Betrachtungen zur Wirkungsweise von Σ-Δ-Modulatoren

von
Dipl.-Math. Bob Metzger

Der Fakultät für
Mathematik und Naturwissenschaften
der Technischen Universität Ilmenau
zur Erlangung des akademischen Grades

Doctor rerum naturalium
(Dr. rer. nat.)

vorgelegte Dissertation.

Berichter : Prof. Dr. Bernd Marx, Technische Universität Ilmenau
Prof. Dr. Armin Hoffmann, Technische Universität Ilmenau
Prof. Dr. Martin Bohner, Florida Institute of Technology, Melbourne, Florida

Tag der wissenschaftlichen Aussprache: 28. November 2001

Jena, am 29. November 2001

Für N.D. und A.N.

Inhaltsverzeichnis

1 Einleitung

Bei der Entwicklung elektronischer Baugruppen spielen neben rein digitalen in zunehmendem Maße auch analoge und gemischt analog-digitale Strukturen eine Rolle. Für die Realisierung digitaler Funktionen existieren bereits leistungsfähige Automatisierungslösungen. Die automatische Generierung analoger Module ist jedoch bisher nur für ausgewählte Funktionen möglich. Unter dem Aspekt der Automatisierung des Entwurfs ist es somit wichtig, solche elektronischen Strukturen zu untersuchen und zu modellieren. Dabei soll die Realisierung einer Lösung ausgehend von den Anforderungen eines technischen Systems generiert werden. Die Anforderungen können folgender Natur sein: Einfachheit im Aufbau, Minimierung der Kosten, hohe Robustheit gegenüber äußeren Einflüssen und hohe Effizienz durch Parameteroptimierung.
Wegen guter Integration in monolithische ICs und den Erfordernissen der Prozeßsignalverarbeitung gut angepaßter Wandlungsverfahren hat sich die Wandlung nach dem Σ-Δ-Prinzip durchgesetzt. Dieses Verfahren kann als ein zeitdiskreter Regelkreis, bei dem der Systemzustand zu festgelegten Zeiten abgefragt wird, aufgefaßt werden. In der Literatur wird es als PCM-Verfahren (pulskodierte Modulation) bezeichnet. Das Eingangssignal wird in eine hochfrequente Bitfolge transformiert. Die ursprüngliche Information (z. B. Periode, Amplitude) ist dann in der Abfolge zweier Werte codiert. Vergleichsweise ist die Informationscodierung beim AM- und beim FM-Band im Radio. Hierbei ist eine hohe Auflösung ohne Einsatz aufwendiger Präzisionselemente möglich, was sicherlich zu ihrer Verbreitung beigetragen hat. Die mit Σ-Δ-Modulatoren erreichbaren hohen Auflösungen beruhen auf einer Kombination der Prinzipien Überabtastung (oversampling), Rauschformung (noise shaping) und Dezimation [28]. Da hier jedoch analoge und digitale Systemteile miteinander rückgekoppelt sind und das System teilweise abgetastet betrieben wird, war bisher eine geschlossene Beschreibung des Σ-Δ-Konverters nicht möglich. Die in der ingenieurtechnischen Literatur angegebenen Modelle formulieren Beschreibungen, welche von einer Linearisierung des Gesamtsystems ausgehen, bei der die Übertragungsfunktion des Konverters in eine Signal- und in eine Rauschsignalübertragungsfunktion im z-Bereich zerlegt wird [3]. Dabei wird das Quantisierungsrauschen ausschließlich als eine von dem Eingangssignal unabhängige Quelle mit speziellen stochastischen Eigenschaften vorausgesetzt. Das exakte zeitliche Verhalten von idealisierten Σ-Δ-Konvertern n-ter Ordnung kann dagegen durch nichtlineare nichtautonome Differenzengleichungen n-ter Ordnung beschrieben werden [30], [60]. Die unstetige Nichtlinearität besteht hierbei in einer Heavisidefunktion (Einheitssprung) auf der rechten Seite der Gleichung, deren Argument selbst Differenzen $(n-1)$-ter Ordnung enthält. In der mathematischen Literatur über Differenzengleichungen bzw. -ungleichungen [1], [2], [6], [19], [40], [44] und [49] bzw. in den Tagungsbänden der Konferenzen zu aktuellen Forschungen über Differenzengleichungen [7] - [10] konnten wir keinen Artikel finden, der eine ähnlich gelagerte Fragestellung zum Inhalt hat. Betrachtete Nichtlinearitäten sind hier differenzierbar oder wenigstens noch Lipschitz-stetig.
Reales Verhalten der Konverter, wie zum Beispiel Nichtlinearitäten der Bauelemente im Arbeitsbereich oder durch Temperaturänderungen, werden in dieser Arbeit nicht berücksichtigt. Die Ordnung der Differenzengleichungen entspricht der Anzahl der Integratoren. Der Vorteil einer Modellierung im Zeitbereich liegt darin, daß keine Linearisierung des Systems vorgenommen zu werden braucht. Der Nachteil liegt jedoch im hohen Aufwand, der zur Lösung bzw. Näherungslösung dieser Differenzengleichungen nötig ist (vgl. Kapitel 4). In der ingenieurtechnischen Literatur basiert die Modellbildung im Zeitbereich auf drei Komponenten: analytsiche Betrachtungen, geometrische Untersuchungen und numerische Simulation, z. B. [22], [66], [64].
In der vorliegenden Arbeit untersuchen wir das Verhalten von Σ-Δ-Konvertern im Zeitbereich. Daran schließt sich eine digitale Filterung (allgemein [26], [54]) durch geeignete Mittelwertbildung (konkret [33], [36]) als Approximation des Eingangssignals durch intervallweise konstante

Funktionen an. Das System Konverter plus Filter nennen wir Σ-Δ-Modulator. Wir unterscheiden zeitlich „lokales Verhalten“, welches seine Entsprechung in der Fehleranalyse des Konverters bzw. des Modulators durch Lösung oder Näherung der Differenzengleichung bzw. abgeleiteter gewöhnlicher Differentialgleichung hat, und „globales Verhalten“, welches in der Untersuchung zur Beschränktheit des Systems und Periodizität sowie Stabilität der Lösung der Differenzengleichung zum Ausdruck kommt. Die Beschreibung des Konverters erster Ordnung ist durch Lösung der zugehörigen Differenzengleichung unter zusätzlicher Beschränktheitsannahme vollständig gelöst [30]. Dagegen kann in dieser Arbeit keine explizite Lösung der Differenzengleichung des Konverters zweiter Ordnung angegeben werden. Darüber hinaus betrachten wir hier keine kaskadierten Konverter, z. B. [52], [60]. Die Arbeit ist wie folgt strukturiert.

Im Kapitel 2 stellen wir die Wirkungsweise der Bauelemente, aus denen Σ-Δ-Konverter aufgebaut sind, vor und leiten die zugehörigen Differenzengleichungen der Konverter erster und zweiter Ordnung her. Diese Herleitung läßt sich auch auf Konverter höherer Ordnung übertragen. Im Falle des Konverters erster Ordnung können wir eine explizite Lösung angeben, die nur vom Eingangssignal und vom Taktindex abhängt. Anschließend diskutieren wir den Zusammenhang zwischen Konverterausgang und dieser expliziten Lösung falls die Randbedingungen, unter denen diese Lösung gilt, verletzt sind [31]. Es zeigt sich, daß diese Regelkreise ein „regeneratives Moment“ in sich tragen, d. h., sie sind in der Lage, in einen beschränkten Zustandsbereich zurückzukehren, falls die entsprechenden Randbedingungen lange genug eingehalten werden. Für den Konverter zweiter Ordnung können wir keine explizite Lösung angeben. Jedoch ergibt sich eine einfachere Differenzengleichung, falls wir Randbedingungen für den inneren Kreis analog der für den Konverter erster Ordnung annehmen. Den Fehler der abgeleiteten Näherungslösung schätzen wir nach oben ab. Diese selbst ist aber aus technischen Gründen für einen Filtervorschlag ungeeignet.

Im dritten Kapitel betrachten wir einen digitalen Filter für einen Konverter erster Ordnung, der auf dem Prinzip der Trapezmittelung basiert [28]. Die Differenz zwischen Komparatoreingang und -ausgang wird als Quantisierungsfehler bezeichnet. In der Modellbildung werden an diesen häufig stochastische Annahmen gestellt, z. B. [46]. Ziel ist es, in Abhängigkeit der Wortlänge der äußeren eine geeignete Wortlänge der inneren gleitenden Mittelung mit minimalem Fehler anzugeben. Gütekriterium hierfür ist die Streuung des Fehlers des Mittelwertbildners bei den verschiedenen stochastischen Annahmen. Wir unterscheiden zwei Fälle. Zum einen ist der Quantisierungsfehler eine Folge stochastisch unabhängiger Zufallsvariablen. Diese Voraussetzung wird in der ingenieurtechnischen Literatur häufig gestellt. Zum anderen sei er eine Folge von Zufallsvariablen, welche von einer einzigen gleichverteilten Zufallsvariablen erzeugt werden [29], [31]. Folgende wesentliche Aussagen lassen sich dann beweisen. Die beste Wahl der inneren Wortlänge ist die nächste ganze Zahl der Hälfte der äußeren Wortlänge. Im zweiten Fall (eine erzeugende Zufallsvariable) ist der Quantisierungsfehler asymptotisch unkorreliert. Darüber hinaus ist die Streuung im zweiten Fall nach unten und oben durch die im Fall der stochastischen Unabhängigkeit und Gleichverteilung abschätzbar.

In dem folgenden Kapitel 4 untersuchen wir das Verhalten von Konvertern erster und zweiter Ordnung, falls die Taktzeit gegen Null konvergiert [36]. Durch diesen schwachen Grenzprozeß „konvergiert“ ein zeitdiskreter Regelkreis gegen ein analoges System, welches durch eine Differentialgleichung beschrieben werden kann. Dabei geht eine Ordnung verloren, d. h., ist der Konverter n-ter Ordnung, dann ist die abgeleitete Differentialgleichung von der Ordnung $n-1$. Der vorgestellte Algorithmus zur Herleitung der Differentialgleichungen ist auf beliebige Konverter anwendbar. Da der Grenzprozeß schwacher - also integraler - Natur ist, verlieren wir die Beschreibungsmöglichkeit des Konverterverhaltens im Kleinen, d. h. punktweise. Dafür erhalten wir aber ein analoges System, aus dem sich leicht grundlegende Eigenschaften [45] des analogen Regelkreises ableiten lassen. Dies ist im Zusammenhang mit den Kapiteln 7.1 und 8.5 zu sehen.

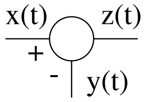
Weiterhin leiten wir im Kapitel 4 durch eine numerische Differentiation - sie ist gleicher Ordnung wie die entsprechende Differenzengleichung - des Komparatoreingangs einen Filtervorschlag ab. Dabei ist der Filterfehler bez. des integralen Mittelwertes des Eingangssignals von der Ordnung $O\left(\frac{1}{m^n}\right)$, wobei m die Wortlänge und n die Ordnung des Konverters bezeichnet. Schließlich leiten wir unter Annahme der Lipschitzstetigkeit des Eingangssignals und der Abschätzung der Lösung einer Volterraschen Integralgleichung zweiter Art hinreichende Bedingungen dafür her, daß der innere Kreis des Konverters zweiter Ordnung wie einer erster Ordnung arbeitet.
Ziel einer Filterung in unserem Fall ist eine digitale Rekonstruktion des Eingangssignals mit minimalem Fehler. Ein erster Versuch im Kapitel 2, dies für einen Konverter zweiter Ordnung über eine Näherungslösung der Differenzengleichung zu tun, erwies sich als ungeeignet. Im Kapitel 5 wollen wir nun eine weitere Näherungslösung und einen zugehörigen Filtervorschlag herleiten. Ausgangspunkt hierfür ist das Argument der Heavisidefunktion, in der ja auch Terme der zweiten Differenzen enthalten sind. Dieses Argument liegt in einem beschränktem Gebiet um Null, falls der Systemzustand des Konverters in seinem Arbeitsbereich ist. Durch die Lösung der homogenen Gleichung kann man von der Struktur der Lösung Rückschlüsse auf die der allgemeinen Differenzengleichung ziehen. Die durch diese Überlegungen gewonnenen Ersetzungsregeln werden auf die ursprüngliche Gleichung übertragen, um eine Näherungslösung für den Konverter zweiter Ordnung zu erhalten. Durch Superposition der Lösung, Abschätzung der Lösung des Problems mit unbestimmter aber beschränkter rechter Seite (erst dadurch kann die Heavisidefunktion aufgelöst werden) und der Annahme der Lipschitzstetigkeit des Eingangssignals ergibt sich eine Näherungslösung der originalen Differenzengleichung, deren Fehler wir abschätzen können. Aus dieser und ihrer zugehörigen Fehlerabschätzung leiten wir Filtervorschläge ab. Dabei nehmen wir an, daß das Eingangssignal auf jedem Intervall über $2m$ Takte näherungsweise linear bzw. quadratisch ist.
Im Kapitel 6 gehen wir zu einer geometrisch motivierten Betrachtungsweise des Konverters zweiter Ordnung über und untersuchen das zu einem konstanten Eingangssignal gehörige U_1-U_2-Phasenporträt, z. B. [55]. Dabei bezeichnen U_1 und U_2 die beiden Integratorausgänge. Ziel dieser Herangehensweise ist es, Aussagen über Beschränktheit des Systemzustandes und Periodizität des Konverterausgangs zu treffen, z. B. [24], [12], [13]. Einleitende Resultate geben darüber Auskunft, wann weder Beschränktheit noch Periodizität auftreten können. Das Hauptresultat besteht in der Angabe eines einfach zusammenhängenden Gebietes in der U_1-U_2-Ebene in Abhängigkeit von der Taktzeit, welches einmal erreicht bei konstantem Eingang nicht mehr verlassen wird. Weiterhin charakterisieren wir n-periodische Lösungen, welche nur bei Erfülltsein einer Rationalitätsbedingung auftreten können. Da die rationalen Zahlen das Maß Null haben, treten periodische Lösungen praktisch nicht auf. Darüber hinaus läßt sich zeigen, daß es in der Nähe n-periodischer Lösungen weitere n-periodische und auch nicht-n-periodische Lösungen liegen. Ob diese nicht-n-periodischen Lösungen überhaupt periodisch sind, haben wir hier nicht untersucht. Der Nachteil dieser Betrachtungen liegt darin, daß nur konstante Funktionen betrachtet werden.
In dem daran anschließenden Kapitel 7 wollen wir die Auswirkungen von Störungen auf das Systemverhalten untersuchen. Wir unterscheiden dabei zwei Arten von Störungen. Zum einen hängt der Konverterausgang von der Anfangslage des Systems ab. Zum anderen können sich die Systemparameter verändern. Den Zusammenhang zwischen der Lösung des ungestörten und des gestörten Systems im ersten Fall bezeichnet man als Stabilität. Und liegt im ungestörten System Stabilität vor, so charakterisiert Robustheit die Art der Störung der Systemparameter, unter denen Stabilität erhalten bleibt. Wir geben einen Weg vor, wie man bei Beschreibung des Modells durch Differentialgleichungen Stabilität [42], [65], [20], [53] des ungestörten Systems und den entsprechenden Stabilitätsradius [57] berechnen kann. Für die Charakterisierung des Modells durch Differenzengleichungen [40], [69] führen wir zwei weitere aber schwächere Begriffe ein,

K-Stabilität und K-Robustheit. Beide sind ein Maß dafür, wie stark die Lösung des ungestörten Systems in Zusammenhang zur Lösung des gestörten Systems steht. Folgende Resultate werden wir hier beweisen. Wegen der Unstetigkeit der rechten Seite in der Differenzengleichung ist kein Σ-Δ-Konverter im allgemeinen stabil noch robust gegenüber Störungen. Für einen Konverter erster Ordnung haben wir jedoch immer K-Stabilität und K-Robustheit wenigstens für integrierbare Eingangssignale. K-Robustheit gilt im allgemeinen nicht. Für einen Konverter zweiter Ordnung gilt weder diese schwächere Stabilität noch Robustheit. Offen hierbei ist die Frage, ob möglicherweise Modulatoren stabil bzw. robust sind. Bisher hatten wir nur das Verhalten von Konvertern untersucht. Ändert die digitale Filterung etwas am qualitativen Verhalten von Σ-Δ-Konvertern bez. Stabilität und Robustheit? Wir vermuten, daß diese Frage bejaht werden kann.
Im letzten Kapitel möchten wir die Modellbeschreibung darstellen, wie sie in der ingenieurtechnischen Literatur häufig vorkommt, z. B. [25], [46], [59]. Diese erhebt keinen Anspruch auf Vollständigkeit, wenn man die Fülle von Publikationen auf diesem Gebiet berücksichtigt. Eine Referenz zu dieser Thematik ist [52]. Wir wollen die in diesem Zusammenhang auftretenden Begriffe und Vorgehensweisen zur Modellbildung erläutern. Grundsätzlich gibt es zwei Möglichkeiten des Ansatzes, die Darstellung im Zeit- [55] und im Frequenzbereich [25]. Jede von ihnen hat Vor- und Nachteile [51]. Beim Modell im Zeitbereich bleiben die Nichtlinearitäten analog zu unserem Modell erhalten. Durch die dadurch auftretenden Schwierigkeiten tritt die analytische Untersuchung gegenüber numerischer Simulation in den Hintergrund. Der Zugang im Frequenzbereich dagegen benutzt folgende Linearisierung des Modells [3]. Der Komparator wird durch eine Summierstelle ersetzt, wobei der Quantisierungsfehler als eine äußere Quelle, die stochastisch unabhängig vom Eingangssignal ist, aufgefaßt wird. Durch weitere spezielle stochastische Annahmen an den Quantisierungsfehler steht hier nun die Analyse im Vordergrund vor der Simulation, z. B. [16]. Wir haben in diesem Kapitel noch einen weiteren Abschnitt aufgenommen, der in der technischen Literatur noch nicht in Zusammenhang mit Σ-Δ-Konvertern gebracht wurde. In Kapitel 4 hatten wir Differentialgleichungen hergeleitet, so daß Σ-Δ-Konverter auch als analoges System beschrieben werden können. Für solche analogen Regelkreise haben wir hier einige Begriffe der Systemtheorie [20], [45] definiert und ihr Wirken bei Konvertern dargestellt.
Diese Arbeit richtet sich gleichermaßen an Mathematiker, die sich für nichtlineare Differenzengleichungen interessieren, wie auch an Ingenieure, die sich im Bereich der Signalprozeßverarbeitung mit Σ-Δ-Modulatoren beschäftigen.
Mein Dank richtet sich zu allererst an meine beiden Betreuer, Prof. Marx und Prof. Hoffmann, für ihre fachlichen Tips und ihre immerwährende Bereitschaft, meine Problemchen in bezug auf diese Arbeit anzuhören und mit mir darüber zu diskutieren. Für die fachlichen Gespräche, Ideen und Anregungen danke ich allen Angehörigen des Fachbereiches Analysis und Systemtheorie, hier insbesondere Boris Thomaschewski und Thomas Wagenknecht. Darüber hinaus gilt mein Dank Ulrich Steinmetz, Daniel Görsch, Frank Schilder, Norbert Hirt und Thomas Neubert. Für die finanzielle Unterstützung, ohne die diese Arbeit nicht möglich gewesen wäre, danke ich der Deutschen Forschungsgemeinschaft für das mir gewährte Stipendium, meinen Eltern und Prof. Ilchmann.

2 Modellbeschreibung durch Differenzengleichungen

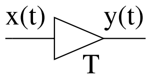
Bei der Beschreibung von Analog-Digital-Konvertern ist eine symbolische Darstellung des Aufbaus der Schaltungen zweckmäßig. In Anlehnung an die Symbolik in der ingenieurtechnischen Literatur seien im folgenden die einzelnen Elemente von Σ-Δ-Konverter und ihre Wirkungsweise erklärt:

Summierer



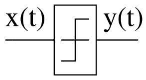
$$z(t) = x(t) - y(t)$$

Integrierer



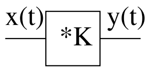
$$y(t) = \frac{1}{T} \int_{t_0}^{t} x(\tau)\mathrm{d}\tau + y(t_0)$$

Komparator



$$y(t) = \begin{cases} +1, & x(t_k) \geq s, \\ -1, & x(t_k) < s, \end{cases} \quad \text{für } t \in [t_k, t_{k+1})$$

Multiplizierer



$$y(t) = K \cdot x(t)$$

Dies sind alle Elemente, aus denen Σ-Δ-Konverter aufgebaut sind. Ihr grundlegender Aufbau und ihre Funktionsweise im Zeitbereich ist Inhalt dieses Kapitels. Die Modellbildung von Σ-Δ-Konvertern ist z. B. in [52] zu finden, welches als die Referenz auf dem Gebiet bezeichnet werden kann. Die Grundzüge der Darstellung in diesem Kapitel orientieren sich an [30], [60], [34], [31], [48] und [35].

2.1 Σ-Δ-Konverter 1. Ordnung

Wir betrachten nun den Σ-Δ-Konverter, wie er in Abbildung 1 dargestellt ist.

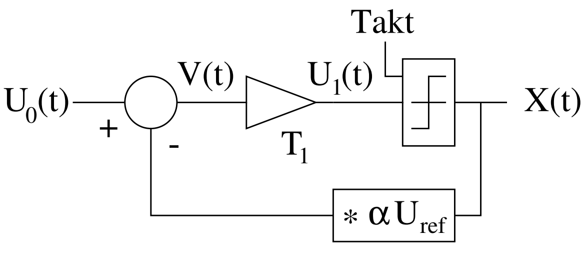


Abbildung 1: Σ-Δ-Konverter 1. Ordnung

Dabei bezeichnet $U_0(t)$ das analoge Eingangssignal und $X(t)$ das Ausgangssignal des Konverters. $X(t)$ ist eine Treppenfunktion, welche nur die Werte -1 und $+1$ annehmen kann. Wie in der Einleitung zu Kapitel 2 beschrieben, gelten dabei folgende Abhängigkeiten:

$$V(t) = U_0(t) - \alpha U_{ref} X(t), \tag{2.1}$$

$$U_1(t) = \frac{1}{T_1} \int_{t_0}^{t} V(\tau)\mathrm{d}\tau + U_1(t_0). \tag{2.2}$$

Wir bezeichnen mit

$$\sigma(r) := \begin{cases} 1, & \text{falls } r \geq 0, \\ 0, & \text{falls } r < 0 \end{cases} \tag{2.3}$$

den Einheitssprung (Heavisidefunktion). Der getaktete Komparator vergleicht nach festen Zeitabständen Δt den Wert der Funktion U_1 mit einem Vergleichswert $s = const.$ und schaltet den Binärausgang X nach folgender Abhängigkeit

$$X(t_k) := 2\sigma(U_1(t_k) - s) - 1,\ t_k = t_0 + k\Delta t,\ k = 0, 1, 2, \dots \tag{2.4}$$

Durch

$$X(t) := X(t_k),\ t \in [t_k, t_{k+1}),\ k = 0, 1, 2... \tag{2.5}$$

erhalten wir eine stückweise konstante Funktion. Diese Funktion entspricht einer Einbitwandlung des Eingangssignals U_0. Aufgrund von (2.5) ist eine Untersuchung der Werte $X(t_k)$ oder der Funktion X gleichbedeutend. Wir nehmen o. B. d. A. an, daß $U_1(t_0) = 0$ ist. Ansonsten ergeben sich alle folgenden Resultate auf die gleiche Weise, indem s durch $s - U_1(t_0)$ ersetzt wird.

2.1.1 Differenzengleichung

Wir wollen nun die Differenzengleichung herleiten, welche das Verhalten des Konverters 1. Ordnung im Zeitbereich widerspiegelt. Aus (2.4), (2.1) und (2.2) ergibt sich für $t \in [t_k, t_{k+1}), k \in \mathbb{N}_0$ dann

$$\begin{aligned} X(t) &= 2\sigma(U_1(t_k) - s) - 1 \\ &= 2\sigma\left(\frac{1}{T_1}\int_{t_0}^{t_k} V(\tau)\mathrm{d}\tau - s\right) - 1 \\ &= 2\sigma\left(\frac{1}{T_1}\int_{t_0}^{t_k} [U_0(\tau) - \alpha U_{ref} X(\tau)]\mathrm{d}\tau - s\right) - 1 \\ &= 2\sigma\left(\frac{1}{T_1}\int_{t_0}^{t_k} U_0(\tau)\mathrm{d}\tau - \frac{\alpha U_{ref}}{T_1}\int_{t_0}^{t_k} X(\tau)\mathrm{d}\tau \quad - s\right) - 1. \end{aligned} \tag{2.6}$$

Insbesondere gilt für $t = t_k$ und $\Delta t = t_{k+1} - t_k$ dann

$$X(t_k) = 2\sigma\left(\frac{1}{T_1}\int_{t_0}^{t_k} U_0(\tau)\mathrm{d}\tau - \frac{\alpha U_{ref}}{T_1}\sum_{j=0}^{k-1} X(t_j)\Delta t \quad - s\right) - 1. \tag{2.7}$$

Mit den Definitionen

$$\begin{aligned} X_k &:= X(t_k), \quad k \in \mathbb{N}_0 \\ Y_{k+1} - Y_k &:= X_k, \quad k \in \mathbb{N}_0 \\ Y_0 &:= 0 \end{aligned}$$

und (2.7) läßt sich das Verhalten des Konverters durch eine nichtlineare, nichtautonome Differenzengleichung erster Ordnung mit Anfangswert darstellen

$$\begin{aligned} Y_{k+1} &= Y_k + 2\sigma\left(\frac{1}{T_1}\int_{t_0}^{t_k} U_0(\tau)\mathrm{d}\tau - \frac{\alpha U_{ref}\Delta t}{T_1} Y_k - s\right) - 1, \quad k \in \mathbb{N}_0 \\ Y_0 &= 0. \end{aligned} \tag{2.8}$$

2.1.2 Lösung der Differenzengleichung

Wir betrachten die Differenzengleichung aus (2.8) und formen diese wie folgt um

$$\begin{aligned} Y_{k+1} &= Y_k + 2\sigma\left(\frac{1}{T_1}\int_{t_0}^{t_k} U_0(\tau)\mathrm{d}\tau - \frac{\alpha U_{ref}\Delta t}{T_1}Y_k - s\right) - 1 \\ &= Y_k + 2\sigma\left(\frac{\alpha U_{ref}\Delta t}{T_1}\left(\frac{1}{\alpha U_{ref}\Delta t}\int_{t_0}^{t_k} U_0(\tau)\mathrm{d}\tau - Y_k - \frac{T_1 s}{\alpha U_{ref}\Delta t}\right)\right) - 1 \end{aligned} \tag{2.9}$$

für $k \in \mathbb{N}_0$. Gilt für den Vergleichswert $s = 0$, so ist $Y_1 = 1$ unabhängig von U_0. Indem wir die positive Homogenität des Einheitssprungs

$\sigma(at) = \sigma(t)$ für $a > 0$

ausnutzen, erhalten wir für $k \in \mathbb{N}_0$ die Differenzengleichung

$$Y_{k+1} = Y_k + 2\sigma\left(\frac{1}{\alpha U_{ref}\Delta t}\int_{t_0}^{t_k} U_0(\tau)\mathrm{d}\tau - Y_k - \frac{T_1 s}{\alpha U_{ref}\Delta t}\right) - 1. \tag{2.10}$$

Eine Lösung von (2.10) zu gewinnen, basiert auf folgender Idee. Es sind Bedingungen zu finden, unter denen das Argument der Heavisidefunktion auf der rechten Seite von (2.10) im Intervall $[-1, 1)$ beschränkt bleibt. Unter diesen Bedingungen ist es nun möglich, $\sigma(.)$ durch $\lfloor . \rfloor$ zu ersetzen. Dabei ist $\lfloor x \rfloor$ der größte ganze Teil von x. Da $Y_k \in \mathbb{N}_0$ ist, kann Y_k aus $\lfloor . \rfloor$ herausgelöst werden. Die Differenzengleichung ist nun linear.

Bemerkung 2.1.1 Aus beweistechnischen Gründen werden wir das ursprüngliche Anfangswertproblem (2.8) bzw. nur die Differenzengleichung (2.10) einer Transformation unterziehen. Wir untersuchen das nun gewonnene neue Problem. Bewiesene Resultate lassen sich dann durch die Rücktransformation leicht auf die ursprünglichen Fragestellungen übertragen. Die Transformation lautet

$$W_k := \frac{Y_k + k}{2}, \quad k \in \mathbb{N}_0. \tag{2.11}$$

Wegen (2.11) und der positiven Homogenität des Einheitssprungs σ ist dann das Anfangswertproblem (2.8) bzw. die Differenzengleichung (2.10) äquivalent zu

$$\begin{aligned} W_{k+1} &= W_k + \sigma\left(\frac{1}{2\alpha U_{ref}\Delta t}\int_{t_0}^{t_k} U_0(\tau)\mathrm{d}\tau - W_k + \frac{k}{2} - \frac{T_1 s}{2\alpha U_{ref}\Delta t}\right), \quad k \in \mathbb{N}_0 \\ W_0 &= 0. \end{aligned} \tag{2.12}$$

Setzt man die Funktion

$$\varphi(k) := \begin{cases} -\dfrac{T_1 s}{2\alpha U_{ref}\Delta t}, & \text{für } k = 0, \\ \dfrac{1}{2\alpha U_{ref}\Delta t}\displaystyle\int_{t_0}^{t_k} U_0(\tau)\mathrm{d}\tau + \frac{k}{2} - \frac{T_1 s}{2\alpha U_{ref}\Delta t}, & \text{für } k \in \mathbb{N} \end{cases} \tag{2.13}$$

in Formel (2.12) ein, so ergibt sich die einfachere Darstellung

$$W_{k+1} = W_k + \sigma(\varphi(k) - W_k).$$

Untersuchen wir nun, welche Eigenschaften die Funktion φ hat. Entscheidend dabei ist die Forderung

$$|U_0(t)| \leq \alpha U_{ref} \quad \forall t \geq t_0. \tag{2.14}$$

Ist (2.14) nicht erfüllt, so kann U_1 nach (2.2) unbeschränkt werden. Eine Konsequenz ist, daß das Argument der Heavisidefunktion in (2.10) nicht beschränkt ist. Eine andere ist die physische Zerstörung des Konverters, wenn dieses als physikalisches Modell betrachtet wird.

Lemma 2.1.2 *Ist die Bedingung* (2.14) *erfüllt, dann gilt*

a) $0 \leq \dfrac{1}{2k\alpha U_{ref}\Delta t} \displaystyle\int_{t_0}^{t_k} U_0(\tau)\mathrm{d}\tau + \frac{1}{2} \leq 1,$

b) $0 \leq \varphi(k) - \varphi(k-1) \leq 1.$

Beweis: a) Aus (2.14) folgt unmittelbar

$$\begin{aligned} 0 &= \frac{1}{2k\alpha U_{ref}\Delta t} \int_{t_0}^{t_k} (-\alpha U_{ref})\mathrm{d}\tau + \frac{1}{2} \\ &\leq \frac{1}{2k\alpha U_{ref}\Delta t} \int_{t_0}^{t_k} U_0(\tau)\mathrm{d}\tau + \frac{1}{2} \\ &\leq \frac{1}{2k\alpha U_{ref}\Delta t} \int_{t_0}^{t_k} \alpha U_{ref}\mathrm{d}\tau + \frac{1}{2} = 1. \end{aligned}$$

b) Es gilt

$$\begin{aligned} \varphi(k) - \varphi(k-1) &= \frac{1}{2\alpha U_{ref}\Delta t} \int_{t_0}^{t_k} U_0(\tau)\mathrm{d}\tau + \frac{k}{2} - \frac{1}{2\alpha U_{ref}\Delta t} \int_{t_0}^{t_{k-1}} U_0(\tau)\mathrm{d}\tau - \frac{k-1}{2} \\ &= \frac{1}{2\alpha U_{ref}\Delta t} \int_{t_{k-1}}^{t_k} U_0(\tau)\mathrm{d}\tau + \frac{1}{2} \in [0,1]. \end{aligned}$$

□

Kommen wir nun zur zentralen Aussage des Kapitels 2.1, die die eindeutige Lösung des Anfangswertproblems (2.12) darstellt.

Satz 2.1.3 *Die Funktion* $\varphi : \mathbb{N}_0 \longrightarrow \mathbb{R}$ *erfülle die Bedingungen*

$$\varphi(0) \in [-1, 1), \tag{2.15}$$

$$\varphi(k) - \varphi(k-1) \in [0, 1], \quad \forall k \in \mathbb{N}. \tag{2.16}$$

Dann besitzt das Anfangswertproblem

$$\begin{aligned} W_{k+1} &= W_k + \sigma(\varphi(k) - W_k), \\ W_0 &= 0 \end{aligned} \tag{2.17}$$

genau eine Lösung der Form

$$W_k = \begin{cases} 0, & \text{für } k = 0 \\ \lfloor \varphi(k-1) \rfloor + 1, & \text{für } k \in \mathbb{N}, \end{cases} \tag{2.18}$$

wobei $\lfloor x \rfloor$ *den ganzen Teil von* x *bezeichnet.*

Beweis: In den Formeln (2.17) und (2.18) ist jeweils $W_0 = 0$.
Nun zeigen wir, daß in beiden Fällen W_1 übereinstimmt. Es sei zunächst $\varphi(0) \in [0,1)$. Dann folgt aus der Differenzengleichung (2.17)

$$W_1 = W_0 + \sigma(\varphi(0) - W_0) = \sigma(\varphi(0)) = 1.$$

Dies stimmt mit (2.18) überein, da dort gilt

$$W_1 = \lfloor \varphi(0) \rfloor + 1 = 1.$$

Es sei nun $\varphi(0) \in [-1,0)$. Dann folgt aus der Differenzengleichung (2.17)

$$W_1 = W_0 + \sigma(\varphi(0) - W_0) = \sigma(\varphi(0)) = 0.$$

Dies stimmt mit (2.18) überein, da dort gilt

$$W_1 = \lfloor \varphi(0) \rfloor + 1 = -1 + 1 = 0.$$

Nun wollen wir den Induktionsschritt beweisen. Hierzu sei $k \in \mathbb{N}$, und für $n \in \{1, \ldots, k\}$ stimmen die Lösungen W_n von (2.17) und (2.18) überein. Aus $0 \leq \varphi(k) - \varphi(k-1) \leq 1$ für $k \in \mathbb{N}$ ergeben sich dann die beiden folgenden Fälle.

