erscheinen lässt. Auch in (4.73) wurde die Bedingung von $\mathfrak{J}_k, \mathbf{D}_k$ bei Spall auf $\mathbf{X}_k, \mathbf{D}_k$ geändert.
- Die Bedingung $\mathbb{E}\Big(\big(\frac{\tilde{f}(\mathbf{X}_k \pm h_k \mathbf{D}_k)}{\mathbf{D}_{ki}}\big)^2\Big) \leq c \; \forall k, i$ in **B.4″** wurde als $\frac{\tilde{f}(\mathbf{X}_k \pm h_k \mathbf{D}_k)}{\mathbf{D}_k}$ sei $\mathcal{L}^2$-beschränkt eingefügt.

- $\frac{1}{\mathbf{D}_{ki}}$ sei $\mathcal{L}^2$-beschränkt wurde zu **B.6″** hinzugefügt. Dies wird für das SPSA-Konvergenztheorem benötigt. Erreicht man **B.4″** durch die hinreichende Bedingung im unten angegebenen Satz 4.43 ist dies ohnehin erfüllt.
- In **B.5″** wird in [Spa05] die Beschränktheit von f gefordert. Für Lemma 4.46 wird dann aber die Beschränktheit der dritten Ableitungen verwendet, so dass dies hier direkt so in **B.5″** aufgenommen wurde.

Zur Beschränktheit der Iterierten siehe Bemerkung 4.31.

Mit der HÖLDERschen Ungleichung (Satz 3.37) lässt sich die Bedingung **B.4″** in Bedingungen für die (wählbare) Verteilung und die durch die Aufgabe vorgegebene Zielfunktion aufteilen:

Satz 4.43. *Ist $\tilde{f}(\mathbf{X}_k \pm \overline{\mathbf{D}}_k)$ gleichmäßig $\mathcal{L}^{2p}$-beschränkt und $\frac{1}{\mathbf{D}_{ki}}$ gleichmäßig $\mathcal{L}^{2q}$-beschränkt, $p, q \in [1, \infty]$, $\frac{1}{p} + \frac{1}{q} = 1$, so gilt die Bedingung (4.74) (2. Teil von **B.4″**).*

Beweis. Nach der HÖLDERschen Ungleichung gilt:

$$\mathbb{E}\Big(\frac{\big(\tilde{f}(\mathbf{X}_k \pm \overline{\mathbf{D}}_k)\big)^2}{{\mathbf{D}_{ki}}^2}\Big) \quad \leq \Big(\mathbb{E}\Big(\big(\tilde{f}(\mathbf{X}_k \pm \overline{\mathbf{D}}_k)\big)^{2p}\Big)\Big)^{1/p} \Big(\mathbb{E}\Big(\big(\tfrac{1}{\mathbf{D}_{ki}}\big)^{2q}\Big)\Big)^{1/q}.$$

□

Bemerkung 4.44. Aus Bedingung **B.4″** folgt, dass die Bedingung

$$T_k \coloneqq \frac{\tilde{f}(\mathbf{X}_k + h_k \mathbf{D}_k) - \tilde{f}(\mathbf{X}_k - h_k \mathbf{D}_k)}{2\mathbf{D}_k} \text{ in } \mathcal{L}^2 \text{ beschränkt} \tag{4.75}$$

erfüllt ist. Beachte dazu, dass nach Lemma 3.48 Summen $\mathcal{L}^2$-beschränkter Zufallsgrößen wieder $\mathcal{L}^2$-beschränkt sind.

Wegen (4.75) ist auch die Gradientenschätzung $\hat{\mathbf{g}}^{\mathrm{SP}}$ integrierbar:
$U_k = u(\mathbf{X}_k, Z_k) = \frac{1}{h_k} T_k$ ist eine $\mathcal{L}^1$-Zufallsgröße, da T_k $\mathcal{L}^2$-beschränkt, also insbesondere integrierbar ist.

Bemerkung 4.45. Wählt man $\mathbf{D}_{ki}$ +1-−1-BERNOULLI-verteilt, d.h.

$$\mathbf{D}_{ki}(\omega) = \begin{cases} +1 & \text{mit Wahrscheinlichkeit } \frac{1}{2} \\ -1 & \text{mit Wahrscheinlichkeit } \frac{1}{2}, \end{cases} \tag{4.76}$$

dann ist zunächst $\mathbf{D}_{ki} = \frac{1}{\mathbf{D}_{ki}}$ und $\frac{1}{\mathbf{D}_{ki}}$ in $\mathcal{L}^q$ beschränkt für $q \in [1, \infty]$. Das heißt, für das Erfülltsein von (4.74) reicht

$$\tilde{f}(\mathbf{X}_k \pm \overline{\mathbf{D}}_k) \text{ gleichmäßig varianzbeschränkt,} \tag{4.77}$$

und gleichzeitig ist **B.6″** erfüllt. In diesem Fall ist hinreichend für (4.77), dass $R_k^{\pm}$ und $f(\mathbf{X}_k \pm \overline{\mathbf{D}}_k)$ in $\mathcal{L}^2$ beschränkt sind. Durch Lemma 3.48 ist (4.77) dann wieder erfüllt.

Beweis. (1) Es wird gezeigt, dass $\frac{1}{\mathbf{D}_{ki}}$ für alle $q \in [1, \infty]$ $\mathcal{L}^q$-beschränkt ist:
(i) $\mathbb{E}\Big(\Big|\frac{1}{\mathbf{D}_{ki}}\Big|^q\Big) = 1$, da $\Big|\frac{1}{\mathbf{D}_{ki}}\Big| = 1$.
(ii) Für $q = \infty$. $\operatorname{ess\,sup}\Big|\frac{1}{\mathbf{D}_{ki}}\Big| = \sup\Big\{c \geq 0 : \mathbb{P}\Big(\Big|\frac{1}{\mathbf{D}_{ki}}\Big| \geq c\Big) > 0\Big\} = 1$, da $\Big|\frac{1}{\mathbf{D}_{ki}}\Big| = 1$ und damit $\mathbb{P}\Big(\Big|\frac{1}{\mathbf{D}_{ki}}\Big| > c\Big) = 0 \ \forall c > 1$.

(2) Um (4.74) aus (4.77) zu folgern, benutzt man Satz 4.43 mit $q = \infty$ und $p = 1$.

□

Konvergenz- und Bias-Lemma

Als ersten Schritt zur Herleitung der Konvergenz des SPSA-Verfahrens kontrolliert man die Verzerrung des Gradientenschätzers $\hat{\mathbf{g}}_k$. Das folgende Lemma, eine Variante von Lemma 1 aus [Spa92], garantiert, dass $\hat{\mathbf{g}}_k$ asymptotisch erwartungstreuer Schätzer für $\nabla f(\mathbf{X}_k)$ ist, gemäß der angepassten Definition mit Bedingung auf $\mathbf{X}_k$.

Lemma 4.46 (Bias-Lemma)**.**
Es existiere ein $\overline{k} \in \mathbb{N}$, so dass für alle $k \geq \overline{k}$ gilt: $\{\mathbf{D}_{ki}\}_{\substack{i=1,\dots,m \\ k=1,2,\dots}}$ seien unabhängige Zufallsgrößen und

(i) $\{\mathbf{D}_{ki}\}_{(i=1,\dots,m)}$ *haben die gleiche, um* 0 *symmetrische Verteilung,*

(ii) $|\mathbf{D}_{ki}| \leq \alpha_0$ *f.s. sowie*

(iii) $\mathbb{E}\left(\left|{\mathbf{D}_{ki}}^{-1}\right|\right) \leq \alpha_1$.

*f sei dreimal stetig differenzierbar mit beschränkten dritten Ableitungen (Bedingung **B.5″**).*[6] *Man nehme **B.4″** für die Messfehler $R_k^{\pm}$ an und außerdem **B.7″** sowie, dass $\mathbf{X}_k$ und $\mathbf{D}_k$ unabhängig sind für jedes k.*[7] *Dann gilt fast sicher:*

$$\|\mathbf{B}_k(\mathbf{X}_k)\| < \infty \quad \forall k \geq \overline{k} \quad \textit{und} \quad \mathbf{B}_k(\mathbf{X}_k) = O(h_k^2) \quad \textit{für } k \to \infty.$$

Bemerkung 4.47. Das Lemma gilt unter den Voraussetzungen **B.1″** mit **B.4″** – **B.7″**. Dabei folgt (iii) aus Lemma 3.45, da $\frac{1}{\mathbf{D}_k}$ $\mathcal{L}^2$-beschränkt ist. Die weiteren Bedingungen an $\mathbf{D}_{ki}$ stehen genauso in **B.6″**.

Wegen der Forderung $h_k \to 0$, $k \to \infty$ in **B.1″** folgt, dass $\hat{\mathbf{g}}_k^{\mathrm{SP}}$ ein asymptotisch unverzerrter Schätzer gemäß der angepassten Verzerrungsdefinition (Definition 4.17) ist.

Beweis. Sei $l \in \{1, \dots, m\}$. Wegen der Linearität des bedingten Erwartungswerts gilt:

$$\mathbb{E}\Big[\tfrac{R_k^+ - R_k^-}{2h_k \mathbf{D}_{kl}} \,\Big|\, \mathbf{X}_k\Big] \doteq \frac{1}{2h_k}\Big(\mathbb{E}\Big[\tfrac{R_k^+}{\mathbf{D}_{kl}} \,\Big|\, \mathbf{X}_k\Big] - \mathbb{E}\Big[\tfrac{R_k^-}{\mathbf{D}_{kl}} \,\Big|\, \mathbf{X}_k\Big]\Big).$$

Da $\frac{1}{\mathbf{D}_{kl}} \in \mathcal{L}^1$, ist $\frac{1}{\mathbf{D}_{kl}} \neq 0$ nach Hilfssatz 3.13. Nach Hilfssatz 3.27 folgt dann aus der Unabhängigkeit von R_k^+ und $(\mathbf{D}_{kl}, \mathbf{X}_k)$ die von R_k^+ und $(\frac{1}{\mathbf{D}_{kl}}, \mathbf{X}_k)$ (analog für R_k^-). Zur Notation $(\mathbf{D}_k, \mathbf{X}_k)$ beachte man Lemma 3.19. Durch Anwenden

[6] In [Spa92] wird von fast überall beschränkten dritten Ableitungen gesprochen, da diese aber stetig sind, folgt die Gültigkeit überall statt nur fast überall.

[7] Die Modellierungsvoraussetzungen **B.7″** und $(\mathbf{X}_k, \mathbf{D}_k)$ unabhängig $\forall k$ wurden hinzugefügt.

von Lemma 3.34(i) erhält man:

$$\begin{aligned}\mathbb{E}\Big[\tfrac{R_k^+-R_k^-}{2h_k\mathbf{D}_{kl}}\,\Big|\,\mathbf{X}_k\Big] &\doteq \frac{1}{2h_k}\Big(\mathbb{E}\big(R_k^+\big)\mathbb{E}\big[\tfrac{1}{\mathbf{D}_{kl}}\,|\,\mathbf{X}_k\big]-\mathbb{E}\big(R_k^-\big)\mathbb{E}\big[\tfrac{1}{\mathbf{D}_{kl}}\,|\,\mathbf{X}_k\big]\Big)\\ &\doteq \frac{1}{2h_k}\underbrace{\mathbb{E}\big(R_k^+-R_k^-\big)}_{=0}\underbrace{\mathbb{E}\Big[\tfrac{1}{\mathbf{D}_{kl}}\,\Big|\,\mathbf{X}_k\Big]}_{\substack{\text{f.s.}\\ <\infty \text{ n. Def.}}} \doteq 0.\end{aligned}$$

Die letzte Gleichheit folgt daraus, dass Bedingung **B.4″**, $\mathbb{E}[R_k^+ - R_k^- \,|\, \mathbf{X}_k, \mathbf{D}_k] \doteq 0$, vermöge Lemma 3.33 $\mathbb{E}(R_k^+ - R_k^-) = 0$ impliziert. Weiter gilt:

$$\begin{aligned}\mathbf{B}_{kl}(\mathbf{X}_k) &\doteq \mathbb{E}\Big[\hat{\mathbf{g}}_{kl}(\mathbf{X}_k)-\mathbf{g}_l(\mathbf{X}_k)\Big|\mathbf{X}_k\Big]\\ &\doteq \mathbb{E}\left[\frac{\tilde f(\mathbf{X}_k+h_k\mathbf{D}_k)-\tilde f(\mathbf{X}_k-h_k\mathbf{D}_k)}{2h_k\mathbf{D}_{kl}}-\mathbf{g}_l(\mathbf{X}_k)\middle|\mathbf{X}_k\right]\\ &\overset{\text{Lin.}}{\doteq} \mathbb{E}\left[\frac{f(\mathbf{X}_k+h_k\mathbf{D}_k)-f(\mathbf{X}_k-h_k\mathbf{D}_k)}{2h_k\mathbf{D}_{kl}}-\mathbf{g}_l(\mathbf{X}_k)\middle|\mathbf{X}_k\right]+\underbrace{\mathbb{E}\left[\frac{R_k^+-R_k^-}{2h_k\mathbf{D}_k}\middle|\mathbf{X}_k\right]}_{\doteq 0 \text{ nach Schritt 1}}.\end{aligned}$$

Durch Entwickeln nach Taylor um $\mathbf{X}_k$ (siehe Satz A.5 mit Beschreibung der Multiindexnotation), erhält man mit auf der Verbindungsstrecke von $\mathbf{X}_k$ und $\mathbf{X}_k \pm h_k\mathbf{D}_k$ liegenden $\widetilde{\mathbf{X}}^{\pm}$:

$$\begin{aligned}\mathbf{B}_{kl}(\mathbf{X}_k) &\doteq \mathbb{E}\left[\frac{f(\mathbf{X}_k)+\nabla f(\mathbf{X}_k)^T(h_k\mathbf{D}_k)+\frac12 h_k^2{\mathbf{D}_k}^T\nabla^2 f(\mathbf{X}_k)\mathbf{D}_k}{2h_k\mathbf{D}_{kl}}\right.\\ &\quad-\frac{f(\mathbf{X}_k)-\nabla f(\mathbf{X}_k)^T(h_k\mathbf{D}_k)+\frac12 h_k^2{\mathbf{D}_k}^T\nabla^2 f(\mathbf{X}_k)\mathbf{D}_k}{2h_k\mathbf{D}_{kl}}\\ &\quad\left.+\frac{\sum_{|\mathbf{j}|=3}\frac{1}{\mathbf{j}!}D^{\mathbf{j}}f(\widetilde{\mathbf{X}}^+)(h_k\mathbf{D}_k)^{\mathbf{j}}}{2h_k\mathbf{D}_{kl}}-\frac{\sum_{|\mathbf{j}|=3}\frac{1}{\mathbf{j}!}D^{\mathbf{j}}f(\widetilde{\mathbf{X}}^-)(-h_k\mathbf{D}_k)^{\mathbf{j}}}{2h_k\mathbf{D}_{kl}}-\mathbf{g}_l(\mathbf{X}_k)\,\middle|\,\mathbf{X}_k\right]\\ &\doteq \mathbb{E}\left[\frac{\nabla f(\mathbf{X}_k)^T\mathbf{D}_k}{\mathbf{D}_{kl}}-\mathbf{g}_l(\mathbf{X}_k)+\text{ Terme 3. Ordnung }\,\middle|\,\mathbf{X}_k\right]\\ &\doteq \underbrace{\mathbb{E}\Big[\tfrac{\nabla f(\mathbf{X}_k)^T\mathbf{D}_k}{\mathbf{D}_{kl}}-\mathbf{g}_l(\mathbf{X}_k)\,\Big|\,\mathbf{X}_k\Big]}_{=:s_1}+\underbrace{\mathbb{E}[\text{Terme 3. Ordnung}\,|\,\mathbf{X}_k]}_{=:s_2}.\end{aligned}$$

Für den ersten Summanden s_1 gilt:

$$\begin{aligned}s_1 &\doteq \mathbb{E}\Big(\tfrac{\mathbf{g}_1(\mathbf{X}_k)\mathbf{D}_{k1}+\ldots+\mathbf{g}_l(\mathbf{X}_k)\mathbf{D}_{kl}+\ldots+\mathbf{g}_l(\mathbf{X}_k)\mathbf{D}_{km}}{\mathbf{D}_{kl}}-\mathbf{g}_l(\mathbf{X}_k)\,\Big|\,\mathbf{X}_k\Big)\\ &\doteq \mathbb{E}\Big(\underbrace{\mathbf{g}_l(\mathbf{X}_k)\tfrac{\mathbf{D}_{kl}}{\mathbf{D}_{kl}}-\mathbf{g}_l(\mathbf{X}_k)}_{\doteq 0}\,\Big|\,\mathbf{X}_k\Big)+\sum_{\substack{i\neq l\\ i=1,\ldots,m}}\mathbb{E}\Big(\tfrac{\mathbf{g}_l(\mathbf{X}_k)\mathbf{D}_{ki}}{\mathbf{D}_{kl}}\,\Big|\,\mathbf{X}_k\Big).\end{aligned}$$

Verwende Lemma 3.34 Regel (ii) mit $f = \mathbf{g}_l$, $Z = \mathbf{X}_k$, $X = \mathbf{D}_k$.

Aus Voraussetzung (iii) folgt $X \neq 0$ (mit Hilfssatz 3.13). Dann wählt man eine Version von X, die ungleich 0 ist nach Bemerkung 3.14.

$$s_1 \doteq \sum_{i\neq l} \underbrace{\mathbb{E}[\mathbf{g}_i(\mathbf{X}_k)\,|\,\mathbf{X}_k]}_{<\infty \text{ n. Def.}} \mathbb{E}\Big(\frac{\mathbf{D}_{ki}}{\mathbf{D}_{kl}}\Big).$$

Da $\mathbf{D}_{ki}$ und $\mathbf{D}_{kl}$ unabhängig sind ($i \neq l$), folgt

$$s_1 \doteq c \sum_{i\neq l} \mathbb{E}\Big(\frac{\mathbf{D}_{ki}}{\mathbf{D}_{kl}}\Big) = c \sum_{i\neq l} \underbrace{\mathbb{E}(\mathbf{D}_{ki})}_{=0} \underbrace{\mathbb{E}(\tfrac{1}{\mathbf{D}_{kl}})}_{<\infty} = 0$$

nach Voraussetzung (iii) und da $\mathbf{D}_{ki}$ symmetrisch verteilt ist. Damit ist der erste Summand s_1 in der TAYLOR-Entwicklung 0 und $\mathbf{B}_{kl} = s_2$. Man betrachtet nun also den auf $\mathbf{X}_k$ bedingten Erwartungswert der sich aus der oben genannten TAYLOR-Entwicklung ergebenden Terme dritter Ordnung.

Hinreichend für $\mathbf{B}_{kl} \in O(h_k^2)$ f.s., $k \to \infty$ ist $|\mathbf{B}_{kl}| \leq c\, h_k^2$ (vergleiche Bemerkung A.2). Dies soll im Folgenden gezeigt werden.

Verwendet wird Regel (vii) aus Lemma 3.34.

$$|\mathbf{B}_{kl}| = \left|\mathbb{E}\left[\frac{\sum_{|\mathbf{j}|=3} \frac{1}{\mathbf{j}!} D^{\mathbf{j}} f(\widetilde{\mathbf{X}}^+)(h_k\mathbf{D}_k)^{\mathbf{j}} - \sum_{|\mathbf{j}|=3} \frac{1}{\mathbf{j}!} D^{\mathbf{j}} f(\widetilde{\mathbf{X}}^-)(-h_k\mathbf{D}_k)^{\mathbf{j}}}{2h_k\mathbf{D}_{kl}} \,\Big|\, \mathbf{X}_k\right]\right|$$
$$\leq \mathbb{E}\left[\left|\frac{\sum_{|\mathbf{j}|=3} \frac{1}{\mathbf{j}!} D^{\mathbf{j}} f(\widetilde{\mathbf{X}}^+)(h_k\mathbf{D}_k)^{\mathbf{j}} - \sum_{|\mathbf{j}|=3} \frac{1}{\mathbf{j}!} D^{\mathbf{j}} f(\widetilde{\mathbf{X}}^-)(-h_k\mathbf{D}_k)^{\mathbf{j}}}{2h_k\mathbf{D}_{kl}}\right| \,\Big|\, \mathbf{X}_k\right] =: \mathbb{E}[I\,|\,\mathbf{X}_k].$$

Für den Term I gilt:

$$I \leq \Big|\sum_{|\mathbf{j}|=3} D^j f(\widetilde{\mathbf{X}}^+) \tfrac{(h_k\mathbf{D}_k)^{\mathbf{j}}}{2h_k\mathbf{D}_{kl}\mathbf{j}!}\Big| + \Big|\sum_{|\mathbf{j}|=3} D^j f(\widetilde{\mathbf{X}}^-) \tfrac{(-h_k\mathbf{D}_k)^{\mathbf{j}}}{2h_k\mathbf{D}_{kl}\mathbf{j}!}\Big|$$
$$\leq \sum_{|\mathbf{j}|=3} \tfrac{1}{\mathbf{j}!} \underbrace{\Big|D^j f(\widetilde{\mathbf{X}}^+)\Big|}_{\leq\alpha_2} \underbrace{\Big|\tfrac{(h_k\mathbf{D}_k)^{\mathbf{j}}}{2h_k\mathbf{D}_{kl}}\Big|}_{=\left|\frac{h_k^{\mathbf{j}}}{2h_k}\right|\left|\frac{\mathbf{D}_k{}^{\mathbf{j}}}{\mathbf{D}_{kl}}\right|} + \sum_{|\mathbf{j}|=3} \tfrac{1}{\mathbf{j}!} \underbrace{\Big|D^j f(\widetilde{\mathbf{X}}^-)\Big|}_{\leq\alpha_2} \Big|\tfrac{(-h_k\mathbf{D}_k)^{\mathbf{j}}}{2h_k\mathbf{D}_{kl}}\Big|.$$

Dabei wurde die Regularitätsbedingung an f (Voraussetzung **B.5″**) benutzt. Das folgende Argument wurde unter Zuhilfenahme von [Msea] entwickelt. Wegen $h_k^{\mathbf{j}} = h_k^{j_1} \cdots h_k^{j_m} = h_k^{j_1+\cdots+j_m} = h_k^{|\mathbf{j}|}$ und mit $|\mathbf{j}| = 3$ ist $\left|\frac{(\pm h_k)^{\mathbf{j}}}{2h_k}\right| = \left|\frac{\pm h_k^3}{2h_k}\right| = \frac{1}{2}h_k^2$, und für den Term I ergibt sich:

$$I \leq \sum_{|\mathbf{j}|=3} \alpha_2 \tfrac{1}{2} h_k^2 \tfrac{1}{\mathbf{j}!} \Big|\tfrac{\mathbf{D}_k{}^{\mathbf{j}}}{\mathbf{D}_{kl}}\Big| + \sum_{|\mathbf{j}|=3} \alpha_2 \tfrac{1}{2} h_k^2 \tfrac{1}{\mathbf{j}!} \Big|\tfrac{\mathbf{D}_k{}^{\mathbf{j}}}{\mathbf{D}_{kl}}\Big| = h_k^2 \alpha_2 \sum_{|\mathbf{j}|=3} \tfrac{1}{\mathbf{j}!} \Big|\tfrac{\mathbf{D}_k{}^{\mathbf{j}}}{\mathbf{D}_{kl}}\Big|.$$

Daher gilt für den Betrag des Bias-Terms $\mathbf{B}_{kl}$:

$$|\mathbf{B}_{kl}| \leq \mathbb{E}\left[h_k^2 \alpha_2 \sum_{|\mathbf{j}|=3} \tfrac{1}{\mathbf{j}!} \Big|\tfrac{\mathbf{D}_k{}^{\mathbf{j}}}{\mathbf{D}_{kl}}\Big| \,\Big|\, \mathbf{X}_k\right] \doteq h_k^2 \alpha_2 \sum_{|\mathbf{j}|=3} \tfrac{1}{\mathbf{j}!} \mathbb{E}\left[\Big|\tfrac{\mathbf{D}_k{}^{\mathbf{j}}}{\mathbf{D}_{kl}}\Big| \,\Big|\, \mathbf{X}_k\right] =: h_k^2 \alpha_2 S.$$

Zu zeigen bleibt, dass die Summe S beschränkt ist. Dazu partioniert man sie gemäß

$$\sum_{|\mathbf{j}|=3} \frac{1}{\mathbf{j}!} \mathbb{E}\left[\Big|\tfrac{\mathbf{D}_k{}^{\mathbf{j}}}{\mathbf{D}_{kl}}\Big| \,\Big|\, \mathbf{X}_k\right] = \sum_{\substack{|\mathbf{j}|=3 \\ j_l=0}} \frac{1}{\mathbf{j}!} \mathbb{E}\left[\Big|\tfrac{\mathbf{D}_k{}^{\mathbf{j}}}{\mathbf{D}_{kl}}\Big| \,\Big|\, \mathbf{X}_k\right] + \sum_{\substack{|\mathbf{j}|=3 \\ j_l\neq 0}} \frac{1}{\mathbf{j}!} \mathbb{E}\left[\Big|\tfrac{\mathbf{D}_k{}^{\mathbf{j}}}{\mathbf{D}_{kl}}\Big| \,\Big|\, \mathbf{X}_k\right].$$

Man schreibt nun $\mathbf{D}_k{}^{\mathbf{j}} = \mathbf{D}_{k\,1}{}^{\mathbf{j}_1} \cdots \mathbf{D}_{k\,m}{}^{\mathbf{j}_m}$ für $|\mathbf{j}| = 3$ als $\mathbf{D}_{k\,i_1}\mathbf{D}_{k\,i_2}\mathbf{D}_{k\,i_3}$, da höchstens drei und per Wahl auch genau drei j_i verschieden von 0 sind. Für $j_i = 3$ z.B. ist $\mathbf{D}_{ki}{}^{j_i} = \mathbf{D}_{ki_1}\mathbf{D}_{ki_2}\mathbf{D}_{ki_3}$ für $i_1 = i_2 = i_3 \coloneqq i$ und $j_{i_1} = j_{i_2} = j_{i_3} = 1$. Es gilt

$$\frac{\mathbf{D}_k{}^{\mathbf{j}}}{\mathbf{D}_{kl}} = \frac{\mathbf{D}_{ki_1}\mathbf{D}_{ki_2}\mathbf{D}_{ki_3}}{\mathbf{D}_{kl}} = \begin{cases} \frac{\mathbf{D}_{ki_1}\mathbf{D}_{ki_2}\mathbf{D}_{ki_3}}{\mathbf{D}_{kl}}, \quad i_1, i_2, i_3 \neq l & \text{falls } j_l = 0 \\ \mathbf{D}_{ki_1}\mathbf{D}_{ki_2}, \quad i_1, i_2 \neq l & \text{falls } j_l = 1 \\ \mathbf{D}_{kl}\mathbf{D}_{ki_1}, \quad i_1 \neq l & \text{falls } j_l = 2 \\ \mathbf{D}_{kl}{}^2 & \text{falls } j_l = 3\,, \end{cases} \tag{4.78}$$

also

$$\mathbb{E}\left[\left|\tfrac{\mathbf{D}_k{}^{\mathbf{j}}}{\mathbf{D}_{kl}}\right| \,\middle|\, \mathbf{X}_k\right] \eqsim \begin{cases} \mathbb{E}\Big[\Big|\frac{\mathbf{D}_{ki_1}\mathbf{D}_{ki_2}\mathbf{D}_{ki_3}}{\mathbf{D}_{kl}}\Big| \,\Big|\, \mathbf{X}_k\Big] & \leq \alpha_0^3\alpha_1, \text{ falls } j_l = 0, \\ \mathbb{E}\Big[\Big|\mathbf{D}_{ki_1}\mathbf{D}_{ki_2}\Big| \,\Big|\, \mathbf{X}_k\Big] & \leq \alpha_0^2, \text{ falls } j_l \neq 0. \end{cases}$$

Dabei entspricht in der Fallunterscheidung $j_k = 0$ der Forderung $i_1, i_2, i_3 \neq l$ für den 1. Fall, während im 2. Fall i_1 und i_2 beliebig sind. Die zweite Ungleichung gilt wegen der Monotonie des bedingten Erwartungswerts nach Lemma 3.32(iv) mit $|\mathbf{D}_{ki}|^2 \leq \alpha_0^2$ f.s. In der ersten Ungleichung beachte man

$$\begin{aligned} &\mathbb{E}\Big[\Big|\mathbf{D}_{ki_1}\mathbf{D}_{ki_2}\mathbf{D}_{ki_3}\tfrac{1}{\mathbf{D}_{kl}}\Big| \,\Big|\, \mathbf{X}_k\Big] \eqsim \mathbb{E}\Big[\,|\mathbf{D}_{ki_1}\mathbf{D}_{ki_2}\mathbf{D}_{ki_3}|\,\Big|\tfrac{1}{\mathbf{D}_{kl}}\Big| \,\Big|\, \mathbf{X}_k\Big] \\ &\leq \alpha_0^3\mathbb{E}\Big[\Big|\tfrac{1}{\mathbf{D}_{kl}}\Big| \,\Big|\, \mathbf{X}_k\Big] \eqsim \alpha_0^3\mathbb{E}\Big(\Big|\tfrac{1}{\mathbf{D}_{kl}}\Big|\Big) \leq \alpha_0^3\alpha_1 \end{aligned}$$

nach Korollar 3.22(iv).
Die Anzahl der Summanden ergeben sich wie folgt (dieses Argument wurde unter Zuhilfenahme von [Mseb] entwickelt):

- $j_l = 0$: „Ziehe“ $k = 3$ Indices aus $n-1$ möglichen, $\binom{n+1}{3}$ Kombinationen (mit Wiederholung).
- $j_l \neq 0$: „Ziehe“ $k = 2$ Indices aus n möglichen, entspricht $\binom{n+1}{2}$ Möglichkeiten. Alternativ zerlegt man dies nach $j_l = 1$ ($k = 2$ aus $n-1$ möglichen, $\binom{n}{2}$), $j_l = 2$ ($k = 1$ aus $n-1$ möglichen, $\binom{n-1}{1}$) und einem Element bei der zu $j_l = 3$ gehörigen Menge (Kombination mit Wiederholung, $\binom{n+k-1}{k}$).

Daher gibt es im Fall $j_l = 0$ $\binom{n+1}{3}$ Summanden, für $j_l \neq 0$ $\binom{n+1}{2}$ Summanden. Also folgt für die partionierte Summe:

$$\begin{aligned} S &= \sum_{\substack{1\leq i_1,i_2,i_3\leq m \\ i_1\neq l, i_2\neq l, i_3\neq l}} \tfrac{1}{\mathbf{j}!}\, \underbrace{\mathbb{E}\Big[\Big|\mathbf{D}_{ki_1}\mathbf{D}_{ki_2}\mathbf{D}_{ki_3}\tfrac{1}{\mathbf{D}_{kl}}\Big| \,\Big|\, \mathbf{X}_k\Big]}_{\leq \alpha_0^3\alpha_1} + \sum_{1\leq i_1,i_2\leq m} \tfrac{1}{\mathbf{j}!}\, \underbrace{\mathbb{E}\Big[\Big|\mathbf{D}_{ki_1}\mathbf{D}_{ki_2}\Big| \,\Big|\, \mathbf{X}_k\Big]}_{\leq \alpha_0^2}, \\ &\leq \underbrace{\Big(-\tfrac{1}{6}m + \tfrac{m^3}{6}\Big)}_{=:C_1}\, \alpha_0^3\alpha_1 + \underbrace{\Big(\tfrac{m(m+1)}{2}\Big)}_{=:C_2}\, \alpha_0^2. \end{aligned} \tag{4.79}$$

Man erhält:

$$|\mathbf{B}_{kl}(\mathbf{X}_k)| \leq h_k^2\alpha_2\Big(C_1\,\alpha_0^3\alpha_1 + C_2\,\alpha_0^2\Big) \text{ f.s.} \tag{4.80}$$

und $\mathbf{B}_{kl} < \infty$ für alle $k \geq K$. □

Der folgende Satz ist an Satz 1 aus [Spa92] angelehnt:

Theorem 4.48 (SPALL's SPSA-Konvergenzsatz)**.**
Es gelten alle Voraussetzungen des Lemmas 4.46 sowie die folgenden:

A1 *Es gelte* $a_k > 0, h_k > 0, a_k \to 0, h_k \to 0, \sum_{k=0}^{\infty} a_k = \infty$ *und* $\sum\limits_{k=0}^{\infty} \frac{a_k^2}{h_k^2} < \infty$.

A2 *Es gelte*

$$\mathbb{E}[R_k^+ - R_k^- \mid \mathbf{X}_k, \mathbf{D}_k] \eqsim 0 \text{ für alle } k.$$

$\frac{\tilde{f}(\mathbf{X}_k \pm h_k \mathbf{D}_k)}{\mathbf{D}_k}$ *seien* $\mathcal{L}^2$*-beschränkt, das heißt,* (4.74) *gilt.*

A3 *Es gelte* $\sup\limits_{k \geq 0} \|\mathbf{X}_k\| < \infty$ *f.s.*

A4 *Nehme an, dass die Differentialgleichung* $\dot{\mathbf{x}} = -\nabla f(\mathbf{x})$ *einen asymptotisch stabilen Gleichgewichtspunkt* $\mathbf{x}^*$ *hat.*

A5 *Der Einzugsbereich der DGL aus A4 habe eine kompakte Teilmenge* $\mathbb{S}$*, so dass* $\mathbf{X}_k \in \mathbb{S}$ *unendlich oft für fast alle* $\omega \in \Omega$.

Es gelte ***B.7′****. Dann gilt:*

$$\mathbf{X}_k \xrightarrow[f.s.]{k \to \infty} \mathbf{x}^*. \tag{4.81}$$

Bemerkung. Das Theorem gilt insbesondere unter den Voraussetzungen **B.1″** bis **B.7″**.

Beweis der Bemerkung. Nach Bemerkung 4.47 sind die Voraussetzungen von Lemma 4.46 erfüllt.

- A1 entspricht **B.1″**.
- A2 entspricht **B.4″** und einer Forderung aus **B.6″**.
- A3 und A5 entsprechen **B.3″**.
- A4 entspricht **B.2″**.

□

Bemerkung. Im Vergleich zu den Bedingungen in [Spa92] wurde hier in A2 für die Bedingung an das Rauschen der bedingte Erwartungswert gegeben $\mathbf{X}_k, \mathbf{D}_k$ statt $\mathfrak{J}_k, \mathbf{D}_k$ verwendet. Die Forderung wurde als $\mathcal{L}^2$-Beschränktheit von $\frac{1}{\mathbf{D}_k}$ aufgenommen.

Die Bedingungen in [Spa92],

$$\begin{aligned} \mathbb{E}({\mathbf{D}_{ki}}^{-2}) &\leq \alpha_2 \, \forall k, \\ \mathbb{E}\big((R_k^{\pm})^2\big) &\leq \alpha_0 \, \forall k \text{ und} \\ \mathbb{E}\Big(\big(f(\mathbf{X}_k \pm h_k \mathbf{D}_k)\big)^2\Big) &\leq \alpha_1 \, \forall k \end{aligned}$$

würden, wenn man sie durch die Forderung der $\mathcal{L}^2$-Beschränktheit von $\frac{R_k}{\mathbf{D}_k}$ und $\frac{f(\mathbf{X}_k \pm h_k \mathbf{D}_k)}{\mathbf{D}_k}$ ersetzt, die (4.74) entsprechende Bedingungaus A2 liefern, da die

Summe $\frac{\tilde{f}(\mathbf{X}_k \pm h_k \mathbf{D}_k)}{\mathbf{D}_k} = \frac{f(\mathbf{X}_k \pm h_k \mathbf{D}_k)}{\mathbf{D}_k} + \frac{R_k^{\pm}}{\mathbf{D}_k}$ nach Lemma 3.48 wieder $\mathcal{L}^2$-beschränkt ist.

Würde man die Bedingungen in A2 näher an den von SPALL angegebenen formulieren wollen, und bedenkt man, dass $\frac{1}{\mathbf{D}_k}$ ohnehin als $\mathcal{L}^2$-beschränkt vorausgesetzt wird, benötigte man ein Resultat, das für zwei $\mathcal{L}^2$-beschränkte Folgen von Zufallsgrößen X_k und Y_k, $X_k Y_k$ wieder $\mathcal{L}^2$-beschränkt ist (und würde dies auf $X_k = \tilde{f}(\mathbf{X}_k \pm h_k \mathbf{D}_k)$ und $Y_k = \frac{1}{\mathbf{D}_{ki}}$ anwenden). Für den Fall BERNOULLI-verteilter Perturbationen $\mathbf{D}_k$ nach Bemerkung 4.45 hat sich dort auch gezeigt, dass die Forderungen $R_k^{\pm}$ und $f(\mathbf{X}_k \pm \overline{\mathbf{D}}_k)$ in $\mathcal{L}^2$ beschränkt gerade wieder hinreichend sind.

Beweis des Theorems. Das KUSHNER-CLARK-Theorem 4.30 soll angewendet werden, daher werden die Voraussetzungen überprüft:

- $\mathbf{g} = \nabla f$ ist stetig, da f sogar als dreimal stetig differenzierbar vorausgesetzt wird.
- $\|\mathbf{B}_k(\mathbf{X}_k)\| < \infty$ f.s. $\forall k$ und $\mathbf{B}_k(\mathbf{X}_k) \to 0$ f.s. folgt aus dem Bias-Lemma (Lemma 4.46) mit $h_k \to 0$, $k \to \infty$ nach **B.1″**.
- Die Bedingungen an a_k gelten nach A1 (entspricht **B.1″**).
- Die Beschränktheit von $\mathbf{X}_k$ gilt nach A3 (entspricht **B.3″**).

Damit folgt die Behauptung (4.81), wenn Bedingung ((iv)) aus Theorem 4.30 gezeigt werden kann.