1. $\lfloor \varphi(k) \rfloor \leq \varphi(k-1) \leq \varphi(k)$, d. h. $\lfloor \varphi(k-1) \rfloor = \lfloor \varphi(k) \rfloor$.
   Somit erhält man einerseits
   $$W_{k+1} - W_k = \lfloor \varphi(k) \rfloor - \lfloor \varphi(k-1) \rfloor = 0$$
   und andererseits
   $$\sigma(\varphi(k) - W_k) = \sigma(\varphi(k) - \lfloor \varphi(k-1) \rfloor - 1) = \sigma(\varphi(k) - \lfloor \varphi(k) \rfloor - 1) = 0.$$

2. $\varphi(k-1) < \lfloor \varphi(k) \rfloor \leq \varphi(k)$, d. h. $\lfloor \varphi(k-1) \rfloor + 1 = \lfloor \varphi(k) \rfloor$.
   Analog wie oben ist dann
   $$W_{k+1} - W_k = \lfloor \varphi(k-1) \rfloor + 1 - \lfloor \varphi(k-1) \rfloor = 1,$$
   und
   $$\sigma(\varphi(k) - W_k) = \sigma(\varphi(k) - \lfloor \varphi(k-1) \rfloor - 1) = \sigma(\varphi(k) - \lfloor \varphi(k) \rfloor) = 1.$$

Da in beiden Fällen die Differenzen $W_{k+1} - W_k$ bez. (2.17) und (2.18) übereinstimmen, folgt aus der Induktionsvoraussetzung die Behauptung. Damit ist alles bewiesen. □

Folgerung 2.1.4 Es seien die Voraussetzungen aus Satz 2.1.3 erfüllt, und die Funktion φ sei wie in (2.13) erklärt. Dann gilt

$$W_{k+1} - W_k = \begin{cases} \lfloor \varphi(0) \rfloor + 1, & \text{für } k = 0 \\ \lfloor \varphi(k) \rfloor - \lfloor \varphi(k-1) \rfloor, & \text{für } k \in \mathbb{N} \end{cases} \in \{0,1\}.$$

Bemerkung 2.1.5 Die Funktion $\varphi : \mathbb{N}_0 \longrightarrow \mathbb{R}$ aus Satz 2.1.3 muß nicht notwendig von der Form (2.13) sein. Die Aussage des Satzes ist für beliebiges φ, welches (2.15) und (2.16) erfüllt, gültig. Mit φ wie in (2.13) ergibt sich die Voraussetzung (2.16) des Satzes aus Lemma 2.1.2. Die Bedingung (2.15) ist schließlich äquivalent zur Forderung

$$-\frac{2\alpha\Delta t U_{ref}}{T_1} < s \leq \frac{2\alpha\Delta t U_{ref}}{T_1}. \tag{2.19}$$

Übertragen wir nun diese Ergebnisse auf das Originalmodell, indem wir die Rücktransformation

$$Y_k = 2W_k - k \tag{2.20}$$

benutzen, erhalten wir folgende Aussage.

Satz 2.1.6 *Es sei* $|U_0(t)| \leq \alpha U_{ref}$ *für* $t \geq t_0$ *erfüllt. Dann besitzt das Anfangswertproblem*

$$\begin{aligned} Y_{k+1} &= Y_k + 2\sigma\left(\frac{1}{\alpha U_{ref}\Delta t}\int\limits_{t_0}^{t_k} U_0(\tau)\mathrm{d}\tau - Y_k - \frac{T_1 s}{\alpha U_{ref}\Delta t}\right) - 1, \quad k \in \mathbb{N}_0, \\ Y_0 &= 0 \end{aligned}$$

die eindeutige Lösung

$$Y_k = \begin{cases} 0, & \text{für } k = 0, \\ 2\left(\left\lfloor \dfrac{1}{2\alpha U_{ref}\Delta t}\displaystyle\int\limits_{t_0}^{t_{k-1}} U_0(\tau)\mathrm{d}\tau + \dfrac{k-1}{2} - \dfrac{T_1 s}{2\alpha U_{ref}\Delta t}\right\rfloor + 1\right) - k, & \text{für } k \in \mathbb{N}. \end{cases} \tag{2.21}$$

Bemerkung 2.1.7 In allen folgenden Betrachtungen legen wir entsprechend der Zweckmäßigkeit eines der beiden Modelle - das $\{0,1\}$- oder $\{-1,1\}$-Modell - zugrunde, da wir beide wegen (2.11) identifizieren können.

2.1.3 Gültigkeit der Lösung bei Veränderung der Annahmen

Eine zentrale Rolle, um Satz 2.1.6 anwenden zu können, spielt die Voraussetzung $|U_0(t)| \leq \alpha U_{ref}$ für alle $t \geq t_0$. Für die Benutzung der exakten Lösung (2.21) bei der Untersuchung von Konvertern höherer Ordnung ist es notwendig, daß das Eingangssignal der innersten Schleife dieser Beschränkung genügt.
Im weiteren werden wir die transformierte Differenzengleichung (2.12) untersuchen. Dabei diskutieren wir die Beziehungen zwischen den Formeln (2.17) und (2.18), falls die Voraussetzungen (2.15) und (2.16) verletzt sind. In den weiteren Überlegungen ist φ wieder von der speziellen Gestalt (2.13). Die erzielten Ergebnisse lassen sich wiederum leicht durch die Rücktransformation (2.20) auf das Originalproblem übertragen. Ist $|U_0(t)| > \alpha U_{ref}$ für $t > t_1 \geq t_0$, dann sind zwei Fälle zu unterscheiden:

1. Nach einer gewissen Zeit überschreitet die Kapazität des Kondensators, der den Integrator darstellt, den Bereich, in dem er sich linear verhält. In diesem Fall versagt unser mathematisches Modell. Hier sind nur noch Simulationen denkbar, in denen das nichtlineare Verhalten einfließt.

2. Ist dagegen der Bereich des linearen Verhaltens des Integrators hinreichend groß, so „schneidet“ der Konverter alle Werte von $U_0(t)$ oberhalb von αU_{ref} bzw. unterhalb von $-\alpha U_{ref}$ ab. D. h., es ist $X_k = 1$ bzw. $X_k = -1$ für $t > t_1$. Dieses mögliche Verhalten ist durch unser Modell beschreibbar.

Wir nehmen nun an, daß (2.1) und (2.2) für $U_0(t) \in \mathbb{R}$ erfüllt ist. Damit besitzt der Kondensator eine unendliche Kapazität und ein lineares Verhalten über ganz $\mathbb{R}$. Zunächst untersuchen wir das Anfangswertproblem (2.17) für den Fall, daß der Anfangswert W_0 von Null verschieden ist. In den folgenden Betrachtungen seien die Größen α, U_{ref} sowie das Eingangssignal $U_0(t)$ fest vorgegeben. Wir definieren nun die Funktionen

$$\Phi(m,k) := \frac{1}{2\alpha U_{ref}\Delta t}\int_{t_m}^{t_k} U_0(\tau)\mathrm{d}\tau + \frac{(k-m)}{2},\ m \in \{0,1,2,...,k\},\ k \in \mathbb{N}_0.$$

Nun eine vorbereitende Aussage, die wir für die Beantwortung der oben aufgeführten Fragestellung benötigen.

Lemma 2.1.8 *Für eine beliebige Folge natürlicher Zahlen $m_j,\ j = 1,2,...,r$ mit $0 \le m_1 \le ... \le m_r \le k$ gilt*

$$\varphi(k) = \varphi(m_1) + \Phi(m_1,m_2) + \cdots + \Phi(m_{r-1},m_r) + \Phi(m_r,k).$$

Beweis: Der Beweis ist elementar. □

Folgerung 2.1.9 Es sei $0 \le \varphi(k) - \varphi(k-1) \le 1$, für $k \in \mathbb{N}$ und $\varphi(0) \in \mathbb{R}$ beliebig. Dann ist

$$W_k = \begin{cases} \lfloor\varphi(0)\rfloor, & \text{für } k = 0 \\ \lfloor\varphi(k-1)\rfloor + 1, & \text{für } k \in \mathbb{N}. \end{cases} \tag{2.22}$$

die eindeutige Lösung des Anfangswertproblems

$$\begin{aligned} W_{k+1} &= W_k + \sigma(\varphi(k) - W_k) \\ W_0 &= \lfloor\varphi(0)\rfloor. \end{aligned} \tag{2.23}$$

Beweis: Unter Nutzung des Lemmas 2.1.8 schreiben wir (2.23) in der Form

$$W_{k+1} = W_k + \sigma(\varphi(0) + \Phi(0,k) - W_k).$$

Hieraus erhalten wir nun

$$W_{k+1} - \lfloor\varphi(0)\rfloor = W_k - \lfloor\varphi(0)\rfloor + \sigma(\varphi(0) - \lfloor\varphi(0)\rfloor + \Phi(0,k) - (W_k - \lfloor\varphi(0)\rfloor)).$$

Mit Hilfe der Transformationen

$$\tilde{Y}_k := W_k - \lfloor\varphi(0)\rfloor \quad \text{und} \quad \tilde{\varphi}(k) := \varphi(0) - \lfloor\varphi(0)\rfloor + \Phi(0,k) = \varphi(k) - \lfloor\varphi(0)\rfloor$$

ergibt sich die Differenzengleichung (2.17) bez. $\tilde{Y}_k$ und $\tilde{\varphi}$. Diese erfüllt die Bedingungen (2.15), (2.16) und (2.17) des Satzes 2.1.3 und besitzt die Lösung (2.18). Die Rücktransformation liefert schließlich (2.22). □

Da die Differenzen $W_{k+1} - W_k$ bzw. $W_{k+1} - W_0$ für $k \in \mathbb{N}$ nicht durch $\lfloor\varphi(0)\rfloor$ beeinflußt werden, können wir die Rekursion mit $Y_0 = \lfloor\varphi(0)\rfloor$ beginnen. Das bedeutet, daß der Kondensator zum Zeitpunkt t_0 bereits etwas aufgeladen ist. In praxi ist dieser Wert unbekannt. Aus diesem Grund sind wir an einer Lösung von (2.17) interessiert, falls (2.15) nicht erfüllt ist.

Folgerung 2.1.10 Die Bedingung (2.15) des Satzes 2.1.3 sei verletzt. Weiterhin existiere ein $\varepsilon > 0$ so, daß $|U_0(t)| \leq \alpha U_{ref} - \varepsilon$ für alle $t \in [t_0, t_1]$, t_1 genügend groß und $0 < \varepsilon < \alpha U_{ref}$ fest. Dann gibt es einen Index $k_0 \in \mathbb{N}$, so daß die Lösung W_k, $k \in \mathbb{N}$ des Anfangswertproblems (2.17) der Formel (2.18) für alle $k \geq k_0$ genügt. Darüber hinaus lautet die eindeutige Lösung im Fall $\varphi(0) \geq 1$

$$W_k = \begin{cases} k, & \text{für} \quad k = 0, 1, ..., k_0 - 1, \\ k_0 = \lfloor \varphi(k_0) \rfloor, & \text{für} \quad k = k_0 \\ \lfloor \varphi(k-1) \rfloor + 1, & \text{für} \quad k = k_0 + 1, k_0 + 2, ... \end{cases}$$

und im Fall $\varphi(0) < -1$

$$W_k = \begin{cases} 0, & \text{für} \quad k = 0, 1, ..., k_0 - 1, \\ -1 = \lfloor \varphi(k_0) \rfloor, & \text{für} \quad k = k_0 \\ \lfloor \varphi(k-1) \rfloor + 1, & \text{für} \quad k = k_0 + 1, k_0 + 2, ... \end{cases}$$

wobei $k_0 + 1$ der erste Index k ist, in welchem $\sigma(\varphi(k) - W_k)$ einen Sprung hat.

Bemerkung 2.1.11 Falls die Bedingung $\varphi(0) \in [-1, 1)$ nicht erfüllt ist, so müssen wir sicherstellen, daß die Folge $(W_k - \varphi(k))_{k \in \mathbb{N}_0}$ die Möglichkeit hat, das Intervall $[-1, 1)$ für endliches $k \geq 0$ zu erreichen. Dies geschieht, falls $U_0(t) \in [-\alpha U_{ref} + \varepsilon, \alpha U_{ref} - \varepsilon]$ hinreichend lange gilt.

Beweis: Fall 1: $\varphi(0) \geq 1$.
Wegen $W_0 = 0$ ist $\sigma(\varphi(0) - W_0) = 1$ and $W_1 = 1$. So lange $\varphi(k) - W_k \geq 0$ ist, gilt $\sigma(\varphi(k) - W_k) = 1$ für $k = 0, 1, 2, ..., k_0$. Dabei bezeichne k_0 den kleinsten Index mit

$$\varphi(k_0) - W_{k_0} \geq 0 \tag{2.24}$$

und

$$\varphi(k_0 + 1) - W_{k_0+1} < 0. \tag{2.25}$$

Dann gilt sicherlich

$$W_k = k, \; k = 0, 1, ..., k_0 + 1. \tag{2.26}$$

Nun wollen wir zeigen, daß ein solcher Index k_0 existiert. Nach obiger Konstruktion haben wir zu prüfen, ob die Ungleichung $\varphi(k_0 + 1) < k_0 + 1$ für einen Index $k_0 < \infty$ gilt. Dabei ist

$$\begin{aligned} \varphi(k_0 + 1) &= \varphi(0) + \Phi(0, k_0 + 1) = \varphi(0) + \frac{1}{2\alpha U_{ref} \Delta t} \int_{t_0}^{t_0 + (k_0+1)\Delta t} U_0(\tau) \mathrm{d}\tau + \frac{k_0 + 1}{2} \\ &\leq \varphi(0) + k_0 + 1 - \varepsilon \frac{(k_0 + 1)}{2\alpha U_{ref}} < k_0 + 1 \end{aligned}$$

wenigstens für $k_0 > \frac{2\alpha U_{ref} \varphi(0)}{\varepsilon} - 1$ erfüllt. Aus (2.24) folgt dann $\lfloor \varphi(k_0) \rfloor \geq W_{k_0}$. Nehmen wir nun an, es gelte $\lfloor \varphi(k_0) \rfloor \neq W_{k_0}$. Dies führt zu $\varphi(k_0) - W_{k_0} \geq 1$, und mit (2.25) erhalten wir den Widerspruch $k_0 + 1 > \varphi(k_0) \geq k_0 + 1$. Also ist $\lfloor \varphi(k_0) \rfloor = W_{k_0}$. Mit Hilfe der Transformationen für $k, m \in \mathbb{N}_0$

$$\begin{aligned} \tilde{W}_k &:= W_{k+k_0}, \\ \tilde{\varphi}(k) &:= \varphi(k_0 + k), \\ \tilde{\Phi}(m, k) &:= \Phi(k_0 + m, k_0 + k), \end{aligned}$$

erhalten wir

$$\tilde{\varphi}(k) = \varphi(k_0 + k) = \varphi(k_0 + m) + \Phi(k_0 + m, k_0 + k) = \tilde{\varphi}(m) + \tilde{\Phi}(m, k).$$

Nach Folgerung 2.1.9 besitzt das Anfangswertproblem

$$\begin{aligned} \tilde{W}_{k+1} &= \tilde{W}_k + \sigma\left(\tilde{\varphi}(k) - \tilde{W}_k\right), k \in \mathbb{N}_0 \\ \tilde{W}_0 &= \lfloor \tilde{\varphi}(0) \rfloor . \end{aligned}$$

die eindeutige Lösung

$$\tilde{W}_k = \begin{cases} \lfloor \tilde{\varphi}(0) \rfloor, & \text{für} \quad k = 0, \\ \lfloor \tilde{\varphi}(k-1) \rfloor + 1, & \text{für} \quad k \in \mathbb{N}. \end{cases} \tag{2.27}$$

Nach der Rücktransformation erhalten wir unter Benutzung von (2.26) dann die Lösung von (2.17)

$$W_k = \begin{cases} k, & \text{für} \quad k = 0, 1, ..., k_0 \\ \lfloor \varphi(k-1) \rfloor + 1, & \text{für} \quad k = k_0 + 1, k_0 + 2, ... \end{cases} \tag{2.28}$$

Fall 2: $\varphi(0) < -1$.
Der Beweis ist analog wie im ersten Fall zu führen. Wir erhalten $W_k = 0$ für $k = 0, 1, 2, ..., k_0 + 1$, wobei k_0 der erste Index ist, für den $\varphi(k_0) - W_{k_0} < 0$ und $\varphi(k_0 + 1) - W_{k_0+1} \geq 0$ gilt. Der Schluß auf $\lfloor \varphi(k_0) \rfloor = -1$ und auf die Lösung

$$W_k = \begin{cases} 0, & \text{für} \quad k = 0, 1, ..., k_0, \\ \lfloor \varphi(k-1) \rfloor + 1, & \text{für} \quad k = k_0 + 1, k_0 + 2, ... \end{cases}$$

des Anfangswertproblems (2.17) ist gleich. □

Nun wollen wir das Lösungsverhalten der Differenzengleichung (2.17) untersuchen, wenn die Bedingung $|U_0(t)| \leq \alpha U_{ref}$ in einem Intervall (t_{k_0}, t_{k_1}) verletzt ist. In Anlehnung an die Bemerkung 2.1.11 nehmen wir an, daß W_k durch (2.18) für $k = k_{-1}, k_{-1} + 1, ..., k_0$ beschrieben werden kann.

Folgerung 2.1.12 Für die Indizes $k_{-1} < k_0 < k_1$ gelte

1. $U_0(t) > \alpha U_{ref}$ im Intervall (t_{k_0}, t_{k_1}),
2. $W_k = \lfloor \varphi(k-1) \rfloor + 1$ für $k = k_{-1},\ k_{-1} + 1, ..., k_0$,
3. $-\alpha U_{ref} \leq U_0(t) \leq \alpha U_{ref} - \varepsilon$ für $t_{k_1} \leq t \leq t_{k_3},\ t_{k_3}$ hinreichend groß, $2\alpha U_{ref} > \varepsilon > 0$, ε fest,
4. $\varphi(k_0) - W_{k_0} \geq 0$.

Dann gibt es einen Index k_2 mit $k_3 \geq k_2 \geq k_1$ so, daß W_k durch die Formel (2.18) für $k = k_2, k_2 + 1, ...$ beschrieben werden kann. Insbesondere lautet die Lösung von (2.17) dann

$$W_k = \begin{cases} \lfloor \varphi(k-1) \rfloor + 1, & \text{für } k = k_{-1}, k_{-1} + 1, ..., k_0, \\ \lfloor \varphi(k_0 - 1) \rfloor + 1 + k - k_0, & \text{für } k = k_0 + 1, ..., k_2, \\ \lfloor \varphi(k-1) \rfloor + 1, & \text{für } k = k_2 + 1, ... \end{cases}$$

Dabei ist k_2 der erste Index $k > k_1$, welcher $\varphi(k+1) - W_{k+1} < 0$ erfüllt.
Weiterhin gilt $\lfloor \varphi(k_0) \rfloor = \lfloor \varphi(k_0 - 1) \rfloor + 1,\ W_{k_2} = \lfloor \varphi(k_2) \rfloor = \lfloor \varphi(k_0 - 1) \rfloor + 1 + k_2 - k_0$.

Beweis: Aus der Voraussetzung 1 folgt für $k_0 + k \in [k_0, k_1]$ die Ungleichung

$$\begin{aligned}\varphi(k_0+k) &= \varphi(k_0) + \Phi(k_0, k_0+k) = \varphi(k_0) + \frac{1}{2\alpha U_{ref}\Delta t}\int\limits_{t_{k_0}}^{t_{k_0}+k\Delta t} U_0(\tau)\mathrm{d}\tau + \frac{k}{2} \\ &\geq \varphi(k_0) + k.\end{aligned}$$

Aus der Gültigkeit der Ungleichung $W_{k_0+k} \leq W_{k_0} + k$ für jede Lösung von (2.17) ergibt sich

$$\varphi(k_0+k) \geq \varphi(k_0) + k \geq \varphi(k_0) + W_{k_0+k} - W_{k_0}.$$

Somit gilt

$$\varphi(k_0+k) - W_{k_0+k} \geq \varphi(k_0) - W_{k_0} \geq 0,$$

was für alle Indizes k mit $k_0 + k \leq k_1$ dann

$$\sigma(\varphi(k_0+k) - W_{k_0+k}) = 1$$

impliziert. Nutzt man

$$\begin{aligned}&0 \leq \varphi(k_0) - \varphi(k_0-1) < 1, \\ &W_{k_0} = \lfloor\varphi(k_0-1)\rfloor + 1, \\ &\lfloor\varphi(k_0)\rfloor \geq W_{k_0}\end{aligned}$$

aus, so ergibt sich

$$\lfloor\varphi(k_0)\rfloor = \lfloor\varphi(k_0-1)\rfloor + 1.$$

Sei nun k_2 der erste Index größer als k_1 mit den Eigenschaften

$$\varphi(k_2) - W_{k_2} \geq 0 \quad \text{und} \quad \varphi(k_2+1) - W_{k_2+1} < 0.$$

Die Existenz eines solchen endlichen Indexes folgt aus der obigen Annahme, daß $U_0(t) \leq \alpha U_{ref} - \varepsilon$ in (t_{k_1}, t_{k_3}) ist, wobei t_{k_3} genügend groß sei. Da andererseits k_1 so festgelegt ist wie in Folgerung 2.1.10 im Fall $\varphi(0) \geq 1$ (der Index k_0 spielte dort die gleiche Rolle wie hier der Index k_2), erhalten wir mit ähnlichen Überlegungen und unter Nutzung von Folgerung 2.1.9 die Behauptungen dieser Bemerkung. □

Im Vergleich zur Folgerung 2.1.12 sei die vierte Voraussetzung verändert.

Folgerung 2.1.13 Für die Indizes $k_{-1} < k_0 < k_1$ gelte

1. $U_0(t) > \alpha U_{ref}$ im Intervall (t_{k_0}, t_{k_1}), $k_1 \geq k_0 + 2$,
2. $W_k = \lfloor\varphi(k-1)\rfloor + 1$ für $k = k_{-1}, k_{-1}+1, ..., k_0$,
3. $-\alpha U_{ref} \leq U_0(t) \leq \alpha U_{ref} - \varepsilon$ für $t_{k_1} \leq t \leq t_{k_3}$, t_{k_3} hinreichend groß, $2\alpha U_{ref} > \varepsilon > 0$, ε fest,
4. $\varphi(k_0) - W_{k_0} < 0$.

Dann gibt es einen Index k_2 mit $k_3 \geq k_2 \geq k_1$ so, daß W_k durch die Formel (2.18) für $k = k_2, k_2+1, ...$ beschrieben werden kann. Insbesondere lautet die Lösung von (2.17) dann

$$W_k = \begin{cases} \lfloor \varphi(k-1) \rfloor + 1, & \text{für } k = k_{-1}, k_{-1}+1, ..., k_0+1, \\ \lfloor \varphi(k_0) \rfloor + k - k_0, & \text{für } k = k_0+2, ..., k_2, \\ \lfloor \varphi(k-1) \rfloor + 1 & \text{für } k = k_2+1, ... \end{cases}$$

Dabei ist k_2 der erste Index $k > k_1$, welcher $\varphi(k+1) - W_{k+1} < 0$ erfüllt.
Weiterhin gilt $W_{k_2} = \lfloor \varphi(k_2) \rfloor = \lfloor \varphi(k_0) \rfloor + k_2 - k_0$.

Beweis: Die Eigenschaft $W_{k+1} = W_k + \sigma(\varphi(k) - W_k)$ liefert zusammen mit der Bedingung $\varphi(k_0) - W_{k_0} < 0$ dann $W_{k_0+1} = W_{k_0}$. Weiterhin sind die Formeln

$$W_{k_0} = \lfloor \varphi(k_0-1) \rfloor + 1 \quad \text{und} \quad 0 \leq \varphi(k_0) - \varphi(k_0-1) < 1$$

für $k \leq k_0$ gültig. Daraus folgt

$$\varphi(k_0-1) \leq \varphi(k_0) < \lfloor \varphi(k_0-1) \rfloor + 1 \quad \text{und} \quad \lfloor \varphi(k_0-1) \rfloor = \lfloor \varphi(k_0) \rfloor.$$

Das bedeutet aber $W_{k_0+1} = \lfloor \varphi(k_0) \rfloor + 1$. Somit ist die Formel (2.18) bis k_0+1 gültig. Danach haben wir es mit einer Situation zu tun, wie sie schon in Folgerung 2.1.12 untersucht wurde. □

Ähnliche Resultate ergeben sich im Fall $U_0(t) < -\alpha U_{ref}$ für $t \in (t_{k_0}, t_{k_1})$. Wir wollen aus diesem Grund nur die Ergebnisse formulieren. Die Beweise sind völlig analog.

Folgerung 2.1.14 Für die Indizes $k_{-1} < k_0 < k_1$ gelte

1. $U_0(t) < -\alpha U_{ref}$ im Intervall (t_{k_0}, t_{k_1}),
2. $W_k = \lfloor \varphi(k-1) \rfloor + 1$ für $k = k_{-1}, k_{-1}+1, ..., k_0$,
3. $\alpha U_{ref} \geq U_0(t) > -\alpha U_{ref} + \varepsilon$ für $t_{k_1} \leq t \leq t_{k_3}$, t_{k_3} hinreichend groß, $2\alpha U_{ref} > \varepsilon > 0$, ε fest,
4. $\varphi(k_0) - W_{k_0} < 0$.

Dann gibt es einen Index k_2 mit $k_3 \geq k_2 \geq k_1$ so, daß W_k durch die Formel (2.18) für $k = k_2, k_2+1, ...$ beschrieben werden kann. Insbesondere lautet die Lösung von (2.17) dann

$$W_k = \begin{cases} \lfloor \varphi(k-1) \rfloor + 1, & \text{für } k = k_{-1}, k_{-1}+1, ..., k_0, \\ \lfloor \varphi(k_0-1) \rfloor + 1, & \text{für } k = k_0+1, ..., k_2, \\ \lfloor \varphi(k-1) \rfloor + 1, & \text{für } k = k_2+1, ... \end{cases}$$

Dabei ist k_2 der erste Index $k > k_1$, welcher $\varphi(k+1) - W_{k+1} \geq 0$ erfüllt.
Weiterhin gilt $W_{k_2} = \lfloor \varphi(k_2) \rfloor = \lfloor \varphi(k_0-1) \rfloor + 1$.

Folgerung 2.1.15 Für die Indizes $k_{-1} < k_0 < k_1$ gelte

1. $U_0(t) < -\alpha U_{ref}$ im Intervall (t_{k_0}, t_{k_1}),
2. $W_k = \lfloor \varphi(k-1) \rfloor + 1$ für $k = k_{-1}, k_{-1}+1, ..., k_0$,
3. $\alpha U_{ref} \geq U_0(t) > -\alpha U_{ref} + \varepsilon$ für $t_{k_1} \leq t \leq t_{k_3}$, t_{k_3} hinreichend groß, $2\alpha U_{ref} > \varepsilon > 0$, ε fest,
4. $\varphi(k_0) - W_{k_0} \geq 0$.

Dann gibt es einen Index k_2 mit $k_3 \geq k_2 \geq k_1$ so, daß W_k durch die Formel (2.18) für $k = k_2, k_2+1, ...$ beschrieben werden kann. Insbesondere lautet die Lösung von (2.17) dann

$$W_k = \begin{cases} \lfloor \varphi(k-1) \rfloor + 1, & \text{für } k = k_{-1}, k_{-1}+1, ..., k_0, \\ \lfloor \varphi(k_0) \rfloor + 1, & \text{für } k = k_0+2, ..., k_2, \\ \lfloor \varphi(k-1) \rfloor + 1, & \text{für } k = k_2+1, ... \end{cases}$$

Dabei ist k_2 der erste Index $k > k_1$, welcher $\varphi(k+1) - W_{k+1} \geq 0$ erfüllt.
Weiterhin gilt $W_{k_2} = \lfloor \varphi(k_2) \rfloor = \lfloor \varphi(k_0-1) \rfloor + 1$.

Die Bedingung $|U_0(t)| \leq \alpha U_{ref}$ ist nur hinreichend für das Erfülltsein von (2.18). Allgemeiner dagegen ist die integrale Bedingung (2.16), damit (2.18) gilt. Bei dieser Bedingung braucht $|U_0(t)| \leq \alpha U_{ref}$ in einigen Takten Δt nicht erfüllt sein.

Folgerung 2.1.16 Für die Indizes $k_{-1} < k_0 < k_1$ gelte

1. $U(t) > \alpha U_{ref}$ im Intervall (t_{k_0}, t_{k_1}),
2. $\varphi(k_0) - \lfloor \varphi(k_0) \rfloor + \Phi(k_0, k) < k - k_0 + 1, k = k_0, k_0+1, ..., k_1$,
3. $W_k = \lfloor \varphi(k-1) \rfloor + 1, k = k_{-1}, k_{-1}+1, ..., k_0$,
4. $-\alpha U_{ref} \leq U(t) \leq \alpha U_{ref} - \varepsilon$ for $t_{k_1} \leq t \leq t_{k_3}, t_{k_3}$ hinreichend groß, $2\alpha U_{ref} > \varepsilon > 0$, ε fest.

Dann bleibt die Formel (2.18) gültig für alle Indizes $k \geq k_0$.

Beweis: Fall 1: $\varphi(k_0) - W_{k_0} \geq 0$.
Mit Berücksichtigung der Folgerung 2.1.12 verbleibt nur noch zu zeigen, daß

$$W_k = \lfloor \varphi(k-1) \rfloor + 1 = \lfloor \varphi(k_0-1) \rfloor + 1 + k - k_0 \quad \forall k = k_0+1, ..., k_2$$

erfüllt ist. Bekannt ist

$$W_{k_0} = \lfloor \varphi(k_0-1) \rfloor + 1 \quad \text{ und } \quad \varphi(k) - W_k \geq 0 \quad \text{für} \quad k = k_0, k_0+1, ..., k_1.$$

Weiterhin gilt

$$\begin{aligned} k - k_0 + 1 &> \varphi(k_0) - \lfloor \varphi(k_0) \rfloor + \Phi(k_0, k) \\ &= \varphi(k) - \lfloor \varphi(k_0) \rfloor \\ &\geq k - k_0 + \varphi(k_0) - \lfloor \varphi(k_0) \rfloor \quad \text{für} \quad k = k_0, k_0+1, ..., k_1. \end{aligned}$$

Hieraus ergibt sich dann

$$k - k_0 + 1 + \lfloor \varphi(k_0) \rfloor > \lfloor \varphi(k) \rfloor \geq k - k_0 + \lfloor \varphi(k_0) \rfloor \quad \text{für} \quad k = k_0, k_0+1, ..., k_1.$$

Damit ist dann

$$\lfloor \varphi(k) \rfloor = k - k_0 + \lfloor \varphi(k_0) \rfloor \quad \text{für} \quad k = k_0, k_0+1, ..., k_1.$$

Aus der Beziehung $\lfloor \varphi(k_0) \rfloor = \lfloor \varphi(k_0-1) \rfloor + 1$ folgt dann die Behauptung, da für $k \geq k_1$ die Folgerung 2.1.9 angewandt werden kann, wobei $W_{k_1} = \lfloor \varphi(k_1) \rfloor$ die Anfangsbedingung ist.
Fall 2: $\varphi(k_0) - W_{k_0} < 0$.
Unter Benutzung von Folgerung 2.1.13 folgt die Behauptung durch analoge Überlegungen wie im Fall 1. □

Folgerung 2.1.17 Es gelte $U_0(t) > \alpha U_{ref}$ in einem Intervall (t_{k_0}, t_{k_1}). Weiterhin sei $W_k = \lfloor \varphi(k-1) \rfloor + 1$. Dann ist leicht zu sehen, daß

$$\varphi(k_0) - \lfloor \varphi(k_0) \rfloor + \Phi(k_0, k) < k - k_0 + 1 \quad \text{für} \quad k = k_0, k_0 + 1, ..., k_1 \tag{2.29}$$

erfüllt ist. D. h., die Bedingung (2.29) ist im Fall $U_0(t) > \alpha U_{ref}$ in einem Intervall (t_{k_0}, t_{k_1}) hinreichend und notwendig für die Gültigkeit von (2.18).

Beweis: Aus der Beziehung

$$\lfloor \varphi(k-1) \rfloor + 1 = \lfloor \varphi(k_0 - 1) \rfloor + 1 + k - k_0$$

folgt sofort

$$\begin{aligned} \varphi(k) - \lfloor \varphi(k_0) \rfloor &= \varphi(k) - \lfloor \varphi(k_0 - 1) \rfloor - 1 \\ &= \lfloor \varphi(k_0 - 1) \rfloor + 1 + k - k_0 - \lfloor \varphi(k_0 - 1) \rfloor - 1 + \varphi(k) - \lfloor \varphi(k) \rfloor \\ &= k - k_0 + \varphi(k) - \lfloor \varphi(k) \rfloor < k - k_0 + 1. \end{aligned}$$

□

Folgerung 2.1.18 Analog zur Folgerung 2.1.17 gilt folgende Aussage. Im Fall $U_0(t) < -\alpha U_{ref}$ in einem Intervall (t_{k_0}, t_{k_1}) ist die Bedingung

$$0 \leq \varphi(k_0) - \lfloor \varphi(k_0) \rfloor + \Phi(k_0, k) \quad \text{für} \quad k = k_0, k_0 + 1, ..., k_1$$

hinreichend und notwendig für die Gültigkeit von (2.18).

Abschließend wollen wir unsere Überlegungen in einer Aussage zusammenfassen, welche noch einmal den Zusammenhang zwischen der Lösung des Anfangswertproblems (2.17) und der Formel (2.18) verdeutlicht.

Fazit 2.1.19 *Es sei die Bedingung* (2.15) *verletzt, oder es gelte* $|U_0(t)| > \alpha U_{ref}$. *Es bezeichne* W_k *bzw.* $\hat{W}_k$ *die Lösung des Anfangswertproblems* (2.17) *bzw. den Wert von* (2.18). *Stimmen beide zum Zeitpunkt* t_{k^*} *überein und ist zusätzlich* $|U_0(t)| \leq \alpha U_{ref}$ *für alle* $k \in [k^*, k^{**}]$, *dann stimmen beide Lösungen im Intervall* $[k^*, k^{**}]$ *überein, d. h., die Lösung des Anfangswertproblems* (2.17) *ist in* $[k^*, k^{**}]$ *explizit durch* (2.18) *gegeben.*

2.2 Σ-Δ-Konverter 2. Ordnung

Wir betrachten nun den Σ-Δ-Konverter, wie er in Abbildung 2 dargestellt ist.

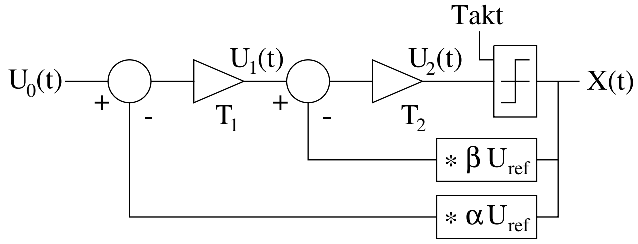


Abbildung 2: Σ-Δ-Konverter 2. Ordnung

Wir geben zunächst die analytische Beschreibung der einzelnen Elemente an. Der erste Integrator ordnet über eine einfache Integration mit der Skalierung $\frac{1}{T_1}$ dem Eingangssignal $U_0(t)$ gemäß

$$U_1(t) = \frac{1}{T_1}\int_{t_0}^{t}(U_0(\tau) - \alpha U_{ref}X(\tau))\mathrm{d}\tau + U_1(t_0) \tag{2.30}$$

die Funktion $U_1(t)$ und entsprechend der 2. Integrator mit der Skalierung $\frac{1}{T_2}$ der Funktion $U_1(t)$ gemäß

$$U_2(t) = \frac{1}{T_2}\int_{t_0}^{t}(U_1(\tau) - \beta U_{ref}X(\tau))\mathrm{d}\tau + U_2(t_0) \tag{2.31}$$

die Funktion $U_2(t)$ zu. Wir bezeichnen mit σ wie in (2.3) den Einheitssprung. Der getaktete Komparator vergleicht nach festen Zeitabständen Δt den Wert der Funktion U_2 mit einem Vergleichswert $s = const.$ und schaltet den Binärausgang X nach folgender Abhängigkeit

$$X(t_k) := 2\sigma(U_2(t_k) - s) - 1,\ t_k = t_0 + k\Delta t,\ k = 0, 1, 2, \ldots \tag{2.32}$$

Durch

$$X(t) := X(t_k),\ t \in [t_k, t_{k+1}),\ k = 0, 1, 2... \tag{2.33}$$

erhalten wir eine stückweise konstante Funktion. Diese Funktion entspricht einer Einbitwandlung des Eingangssignals U_0. Aufgrund von (2.33) ist eine Untersuchung der Werte $X(t_k)$ oder der Funktion X gleichbedeutend. Wir nehmen o. B. d. A. an, daß $U_2(t_0) = 0$ ist. Ansonsten ergeben sich alle folgenden Resultate auf die gleiche Weise, indem s durch $s - U_2(t_0)$ ersetzt wird.

2.2.1 Differenzengleichung

Für alle folgenden Rechnungen sei $k \in \mathbb{N}_0$ und $t \in [t_k, t_{k+1})$. Weiterhin sei $U_1(t_0) = 0$. Dies vereinfacht die Herleitung der entsprechenden Differenzengleichung beginnend bei (2.34). Bei der Ableitung eines Filtervorschlags in Kapitel 4.2.2 bzw. der Approximation einer Näherungslösung des allgemeineren DAWP in Kapitel 5 lassen wir diese Annahme fallen. Aus dem Modell wie in Abbildung 2 dargestellt und den Formeln (2.30) bis (2.33) folgt dann

$$\begin{aligned}
X(t) &= 2\sigma\left(U_2(t_k) - s\right) - 1 \\
&= 2\sigma\left(\frac{1}{T_2}\int_{t_0}^{t_k}\left[U_1(\tau) - \beta U_{ref}X(\tau)\right]\mathrm{d}\tau - s\right) - 1 \\
&= 2\sigma\left(\frac{1}{T_2}\int_{t_0}^{t_k}\left[\frac{1}{T_1}\int_{t_0}^{\tau}\left[U_0(a) - \alpha U_{ref}X(a)\right]\mathrm{d}a - \beta U_{ref}X(\tau)\right]\mathrm{d}\tau - s\right) - 1 \\
&= 2\sigma\left(\frac{1}{T_1T_2}\int_{t_0}^{t_k}\int_{t_0}^{\tau}U_0(a)\mathrm{d}a\mathrm{d}\tau - \frac{\alpha U_{ref}}{T_1T_2}\int_{t_0}^{t_k}\int_{t_0}^{\tau}X(a)\,\mathrm{d}a\mathrm{d}\tau\right. \\
&\qquad\left. - \frac{\beta U_{ref}}{T_2}\int_{t_0}^{t_k}X(\tau)\,\mathrm{d}\tau - s\right) - 1.
\end{aligned} \tag{2.34}$$

Die Doppelintegrale werden gemäß

$$\int_{t_0}^{t}\int_{t_0}^{\tau} f\left(a\right) \mathrm{d}a\mathrm{d}\tau = \int_{t_0}^{t}\int_{a}^{t} f\left(a\right) \mathrm{d}\tau\mathrm{d}a = \int_{t_0}^{t} \left(t-a\right) f\left(a\right) \mathrm{d}a \tag{2.35}$$

umgeformt. Damit ergibt sich

$$\begin{aligned} X\left(t\right) &= 2\sigma\left(\frac{1}{T_1T_2}\int_{t_0}^{t_k}(t_k-a)U_0(a)\mathrm{d}a - \frac{\alpha U_{ref}}{T_1T_2}\int_{t_0}^{t_k}(t_k-a)X(a)\mathrm{d}a\right. \\ &\qquad \left. -\frac{\beta U_{ref}}{T_2}\int_{t_0}^{t_k}X(a)\mathrm{d}a - s\right) - 1. \end{aligned}$$

Zu den Zeitpunkten t_k ergeben sich mit den Bezeichnungen

$$\begin{aligned} \Delta t &:= t_{k+1} - t_k \\ X_k &:= X\left(t_k\right) \\ A_k &:= \int_{t_0}^{t_k} X(a)\mathrm{d}a \\ B_k &:= \int_{t_0}^{t_k} (t_k-a)X(a)\mathrm{d}a \\ C_k &:= \int_{t_0}^{t_k} (t_k-a)U_0(a)\mathrm{d}a \end{aligned}$$

für $k \in \mathbb{N}_0$ die folgenden Formeln

$$X_k = 2\sigma\left(\frac{1}{T_1T_2}C_k - \frac{\alpha U_{ref}}{T_1T_2}B_k - \frac{\beta U_{ref}}{T_2}A_k - s\right) - 1. \tag{2.36}$$

Wir berechnen zunächst nur die Größen A_k und B_k. Es folgt für

$$A_k = \int_{t_0}^{t_k} X(\tau)\mathrm{d}\tau = \sum_{j=0}^{k-1}\int_{t_j}^{t_{j+1}} X_j\mathrm{d}\tau = \sum_{j=0}^{k-1} X_j(t_{j+1}-t_j) = \Delta t\sum_{j=0}^{k-1}X_j \tag{2.37}$$

und

$$\begin{aligned} B_k &= \int_{t_0}^{t_k} X(a)(t_k-a)\mathrm{d}a = \sum_{j=0}^{k-1} X_j \int_{t_j}^{t_{j+1}} (t_k-a)\mathrm{d}a \\ &= \sum_{j=0}^{k-1} X_j\left(t_kt_{j+1} - \frac{1}{2}t_{j+1}^2 - t_kt_j + \frac{1}{2}t_j^2\right) \\ &= \sum_{j=0}^{k-1} X_j\left(t_{j+1}-t_j\right)\left(-\frac{1}{2}t_{j+1} + t_k - \frac{1}{2}t_j\right) \\ &= \Delta t\sum_{j=0}^{k-1} X_j\left(t_k - \frac{1}{2}\left(t_{j+1}+t_j\right)\right). \end{aligned}$$

Zur weiteren Vereinfachung führen wir erste Differenzen der Form

$$Y_{k+1} - Y_k := X_k \quad \text{mit} \quad Y_0 = 0$$

und zweite Differenzen der Form

$$Z_{k+1} - Z_k := Y_k \quad \text{mit} \quad Z_0 = Z_1 = 0$$

ein. Es ergibt sich für A_k

$$A_k = \Delta t \sum_{j=0}^{k-1} X_j = \Delta t \sum_{j=0}^{k-1} (Y_{j+1} - Y_j) = \Delta t (Y_k - Y_0) = Y_k \cdot \Delta t = (Z_{k+1} - Z_k)\Delta t$$

und für B_k

$$\begin{aligned} B_k &= \Delta t \left[t_k \sum_{j=0}^{k-1} X_j - \frac{1}{2} \sum_{j=0}^{k-1} X_j \, (t_{j+1} + t_j) \right] \\ &= \Delta t \left[t_k \cdot Y_k - \frac{1}{2} \sum_{j=0}^{k-1} (Y_{j+1} - Y_j) \, (t_{j+1} + t_j) \right] \\ &= \Delta t \left[t_k \cdot Y_k - \frac{1}{2} \sum_{j=0}^{k-1} (Y_{j+1} t_{j+1} - Y_j t_j) - \frac{1}{2} \sum_{j=0}^{k-1} (Y_{j+1} t_j - Y_j t_{j+1}) \right] . \end{aligned} \tag{2.38}$$

Der erste und zweite Term in (2.38) vereinfachen sich zu $\frac{1}{2} t_k Y_k$. Wir führen jetzt die zweiten Differenzen ein und erhalten

$$B_k = \Delta t \left[\frac{1}{2} t_k Y_k - \frac{1}{2} \underbrace{\sum_{j=0}^{k-1} (Y_{j+1} t_j - Y_j t_{j+1})}_{=:Q} \right],$$

wobei sich Q in nachfolgender Weise berechnen läßt:

$$\begin{aligned} Q &= \sum_{j=0}^{k-1} (Z_{j+2} - Z_{j+1}) \, t_j - \sum_{j=0}^{k-1} (Z_{j+1} - Z_j) \, t_{j+1} \\ &= \sum_{j=1}^{k} (Z_{j+1} - Z_j) \, t_{j-1} - \sum_{j=0}^{k-1} (Z_{j+1} - Z_j) \, t_{j+1} \\ &= -(Z_1 - Z_0) \, t_1 + (Z_{k+1} - Z_k) \, t_{k-1} + \sum_{j=1}^{k-1} \left[(Z_{j+1} - Z_j) \, t_{j-1} - (Z_{j+1} - Z_j) \, t_{j+1} \right] \\ &= -(Z_1 - Z_0) \, t_1 + (Z_{k+1} - Z_k) \, t_{k-1} - 2\Delta t \sum_{j=1}^{k-1} (Z_{j+1} - Z_j) \\ &= -(Z_1 - Z_0) \, t_1 + (Z_{k+1} - Z_k) \, t_{k-1} - 2\Delta t \, (Z_k - Z_1) \\ &= (Z_{k+1} - Z_k) \, t_{k-1} - 2\Delta t Z_k. \end{aligned}$$

Somit ergibt sich für B_k

$$\begin{aligned}
B_k &= \Delta t \left[\frac{1}{2}(Z_{k+1} - Z_k)\, t_k + \frac{1}{2}(2\Delta t Z_k - (Z_{k+1} - Z_k)\, t_{k-1})\right] \\
&= \Delta t^2 \left[\frac{1}{2}(Z_{k+1} - Z_k) + Z_k\right] \\
&= \frac{\Delta t^2}{2}(Z_{k+1} + Z_k)\,. \qquad (2.39)
\end{aligned}$$

Wir erhalten mit den berechneten Größen A_k, B_k die folgende Rekursionsformel

$$\begin{aligned}
X_k &= Y_{k+1} - Y_k = Z_{k+2} - Z_{k+1} - (Z_{k+1} - Z_k) \\
&= Z_{k+2} - 2Z_{k+1} + Z_k \\
&= 2\sigma\left(\frac{1}{T_1T_2}C_k - \frac{\alpha U_{ref}}{T_1T_2}B_k - \frac{\beta U_{ref}}{T_2}A_k - s\right) - 1 \\
&= 2\sigma\left(\frac{1}{T_1T_2}C_k - s - \frac{\alpha U_{ref}\Delta t^2}{2T_1T_2}(Z_{k+1} + Z_k) - \frac{\beta U_{ref}\Delta t}{T_2}(Z_{k+1} - Z_k)\right) - 1 \\
&= 2\sigma\left(\frac{U_{ref}\Delta t}{T_2}\left[\frac{C_k}{U_{ref}\Delta t T_1} - \frac{\alpha\Delta t}{2T_1}(Z_{k+1} + Z_k) - \beta(Z_{k+1} - Z_k) - \frac{T_2 s}{U_{ref}\Delta t}\right]\right) - 1.
\end{aligned}$$

Unter Benutzung der positiven Homogenität des Einheitssprunges läßt sich somit das Verhalten eines Σ-Δ-Konverters zweiter Ordnung durch ein nichtlineares, nichtautonomes Differenzenanfangswertproblem zweiter Ordnung

$$\begin{aligned}
Z_{k+2} - 2Z_{k+1} + Z_k &= 2\sigma\left(\frac{1}{U_{ref}\Delta t T_1}\int_{t_0}^{t_k}(t_k - \tau)U_0(\tau)\mathrm{d}\tau - \frac{T_2}{U_{ref}\Delta t}s\right. \\
&\quad \left. - \left(\frac{\alpha\Delta t}{2T_1} + \beta\right)Z_{k+1} + \left(\beta - \frac{\alpha\Delta t}{2T_1}\right)Z_k\right) - 1, \quad k \in \mathbb{N}_0, \qquad (2.40) \\
Z_0 &= Z_1 = 0
\end{aligned}$$

beschreiben. Wir wollen wie im Fall des Konverters erster Ordnung eine Transformation des Anfangswertproblems durchführen. Denn im Kapitel 5.3 werden wir für dieses neue Anfangswertproblem eine Näherungslösung berechnen und deren Fehler bestimmen.

Bemerkung 2.2.1 Durch die Vorschrift

$$Z_k = 2H_k + \frac{k - k^2}{2} \qquad (2.41)$$

erhalten wir folgendes Anfangswertproblem, welches zu (2.40) äquivalent ist

$$\begin{aligned}
&H_{k+2} - 2H_{k+1} + H_k = \\
&\quad \sigma\left(\frac{1}{U_{ref}\Delta t T_1}\int_{t_0}^{t_k}(t_k - \tau)U_0(\tau)\mathrm{d}\tau + \frac{\alpha\Delta t}{2T_1}k^2 + \beta k\right. \\
&\qquad \left. - \left(\frac{\alpha\Delta t}{T_1} + 2\beta\right)H_{k+1} + \left(2\beta - \frac{\alpha\Delta t}{T_1}\right)H_k - \frac{T_2}{U_{ref}\Delta t}s\right) \in \{0, 1\}, \qquad (2.42) \\
&\quad H_0 = H_1 = 0.
\end{aligned}$$

2.2.2 Erste Betrachtungen und Abschätzungen

Nun wollen wir das DAWP (2.42) einer ersten Untersuchung unterziehen. Entscheidend dabei ist die Annahme, daß $|U_1(t)| \leq \beta U_{ref}$ erfüllt ist. Dies ist gleichbedeutend, daß der innere Regelkreis wie ein Konverter erster Ordnung arbeitet, wobei das Eingangssignal U_1 ist. Zunächst werden wir das DAWP (2.42) unter dieser Annahme neu darstellen.

Lemma 2.2.2 *Es sei* $|U_1(t)| \leq \beta U_{ref}$, $\lambda = \dfrac{\alpha \Delta t}{2\beta T_1}$ *und* $s = 0$ *sowie*

$$\psi(k) := \frac{1}{2\beta T_1 \Delta t U_{ref}} \int_{t_0}^{t_k} (t_k - \tau) U_0(\tau) \mathrm{d}\tau + \frac{\lambda}{2} k^2 + \frac{k}{2} + 1.$$

Dann ist das DAWP (2.42) *äquivalent zum DAWP*

$$\begin{aligned} H_{k+2} &= \lfloor \psi(k) + (1-\lambda) H_{k+1} - \lambda H_k \rfloor, \\ H_0 &= H_1 = 0. \end{aligned} \tag{2.43}$$

Beweis: Es gelte o. B. d. A. für den Vergleichswert des Komparators $s = 0$. Es sei weiterhin

$$\varphi_{U_1}(k) := \begin{cases} 0, & \text{für } k = 0, \\ \dfrac{1}{2\beta \Delta t U_{ref}} \displaystyle\int_{t_0}^{t_k} U_1(\tau) \mathrm{d}\tau + \dfrac{k}{2}, & \text{für } k \in \mathbb{N}. \end{cases}$$

Nach obigen Voraussetzungen und den gleichen Überlegungen wie zu den Aussagen von Lemma 2.1.2 und Satz 2.1.3 folgt für die ersten Differenzen

$$W_{k+1} - W_k = X_k \quad \text{mit} \quad W_0 = 0$$

dann

$$W_k = \begin{cases} 0, & \text{für } k = 0, \\ \lfloor \varphi_{U_1}(k-1) \rfloor + 1, & \text{für } k \in \mathbb{N}. \end{cases}$$

Hieraus ergibt sich für die zweiten Differenzen H_k von X_k schließlich $H_0 = H_1 = 0$ und

$$\begin{aligned} H_{k+2} - H_{k+1} &= \lfloor \varphi_{U_1}(k) \rfloor + 1 \\ &= \left\lfloor \frac{1}{2\beta T_1 \Delta t U_{ref}} \int_{t_0}^{t_k} (t_k - \tau)[U_0(\tau) - \alpha U_{ref}(2X(\tau) - 1)] \mathrm{d}\tau + \frac{k}{2} \right\rfloor + 1 \\ &= \left\lfloor \frac{1}{2\beta T_1 \Delta t U_{ref}} \int_{t_0}^{t_k} (t_k - \tau) U_0(\tau) \mathrm{d}\tau + \frac{\lambda}{2} k^2 + \frac{k}{2} + 1 - \frac{\alpha}{\beta T_1 \Delta t} \int_{t_0}^{t_k} (t_k - \tau) X(\tau) \mathrm{d}\tau \right\rfloor \\ &= \lfloor \psi(k) - \lambda (H_{k+1} + H_k) \rfloor. \end{aligned}$$

Hieraus folgt wegen $H_{k+1} \in \mathbb{Z}$ unmittelbar die Behauptung. □

Nun wollen wir eine Näherungslösung des DAWP (2.43) angeben und den Fehler zur wahren Lösung abschätzen.

Lemma 2.2.3 *Es sei* $|U_1(t)| \le \beta U_{ref}$ *und* $\lambda \in (0, 3-\sqrt{8})$. *Weiterhin sei* $M = \sqrt{\lambda^2 - 6\lambda + 1}$ *Dann ist*

$$H_k^* := \frac{2\lambda}{M(-1+\lambda-M)} \sum_{n=2}^{k} \left(\left(\frac{-2\lambda}{-1+\lambda-M} \right)^{k-n} - \left(\frac{-2\lambda}{-1+\lambda+M} \right)^{k-n} \right) \psi(n-2)$$

eine Näherungslösung des DAWP (2.43) *mit einem Fehler zur wahren Lösung*

$$\begin{aligned} &|H_k - H_k^*| \le \\ &\quad -\frac{1}{2}\left(\frac{M^2+3M-\lambda M}{M^2(-1+\lambda-M)} \left(\frac{-2\lambda}{-1+\lambda-M} \right)^k + \frac{-3M+M^2+\lambda M}{M^2(-1+\lambda+M)} \left(\frac{-2\lambda}{-1+\lambda+M} \right)^k \right) + \frac{1}{2\lambda}. \end{aligned}$$

Darüber hinaus gilt

$$\lim_{k\to\infty} |H_k - H_k^*| \le \frac{1}{2\lambda}.$$

Beweis: Nach den Voraussetzungen ist das DAWP (2.43) äquivalent zu

$$H_{k+2} = \psi(k) + (1-\lambda)H_{k+1} - \lambda H_k - p_k \quad \text{mit} \quad p_k \in [0,1), \tag{2.44}$$

da φ_{U_1} den analogen Bedingungen zu (2.15) und (2.16) genügt (vgl. Lemma 2.2.2). Für die Lösung des Originalproblems folgt aus der Lösungsformel für die vollständige Lösung des linearen DAWP (2.44) mit der rechten Seite $\psi(k) - p_k$ (vgl. [19, Ch.3]) dann

$$H_k \in [H_k^1, H_k^0],$$

wobei H_k^0 bzw. H_k^1 die Lösungen der DAWP

$$\begin{aligned} H_{k+2}^0 &= \psi(k) + (1-\lambda)H_{k+1}^0 - \lambda H_k^0, \\ H_0^0 &= H_1^0 = 0 \end{aligned} \tag{2.45}$$

bzw.

$$\begin{aligned} H_{k+2}^1 &= \psi(k) + (1-\lambda)H_{k+1}^1 - \lambda H_k^1 - 1, \\ H_0^1 &= H_1^1 = 0 \end{aligned} \tag{2.46}$$

bezeichne. Die Lösungen von (2.45) bzw. (2.46) lauten (vgl. z.B. [19, Ch.3])

$$H_k^0 = \frac{2\lambda}{M(-1+\lambda-M)} \sum_{n=2}^{k} \left(\left(\frac{-2\lambda}{-1+\lambda-M} \right)^{k-n} - \left(\frac{-2\lambda}{-1+\lambda+M} \right)^{k-n} \right) \psi(n-2)$$

bzw.

$$\begin{aligned} H_k^1 = &\frac{1}{2}\left(\frac{M^2+3M-\lambda M}{M^2(-1+\lambda-M)} \left(\frac{-2\lambda}{-1+\lambda-M} \right)^k + \frac{-3M+M^2+\lambda M}{M^2(-1+\lambda+M)} \left(\frac{-2\lambda}{-1+\lambda+M} \right)^k \right) \\ &- \frac{1}{2\lambda} + \frac{2\lambda}{M(-1+\lambda-M)} \sum_{n=2}^{k} \left(\left(\frac{-2\lambda}{-1+\lambda-M} \right)^{k-n} - \left(\frac{-2\lambda}{-1+\lambda+M} \right)^{k-n} \right) \psi(n-2). \end{aligned}$$

M ist für

$$\lambda \in (0, 3-\sqrt{8}) \cup (3+\sqrt{8}, \infty)$$

reell. Nach der Wahl von H_k^* gilt also $H_k^* = H_k^0$. Wegen $H_k \in [H_k^1, H_k^0]$ gilt für den Fehler der Näherungslösung zur Originallösung

$$\begin{aligned}
|H_k - H_k^*| &\leq H_k^0 - H_k^1 \\
&= -\frac{1}{2}\left(\frac{M^2+3M-\lambda M}{M^2(-1+\lambda-M)}\left(\frac{-2\lambda}{-1+\lambda-M}\right)^k + \frac{-3M+M^2+\lambda M}{M^2(-1+\lambda+M)}\left(\frac{-2\lambda}{-1+\lambda+M}\right)^k\right) \\
&\quad + \frac{1}{2\lambda}.
\end{aligned}$$

Wegen $\lambda \in (0, 3-\sqrt{8})$ ist

$$\frac{-2\lambda}{-1+\lambda-M} \in (0,1) \quad \text{und} \quad \frac{-2\lambda}{-1+\lambda+M} \in (0,1).$$

Hieraus folgt sofort die letzte Behauptung. Damit ist alles bewiesen. $\square$

Bemerkung 2.2.4 Die Näherungslösung H_k^* ist wenig geeignet, das Eingangssignal durch eine Filterung wieder zu rekonstruieren. Dieser praktischen Fragestellung widmen wir uns später im Kapitel 5.
Die Wahl $\lambda \in (3+\sqrt{8}, \infty)$ ist wegen

$$\lim_{k\to\infty} H_k^0 - H_k^1 = +\infty$$

als Grundlage für einen Filtervorschlag ungeeignet.

3 Zur Wirkungsweise eines digitalen Filters

Der Konverterausgang X entspricht einer Einbitquantisierung des Eingangssignals U_0. Dem schließt sich eine digitale Filterung an, bei der U_0 durch eine Treppenfunktion, die aus X gebildet wird, angenähert ist. Wir untersuchen hier nur einen Typ - Filterung durch Mittelwertbildung. Darüber hinaus gibt es andere Typen von Filtern (vgl. Kapitel 8.4.2). Daran kann sich nun eine weitere digitale Verarbeitung der gewonnenen Daten entsprechend der Aufgabenstellung anschließen [54], [39], [26] und [56]. Die Darstellung in diesem Kapitel orientiert sich an [29], [33] und [31].

3.1 Vorbetrachtungen

Zunächst möchten wir die Funktionsweise eines Mittelwertbildners darstellen und einige grundlegende Eigenschaften dieses Filters für einen Σ-Δ-Modulator formulieren. Wir beginnen mit einem vorbereitenden Beispiel.

3.1.1 Beispiel

Bei der Filterung durch Mittelwertbildung werden die Wörter $\mathcal{W}^j$, $j = 0,1,2,\ldots$ (Bitfolgen der Länge n) wie folgt konstruiert

$$\mathcal{W}^j = (X_j, X_{j+1}, \ldots, X_{j+n-1}), \quad j = 0,1,2,\ldots \tag{3.1}$$

Wir betrachten zwei Mittelwertbildungen für jedes Wort $\mathcal{W}^j$. Die erste, welche die Signalgruppen generiert, ergibt sich durch gleitende Mittelung über $m-$Bits innerhalb des Wortes $\mathcal{W}^j$ gemäß

$$\begin{aligned} S_m^{j+k} &:= \frac{1}{m} \sum_{\mu=k+j}^{k+j+m-1} X_\mu \\ &= \frac{1}{m}(Y_{k+j+m} - Y_{k+j}) \quad k = 0,\, 1, \ldots, n-m,\; j = 0,1,2,\ldots \end{aligned} \tag{3.2}$$

Durch erneute Mittelung über die zu dem Wort $\mathcal{W}^j$ gehörenden $n-m+1$ Signalgruppen S_m^{j+k} entsteht der Signalgruppenmittelwert

$$S_{mn}^j := \frac{1}{n-m+1} \sum_{k=0}^{n-m} S_m^{j+k}, \quad j = 0,1,2,\ldots \tag{3.3}$$

Nach Konstruktion gilt

$$S_{mn}^j,\; S_m^{j+k} \in [-1,1], \quad k = 0,\, 1,2,\ldots, n-m,\; j = 0,1,2,\ldots.$$

Im folgenden betrachten wir bezüglich S_m^{j+k} und S_{mn}^j die auf das Intervall $[-\alpha U_{ref}, \alpha U_{ref}]$ relativierten Größen

$$\begin{aligned} R_m^{j+k-1} &:= \alpha U_{ref} S_m^{j+k}, \\ R_{mn}^{j-1} &:= \alpha U_{ref} S_{mn}^j = \frac{1}{n-m+1} \sum_{k=0}^{n-m} R_m^{j+k-1}. \end{aligned} \tag{3.4}$$

Wir verwenden im folgenden für R_m^{j+k-1} bzw. R_{mn}^{j-1} die Bezeichnungen (digitales) Signal bzw. (digitales) Mittelwertsignal. Analog verfahren wir bei den Fehlern F_m^{j+k-1} bzw. F_{mn}^{j-1} zwischen den Signalen R_m^{j+k-1} bzw. R_{mn}^{j-1} und noch zu definierenden geeigneten analogen Eingangsgrößen.

3.1.2 Fehler und Mittelwerte

Die Werte R_m^{j+k-1} kann man als digitale Approximation eines geeigneten Mittelwerts des analogen Eingangssignals interpretieren. Zunächst beschreiben wir den Quantisierungsfehler (vgl. Kapitel 8.2) im k-ten Abtastintervall. Unter Berücksichtigung des Anfangswertes $U_1(t_0)$ am Integratorausgang zum Zeitpunkt t_0 erhalten wir für die Funktion U_1 zum Zeitpunkt t_k die rekursive Beziehung

$$U_1(t_k) = \frac{1}{T_1} \int\limits_{t_{k-1}}^{t_k} \left(U_0(\tau) - \alpha U_{ref} X(\tau) \right) \mathrm{d}\tau + U_1(t_{k-1}) . \tag{3.5}$$

Definieren wir im k-ten Abtastintervall die quantisierte Analogantwort q_{k-1} durch

$$q_{k-1} := \frac{1}{T_1} \int\limits_{t_{k-1}}^{t_k} \alpha U_{ref} X(\tau) \mathrm{d}\tau = \frac{\Delta t}{T_1} \alpha U_{ref} X_{k-1}$$

und die Quantisierungsgröße e_k gemäß

$$e_k := U_1(t_k) - q_k \tag{3.6}$$

dann erhalten wir aus der Rekursion (3.5)

$$\begin{aligned} U_1(t_k) &= U_1(t_{k-1}) - \frac{1}{T_1} \int\limits_{t_{k-1}}^{t_k} \alpha U_{ref} X(\tau) \mathrm{d}\tau + \frac{1}{T_1} \int\limits_{t_{k-1}}^{t_k} U_0(\tau)\, \mathrm{d}\tau \\ &= e_{k-1} + \frac{1}{T_1} \int\limits_{t_{k-1}}^{t_k} U_0(\tau)\, \mathrm{d}\tau. \end{aligned}$$

Aus der letzten Formel folgt mit der Definition der e_k für $r = 0, 1, 2, ..., n-1$ die Rekursion

$$\begin{aligned} e_n &= U_1(t_n) - q_n \\ &= e_{n-1} + \frac{1}{T_1} \int\limits_{t_{n-1}}^{t_n} U_0(\tau) \mathrm{d}\tau - q_n \\ &= U_1(t_{n-1}) + \frac{1}{T_1} \int\limits_{t_{n-1}}^{t_n} U_0(\tau)\, \mathrm{d}\tau - q_{n-1} - q_n \\ &\vdots \\ &= e_r + \frac{1}{T_1} \int\limits_{t_r}^{t_n} U_0(\tau)\, \mathrm{d}\tau - \sum_{k=r+1}^{n} q_k. \end{aligned} \tag{3.7}$$

Mit

$$q_{k-1} = \frac{\Delta t}{T_1} \alpha U_{ref} X_{k-1} = \frac{\Delta t}{T_1} \alpha U_{ref} \left(Y_k - Y_{k-1} \right)$$

und (3.3) sowie (3.4) folgt

$$\begin{aligned} T_1 \frac{e_s - e_{s-m}}{m\Delta t} &= \frac{1}{m\Delta t} \int\limits_{t_{s-m}}^{t_s} U_0(\tau)\,\mathrm{d}\tau - \frac{T_1}{m\Delta t} \sum_{k=s-m+1}^{s} q_k \\ &= \frac{1}{m\Delta t} \int\limits_{t_{s-m}}^{t_s} U_0(\tau)\,\mathrm{d}\tau - \frac{\alpha}{m} U_{ref}\,(Y_{s+1} - Y_{s-m+1}) \\ &= \frac{1}{t_s - t_{s-m}} \int\limits_{t_{s-m}}^{t_s} U_0(\tau)\,\mathrm{d}\tau - R_m^{s-m} \end{aligned} \tag{3.8}$$

für $s, m \in \mathbb{N}$ mit $s > m$. Wir bezeichnen mit

$$U_m^{s-m} := \frac{1}{t_s - t_{s-m}} \int\limits_{t_{s-m}}^{t_s} U_0(\tau)\,\mathrm{d}\tau \tag{3.9}$$

den integralen Mittelwert von U_0 über m Takte. Unter Benutzung von (3.8) und (3.9) definieren wir durch

$$F_m^{s-m} := U_m^{s-m} - R_m^{s-m} = T_1 \frac{e_s - e_{s-m}}{m\Delta t} \tag{3.10}$$

den Fehler zwischen dem integralen Mittel des analogen Eingangssignals und des digitalen Signals (erste Mittelung) bezogen auf den Zeitraum $[t_{s-m}, t_s)$.

Wir geben jetzt Abschätzungen für den Fehler F_m^{s-m} unter der Bedingung an, daß $U_1(t_0) = 0$ gilt, d.h. der Kondensator (Integrierer) war bei Beginn des Prozesses nicht vorgeladen. Wir benutzen die expliziten Formeln (2.18) bzw. (2.21).