Verwende dazu Lemma 4.32 und Lemma 4.33, mit

$$\begin{aligned} \mathbf{T}_k &= \frac{1}{2\mathbf{D}_k}\Big(\tilde{f}(\mathbf{X}_k + \overline{\mathbf{D}}_k) - \tilde{f}(\mathbf{X}_k - \overline{\mathbf{D}}_k)\Big) \\ Z_k &= (R_k, \overline{\mathbf{D}}_k),\ \overline{\mathbf{D}}_k = h_k \mathbf{D}_k \\ u = \hat{\mathbf{g}}_k^{\mathrm{SP}} \text{ d.h. } u(\mathbf{X}_k, Z_k) &\coloneqq \frac{f(\mathbf{X}_k + \overline{\mathbf{D}}_k) - f(\mathbf{X}_k - \overline{\mathbf{D}}_k)}{2\overline{\mathbf{D}}_k} + \frac{R_k}{2\overline{\mathbf{D}}_k}, \end{aligned}$$

$\mathbf{T}_k$ ist dabei $\mathcal{L}^2$-beschränkt nach A2 und Bemerkung 4.44 für Lemma 4.32 und es folgt $\mathbb{E}(\|\mathbf{E}_k\|^2) \leq c h_k^{-2}$. Für Lemma 4.33 sind $(R_k, \overline{\mathbf{D}}_k)$, $k \geq 0$ unabhängig nach **B.7″**. $U_k \in \mathcal{L}^2$, da $U_k = \frac{1}{h_k}\mathbf{T}_k$ und $\mathbf{T}_k$ $\mathcal{L}^2$-beschränkt. □

Bemerkung 4.49. Die Bemerkung 4.38 auf SPSA erweiternd, bietet sich auch hier die Wahl $h^{\mathrm{SP}} = \sqrt[3]{r}$ an.

Einordnung der Bedingungen für die Konvergenz

Bemerkung 4.50. An dieser Stelle werden die Voraussetzungen gewürdigt.

- **B.1″** kann durch geeignete Wahl der Schrittweitenfolgen erfüllt werden.
- Im Falle der modellfreien Optimierung, ohne Kenntnis von f, kann man über **B.2″** kaum Angaben machen, und auch im Allgemeinen ist es meines Erachtens schwer überprüfbar. Zum zweite Teil von **B.3″** kann ebenso, vor allem im modellfreien Fall, eher keine Aussage getroffen werden.
- Ob die Beschränktheit von $\mathbf{X}_k$ als sinnvolle Annahme erscheint, hängt stark von der Anwendung ab.

- Die Bedingung an das Rauschen in **B.4″** erscheint sehr natürlich, da die Zufallsgrößen $\mathbf{X}_k, \mathbf{D}_k$ gerade bestimmen, an welcher Stelle die Funktionsauswertungen durchgeführt werden und der bedingte Erwartungswert der entsprechenden Rausch-Terme dann gegeben $\mathbf{X}_k, \mathbf{D}_k$ null sein soll.

- Hat man $\mathbf{D}_k$ BERNOULLI-verteilt gewählt nach Bemerkung 4.45, so reicht es für das Erfülltsein von (4.74), wenn $\tilde{f}(\mathbf{X}_k \pm h_k \mathbf{D}_k)$ in $\mathcal{L}^2$ beschränkt ist. Wie in der gleichen Bemerkung erwähnt, ist dies aber bereits erfüllt, wenn $f(\mathbf{X}_k \pm h_k \mathbf{D}_k)$ und $R_k^{\pm}$ in $\mathcal{L}^2$ beschränkt sind. Letzteres ist dabei eine auch schon für das FDSA-Verfahren getroffene sinnvolle Bedingung an das Rauschen. Inwiefern die Bedingung an f plaubsibel wirkt, hängt wieder von der Anwendung ab.

- Wie auch im deterministischen ableitungsfreien Fall in Lemma 4.11 ist eine Regularitätsvoraussetzung nötig und sinnvoll, um Aussagen über die Güte der ableitungsfreien Näherung an den Gradienten zu treffen. $f \in C^3$ aus **B.5″** erscheint dabei sehr passend. Inwiefern die geforderte Beschränktheit plausibel ist, hängt stark von der Anwendung ab.

- Die Bedingungen an die Perturbationen können einfach durch eine geeignete Wahl der $\mathbf{D}_k$ erfüllt werden. Mit der BERNOULLI-verteilten Wahl nach Bemerkung 4.45 liegt ein sinnvoller Vorschlag vor, wie die Bedingungen erfüllt werden können.

- Die Bedingungen aus der wahrscheinlichkeitstheoretischen Modellbildung erscheinen natürlich: Dass die das Rauschen beschreibenden Zufallsgrößen unabhängig voneinander und von den Iterierten und Perturbationen sind, wirkt plausibel (bzw. kann man im Fall der Perturbationen dafür sorgen, sie unabhängig zu wählen). Die Bedingung an den Startpunkt der Iteration $\mathbf{X}_0$ ist nicht sehr einschränkend und z.B. durch deterministische Wahl des Startpunkts zu erfüllen.

Für den Anwendungsfall werden die Bedingungen in der hierauf aufbauenden Bemerkung 6.2 eingeordnet.

SPSA- und GSD-Verfahren im rauschfreien Fall

Durch die geringere Anzahl von benötigten Funktionsauswertungen in der Größenordnung $O(1)$ statt $O(m)$ (siehe z.B. Tabelle 4.1), erscheint es auch gewinnbringend, das SPSA-Verfahren im deterministischen Fall anwenden zu wollen.

Ist es, wie in der Anwendung für die adaptive Optik vor allem von Interesse, dass der Zielfunktionswert schnell fällt, erscheint es ausgesprochen günstig, dass SPSA-Verfahren in Erwägung zu ziehen. Wenn dies nicht so entscheidend ist, dann stellt sich die Frage, welches Verfahren, bezogen auf die Anzahl der Funktionsauswertungen, effizienter ist.

Da für die Gradientenschätzung/-näherung weniger Informationen verwendet werden, muss man natürlich davon ausgehen, dass $\hat{\mathbf{g}}_k^{\mathrm{SP}}(\mathbf{x})$ schlechter den Gradienten $\nabla f(\mathbf{x})$ schätzt, als $\hat{\mathbf{g}}_k^{\mathrm{GN}}(\mathbf{x})$ ihn nähert. In Abschnitt 4.2 hat man gesehen, dass neben dem Verfahren des steilsten Abstiegs eine ganze Klasse von Abstiegsverfahren konvergiert, bei denen die Abstiegsrichtung von $\mathbf{p}_k = -\nabla f(\mathbf{x})$ abweicht. Das Verfahren muss kein schlechteres Verhalten aufweisen,

nur weil seine Abstiegsrichtung der Richtung des steilsten Abstiegs nicht möglichst nahe kommt.

Als Maß für die Effizienz der Verfahren kann man den zurückgelegten Weg im Zielfunktionsbereich pro Funktionsauswertungen heranziehen. Es stellt sich aber immer die Frage, ob man für beide Verfahren wirklich optimale Parameter gefunden hat, ob und mit welcher Liniensuche man GSD verwendet, so dass aus einzelnen Tests nicht ohne weiteres allgemeiner belastbare Aussagen erhalten werden können. In [Spa05] geht SPALL auf den Vergleich zwischen FDSA- und SPSA-Verfahren ein.

Bemerkung (Einordnung der SPSA-Bedingungen im rauschfreien Fall). Alle Bedingungen an das Rauschen fallen zusätzlich zu dem in Bemerkung 4.50 Gesagten weg. Wählt man die Perturbationen $\mathbf{D}_k$ nach Bemerkung 4.45 BERNOULLI-verteilt, so ist neben $\mathbf{B.6}''$ auch $\mathbf{B.4}''$ schon erfüllt, wenn $f(\mathbf{X}_k \pm \overline{\mathbf{D}}_k)$ in $\mathcal{L}^2$-beschränkt ist. Wählt man $\mathbf{X}_0$ geeignet, ist auch $\mathbf{B.7}''$ erfüllt. Für alle weiteren Bedingungen siehe Bemerkung 4.50.

4.3.8 Effizienztheorie des SPSA-Verfahrens

Einleitung

In dieser Arbeit wird kaum auf die asymptotische Theorie eingegangen, weil man für die Anwendung in erster Linie an der Konvergenz der $\mathbf{X}_k$ interessiert ist und nur mittelbar daran, inwieweit der Gradientenschätzer $\hat{\mathbf{g}}_k(\mathbf{X}_k)$ ein asymptotisch normaler Schätzer des Gradienten ist. Die Fixierung auf $-\nabla f(\mathbf{x}_k)$ erscheint nicht nötig, wenn man bedenkt, dass es statt des SD-Verfahrens ohnehin noch Verfahren höherer Ordnung gibt, bei denen $\mathbf{p}_k$ von der Richtung des steilsten Abstiegs $-\nabla f(\mathbf{x}_k)$ abweicht.

Schrittweiten und Effizienz

Im Rahmen der Umsetzung des Verfahrens in einen Algorithmus sind unter anderem die Schrittweitenfolgen a_k und h_k festzulegen. Die klassische Wahl ist

$$a_k = \frac{a}{(k+1)^\alpha} \text{ und} \tag{4.82}$$

$$h_k = \frac{h}{(k+1)^\gamma}. \tag{4.83}$$

Unter Beibehaltung der Vorschrift $\mathbf{x}_{k+1} = \mathbf{x}_k - a_k \hat{\mathbf{g}}_k(\mathbf{x}_k)$ und wenn $\mathbf{x}_0$ der Startpunkt sein soll, ist hier der Shift $k \hookrightarrow k+1$ erfolgt.

Lemma 4.51. *Wählt man a_k und h_k gemäß* (4.82), (4.83), *so sind*

$$\gamma > 0,\ \alpha \in [\tfrac{1}{2} + \gamma, 1] \tag{4.84}$$

notwendige Bedingungen für die Konvergenz des SPSA-Verfahrens nach Theorem 4.48.

Der folgende Satz ist an Satz 2 aus [Spa92] angelehnt:

Satz 4.52 (Asymptotische Normalverteilung von $\mathbf{X}_k$ um $\mathbf{x}^*$)**.**
Es gelten die Bedingungen von Theorem 4.48 und Lemma 4.46 jeweils in der Form von SPALL *mit der verschärften Voraussetzung:*

A.2′ Es gebe $\delta, \alpha_0, \alpha_1, \alpha_2 > 0$, mit denen für alle k gelte:

$$\begin{aligned} \mathbb{E}\left(|R_k^{\pm}|^{2+\delta}\right) &\le \alpha_0, \\ \mathbb{E}\left(|f(\mathbf{X}_k \pm h_k \mathbf{D}_k)|^{2+\delta}\right) &\le \alpha_1 \text{ und} \\ \mathbb{E}\left(|\mathbf{D}_{ki}|^{-2-\delta}\right) &\le \alpha_2,\ i = 1, \dots, m. \end{aligned}$$

Sei σ^2, ρ^2, ξ^2, so dass

$$\mathbb{E}\left[\left(R_k^+ - R_k^-\right)^2 \,\middle|\, \mathbf{X}_0, \dots, \mathbf{X}_k\right] \to \sigma^2 \quad \text{f.s.} \tag{4.85}$$

$$\mathbb{E}({\mathbf{D}_{ki}}^{-2}) \to \rho^2 \quad \mathbb{E}({\mathbf{D}_{ki}}^{2}) \to \xi^2 \quad k \to \infty \quad \forall l. \tag{4.86}$$

Außerdem sei für alle genügend großen k die Folge $\left\{\mathbb{E}\left[R_k^2 \,\middle|\, \mathbf{X}_0, \dots, \mathbf{X}_k; h_k \mathbf{D}_k = \boldsymbol{\eta}\right]\right\}$ f.s. gleichgradig stetig bei $\boldsymbol{\eta} = \mathbf{0}$ und stetig in $\boldsymbol{\eta}$ in einer kompakten zusammenhängenden Menge, die f.s. $\overline{\mathbf{D}}_k$ enthält. Des Weiteren sei $\beta \coloneqq \alpha - 2\gamma > 0$, $3\gamma - \alpha/2 \ge 0$ und P orthogonal mit

$$P\ \nabla^2 f(\mathbf{x}^*)\ P^T = a^{-1} \operatorname{diag}(\lambda_1, \dots, \lambda_m). \tag{4.87}$$

Dann gilt:

$$k^{\beta/2}(\mathbf{X}_k - \mathbf{x}^*) \xrightarrow{\text{in Verteilung}} \mathcal{N}(\boldsymbol{\mu}, PMP^T), \quad k \to \infty, \tag{4.88}$$

wobei $M = \frac{1}{4} a^2 c^{-2} \sigma^2 \rho^2 \operatorname{diag}\left((2\lambda_1 - \beta_+)^{-1}, \dots, (2\lambda_m - \beta_+)^{-1}\right)$ mit

$$\beta_+ = \begin{cases} \beta < 2 \min_i \lambda_i & \text{falls } \alpha = 1 \\ 0 & \text{falls } \alpha < 1 \end{cases} \quad \text{und}$$

$$\boldsymbol{\mu} = \begin{cases} \mathbf{0} & \text{falls } 3\gamma - \alpha/2 > 0 \\ (a\nabla^2 f(\mathbf{x}^*) - \frac{1}{2}\beta_+ Id)^{-1}\mathbf{T} & \text{falls } 3\gamma - \alpha/2 = 0 \end{cases}$$

und für das l-te Element von $\mathbf{T}$ gilt: $\mathbf{T}_l = -\frac{1}{6} a c^2 \xi^2 \left(f_{lll}^{(3)}(\mathbf{x}^*) + 3 \sum_{\substack{i=1 \\ i \ne l}}^{m} f_{iil}^{(3)}(\mathbf{x}^*)\right)$.

Bemerkung. Wählt man $\mathbf{D}_k$ nach Bemerkung 4.45 BERNOULLI-verteilt, so ist $\sigma^2 = \xi^2 = 1$ in (4.86). Für den Beweis siehe [Spa92], der dazu ein Resultat in [Fab68] verwendet.

Korollar 4.53. *Soll $\mathbf{X}_k$ asymptotisch normalverteilt um $\mathbf{x}^*$ sein nach Satz 4.52 und sollen die Voraussetzungen des Theorems 4.48 erfüllt sein bei der Wahl (4.82), (4.83) für die Schrittweitenfolgen a_k, h_k, so muss notwendig gelten:*

$$\gamma > 0,\ \alpha \in [\tfrac{1}{2} + \gamma, 1],\ \alpha < 6\gamma. \tag{4.89}$$

Soll alternativ zusätzlich zu den Konvergenzbedingungen β maximal sein, so ergibt sich die asymptotisch optimale Wahl

$$\alpha = 1 \text{ und } \gamma = \tfrac{1}{6}. \tag{4.90}$$

Beweis. Wegen der Voraussetzungen des Satzes 4.52 muss $\gamma \geq \frac{\alpha}{6}$ gelten. Da $\mu = 0$ in Satz 4.52 sein soll, muss zusätzlich zu den in Lemma 4.51 genannten notwendigen Bedingungen gelten $3\gamma - \frac{\alpha}{2} > 0$, d.h. $\gamma > \frac{\alpha}{6}$.

Für die zweite Aussage beachte man, dass $\beta \equiv \alpha - 2\gamma$ maximal wird unter a) α maximal, d.h. $\alpha = 1$ wegen der Randbedingung $\alpha \in [\frac{1}{2} + \gamma, 1]$, und b) γ minimal, d.h. $\gamma = \frac{1}{6}$ wegen der Bedingung $\gamma \geq \frac{\alpha}{6}$. □

Auch wegen des erwähnten Effekts eines zu kleinen h in der Gradientenschätzung (im Sinne einer Erweiterung von Bemerkung 4.38 auf das SPSA-Verfahren) ist es sinnvoll, γ so klein wie möglich zu wählen.

In [Spa98] und darauf zurückgreifend in [Spa05] werden Empfehlungen zur Wahl der Algorithmus-Parameter gegeben.

Bemerkung 4.54. Spall, [Spa05, S.164] schlägt vor, α und γ so klein wie möglich zu wählen unter Einhaltung der Bedingungen von Satz 4.52 und Theorem 4.48, damit die Schrittweiten auch bei großen k noch groß sind. Er schlägt

$$\alpha_{\text{Spall}} = 0.602 \text{ und} \tag{4.91}$$

$$\gamma_{\text{Spall}} = 0.101 \tag{4.92}$$

vor. Mit dieser Wahl ist $\beta_+ = 0$ und $\mu = 0$ in der asymptotischen Verteilung.

Beweis. Für $\alpha \to \min!$ folgt $\alpha > 0.6$ wegen $\alpha = \frac{1}{2} + \gamma > \frac{1}{2} + \frac{\alpha}{6}$. Für $\gamma \to \min!$ dann $\gamma > \frac{\alpha}{6} > 0.1$ (Forderung der asymptotischen Normalverteilung um $\mathbf{x}^*$ nach Korollar 4.53). □

Bemerkung 4.55. Um im Sinne von Bemerkung 4.49 h in der Größenordnung $h^{\text{SP}} := \sqrt[3]{r}$ zu wählen, aber bei der üblichen Form $h_k = \frac{h}{k^\gamma}$ zu bleiben, kann man die Anzahl der Iterierten soweit einschränken, dass h_k nahe bei h^{SP} bleibt. Aufgrund des exponentiellen Abfalls ist dies aber nicht unbedingt praktikabel. Die Wahl $\gamma = 0$ entspricht einer konstanten Schrittweitenfolge $h_k = h^{\text{SP}}$.

Mit der asymptotischen Verteilung und deren Verbesserung beschäftigt sich Dippon u.a. in [DR97].

Bemerkung 4.56 (Das Vorontsovsche SPGD-Verfahren als SPSA-Verfahren). Die Iterationsvorschrift $u_l^{(m+1)} = u_l^{(m)} - \mu \delta J^{(m)} \delta u_l^{(m)}$ des SPGD nach [VS98, (6)] lautet in der Notation dieser Arbeit

$$\mathbf{X}_{k+1} = \mathbf{X}_k - \mu \; \delta f_k \; \delta \mathbf{x}_k \tag{4.93}$$

mit $\mu > 0$, $\delta f_k = \tilde{f}(\mathbf{x} + \delta \mathbf{x}_k) - \tilde{f}(\mathbf{x})$, wobei die Perturbationen $\delta \mathbf{x}_k$ als Zufallsgrößen mit gleicher Varianz und symmetrisch um 0 verteilt gewählt werden.

Wählt man die Perturbationen gemäß $\delta \mathbf{x}_k = h_k \mathbf{D}_k$, $\mathbf{D}_k$ +1-−1-Bernoulli-verteilt, dann entspricht dieses Verfahren mit $\mu_k = \frac{a_k}{h_k}$ einem SPSA-Verfahren mit Gradientenschätzung erster Ordnung. Will man an $\mu_k = \mu$ festhalten, muss also $\frac{a_k}{h_k}$ konstant gewählt werden.

Literaturhinweise

Während der Recherche für diese Arbeit stieß ich auf eine ganze Reihe von Veröffentlichungen, die nur zu einem Teil direkt zur Arbeit beigetragen haben. Ich möchte einige dieser Quellen für diejenigen, die sich mit den *Stochastic-Approximation*-Verfahren beschäftigen wollen, aber kurz anführen:

Spall entwickelt die Idee des SPSA-Verfahrens in [Spa87], [Spa92] und verweist beim Beweis im Buch [Spa05] auf seinen Beweis in [Spa92].
In [Spa] wird ein Überblick über das SPSA-Verfahren gegeben. Die Frage der Parameterwahl wird in [Spa98] erörtert.
Weitere Veröffentlichungen von Spall über SPSA sind [SC94] und [SC98].
Ein Vergleich von FDSA und SPSA findet sich neben der Erörterung in [Spa05] in [Chi97]. Dort wird auch eine Korrektur für den Beweis der Konvergenz des RDSA-Verfahrens in [KC78] angegeben. Ein Vergleich zwischen RDSA und SPSA findet sich auch in [WC98]. Eine wichtige Quelle, die auch das in dieser Arbeit angegebene Kushner-Clark-Theorem enthält, ist [KC78]. Daneben gibt es aber auch ein neueres Buch von Kushner und Yin: [KY03].
Mit Aussagen über die asymptotische Verteilung von *Stochastic-Approximation*-Verfahren beschäftigt sich [Fab68]. In [Ger99] findet sich eine weitere Art von SA-Konvergenzbeweis. In [WCK96] werden die Bedingungen in [KC78] und Alternativen untersucht und verglichen. Dippon, Privatdozent an der Universität Stuttgart, beschäftigt sich in [Dip02] mit der asymptotischen Verteilung mit *Randomized Stochastic Approximation*-Verfahren, die Spall's SPSA und Kushner und Clark's RDSA einschließt, und untersucht den Effekt von *Iterate Averaging*, auf den in [DR97] eingegangen wird. Auch seine Habilitationsschrift [Dip98] beschäftigt sich mit der asymptotischen Entwicklung des RM-SA-Prozesses.