Lemma 3.1.1 *Es sei*

$$\chi_s := \varphi(s) - \lfloor \varphi(s) \rfloor \in [0, 1), \quad s \in \mathbb{N} \tag{3.11}$$

mit φ wie in Formel (2.13). *Dann gilt*

a.) $Y_{s+1} - Y_{s-m+1} = \dfrac{1}{\alpha U_{ref} \Delta t} \displaystyle\int\limits_{t_{s-m}}^{t_s} U_0(\tau)\mathrm{d}\tau - 2(\chi_s - \chi_{s-m}),$

b.) $F_m^{s-m} = \dfrac{2\alpha U_{ref}}{m}(\chi_s - \chi_{s-m}).$

Beweis: a.) Die Behauptung folgt aus (2.20), (2.13), (3.11) und (2.18)

$$\begin{aligned} Y_{s+1} - Y_{s-m+1} &= 2W_{s+1} - (s+1) - 2W_{s-m+1} + (s-m+1) \\ &= 2\lfloor \varphi(s) \rfloor - 2\lfloor \varphi(s-m) \rfloor - m \\ &= 2(\varphi(s) - \varphi(s-m)) - m - 2(\chi_s - \chi_{s-m}) \\ &= \frac{1}{\alpha U_{ref} \Delta t} \left(\int\limits_{t_0}^{t_s} U_0(\tau)\mathrm{d}\tau - \int\limits_{t_0}^{t_{s-m}} U_0(\tau)\mathrm{d}\tau \right) - 2(\chi_s - \chi_{s-m}) \\ &= \frac{1}{\alpha U_{ref} \Delta t} \int\limits_{t_{s-m}}^{t_s} U_0(\tau)\mathrm{d}\tau - 2(\chi_s - \chi_{s-m}). \end{aligned}$$

b.) Unter Benutzung von (3.10), eines Terms aus der Herleitung von (3.8) und des Ergebnisses in a.) ergibt sich

$$
\begin{aligned}
F_m^{s-m} &= \frac{1}{m\Delta t}\int\limits_{t_{s-m}}^{t_s} U_0(\tau)\mathrm{d}\tau - \frac{\alpha U_{ref}}{m}\left(Y_{s+1} - Y_{s-m+1}\right) \\
&= \frac{1}{m\Delta t}\int\limits_{t_{s-m}}^{t_s} U_0(\tau)\mathrm{d}\tau - \frac{\alpha U_{ref}}{m}\left(\frac{1}{\alpha U_{ref}\Delta t}\int\limits_{t_{s-m}}^{t_s} U_0(\tau)\mathrm{d}\tau - 2(\chi_s - \chi_{s-m})\right) \\
&= \frac{2\alpha U_{ref}}{m}(\chi_s - \chi_{s-m}).
\end{aligned}
$$

□

Folgerung 3.1.2 Der absolute Fehler F_m^{s-m} unterliegt für $m \geq 1$ den Schranken

$$\left|F_m^{s-m}\right| \leq \frac{2\alpha U_{ref}}{m}, \quad \forall s \geq m. \tag{3.12}$$

Umgekehrt läßt sich die Abschätzung nicht verbessern, d.h. zu jedem $\varepsilon > 0$ - hinreichend klein - existieren für alle Zeiten $t \geq t_0$ ein analoges Signal $U_{\varepsilon m}(t) \in [-\alpha U_{ref}, \alpha U_{ref}]$ und eine natürliche Zahl $s_0(\varepsilon, m) > m$, so daß

$$\left|F_m^{s_0-m}\right| \geq \frac{2\alpha U_{ref}}{m}(1-\varepsilon). \tag{3.13}$$

erfüllt ist.

Beweis: Die erste Behauptung folgt unmittelbar aus Lemma 3.1.1.
Nun wollen wir (3.13) beweisen. Es sei $\varepsilon \in \left(0, \frac{1}{2}\right)$ hinreichend klein. Weiterhin sei $m \in \mathbb{N}$ gegeben. Es sei $u(\varepsilon, m)$ so gewählt, daß

$$\frac{mu\left(\varepsilon, m\right)}{2\alpha U_{ref}} \in (0, \varepsilon) \quad \text{und} \quad s_0\left(\varepsilon, m\right) = \min_{s\in\mathbb{N}}\left\{s > \frac{2\alpha U_{ref}}{u\left(\varepsilon, m\right)}\right\}$$

erfüllt sind. Dann folgt

$$\frac{s_0(\varepsilon, m)u(\varepsilon, m)}{2\alpha U_{ref}} > 1 \quad \text{und} \quad (s_0(\varepsilon, m) - m)\frac{u(\varepsilon, m)}{2\alpha U_{ref}} \in \left(0, \frac{\varepsilon}{m}\right).$$

Hieraus ergibt sich schließlich durch Lemma 3.1.1 für $U_{\varepsilon m}(t) := \alpha U_{ref} - u(\varepsilon, m)$

$$
\begin{aligned}
F_m^{s_0-m} &= \frac{2\alpha U_{ref}}{m}(\chi_s - \chi_{s_0-m}) \\
&= \frac{2\alpha U_{ref}}{m}\left(\frac{1}{2\alpha U_{ref}\Delta t}\int\limits_{t_0}^{t_{s_0}}(\alpha U_{ref} - u(\varepsilon, m))\mathrm{d}\tau + \frac{s_0(\varepsilon, m)}{2}\right. \\
&\qquad \left. - \left\lfloor \frac{1}{2\alpha U_{ref}\Delta t}\int\limits_{t_0}^{t_{s_0}}(\alpha U_{ref} - u(\varepsilon, m))\mathrm{d}\tau + \frac{s_0(\varepsilon, m)}{2}\right\rfloor\right.
\end{aligned}
$$

$$
\begin{aligned}
&-\frac{1}{2\alpha U_{ref}\Delta t}\int\limits_{t_0}^{t_{s_0-m}}(\alpha U_{ref}-u(\varepsilon,m))\mathrm{d}\tau+\frac{s_0(\varepsilon,m)-m}{2}\\
&\quad+\left\lfloor\frac{1}{2\alpha U_{ref}\Delta t}\int\limits_{t_0}^{t_{s_0-m}}(\alpha U_{ref}-u(\varepsilon,m))\mathrm{d}\tau+\frac{s_0(\varepsilon,m)-m}{2}\right\rfloor\Bigg)\\
&=\frac{2\alpha U_{ref}}{m}\left(m\left(1-\frac{u(\varepsilon,m)}{2\alpha U_{ref}}\right)-\left\lfloor s_0(\varepsilon,m)\left(1-\frac{u(\varepsilon,m)}{2\alpha U_{ref}}\right)\right\rfloor\right.\\
&\quad\left.+\left\lfloor(s_0(\varepsilon,m)-m)\left(1-\frac{u(\varepsilon,m)}{2\alpha U_{ref}}\right)\right\rfloor\right)\\
&=\frac{2\alpha U_{ref}}{m}\left(-\frac{mu(\varepsilon,m)}{2\alpha U_{ref}}-\left\lfloor-\frac{s_0(\varepsilon,m)\,u(\varepsilon,m)}{2\alpha U_{ref}}\right\rfloor+\left\lfloor-\frac{(s_0(\varepsilon,m)-m)\,u(\varepsilon,m)}{2\alpha U_{ref}}\right\rfloor\right)\\
&\geq\frac{2\alpha U_{ref}}{m}\left(-\varepsilon+2-1\right).
\end{aligned}
$$

Die Behauptung $-\frac{2\alpha U_{ref}}{m}\leq F_m^{s_0-m}\leq-\frac{2\alpha U_{ref}}{m}(1-\varepsilon)$ wird analog bewiesen. □

Wir wollen nun den Zusammenhang zwischen dem Mittelwertsignal R_{mn}^j, d. h. der zweiten gleitenden Mittelung und einem geeigneten Mittelwert des analogen Eingangssignals betrachten.

Bemerkung 3.1.3 Analog zu den obigen Fehlerbetrachtungen ist

$$
U_{mn}^j:=\frac{1}{n-m+1}\sum_{k=0}^{n-m}\frac{1}{m\Delta t}\int\limits_{t_{k+j}}^{t_{k+m+j}}U_0\left(\tau\right)\mathrm{d}\tau
$$

die zu betrachtende Mittelwertbildung bezüglich R_{mn}^j. Mit ihr ist es jedoch schwer, geeignete Aussagen über den Fehler

$$
F_{mn}^j:=U_{mn}^j-R_{mn}^j=\frac{1}{n-m+1}\sum_{k=0}^{n-m}(U_m^{j+k}-R_m^{j+k})=\frac{1}{n-m+1}\sum_{k=0}^{n-m}F_m^{j+k}\tag{3.14}
$$

zu treffen. Deshalb wollen wir eine zusätzliche vereinfachende Annahme an U_0 treffen. Es sei $U_0(t)$ in den Intervallen $[t_j,t_{j+n})$ als Polynom darstellbar. Es ist leicht nachzurechnen, daß für linear affine Funktionen $U_0(t)=a+bt$ die Gleichung

$$
\frac{1}{n-m+1}\sum_{k=0}^{n-m}\frac{1}{m\Delta t}\int\limits_{t_k}^{t_{k+m}}U_0\left(\tau\right)\mathrm{d}\tau=\frac{1}{n\Delta t}\int\limits_{t_0}^{t_n}U_0\left(\tau\right)\mathrm{d}\tau
$$

gilt. Dies ist jedoch nicht wahr für Polynome höherer Ordnung. Aus diesem Grund nehmen wir an, daß der nichtlineare Anteil von U_0 im Intervall $[t_j,t_{j+n})$ klein bez. höherer Ordnung ist und somit vernachlässigt werden kann. Dann gilt wegen (3.14), (3.12) und $U_{mn}^j\approx U_n^j$ für den Mittelwertfehler F_{mn}^j bei einmaliger Mittelung, d. h. $m=1$ oder $m=n$

$$
\left|F_{mn}^j\right|\leq\frac{2\alpha U_{ref}}{n}.
$$

Dieser Wert ist eine obere Abschätzung von $|F_{mn}^j|$, da (3.12) zwar unabhängig von j,n ist, aber bei der nochmaligen Mittelung über $k=0,1,...,n-m$ erhalten wir durch Fehlerauslöschung Verbesserungen.

3.1.3 Asymptotische Aussagen

Ist die Funktion $U_0(t)$ bekannt, so sind die Werte χ_s wohlbestimmt. Im allgemeinen ist die Funktion $t \mapsto U_0(t)$ ein stochastischer Prozeß. In diesem Sinn kann χ_s für festes $s \in \mathbb{N}_0$ als Zufallsvariable mit Werten in $[0,1)$ mit unbekannter Verteilung aufgefaßt werden.
Wir normieren die Fehler F_m^{j+k} und F_{mn}^j, indem wir sie durch $2\alpha U_{ref}$ dividieren. Wir setzen $s := j+k$ und benutzen die gleiche Notation wie oben. Dann gilt

$$F_m^s := \frac{\chi_{s+m} - \chi_s}{m}, \tag{3.15}$$

$$F_{mn}^j := \frac{1}{n-m+1} \sum_{k=0}^{n-m} F_m^{j+k}. \tag{3.16}$$

In der ingenieurtechnischen Literatur (vgl. [52, Ch. 2] mit vielen Referenzen, [3] und die Kapitel 8.2, 8.3) werden Untersuchungen unter der Annahme geführt, daß die Fehler F_m^s stochastisch unabhängig und identisch gleichverteilt sind. Wir wollen zeigen, daß diese Annahme approximativ für hinreichend große j gerechtfertigt ist.
Im weiteren nehmen wir an, daß $(\chi_s)_{s\in\mathbb{N}_0}$ eine Folge von Zufallsvariablen auf $[0,1)$ mit unbekannten Verteilungen ist. Wir werden nun die Zufallsvariablen F_m^s und F_{mn}^j mit $s,j \in \mathbb{N}_0$, $n \in \mathbb{N}$ und $m \in \{1,2,\dots,n\}$ untersuchen.

Lemma 3.1.4 *Es seien $n = am + r$ mit $r \in \{0,1,2,\dots,m-1\}$, $a \in \mathbb{N}$ und χ_n seien Zufallsvariablen auf $[0,1)$. Dann gelten folgende Aussagen:*

$$a) \quad F_{mn}^j = \frac{1}{(n-m+1)\,m}\left[\sum_{k=0}^{r}(\chi_{j+k+am} - \chi_{j+k}) + \sum_{k=r+1}^{m-1}\left(\chi_{j+k+(a-1)m} - \chi_{j+k}\right)\right].$$

$$b) \quad \left|F_{mn}^j\right| \le \begin{cases} \dfrac{1}{n-m+1}, & \textit{für } m \le \dfrac{n}{2} \ (\textit{d. h. } a \ge 2) \\[2mm] \dfrac{1}{m}, & \textit{für } m > \dfrac{n}{2} \ (\textit{d. h. } a = 1). \end{cases}$$

Beweis: Wegen (3.15) und (3.16) gilt

$$F_{mn}^j = \frac{1}{n-m+1}\sum_{k=0}^{n-m}\frac{\chi_{k+m} - \chi_k}{m}.$$

In dieser Summe kann es nur Auslöschung von Summanden gleicher Indizes geben, wenn $a \ge 2$ ist. Schreibt man die obige Summe tabellarisch auf, so löschen sich zu übereinander stehenden Indizes gehörende Größen χ_n aus. Zur Vereinfachung wird dies im folgenden für $n = 2m+2$ tabellarisch dargestellt.

$$\begin{array}{ccccccccccc}
0 & & & \cdots & m & & & & & & \\
 & 1 & & \cdots & & m+1 & & & & & \\
 & & 2 & & & & m+2 & & & & \\
 & & & \ddots & & & & m+3 & & & \\
 & & & & m & & & \cdots & 2m & & \\
 & & & & & m+1 & & \cdots & & 2m+1 & \\
 & & & & & & m+2 & \cdots & & & 2m+2
\end{array}$$

Aus $|\chi_m - \chi_n| \le 1$ und a) folgt unmittelbar die Aussage in b). $\square$

Das folgende Lemma gibt eine typische Aussage über die Verteilung von Folgen von Zufallsvariablen, welche hilfreich bei der Bildung von Zufallszahlen sind. Eine leicht modifizierte Aussage ist in [62, Th.2] formuliert (vgl. auch [29]).

Lemma 3.1.5 *Es sei $(V_n)_{n\in\mathbb{N}_0}$ eine Folge von Zufallsvariablen auf $[0,1)$, wobei deren zugehörige Folge von Dichten $(g_n)_{n\in\mathbb{N}_0}$ gleichgradig stetig sei. Weiterhin sei durch*

$$\chi_n := nV_n - \lfloor nV_n \rfloor \in [0,1)$$

eine Folge von Zufallsvariablen definiert. Dann ist $(\chi_n)_{n\in\mathbb{N}_0}$ asymptotisch gleichverteilt auf $[0,1)$. Insbesondere gilt für $\lambda \in [0,1)$ und $\gamma \in (0,1]$ dann

$$P(\chi_n < \lambda) = \begin{cases} \lambda, & \text{falls } g_n \equiv 1 \quad \forall n \in \mathbb{N}_0, \\ \lambda + O\left(\dfrac{1}{n^\gamma}\right), & \text{falls } g_n \text{ gleichgradig Hölder-stetig sind.} \end{cases}$$

Dabei ist $\left|O\left(\dfrac{1}{n^\gamma}\right)\right| \le \dfrac{C}{n^\gamma}$ falls $|g_n(t) - g_n(t')| \le C\,|t-t'|^\gamma$ für alle $n \in \mathbb{N}_0$ und alle $t, t' \in [0,1)$ erfüllt ist.

Beweis: Es sei $\lambda \in [0,1)$. Dann gilt (vgl. [68, p.6])

$$\begin{aligned} P(\chi_n < \lambda) &= P\left(V_n \in \bigcup_{k=0}^{n-1}\left[\frac{k}{n}, \frac{k}{n} + \frac{\lambda}{n}\right)\right) \\ &= \sum_{k=0}^{n-1}\int_{\frac{k}{n}}^{\frac{k+\lambda}{n}} g_n(t)\mathrm{d}t \;=\; \sum_{k=0}^{n-1}\frac{1}{n}\int_0^\lambda g_n\left(\frac{k+t}{n}\right)\mathrm{d}t \\ &= \lambda\int_0^1 g_n(t)\,\mathrm{d}t + \lambda\left(\sum_{k=0}^{n-1}\frac{1}{n}g_n\left(\frac{k}{n}\right) - \int_0^1 g_n(t)\,\mathrm{d}t\right) + \sum_{k=0}^{n-1}\frac{1}{n}\int_0^\lambda\left(g_n\left(\frac{k+t}{n}\right) - g_n\left(\frac{k}{n}\right)\right)\mathrm{d}t \\ &=: \lambda + \lambda I + J. \end{aligned} \tag{3.17}$$

Wir untersuchen nun

$$I = \sum_{k=0}^{n-1}\frac{1}{n}g_n\left(\frac{k}{n}\right) - \int_0^1 g_n(t)\,\mathrm{d}t \quad \text{und} \quad J = \sum_{k=0}^{n-1}\frac{1}{n}\int_0^\lambda\left(g_n\left(\frac{k+t}{n}\right) - g_n\left(\frac{k}{n}\right)\right)\mathrm{d}t.$$

unter verschiedenen Annahmen an $(g_n)_{n\in\mathbb{N}_0}$.
1. Fall: Sei $(V_n)_{n\in\mathbb{N}_0}$ gleichverteilt auf $[0,1)$. Dann ist $g_n \equiv 1$ für alle $n \in \mathbb{N}_0$. Somit folgt unmittelbar $I = J = 0$ und daraus die Behauptung.
2. Fall: Sei $(g_n)_{n\in\mathbb{N}_0}$ gleichgradig stetig auf $[0,1)$. Wegen der gleichgradigen Stetigkeit gibt es dann zu jedem $\varepsilon > 0$ ein $\delta(\varepsilon) > 0$, so daß

$$\forall n \in \mathbb{N}\ \forall t, t' \in [0,1) \text{ mit } \left|t - t'\right| < \delta \Longrightarrow \left|g_n(t) - g_n(t')\right| < \varepsilon$$

erfüllt ist. Für $\dfrac{1}{n} < \delta(\varepsilon)$ mit $n \in \mathbb{N}$ ergibt sich dann

$$|I| \;=\; \left|\sum_{k=0}^{n-1}\frac{1}{n}g_n\left(\frac{k}{n}\right) - \int_0^1 g_n(t)\,\mathrm{d}t\right| = \left|\sum_{k=0}^{n-1}\int_{\frac{k}{n}}^{\frac{k+1}{n}}\left(g_n\left(\frac{k}{n}\right) - g_n(t)\right)\mathrm{d}t\right|$$

$$< \sum_{k=0}^{n-1} \int_{\frac{k}{n}}^{\frac{k+1}{n}} \varepsilon \mathrm{d}t = \varepsilon \tag{3.18}$$

und

$$|J| = \left| \sum_{k=0}^{n-1} \frac{1}{n} \int_0^\lambda \left(g_n\left(\frac{k+t}{n}\right) - g_n\left(\frac{k}{n}\right)\right) \mathrm{d}t \right| < \sum_{k=0}^{n-1} \frac{1}{n} \int_0^\lambda \varepsilon \mathrm{d}t = \sum_{k=0}^{n-1} \frac{1}{n} \lambda\varepsilon = \lambda\varepsilon. \tag{3.19}$$

Mit der Wahl $\frac{1}{n} < \delta(\varepsilon)$ - hier für $n \in \mathbb{N}$ - folgt aus (3.17) bis (3.19) die asymptotische Gleichverteilung von $(g_n)_{n\in\mathbb{N}_0}$.
3. Fall: Sei nun $(g_n)_{n\in\mathbb{N}}$ gleichgradig Hölder-stetig auf $[0,1)$ mit der Hölderkonstanten C zum Exponenten γ. Dann folgt analog zur gleichgradigen Stetigkeit auch

$$|I| \leq \frac{C}{n^\gamma} \quad \text{und} \quad |J| \leq \lambda \frac{C}{n^\gamma}$$

Hieraus ergibt sich unmittelbar die Behauptung. □

Bemerkung 3.1.6 In der Arbeit von Wilms [68] sind viele Ergebnisse dargestellt, die das Verhalten des ganzahligen Anteils von Summen oder Produkten von stochastisch unabhängigen und identisch gleichverteilten Zufallsvariablen untersuchen. Diese Resultate können leider nicht auf unsere Fragestellungen angewandt werden, da wir die stochastische Unabhängigkeit im unten aufgeführten zweiten Fall nicht voraussetzen können und da wir in beiden Fällen die Summe von ganzzahligen Anteilen von Zufallsvariablen betrachten.

Wir betrachten jetzt im folgenden zwei verschiedene Fälle:

A $(\chi_n)_{n\in\mathbb{N}_0}$ ist eine Folge stochastisch unabhängiger Zufallsvariablen, welche auf $[0,1)$ gleichverteilt sind.

B Alle Zufallsvariablen χ_n seien von der gleichen über $[0,1)$ gleichverteilten Zufallsvariablen V erzeugt. D. h., es ist $\chi_n = nV - \lfloor nV \rfloor$ für alle $n \in \mathbb{N}_0$.

Um den ersten Fall zu untersuchen, nutzen wir bekannte stochastische Argumente. Für den zweiten Fall benötigen wir einige spezielle algebraische Berechnungen, um die Struktur von χ_n zu beleuchten.
Aber welche rationalen Argumente gibt es von unserem Standpunkt aus, den zweiten Fall zu betrachten?

1. Da V_n durch (vgl. (2.13) dort mit $s \equiv 0$)

   $$V_n = \frac{1}{2n\alpha U_{ref}\Delta t} \int_{t_0}^{t_n} U_0(\tau)\, \mathrm{d}\tau + \frac{1}{2}$$

   definiert ist, wobei die Funktion $t \mapsto U_0(t)$ zu einem Wahrscheinlichkeitsraum gehört, werden wir sehen, daß für jedes $k \in \{1, 2, \ldots, n\}$ bei festem $n \in \mathbb{N}$ gilt

   $$\lim_{j\to\infty} (V_{j+k} - V_j) = 0 \text{ fast sicher.}$$

   Aus diesem Grund ist es berechtigt, alle Zufallsvariablen $V_j, V_{j+1}, ..., V_{j+n}$ zu identifizieren, falls j nur groß genug ist.

2. Der bezüglich der stochastischen Unabhängigkeit ungünstigste Fall scheint gerade der zu sein, bei dem alle χ_n durch die gleiche Zufallsvariable erzeugt werden.

3. Falls wir die maximale Frequenz von $U_0(t)$ im betrachteten Zeitintervall kennen, so können wir annehmen, daß die Ungleichung $|U_0(t_{n+1}) - U_0(t_n)| \leq K\Delta t$ für ein $K > 0$ erfüllt ist und daß die Differenz $U_0(t_{n+1}) - U_0(t_n)$ stochastisch unabhängig ist. Wir nehmen vereinfachend an, daß die V_n gleichverteilt auf $[0,1)$ sind.

4. Schließlich nehmen wir den einfachsten Fall an, daß $V_n \equiv V$ ist und daß V gleichverteilt auf $[0,1)$ ist.

3.2 Fall A: Unabhängigkeit und Gleichverteilung

Wir werden in diesem Abschnitt einige elementare Aussagen für den Fall treffen, daß $(\chi_n)_{n\in\mathbb{N}_0}$ eine Folge stochastisch unabhängiger auf $[0,1)$ gleichverteilter Zufallsvariablen ist. Die Resultate sind leicht beweisbar. Wir werden sie anschließend diskutieren.

Lemma 3.2.1 *Es sei* $(\chi_n)_{n\in\mathbb{N}_0}$ *eine Folge stochastisch unabhängiger auf* $[0,1)$ *gleichverteilter Zufallsvariablen. Dann gilt*

$$\begin{aligned} \mathrm{E}(F_m^s) &= 0, \\ \mathrm{D}(F_m^s) &= \frac{1}{m}\sqrt{\frac{1}{6}}, \end{aligned}$$

wobei $\mathrm{E}(F_m^s)$ *bzw.* $\mathrm{D}(F_m^s)$ *den Erwartungswert bzw. die Streuung von* F_m^s *bezeichnen.*

Beweis: Mit $\mathrm{E}(\chi_n) = \frac{1}{2}$ und $\mathrm{VAR}(\chi_n) = \mathrm{D}^2(\chi_n) = \frac{1}{12}$ für jedes $n \in \mathbb{N}_0$ (vgl. [18]) sowie der stochastischen Unabhängigkeit der χ_n und (3.15) folgen die obigen Beziehungen. □

Lemma 3.2.2 *Es sei* $(\chi_n)_{n\in\mathbb{N}}$ *eine Folge stochastisch unabhängiger auf* $[0,1)$ *gleichverteilter Zuvallsvariablen. Dann gilt*

$$\mathrm{E}\left(F_{mn}^j\right) = 0, \tag{3.20}$$

$$\mathrm{D}\left(F_{mn}^j\right) = \begin{cases} \dfrac{1}{\sqrt{m}(n-m+1)}\sqrt{\dfrac{1}{6}}, & \text{für } m \leq \dfrac{n+1}{2} \\ \dfrac{1}{m\sqrt{n-m+1}}\sqrt{\dfrac{1}{6}}, & \text{für } m > \dfrac{n+1}{2}. \end{cases} \tag{3.21}$$

Für die bestmögliche Streuung bei variablem m *und festem* n *gilt*

$$\min_{n\geq m\geq \frac{n+1}{2}} \mathrm{D}\left(F_{mn}^j\right) = \sqrt{\frac{1}{6}}\min\left\{\frac{1}{m_1\sqrt{n-m_1+1}}, \frac{1}{m_2\sqrt{n-m_2+1}}\right\}$$

mit $m_1 = \left\lfloor \frac{2}{3}(n+1) \right\rfloor$ *und* $m_2 = m_1 + 1$. *Wegen der Symmetrie bez.* $m = \frac{n+1}{2}$ *ergibt sich der gleiche Wert noch einmal bei* $m_3 = \left\lfloor \frac{1}{3}(n+1) \right\rfloor$ *und* $m_4 = m_3 + 1$.

Beweis: Die 1. und 2. Beziehung folgen sofort aus den Additionsgesetzen für Erwartungswerte und Streuungen bei unabhängigen und gleichverteilten Zufallsgrößen (vgl. [18]) sowie Teil a.) in Lemma 3.1.4.
Durch die folgende Extremwertdiskussion für differenzierbare Funktionen folgt für $m \in (0,n]$ die letzte Beziehung. Aus

$$\frac{\partial}{\partial m}\left(m\sqrt{n-m+1}\right)^{-1} = 0$$

folgt

$$m = \frac{2}{3}(n+1).$$

Wegen

$$\frac{\partial^2}{\partial m^2}\left(m\sqrt{n-m+1}\right)^{-1} > 0$$

für alle $m \in (0, n]$ folgt die Behauptung. $\square$

Bemerkung 3.2.3 Die stetige Abbildung $m \longmapsto \mathrm{D}F^j_{mn}$ (wie (3.21)) ist auf $\left[1, \frac{n+1}{2}\right]$ und $\left[\frac{n+1}{2}, n\right]$ konvex (vgl. Abb. 6, dort steht der Index u für Unabhängigkeit). In $m = \frac{n+1}{2}$ besitzt sie ein relatives Maximum. Den gleichen Wert erreicht sie noch zweimal, in $\left[1, \frac{n}{2}\right]$ für $m_l = \frac{1}{4}(3-\sqrt{5})(1+n) \approx 0.191(1+n)$ und in $\left[\frac{n+1}{2}, n\right]$ für $m_r = \frac{1}{4}\left(\sqrt{5}+1\right)(1+n) \approx 0.809\,(1+n)$. Bezogen auf den minimalen Wert $\mathrm{D}F^j_{\frac{2}{3}(n+1),n}$ ergibt sich folgende von n unabhängige relative Abweichung zur Streuung bei $m = \frac{n+1}{2}$

$$\frac{\mathrm{D}F^j_{\frac{n+1}{2},n} - \mathrm{D}F^j_{\frac{2}{3}(n+1),n}}{\mathrm{D}F^j_{\frac{2}{3}(n+1),n}} = \frac{\sqrt{2}}{18}\left(8\sqrt{3} - 9\sqrt{2}\right) \approx 0.0887.$$

Damit liegt diese relative Abweichung bei rund 9%. Deshalb ist es praktisch kaum von Bedeutung, welcher Wert m aus dem Intervall $[m_l, m_r]$ gewählt wird, um ein besseres Ergebnis durch zweimalige Mittelung zu erzielen, sofern die Unabhängigkeit erfüllt ist. Durch die vorangegangenen Betrachtungen erscheint die Wahl $m = \lfloor\frac{2}{3}(n+1)\rfloor$ als bestmögliche. Im Kapitel 3.3 werden wir Argumente gegen diese spezielle Wahl geben.

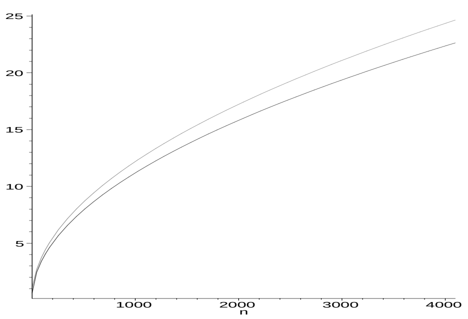


Abbildung 3: Faktor (Ordinate) für die Verbesserung nach der 2. Mittelung im Filter über die Wortlänge n, obere Kurve $m = \frac{2}{3}(n+1)$, untere Kurve $m = \frac{n}{2}$

Bemerkung 3.2.4 Unter der Annahme der stochastischen Unabhängigkeit ergibt sich durch die beste (wir setzen $m = \frac{2}{3}(n+1)$ ein) zweite Mittelung bezogen auf die bestmögliche erste Mittelung ($m = n$) eine Verbesserung der Streuung um den Faktor

$$\kappa_n = \frac{\mathrm{D}F^j_{1,n}}{\mathrm{D}F_{\frac{2}{3}(n+1),n}} = \frac{2\sqrt{3}\,(n+1)^{\frac{3}{2}}}{9n}.$$

Wir stellen die Verbesserung graphisch dar (vgl. Abb. 3). Gleichzeitig tragen wir die Verbesserung ab, die sich ergibt, wenn man anstelle von $m = \frac{2}{3}(n+1)$ den Wert $m = \frac{n}{2}$ nimmt

$$\omega_n = \frac{\mathrm{D}F^j_{1,n}}{\mathrm{D}F_{\frac{n}{2},n}} = \frac{n+2}{2\sqrt{2n}}.$$

3.3 Fall B: Identische Gleichverteilung

In diesem Abschnitt nehmen wir stets an, daß $V_n \equiv V$ für alle n gilt und V eine auf $[0,1)$ gleichverteilte Zufallsvariable ist. Dieses Kapitel orientiert sich stark an den Artikel [29].

3.3.1 Algebraische Vorbetrachtungen

Bevor wir im Fall B stochastische Kenngrößen berechnen können, müssen wir eine geeignete Darstellung der zugehörigen Integrale herleiten. Dafür benötigen wir zunächst einige algebraische Aussagen, die in diesem Kapitel bereitgestellt werden.

Lemma 3.3.1 *Es sei $v \in [0,1)$ Dann hat die stückweise lineare Funktion*

$$\delta(v) := (\chi_{n+m} - \chi_n)(v) = mv - \lfloor (m+n)v \rfloor + \lfloor nv \rfloor \tag{3.22}$$

die Darstellung

$$\delta(v) = mv - \sum_{k=0}^{n+m-1} \sigma\left(v - \frac{k+1}{n+m}\right) + \sum_{k=0}^{n-1} \sigma\left(v - \frac{k+1}{n}\right).$$

Es gilt für $v_0, w \geq 0$ und $0 \leq v_0 + w < 1$

$$\begin{aligned}\delta(v_0 + w) &= \delta(v_0) + mw - \sum_{\substack{k=0 \\ v_0 \leq \frac{k+1}{n+m}}}^{n+m-1} \sigma\left(v_0 + w - \frac{k+1}{n+m}\right) \\ &\quad + \sum_{\substack{k=0 \\ v_0 \leq \frac{k+1}{n}}}^{n-1} \sigma\left(v_0 + w - \frac{k+1}{n}\right).\end{aligned}$$

Beweis: Bekanntlich ist $\lfloor x \rfloor = \sum_{k=0}^{n-1} \sigma(x - (k+1)), 0 \leq x < n$, wobei $\sigma(x)$ den Einheitssprung bei $x = 0$ darstellt (vgl. (2.3)). Außerhalb der Sprungstellen ergibt sich für δ lineares Verhalten mit dem Anstieg m. In der letzten Formel sind nur noch die Sprünge zu berücksichtigen, die an Stellen größer oder gleich v_0 stattfinden. Die anderen wurden bereits in $\delta(v_0)$ berücksichtigt. □

Bemerkung 3.3.2 Die Zufallsvariable $\delta = (\chi_{n+m} - \chi_n)$ hängt von der Zufallsvariable V ab. Um die Streuung berechnen zu können, benötigen wir eine geschickte analytische Darstellung der unstetigen Funktion $V \longmapsto \delta(V)$, die es uns erlaubt, das Integral

$$\int_0^1 f(\delta(V))\, \mathrm{d}V$$

für eine gegebene stetige Funktion f zu berechnen. Wir haben bei einer analytischen Berechnung des Integrales das Integrationsintervall in Bezug auf die Unstetigkeitsstellen von δ zu zerlegen. Die Sprungstellen von δ hängen von zahlentheoretischen Größen ab, die durch die Parameter m, n bestimmt werden. Es stellt sich heraus, daß das analytische Verhalten von δ zwischen den Sprungstellen allgemein erfaßbar ist. Es kann gezeigt werden, daß eindeutig bestimmbare

Teilstücke auftreten. Jedoch ist nicht zu zeigen, in welcher Reihenfolge sie sich über $[0,1)$ anordnen. Da die Verteilung von V als Gleichverteilung angenommen wird, ist aber die Anordnung dieser linear affinen Teilstücke unerheblich. Wir wählen daher eine Anordnung dieser Teilstücke, die eine Berechnung des obigen Integrales ermöglicht.
Zunächst bestimmen wir die Formeln für alle auftretenden linear affinen Teilstücke unabhängig von ihrer real sich ergebenden Anordnung über $[0,1)$ unter Berücksichtigung eventueller Vielfachheiten.

Lemma 3.3.3 *Es sei* $I_k := \left[\frac{k}{n}, \frac{k+1}{n}\right)$, $k = 0,1,...,n-1$. *Weiterhin seien* $v_k \in [0,1)$ *und* $j_k \in \{0,1,\ldots,n+m-1\}$ *durch*

$$v_k := \min\left\{\frac{j}{n+m}\,\middle|\,\frac{j}{n+m} \in I_k\right\} \quad \textit{und} \quad j_k := (n+m)v_k$$

definiert. Dann gilt

$$m_k := \sup_{v\in I_k} \delta(v) = \lim_{v\to v_k-0} \delta(v) = 1 - nv_k + \lfloor nv_k \rfloor ,$$

Beweis: δ ist stückweise linear und auf jedem linearen Teilstück mit dem Anstieg m wachsend. Nur in den rationalen Punkten $\frac{k}{n}$ gibt es einen Sprung um 1 und nur in den Punkten $\frac{i}{n+m}$ gibt es einen Sprung -1 (vgl. Abb. 4). Natürlich können diese Sprünge zusammenfallen, und es ergibt sich eine stetige Fortsetzung (siehe Abb. 4 bei $v = 0.5$). Damit ergibt sich das Supremum in jedem Teilintervall I_k als linksseitiger Grenzwert an der ersten Stelle $\frac{i}{n+m}$ in dem betrachteten Intervall. Der Grenzwert berechnet sich wegen

$$\lim_{x\to n-0} \lfloor x \rfloor = n-1$$

gemäß Formel (3.22). □

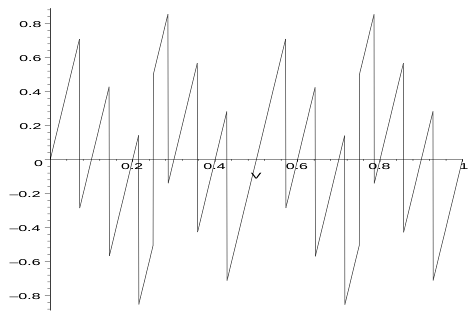


Abbildung 4: $\delta(v) = 10v - \lfloor 14v \rfloor + \lfloor 4v \rfloor$, also $m = 10$ und $n = 4$

Im weiteren nutzen wir die zahlentheoretische Zerlegung für $m, n \in \mathbb{Z}$

$$\mathrm{ggT}(m,n) = \gamma, m = a\gamma, n = b\gamma, a+b = cb+d, r = d\gamma, \tag{3.23}$$

wobei ggT den größten gemeinsamen Teiler der beiden ganzen Zahlen m und n bezeichnet. Die Zerlegung (3.23) ist eindeutig und existiert immer.

Lemma 3.3.4 *Es gelten die Bezeichnungen von* (3.23), *und es sei für* $k = 0,1,2,...,n-1$

$$\varrho_k := n\ \delta\left(\frac{k}{n}\right).$$

Dann gelten die folgenden Aussagen.
a) Die lokalen Supremumstellen von δ haben die Koordinaten

$$(v_k, m_k) = \left(\frac{k}{n} + \frac{n-\varrho_k}{n\,(n+m)}, \frac{m+\varrho_k}{n+m}\right), k = 0,1,2,...,n-1. \tag{3.24}$$

Weiterhin ist δ rechtsseitig stetig in den Punkten $\frac{k}{n}$, d.h.

$$\delta\left(\frac{k}{n}+0\right) = \frac{\varrho_k}{n},\ k = 0,1,2,...,n-1.$$

b) Die Zahlen ϱ_k, $k = 0,1,2,...,n-1$ durchlaufen periodisch eine Permutation der Menge

$$R := \{j\gamma\,|j = 0,1,2,...,b-1\}$$

mit der Vielfachheit γ.
c) Für die Anzahl J_k der Sprünge der Höhe -1 im Intervall I_k , $k = 0,1,2,...,n-1$ gilt

$$J_k = \begin{cases} c+1, & \textit{falls } r+\varrho_k > n, \\ c, & \textit{falls } r+\varrho_k \leq n. \end{cases} \tag{3.25}$$

d) Das letzte unvollständige Intervall hat die Länge

$$\beta_k = \begin{cases} \dfrac{r+\varrho_k-n}{n\,(n+m)}, & \textit{falls } r+\varrho_k > n, \\ \dfrac{r+\varrho_k}{n\,(n+m)}, & \textit{falls } r+\varrho_k \leq n. \end{cases} \tag{3.26}$$

Beweis: Als Hilfestellung für die nachfolgenden Teilbeweise kann die Abb. 5 für die Funktion δ mit $n = 6, m = 4, \mathrm{ggT}\,(n,m) = 2$ herangezogen werden.

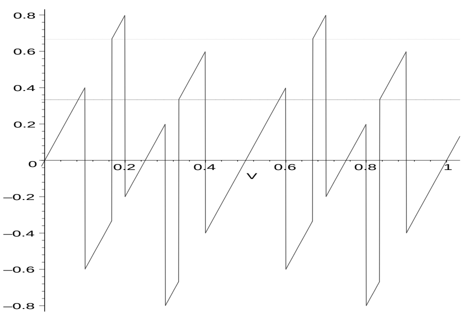


Abbildung 5: $\delta\,(v) = 4v - \lfloor 10v \rfloor + \lfloor 6v \rfloor$

a) Sei $v_k\,(q) := \frac{k}{n} + \frac{q}{m+n}$ für ein gewisses $q \in [0,1)$. Aus (3.22) und $\frac{nq}{m+n} \in [0,1)$ ergibt sich dann

$$\begin{aligned} \delta\,(v_k\,(q)) &= m\left(\frac{k}{n}+\frac{q}{m+n}\right) - \left\lfloor (m+n)\left(\frac{k}{n}+\frac{q}{m+n}\right)\right\rfloor + \left\lfloor n\left(\frac{k}{n}+\frac{q}{m+n}\right)\right\rfloor \\ &= \frac{mk}{n} + \frac{mq}{m+n} - \left\lfloor m\frac{k}{n}+k+q\right\rfloor + \left\lfloor k+\frac{nq}{m+n}\right\rfloor \\ &= \frac{mk}{n} + \frac{mq}{m+n} - \left\lfloor m\frac{k}{n}+q\right\rfloor. \end{aligned} \tag{3.27}$$

Speziell gilt für $q = 0$ wegen (3.27) dann

$$\frac{\varrho_k}{n} = \delta\left(\frac{k}{n}\right) = \frac{mk}{n} - \left\lfloor \frac{mk}{n} \right\rfloor. \tag{3.28}$$

Aus der rechtsseitigen Stetigkeit der Funktion $\lfloor . \rfloor$ folgt wegen (3.28) die von δ in den Punkten $\frac{k}{n}$. Weiterhin folgt wegen $x - \lfloor x \rfloor \in [0,1)$ aus (3.28) dann

$$\varrho_k \in \{0, 1, 2, ..., n-1\}\,. \tag{3.29}$$

Dann ergibt sich aus (3.27) und (3.29)

$$\begin{aligned}
\delta\left(v_k\left(q\right)\right) &= \frac{mk}{n} + \frac{mq}{m+n} - \left\lfloor m\frac{k}{n} + q \right\rfloor \\
&= \frac{mk}{n} - \left\lfloor \frac{mk}{n} \right\rfloor + \left\lfloor \frac{mk}{n} \right\rfloor + \frac{mq}{m+n} - \left\lfloor m\frac{k}{n} + q \right\rfloor \\
&= \frac{mk}{n} - \left\lfloor \frac{mk}{n} \right\rfloor + \frac{mq}{m+n} - \left\lfloor \frac{mk}{n} - \left\lfloor \frac{mk}{n} \right\rfloor + q \right\rfloor \\
&= \frac{\varrho_k}{n} + \frac{mq}{m+n} - \left\lfloor \frac{\varrho_k}{n} + q \right\rfloor.
\end{aligned} \tag{3.30}$$

Lokale Suprema für $\delta\left(v_k\left(q\right)\right)$ erhält man für $\frac{\varrho_k}{n} + q \longrightarrow 1 - 0$. Also folgt für $v_k = v_k\left(1 - \frac{\varrho_k}{n} - 0\right)$ dann

$$v_k - \frac{k}{n} = \frac{q}{m+n} = \frac{1 - \frac{\varrho_k}{n}}{m+n} = \frac{n - \varrho_k}{n\left(m+n\right)} \tag{3.31}$$

Aus (3.30) wiederum ergibt sich

$$m_k = \delta\left(v_k - 0\right) = \frac{\varrho_k}{n} + \frac{m\left(1 - \frac{\varrho_k}{n}\right)}{m+n} = \frac{\varrho_k + m}{m+n}. \tag{3.32}$$

Aus (3.31) und (3.32) folgt die Behauptung in a).
b) Wegen $m = a\gamma, n = b\gamma$ ergibt sich für

$$mk = n\left\lfloor m\frac{k}{n} \right\rfloor + \varrho_k$$

sofort

$$a\gamma k = b\gamma\left\lfloor m\frac{k}{n} \right\rfloor + \varrho_k.$$

Damit ist ϱ_k durch γ teilbar und wir setzen $\varrho_k' := \frac{\varrho_k}{\gamma}$. Es folgt $ak \equiv \varrho_k' \operatorname{mod}\left(b\right)$. Wegen $\operatorname{ggT}\left(a,b\right) = 1$ durchläuft ϱ_k' alle Reste von $0, ..., b-1$ in einer gewissen Permutation und zwar periodisch γ-mal. Damit ergibt sich für $\varrho_k = \gamma\varrho_k'$ die Aussage in b).
c) und d) Aus $m + n = cn + r$ ergibt sich

$$\frac{1}{n} = c\frac{1}{m+n} + \frac{r}{n\left(n+m\right)}.$$

Damit können maximal $(c+1)$ Sprünge der Länge 1 nach unten auftreten. Sei nun $r + \varrho_k > n$ dann ist der Rest $\frac{r}{n(n+m)}$ größer als das erste unvollständige Stück der Länge $\frac{n-\varrho_k}{n(m+n)}$ und $c+1$ Sprünge sind möglich. Die Länge des letzten unvollständigen Intervalles berechnet sich elementar zu $\frac{r}{n(n+m)} - \frac{n-\varrho_k}{n(m+n)}$. Analog folgt im zweiten Fall $\frac{n-\varrho_k}{n(m+n)} \geq \frac{r}{n(n+m)}$, daß ein Sprung nach unten in I_k (halboffenes Intervall !) entfällt. Die letzte unvollständige Intervallänge berechnet sich ebenfalls wieder elementar. □

3.3.2 Erwartungswerte

Wir können jetzt die affinen Teilstücke der Funktion δ umordnen, indem wir die obige natürliche Reihenfolge der ϱ_k benutzen und beachten, daß bei entsprechenden Periodizitäten nur über ein Periodenintervall integriert werden muß. Hieraus ergibt sich folgende Aussage.

Lemma 3.3.5 *Es seien $f : [0,1] \longrightarrow \mathbb{R}$ Lebesgue integrierbar, $V \in [0,1)$ und*

$$\delta_k(V) := \begin{cases} \dfrac{k\gamma}{n} + mV - \displaystyle\sum_{j=0}^{c} \sigma\left(V - \dfrac{n-k\gamma}{n(m+n)} - \dfrac{j}{n+m}\right), & \textit{falls } k > b-d, \\ \dfrac{k\gamma}{n} + mV - \displaystyle\sum_{j=0}^{c-1} \sigma\left(V - \dfrac{n-k\gamma}{n(m+n)} - \dfrac{j}{n+m}\right), & \textit{falls } k \leq b-d, \end{cases}$$

mit $\gamma = \mathrm{ggT}(m,n)$, $m = a\gamma$, $n = b\gamma$, $m+n = cn+r$, $r = d\gamma$. Dann gilt

$$\int_0^1 f(\delta(V))\,\mathrm{d}V = \gamma \sum_{k=0}^{b-1} \int_0^{\frac{1}{n}} f(\delta_k(V))\,\mathrm{d}V.$$

Beweis: Es gilt mit der Transformation $V = W + \frac{k}{n}$ auf den Intervallen I_k, $k = 0,1,2,...,n-1$

$$\int_0^1 f(\delta(V))\,\mathrm{d}V = \sum_{k=0}^{n-1} \int_{\frac{k}{n}}^{\frac{k+1}{n}} f(\delta(V))\,\mathrm{d}V = \sum_{k=0}^{n-1} \int_0^{\frac{1}{n}} f\left(\delta\left(\frac{k}{n} + W\right)\right)\mathrm{d}W.$$

Wegen $\delta\left(\frac{k}{n}\right) = \frac{\varrho_k}{n}$ folgt mit Lemma 3.3.1 für $W \in \left[0, \frac{1}{n}\right)$ die Darstellung

$$\delta\left(\frac{k}{n} + W\right) = \begin{cases} \dfrac{\varrho_k}{n} + mW - \displaystyle\sum_{j=0}^{c} \sigma\left(W - \dfrac{n-\varrho_k}{n(m+n)} - \dfrac{j}{n+m}\right), & \text{falls } r + \varrho_k > n, \\ \dfrac{\varrho_k}{n} + mW - \displaystyle\sum_{j=0}^{c-1} \sigma\left(W - \dfrac{n-\varrho_k}{n(m+n)} - \dfrac{j}{n+m}\right), & \text{falls } r + \varrho_k \leq n. \end{cases}$$

Die Reihenfolge der Summation ist unerheblich. Wir müssen nur alle Summanden abarbeiten. Unter Berücksichtigung der Vielfachheiten gemäß Lemma 3.3.4 und der in R aufgeführten natürlichen Reihenfolge ergibt sich die obige Formel. □

Wir stellen im folgenden Lemma einige aus der Stochastik bekannten Formeln [15] zusammen.

Lemma 3.3.6 *Es sei die Zufallsvariable V auf $[0,1)$ gleichverteilt. Dann gelten für die Zufallsvariablen $\chi_n = nV - \lfloor nV \rfloor$ mit $n \in \mathbb{N}$ folgende Aussagen:*

a) $\mathrm{E}\chi_q = \frac{1}{2}$, $\mathrm{E}\chi_q^2 = \frac{1}{3}$, $\mathrm{VAR}(\chi_q) = \frac{1}{12}$,

b) $\mathrm{E}(\chi_{q+p} - \chi_q) = 0$, $\mathrm{E}(\chi_{q+p} + \chi_q) = 1$,

c) $\mathrm{E}(\chi_q\chi_p) = \frac{1}{3} - \frac{1}{2}\mathrm{E}\left((\chi_q - \chi_p)^2\right)$,

d) $\mathrm{E}\left((\chi_q + \chi_p)^2\right) = \frac{2}{3} + 2\mathrm{E}(\chi_q\chi_p) = \frac{4}{3} - \mathrm{E}\left((\chi_q - \chi_p)^2\right)$,

e) $\mathrm{VAR}(\chi_q + \chi_p) = \frac{1}{3} - \mathrm{E}\left((\chi_q - \chi_p)^2\right)$,

f) $\mathrm{VAR}(\chi_q - \chi_p) = \mathrm{E}\left((\chi_q - \chi_p)^2\right)$.

Damit reduziert sich alles weitere auf die Berechnung des 2. Moments der Zufallsvariable $\chi_{n+m} - \chi_n$. Wir wählen die zuletzt aufgeführte Indizierung gemäß unseren Vorbetrachtungen.

Bemerkung 3.3.7 Für die Einheitssprungfunktion gilt die folgende bemerkenswerte Eigenschaft, auf die wir kurz hinweisen möchten. Es seien $0 \leq a_k \leq a_{k+1}$, $k = 1, 2, 3, ..., n$ und

$$A(V) := \sum_{k=1}^{n} \sigma(V - a_k).$$

Dann gilt

$$A(V)^2 = \sum_{k=1}^{n} (2(k-1)+1)\,\sigma(V - a_k).$$

Man sieht dies leicht, indem man für $a \leq b$ die Identität $\sigma(V-a)\,\sigma(V-b) = \sigma(V-b)$ benutzt. Mit

$$\left(\sum_{k=1}^{n} \sigma(V - a_k)\right)^2 = \sum_{k=1}^{n}\sum_{j=1}^{n} \sigma(V - a_k)\,\sigma(V - a_j)$$

ergibt sich über ein Dreieckssummationsschema, in dem jedes Hauptdiagonalelement einmal und jedes unterhalb der Hauptdiagonalen zweimal so oft auftritt wie Stellen links von der Hauptdiagonalen liegen, sofort die obige Formel.

Lemma 3.3.8 *Es sei $f : [0,1] \longrightarrow \mathbb{R}$ Lebesgue integrierbar. Dann gilt*

$$\begin{aligned}
&\int_0^1 f(\delta(V))\,\mathrm{d}V \\
&= \gamma \sum_{k=0}^{b-d} \left(\int_0^{\frac{n-k\gamma}{n(m+n)}} f\left(\frac{k\gamma}{n} + mV\right) \mathrm{d}V + \sum_{j=1}^{c-1} \int_{\frac{jn-k\gamma}{n(m+n)}}^{\frac{(j+1)n-k\gamma}{n(m+n)}} f\left(\frac{k\gamma}{n} + mV - j\right) \mathrm{d}V \right. \\
&\quad \left. + \int_{\frac{cn-k\gamma}{n(m+n)}}^{\frac{1}{n}} f\left(\frac{k\gamma}{n} + mV - c\right) \mathrm{d}V \right) \\
&\quad + \gamma \sum_{k=b-d+1}^{b-1} \left(\int_0^{\frac{n-k\gamma}{n(m+n)}} f\left(\frac{k\gamma}{n} + mV\right) \mathrm{d}V + \sum_{j=1}^{c} \int_{\frac{jn-k\gamma}{n(m+n)}}^{\frac{(j+1)n-k\gamma}{n(m+n)}} f\left(\frac{k\gamma}{n} + mV - j\right) \mathrm{d}V \right. \\
&\quad \left. + \int_{\frac{(c+1)n-k\gamma}{n(m+n)}}^{\frac{1}{n}} f\left(\frac{k\gamma}{n} + mV - (c+1)\right) \mathrm{d}V \right),
\end{aligned}$$

mit $\gamma = \mathrm{ggT}(m,n)$, $m = a\gamma$, $n = b\gamma$, $r = d\gamma$, $m + n = cn + r$.

Beweis: $\delta_k(V)$ hat die Struktur $q(V) - \sum_{j=0}^{c} \sigma(V - a - jb)$. Aus

$$\int_0^{\frac{1}{n}} f\left(q(V) - \sum_{j=0}^{c} \sigma(V - a - jb)\right) \mathrm{d}V$$
$$= \int_0^a f(q(V))\,\mathrm{d}V + \sum_{j=1}^{c-1} \left(\int_{a+jb}^{a+(j+1)b} f(q(V) - j)\,\mathrm{d}V \right) + \int_{a+cb}^{\frac{1}{n}} f(q(V) - c)\,\mathrm{d}V$$

folgt sofort der erste Teil (Summe über die Sprünge $k = 0, 1, ..., b - d$) der obigen Formel mit $a = \frac{n - k\gamma}{n(n+m)}$, $b = \frac{1}{n+m}$ und $q(V) = \frac{k\gamma}{n} + mV$.
Mit c anstelle von $c - 1$ ergibt sich analog der zweite Teil (Summe über die Sprünge $k = b - d + 1, ..., b - 1$). □

Kommen wir nun zum Hauptresultat dieses Kapitels. Darin werden wir zeigen, daß die Folge $(\chi_n)_{n \in \mathbb{N}_0}$ asymptotisch unkorreliert ist.

Satz 3.3.9 *Mit den folgenden Bezeichnungen* $\gamma = \mathrm{ggT}(m, n)$, $m = a\gamma$, $n = b\gamma$, $a + b = cb + d$ *gilt für die Streuung von* $(\chi_{n+m} - \chi_m)(V)$ *(vgl. Lemma* 3.3.1*) bei angenommener Gleichverteilung für* V *auf* $[0, 1)$

$$\mathrm{VAR}(\chi_{n+m} \pm \chi_m) = \frac{1}{6}\left(1 \pm \frac{\mathrm{ggT}(m+n, n)}{\mathrm{kgV}(m+n, n)}\right)$$

und

$$\mathrm{COV}(\chi_{n+m}, \chi_m) = \frac{1}{12} \frac{\mathrm{ggT}(m+n, n)}{\mathrm{kgV}(m+n, n)}.$$

Darüber hinaus ist $(\chi_n)_{n \in \mathbb{N}_0}$ *asymptotisch unkorreliert, d. h., für feste* $p, q \in \mathbb{N}$ *gilt*

$$\lim_{j \to \infty} \mathrm{COV}(\chi_{p+j}, \chi_{q+j}) = \frac{1}{12}\delta_{pq} = \frac{1}{12}\begin{cases} 1, & \textit{falls } p = q, \\ 0, & \textit{falls } p \neq q. \end{cases}$$

Beweis: Man benutze die Formel in Lemma 3.3.8 für $f(x) = x^2$. Das obige Ergebnis wurde unter Benutzung von Maple V berechnet. Zunächst wurde in der obigen Formel mittels „Replace“ die obigen Ersetzungen vorgenommen, d.h. $m = a\gamma, n = b\gamma, a = (c-1)b + d$. Auf diese so abgeänderte gesamte Formel wurde „Simplify“ angewendet. Wir erhielten das Ergebnis

$$\frac{1}{6} \frac{b^2 c + bd \pm 1}{b(cb + d)} = \frac{1}{6}\left(1 \pm \frac{\mathrm{ggT}(m+n, n)}{\mathrm{kgV}(m+n, n)}\right).$$

Kontrollrechnungen für den Erwartungswert $(f(x) = x)$ lieferten den richtigen Wert 0. Man beachte dabei, daß die Fälle $d \in \{0, 1\}$ getrennt durchgerechnet werden müssen, ebenso der Fall $c = 1$, da einige Summenterme entfallen.
Es sei $q = n, p = m + n$, dann folgt mittels Lemma 3.3.6

$$\mathrm{E}\left(\left(\chi_{n+m} - \frac{1}{2}\right)\left(\chi_m - \frac{1}{2}\right)\right) = \mathrm{E}(\chi_{n+m}\chi_m) - \frac{1}{2}\mathrm{E}(\chi_{n+m} + \chi_m) + \frac{1}{4}$$
$$= \frac{1}{12} - \frac{1}{2}\mathrm{E}\left((\chi_{n+m} - \chi_m)^2\right) = \frac{1}{12} - \frac{1}{12}\frac{b(cb + d) - 1}{b(cb + d)}$$
$$= \frac{1}{12b(cb + d)} = \frac{1}{12}\frac{\gamma^2}{qp} = \frac{1}{12}\frac{\mathrm{ggT}(p, q)}{\mathrm{kgV}(p, q)},$$

mit $\operatorname{ggT}(p,q) = \gamma, b = \frac{q}{\gamma}, p = cq + d\gamma, d = \frac{p-cq}{\gamma}$.
Die Ungleichung

$$\operatorname{ggT}(p+j, q+j) \leq \min\{p+j, q+j, |p-q|\}$$

liefert sofort die asymptotische Eigenschaft. □

3.3.3 Fehler in diesem Fall und Vergleich

Zurück zum Filter des Konverters erster Ordnung. Wir wollen nun den Fehler im Fall B abschätzen und ihn dann mit dem des Falles A vergleichen. Beginnen wollen wir mit einer Folgerung aus dem Satz 3.3.9.

Folgerung 3.3.10 Es sei $V_n \equiv V$ für alle $n \in \mathbb{N}$ und V ist eine Zufallsvariable, welche auf $[0,1)$ gleichverteilt ist. Für den Fehler F_m^s gilt dann

$$\mathrm{E}\left(F_m^s\right) = 0$$

$$\mathrm{D}\left(F_m^s\right) = \frac{1}{m}\mathrm{D}(\chi_{s+m} - \chi_s) = \frac{1}{m}\sqrt{\frac{1}{6}\left(1 - \frac{\operatorname{ggT}(m+s,s)}{\operatorname{kgV}(m+s,s)}\right)} \tag{3.33}$$

und

$$\frac{1}{m}\sqrt{\frac{1}{6}\left(1 - \frac{\left(\min\{s,m\}\right)^2}{(m+s)\,s}\right)} \leq \mathrm{D}\left(F_m^s\right) \leq \frac{1}{m}\sqrt{\frac{1}{6}\left(1 - \frac{1}{(m+s)\,s}\right)}. \tag{3.34}$$

Beweis: Die ersten beiden Aussagen folgen sofort aus Lemma 3.3.6 in Verbindung mit Satz 3.3.9. Die Abschätzungen ergeben sich aus (3.33) und der Ungleichung

$$1 \leq \operatorname{ggT}(m+k,k) \leq \min\{k,m\}$$

sowie der Beziehung

$$\frac{\operatorname{ggT}(m+k,k)}{\operatorname{kgV}(m+k,k)} = \frac{\left(\operatorname{ggT}(m+k,k)\right)^2}{(m+k)\,k}.$$

□

Bemerkung 3.3.11 Ein Vergleich beider Fälle zeigt, daß die Streuung $\mathrm{D}\left(F_m^s\right) = \frac{1}{m}\sqrt{\frac{1}{6}}$, welche unter der Annahme der Unabhängigkeit der Zufallsvariablen χ_n in Lemma 3.2.1 berechnet wurde, eine geeignete Approximation der wirklichen Streuung des stochastischen Fehlers F_m^s ist.
Nun wollen wir Lemma 3.1.4 nutzen, um die Streuung $\mathrm{D}(F_{mn}^j)$ zu bestimmen. Nach Satz 3.3.9 sind die Zufallsvariablen $\chi_j, \chi_{j+1}, \ldots, \chi_{j+k}$ für große $j \in \mathbb{N}$ asymptotisch unkorreliert, da die Kovarianzmatrix gegen die Einheitsmatrix konvergiert. Aus diesem Grund betrachten wir hier nun den ungünstigsten Fall $j = 1$. Nach Lemma 3.1.4 erhalten wir mit $n = am + r$ für $0 \leq r < m$ dann

$$\mathrm{D}\left(F_{mn}\right) = \frac{1}{\left((a-1)\,m + r + 1\right)m}\mathrm{D}\left(\sum_{k=1}^{r+1}\left(\chi_{k+am} - \chi_k\right) + \sum_{k=r+2}^{m}\left(\chi_{k+(a-1)m} - \chi_k\right)\right). \tag{3.35}$$

Betrachten wir den Fall $a = 1$, so führt dies zu

$$\mathrm{D}(F_{mn}) = \frac{1}{(r+1)m}\mathrm{D}\left(\sum_{k=1}^{r+1}(\chi_{k+m} - \chi_k)\right)$$

mit $n \geq m \geq \frac{n+1}{2}$. Wir erhalten somit

$$\begin{aligned}
\mathrm{VAR}\left(\sum_{k=1}^{r+1}(\chi_{k+m}-\chi_k)\right) &= \mathrm{E}\left(\left(\sum_{k=1}^{r+1}(\chi_{k+m}-\chi_k)\right)^2\right) \\
&= \sum_{k=1}^{r+1}\mathrm{E}\left(\chi_{k+m}-\chi_k\right)^2 + 2\sum_{k=1}^{r}\sum_{j=k+1}^{r+1}\mathrm{E}\left(\left(\chi_{k+m}-\chi_k\right)\left(\chi_{j+m}-\chi_j\right)\right)
\end{aligned}$$

mit

$$\begin{aligned}
&2\mathrm{E}\left(\left(\chi_{k+m}-\chi_k\right)\left(\chi_{j+m}-\chi_j\right)\right) \\
&\quad = 2\left(\mathrm{E}\left(\chi_{k+m}\chi_{j+m}\right)+\mathrm{E}\left(\chi_k\chi_j\right)-\mathrm{E}\left(\chi_{k+m}\chi_j\right)-\mathrm{E}\left(\chi_k\chi_{j+m}\right)\right) \\
&\quad = -\mathrm{E}\left(\left(\chi_{k+m}-\chi_{j+m}\right)^2\right)-\mathrm{E}\left(\left(\chi_k-\chi_j\right)^2\right)+\mathrm{E}\left(\left(\chi_{k+m}-\chi_j\right)^2\right)+\mathrm{E}\left(\left(\chi_k-\chi_{j+m}\right)^2\right) \\
&\quad = \frac{\mathrm{ggT}\left(k+m,j+m\right)^2}{\left(k+m\right)\left(j+m\right)}+\frac{\mathrm{ggT}\left(k,j\right)^2}{kj}-\frac{\mathrm{ggT}\left(k+m,j\right)^2}{\left(k+m\right)j}-\frac{\mathrm{ggT}\left(j+m,k\right)^2}{\left(j+m\right)k}
\end{aligned}$$

und

$$\mathrm{E}\left(\chi_{k+m}-\chi_k\right)^2 = \frac{1}{6}\left(1-\frac{\mathrm{ggT}\left(k+m,k\right)^2}{\left(k+m\right)k}\right).$$

Also ergibt sich schließlich

$$\begin{aligned}
\mathrm{VAR}\left(\sum_{k=1}^{r+1}(\chi_{k+m}-\chi_k)\right) &= \frac{1}{6}\left(\sum_{k=1}^{r+1}\left(1-\frac{\mathrm{ggT}\left(k+m,k\right)^2}{\left(k+m\right)k}\right)\right. \\
&\left.+\sum_{k=1}^{r}\sum_{j=k+1}^{r+1}\left(\frac{\mathrm{ggT}\left(k+m,j+m\right)^2}{\left(k+m\right)\left(j+m\right)}+\frac{\mathrm{ggT}\left(k,j\right)^2}{kj}-\frac{\mathrm{ggT}\left(k+m,j\right)^2}{\left(k+m\right)j}-\frac{\mathrm{ggT}\left(j+m,k\right)^2}{\left(j+m\right)k}\right)\right).
\end{aligned} \tag{3.36}$$

Leider waren wir nicht in der Lage, analytisch für (3.36) scharfe obere und untere Schranken zu bestimmen Deshalb untersuchen wir das Verhalten von (3.36) numerisch für $5 \leq n \leq 1024$. Für $r = 0, 1, 2, ..., n$ mit $n \leq 200$ können die folgenden Betrachtungen mit „Mathematica“ unmittelbar nachvollzogen werden. Die Rechenzeiten liegen im Sekundenbereich. Wir werden die Ergebnisse für $n = 1024$ darlegen. Mittels Mathematica 2.22 wurde die Streuung der Fehlerfunktion F_{mn} mit $r = n - m$ gemäß

$$\mathrm{D}(F_{mn}) = \frac{1}{m\left(n-m+1\right)}\mathrm{D}\left(\sum_{k=1}^{n-m+1}\left(\chi_{k+m}-\chi_k\right)\right)$$

im interessanten Bereich $m \in \left[\frac{n}{2}, \frac{5}{8}n\right]$ geplottet (Abb.7). Die Funktion $m \mapsto \mathrm{D}(Fu_{mn})$, welche die Streuung im Fall der stochastischen Unabhängigkeit bezeichne, ist durch Lemma 3.2.2 gegeben. Die Funktion ist ebenfalls in Abbildung 7 dargestellt. Wir werden nun die Unterschiede zwischen beiden Extrema darlegen.
In beiden Fällen finden wir das globale Minimum der Funktionen $m \mapsto \mathrm{D}(F_{m,1024})$ und $m \mapsto \mathrm{D}(Fu_{m,1024})$ symmetrisch rechts und links von $m = \frac{n+1}{2}$. Daher genügt es, die Werte rechts von $m = \frac{n}{2}$ zu betrachten (vgl. Abb. 6, Abb. 7). Im Fall der stochastischen Abhängigkeit vermuten wir aufgrund der numerischen Berechnungen, daß das Minimum für wachsende $n \in \mathbb{N}$ gegen $\frac{n+1}{2}$ strebt. Dies steht im Widerspruch zum Fall der stochastischen Unabhängigkeit, bei dem das Minimum nahe der beiden Werte $m = \frac{n+1}{3}$ und $m = 2\frac{n+1}{3}$ zu finden ist (vgl. Lemma 3.2.2)).