Ein Ansatz zum Erreichen eines Konvergenzresultats durch Betrachtung deterministischer Perturbations-Folgen $\overline{\mathbf{D}}_k$ findet sich in [Bha+03].

[Zin+] beschäftigt sich mit einer Parallelisierung des stochastischen Gradientenabstiegsverfahrens.

4.3.9 Schlusswort

Man kann für SPSA- und FDSA-Verfahren Konvergenz zeigen, und es gibt auch eine Basis für die Behandlung der asymptotischen Verteiltung. Mit dem Erscheinen von [KY03] wäre es sicher wünschenswert, die SPSA-Konvergenztheorie aufbauend auf dem dortigen Konvergenzresultat in einer Arbeit kompakt neu darzustellen (dort kann man auf die eventuell kritische Aussage der Beschränktheit von $\mathbf{X}_k$ verzichten). Inwiefern die Bedingungen an das SPSA-Verfahren in der Praxis als realistisch anzusehen sind, wird im Kapitel 6 betrachtet, nach den nun folgenden Ausführungen zur algorithmischen Umsetzung.

Kapitel 5

Vom Verfahren zum Algorithmus

5.1 Einleitung

An dieser Stelle wird in die schon im Kapitel 1 dargestellte Anwendung übergeleitet. Zunächst wird noch eine Erweiterung des SPSA-Verfahrens eingeführt und Spalls Empfehlung zur semiautomatischen Parameterbestimmung vorgestellt. Wie eine Implementierung aussehen kann, wird danach an einem C++-Codebeispiel vorgestellt.

5.2 Mittelungs-Erweiterungen (q-c-SPSA)

Über mehrfache Funktionsauswertungen $\tilde{f}(\mathbf{x})$ kann man aus der Standardabweichung der gemessenen Werte die Größe des Rauschens $R_{\mathbf{x}}$ schätzen. Verwendet man den Mittelwert der dabei gemachten Funktionsauswertungen, kann man das Verfahren zusätzlich stabilisieren. Man setzt dann als c-fach gemittelten Funktionswert

$$\tilde{f}_c(\mathbf{x}) = \frac{1}{c} \sum_{i=1}^{c} \tilde{f}^{(i)}(\mathbf{x}) \tag{5.1}$$

an, und wählt z.B. $c = 4$.

Als weitere Stabilisierungsmaßnahme im Falle großen Rauschens bietet sich die Mittelung der Gradientenschätzer an [Spa92, S. 333]:

$$\hat{\mathbf{g}}_k^{q\text{-SP}}(\mathbf{X}_k) = \sum_{i=1}^{q} \hat{\mathbf{g}}_k^{\text{SP}\,(i)}(\mathbf{X}_k), \tag{5.2}$$

wobei für jeden der q Summanden neue Perturbationen $\mathbf{D}_k$ bestimmt werden. Mit q-SPSA wird ein SPSA-Verfahren mit Mittelung der Gradientenschätzung aber ohne Mittelung der Funktionsauswertungen bezeichnet.

Die Anzahl der benötigten Funktionsauswertungen pro Iteration ist auch mit dieser Erweiterung noch von der Dimension unabhängig. Tabelle 5.1 fasst die Anzahl der benötigten Funktionsauswertungen pro Iteration für verschiedene Mittelungskombinationen zusammen.

Mittelung von $\hat{\mathbf{g}}_k^{\mathrm{SP}}$	Mittelung von $\tilde{f}$	Funktionsauswertungen pro Iteration
1	1	3
2	1	5
1	2	6
2	**2**	**10**
1	4	12
2	**4**	**20**
1	8	24
2	8	40
q	c	$c + 2 \cdot q \cdot c$

Tabelle 5.1: Anzahl benötigter Funktionsauswertungen pro Iteration bei verschiedenen Mittelungsvarianten von q-c-SPSA.

5.3 Verwerfung von Verschlechterungen

Unter anderem bei starkem Rauschen kann es vorkommen, dass eine neue Iterierte $\mathbf{x}_{k+1}$ eine Verschlechterung darstellt, das heißt, $f(\mathbf{x}_{k+1}) > f(\mathbf{x}_k)$ gilt. Für solche Fälle kann man den Algorithmus noch so erweitern, dass neue Iterierte nur akzeptiert werden, falls

$$\tilde{f}(\mathbf{x}_{k+1}) \leq \tilde{f}(\mathbf{x}_k).$$

Da $\tilde{f}$ mit Rauschen behaftet ist, kann dies zu strikt sein. Man kann stattdessen zu *gelockertem Verwerfen von Verschlechterungen* übergehen, bei dem man erst ab einem gewissen Schwellenwert (*Threshold*) δy die neue Iterierte $\mathbf{x}_{k+1}$ verwirft. In diesem Fall setzt man also $\mathbf{x}_{k+1} = \mathbf{x}_k$, falls

$$\tilde{f}(\mathbf{x}_{k+1}) - \tilde{f}(\mathbf{x}_k) \leq \delta y.$$

5.4 Parameterbestimmung nach Spall

Zur Bestimmung der Parameter nach der Empfehlung in [Spa98] wählt man für a_k in Abwandlung von (4.82) $a_k = \frac{a}{(A+k+1)^\alpha}$. Der Parameter A soll größere a ermöglichen, so dass die Schrittweiten im späten Verlauf groß bleiben, ohne zu Beginn zu Instabilitäten zu führen [Spa98, S. 820]. Gleichzeitig ändert sich das asymptotische Verhalten nicht.

Dort wird vorgeschlagen, A auf 10% der Anzahl der erwarteten Iterationen zu setzen, für die Testläufe im Rahmen dieser Arbeit wurde oft $A = 10$ gesetzt. α wird nach Bemerkung 4.54 auf $\alpha_{\mathrm{Spall}} = 0.602$ gesetzt.

a wählt man dann so, dass zu einer vorgegebenen Änderungsgrößenordnung des Steuersignals in den ersten Iterationen, die man mit c_1 bezeichnet, die Größe von a je nach Größenordnung des Gradienten gewählt wird:

$$c_1 \approx \frac{a}{(A+1)^\alpha} \left\| \hat{\mathbf{g}}_0^{\mathrm{SP}}(\mathbf{X}_0) \right\|. \tag{5.3}$$

Die Norm $\left\|\hat{\mathbf{g}}_0^{\mathrm{SP}}(\mathbf{X}_0)\right\|$ kann dabei auch als Maximumsnorm betrachtet werden und man erhält eine maximale Änderungsgröße der einzelnen Steuersignale. Der Parameter c_1 charakterisiert dann also die maximale gewünschte Änderung in den Komponenten zwischen aufeinanderfolgenden Steuersignalen $\mathbf{x}'$ und $\mathbf{x}''$:

$$\delta\mathbf{x} \coloneqq \|\mathbf{x}' - \mathbf{x}''\|_\infty \leq c_1. \tag{5.4}$$

Ein Beispiel für den Verlauf der Schrittweitenfolge a_k ist in Abbildung 5.1 dargestellt.

Nach Spalls Empfehlung setzt man $h_k = \frac{h}{(k+1)^\gamma}$ mit $\gamma = \gamma_{\mathrm{Spall}} = 0.101$ nach (4.92) und ebenfalls nach dieser h in der Größenordnung der Standardabweichung des Rauschens $R_{\mathbf{x}}$. Es hat sich gezeigt, dass hier größere Werte zu empfehlen sind. Gemäß Bemerkung 4.49 sollte h_k abweichend davon in der Größenordnung $\sqrt[3]{r}$ gewählt werden, wobei r eine obere Grenze des Rauschens ist ($|R_{\mathbf{x}}| \leq r$). Andererseits sollte aber die Beschränkung der maximalen Änderung der Komponenten aufeinanderfolgender Steuersignale, $\delta\mathbf{x} \leq c_1$ eingehalten werden, d.h.

$$\delta\mathbf{x} = \left\|(\mathbf{x}_k + h_k\mathbf{D}_k) - \mathbf{x}_k\right\|_\infty = h_k \leq c_1 \tag{5.5}$$

(denn $|\mathbf{D}_{ki}| = 1$). Dass die Gradientenschätzungs-Schrittweitenfolge h_k in einem gewünschten Bereich bleibt, erreicht man durch Beschränkung der Zahl der Iterierten oder in dem man abweichend $\gamma = 0$ setzt. Informationen für die Wahl von α und γ sind in Tabelle 5.2 zusammengefasst. Ein beispielhafter Verlauf der Schrittweitenfolge h_k wird in Abbildung 5.2 dargestellt.

	α	γ
für Konvergenz nach (4.84)	$\alpha \in [\frac{1}{2} + \gamma, 1]$	$\gamma > 0$
zusätzlich für asymptotische Normalverteilung nach (4.89)	$\alpha < 6\gamma$	
asymptotisch optimal nach (4.90)	1	1/6
praktische Empfehlung nach Spall, (4.91) & (4.92)	0.602	0.101
konstante Schrittweitenfolgen	0	0

Tabelle 5.2: Die Schrittweitenparameter α und γ

.

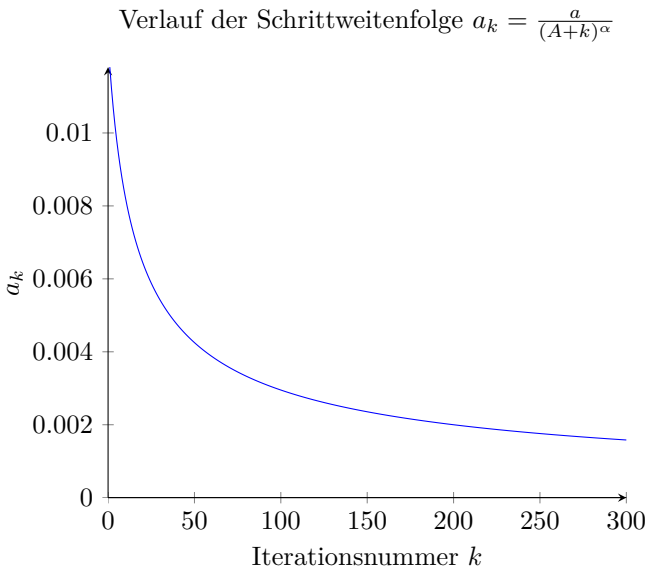


Abbildung 5.1: Der Verlauf der Schrittweitenfolge a_k für einen Startwert von 0.05, dem von SPALL empfohlenen $\alpha_{\text{Spall}} = 0.602$ nach (4.91) mit Stabilitätsparameter $A = 10$ bis zur 300. Iteration

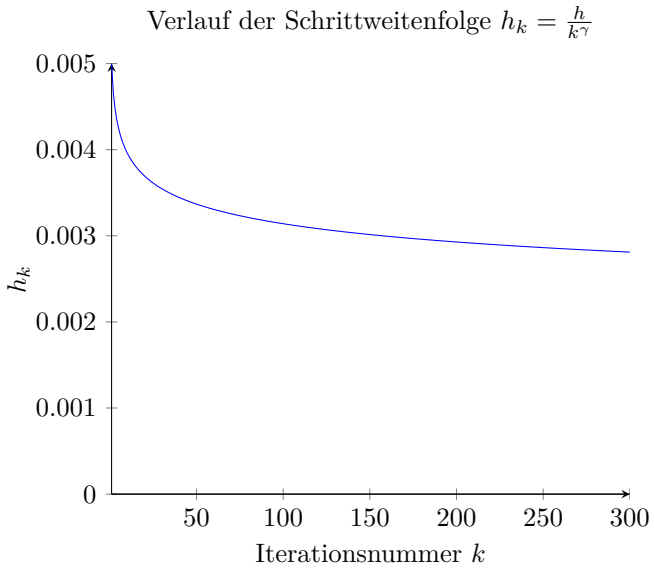


Abbildung 5.2: Der Verlauf der Schrittweitenfolge h_k für einen Startwert von 0.005, dem von SPALL empfohlenen $\gamma_{\text{Spall}} = 0.101$ nach (4.92) bis zur 300. Iteration

5.5 C++ Code

Eine Implementierung mit BERNOULLI-verteilten Perturbationen $\mathbf{D}_k$ nach Bemerkung 4.45 in C++ erhält man durch[1]

```
1  typedef double real;
   int SPSA()
   {
     real ak, hk;
     //Hier initialisieren, Parameter und Startpunkt
         setzen
6
     unsigned long int k = 1;
     ak = a / pow((k + A), alpha);
     for (; true; ++k) {
       hk = h / pow(k, gamma);
11     delta = toss();
       xplus = x + hk * delta;
       xminus = x - hk * delta;
       grad = ((eval(xplus) - eval(xminus)) / (2 * hk)) *
           delta.reziprok();
       p = (-1.0) * grad;
16     ak = a / pow((k + A), alpha);
       x = x + ak * p;
       //Schritt evtl. verwerfen
       //Hier Abbruchbedingung prüfen
     }
21   return 0;
   }
```

`eval` ist dabei eine Funktion, die zu $\mathbf{x}$ gegebenenfalls verrauschte Zielfunktionswerte $\tilde{f}(\mathbf{x})$ zurückliefert und z.B. auch prüft, ob $\mathbf{x}$ im zulässigen Bereich ist. Für `delta`, `x`, `xplus`, `xminus` usw. gibt es eine von `vector<real>` abgeleitete `Vektor`-Klasse, die übliche Vektorfunktionen für eine einfachere Notation des Quelltexts enthält, wie Skalarmultiplikation, Vektoraddition oder die Kehrwert-Funktion in Zeile 14, die angelehnt an Notation (3.6) $(\frac{1}{\mathbf{x}_1}, \ldots, \frac{1}{\mathbf{x}_m})$ zurückgibt. Für die Erzeugung der Perturbation in `toss` wurde das in C++ eingebundene Random-Paket von FOG [Fog] verwendet. Dort liefert die Funktion `IRandomX(a,b)` eine gleichverteilte Pseudo-Integer-Zufallszahl in $[a, b]$. Der Aufruf in der zweiten Zeile im unteren Codeauschnitt erzeugt dabei einen m-dimensionalen Vektor.

```
1  Vektor toss() {
     Vektor val(m);
     for (Vektor::iterator vali = val.begin(); vali != val
         .end(); ++vali) {
       *vali = 2 * round(rndgen.IRandomX(0, 1)) - 1;
5    }
     return val;
   }
```

[1] Dabei ist $k = 1, 2, \ldots$, das heißt $\mathbf{x}_1$ ist der Startpunkt und die Schrittweitenfolgen haben die Form $h_k = \frac{h}{k^\gamma}$, man erhält die Formen (4.82) und (4.83) ohne den dortigen Shift.

Für die Tests am Live-System wurde das Verfahren neu in Labview 2010 implementiert, da eine Anbindung des Laborequipments über Labview bereits funktionstüchtig bestand und eine Einbindung in Labviews C-Code-Schnittstelle unpraktikabel erschien (u.a. wegen Datentyp-Umwandlungsfragen zwischen C, C++ und Labview und der Beschränktheit auf bestimmte C-Compiler).

Kapitel 6

Anwendung in der adaptiven Optik

6.1 Einleitung

Zum Testen des Verhaltens der betrachteten Optimierungsverfahren für die adaptive Optik in der Praxis wurde der Testaufbau von Peter Becker übernommen, der bis März 2011 die Masterarbeit „Korrektur von Leichtbau-Membranspiegeln mittels aktiver Optik“ [Bec11] in der Abteilung aktive optische Systeme am Institut für Technische Physik des Deutschen Zentrums für Luft- und Raumfahrt e.V. (DLR), Standort Stuttgart, geschrieben hat. Konkrete Angaben über das Testsystem basieren auf entsprechenden Abschnitten in dessen Arbeit. Um die Implementierung der Verfahren mit dem physikalischen System zu verbinden, wurde das von Peter Becker im Rahmen seiner Masterarbeit und seiner Arbeit am *DLR* entwickelte Labview-Programm um ein Modul *allgemeine Algorithmen* `ga_main.vi` erweitert und die bestehenden Module zur Kameraansteuerung `cam_main.vi` und zur Spiegelsteuerung `mrrctrl_main.vi` für die vorliegende Arbeit verwendet und weiterentwickelt.

6.2 Praktische Parameterwahl

Durch die halbautomatische Parameterbestimmung ist der Parameter c_1 die Haupteinflussgröße auf den Algorithmus. Bei Kenntnis des Spiegels kann man hierfür einen sinnvollen Bereich schätzen: So wurde für den verwendeten Spiegel `Mirao52d` mit einem Aktuatorspannungsbereich von $[-0.5\,\mathrm{V}, 0.5\,\mathrm{V}]$ $c_1 = 0.005$ festgelegt. Es bräuchte also mindestens 100 Schritte, um den gesamten Spannungsbereich zu durchlaufen.

Die Größenordnung des Rauschens $R_{\mathbf{x}} = \tilde{f}(\mathbf{x}) - f(\mathbf{x})$ wird aus der Standardabweichungen von 4 bis 5 Auswertungen von $\tilde{f}(\mathbf{x})$ bestimmt. Mögliche Quellen des Rauschens sind unter anderem, wie schon in Abschnitt 1.4 erwähnt, das Nachschwingen des Spiegels, eine schwankende Strahllage des Laserstrahls und das Rauschen der Kamera.

Im Sinne von (5.5) ergibt sich, wenn man $h_k = h^{\mathrm{SP}}$ wählen möchte, die Bedingung $h^{\mathrm{SP}} = \sqrt[3]{r} \leq 0.005$. Dies wäre die zu wählende Gradientenschätzungs-

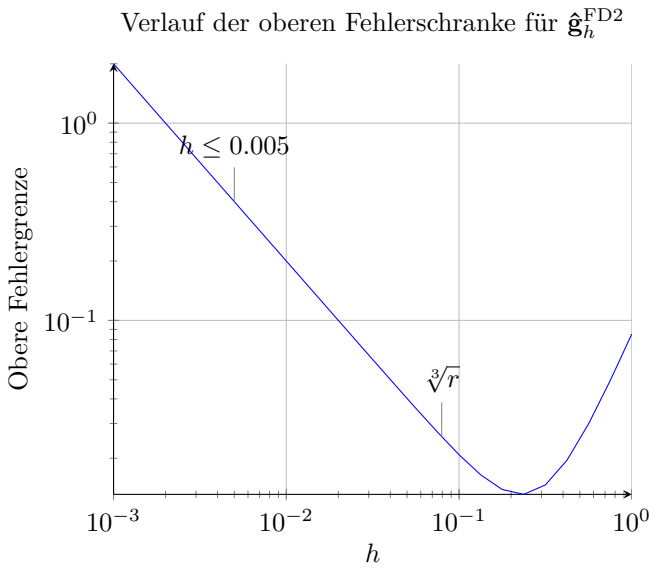


Abbildung 6.1: Das Rauschen wird hier als mit $r = 0.0005$ beschränkt angenommen. Die obere Fehlerschranke ist für $h = 0.005$ etwa 15-mal höher als bei der empfohlenen Wahl $\sqrt[3]{r} = 0.079$. Die Werte der Fehlerabschätzung sind dabei: 0.4 bei $h = 0.005$ und 0.026 bei $h = 0.079$. Der geringste Wert liegt bei 0.013 für $h = 0.2289$.