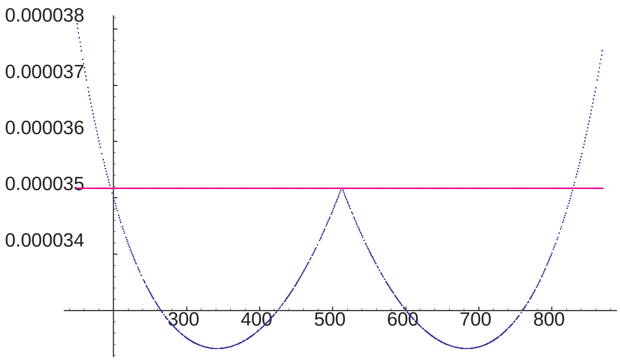


Abbildung 6: $\mathrm{D}Fu_{m,1024}$, $150 \leq m \leq 875$

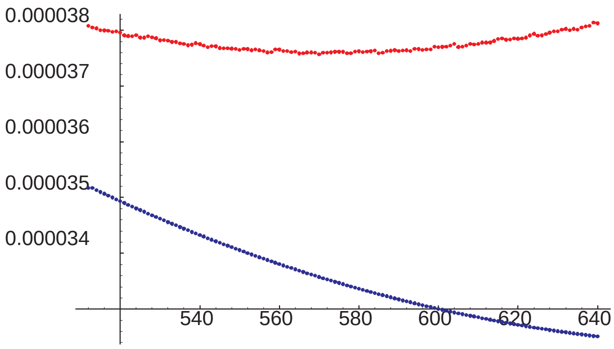


Abbildung 7: $\mathrm{D}F_{m,1024}$ (oben), $\mathrm{D}Fu_{m,1024}$ (unten)

In Abb. 7 ist nur das rechte Minimum von $\mathrm{D}(F_{m,1024})$ bei $m = 570$ zu sehen, das linke Minimum liegt bei $m = 456$. Beide Streuungen haben für $m = \lfloor \frac{n+1}{2} \rfloor = 512$ ein lokales Maximum. Die globalen Maxima liegen bei $m = 0$ und $m = n$ (vgl. Abb. 7). Die Funktion $m \mapsto \mathrm{D}(F_{m,1024})$ hat insgesamt auf $[1, \frac{n+1}{2}]$ und auf $[\frac{n+1}{2}, n]$ eine „grob konvexe“ bez. $m = \frac{n+1}{2}$ symmetrische Struktur, womit wir andeuten wollen, daß bei niedriger Auflösung die Funktion auf dem Bildschirm „konvexes“ Verhalten links und rechts von $m = \frac{n+1}{2}$ zeigt. Bei höherer Auflösung zeigt sich die lokal verrauschte Struktur (Abb. 8) der Fehlerfunktion $m \mapsto \mathrm{D}(F_{m,1024})$. Bemerkenswert dabei ist, daß der Rechenaufwand im interessanten Bereich $\frac{3}{8}n \leq m \leq \frac{n}{2}$ mit der dritten Potenz bez. n wächst (vgl. Abb.9). Für Werte von $n = 1$ bis 200 kann dies leicht auf dem Rechner nachvollzogen werden.
Die Funktionswertunterschiede zwischen dem lokalen Maximum bei $m = \lfloor \frac{n+1}{2} \rfloor$ mit

$$\mathrm{D}(F_{512,1024}) = 3.80830 \times 10^{-5}$$

und dem in der „Nähe“ liegenden globalen Minimum

$$\mathrm{D}(F_{570,1024}) = 3.75681 \times 10^{-5}$$

sind für die Fehlerbetrachtung ohne Bedeutung. Für $m \in [512, 640]$ gilt

$$\mathrm{D}(F_{m,1024}) \leq 3.81401 \times 10^{-5}.$$

Für die Berechnung wurde auf einer Sun - Sparc - Workstation (Time sharing) mit Mathematica ca. 81688 s CPU-Time benötigt. Unter der Annahme der Unabhängigkeit erhielten wir für $m \geq$

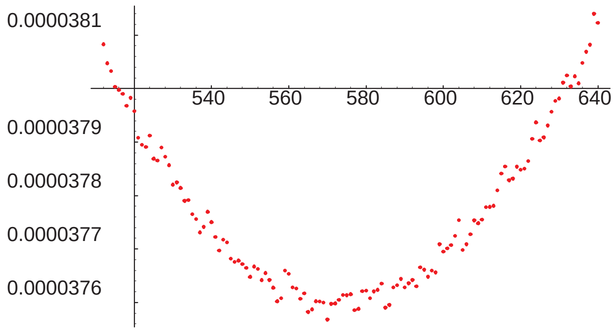


Abbildung 8: $\mathrm{D}(F_{m,1024})$, $m = 512, 513, ..., 640$

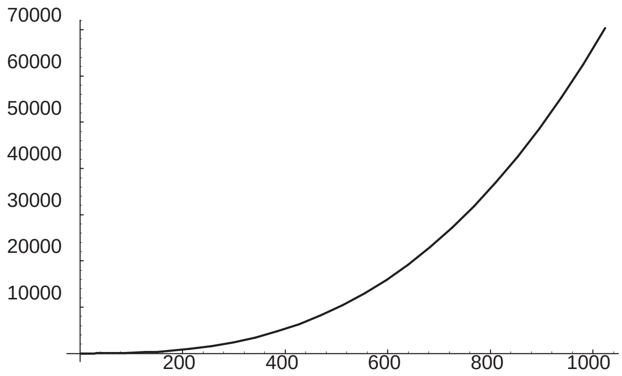


Abbildung 9: Abszisse Variable n, Ordinate CPU Zeit in Sekunden

$\lfloor \frac{n+1}{2} \rfloor$ den Fehler (vgl. Lemma 3.2.2)

$$\mathrm{D}(Fu_{m,n}) = \frac{1}{m\sqrt{n-m+1}}\sqrt{\frac{1}{6}}$$

für

$$m = \frac{2}{3}(n+1) = \frac{2050}{3}$$

den optimalen Wert für Minimum und Maximum

$$\begin{aligned} \mathrm{D}(Fu_{683,1024}) &\approx 3.2321 \times 10^{-5} \\ \mathrm{D}(Fu_{512,1024}) &\approx 3.5204 \times 10^{-5}. \end{aligned}$$

Der optimale Fehler unter Berücksichtigung der Abhängigkeit ist also um den Faktor $\frac{3.75681}{3.23210} \approx 1.16$ und der Fehler unter Berücksichtigung des Wertes $m = 512$ ist sogar nur um den Faktor $\frac{3.8083}{3.5204} \approx 1.08$ schlechter. In Bezug auf die erste Mittelung ergibt sich mit Bemerkung 3.2.4 im Schnitt bei $m = \lfloor \frac{n}{2} \rfloor$ eine Verbesserung mit dem Faktor $\frac{n+2}{2\sqrt{2n}}$.

Wie wir also gesehen haben, ist es wenig von Bedeutung, welchen Fall wir betrachten, sofern wir nur nahe $\frac{n+1}{2}$ sind. Der Unterschied ist recht gering. Falls wir aber den Punkt des Minimums bei Unabhängigkeit als Grundlage nehmen und dann den Wert in diesem Punkt für den Fall der stochastischen Abhängigkeit berechnen, so wird die Differenz größer. Deshalb gelangen wir zu folgendem verbalen Schluß.

Folgerung 3.3.12 Da die genaue Verteilung der Zufallsvariable χ_n für $n \in \mathbb{N}$ unbekannt ist, ist die beste Wahl, den Filterfehler bei zweifacher Mittelung zu minimieren, durch $m = \lfloor \frac{n+1}{2} \rfloor$ gegeben.

Abschließend wenden wir uns einem weiteren asymptotischen Aspekt zu, dem eine gewisse Bedeutung zukommt. Wir untersuchen das asymptotische Verhalten der Funktion $m \mapsto \mathrm{D}(F^{\tau}_{mn}) - \mathrm{D}(Fu_{mn})$ falls $\tau \gg n$ ist. Ist τ hinreichend groß, so kann man zeigen, daß die Formel (3.21) eine gute Näherung von (3.35) ist. Dies zeigt zusätzlich, daß die Annahme der stochastischen Unabhängigkeit der Fehler bei der Modellbildung bei Σ-Δ-Modulatoren in der ingenieurtechnischen Literatur eine gewisse Berechtigung besitzt.

Satz 3.3.13 *Es sei $\tau > n \geq 3$ und $1 \leq m \leq n$. Dann gilt*

$$\mathrm{D}(Fu_{mn})\sqrt{1 - \frac{n^3}{5\tau^2}} \leq \mathrm{D}(F^{\tau}_{mn}) \leq \mathrm{D}(Fu_{mn})\sqrt{1 + \frac{n^3}{36\tau^2}}.$$

Beweis: Es sei $\bar{k} = \tau + k$ und $\bar{j} = \tau + j$ mit $\tau > n$ und $\frac{n+1}{2} \leq m \leq n$. Dann ist

$$\mathrm{D}F^{\tau}_{mn} = \frac{1}{m\,(n-m+1)}\mathrm{D}\left(\sum_{k=1}^{n-m+1}\left(\chi_{\bar{k}+m} - \chi_{\bar{k}}\right)\right).$$

Unter Benutzung von

$$1 \leq \mathrm{ggT}(p+s, q+st) \leq \min\{p+s, q+s, |p-q|\}$$

und den groben Abschätzungen der Summen in (3.36) erhalten wir die Behauptungen.
Im Fall $\frac{n+1}{2} \leq m \leq n$ folgt aus (3.36) folgende obere Abschätzung

$$\begin{aligned}
m\,(n-m+1)\,\mathrm{VAR}\,(F^{\tau}_{mn}) &= m(n-m+1)\mathrm{VAR}\left(\sum_{k=1}^{n-m+1}\left(\chi_{\bar{k}+m} - \chi_{\bar{k}}\right)\right) \\
&\leq \frac{1}{6}\left(\sum_{k=1}^{n-m+1}\left(1 - \frac{1}{(k+m+\tau)\,(k+\tau)}\right)\right. \\
&\quad + \sum_{k=1}^{n-m}\sum_{j=k+1}^{n-m+1}\left(\frac{(k-j)^2}{(k+\tau+m)\,(j+m+\tau)} + \frac{(k-j)^2}{(k+\tau)\,(j+\tau)}\right. \\
&\quad \left.\left. - \frac{1}{(k+m+\tau)\,(j+\tau)} - \frac{1}{(j+m+\tau)\,(k+\tau)}\right)\right) \\
&= \frac{A+B}{6}.
\end{aligned}$$

Zunächst wollen wir A für $\tau > n$ abschätzen. Dabei ist

$$\begin{aligned}
A &= n - m + 1 - \sum_{k=1}^{n-m+1}\frac{1}{(k+m+\tau)\,(k+\tau)} \\
&\leq n - m + 1 - \int_{k=1}^{n-m+2}\frac{1}{(k+m+\tau)\,(k+\tau)}\mathrm{d}k \\
&= n - m + 1 + \frac{1}{m}\ln\left(1 - \frac{(n+1-m)\,m}{(n-m+2+\tau)\,(1+m+\tau)}\right)
\end{aligned}$$

$$\begin{aligned}
&\leq \quad (n-m+1)\left(1-\frac{1}{(n-m+2+\tau)(1+m+\tau)}\right) \\
&\leq \quad (n-m+1)\left(1-\frac{1}{4\tau^2}\right). \qquad (3.37)
\end{aligned}$$

Nun schätzen wir B für $\tau > n \geq m \geq \frac{n+1}{2}, j = k+1, ..., n-m+1, k = 1, ..., n-m$ ab. Es gilt

$$\frac{(k-j)^2}{(k+\tau+m)(j+m+\tau)} \leq \frac{(k-j)^2}{(k+\tau)(j+\tau)} \leq \frac{(k-j)^2}{\tau^2}.$$

Mit

$$\begin{aligned}
j+m &\leq n+1, \\
k+m &\leq n
\end{aligned}$$

haben wir

$$\begin{aligned}
\frac{1}{(j+m+\tau)(k+\tau)} &\geq \frac{1}{(n+1+\tau)(n-m+\tau)} \geq \frac{1}{4\tau^2}, \\
\frac{1}{(k+m+\tau)(j+\tau)} &\geq \frac{1}{(n+\tau)(n-m+1+\tau)} \geq \frac{1}{4\tau^2}.
\end{aligned}$$

Schließlich erhalten wir für B

$$\begin{aligned}
B \quad &\leq \quad \frac{1}{\tau^2}\sum_{k=1}^{n-m}\sum_{j=k+1}^{n-m+1}\left(2(k-j)^2-\frac{1}{2}\right) \\
&= \quad \frac{1}{\tau^2}\frac{1}{12}(n-m+1)(n-m)(2n(n+3)+(-2m(-m+2n+3))+1) \\
&\leq \quad \frac{1}{48\tau^2}(n-m+1)(n-1)\left(n^2+4n-3\right). \qquad (3.38)
\end{aligned}$$

Die obere Abschätzung der Summe von A und B lautet wegen (3.37) und (3.38) dann

$$A+B \leq (n-m+1)\left(1+\frac{1}{\tau^2}\left(-\frac{1}{4}+\frac{1}{48}(n-1)\left(n^2+4n-3\right)\right)\right).$$

Hieraus ergibt sich jetzt

$$\begin{aligned}
\mathrm{D}\left(F_{mn}^{\tau}\right) \quad &\leq \quad \frac{1}{m\sqrt{n-m+1}}\sqrt{\frac{1}{6}\left(1+\frac{1}{\tau^2}\left(-\frac{1}{4}+\frac{1}{48}(n-1)(n^2+4n-3)\right)\right)} \\
&\leq \quad \frac{1}{m\sqrt{n-m+1}}\sqrt{\frac{1}{6}\left(1+\frac{n^3}{36\tau^2}\right)}. \qquad (3.39)
\end{aligned}$$

Im Fall $\frac{n+1}{2} \leq m \leq n$ folgt aus (3.36) folgende untere Abschätzung

$$\begin{aligned}
&m(n-m+1)\operatorname{Var}\left(F_{mn}^{\tau}\right)^2 \\
&\geq \frac{1}{6}\Bigg(\sum_{k=1}^{n-m+1}\left(1-\frac{m^2}{(k+m+\tau)(k+\tau)}\right) \\
&\qquad +\sum_{k=1}^{n-m}\sum_{j=k+1}^{n-m+1}\Bigg(\frac{1}{(k+\tau+m)(j+m+\tau)}+\frac{1}{(k+\tau)(j+\tau)} \\
&\qquad\qquad -\frac{(k+m-j)^2}{(k+m+\tau)(j+\tau)}-\frac{(j+m-k)^2}{(j+m+\tau)(k+\tau)}\Bigg)\Bigg) \\
&= \frac{C+D}{6}.
\end{aligned}$$

Zunächst wollen wir C für $\tau > n$ abschätzen. Wegen

$$\ln(1-x) \geq 1 - \frac{1}{1-x} \quad \text{für} \quad 0 \leq x < 1$$

gilt dabei

$$\begin{aligned}
C &= n-m+1-m^2 \sum_{k=1}^{n-m+1} \frac{1}{(k+m+\tau)(k+\tau)} \\
&\geq n-m+1-\int\limits_{k=0}^{n-m+1} \frac{1}{(k+m+\tau)(k+\tau)} \mathrm{d}k \\
&= n-m+1+\frac{1}{m}\ln\left(1-\frac{(n+1-m)\,m}{(n-m+1+\tau)(m+\tau)}\right) \\
&\geq (n-m+1)+\frac{1}{m}\left(-(n-m+1)\frac{m}{\tau(n+1+\tau)}\right) \\
&\geq (n-m+1)\left(1-\frac{1}{2\tau^2}\right). \qquad (3.40)
\end{aligned}$$

Nun schätzen wir D für $\tau > n \geq m \geq \frac{n+1}{2}, j = k+1, ..., n-m+1, k = 1, ..., n-m$ ab. Es gilt

$$\begin{aligned}
\frac{(k+m-j)^2}{(k+\tau+m)(j+\tau)} &\leq \frac{(k+m-j)^2}{\tau^2} \\
\frac{(j+m-k)^2}{(j+m+\tau)(k+\tau)} &\leq \frac{(j+m-k)^2}{\tau^2}.
\end{aligned}$$

Mit

$$\begin{aligned}
j+m &\leq n+1, \\
k+m &\leq n
\end{aligned}$$

haben wir

$$\frac{1}{(j+m+\tau)(k+m+\tau)} \geq \frac{1}{(n+1+\tau)(n+\tau)} \geq \frac{1}{4\tau^2},$$

$$\frac{1}{(k+\tau)(j+\tau)} \geq \frac{1}{(n-m+\tau)(n-m+1+\tau)} \geq \frac{1}{4\tau^2}.$$

Schließlich erhalten wir für D

$$\begin{aligned}
D &\geq \frac{1}{\tau^2}\sum_{k=1}^{n-m}\sum_{j=k+1}^{n-m+1}\left(-(k+m-j)^2-(j+m-k)^2+\frac{1}{2}\right) \\
&= -\frac{1}{\tau^2}\frac{1}{12}(n-m+1)(n-m)\left(2n^2+6n-4nm-6m+1+14m^2\right) \\
&\geq -\frac{1}{48\tau^2}(n-m+1)(n-1)\left(7n^2+16n+3\right). \qquad (3.41)
\end{aligned}$$

Die untere Abschätzung der Summe von C und D lautet wegen (3.40) und (3.41) dann

$$\begin{aligned}
C+D &\geq (n-m+1)\left(1-\frac{1}{\tau^2}\left(\frac{1}{2}+\frac{1}{48}(n-1)\left(7n^2+16n+3\right)\right)\right) \\
&\geq (n-m+1)\left(1-\frac{1}{\tau^2}\left(\frac{1}{2}+\frac{n^3}{5}\right)\right).
\end{aligned}$$

Hieraus ergibt sich jetzt für $n \geq 3$

$$\mathrm{D}\left(F_{mn}^{\tau}\right) \geq \frac{1}{m\sqrt{n-m+1}}\sqrt{\frac{1}{6}\left(1-\frac{n^3}{5\tau^2}\right)}. \tag{3.42}$$

Aus Lemma 3.2.2, aus (3.39) und (3.42) folgen die beiden Ungleichungen. Damit ist alles bewiesen. □

Bemerkung 3.3.14 Die Abschätzungen im Beweis des Satzes 3.3.13 sind in folgender Weise verbesserungswürdig. Es sei $\tau > n$ und $1 \leq m \leq n$. Dann existieren Konstanten $C_1, C_2 \in (0,1)$ mit

$$\mathrm{D}(Fu_{mn})\sqrt{1-\frac{C_1 n^3}{\tau^2}} \leq \mathrm{D}(F_{mn}^{\tau}) \leq \mathrm{D}(Fu_{mn})\sqrt{1+\frac{C_2 n^3}{\tau^2}}.$$

Dabei gilt $C_1 \leq \frac{1}{5}$ für $n \geq 3$ und $C_2 \leq \frac{1}{36}$ für $n \geq 1$.

3.4 Fazit und Aussicht

Benutzen wir die explizite Lösung (2.18) für einen Σ-Δ-Konverter erster Ordnung unter der Annahme $|U_0(t)| \leq \alpha U_{ref}$, so konnten wir zeigen, daß die allgemein benutzte Annahme der stochastischen Unabhängigkeit des Quantisierungsfehlers in einem gewissen Sinn berechtigt ist. Wir haben gesehen, daß die Wahl der Wortlänge bei zweifacher gleitender Mittelung für $m \approx \frac{n+1}{3}$ oder $m \approx 2\frac{n+1}{3}$ im allgemeinen nicht die kleinste Streuung ergibt. Denn dieser Wert für m kann große Streuungen bei bestimmten Verteilungen der Eingangsfunktion U_0 ergeben. Aus diesem Grund ist die Wahl $m \approx \frac{n+1}{2}$ die geeignetere.
Σ-Δ-Konverter höherer Ordnung können in ähnlicher Weise untersucht werden, wenn wir Bedingungen angeben, unter denen die innerste Schleife wie ein Konverter erster Ordnung arbeitet. In diesem Fall ersetzt man die Heavisidefunktion aus der Differenzengleichung durch die oben betrachtete explizite Lösung für den letzten Integratorausgang (vgl. Kapitel 2.2.2). In diesem Fall erhalten wir lineare Differenzengleichungen, deren Lösungen die ursprünliche nach oben und unten beschränken. Eine andere Möglichkeit ist die in Kapitel 4 aufgezeigte. Dort werden wir durch Herleitung von Differentialgleichungen Filtervorschläge angeben.

Die folgenden Vermutungen fassen noch einmal unsere obigen Betrachtungen zusammen. Diese rein algebraischen, analytischen und geometrischen Fragestellungen besitzen eine eigenständige Bedeutung unabhängig der Untersuchungen, die wir im Zusammenhang mit Σ-Δ-Konvertern geführt haben. Aus diesem Grund formulieren wir unsere experimentellen Beobachtungen als Vermutungen.

Vermutung 3.4.1 Die Abbildung $m \mapsto DF_{mn}$ ist symmetrisch bez. $m = \frac{n+1}{2}$.

Vermutung 3.4.2 Die Abbildung $m \mapsto DF_{mn}$ besitzt in einem verallgemeinerten Sinn konvexes („schwach grob konvexes") Verhalten für alle m mit $m \in \left\{1, 2, \ldots, \left\lfloor \frac{n+1}{2} \right\rfloor\right\}$ und für alle m mit $m \in \left\{\left\lceil \frac{n+1}{2} \right\rceil, \left\lceil \frac{n+1}{2} \right\rceil + 1, \ldots, n\right\}$.

Vermutung 3.4.3 Die Abbildung $m \mapsto DF_{mn}$ besitzt lokale Maxima in $m \in \left\{\left\lfloor \frac{n+1}{2} \right\rfloor, \left\lceil \frac{n+1}{2} \right\rceil\right\}$.

Vermutung 3.4.4 Es sei $m_n \geq \frac{n+1}{2}$ das Argument des globalen Minimums der Abbildung $m \mapsto DF_{mn}$. Dann gelten folgende Aussagen:
a) m_n ist eindeutig bestimmt.
b) Es ist $\frac{n+1}{2} \leq m_{n+1} < m_n$.
c) Es ist $\lim\limits_{n\to\infty} \left(m_{n+1} - \frac{n+1}{2}\right) = 0$.

4 Modellierung mittels Differentialgleichungen und Betrachtungen zur Filterung

In diesem Kapitel untersuchen wir das Grenzverhalten der Differenzengleichungen bei einer Folge von Σ-Δ-Konvertern, bei denen die Taktzeit gegen Null konvergiert. Da diese Grenzprozesse integraler Natur sind, gelangen wir somit zu schwachen Grenzwerten. Die Differenzengleichungen „konvergieren“ für $\Delta t \to 0$ schwach gegen Differentialgleichungen, d. h. also zu Differentialgleichungen im distributiven Sinn. Die durch diese Überlegungen gewonnenen Ergebnisse werden dann bei der Herleitung entsprechender Filtervorschläge genutzt. Die Darstellung orientiert sich stark an [36]. Darüber hinaus geben wir hinreichende Bedingungen an, so daß die innere Schleife eines Konverters zweiter Ordnung durch die explizite Lösung eines Konverters erster Ordnung beschrieben werden kann. Ein erster gößerer Beweis mit Hilfe einer induktiven Fixpunktbetrachtung [35] scheiterte leider an einer im nachhinein enteckten Lücke.

4.1 Herleitung der Differentialgleichungen

In diesem Abschnitt wollen wir die Frage beantworten, wie sich eine Folge von Σ-Δ-Konvertern verhält, bei der die Taktzeit gegen Null konvergiert. Hierzu benötigen wir jedoch einige funktionalanalytische Aussagen, die wir im folgenden Abschnitt bereitstellen wollen.

4.1.1 Einige funktionalanalytische Aussagen

Zunächst wollen wir einen Konvergenzbegriff angeben, der weniger stark als die Konvergenz der Norm nach ist, aber für unsere Zwecke ausreicht. Es handelt sich hierbei um die sogenannte schwache Konvergenz.

Definition 4.1.1 [5] *Es sei X ein Banachraum und X^* der zugehörige duale Banachraum.*
a.) Eine Folge $(x^k)_{k\in\mathbb{N}}$ von Elementen aus X heißt schwach konvergent, falls es ein $x^0 \in X$ gibt, für welches

$$\lim_{k\to\infty} x^*(x^k) = x^*(x^0) \quad \textit{für alle} \quad x^* \in X^*$$

erfüllt ist. Dieses $x^0 \in X$ heißt dann schwacher Grenzwert von $(x^k)_{k\in\mathbb{N}}$. Bezeichnung $x^k \xrightarrow{w} x^0$.
b.) Eine Folge $(x^k)_{k\in\mathbb{N}}$ aus X, für die $(x^*(x^k))_{k\in\mathbb{N}}$ eine skalare Cauchyfolge für jedes $x^* \in X^*$ ist, heißt schwache Cauchyfolge in X.
c.) Ein Banachraum X heißt schwach vollständig, falls jede schwache Cauchyfolge einen schwachen Grenzwert in X besitzt.

Die folgende Aussage gibt Auskunft darüber, welche Lebesgueräume L^p schwach vollständig sind.

Satz 4.1.2 [5] *Es sei $0 < T - t_0 < +\infty$. Bezeichne λ das Lebesguemaß auf der borelschen σ-Algebra $\mathcal{B}^1[t_0,T]$. Dann ist $L^p([t_0,T],\mathcal{B}^1[t_0,T],\lambda)$ mit $p \in [1,+\infty)$ schwach vollständig.*

Der nächste Satz charakterisiert die schwachen Cauchyfolgen in den Lebesgueräumen.

Satz 4.1.3 [5, E.23, p.342] *Es sei $p \in [1,+\infty)$, $0 < T - t_0 < +\infty$ und $(I^k)_{k\in\mathbb{N}}$ eine Folge aus $L^p[t_0,T]$. $(I^k)_{k\in\mathbb{N}}$ ist genau dann eine schwache Cauchyfolge, wenn gilt*

$$\sup_{k\in\mathbb{N}} \int_{t_0}^{T} |I^k(\tau)|^p \mathrm{d}\tau < +\infty \quad \textit{und} \quad \forall t \in [t_0,T] \quad \exists \lim_{k\to\infty} \int_{t_0}^{t} I^k(\tau)\mathrm{d}\tau.$$

Nun geben wir speziell eine Eigenschaft für Folgen des L^1 an, die äquivalent zur schwachen Konvergenz dieser Folgen ist.

Satz 4.1.4 [5, S.7, p.291] *Es sei $0 < T - t_0 < +\infty$. Eine Folge $(I^k)_{k\in\mathbb{N}}$ aus $L^1[t_0, T]$ konvergiert genau dann schwach gegen ein $I^0 \in L^1[t_0, T]$, wenn $(I^k)_{k\in\mathbb{N}}$ beschränkt ist und*

$$\forall E \in \mathcal{B}^1[t_0, T] : \int_E I^0(\tau)\mathrm{d}\tau = \lim_{k\to\infty} \int_E I^k(\tau)\mathrm{d}\tau$$

gilt.

Der folgende Satz besagt, daß schwache Konvergenz unter stetigen linearen Abbildungen erhalten bleibt.

Satz 4.1.5 [41] *Es seien X und Y zwei Banachräume und $F : X \to Y$ sei eine lineare und stetige Abbildung. Weiterhin sei $(x^k)_{k\in\mathbb{N}}$ eine Folge aus X und $x^0 \in X$. Gilt $x^k \overset{w}{\to} x^0$ so auch $Fx^k \overset{w}{\to} Fx^0$.*

Nun zu einem der Fundamentalsätze der linearen Funktionalanalysis und einer wichtigen Folgerung daraus.

Satz 4.1.6 [70],[47] *(open mapping theorem (S. Banach)) X, Y seien zwei Banachräume, und sei S eine stetige lineare Abbildung von X auf Y. Dann bildet S jede offene Teilmenge von X auf eine offene Teilmenge von Y ab.*

Folgerung 4.1.7 [70],[47] (inverse mapping theorem (S. Banach)) X, Y seien zwei Banachräume, und sei S eine stetige lineare Abbildung von X auf Y. Ist S darüber hinaus injektiv, dann ist auch die inverse Abbildung S^{-1} ein linearer stetiger Operator.

Den kleinen Vorbereitungskurs schließt ein Resultat aus der Maßtheorie ab.

Satz 4.1.8 [11, S.4.4, p.140] *Es sei $(\Omega, \mathcal{A}, \mu)$ ein nichtleerer Maßraum und $\overline{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$. Sind $f, g : \Omega \to \overline{\mathbb{R}}$ integrierbar mit*

$$\forall E \in \mathcal{A} : \int_E f\mathrm{d}\mu \leq \int_E g\mathrm{d}\mu,$$

so ist $f \leq g$ μ-f.ü. Gilt insbesondere das Gleichheitszeichen in obigem Ausdruck, so ist $f = g$ μ-f.ü.

4.1.2 Differentialgleichung für Konverter erster Ordnung

Wir betrachten nun das Modell eines Σ-Δ-Konverters erster Ordnung, wie es in Kapitel 2.1 dargestellt wurde. In allen folgenden Überlegungen seien die Systemparameter $\alpha, T_1 > 0$ sowie das Eingangssignal $U_0(t)$ für $t \in [t_0, T]$ mit $T - t_0 \in (0, +\infty)$ fest gegeben. Wir untersuchen das Verhalten einer Folge von Konvertern, die sich durch die Taktzeit unterscheiden. Dabei sei $(\Delta t_k)_{k\in\mathbb{N}}$ eine positive monoton fallende Nullfolge, wobei $k \in \mathbb{N}$ als Indizierung für die zugehörigen Funktionen U_1^k sowie X^k diene. Im weiteren verstehen wir unter der Meßbarkeit einer Funktion ihre Lebesgue-Meßbarkeit.
Wir wollen nun die Frage beantworten, wie sich der Grenzwert von X^k für $k \to +\infty$ verhält. Zunächst aber ein vorbereitendes Lemma.

Lemma 4.1.9 *Es sei U_0 eine meßbare Funktion mit $|U_0(t)| \leq \alpha U_{ref}$ für alle Zeiten $t \in [t_0, T]$ und $U_1^k(t_0) = U_1(t_0) = 0$ für alle $k \in \mathbb{N}$. Dann gelten folgende Aussagen.*

a) $|U_1^k(t + \Delta t_k) - U_1^k(t)| \leq \dfrac{2\alpha \Delta t_k U_{ref}}{T_1}$ *für alle* $t \in [t_0, T - \Delta t_k], k \in \mathbb{N}$;

b) $|U_1^k(t)| \leq \dfrac{2\alpha \Delta t_k U_{ref}}{T_1}$ *für alle* $t \in [t_0, T], k \in \mathbb{N}$;

c) $X^k \in L^p[t_0, T]$ *für alle* $k \in \mathbb{N}$, $p \in [1, +\infty]$;

d) $U_1^k \to 0$ *für* $k \to +\infty$ *in* $C^0[t_0, T]$;

e) $\left| \int\limits_{t_0}^{t} \dfrac{U_0(\tau)}{\alpha U_{ref}} \mathrm{d}\tau - \int\limits_{t_0}^{t} X^k(\tau) \mathrm{d}\tau \right| \leq 2\Delta t_k$ *für alle* $k \in \mathbb{N}$ *und alle* $t \in [t_0, T]$.

Beweis: Wegen $U_1(t_0) = 0$ und (2.2) gilt

$$U_1^k(t) = \frac{1}{T_1} \int\limits_{t_0}^{t} (U_0(\tau) - \alpha U_{ref} X^k(\tau)) \mathrm{d}\tau. \tag{4.1}$$

a) Es sei $k \in \mathbb{N}$ und $t \in [t_0, T - \Delta t_k]$. Aus (4.1) und (2.4) folgt sofort

$$\begin{aligned}
\left| U_1^k(t + \Delta t_k) - U_1^k(t) \right| &= \left| \frac{1}{T_1} \int\limits_{t}^{t+\Delta t_k} (U_0(\tau) - \alpha U_{ref} X^k(\tau)) \mathrm{d}\tau \right| \\
&\leq \frac{1}{T_1} \left[\int\limits_{t}^{t+\Delta t_k} |U_0(\tau)| \mathrm{d}\tau + \alpha U_{ref} \int\limits_{t}^{t+\Delta t_k} |X^k(\tau)| \mathrm{d}\tau \right] \\
&\leq \frac{2\alpha U_{ref}}{T_1} \int\limits_{t}^{t+\Delta t_k} \mathrm{d}\tau \\
&= \frac{2\alpha \Delta t_k U_{ref}}{T_1}.
\end{aligned}$$

b) Der Beweis erfolgt induktiv. Wir zeigen hier nur den Induktionsschritt. Es sei $n \in \mathbb{N}$, $t \in [t_n, t_{n+1})$ und $|U_1^k(t_n)| \leq \dfrac{2\alpha \Delta t_k U_{ref}}{T_1}$. O. B. d. A. sei $U_1^k(t_n) \geq 0$. Nach (2.3) und (2.4) ist dann $X^k(\tau) = 1$ für $\tau \in [t_n, t_{n+1})$. Nach (4.1) gilt dann

$$U_1^k(t) = \frac{1}{T_1} \int\limits_{t_0}^{t} (U_0(\tau) - \alpha U_{ref} X^k(\tau)) \mathrm{d}\tau = \frac{1}{T_1} \int\limits_{t_n}^{t} (U_0(\tau) - \alpha U_{ref}) \mathrm{d}\tau + U_1^k(t_n). \tag{4.2}$$

Wegen $U_0(\tau) - \alpha U_{ref} \in [-2\alpha U_{ref}, 0]$ und der Induktionsvoraussetzung gilt somit

$$\frac{1}{T_1} \int\limits_{t_n}^{t} (U_0(\tau) - \alpha U_{ref}) \mathrm{d}\tau \in \left[-\frac{2\alpha U_{ref}}{T_1}(t - t_n), 0 \right] \;\text{ und }\; U_1^k(t_n) \in \left[0, \frac{2\alpha \Delta t_k U_{ref}}{T_1} \right]. \tag{4.3}$$

Aus (4.2) und (4.3) ergibt sich schließlich

$$U_1^k(t) \in \left[-\frac{2\alpha \Delta t_k U_{ref}}{T_1}, \frac{2\alpha \Delta t_k U_{ref}}{T_1} \right] \quad \forall t \in [t_n, t_{n+1}).$$

Der Fall $U_1^k(t_n) < 0$ wird analog bewiesen. Die Induktionsvoraussetzung ist beim Induktionsanfang durch $U_1(t_0) = 0$ bereits gegeben.

c) Sei $k \in \mathbb{N}$ beliebig aber fest. Nach (2.4) und (2.5) ist X^k auf $[t_0, T]$ wohldefiniert und nimmt nur die Werte -1 und $+1$ an. Also ist (vgl. [5, D.11, p.100])

$$||X^k||_{L^\infty[t_0,T]} = \operatorname*{ess\ sup}_{t\in[t_0,T]} |X^k(t)| = \sup_{t\in[t_0,T]} |X^k(t)| = 1.$$

Somit ist für jedes $k \in \mathbb{N}$

$$X^k \in L^\infty[t_0, T].$$

Hieraus folgt wegen $T - t_0 < +\infty$ für jedes $k \in \mathbb{N}$ unmittelbar

$$X^k \in L^p[t_0, T]$$

für alle $p \in [1, +\infty)$ (vgl. [27]).

d) Diese Aussage folgt unmittelbar aus der von b), da

$$||U_1^m|| := \sup_{t\in[t_0,T]} |U_1^m(t)| \leq \frac{2\alpha\Delta t_m U_{ref}}{T_1}$$

mit $\lim\limits_{m\to+\infty} \Delta t_m = 0$ sofort

$$\lim_{m\to+\infty} U_1^m = 0$$

in der Norm der gleichmäßigen Konvergenz impliziert.

e) Die Behauptung ergibt sich unmittelbar aus (4.1) und b). □

Nun kommen wir zum Hauptresultat dieses Abschnitts, welches die oben gestellte Frage hinreichend beantwortet.