Schrittweite bei einer Größenordung des Rauschens von 10^{-7} oder kleiner. Meist lag der Wert für das Rauschen der Zielfunktion in der Größenordnung 10^{-4}, so dass eine gute Wahl für h in der Größenordnung von 0.05 liegt. Der Verlauf der oberen Fehlergrenze für $\hat{\mathbf{g}}_k^{\mathrm{FD2}}$ wird an einem Beispiel in Abbildung 6.1 dargestellt und kann qualitativ auf den Fall des SPSA-Verfahrens übertragen werden. In Abbildung 6.2 wird ein Beispiel für den Effekt von addiertem normalverteilten Rauschen gegeben.

Die Beschränkung auf $c_1 = 0.005$ entspringt dabei auch dem Gedanken, dass die Steuersignale bei den Funktionsauswertungen zur Gradientenschätzung keine größeren Sprünge machen sollen als beim eigentlichen Iterationsschritt. Für den Iterationsschritt $\mathbf{x}_k \hookrightarrow \mathbf{x}_{k+1}$ ist der Schrittweitenparameter a in der Größenordnung von 0.05 ist so gewählt, dass die Steuersignaländerungen $\delta\mathbf{x}$ in der Größenordnung von c_1 liegen. Diese wurde erreicht, in dem bei der automatischen Parameterdetektion $a = \frac{c_1}{g} \cdot (A+1)^{\alpha_{\mathrm{Spall}}}$ gesetzt wurde. Für die Größenordnung g des Gradientenschätzers wurde dabei der Mittelwert von $|\hat{\mathbf{g}}_{ki}(\mathbf{x}_k)|$ über 10 Test-Gradientenschätzungen $\hat{\mathbf{g}}_k$ gemittelt.

6.3 Laboraufbau

Um das Verhalten des Optimierungsverfahrens SPSA unter Praxisbedingungen darzustellen, wurde der nachjustierte Labor-Testaufbau (Laboraufbau B) verwendet. Im Vergleich zu dem schematisch in Abbildung 6.4 beschriebenen und im Foto der Abbildung 6.5 dargestellten Laboraufbau A wurde die Lin-

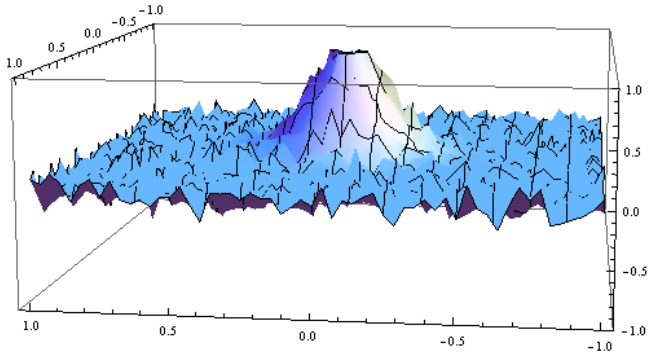


Abbildung 6.2: Der Effekt von Rauschen auf die Strehl-Funktion $\exp\big(-\big(2\pi\sqrt{\frac{x^2+y^2}{2}}\big)^2\big)$, vergleiche (1.10), anhand von addierten $\mathcal{N}(0, 0.2^2)$-verteilten Pseudozufallszahlen. Plot erzeugt mit Mathematica.

se $L1$ direkt durch einen Planspiegel ersetzt. Das heißt, die hauptsächlichen Aberrationen des Systems kommen aus der nichtplanen Grundstellung des adaptiven Spiegels selbst und im geringeren Maß von den weiterhin vorhandenen Komponenten wie dem Strahlaufweitungsteleskop, der Linse L2 und den Planspiegeln. Dies stellt eine Vereinfachung des Versuchsaufbaus dar. Der Nachweis eines funktionierenden Einsatzes des Optimierungsverfahrens auch für andersartige optische Störungen bleibt damit aber weiterhin möglich. Der in Abbildung 6.4 gezeigte Wellenfrontsensor wurde für die adaptive Optik nicht verwendet. Der verwendete adaptive Spiegel ist in Abbildung 6.3 dargestellt und in Tabelle 6.1 werden einige seiner Kenngrößen angegeben.
Zur Kennzeichnung des optischen Systems soll der Radius des Beugungsscheibchens angegeben werden. Der Durchmesser des 1. Beugungsscheibchens im Testaufbau ergibt sich gemäß

$$d = 2.4392\frac{\lambda f}{D}, \tag{6.1}$$

wobei $\lambda = 632\,\text{nm}$ die Wellenlänge ist, $f = 250\,\text{mm}$ die Brennweite der Fokussierlinse ist (Linse L2 in Abbildung 6.4) und $d = 15\,\text{mm}$ der Blendendurchmesser ist (vor R1, siehe Tabelle 6.1). Mit den angegebenen Werten ergibt sich $d = 25.7\,\mu\text{m}$. Mit der Pixelgröße $4.4\,\mu\text{m}$ der verwendeten Kamera entspricht dies einem Kreis mit einem Radius von 2.9 Pixeln.

Laserlicht zeichnet sich im Allgemeinen aus durch

- Monochromatismus (enger Wellenlängenbereich),
- Parallelität des Strahls und
- große Kohärenzlänge.

Der verwendete HeNe-Laser arbeitet bei einer Wellenlänge von 632nm.

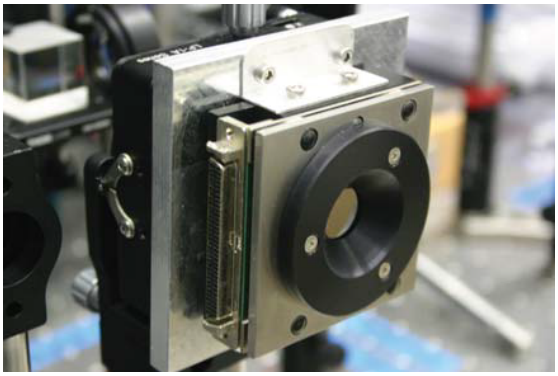


Abbildung 6.3: Adaptiver Spiegel im Laboraufbau. Entnommen aus [Bec11, S.12]

Apertur-Durchmesser	15 mm
Anzahl Aktuatoren	52
Abstand zwischen den Aktuatoren	2.5 mm
Stellfrequenz	> 200 Hz
Maximale/minimale Aktuatorspannung	± 0.5 V
Maximale Wellenfront-Amplitude	50 μm

Tabelle 6.1: Kenngrößen des adaptiven Spiegels, entnommen aus [Bec11, Tabelle 1]

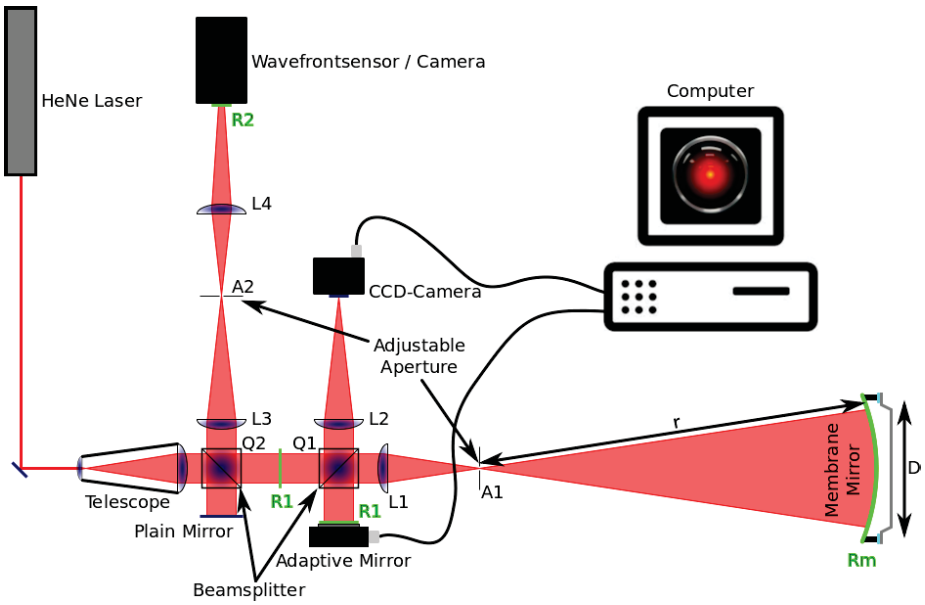


Abbildung 6.4: Schema des Laboraufbaus A, aus [Bec11, S. 23] entnommen. Der Wellenfrontsensor wird für die in dieser Arbeit betrachtete Optimierungsanwendung nicht verwendet.

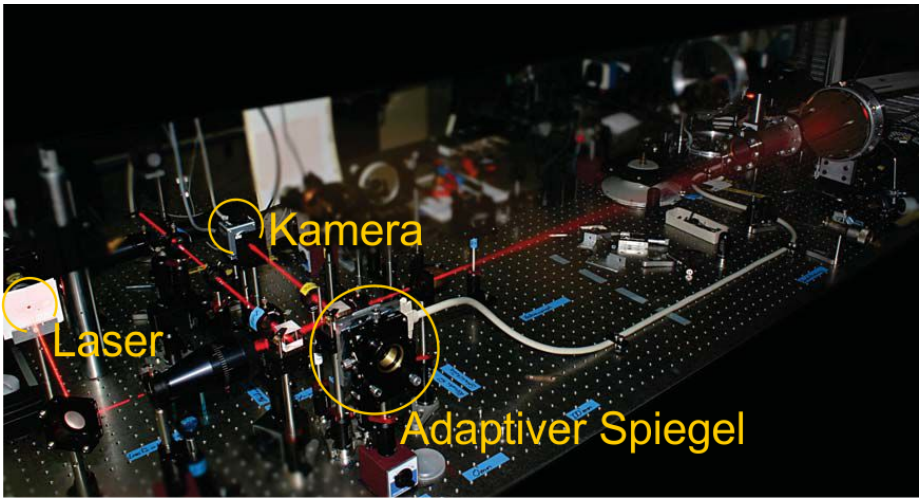


Abbildung 6.5: Foto des Laboraufbaus A, entnommen aus [Bec11, S.23], Beschriftung geändert.

6.4 Metrik

Grundlegend wurde als Metrik die *Power-in-the-Bucket*-Funktion in (1.14) betrachtet. Der Mittelwert verschiedener *Power-in-the-Bucket*-Funktionen stellt aber ebenfalls eine Metrik für das System dar, so dass die einzelnen in diesen Mittelwert eingehenden *Power-in-the-Bucket*-Funktionen als *Submetriken* bezeichnet werden, da sie alleine auch schon die Systemleistung charakterisieren würden. Den *Power-in-the-Bucket*-Wert für multiple virtuelle Aperturen berechnet man gemäß

$$P_{r_1,r_2,\ldots,r_n}(C) = \frac{1}{n}\sum_{i=1}^{n} P_{r_i}(C). \tag{6.2}$$

Eine beispielhafte Darstellung einer solchen *Power-in-the-Bucket*-Metrik mit mehrfachen virtuellen Aperturen auf dem Kamerabild des Laborsystems findet man in Abbildung 6.6.

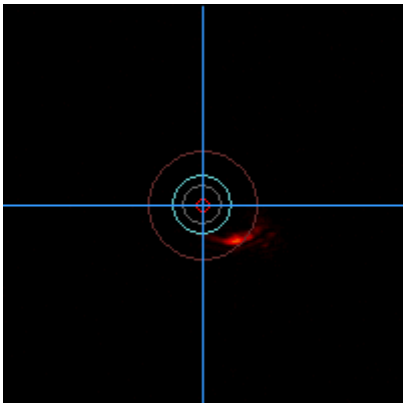


Abbildung 6.6: Das Kamerabild des Spots (rot, im unteren rechten Quadranten) in der Nullstellung des adaptiven Spiegels vor Beginn der Optimierung. 400px×400px -Hardware-*Region of Interest* in *Gradient*-Farbpalette. Die für die eingehenden Submetriken relevanten kreisförmigen virtuellen Aperturen mit Radien von 55, 30, 20 und 7 Pixeln sind sichtbar (letzterer hellrot).

Bemerkung 6.1. Es gilt $P_r \in [0,1]$, denn im schlechtesten Fall trifft (fast) überhaupt kein Licht auf die Kamera, und man erhält einen *Power-in-the-Bucket*-Wert von 0 %. Im besten Fall trifft das komplette Licht in A_r, d.h. $P_r = 100\,\%$. Diese Eigenschaft überträgt sich ebenfalls auf den Fall mehrfacher virtueller Aperturen wie in (6.2). Zu beachten ist, dass dies keine Aussage darüber ist, inwiefern die Werte tatsächlich erreicht werden können, sondern nur zeigt, dass der Zielfunktionswert beschränkt ist. Um ein Minimierungsproblem vorliegen zu haben, setzt man dann $\tilde{f} := -P_r \circ f_1$, vergleiche (1.11). Für *Power-in-the-Bucket*-Werte von einfachen oder mehrfachen virtuellen Aperturen gilt also $f(\mathbf{x}), \tilde{f}(\mathbf{x}) \in [-1,0]\ \forall \mathbf{x}$.

Wählt man den beugungsbegrenzten Radius von 2.9 px als Bucket-Radius, so ergibt sich der maximale *Power-in-the-Bucket*-Wert wie folgt: Da die Intensität

durch

$$I(r) = I_0 \left(\frac{J_1(2\pi r)}{\pi r} \right)^2 \tag{6.3}$$

unter Verwendung der Bessel-Funktion erster Art (nach Definition A.7) gegeben ist und für $f(r,\phi) = f(r)$ gilt $\int_0^{2\pi} \int_0^r f(r,\phi) dr d\phi = 2\pi \int_0^r f(r) dr$, ist das Verhältnis von Intensität innerhalb des Beugungsscheibchens zur Gesamtintensität gegeben durch

$$\text{pitb}_{\text{Th. Max}} = \frac{\cancel{2\pi} \int_0^{0.6098} \cancel{I_0} \left(\frac{J_1(2\pi r)}{\pi r} \right)^2 dr}{\cancel{2\pi} \int_0^{\infty} \cancel{I_0} \left(\frac{J_1(2\pi r)}{\pi r} \right)^2 dr} = 97.6\%. \tag{6.4}$$

Bei größeren Virtuellen-Apertur-Kreisen kann wie oben beschrieben ein Wert von 100% erreicht werden.

6.5 Die Bedingungen in der Anwendung

Durch die Wahl der einfachen *Power-in-the-Bucket*-Funktion (1.14) oder der *Power-in-the-Bucket*-Funktion mit multiplen virtuellen Aperturen (6.2) als Systemleistungsmetrik ($f_2(C) \coloneqq -P(C)$, vergleiche (1.13)) ist f beschränkt: $|f(\mathbf{x})| \leq 1$. Gleiches gilt für $\tilde{f}$. Das Optimierungsverfahren wird zwar unrestringiert durchgeführt, d.h. mit $\mathbb{X} = \mathbb{R}^m$ bzw. $\mathbb{H} = \mathbb{R}^m$, dabei aber darauf geachtet, die Steuersignale komponentenweise auf $[-x_{\min}, x_{\max}]$ einzuschränken. Das heißt, das Verfahren wird abgebrochen und mit anderen Parametern neu gestartet, falls die Aktuatorspannungen zu nahe an die Grenzen kommen.

Einordnung der SPSA-Bedingungen in der Anwendung

In Bemerkung 4.50 wurde schon begonnen, die Bedingungen für die Konvergenz des SPSA-Verfahrens einzuordnen. Für den Anwendungsfall der modellfreien Metrik-basierten adaptiven Optik soll dies an dieser Stelle fortgeführt werden:

Bemerkung 6.2. Durch die Wahl von geeigneten Schrittweitenfolgen und dem Nutzen der BERNOULLI-verteilten Perturbationen $\mathbf{D}_k$ nach Bemerkung 4.45 sind **B.1″** und **B.6″** erfüllt. Auch der Startpunkt ist so zu wählen, dass Bedingung **B.7″** erfüllt werden kann, dies wurde schon in Bemerkung 4.50 erwähnt.

- Ohne Kenntnis der genauen Funktion $\mathbf{x} \mapsto y$ hat man wohl keine Chance, etwas über die Erfülltheit der Bedingungen **B.2″** und des 2. Teils von **B.3″** auszusagen.
- Die Beschränktheit der Iterierten in **B.3″** erscheint nicht sehr problematisch. Da die Aktuatorspannungen auf einen Minimal- und Maximalwert geprüft werden und diesen nicht überschreiten sollen wird der Algorithmus so konfiguriert, dass diese im zulässigen Bereich bleiben.
- Der erste Teil von **B.4″** ist eine typische Rausch-Bedingung, die wie gesagt als sehr plausibel gelten kann. Wegen Bemerkung 4.45 ist für die zweite Bedingung hinreichend, dass $R_k^{\pm}$ und $f(\mathbf{X}_k \pm h_k \mathbf{D}_k)$ gleichmäßig varianzbeschränkt sind. Die Annahme an das Rauschen wurde schon in Bemerkung 4.50 als plausibel charakterisiert, und auch den zweiten Teil der Annahme kann man als erfüllt annehmen.

- f als dreimal stetig differenzierbar anzunehmen, erscheint vertretbar und auch die Beschränktheit der dritten Ableitung wirkt nicht problematisch. **B.5″** kann also als erfüllt angenommen werden.
- Mit der Wahl $\mathbf{D}_k$ BERNOULLI-verteilt, die man mit einem Zufallsgenerator erzeugt, während R_k aus dem physikalischen System stammendes Rauschen ist, scheint es sehr einsichtig, diese beiden Zufallsgrößen als unabhängig anzusehen.

Bis auf die Bedingungen, die man ohne die Kenntnis von f im Prinzip nicht nachprüfen kann, sind also alle Bedingungen erfüllt oder können vertretbar als erfüllt angenommen werden. Um die Aussage über die Erfülltheit der Konvergenzbedingungen wesentlich zu verbessern, müsste man f kennen. Dies ist also mit dem modellfreien Ansatz nicht möglich. Er ist aber nötig, da die Kenntnis von f aufgrund der physikalischen Gegebenheiten beim Wellenfrontsensor-losen Ansatz praktisch unmöglich ist.

Einordnung der FDSA-Bedingungen in der Anwendung

Für die Bedingungen des Konvergenzresultats für das FDSA-Verfahrens gilt zunächst – neben dem offensichtlichen Wegfallen der Bedingungen an die Perturbationen – das Gleiche wie für die Bedingungen an das SPSA-Verfahren. Zu den beiden Punkten, in denen die Bedingungen des FDSA-Verfahrens unwesentlich einfacher sind (vergleiche Bemerkung 4.40), ist noch zu bemerken, dass die Bedingungen an die Regularität für f_i''' natürlich ebenfalls unproblematisch sind, wenn sie es schon für alle dritten Ableitungen waren. Die Bedingung, dass das Rauschen in $\mathcal{L}^2$ beschränkt ist, kann als erfüllt angenommen werden.

Nun wird ein Optimierungsdurchgang dargestellt anhand dessen das Verhalten des SPSA-Verfahrens für die Anwendung am Beispiel des Laborsystems beschrieben wird.

6.6 Test-Optimierungsdurchgang

In diesem Abschnitt wird zunächst ein Optimierungsdurchgang dargestellt, dann bewertet und schließlich ein Ausblick auf die Anwendung des SPSA-Verfahrens für die Anwendung in der adaptiven Optik gegeben.

6.6.1 Test-Darstellung

Vor dem Start der Optimierung

Unter dem *Verlauf* der Optimierung soll im Folgenden die zeitliche Entwicklung des Zielfunktionswerts während eines Optimierungslaufs verstanden werden. Dieser kann pro Iteration oder pro Funktionsauswertung dargestellt werden. Letzteres ist sinnvoll, um die Güte des Verfahrens für die Anwendung beurteilen zu können, da diese Darstellung genau den zeitlichen Verlauf zeigt und man die Ergebnisse auf möglicherweise schnellere Hardware skalieren kann wie im Kapitel 1 beschrieben. Ein Optimierungsdurchgang besteht aus mehreren *Optimierungsläufen*, bei dem die Parameter nicht mehr geändert werden, sich das Rauschen und die Perturbation aber unterschieden und zu verschiedenen Ergebnissen führen. Außerdem ist es sinnvoll, dabei jeden Test-Funktionswert und nicht etwa nur den aktuell besten Zielfunktionswert $\tilde{f}(\mathbf{x}_k)$ aufzutragen. Dieses Verhalten entspricht dann dem im „laufenden Betrieb“. Wenn man Verschlechterungen verwirft, ist es sinnvoll, die aus der Anzahl der verworfenen Iterierten `worse` und der Anzahl der Iterationen $\overline{k}$ bestimmte Akzeptanzrate zu definieren:

$$\text{Akzeptanzrate} = 1 - \frac{\text{worse}}{\overline{k}}. \tag{6.5}$$

Damit die Ergebnisse repräsentativ sein können, sollte die Akzeptanzrate bei mindestens 75% liegen.

In den Diagrammen wird der Zielfunktionswert f als negativer Systemleistungs-/Metrikwert aufgetragen im Sinne der Umwandlung in ein Minimierungsproblem, siehe Lemma 4.1. Ansonsten wird aber vom Betrag des Systemleistungswerts gesprochen, der dann wieder anschaulich beschreibt, wieviel Prozent der Intensität im *Bucket* ist.

Start des Optimierungsdurchgangs

Als Systemleistungsmetrik wurde der Mittelwert der *Power-in-the-Bucket*-Werte (6.2) für virtuelle Aperturen mit Radien von 3, 7, 20, 30 und 55 Pixeln verwendet, das heißt die Zielfunktion

$$f(\mathbf{x}) \coloneqq -(P_{3,7,20,30,55\,\mathrm{px}} \circ f_2)(\mathbf{x}) \tag{6.6}$$

wird verwendet.

Der kleinste Radius ist dabei so gewählt, dass ein beugungsbegrenztes Bild prinzipiell möglich ist. Der theoretische Maximalwert des Mittelwerts der genannten *Power-in-the-Bucket*-Submetriken liegt dementsprechend zwischen 0.976 und 1 (wenn der Spiegel alle Aberrationen des Systems ausgleichen kann). Die Größe der Radien wurde so gewählt, dass der Spot auf dem Kamerabild zur Hälfte innerhalb und zur Hälfte außerhalb des Kreises liegt, was sich auch bei anderen Optimierungsdurchgängen als sinnvoll erwiesen hat, wie in Punkt I bei der Beschreibung der Abhängigkeit der Güte des Verlaufs erläutert wird.

Der Startwert ist jeweils das **0**-Spannungsmuster mit einem Startniveau von $\tilde{f}(\mathbf{x}) \approx -0.11$. Gewählt wurde die $q = 2$-4-SPSA-Variante des SPSA-Verfahrens und es werden fünf Optimierungsläufe verglichen.

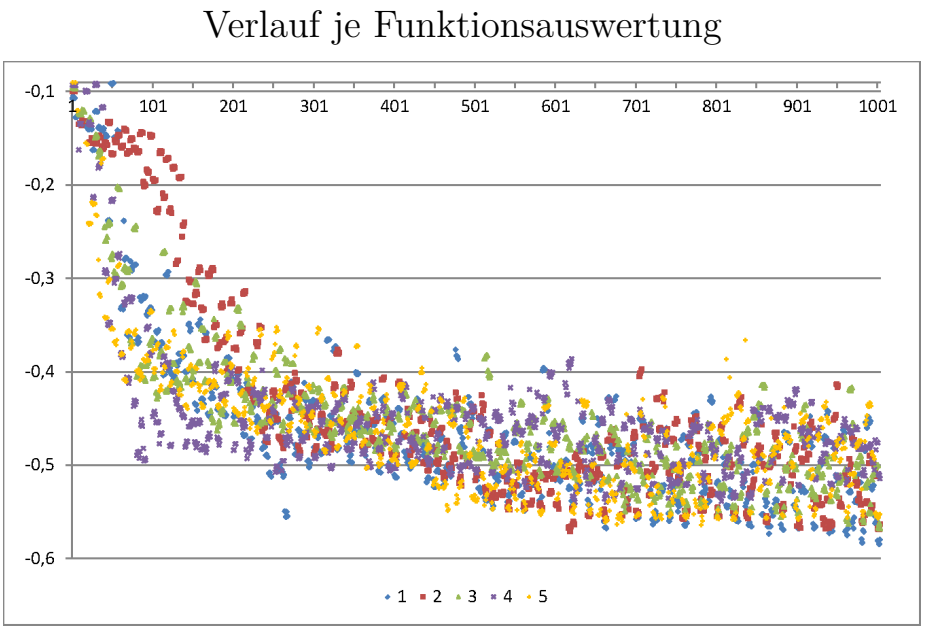
Verlauf und Endzustände des Optimierungsdurchgangs

Der Verlauf des Zielfunktionswerts während der Optimierung ist auf den Diagrammen in Tabelle 6.2 dargestellt, pro Funktionsauswertung und pro Iteration. In Tabelle 6.3 werden die Endzustände der fünf Läufe verglichen. Die Abbildungen zeigen dort einen Ausschnitt um den hellsten Bildbereich in der *Gradient*-3D-Darstellung. Zusätzlich sind die Metrikwerte am Ende der Optimierung angegeben sowie die Submetrik-Werte der zu den drei innersten Kreisen gehörigen virtuellen Aperturen mit Radien von 3, 7 und 20 Pixeln. Zum Vergleich ist auch die Akzeptanzrate angegeben.

In Tabelle 6.4 ist der Verlauf des 2. Optimierungslaufs näher dargestellt und die Entwicklung des Spots exemplarisch durch die Momentaufnahmen nach der 0., 5., 16. und 50. Iteration dargestellt. Tabelle 6.5 veranschaulicht für diesen Lauf auch die Entwicklung der Metrikwerte und der genannten Submetriken.

Die Zielfunktionswerte fallen in den ersten 13 Iterationen relativ zügig, danach sinken sie nur noch mit einem flachen Trend bei starken Schwankungen. Bei der Betrachtung der Kamerabilder, die die zeitliche Entwicklung wiedergeben, fällt auf, dass der „steile“ Bereich der Bewegung des Spots in die Bildmitte entspricht und der „flache“ Bereich der Verformung des Spots. Bei der Verformung des Spots, wenn dieser bereits in der Bildmitte ist, ändern sich die äußeren virtuellen Aperturen kaum noch, so dass visuell verschieden gut geformte Spots sich in ihren Metrikwerten nicht so drastisch unterscheiden. Letztlich sind die hier gewählte Metrik und die Algorithmusparameter mehr dafür geeignet, diese Positionierung des Spots in der Bildmitte und die grundlegende Formung des Spots zu erreichen. Für eine optimale Form des Spots müsste aber ein besser passender Durchgang mit entsprechend gewählter Metrik und Parametern durchgeführt werden. Dort sollte man dann eventuell auch mehr als 50 Iterationen betrachten. Das damit erreichbare Niveau kann dann noch vom Spiegel beschränkt sein, der die optimal korrigierende Spiegelform vielleicht nicht genau genug einstellen kann.

Die 3D-Bilder in den Tabellen 6.3 und 6.4 wurden mit dem entsprechenden National Instrument Labview 2010 Vision-Befehl erzeugt und sind in der Labview-Farbpalette *Gradient* wiedergegeben. Wie die Grauwerte in Farben umgesetzt werden, ist in [Tiv, S. 2-4] dokumentiert.



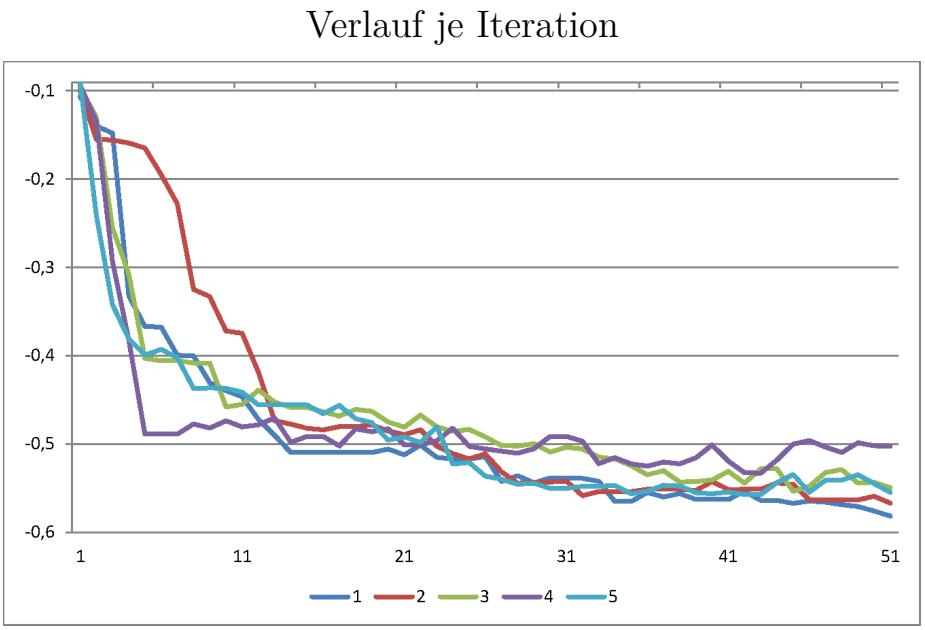


Tabelle 6.2: Verlauf des Zielfunktionswerts je Funktionsauswertung und je Iteration. Auf der y-Achse ist jeweils der Zielfunktionswert, also der negative gemessene Systemleistungswert aufgetragen. Auf der x-Achse des oberen Diagramms ist die Nummer der Funktionsauswertung, wobei 1 den Startwert bezeichnet aufgetragen. Im unteren Diagramm auf der x-Achse die Nummer der Iteration aufgetragen, wobei 1 den Startwert bezeichnet, 2 die erste Iterierte usw. Die verschiedenen Farben kennzeichnen die fünf unterschiedlichen Optimierungsläufe.

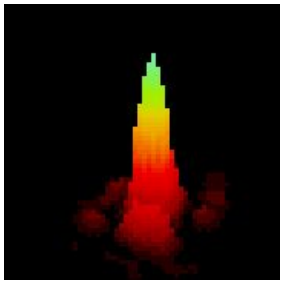
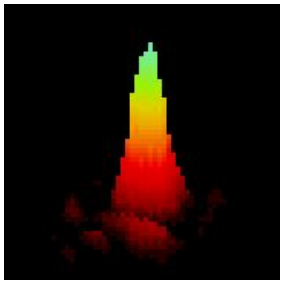
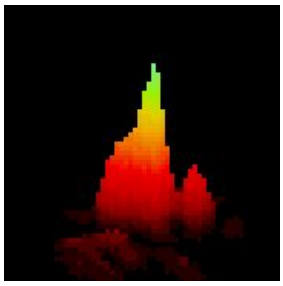
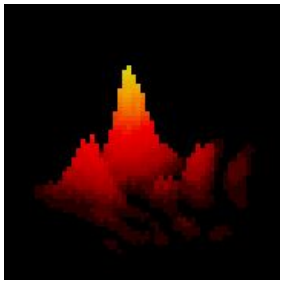
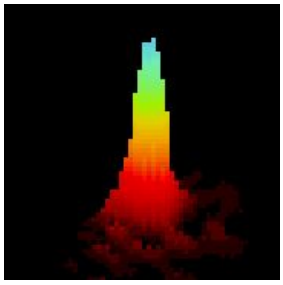
Lauf	Metrik	3	7	20		Akzeptanz-rate
1	0.58	0.22	0.42	0.7		78 %
2	0.57	0.19	0.47	0.68		88 %
3	0.55	0.15	0.43	0.69		92 %
4	0.51	0.06	0.28	0.68		92 %
5	0.56	0.2	0.43	0.67		82 %

Tabelle 6.3: Vergleich der Endzustände nach 50 Iterationen. Dargestellt werden jeweils die Metrikwerte $P_{3,7,20,30,55\,\mathrm{px}}$, die Werte der Submetriken von virtuellen Aperturen mit Radien von 3, 7 und 20 Pixeln, Ausschnitte aus dem Kamerabild des Laserspots und die Akzeptanzraten.

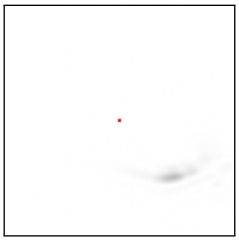
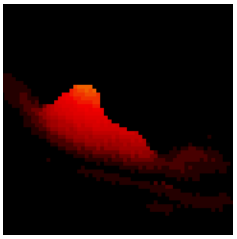
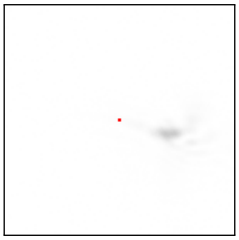
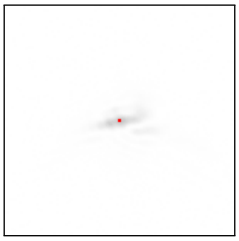
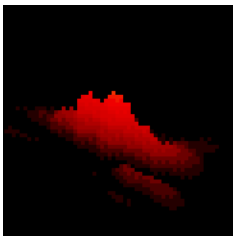
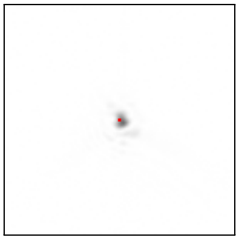
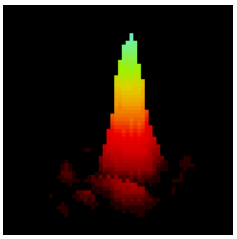
It.	Kamerabild	Negativ	3D-*Gradient*-Darstellung
0			
5			
16			
50			

Tabelle 6.4: Verlauf der Optimierung beim zweiten Lauf am Beispiel des Spots auf dem Kamerabild vor der Optimierung, nach der 5., 16. und 50. Iteration. Die ersten beiden Spalten sind in der Mitte des Kamerabildes zentriert, in der letzten ist nur die Form des Spots dargestellt.

Iteration	Metrikwert	$P_{3\,\mathrm{px}}$	$P_{7\,\mathrm{px}}$	$P_{20\,\mathrm{px}}$
0	0.11	0.00	0.00	0.00
5	0.19	0.01	0.02	0.07
16	0.48	0.09	0.26	0.59
50	0.57	0.19	0.47	0.68
Planspiegel		0.52	0.68	0.69

Tabelle 6.5: Verlauf des Metrikwerts und der *Power-in-the-Bucket*-Werte der innersten drei virtuellen Aperturen des in den Abbildungen von Tabelle 6.4 angegebenen Laufs 2. Durch Einsetzen eines Planspiegels mit einer ebenso großen Blende wie der des adaptiven Spiegels erhält man die unter Planspiegel zusätzlich angegebenen Werte.

Rauschen im Verlauf des Optimierungsdurchgangs

Die Größe des Rauschens während des Optimierungslaufs wurde untersucht. Ihre Entwicklung ist in Abbildung 6.7 dargestellt. Sie wird als Standardabweichung über 4 gemittelte Zielfunktionsauswertungen bei jeder neuen Iterierten bestimmt. Man erkennt, dass für den schlechtesten Lauf 4 auch das Rauschen im Vergleich hohe Werte aufweist. Insgesamt bewegt sich das Rauschen etwa in der Größenordnung $0 \ldots 0.04$, wobei der typische Wert in der Größenordnung 10^{-4} liegt.

Für einen weiteren Test, mit dem die Stärke der Schwankungen des Systems über den Zeitraum des Optimierungsdurchgangs hinweg geschätzt werden kann, legt man die optimale Spiegelstellung des ersten Laufs nach dem letzten Lauf noch einmal an. Der so am Ende neu ermittelte Metrikwert kann dann mit dem ursprünglich für diese Spiegelstellung ermittelten Wert verglichen werden.

Dabei ergab sich:

	Metrikwert	$P_{3\,\mathrm{px}}$	$P_{7\,\mathrm{px}}$	$P_{20\,\mathrm{px}}$
ursprünglich	0.58	0.22	0.42	0.70
am Ende neu ermittelt	0.55	0.20	0.41	0.67

Die Schwankung liegt also auf einem ähnlichem Niveau, wie es auch in Abbildung 6.7 an einigen Iterierten erreicht wird. Eine Schwankung im Bereich 0.03 im Zielfunktionsniveau kann also einer zeitlichen Veränderung des Laborsystems geschuldet sein. Man kann also z.B. die Güte von Lauf 1 und 5 aus Sicht der numerischen Optimierung als gleich ansehen.

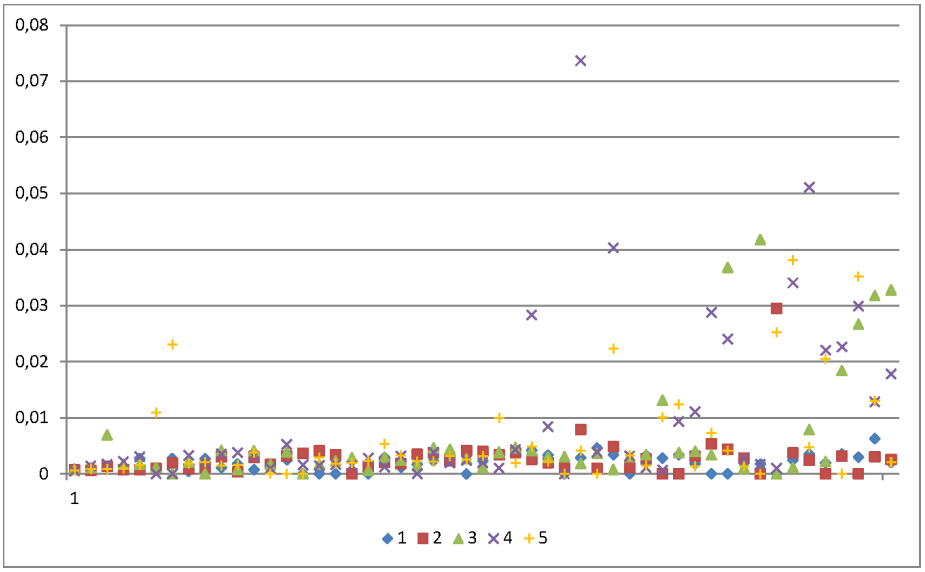


Abbildung 6.7: Auf der y-Achse ist die Größe des Rauschens, auf der x-Achse die Anzahl der Iterationen dargestellt. Das Rauschen wird je Iteration aufgezeichnet und aus der Standardabweichung der gemittelten Funktionswerte berechnet. Für den Fall verworfener Iterierter wird statt des Rauschens 0 aufgetragen.