Satz 4.1.10 *Es sei U_0 eine meßbare Funktion mit $|U_0(t)| \leq \alpha U_{ref}$ für alle Zeiten $t \in [t_0, T]$. Weiterhin sei $U_1^k(t_0) = U_1(t_0) = 0$ für alle $k \in \mathbb{N}$. Dann gelten folgende Aussagen.*
a) Für jedes $p \in [1, +\infty)$ existiert ein $X^0 \in L^p[t_0, T]$ mit $X^k \xrightarrow{w} X^0$ in $L^p[t_0, T]$.
b) Für diesen schwachen Grenzwert X^0 gilt

$$X^0(t) = \frac{U_0(t)}{\alpha U_{ref}} \text{ f.ü. in } [t_0, T]. \tag{4.4}$$

Beweis: a) Wir zeigen, daß $(X^k)_{k\in\mathbb{N}}$ eine schwache Cauchyfolge in $L^p[t_0, T]$ ist. Aufgrund der schwachen Vollständigkeit dieser Lebesgueräume folgt dann nach Satz 4.1.2 die Existenz eines schwachen Grenzwertes $X^0 \in L^p[t_0, T]$. Das Kriterium, ob $(X^k)_{k\in\mathbb{N}}$ eine schwache Cauchyfolge in $L^p[t_0, T]$ ist, liefert Satz 4.1.3.

Sei nun $\varepsilon > 0$. Nach Teil e) von Lemma 4.1.9 existiert dann ein $n_0 \in \mathbb{N}$, so daß für alle natürlichen Zahlen $m, n \geq n_0$ und alle Zeiten $t \in [t_0, T]$ gilt

$$\begin{aligned}
\left|\int_{t_0}^{t} (X^m - X^n)(\tau)\mathrm{d}\tau\right| &= \left|\int_{t_0}^{t} X^m(\tau)\mathrm{d}\tau - \int_{t_0}^{t} X^n(\tau)\mathrm{d}\tau\right| \\
&= \left|\int_{t_0}^{t} \left(X^m(\tau) - \frac{U_0(\tau)}{\alpha U_{ref}}\right)\mathrm{d}\tau - \int_{t_0}^{t} \left(X^n(\tau) - \frac{U_0(\tau)}{\alpha U_{ref}}\right)\mathrm{d}\tau\right| \\
&\leq \left|\int_{t_0}^{t} \left(X^m(\tau) - \frac{U_0(\tau)}{\alpha U_{ref}}\right)\mathrm{d}\tau\right| + \left|\int_{t_0}^{t} \left(X^n(\tau) - \frac{U_0(\tau)}{\alpha U_{ref}}\right)\mathrm{d}\tau\right| \\
&\leq 4\Delta t_{n_0} \\
&< \varepsilon.
\end{aligned} \tag{4.5}$$

Aus der Vollständigkeit von $\mathbb{R}$ bzw. dem Cauchykriterium für Zahlenfolgen folgt mit (4.5) dann

$$\forall t \in [t_0, T] \quad \exists \lim_{k\to\infty} \int_{t_0}^{t} X^k(\tau)\mathrm{d}\tau. \tag{4.6}$$

Nach Teil c) aus Lemma 4.1.9 ist $(X^k)_{k\in\mathbb{N}}$ eine Folge aus $L^p[t_0, T]$ mit $p \in [1, +\infty]$. Insbesondere gilt für jedes $p \in [1, +\infty)$

$$\sup_{k\in\mathbb{N}} \int_{t_0}^{T} |X^k(\tau)|^p \mathrm{d}\tau = \sup_{k\in\mathbb{N}} \int_{t_0}^{T} 1^p \mathrm{d}\tau = T - t_0 < +\infty. \tag{4.7}$$

Wegen (4.6) und (4.7) sind die Voraussetzung des Satzes 4.1.3 erfüllt. Damit gilt die Aussage in a).
b) Wegen Satz 4.1.4 und der Tatsache, daß für alle $t \in [t_0, T]$ dann $[t_0, t] \in \mathcal{B}^1[t_0, T]$ ist, folgt aus a) schließlich

$$\forall t \in [t_0, T] : \int_{t_0}^{t} X^0(\tau)\mathrm{d}\tau = \lim_{k\to\infty} \int_{t_0}^{t} X^k(\tau)\mathrm{d}\tau. \tag{4.8}$$

Andererseits ist wegen $\lim\limits_{k\to+\infty} \Delta t_k = 0$

$$\forall t \in [t_0, T] : \lim_{k\to\infty} \left(\int_{t_0}^{t} \frac{U_0(\tau)}{\alpha U_{ref}} \mathrm{d}\tau \pm 2\Delta t_k \right) = \int_{t_0}^{t} \frac{U_0(\tau)}{\alpha U_{ref}} \mathrm{d}\tau. \tag{4.9}$$

Aus Teil e) von Lemma 4.1.9, (4.8) und (4.9) ergibt sich für alle $t \in [t_0, T]$

$$\begin{array}{ccccc}
\lim\limits_{k\to\infty} \left(\int_{t_0}^{t} \frac{U_0(\tau)}{\alpha U_{ref}} \mathrm{d}\tau - 2\Delta t_k \right) & \leq & \lim\limits_{k\to\infty} \int_{t_0}^{t} X^k(\tau)\mathrm{d}\tau & \leq & \lim\limits_{k\to\infty} \left(\int_{t_0}^{t} \frac{U_0(\tau)}{\alpha U_{ref}} \mathrm{d}\tau + 2\Delta t_k \right) \\
\| & & \| & & \| \\
\int_{t_0}^{t} \frac{U_0(\tau)}{\alpha U_{ref}} \mathrm{d}\tau & \leq & \int_{t_0}^{t} X^0(\tau)\mathrm{d}\tau & \leq & \int_{t_0}^{t} \frac{U_0(\tau)}{\alpha U_{ref}} \mathrm{d}\tau.
\end{array}$$

Also gilt

$$\forall t \in [t_0, T] : \int_{t_0}^{t} \frac{U_0(\tau)}{\alpha U_{ref}} \mathrm{d}\tau = \int_{t_0}^{t} X^0(\tau)\mathrm{d}\tau.$$

Dies ist äqivalent zur Behauptung in b). □

Bemerkung 4.1.11 Nach Satz 4.1.10 besitzt die Folge $(X^k)_{k\in\mathbb{N}}$ einen schwachen Grenzwert $X^0 \in L^p[t_0, T]$, welcher sich nach Aussage b) fast überall gleich verhält wie die Eingangsfunktion U_0, sofern diese durch αU_{ref} dividiert wird. Bei Meßprozessen in der makroskopischen Natur, die sich ja stetig verhält, ist es daher sinnvoll, zusätzlich anzunehmen, daß U_0 stetig ist. In diesem Fall existiert dann ein stetiger Repräsentant $\varphi \in C^0[t_0, T]$ des schwachen Grenzwertes X^0, für den

$$\forall t \in [t_0, T] : \varphi(t) = \frac{U_0(t)}{\alpha U_{ref}} \tag{4.10}$$

gilt. Wir bezeichnen (4.10) nicht ohne Grund als eine Differentialgleichung nullter Ordnung. Diese beschreibt ein analoges Verhalten eines nichtlinearen Systems, wenn die Taktzeit verschwindet. Bemerkenswert dabei ist, daß das System bei endlicher Taktzeit durch eine Differenzengleichung erster Ordnung beschrieben wird (vgl. (2.10)) und bei verschwindender Taktzeit geht eine Ordnung bei der Differentialgleichung „verloren".

4.1.3 Differentialgleichung für Konverter zweiter Ordnung

Wir betrachten nun das Modell eines Σ-Δ-Konverters zweiter Ordnung, wie es in Kapitel 2.2 dargestellt wurde. In allen folgenden Überlegungen seien die Systemparameter $\alpha, \beta, T_1, T_2 > 0$ sowie das Eingangssignal $U_0(t)$ für $t \in [t_0, T]$ mit $T - t_0 < +\infty$ fest gegeben. Wir untersuchen das Verhalten einer Folge von Konvertern, die sich durch die Taktzeit unterscheiden. Dabei sei $(\Delta t_k)_{k\in\mathbb{N}}$ eine positive monoton fallende Nullfolge, wobei $k \in \mathbb{N}$ als Indizierung für die Funktionen U_1^k, U_2^k sowie X^k diene. Zusätzlich nehmen wir an, daß der Anfangszustand aller Konverter gleich sei, d. h., es ist $U_1^k(t_0) = U_1(t_0)$ und $U_2^k(t_0) = U_2(t_0)$.
Wir wollen nun die Frage beantworten, wie sich der Grenzwert von X^k für $k \to +\infty$ verhält. Zunächst aber ein vorbereitendes Lemma.

Lemma 4.1.12 *Es sei $U_2(t_0) = 0$ und U_0 eine meßbare Funktion. Weiterhin sei $|U_1^k(t)| \le \beta U_{ref}$ für alle Zeiten $t \in [t_0, T]$ und für alle $k \in \mathbb{N}$. Dann gelten folgende Aussagen.*
a) $|U_2^k(t + \Delta t_k) - U_2^k(t)| \le \dfrac{2\beta\Delta t_k U_{ref}}{T_2}$ *für alle $t \in [t_0, T - \Delta t_k]$, $k \in \mathbb{N}$;*
b) $|U_2^k(t)| \le \dfrac{2\beta\Delta t_k U_{ref}}{T_2}$ *für alle $t \in [t_0, T]$, $k \in \mathbb{N}$;*
c) $$\left| \int_{t_0}^{t}\int_{t_0}^{\tau} \frac{U_0(r)}{U_{ref}} \mathrm{d}r\mathrm{d}\tau - \int_{t_0}^{t} \left[\int_{t_0}^{\tau} \alpha X^k(r)\mathrm{d}r + T_1\beta X^k(\tau) - \frac{U_1(t_0)}{U_{ref}} \right] \mathrm{d}\tau \right| \le 2\beta T_1 \Delta t_k,$$
für alle $k \in \mathbb{N}$ und alle $t \in [t_0, T]$;
d) Die Funktion $I^k(\tau) := \int_{t_0}^{\tau} \alpha X^k(r)\mathrm{d}r + T_1\beta X^k(\tau)$, $\tau \in [t_0, T]$, *ist Element von $L^p[t_0, T]$ für jedes $k \in \mathbb{N}$ und jedes $p \in [1, +\infty]$;*
e) $U_2^k \to 0$ *in* $C^0[t_0, T]$ *;*
f) Für jedes $p \in [1, +\infty)$ existiert ein $I^0 \in L^p[t_0, T]$ mit $I^k \overset{w}{\to} I^0$ in $L^p[t_0, T]$;
g) Für diesen schwachen Grenzwert gilt

$$I^0(\tau) = \int_{t_0}^{\tau} \frac{U_0(r)}{U_{ref}} \mathrm{d}r + \frac{U_1(t_0)}{U_{ref}} \quad \textit{f.ü. in } [t_0, T].$$

Beweis: Wegen $U_2(t_0) = 0$, (2.30) und (2.31) gilt

$$U_1^k(t) = \frac{1}{T_1} \int_{t_0}^{t} \Big(U_0(\tau) - \alpha U_{ref} X^k(\tau) \Big) \mathrm{d}\tau + U_1(t_0), \tag{4.11}$$

$$U_2^k(t) = \frac{1}{T_2} \int_{t_0}^{t} \Big(U_1^k(\tau) - \beta U_{ref} X^k(\tau) \Big) \mathrm{d}\tau. \tag{4.12}$$

a) und b) lassen sich analog der Aussagen a) und b) des Lemmas 4.1.9 beweisen, indem man die Indizierung und α mit β tauscht und Formel (4.12) benutzt.

c) Setzt man (4.11) in (4.12) ein, so ergibt sich unter Berücksichtigung von b) nach Multiplikation mit $\dfrac{T_1T_2}{U_{ref}}$ die Behauptung.
d) Nach (2.32) und (2.33) ist X^k auf $[t_0,T]$ wohldefiniert, beschränkt und als Treppenfunktion auch meßbar. Nach Definition ist dann auch I^k eine meßbare Funktion auf $[t_0,T]$. Deshalb gilt

$$\begin{aligned}||I^k||_{L^\infty[t_0,T]} &= \operatorname*{ess\,sup}_{\tau\in[t_0,T]} |I^k(\tau)| = \sup_{\tau\in[t_0,T]} \left|\int_{t_0}^{\tau} \alpha X^k(r)\mathrm{d}r + T_1\beta X^k(\tau)\right| \\ &\le \sup_{\tau\in[t_0,T]} \alpha \int_{t_0}^{\tau} |X^k(r)|\mathrm{d}r + |T_1\beta X^k(\tau)| \\ &\le \alpha(T-t_0) + T_1\beta < +\infty.\end{aligned}$$

Damit gilt für $k \in \mathbb{N}$

$$I^k \in L^\infty[t_0,T]. \tag{4.13}$$

Sei nun $p \in [1,+\infty)$. Dann gilt mit Hilfe der Minkowskiungleichung

$$\begin{aligned}||I^k||_{L^p[t_0,T]} &= \left(\int_{t_0}^{T}\left|\int_{t_0}^{\tau} \alpha X^k(r)\mathrm{d}r + T_1\beta X^k(\tau)\right|^p \mathrm{d}\tau\right)^{\frac{1}{p}} \\ &\le \left(\int_{t_0}^{T}\left|\int_{t_0}^{T} \alpha X^k(r)\mathrm{d}r\right|^p \mathrm{d}\tau\right)^{\frac{1}{p}} + \left(\int_{t_0}^{T} |T_1\beta X^k(\tau)|^p\mathrm{d}\tau\right)^{\frac{1}{p}} \\ &\le \alpha\left(\int_{t_0}^{T}\left(\int_{t_0}^{T} |X^k(r)|\mathrm{d}r\right)^p \mathrm{d}\tau\right)^{\frac{1}{p}} + T_1\beta\left(\int_{t_0}^{T} |X^k(\tau)|^p\mathrm{d}\tau\right)^{\frac{1}{p}} \\ &= \alpha\left(\int_{t_0}^{T}\left(\int_{t_0}^{T} 1\mathrm{d}r\right)^p \mathrm{d}\tau\right)^{\frac{1}{p}} + T_1\beta\left(\int_{t_0}^{T} 1^p\mathrm{d}\tau\right)^{\frac{1}{p}} \\ &= \alpha(T-t_0)^{\frac{p+1}{p}} + T_1\beta(T-t_0)^{\frac{1}{p}} < +\infty.\end{aligned} \tag{4.14}$$

Damit gilt für $k \in \mathbb{N}$ und $p \in [1,+\infty)$

$$I^k \in L^p[t_0,T]. \tag{4.15}$$

Aus (4.13) und (4.15) folgt dann die Aussage in d).
e) Der Beweis erfolgt analog zu dem der Aussage d) des Lemmas 4.1.9, wenn man hier U_1^k durch U_2^k, T_1 durch T_2 und α durch β ersetzt.
f) Mit den gleichen Argumenten wie beim Beweis der Aussage a) des Satzes 4.1.10 zeigen wir f). Sei nun $p \in [1,+\infty)$ und $\varepsilon > 0$. Dann existiert nach c) ein $n_0 \in \mathbb{N}$, so daß für alle natürlichen Zahlen $m,n \ge n_0$ und für alle Zeiten $t \in [t_0,T]$ gilt

$$\left|\int_{t_0}^{t} (I^m - I^n)(\tau)\mathrm{d}\tau\right| = \left|\int_{t_0}^{t} I^m(\tau)\mathrm{d}\tau - \int_{t_0}^{t}\int_{t_0}^{\tau} \frac{U_0(r)}{\alpha U_{ref}}\mathrm{d}r\mathrm{d}\tau - \int_{t_0}^{t} \frac{U_1(t_0)}{U_{ref}}\mathrm{d}\tau\right.$$

$$
\begin{aligned}
&\quad - \left[\int_{t_0}^{t} I^n(\tau)\mathrm{d}\tau - \int_{t_0}^{t}\int_{t_0}^{\tau} \frac{U_0(r)}{\alpha U_{ref}}\mathrm{d}r\mathrm{d}\tau - \int_{t_0}^{t} \frac{U_1(t_0)}{U_{ref}}\mathrm{d}\tau \right] \Bigg| \\
&\leq \left| \int_{t_0}^{t} I^m(\tau)\mathrm{d}\tau - \int_{t_0}^{t}\int_{t_0}^{\tau} \frac{U_0(r)}{\alpha U_{ref}}\mathrm{d}r\mathrm{d}\tau - \frac{(t-t_0)U_1(t_0)}{U_{ref}} \right| \\
&\quad + \left| \int_{t_0}^{t} I^n(\tau)\mathrm{d}\tau - \int_{t_0}^{t}\int_{t_0}^{\tau} \frac{U_0(r)}{\alpha U_{ref}}\mathrm{d}r\mathrm{d}\tau - \frac{(t-t_0)U_1(t_0)}{U_{ref}} \right| \\
&\leq 2\beta T_1(\Delta t_m + \Delta t_n) \\
&\leq 4\beta T_1 \Delta t_{n_0} \\
&\leq \varepsilon. \qquad (4.16)
\end{aligned}
$$

Aus der Vollständigkeit von $\mathbb{R}$ bzw. dem Cauchykriterium für Zahlenfolgen folgt mit (4.16) dann

$$
\forall t \in [t_0, T] \quad \exists \lim_{k\to\infty} \int_{t_0}^{t} I^k(\tau)\mathrm{d}\tau. \qquad (4.17)
$$

Aus Formel (4.14) ergibt sich

$$
\begin{aligned}
\sup_{k\in\mathbb{N}} \int_{t_0}^{T} |I^k(\tau)|^p \mathrm{d}\tau &\leq \sup_{k\in\mathbb{N}} \left(\alpha(T-t_0)^{\frac{p+1}{p}} + T_1\beta(T-t_0)^{\frac{1}{p}} \right)^p \\
&= \left(\alpha(T-t_0)^{\frac{p+1}{p}} + T_1\beta(T-t_0)^{\frac{1}{p}} \right)^p < +\infty. \qquad (4.18)
\end{aligned}
$$

Da durch (4.17) und (4.18) die Bedingungen aus Satz 4.1.3 erfüllt sind, ist $(I^k)_{k\in\mathbb{N}}$ eine schwache Cauchyfolge in $L^p[t_0,T]$ für $p \in [1,+\infty)$. Da diese Lebesgueräume nach Satz 4.1.2 schwach vollständig sind, existiert dann ein $I^0 \in L^p[t_0,T]$, für welches $I^k \stackrel{w}{\to} I^0$ in $L^p[t_0,T]$ gilt.
g) Wegen Satz 4.1.4 und der Tatsache, daß für alle $t \in [t_0,T]$ dann $[t_0,t] \in \mathcal{B}^1[t_0,T]$ ist, folgt aus f) schließlich

$$
\forall t \in [t_0,T] : \int_{t_0}^{t} I^0(\tau)\mathrm{d}\tau = \lim_{k\to\infty} \int_{t_0}^{t} I^k(\tau)\mathrm{d}\tau. \qquad (4.19)
$$

Andererseits ist nach Voraussetzung an $(\Delta t_k)_{k\in\mathbb{N}}$ für alle $t \in [t_0,T]$ auch

$$
\lim_{k\to\infty} \int_{t_0}^{t}\int_{t_0}^{\tau} \frac{U_0(r)}{\alpha U_{ref}}\mathrm{d}r\mathrm{d}\tau + \int_{t_0}^{t} \frac{U_1(t_0)}{U_{ref}}\mathrm{d}\tau \pm 2\beta T_1 \Delta t_k = \int_{t_0}^{t}\int_{t_0}^{\tau} \frac{U_0(r)}{\alpha U_{ref}}\mathrm{d}r\mathrm{d}\tau + \int_{t_0}^{t} \frac{U_1(t_0)}{U_{ref}}\mathrm{d}\tau. \qquad (4.20)
$$

Aus c), (4.19) und (4.20) ergibt sich für alle $t \in [t_0,T]$ dann

$$
\begin{array}{ccc}
\lim\limits_{k\to\infty} \displaystyle\int_{t_0}^{t}\int_{t_0}^{\tau} \frac{U_0(r)}{\alpha U_{ref}}\mathrm{d}r\mathrm{d}\tau + \int_{t_0}^{t} \frac{U_1(t_0)}{U_{ref}}\mathrm{d}\tau - 2\beta T_1 \Delta t_k & \leq & \lim\limits_{k\to\infty} \displaystyle\int_{t_0}^{t} I^k(\tau)\mathrm{d}\tau \\
\| & & \| \\
\displaystyle\int_{t_0}^{t}\int_{t_0}^{\tau} \frac{U_0(r)}{\alpha U_{ref}}\mathrm{d}r\mathrm{d}\tau + \int_{t_0}^{t} \frac{U_1(t_0)}{U_{ref}}\mathrm{d}\tau & \leq & \displaystyle\int_{t_0}^{t} I^0(\tau)\mathrm{d}\tau
\end{array} \qquad (4.21)
$$

sowie

$$
\begin{array}{ccc}
\lim\limits_{k\to\infty}\int\limits_{t_0}^{t} I^k(\tau)\mathrm{d}\tau & \leq & \lim\limits_{k\to\infty}\int\limits_{t_0}^{t}\int\limits_{t_0}^{\tau}\frac{U_0(r)}{\alpha U_{ref}}\mathrm{d}r\mathrm{d}\tau + \int\limits_{t_0}^{t}\frac{U_1(t_0)}{U_{ref}}\mathrm{d}\tau + 2\beta T_1\Delta t_k \\
\| & & \| \\
\int\limits_{t_0}^{t} I^0(\tau)\mathrm{d}\tau & \leq & \int\limits_{t_0}^{t}\int\limits_{t_0}^{\tau}\frac{U_0(r)}{\alpha U_{ref}}\mathrm{d}r\mathrm{d}\tau + \int\limits_{t_0}^{t}\frac{U_1(t_0)}{U_{ref}}\mathrm{d}\tau.
\end{array}
\tag{4.22}
$$

Also gilt wegen (4.21) und (4.22)

$$
\forall t\in[t_0,T]: \int\limits_{t_0}^{t}\int\limits_{t_0}^{\tau}\frac{U_0(r)}{\alpha U_{ref}}\mathrm{d}r\mathrm{d}\tau + \int\limits_{t_0}^{t}\frac{U_1(t_0)}{U_{ref}}\mathrm{d}\tau = \int\limits_{t_0}^{t} I^0(\tau)\mathrm{d}\tau.
$$

Dies liefert wegen Satz 4.1.8 die Behauptung in g). □

Nun wissen wir, wie sich $(I^k)_{k\in\mathbb{N}}$ bei immer kleiner werdender Taktzeit verhält. Interessanter dagegen ist aber das Verhalten von $(X^k)_{k\in\mathbb{N}}$. Um dieses zu untersuchen, benötigen wir zunächst einen Integraloperator, wie er im folgenden Lemma definiert wird.

Lemma 4.1.13 [23] *Es sei $J=[a,b]\subset\mathbb{R}$ ein abgeschlossenes Intervall. Weiterhin sei $k(x,y,z)$ eine gegebene Funktion auf der Definitionsmenge $\{(x,y,z): x\in J, y\in[a,x], z\in\mathbb{R}\}\subset J\times J\times\mathbb{R}$. Es sei*

$$
G := \{(x,y): x\in J, y\in[a,x]\}\subset J\times J.
$$

Ferner sei $k\in C^0(G\times\mathbb{R})$ und erfülle folgende Lipschitzbedingung. Es gibt eine Funktion $L\in C^0(G)$ mit

$$
|k(x,y,z)-k(x,y,\tilde z)| \leq L(x,y)|z-\tilde z| \textit{ für alle } (x,y)\in G, z,\tilde z\in\mathbb{R}.
$$

Ist dann $g\in L^p(J)$ mit $p\in[1,+\infty]$, dann besitzt die Volterrasche Integralgleichung 2. Art

$$
f(x) = g(x) + \int\limits_a^x k(x,y,f(y))\mathrm{d}y \quad \textit{f.ü.,} \tag{4.23}
$$

mit $x\in J$ genau eine Lösung $f\in L^p(J)$.

Beweis: Den Beweis findet man in [23, Kap.2.1]. Dort wurde der Satz 2. 1. 1 unter der Annahme $g\in C^0(J)$ bewiesen. Der Teil (a) der dortigen Bemerkung 2. 1. 5 erweitert diese Annahme auf $g\in L^\infty(J)$. Der Nachweis für $p\in[1,+\infty)$ erfolgt mit den gleichen Argumenten. □

Lemma 4.1.14 *Es sei $p\in[1,+\infty)$, $\varphi\in L^p[t_0,T]$, und die Abbildung $S: L^p[t_0,T]\to L^p[t_0,T]$ sei definiert durch*

$$
(S\varphi)(t) := T_1\beta\varphi(t) + \alpha\int\limits_{t_0}^{t}\varphi(r)\mathrm{d}r =: \psi(t) \quad \textit{f.ü.} \tag{4.24}
$$

Dann ist S ein linearer Homöomorphismus auf $L^p[t_0,T]$, und es gilt

$$
(S^{-1}\psi)(t) = \frac{\psi(t)}{\beta T_1} - \frac{\alpha}{\beta^2T_1^2}\int\limits_{t_0}^{t}\psi(r)\mathrm{e}^{\frac{\alpha}{\beta T_1}(r-t)}\mathrm{d}r = \varphi(t) \quad \textit{f.ü.} \tag{4.25}
$$

Beweis: Wir wollen zunächst die Stetigkeit von S auf $L^p[t_0,T]$ zeigen. Sei nun $\varphi \in L^p[t_0,T]$. Dann gilt nach (4.24), den Ungleichungen von Minkowski und Hölder mit $q = \frac{p}{p-1}$

$$\begin{aligned}
||S\varphi||_{L^p} &= \left(\int_{t_0}^{T} \left|\beta T_1\varphi(t) + \alpha\int_{t_0}^{t}\varphi(r)\mathrm{d}r\right|^p \mathrm{d}t\right)^{\frac{1}{p}} \\
&\leq \left(\int_{t_0}^{T} |\beta T_1\varphi(t)|^p\mathrm{d}t\right)^{\frac{1}{p}} + \left(\int_{t_0}^{T} \left|\alpha\int_{t_0}^{t}\varphi(r)\mathrm{d}r\right|^p \mathrm{d}t\right)^{\frac{1}{p}} \\
&= \beta T_1\left(\int_{t_0}^{T} |\varphi(t)|^p\mathrm{d}t\right)^{\frac{1}{p}} + \alpha\left(\int_{t_0}^{T} \left|\int_{t_0}^{T}\varphi(r)\mathrm{d}r\right|^p \mathrm{d}t\right)^{\frac{1}{p}} \\
&\leq \beta T_1||\varphi||_{L^p} + \alpha\left(\int_{t_0}^{T}\left[\left(\int_{t_0}^{T} 1^q\mathrm{d}r\right)^{\frac{1}{q}}\left(\int_{t_0}^{T}|\varphi(r)|^p\mathrm{d}r\right)^{\frac{1}{p}}\right]^p \mathrm{d}t\right)^{\frac{1}{p}} \\
&= \beta T_1||\varphi||_{L^p} + \alpha\left(\int_{t_0}^{T}\left[\sqrt[q]{T-t_0}\cdot||\varphi||_{L^p}\right]^p \mathrm{d}t\right)^{\frac{1}{p}} \\
&= \beta T_1||\varphi||_{L^p} + \alpha\sqrt[q]{T-t_0}\,\sqrt[p]{T-t_0}||\varphi||_{L^p} \\
&= (\beta T_1 + \alpha(T-t_0))||\varphi||_{L^p} < +\infty. \qquad (4.26)
\end{aligned}$$

Also ist $S\varphi \in L^p[t_0,T]$ für $p \in [1,+\infty)$, und S ist wegen (4.26) stetig.
Nach Voraussetzung gilt

$$\varphi(t) = \frac{\psi(t)}{\beta T_1} - \frac{\alpha}{\beta T_1}\int_{t_0}^{t}\varphi(r)\mathrm{d}r \quad \text{f.ü.} \qquad (4.27)$$

Somit ist (4.27) eine lineare Volterrasche Integralgleichung 2. Art mit einem Kern, der identisch eins ist. Nach Lemma 4.1.13 ist (4.27) in $L^p[t_0,T]$ eindeutig lösbar. Also ist S eine bijektive Abbildung von $L^p[t_0,T]$ auf $L^p[t_0,T]$. Die Linearität von S selbst folgt unmittelbar aus ihrer Definition (4.24). Damit sind alle Voraussetzungen des inverse mapping theorem (vgl. Folgerung 4.1.7) erfüllt, und S ist ein linearer Homöomorphismus auf $L^p[t_0,T]$.
Wir wollen nun die Gleichung $S\varphi = \psi$ in $L^p[t_0,T]$ lösen. Wir wenden hier die Methode des lösenden Kerns an, eine zweite ist eine sukzessive Approximation (vgl. [14, Kap.5.3]). Die Lösung von (4.27) stellt sich dann wie folgt dar

$$\varphi(t) = -\frac{\alpha}{\beta T_1}\int_{t_0}^{t} R\left(t,r,-\frac{\alpha}{\beta T_1}\right)\cdot\frac{\psi(r)}{\beta T_1}\mathrm{d}r + \frac{\psi(t)}{\beta T_1} \quad \text{f.ü.} \qquad (4.28)$$

mit der absolut und gleichmäßig auf $[t_0,T]$ konvergenten Neumannschen Reihe - „lösender Kern" - (vgl. [14, Kap.5.3.1])

$$R\left(t,r,-\frac{\alpha}{\beta T_1}\right) = \sum_{m=1}^{\infty} K^{(m)}(t,r)\cdot\left(-\frac{\alpha}{\beta T_1}\right)^{m-1}$$

und mit den iterierten Kernen

$$
\begin{aligned}
K^{(1)}(t,r) &= 1,\\
K^{(2)}(t,r) &= \int_r^t 1\mathrm{d}t = t-r,\\
K^{(3)}(t,r) &= \int_r^t (t-r)\mathrm{d}t = \frac{(t-r)^2}{2},\\
&\vdots\\
K^{(m)}(t,r) &= \int_r^t \frac{(t-r)^{m-2}}{(m-2)!}\mathrm{d}t = \frac{(t-r)^{m-1}}{(m-1)!}.
\end{aligned}
$$

Somit gilt wegen (4.28)

$$
\begin{aligned}
\varphi(t) &= -\frac{\alpha}{\beta T_1}\int_{t_0}^t \sum_{m=1}^{\infty} \frac{\left((t-r)\left(-\frac{\alpha}{\beta T_1}\right)\right)^{m-1}}{(m-1)!}\cdot\frac{\psi(r)}{\beta T_1}\mathrm{d}r + \frac{\psi(t)}{\beta T_1} \qquad \text{f.ü.}\\
&= -\frac{\alpha}{\beta^2 T_1^2}\int_{t_0}^t \mathrm{e}^{\frac{\alpha}{\beta T_1}(r-t)}\psi(r)\mathrm{d}r + \frac{\psi(t)}{\beta T_1} \qquad \text{f.ü.}
\end{aligned}
$$

Damit ist alles bewiesen. □

Nun kommen wir zum Hauptresultat dieses Abschnitts, welches die oben gestellte Frage hinreichend beantwortet.

Satz 4.1.15 *Es sei $U_2(t_0) = 0$ und U_0 eine meßbare Funktion. Weiterhin sei $|U_1^k(t)| \leq \beta U_{ref}$ für alle Zeiten $t \in [t_0, T]$ und alle $k \in \mathbb{N}$. Es seien I^k sowie I^0 gemäß Lemma 4.1.12 gegeben, und es sei X^0 die nach Lemma 4.1.13 eindeutig bestimmte Lösung der Gleichung $I^0 = S(X)$. Dann gelten folgende Aussagen.*
a) Es ist $X^0 \in L^p[t_0,T]$ mit $X^k \overset{w}{\to} X^0$ in $L^p[t_0,T]$ für jedes $p \in [1,+\infty)$.
b) Ist zusätzlich $U_0 \in C^0[t_0,T]$, dann existiert eine eindeutig bestimmte Funktion $\varphi \in C^1[t_0,T]$ mit $\varphi = X^0$ f.ü. in $[t_0,T]$, welche dem Anfangswertproblem

$$\varphi'(t) + \frac{\alpha}{T_1\beta}\varphi(t) = \frac{U_0(t)}{T_1\beta U_{ref}}, \quad \varphi(t_0) = \frac{U_1(t_0)}{\beta U_{ref}} \tag{4.29}$$

genügt.