6.6.2 Bewertung

Im Folgenden wird der zuvor vorgestellte Verlauf kurz bewertet.

(i) Der Algorithmus hat den Spot in die Mitte gezogen und dort dann Leistung konzentriert. Die breite Schwankung am Ende legt aber nahe, dass dort die Schrittweitenparameter noch verbessert werden könnten bzw. eine Metrikänderung angezeigt ist.

Für die Änderung der Schrittweitenparameter scheint sich eine Verkleinerung von h_k anzubieten. Eine zu starke Verkleinerung könnte aber wieder größere Instabilitäten bringen. Bei der Metrikänderung könnte man ein schrittweises Weglassen der äußeren Virtuellen-Apertur-Kreise durchführen.

(ii) Die wesentlichen Verbesserungen im Zielfunktionswert werden zu Beginn bis zur 300ten Test-Funktionsauswertung erzielt.

(iii) Dieser Test zeigt auch, dass die Läufe schon im Wesentlichen repräsentativ sind, alle auf ein ähnliches Niveau sinken (leider mit hoher Varianz dort) und im ersten Drittel des Verlaufs stark fallen.

6.7 Fazit

6.7.1 Abhängigkeit der Güte des Optimierungsverlaufs

Die Güte des Verlaufs der Optimierung hängt wesentlich von den drei folgenden Größen ab, die für ein gutes Optimierungsverhalten passend gewählt werden sollten: Startpunkt, Metrik und Schrittweitenfolgen. Dies wird nun näher ausgeführt:

I Startpunkt.
Für die Optimierung scheint es am günstigsten, wenn der Spot halb innerhalb und halb außerhalb einer kreisförmigen virtuellen Apertur liegt. Dies liegt wohl darin begründet, dass es einerseits kaum noch Funktionswertänderungen gibt, wenn der gesamte Spot bereits innerhalb der virtuellen Aperturen liegt und andererseits bei weit von der virtuellen Apertur entferntem Spot kleine Perturbationen auch nicht mehr Intensität in die Apertur bringen. Wegen dieser Überlegung entstand auch die Idee der multiplen virtuellen Aperturen. Mit diesen kann man Submetriken hinzufügen und die Radien passend wählen.

II Wahl der Metrik.
In der Anwendung im Labor-System war das Optimierungsziel vorgegeben, die Intensität des Spots in einer kreisförmigen virtuellen Apertur mit beugungsbegrenztem Radius zu konzentrieren. Es hat sich gezeigt, dass es auch bei fest vorgegebenem Optimierungsziel (wie der Maximierung des *Power-in-the-Bucket*-Werts $P_{2.9\,\mathrm{px}}$ hier) sinnvoll sein kann, andere Metriken zu wählen. Dies lässt die Wahl anderer Optimierungsziele, wie auf S. 14 erwähnt wurde unberührt.

Mit dem schon im Punkt I Gesagten kann man insbesondere neben dieser ursprünglichen Metrik weitere Submetriken hinzufügen, um ein gutes Verhalten des Algorithmus bis zur Zentrierung des Spots am gewünschten Punkt zu erreichen. Da der Einfluss der ursprünglichen Metrik durch Hinzunahme der Submetriken abnimmt, scheint man am Ende keine ausreichende Verbesserung erreichen zu können, was wiederum durch eine Metrikänderung mit Reinitialsierung oder durch geeignetere Schrittweitenwahlen erreicht werden könnte.

III Wahl der Schrittweitenparameter.

(a) Wahl der Iterierten-Schrittweitenfolge a_k
Eine grundlegende Frage bei den gradientenartigen Verfahren ist, wie weit man von der aktuellen Iterierten $\mathbf{x}_k$ aus der Abstiegsrichtung folgen sollte. Dies setzt sich aus der Größe der Gradientenschätzung $\hat{\mathbf{g}}_k(\mathbf{X}_k)$ und der Schrittweite a_k der k-ten Iteration zusammen: $\mathbf{X}_{k+1} - \mathbf{X}_k = a_k \hat{\mathbf{g}}_k^{\mathrm{SP}}(\mathbf{X}_k)$. Die Größe der Gradientenschätzung kann stark schwanken und kann jeweils stark vom Gradienten abweichen. Die Wahl einer sinnvollen Schrittweite a_k wird durch die semiautomatische Parameterbestimmung vereinfacht, da man dort nur den in der Anwendung anschaulicheren Wert c_1 bestimmen muss. Die Annahme, dass diese sinnvoll funktioniert, geht aber gerade wieder davon aus, dass sich die Größenordnung der Gradientenschätzung während

des gesamten Optimierungslaufs im selben Bereich bewegt, denn a_k sollte für jeden Iterationsschritt k weder zu groß noch zu klein sein. Gegebenenfalls muss man daher innerhalb des als sinnvoll angenommenen Bereichs von c_1 noch Feinabstimmungen vornehmen. c_1 muss auf verschiedene Metrikwahlen, Änderungen im physikalischen Aufbau usw. angepasst werden.

(b) Wahl der Gradientenschätzungs-Schrittweite h_k
Wählt man h_k zu klein, so ist die Gradientenschätzung sehr ungenau. Dieser Effekt ist stärker, als man vielleicht zunächst annimmt.

Aus der mathematischen Theorie der numerischen Differenation weiß man, dass man h nich zu klein wählen darf, um die relative Ungenauigkeit nicht zu verstärken, siehe Bemerkung 4.15. Gemäß Bemerkung 4.49 sollte man $h \approx \sqrt[3]{r}$ wählen, wobei r eine obere Schranke für das Rauschen ist

Das Rauschen lag im Laborsystem in der Größenordnung 10^{-4}, wozu es zunächst passend scheint, $h \approx 0.05$ zu setzen. Dies ist aber viel zu hoch, da h auf jeden Fall kleiner sein sollte als c_1 (die Steuersignaländerungen von einer Iteration zur nächsten sollen größer sein als die Änderungen bei den Testmessungen zur Gradientenschätzung). Damit bietet sich an, h auf höchstens c_1 oder einen gewissen Bruchteil davon zu setzen.

Auch das Verwenden des Verfahrens mit gelockertem Verwerfen von Verschlechterungen nach Abschnitt 5.3 nimmt etwas die Problematik aus der Schrittweitenwahl a_k heraus, da dann nur viele Schritte verworfen werden, anstatt dass das Verfahren divergiert.

Zusätzlich hängt die Güte des Verlaufs auch ab von:

1. der Größenordnung des Rauschens und

2. der Gutartigkeit der Abbildung f, die ja u.a. $\mathcal{C}^3$ sein sollte und beschränkte dritte Ableitungen haben sollte.

In einem festen System muss man das Rauschen und die Funktion f_1 als nicht beeinflussbar ansehen. Bei der Test-Darstellung in Abschnitt 6.6.1 wird auf das Rauschen beim Test-Optimierungsdurchgang unter anderem mit der Darstellung des Rauschens in Abbildung 6.7 eingegangen.

6.7.2 Vergleich mit anderen Verfahren

Der wesentliche Vorteil des SPSA- gegenüber des FDSA-Verfahrens ist, dass schneller neue Iterierte produziert werden und ohne langes Verharren an einer Iterierten zu besseren Funktionswerten an neuen Iterierten übergegangen wird. Die folgenden Zahlen verdeutlichen dies: Das FDSA-Verfahren würde bei ebenfalls 4-facher Mittelung der Funktionswerte für 212 (FD1) – 420 (FD2) Funktionsauswertungen an derselben Spiegelstellung ($\pm h_k$) verharren. Ein $q = 2$-4-SPSA-Verfahren hätte im selben Zeitraum die 10. bzw. 21. Iteratierte erreicht. Selbst bei einem FDSA-Verfahren ohne Mittelung der Funktionswerte würde das $q = 2$-4-SPSA-Verfahren Iteration 2 bzw. 5 erreichen, während das FDSA-Verfahren an einer Stelle verharrt.

Das SPSA-Verfahren eignet sich aufgrund der konstanten Anzahl von Funktionsauswertungen pro Iteration besser für höherdimensionale Probleme als das FDSA-Verfahren, bei dem die benötigten Anzahl linear mit der Dimension wächst. Bei der üblichen Wahl $q = 2$-SPSA verwendet man 4 Testmesspunkte. Damit zusammenhängend kann man bei „kleinen Problemen“ bis Dimension 5 das FDSA-Verfahren vorziehen, da die Gradientenschätzung besser ist und noch nicht viel kostet, und erst bei größeren Dimensionen auf das SPSA-Verfahren wechseln, vergleiche die Darstellung in Abbildung 6.8.

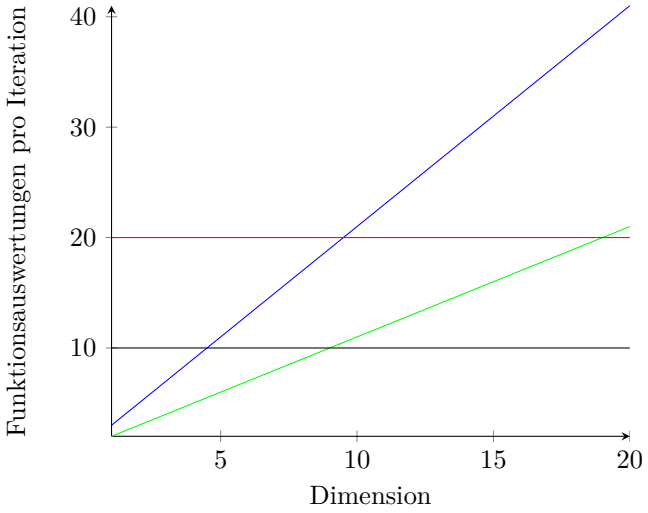


Abbildung 6.8: Dimensionsabhängiger Vergleich der benötigten Funktionsauswertungen pro Iteration von FDSA- und SPSA-Verfahren: Die blaue Linie stellt die $2m + 1$ nötigen Funktionsauswertungen pro Iteration des FDSA-Verfahrens mit FD2-Gradientenschätzer, die grüne die bei Verwendung von FD1 anfallenden Funktionsauswertungen pro Iteration dar. Die beiden konstanten Linien entsprechen dem SPSA-Verfahren mit $q = 2$ bei 2facher Mittelung der Funktionsauswertungen (schwarz) und für $q = 2$ bei 4facher Mittelung der Funktionsauswertungen (rot). Es werden also SPSA-Verfahren mit Mittelungen der Test-Zielfunktionsauswertungen verglichen mit FDSA-Verfahren ohne Mittelungen der Funktionsauswertungen.

6.7.3 Weitere mögliche Tests

In der Anwendung wäre noch sehr interessant zu prüfen, wie sich die theoretische Dimensionsunabhängigkeit des SPSA-Verfahrens in die Praxis überträgt, z.B. durch Austausch des vorhandenen Spiegels mit einem MEMS-Spiegel mit 1000 Aktuatoren.

Außerdem könnte man testen, ob das Verfahren auch mit weniger Funktionsauswertungen pro Iteration bei der gegebenen Höhe des Rauschens ein stabiles Verhalten aufweist, indem man z.B. $q = 2$-2-SPSA und $q = 1$-4-SPSA gegen das dargestellte $q = 2$-4-SPSA-Verfahren antreten lässt.

6.8 Ausblick für die Anwendung

6.8.1 Mögliche Erweiterungen des SPSA-Verfahrens

An dieser Stelle werden Ideen vorgestellt, wie man das SPSA-Verfahren als solches erweitern könnte.

Schrittweitensuche

Bei deterministischen Gradientenverfahren benutzt man üblicherweise statt einer festen Schrittweitenfolge eine Liniensuche, um die angesprochene Problematik der Wahl der Schrittweitenfolge a_k zu entschärfen. Nun ist gerade beim SPSA-Verfahren die Gradientenschätzung möglicherweise eine so ungünstige Richtung, dass es nicht lohnenswert erscheint, darauf noch viele Funktionsauswertungen zu verwenden. Das Verfahren mit einer einfachen Liniensuche und begrenzter Anzahl dafür verwendeter Funktionsauswertungen zu verbinden, wäre eine interessante Erweiterung.

Einbeziehung der „Geschichte“

Es finden sich die beiden folgenden Ideen, Informationen aus den vorangegangenen Iterationen weiterzuverwenden, zum einen

- die alten Gradientenschätzungen weiter zu benutzen, in der Form

$$\hat{\mathbf{g}}_k^H(\mathbf{X}_k) = \frac{1}{l+1}\big(\hat{\mathbf{g}}_k(\mathbf{X}_k) + \cdots + \hat{\mathbf{g}}_k(\mathbf{X}_{k-l})\big), \tag{6.7}$$

  und zum anderen

- über die Iterationen zu mitteln (*Iterate Averaging*, siehe u.a. [DR97]):

$$\widehat{\mathbf{X}}_k = \frac{1}{l+1}\big(\mathbf{X}_k + \ldots + \mathbf{X}_{k-l}\big). \tag{6.8}$$

Beim ersten Ansatz könnte $\hat{\mathbf{g}}_k^H(\mathbf{X}_k)$ ein „besserer“ Korrekturterm als die Gradientenschätzung $\hat{\mathbf{g}}_k^{\mathrm{SP}}(\mathbf{X}_k)$ sein, beim zweiten könnte die Iterierte $\widehat{\mathbf{X}}_k$ „besser“ sein als $\mathbf{X}_k$. Beides würde keine zusätzlichen Funktionsauswertungen kosten, macht aber eine sinnvolle Wahl von l nötig.

6.8.2 Mögliche andere Verfahren

Neben dem SPSA-Verfahren gibt es noch andere Ansätze, die man für diese Anwendung in der adaptiven Optik benutzen könnte.

Für eine modale Ansteuerung über Zernike-Koeffizienten (siehe S. 16) kann auch das FDSA-Verfahren verwendet werden. Erst bei höherdimensionalen Problemen lohnt es sich meines Erachtens zum SPSA-Verfahren zu wechseln (je nach Mittelung von FDSA und SPSA z.B. ab 5-10 Zernike-Moden, siehe auch Abbildung 6.8).

Sehr interessant wäre auch die Benutzung des *Simulated-Annealing*-Verfahrens. Dort sind die maximalen Änderungen zwischen den Schritten $\delta\mathbf{x}$ relativ

fest, es wird andererseits aber keine Gradienteninformation verwendet. Ich vermute daher, dass es weniger kritisch auf Parameterwahlen reagiert, dafür aber auch langsamer zu guten Metrikwerten konvergiert.

Eine weitere Frage wäre, ob man z.B. nichtlineare konjugierte Gradientenverfahren über eine Art *Simultaneous-Perturbation*-Gradientenschätzung für die zonale Ansteuerung nutzbar macht.

6.8.3 Schlusswort und mögliche Herangehensweise für ein adaptiv-optisches System

Die mathematische Theorie wurde behandelt, ein Optimierungsdurchgang und die Erkenntnisse aus den Laborsystem-Experimenten für die Anwendung dargestellt. Diese Arbeit schließt nun damit ab, darauf basierend einen Ausblick für eine mögliche Verwendung des Verfahrens in einem adaptiv-optischen System zu geben.

Zur Einbeziehung des SPSA-Verfahrens in ein solches System bei zonaler Ansteuerung scheint sich ein kombinierter Lauf aus

- einer Tip-Tilt-Vorkorrektur (d.h. Verschiebung des Spots in die Mitte), manuell, durch einfaches Durchtesten oder einen 3-dimensionalen modalen FDSA-Lauf,
- einem SPSA-Lauf mit semiautomatischer Parameterbestimmung
  (bei dem nur c_1 vorzugeben ist, eventuell kann man dies auch durch eine Meta-Optimierung bestimmen)
  und aus
- einer Reinitialisierung mit Metrikänderung
  (schrittweise Entfernung der äußeren virtuellen Aperturen)

anzubieten. Die Reinitialisierung kann dabei neben der festen Kopplung an die Iterationszahl auch adaptiv an den Verlauf des Zielfunktionswerts geknüpft werden.

Anhang A

Zusätzliche Notation

Weitere verwendete Notation wird angegeben.

Def. A.1 (Groß-O-Notation)**.** Seien a_k und b_k zwei reelle Zahlenfolgen. a_k ist eine asymptotische obere Schranke von b_k, Notation: $a_k \in O(b_k)$, wenn es ein $c > 0$ und ein $k_0 \in \mathbb{N}$ gibt, so dass $|a_k| \leq c\,|b_k|$ für alle $k \geq k_0$.

Für Funktionen gilt: $f \in O(g)$, wenn $\limsup_{x \to a} \left| \frac{f(x)}{g(x)} \right| < \infty$.

Bemerkung A.2. Ist $f(x) \leq cg(x)$, so ist $f(x) \in O(g)$, $x \to \infty$ denn
$\limsup \left| \frac{f(x)}{g(x)} \right| = \limsup \left| \frac{cg(x)}{g(x)} \right| = \limsup |c| = \lim |c| = c < \infty$.

Satz A.3 (Mittelwertsatz)**.** *Sei $f \in \mathcal{C}^1(\mathbb{R}^m, \mathbb{R})$ und $\mathbf{p} \in \mathbb{R}^m$. Dann gibt es ein $a \in (0,1)$, so dass*

$$f(\mathbf{x} + \mathbf{p}) = f(\mathbf{x}) + \nabla f(\mathbf{x} + a\mathbf{p})^T \mathbf{p}. \tag{A.1}$$

Im Folgenden werden der 1- und n-dimensionale Taylorsche Lehrsatz angegeben:

Satz A.4. *Sei $\mathbb{I} \subset \mathbb{R}$ ein Intervall, $f \in \mathcal{C}^3(\mathbb{I}, \mathbb{R})$. Dann gilt für alle $x \in \mathbb{I}$ und Entwicklungsstellen $a \in \mathbb{I}$:*

$$f(x) = f(a) + f'(a)(x-a) + \frac{1}{2} f''(a)(x-a)^2 + \textit{Restglied.} \tag{A.2}$$

Satz A.5. *Sei $f \in \mathcal{C}^3(M, \mathbb{R})$ ($M \subset \mathbb{R}^m$ offen). Dann gibt es für alle $\mathbf{v} \in M$ und Entwicklungspunkte $\mathbf{x} \in M$, für die die Verbindungsstrecke von $\mathbf{x}$ und $\mathbf{x} + \mathbf{v}$ in M liegt, ein $\mathbf{z}$ auf dieser Verbindungsstrecke, so dass*

$$f(\mathbf{x} + \mathbf{v}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^T \ \mathbf{v} + \frac{1}{2} \mathbf{v}^T \ \nabla^2 f(\mathbf{x}) \ \mathbf{v} + \sum_{|\mathbf{j}|=3} \frac{D^{\mathbf{j}} f(\mathbf{z})}{\mathbf{j}!} \mathbf{v}^{\mathbf{j}} \tag{A.3}$$

in Multiindex-Notation, wobei

$$\begin{aligned}
&\mathbf{j} = (j_1, \ldots, j_m) \in \mathbb{N}_0^m, \\
&D^{\mathbf{j}} f = \frac{\partial^{|\mathbf{j}|}}{\partial d_1^{j_1} \cdots \partial d_n^{j_n}} f, \\
&|\mathbf{j}| = j_1 + \cdots + j_m, \\
&\mathbf{j}! = j_1! \cdots j_m!, \\
&\mathbf{x}^j = d_1^{j_1} \ldots d_m^{j_m}.
\end{aligned}$$

Def. A.6. Eine Funktion $f : \mathbb{R}^m \to \mathbb{R}^m$ heißt LIPSCHITZ-stetig in einer offenen Menge $\mathcal{U} \subseteq \mathbb{R}^m$ mit LIPSCHITZ-Konstante L, falls

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L \|\mathbf{x} - \mathbf{y}\| \quad \forall\, \mathbf{x}, \mathbf{y} \in \mathcal{U}. \tag{A.4}$$

Def. A.7. Die Besselfunktionen erster Art $J_n(x)$, $n \in \mathbb{N}$, sind definiert als diejenigen Lösungen der Besselschen Differentialgleichung

$$x^2 \frac{d^2 y}{dx^2} + x \frac{dy}{dx} + (x^2 - n^2) y = 0, \tag{A.5}$$

die im Ursprung nichtsingulär sind.

Def. A.8. Man erklärt die Konvergenz $x \to A$, wobei $x(t) : \mathbb{R} \to \mathbb{R}^m$ eine Funktion und $A \subseteq \mathbb{R}^m$ eine Menge ist, durch

$$\lim_{t \to \infty} \operatorname{dist}\big(x(t), A\big) = 0,$$

wobei dist für $x \in \mathbb{R}^m$ und $A \subseteq \mathbb{R}^m$ wie folgt zu verstehen ist:

$$\operatorname{dist}(x, A) = \min_{y \in A} \|x - y\|.$$

Kurzzusammenfassung

Diese Diplomarbeit befasst sich mit stochastischen Gradientenverfahren als Methoden der numerischen Optimierung für die Anwendung in der adaptiven Optik. Die adaptive Optik wird insbesondere im Fall der Propagation von Licht durch die Atmosphäre zur Verbesserung der Abbildungseigenschaften eines optischen Systems verwendet. Als Verfahren der Wahl wird das *Simultaneous Perturbation Stochastic Approximation* Verfahren (SPSA-Verfahren) betrachtet, das aufgrund seiner geringen Anzahl von benötigten Zielfunktionsauswertungen pro Iteration sehr gut zu den Echtzeit-Anforderungen der Anwendung passt.

Neben der Darstellung des physikalischen Hintergrunds wird Wert auf eine fundierte Verankerung der Analyse der Verfahren in der Wahrscheinlichkeitstheorie gelegt. Die verwendeten Resultate u.a. der Martingaltheorie werden im Rahmen der Arbeit in die benötigte Form gebracht.

Als Hintergrund wird ein Einblick in die Theorie der deterministischen, ableitungsfreien Gradientenverfahren gegeben. Die Konvergenz des SPSA-Verfahrens wird auf ein Konvergenztheorem für *Stochastic-Approximation*-Verfahren von Kushner und Clark zurückgeführt. Dabei wurde in der Konvergenztheorie zunächst Spall gefolgt, aufgrund von wahrscheinlichkeitstheoretischen Überlegungen werden dann aber teilweise geänderte Voraussetzungen verwendet.

Der praktische Teil der Arbeit beschäftigt sich mit der Anwendung des Verfahrens in der adaptiven Optik am Beispiel eines Labor-Testsystems. Die Voraussetzungen der Theorie werden für die Anwendung gewürdigt. Zur Umsetzung des Verfahrens in einen konkreten Algorithmus wird auf die Parameterwahl und mögliche Erweiterungen eingegangen. Die Wahl der Parameter nach Spall wird um eine Überlegung aus der Theorie der numerischen Differentiation erweitert, so dass sich eine andere Empfehlung hinsichtlich der Wahl der Schrittweite für die Gradientenschätzung ergibt. Ein Optimierungsdurchgang wird dargestellt und bewertet und anschließend ein Ausblick für die mögliche Verwendung des Verfahrens in einem adaptiv-optischen System gegeben.

Schlagwörter: ableitungsfreie Verfahren, Abstiegsverfahren, Adaptive Optik, Echtzeit-Anforderungen, *Finite Differences Stochastic Approximation*, Gradientenschätzung, Gradientenverfahren, Kiefer-Wolfowitz-Methode, Martingaltheorie, Numerische Optimierung, *Simultaneous Perturbation Stochastic Approximation*, *Stochastic Approximation*, Stochastische Optimierung, *Stochastic Parallel Gradient Descent*.

Index

Tabellenverzeichnis

Abbildungsverzeichnis

Literatur

Bücher, Skripte und Vorlesungen

[GS06] Geoffrey Grimmett und David Stirzaker: *Probability and random processes*. 3. Aufl. Oxford [u.a.]: Oxford University Press, 2006.
URL: http://www.ulb.tu-darmstadt.de/tocs/185123112.pdf.

[KC78] Harold J. Kushner und Dean S. Clark: *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. New York, 1978.

[Kun07] Peter Kunkel: *Vorlesung Numerik I*. 2007.

[KY03] Harold J. Kushner und G. George Yin: *Stochastic approximation and recursive algorithms and applications*. Hrsg. von Springer. 2. Aufl. New York, Berlin, Heidelberg, 2003.
URL: http://swbplus.bsz-bw.de/bsz107340135kap-1.htm.

[Kön06] Wolfgang König: *Stochastische Prozesse II: Martingale und Brownsche Bewegung*. 2006.
URL: http://www.wias-berlin.de/people/koenig/StPrII.pdf.

[Kön08] Wolfgang König: *MASS- UND INTEGRATIONSTHEORIE*. 2008.
URL: http://www.wias-berlin.de/people/koenig/MIT.pdf.

[Kön09] Wolfgang König: *WAHRSCHEINLICHKEITSTHEORIE I und II*. 2009.
URL: http://www.wias-berlin.de/people/koenig/www/WT.pdf.

[LR79] R. G. Laha und V. K. Rohatgi: *Probability theory*. New York, 1979.

[NW06] Nocedal und Wright: *Numerical Optimization*. 2. Aufl. Springer-Verlag, 2006.

[Rao84] Malempati M. Rao: *Probability theory with applications*. Probability and mathematical statistics. Orlando [u.a.]: Acad. Pr., 1984.

[RW96] M. C. Roggemann und B. M. Welsh: *Imaging through Turbulence*. Boca Raton u.a., 1996.

[Sch79] Hubert Schwetlick: *Numerische Lösung nichtlinearer Gleichungen*. Hrsg. von VEB Deutscher Verlag der Wissenschaften. Berlin, 1979, S. 346.

[Spa05] James C. Spall: *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*. Hrsg. von Wiley. Hoboken, 2005.

[Tys00] Robert K. Tyson: *Introduction To Adaptive Optics*. Bellingham, 2000.

[Wil91] David Williams: *Probability with Martingales*. 1991.

Artikel usw.

[Bec11] Peter Becker: „Korrektur von Leichtbau-Membranspiegeln mittels aktiver Optik“. Magisterarb. Fachhochschule Koblenz, RheinAhr-Campus Remagen, 2011.
URL: http://elib.dlr.de/67963/.

[Bha+03] Shalabh Bhatnagar u. a.: „Two-timescale simultaneous perturbation stochastic approximation using deterministic perturbation sequences“. In: *ACM Trans. Model. Comput. Simul.* 13 (2 Apr. 2003), S. 180–209.
URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.81.8830&rep=rep1&type=pdf.

[Blu54] Julius R. Blum: „Multidimensional Stochastic Approximation Methods“. In: *Ann. Math. Statist.* 25.4 (1954), S. 737–744.
URL: http://projecteuclid.org/euclid.aoms/1177728659.

[Chi97] Daniel C. Chin: „Comparative study of stochastic algorithms for system optimization based on gradient approximations.“ In: *IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics* 27.2 (1997), S. 244–9.
URL: http://www.jhuapl.edu/SPSA/PDF-SPSA/Chin_Comparative_Study.PDF.

[Dip02] Jürgen Dippon: *Accelerated Randomized Stochastic Optimization*. 2002.
URL: http://projecteuclid.org/DPubS/Repository/1.0/Disseminate?view=body&id=pdf_1&handle=euclid.aos/1059655913.

[Dip98] Jürgen Dippon: „Asymptotische Entwicklungen des Robbins-Monro-Prozesses“. Diss. Universität Stuttgart, 1998.
URL: http://www.isa.uni-stuttgart.de/LstStoch/Dippon/Papers/habil.pdf.

[DR97] Jürgen Dippon und J. Renz: „Weighted means in stochastic approximation of minima“. In: *Society for Industrial and Applied Mathematics Journal on Control and Optimization (SICON)* (1997).
URL: http://www.jhuapl.edu/SPSA/PDF-SPSA/Dippon_Weighted_Means.PDF.

[Fab68] V. Fabian: „On asymptotic normality in stochastic approximation“. In: *Annals of Mathematical Statistics* (1968).
URL: http://projecteuclid.org/DPubS/Repository/1.0/Disseminate?view=body&id=pdf_1&handle=euclid.aoms/1177698258.

[Fog] Agner Fog: *Uniform random number generators in C++*.
URL: http://www.agner.org/random/ran-instructions.pdf.

[Ger99] László Gerencsér: „Convergence rate of moments in stochastic approximation with simultaneous perturbation gradient approximation and resetting“. In: *IEEE Transactions on Automatic Control* (1999), S. 894–905.

[Grü10] Karin Grünewald: „Jahresgang der optischen Turbulenz auf der Laserfreistrahlstrecke des DLR in Lampoldshausen“. 2010.
URL: http://elib.dlr.de/64946/.

[KW52] J. Kiefer und J. Wolfowitz: „Stochastic Estimation of the Maximum of a Regression Function“. In: *Annals of Mathematical Statistics* 23.3 (1952), S. 462–466.
URL: http://projecteuclid.org/DPubS?service=UI&version=1.0&verb=Display&handle=euclid.aoms/1177729392.

[Msea] *Bound for multi-index sum.* Mathematics. (Version 2. September 2011).
URL: http://math.stackexchange.com/q/61374.

[Mseb] *Combinatorics for multi-index set, how many elements does* $\{X^j : |j| = 3, j_l \neq 0\}$ *have?* Mathematics. (Version 8. September 2011).
URL: http://math.stackexchange.com/q/62815.

[Msec] *Conditional Expectation of function of two RVs, one measurable, one independent.* Mathematics. (Version 18. Oktober 2011).
URL: http://math.stackexchange.com/q/73353.

[Msed] *Detail in Conditional expectation on more than one random variable.* Mathematics. (Version 4. August 2011).
URL: http://math.stackexchange.com/q/55562.

[Msee] *(How) follows P(X=0)=0 if E(1/X) finite?* Mathematics. (Version 2. August 2011).
URL: http://math.stackexchange.com/q/55088.

[Msef] *How to show Martingale property for sum of* $S_k - E(S_k)$*-summands where* S_k *is a function of two RV's.* Mathematics. (Version 5. Oktober 2011).
URL: http://math.stackexchange.com/q/66619.

[Mseg] *Independence and Conditional Expectations for three random variables, (How) does X independent from (Y,Z) imply E(X/Y | Z)= E(X) E(1/Y|Z)?* Mathematics. (Version 18. August 2011).
URL: http://math.stackexchange.com/q/58265.

[Mseh] *Measurability question for martingale (Is* $S(X_i, Z_i) - E(S(X_i, Z_i) \mid \mathcal{F}_i)$ $\mathcal{F}_i = \{X_1, \ldots, X_i\}$*-measurable?)* Mathematics. (Version 4. Oktober 2011).
URL: http://math.stackexchange.com/q/69773.

[Msei] *Rule with independent random variables and conditional expectations.* Mathematics. (Version 4. August 2011).
URL: http://math.stackexchange.com/q/55524.

[Nol76] Robert J. Noll: „Zernike polynomials and atmospheric turbulence“. In: *Journal of the Optical Society of America* 66.3 (1976), S. 207–211.
URL: http://www.opticsinfobase.org/viewmedia.cfm?uri=josa-66-3-207&seq=0.

[Piaa] Didier Piau: *Independence and Conditional Expectations for three random variables, (How) does X independent from (Y,Z) imply E(X/Y | Z)= E(X) E(1/Y|Z)?* Mathematics. (Version 29. August 2011).
URL: http://math.stackexchange.com/q/60394.

[Piab] Didier Piau: *Is the norm of a martingale a martingale?* Mathematics. (Version 19. Oktober 2011).
URL: http://math.stackexchange.com/q/73980.

[PR07] Piotr Piatrou und Michael Roggemann: „Beaconless stochastic parallel gradient descent laserbeam control: numerical experiments". In: *Optical Society of America* (2007).
URL: http://www.opticsinfobase.org/viewmedia.cfm?uri=ao-46-27-6831&seq=0.

[RM51] Herbert Robbins und Sutton Monro: „A Stochastic Approximation Method". In: *The Annals of Mathematical Statistics* 22.3 (1951), S. 400–407.
URL: http://projecteuclid.org/DPubS/Repository/1.0/Disseminate?view=body&id=pdf_1&handle=euclid.aoms/1177729586.

[SC94] James C. Spall und John A. Cristion: „Nonlinear adaptive control using neural networks: estimation with a smoothed form of simultaneous perturbation gradient approximation". In: *American Control Conference*. Bd. 3. 1994, S. 2560 –2564.
URL: http://www3.stat.sinica.edu.tw/statistica/password.asp?vol=4&num=1&art=1.

[SC98] James C. Spall und John A. Cristion: „Model-free control of nonlinear stochastic systems with discrete-time measurements". In: *IEEE Transactions on Automatic Control* 43 (1998), S. 1198–1210.
URL: http://www.jhuapl.edu/SPSA/PDF-SPSA/SpallCristion_TAC98.pdf.

[Spa] James C. Spall: „An Overview of the Simultaneous Perturbation Method for Efficient Optimization". In: ().
URL: http://www.jhuapl.edu/SPSA/PDF-SPSA/Spall_An_Overview.PDF.

[Spa87] James C. Spall: „A Stochastic Approximation Technique for Generating Maximum Likelihood Parameter Estimates". In: *American Control Conference, 1987*. Juni 1987, S. 1161 –1167.
URL: http://www.jhuapl.edu/spsa/PDF-SPSA/Spall_A_Stochastic_Approximation.PDF.

[Spa92] James C. Spall: „Multivariate Stochastic Approximation Using a Simultaneous Perturbation Gradient Approximation". In: *IEEE Transactions on Automatic Control* 37 (1992), S. 332–341.
URL: http://citeseer.ist.psu.edu/viewdoc/download;jsessionid=11974F120819C42859A52467B2F809FB?doi=10.1.1.19.4562&rep=rep1&type=pdf.

[Spa98] James C. Spall: „Implementation Of The Simultaneous Perturbation Algorithm for Stochastic Optimization“. In: *IEEE Transactions on Aerospace and Electronic Systems* 34.3 (1998), S. 817–823.
URL: http://www.jhuapl.edu/spsa/PDF-SPSA/Spall_Implementation_of_the_Simultaneous.PDF.

[Tiv] *IMAQ Vision User Manual*. May 1999 Edition. National Instruments. 1999.
URL: http://www.ni.com/pdf/manuals/322320b.pdf.

[Vor] Stand 21.11.2011.
URL: http://www.isr.umd.edu/faculty/gateways/vorontsov.htm.

[Vor+00] Mikhail A. Vorontsov u. a.: „Adaptive optics based on analog parallel stochastic optimization: analysis and experimental demonstration“. In: *Journal of the Optical Society of America* 17 (2000), S. 1440–1453.

[VS98] Mikhail A. Vorontsov und V.P. Sivokon: „Stochastic parallel-gradient-descent technique for high-resolution wave-front phase-distortion“. In: *Journal of the Optical Society of America A* 15 (1998), S. 2745–2758.
URL: http://www.opticsinfobase.org/viewmedia.cfm?uri=josaa-15-10-2745&seq=0.

[WC98] I.-J. Wang und Edwin K.P. Chong: „A deterministic analysis of stochastic approximation with randomized directions“. In: *IEEE Transactions on Automatic Control* 43.12 (Dez. 1998), S. 1745 – 1749.
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=736077.

[WCK96] I-Jeng Wang, Edwin K. P. Chong und Sanjeev R. Kulkarni: „Equivalent necessary and sufficient conditions on noise sequences for stochastic approximation algorithms“. In: (1996), S. 784–801.
URL: http://www.cs.jhu.edu/~ijwang/pub/aap.pdf.

[Wik11] Wikipedia: *Seeing — Wikipedia, Die freie Enzyklopädie*. (Stand 19. Mai 2011). 2011.
URL: http://de.wikipedia.org/w/index.php?title=Seeing&oldid=84752178.

[Zin+] Martin A. Zinkevich u. a.: *Parallelized Stochastic Gradient Descent*. Sunnyvale.
URL: http://books.nips.cc/papers/files/nips23/NIPS2010_1162.pdf.

Erklärung

Ich versichere, dass ich die vorliegende Arbeit selbstständig und nur unter Verwendung der angegebenen Quellen und Hilfsmittel angefertigt habe, insbesondere sind wörtliche oder sinngemäße Zitate als solche gekennzeichnet.
Mir ist bekannt, dass Zuwiderhandlung auch nachträglich zur Aberkennung des Abschlusses führen kann.

Ort Datum Unterschrift

Abstract

This diploma thesis deals with Stochastic Gradient Descent methods of Numerical Optimization for applications in Adaptive Optics. Adaptive Optics is used in particular in the case of propagation of light through turbulent atmosphere to improve the performance of an optical system. It is choosen to consider Simultaneous Perturbation Stochastic Approximation (SPSA) more closely as it fits very well to the real-time-requirements of the application with its low amount of required function evaluations per iteration.

In addition to the presentation of the physical background, the well-grounded basis of the theoretical analysis in probability theory is emphasized. The required shapes are given to utilized results of martingale theory and others.

As a background, an insight in the theory of deterministic, gradient-free gradient-descent methods will be presented. Convergence of SPSA will be traced back to a convergence theorem for Stochastic Approximation methods by Kushner and Clark. In doing so, initially Spall's convergence theory was followed. Because of probability theory related considerations however, partially changed conditions have been used.

The practical part of this diploma thesis concerns the application of the regarded method in Adaptive Optics using the example of an laboratory set-up. The conditions of the theory are evaluated for the application. For implementing the method as a practical algorithm we go into detail about choice of parameters and extensions. The parameter choice according to Spall is extended with an consideration from the theory of Numerical Differentiation, which leads to a different recommendation for the choice of the gradient-estimation gain sequence. Optimization runs are presented and evaluated, and subsequently we give an outlook to an possible inclusion of SPSA in an adaptive optics system.

Keywords: ableitungsfreie Verfahren, Abstiegsverfahren, Adaptive Optik, Echtzeit-Anforderungen, gradient estimation Gradientenschätzung, Gradientenverfahren, Kiefer-Wolfowitz-Methode, Kiefer Wolfowitz procedure, Martingaltheorie, Numerische Optimierung, Stochastische Optimierung, Adaptive Optics, descent methods, Finite Differences Stochastic Approximation gradient descent methods, gradient-free methods, martingale theory, Numerical Optimization, real-time requirements, Simultaneous Perturbation Stochastic Approximation, Stochastic Approximation, Stochastic Optimization, Stochastic Parallel Gradient Descent.