Beweis: a) Sei $p \in [1,+\infty)$. Nach Teil f) des Lemmas 4.1.12 gilt

$$I^k \overset{w}{\to} I^0 \text{ in } L^p[t_0,T]. \tag{4.30}$$

Nach Definition der I^k und I^0 sind die Funktionen X^k für $k \in \mathbb{N}_0$ gerade die Lösungen der Gleichungen $I^k = S(X)$. Nach Lemma 4.1.13 sind diese linearen Volterraschen Integralgleichungen eindeutig in $L^p[t_0,T]$ lösbar. Nach (4.30) ist dann

$$S(X^k) = I^k \overset{w}{\to} I^0 = S(X^0). \tag{4.31}$$

Da die Abbildung S^{-1} nach Lemma 4.1.14 auf $L^p[t_0, T]$ für $p \in [1, +\infty)$ stetig ist, folgt aus Satz 4.1.5 und (4.31) dann

$$X^k = S^{-1}(I^k) \overset{w}{\to} S^{-1}(I^0) = X^0.$$

b) Nach Definition von X^0 ist für $p \in [1, +\infty]$

$$I^0(\tau) = \int_{t_0}^{\tau} \alpha X^0(r)\mathrm{d}r + \beta T_1 X^0(\tau) \text{ in } L^p[t_0, T].$$

Hieraus und aus Teil g) von Lemma 4.1.12 ergibt sich nun

$$\beta T_1 X^0(\tau) = \int_{t_0}^{\tau} \left(\frac{U_0(r)}{U_{ref}} - \alpha X^0(r) \right) \mathrm{d}r + \frac{T_1 U_1(t_0)}{U_{ref}} \quad \text{f.ü.} \tag{4.32}$$

Nach Voraussetzung ist U_0 eine meßbare Funktion. Aus der Beschränktheit von U_1^k für alle $k \in \mathbb{N}$ und Formel (4.11) folgt die von U_0. Ansonsten entstünde ein Widerspruch. Somit ist wegen $X^0 \in L^p[t_0, T]$ für $p \in [1, +\infty)$ die linke Seite von (4.32) - ergo X^0 selbst - eine absolutstetige Funktion (vgl. [50, S.8, p.165]). Damit existiert ein $\varphi \in C^0[t_0, T]$ mit $\varphi = X^0$ fast überall. Also gilt

$$\beta T_1 \varphi(\tau) = \int_{t_0}^{\tau} \left(\frac{U_0(r)}{U_{ref}} - \alpha \varphi(r) \right) \mathrm{d}r + \frac{T_1 U_1(t_0)}{U_{ref}}. \tag{4.33}$$

Unter der Annahme, daß $U_0 \in C^0[t_0, T]$ ist, folgt aus dem Hauptsatz der Differential und Integralrechnung die Differenzierbarkeit der rechten Seite von (4.33). Es gilt dann $\varphi \in C^1[t_0, T]$ und

$$\beta T_1 \frac{d}{d\tau}\varphi(\tau) = \frac{U_0(\tau)}{U_{ref}} - \alpha\varphi(\tau) \quad \text{und} \quad \varphi(t_0) = \frac{U_1(t_0)}{\beta U_{ref}}.$$

Damit ist alles bewiesen. □

Bemerkung 4.1.16 Anknüpfend an Bemerkung 4.1.11 können wir nun feststellen, daß ein Σ-Δ-Konverter zweiter Ordnung mit einem Komparator bei endlicher Taktzeit durch eine Differenzengleichung zweiter Ordnung beschrieben wird. Dagegen ist bei verschwindender Taktzeit und der Einschränkung an zu messende Funktionen, stetig zu sein, eine Beschreibung durch eine lineare Differentialgleichung erster Art mit konstanten Koeffizienten möglich (vgl. (4.29)). Durch die Differentialgleichung kann jedoch nur ein qualitatives Verhalten des Konverters im großen widergegeben werden. Jedoch kein qualitatives im kleinen wie durch die Differenzengleichung. Σ-Δ-Konverter arbeiten immer mit endlicher Taktzeit. Wir sind jedoch nicht in der Lage, ein Kriterium anzugeben, wann die Differentialgleichung das Konverterverhalten hinreichend gut approximiert. Wann ist Δt dafür hinreichend klein? An das reale Verhalten eines Konverters mit nichtidealen - sprich nichtlinearen - Integrierern oder der Temperaturabhängigkeit von Bauelementen sei hier nur am Rande erinnert.

Bemerkung 4.1.17 Ohne weitere Konverter höherer oder gar kaskadierter Ordnung zu untersuchen, gelangen wir durch unsere Überlegungen via schwacher Grenzwerte in den Kapiteln 4.1.2 und 4.1.3 zu folgendem Fazit.

A Σ-Δ-Konverter n-ter Ordnung, gleich ob kaskadiert oder nicht, werden bei verschwindender Taktzeit durch eine Differentialgleichung $(n-1)$-ter Ordnung beschrieben.

B Im Gegensatz zu den nichtlinearen, nichtautonomen Differenzengleichungen bei endlicher Taktzeit sind die Differentialgleichungen linear und autonom.

C Erst bei Konverter dritter Ordnung kann entsprechend der Differentialgleichung zweiter Ordnung folgend Eigenschwingung auftreten. Dies steht aber im Widerspruch zum realen Verhalten, bei dem Eigenschwingung bereits bei Konverter zweiter Ordnung auftreten kann. Darauf werden wir im Kapitel 6.3 zurück kommen.

Bemerkung 4.1.18 Wie wir gesehen haben, hat diese Modellbildung Vor- und Nachteile. Auf der einen Seite erhalten wir Differentialgleichungen, die wir sehr gut untersuchen können. Wir werden gleich, ohne Rücksicht auf Existenz oder Eindeutigkeit zu nehmen, einen Algorithmus vorstellen, mit dem man die Differentialgleichung eines beliebigen Σ-Δ-Konverters leicht herleitet. Aber auf der anderen Seite geht die Beschreibung des qualitativen Verhaltens im kleinen verloren. Kommen wir nun zum angekündigten Algorithmus wie ihn Frank Schilder [58] vorgeschlagen hat.

a) Es sei $n \in \mathbb{N}$. Bei gleicher Indizierung wie in den Kapiteln 4.1.2 und 4.1.3, der Beschränktheitsannahme $|U_{n-1}^k(t)| \leq \alpha_n U_{ref}$ für alle Zeiten $t \in [t_0, T]$ und alle $k \in \mathbb{N}$ sowie gleichen Anfangswerten $U_i^k(t_0) = U_i(t_0)$ mit $i \in \{1, \ldots, n\}$ und $U_n(t_0) = 0$ ergeben sich analog zu (4.11) und (4.12) dann

$$\begin{aligned} U_i^k(t) &= \frac{1}{T_i}\int\limits_{t_0}^{t}(U_{i-1}^k(\tau) - \alpha_i U_{ref} X^k(\tau))\mathrm{d}\tau + U_i(t_0), \quad i \in \{1, \ldots, n-1\}, \\ U_n^k(t) &= \frac{1}{T_n}\int\limits_{t_0}^{t}(U_{n-1}^k(\tau) - \alpha_n U_{ref} X^k(\tau))\mathrm{d}\tau. \end{aligned} \tag{4.34}$$

Dabei ist zu beachten, daß bei Konvertern mit $m \leq n$ Komparatoren auch m Beschränktheitsbedingungen notwendig sind.

b) Dann gilt $U_n \to 0$ für $k \to +\infty$ in $C^0[t_0, T]$, und (4.34) wird zu

$$\begin{aligned} U_i(t) &= \frac{1}{T_i}\int\limits_{t_0}^{t}(U_{i-1}(\tau) - \alpha_i U_{ref} X^0(\tau))\mathrm{d}\tau + U_i(t_0), \quad i \in \{1, \ldots, n-1\}, \\ U_n(t) &= 0. \end{aligned} \tag{4.35}$$

Konverter mit m Komparatoren haben m Gleichungssysteme wie (4.35).

c) Die zusätzliche Annahme $U_0 \in C^0[t_0, T]$ liefert die Existenz eines $\varphi \in C^{n-1}[t_0, T]$, und die Differentiation von (4.35) ergibt

$$\dot{U}_i(t) = \frac{1}{T_i}(U_{i-1}(t) - \alpha_i U_{ref}\varphi(t)), \quad i \in \{1, \ldots, n-1\}, \tag{4.36}$$

$$0 = \frac{1}{T_n}(U_{n-1}(t) - \alpha_n U_{ref}\varphi(t)). \tag{4.37}$$

Konverter mit m Komparatoren haben verschiedene φ_l mit $l \in \{1, \ldots, m\}$ und m Systeme bestehend aus Systemen linearer Differentialgleichungen wie (4.36) und (4.37).

d) Rekursives Einsetzen beginnend bei (4.37) in (4.36) liefert dann eine lineare Differentialgleichung $(n-1)$-ter Ordnung mit konstanten Koeffizienten. Konverter mit m Komparatoren haben m solche Differentialgleichungen, wobei die Summe ihrer Ordnungen $n-1$ ist.

Noch einmal wiederholend und ohne die Anfangsbedingung ergeben sich für die Konverter erster bis dritter Ordnung nach diesem Algorithmus mit $\alpha = \alpha_1$, $\beta = \alpha_2$ und $\gamma = \alpha_3$ dann folgende Differentialgleichungen

$$\varphi(t) = \frac{U_0(t)}{\alpha U_{ref}}, \tag{4.38}$$

$$\dot{\varphi}(t) + \frac{\alpha}{\beta T_1}\varphi(t) = \frac{U_0(t)}{\beta T_1 U_{ref}}, \tag{4.39}$$

$$\ddot{\varphi}(t) + \frac{\beta}{\gamma T_2}\dot{\varphi}(t) + \frac{\alpha}{\gamma T_1 T_2}\varphi(t) = \frac{U_0(t)}{\gamma T_1 T_2 U_{ref}}. \tag{4.40}$$

4.1.4 Bemerkungen zu $|U_1(t)| \leq \beta U_{ref}$ für Konverter zweiter Ordnung

Beide Aussagen des Satzes 4.1.15 bez. eines Σ-Δ-Konverters zweiter Ordnung benötigen die Annahme, daß für alle $k \in \mathbb{N}_0$ die Beschränkung $|U_1^k(t)| \leq \beta U_{ref}$ erfüllt ist. Wir wollen nun Bedingungen an die Systemparameter und an die Eingangsfunktion U_0 stellen, unter denen diese Annahme gilt. Die Aussage und den Beweis werden wir jedoch für eine feste Taktzeit Δt führen. Schlußfolgerungen für $\Delta t \to 0$ lassen sich dann leicht ziehen.

Satz 4.1.19 *Es sei $|U_0(t)| \leq \gamma U_{ref}$ für alle Zeiten $t \geq t_0$. Es sei $U_2(t_0) = 0$ sowie $|U_1(t_0)| \leq \frac{\beta}{\alpha}\gamma U_{ref}$. Weiterhin sei U_0 fast überall Lipschitzstetig mit der Lipschitzkonstanten L_U. Gilt dann*

$$\frac{\gamma}{\alpha} + \frac{4\alpha\Delta t}{\beta T_1} + \frac{\beta T_1 L_U}{\alpha^2 U_{ref}} \in (0,1), \tag{4.41}$$

so ist $|U_1(t)| \leq \beta U_{ref}$ für alle Zeiten $t \geq t_0$.

Beweis: Wir wollen hier nur die obere Schranke beweisen, da der Nachweis für die untere analog geführt werden kann.
Die Idee des Beweises liegt darin, anzunehmen, daß U_0 maximal und die Rückkopplung minimal ist, damit $U_1(t)$ maximal wird (vgl. (2.30)). Der Beweis wird induktiv geführt. Als erstes wollen wir zeigen, daß unter den oben gegebenen Voraussetzungen ein Intervall $[t_0, t_0 + \tilde{t}]$ existiert, in dem die Behauptung erfüllt ist. Also sei o. B. d. A. $U_0(t) = \gamma U_{ref}$. Da alle Parameter positiv sind, folgt aus (4.41) und der Voraussetzung $|U_1(t_0)| \leq \frac{\beta}{\alpha}\gamma U_{ref}$ dann

$$|U_1(t_0)| \leq \beta U_{ref}. \tag{4.42}$$

Sei nun $\varepsilon > 0$ so, daß

$$\frac{\gamma}{\alpha}\beta U_{ref} + \frac{4\alpha\Delta t}{T_1}U_{ref} + \frac{\beta^2 T_1 L_U}{\alpha^2} \leq \beta U_{ref} - \varepsilon \tag{4.43}$$

erfüllt ist, d. h., es gilt

$$0 < \varepsilon \leq \beta U_{ref}\left[1 - \frac{\gamma}{\alpha} - \frac{4\alpha\Delta t}{\beta T_1} - \frac{\beta T_1 L_U}{\alpha^2 U_{ref}}\right].$$

Weiterhin nehmen wir an

$$\varrho(t) := U_1(t) - \frac{\beta}{\alpha}U_0(t). \tag{4.44}$$

Wegen (2.30) ist dann für $t \geq t_0$

$$U_1(t) = \frac{1}{T_1} \int_{t_0}^{t} \gamma U_{ref} \mathrm{d}\tau - \frac{\alpha U_{ref}}{T_1} \int_{t_0}^{t} X(\tau) \mathrm{d}\tau + U_1(t_0).$$

Den ungünstigsten Fall annehmend sei nun $X(t) \equiv -1$. Damit erhalten wir

$$U_1(t) \leq (t - t_0) \frac{\gamma U_{ref}}{T_1} + (t - t_0) \frac{\alpha U_{ref}}{T_1} + \frac{\beta}{\alpha} \gamma U_{ref}. \tag{4.45}$$

Sei nun

$$\tilde{t} = \frac{4\alpha^3 \Delta t U_{ref} + \beta^2 T_1^2 L_U}{\alpha^2 (\alpha + \gamma) U_{ref}}. \tag{4.46}$$

Dann gilt für jedes $t \in [t_0, t_0 + \tilde{t}]$

$$(t - t_0) \frac{\gamma U_{ref}}{T_1} + (t - t_0) \frac{\alpha U_{ref}}{T_1} + \frac{\beta}{\alpha} \gamma U_{ref} \leq \frac{\gamma}{\alpha} \beta U_{ref} + \frac{4\alpha \Delta t}{T_1} U_{ref} + \frac{\beta^2 T_1 L_U}{\alpha^2}. \tag{4.47}$$

Wegen (4.45), (4.47) und (4.43) ist dann $U_1(t) \leq \beta U_{ref}$ für jedes $t \in [t_0, t_0 + \tilde{t}]$. Wegen (4.46) ist das Intervall, auf dem die Behauptung erfüllt ist, endlich. Wir benötigen nun ein anderes Argument, um das zulässige Intervall zu erweitern. Dazu benutzen wir (4.44) und (2.31). Nach Teil b) des Lemmas 4.1.12 (ohne Index k) gilt für $t \in [t_0, t_0 + \tilde{t}]$

$$U_2(t) =: \frac{2\Delta t \beta U_{ref}}{T_2} \delta(t) \quad \text{mit} \quad |\delta(t)| \leq 1. \tag{4.48}$$

Wegen $U_2(t_0) = 0$ ist auch $\delta(t_0) = 0$. Aus (2.31) und (4.44) ergibt sich

$$\begin{aligned} \int_{t_0}^{t} X(\tau) \mathrm{d}\tau &= \frac{1}{\beta U_{ref}} \int_{t_0}^{t} U_1(\tau) \mathrm{d}\tau - 2\Delta t \delta(t) \\ &= \frac{1}{\beta U_{ref}} \int_{t_0}^{t} \left(\frac{\beta}{\alpha} U_0(\tau) + \varrho(\tau) \right) \mathrm{d}\tau - 2\Delta t \delta(t). \end{aligned} \tag{4.49}$$

Mit (2.30) und (4.49) erhalten wir

$$\begin{aligned} U_1(t) &= \frac{1}{T_1} \int_{t_0}^{t} \left(U_0(\tau) - \alpha U_{ref} X(\tau) \right) \mathrm{d}\tau + U_1(\tau) \\ &= \frac{1}{T_1} \int_{t_0}^{t} U_0(\tau) \mathrm{d}\tau - \frac{\alpha}{\beta T_1} \int_{t_0}^{t} \left(\frac{\beta}{\alpha} U_0(\tau) + \varrho(\tau) \right) \mathrm{d}\tau + \frac{2\alpha \Delta t U_{ref}}{T_1} \delta(t) + U_1(t_0) \\ &= \frac{1}{T_1} \int_{t_0}^{t} U_0(\tau) \mathrm{d}\tau - \frac{1}{T_1} \int_{t_0}^{t} U_0(\tau) \mathrm{d}\tau - \frac{\alpha}{\beta T_1} \int_{t_0}^{t} \varrho(\tau) \mathrm{d}\tau + \frac{2\alpha \Delta t U_{ref}}{T_1} \delta(t) + U_1(t_0) \\ &= -\frac{\alpha}{\beta T_1} \int_{t_0}^{t} \varrho(\tau) \mathrm{d}\tau + \frac{2\alpha \Delta t U_{ref}}{T_1} \delta(t) + U_1(t_0). \end{aligned} \tag{4.50}$$

Aus (4.44), nach $U_1(t)$ aufgelöst, und (4.50) ergibt sich folgende Volterrasche Integralgleichung zweiter Art

$$\varrho(t) = -\frac{\alpha}{\beta T_1}\int\limits_{t_0}^{t} \varrho(\tau)\mathrm{d}\tau + \frac{2\alpha\Delta t U_{ref}}{T_1}\delta(t) + \left(U_1(t_0) - \frac{\beta}{\alpha}U_0(t)\right). \tag{4.51}$$

Nach Voraussetzung und der Definition von δ sind U_0 und δ absolutstetig (vgl. [50]). Demzufolge ist die Differentiation dieser Funktionen entsprechend des Hauptsatzes der Differental- und Integralrechnung für Lebesgueintegrale (vgl. [11]) sinnvoll. Wegen (4.51), Lemma 4.1.14 und $\varrho(t_0) \leq 0$ gilt dann

$$\begin{aligned}
\varrho(t) &= \varrho(t_0) + \frac{1}{\alpha T_1}\int\limits_{t_0}^{t} \mathrm{e}^{\frac{\alpha}{\beta T_1}(\tau - t)}\left[2\Delta t\alpha^2 U_{ref}\delta'(\tau) - \beta T_1 U_0'(\tau)\right]\mathrm{d}\tau \\
&\leq \frac{\mathrm{e}^{-\frac{\alpha}{\beta T_1}t}}{\alpha T_1}\left[\underbrace{\int\limits_{t_0}^{t} \mathrm{e}^{\frac{\alpha}{\beta T_1}\tau} 2\Delta t\alpha^2 U_{ref}\delta'(\tau)\mathrm{d}\tau}_{} \qquad - \int\limits_{t_0}^{t} \mathrm{e}^{\frac{\alpha}{\beta T_1}\tau}\beta T_1 U_0'(\tau)\mathrm{d}\tau\right] \\
&\qquad\qquad = 2\Delta t\alpha^2 U_{ref}\left[\mathrm{e}^{\frac{\alpha}{\beta T_1}\tau}\delta(\tau)\Big|_{t_0}^{t} - \frac{\alpha}{\beta T_1}\int\limits_{t_0}^{t} \mathrm{e}^{\frac{\alpha}{\beta T_1}\tau}\delta(\tau)\mathrm{d}\tau\right] \\
&\qquad\qquad = 2\Delta t\alpha^2 U_{ref}\left[\mathrm{e}^{\frac{\alpha}{\beta T_1}t}\delta(t) - \frac{\alpha}{\beta T_1}\int\limits_{t_0}^{t} \mathrm{e}^{\frac{\alpha}{\beta T_1}\tau}\delta(\tau)\mathrm{d}\tau\right] \\
&= \frac{2\alpha\Delta t U_{ref}}{T_1}\left[\delta(t) - \frac{\alpha}{\beta T_1}\int\limits_{t_0}^{t} \mathrm{e}^{\frac{\alpha}{\beta T_1}(\tau - t)}\delta(\tau)\mathrm{d}\tau\right] - \frac{\beta}{\alpha}\int\limits_{t_0}^{t} \mathrm{e}^{\frac{\alpha}{\beta T_1}(\tau - t)}U_0'(\tau)\mathrm{d}\tau.
\end{aligned}$$

Hieraus, der Lipschitzbedingung an U_0 und (4.48) ergibt sich folgende Abschätzung

$$\begin{aligned}
\varrho(t) &= \frac{2\alpha\Delta t U_{ref}}{T_1}\left[\delta(t) - \frac{\alpha}{\beta T_1}\int\limits_{t_0}^{t} \mathrm{e}^{\frac{\alpha}{\beta T_1}(\tau - t)}\delta(\tau)\mathrm{d}\tau\right] - \frac{\beta}{\alpha}\int\limits_{t_0}^{t} \mathrm{e}^{\frac{\alpha}{\beta T_1}(\tau - t)}U_0'(\tau)\mathrm{d}\tau \\
&\leq \frac{2\alpha\Delta t U_{ref}}{T_1} + \frac{2\alpha^2\Delta t U_{ref}}{\beta T_1}\int\limits_{t_0}^{t} \mathrm{e}^{\frac{\alpha}{\beta T_1}(\tau - t)}\mathrm{d}\tau + \frac{\beta L_U}{\alpha}\int\limits_{t_0}^{t} \mathrm{e}^{\frac{\alpha}{\beta T_1}(\tau - t)}\mathrm{d}\tau \\
&= \frac{2\alpha\Delta t U_{ref}}{T_1} + \frac{2\alpha\Delta t U_{ref}}{T_1}\left(1 - \mathrm{e}^{\frac{\alpha}{\beta T_1}(t_0 - t)}\right) + \frac{\beta^2 L_U}{\alpha^2}\left(1 - \mathrm{e}^{\frac{\alpha}{\beta T_1}(t_0 - t)}\right) \\
&\leq \frac{4\alpha\Delta t U_{ref}}{T_1} + \frac{\beta^2 T_1 L_U}{\alpha^2}.
\end{aligned} \tag{4.52}$$

Somit haben wir ϱ in $[t_0, t_0 + \tilde{t}]$ abgeschätzt und können nun U_1 mit (4.44) und (4.43) durch

$$U_1(t) \leq \frac{\beta}{\alpha}\gamma U_{ref} + \frac{4\alpha\Delta t U_{ref}}{T_1} + \frac{\beta^2 T_1 L_U}{\alpha^2} \leq \beta U_{ref} - \varepsilon \tag{4.53}$$

nach oben abschätzen. Insbesondere gilt also

$$U_1(t_0 + \tilde{t}) \leq \beta U_{ref} - \varepsilon. \tag{4.54}$$

Bedingung (4.54) entspricht dem Induktionsanfang des Beweises, den wir nun führen wollen. Dabei zeigen wir, daß folgender Schluß unter den oben aufgeführten Voraussetzungen richtig ist. Gilt für ein $t \geq t_0$

$$U_1(t) \leq \beta U_{ref} - \varepsilon,$$

dann existiert ein $z > 0$, welches unabhängig von t ist, so daß

$$\forall \tau \in [t, t+z] : U_1(\tau) \leq U_{ref} - \varepsilon.$$

Diesen Induktionsschritt führen wir hier jedoch nur von $t_0 + \tilde{t}$ nach $t_0 + \tilde{t} + z$.
Sei nun $z := \dfrac{\varepsilon T_1}{(\alpha+\gamma)U_{ref}}$. Dann folgt aus (2.30) für $t \in \left[t_0 + \tilde{t}, t_0 + \tilde{t} + z\right]$

$$U_1(t) \quad = \quad U_1(t_0 + \tilde{t}) + \frac{1}{T_1} \int\limits_{t_0+\tilde{t}}^{t} U_0(\tau) \mathrm{d}\tau - \frac{\alpha U_{ref}}{T_1} \int\limits_{t_0+\tilde{t}}^{t} X(\tau) \mathrm{d}\tau.$$

Hieraus, aus den obigen Annahmen, aus (4.53) und (4.54) folgt für $X(t) \equiv -1$ dann

$$\begin{aligned} U_1(t) &\leq U_1(t_0 + \tilde{t}) + \frac{\gamma U_{ref}}{T_1} z + \frac{\alpha U_{ref}}{T_1} z \\ &= U_1(t_0 + \tilde{t}) + \varepsilon \\ &\leq \beta U_{ref}. \end{aligned} \tag{4.55}$$

Somit sind also wegen (4.55) für $t \in \left[t_0, t_0 + \tilde{t} + z\right]$ alle Voraussetzungen von Lemma 4.1.12 erfüllt. Mit den gleichen Überlegungen beginnend bei (4.48) bis (4.53) ergibt sich

$$\forall t \in \left[t_0, t_0 + \tilde{t} + z\right] : \varrho(t) \leq \frac{4\alpha \Delta t U_{ref}}{T_1} + \frac{\beta^2 T_1 L_U}{\alpha^2}. \tag{4.56}$$

Somit gilt wegen (4.44) nicht nur (4.55) sondern sogar

$$\forall t \in \left[t_0, t_0 + \tilde{t} + z\right] : U_1(t) \leq \beta U_{ref} - \varepsilon. \tag{4.57}$$

Damit ist der Induktionsschritt von $t_0 + \tilde{t}$ nach $t_0 + \tilde{t} + z$ zur Abschätzung der oberen Schranke von U_1 getan. Somit ist alles gezeigt. □

Bemerkung 4.1.20 Der Beweis der Aussage des Satzes 4.1.19 weist folgende Besonderheiten auf.

a) Wir benötigen für $U_1(t_0)$ eine spezielle Abschätzung, welche sich nur im Intervall $[t_0, t_0 + \tilde{t}]$ widerspiegelt. Später geht diese nach der zweiten Abschätzung durch die Integralgleichung in den Wert $U_1(t_0 + \tilde{t})$ ein und taucht danach nicht mehr auf.

b) Die eigentliche Induktion beginnt erst für das Intervall $[t_0 + \tilde{t}, t_0 + \tilde{t} + z]$. Danach „arbeitet“ die Argumentation auf den Intervallen $[t_0 + \tilde{t} + nz, t_0 + \tilde{t} + (n+1)z]$ mit $n \in \mathbb{N}$.

c) Bei der ersten Abschätzung (vgl. (4.55)) in einem Induktionsschritt erreichen wir durch (2.30) mit $U_0(t) \equiv \gamma U_{ref}$ und $X(t) \equiv -1$ nur

$$U_1(t) \leq \beta U_{ref}.$$

Bei der zweiten (vgl. (4.57)) erhalten wir jedoch mit Hilfe des Lemmas 4.1.12 durch Abschätzung der Lösung einer Integralgleichung sogar

$$U_1(t) \leq \beta U_{ref} - \varepsilon.$$

Das gelingt nur dadurch, daß keine spezielle Annahme an $X(t)$ gestellt werden muß. Dies ist hier nicht nötig, da (2.31) nach $\int X(\tau)\mathrm{d}\tau$ aufgelöst und dies in (2.30) eingesetzt wird. Wie sich $X(t)$ verhält, wissen wir nicht. Wir können nur eine Näherungsformel für die zweiten Differenzen wie in Kapitel 5.3.3 angeben.

d) Es ist bemerkenswert, daß die Intervallänge z bei der eigentlichen Induktion (vgl. b)) nicht mit Vielfachen von Δt übereinstimmt. Die Begründung dafür liegt in c).

Folgerung 4.1.21 Es sei $|U_0(t)| \leq \gamma U_{ref}$ für alle Zeiten $t \geq t_0$. Es sei $U_2(t_0) = 0$ sowie $|U_1(t_0)| \leq \frac{\beta}{\alpha}\gamma U_{ref}$. Ist zusätzlich

$$\frac{\gamma}{\alpha} + \frac{4\Delta t\alpha}{\beta T_1} \in (0,1),$$

so gilt für fast überall konstante Funktionen U_0 dann $|U_1(t)| \leq \beta U_{ref}$ für alle Zeiten $t \geq t_0$.

Beweis: Der Beweis wird wie der des Satzes 4.1.19 geführt. Der Unterschied besteht in der Abschätzung der Formeln (4.52) und (4.56). Dort ist dann $L_U = 0$. □

Folgerung 4.1.22 Für $\Delta t \to 0$ werden die beiden Bedingungen aus dem Satz 4.1.19 bzw. aus der Folgerung 4.1.21 zu

$$\frac{\gamma}{\alpha} + \frac{\beta T_1 L_U}{\alpha^2 U_{ref}} \in (0,1) \quad \text{bzw.} \quad \frac{\gamma}{\alpha} \in (0,1).$$

Somit ist der Quotient $\frac{\gamma}{\alpha}$ und die Lipschitzkonstante L_U bei hinreichend kleinem Δt von entscheidender Bedeutung. In praxi ist jedoch $\frac{\Delta t}{T_1} \in \left[\frac{1}{15}, \frac{1}{5}\right]$.

Bemerkung 4.1.23 Die Aussage des Satzes 4.1.19 kann wiederum als Berechtigung der Untersuchung stückweise konstanter Eingangssignale bzw. solcher mit kleiner L-Konstante aufgefaßt werden. Ist nämlich L_U hinreichend groß, so ist (4.41) nicht erfüllt. Dann ist möglicherweise $|U_1| \geq \beta U_{ref}$, was zu Instabilitäten im physikalischen Sinn führt (vgl. Kap. 6.1).

Wir wollen uns nun die zulässigen Parameterbereiche ansehen und dabei die Aussagen an einem Beispiel illustrieren.

Bemerkung 4.1.24 Es sei $\frac{\Delta t}{T_1} = \frac{1}{10}$. Betrachten wir nun $U_0(t) = \gamma U_{ref} \sin \frac{\pi}{500\Delta t} t$ als Beispiel für eine L-stetige Funktion wie in Satz 4.1.19. Dann gilt für die Lipschitzkonstante $L_U = \frac{\gamma U_{ref}\pi}{500\Delta t}$. Dabei ist die Überabtastrate also eintausend, d. h., eine Periode des Eingangssignals U_0 entspricht tausend Zeittakten. In den folgenden Abbildungen werden die zulässigen Parameterbereiche entsprechend Satz 4.1.19 und Folgerung 4.1.21 dargestellt. Dabei ist im linken Bild jeweils α-β-Bereich für $\gamma = 1$ und im rechten der β-γ-Bereich für $\alpha = 1$ aufgezeigt. Zudem ist in jedem Fall $U_{ref} = 1$.

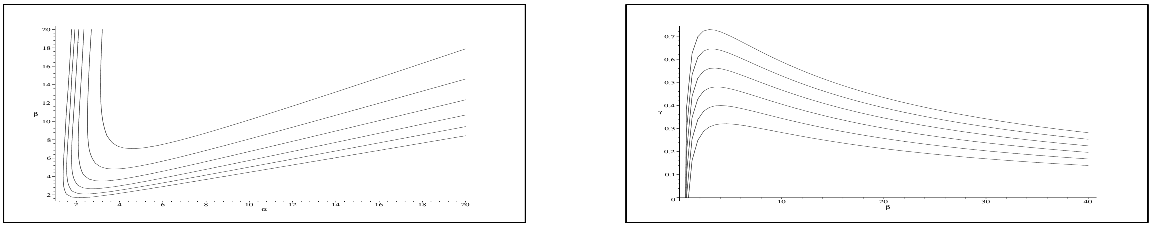


Abbildung 10: $\dfrac{\gamma}{\alpha} + \dfrac{4\alpha\Delta t}{\beta T_1} + \dfrac{\beta\gamma T_1 \pi}{\alpha^2 500 \Delta t} \in \left\{\dfrac{5}{10}, \dots, \dfrac{10}{10}\right\}$

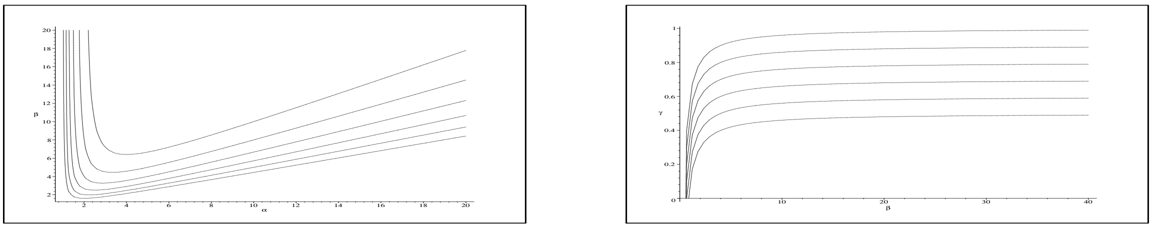


Abbildung 11: $\dfrac{\gamma}{\alpha} + \dfrac{4\alpha\Delta t}{\beta T_1} \in \left\{\dfrac{5}{10}, \dots, \dfrac{10}{10}\right\}$

4.2 Abgeleitete Filtervorschläge

Ausgehend von den Kapiteln 4.1.2 und 4.1.3 wollen wir Filtervorschläge machen und eine zugehörige Fehlerdiskussion führen. Die Filter sind wiederum FIR-Filter [26] und werden durch gleitende Mittelung realisiert. Es sei dabei bemerkt, daß bei der Fehlerabschätzung nur Aussagen bez. integraler Mittelwerte von U_0 getroffen werden können.

4.2.1 Filtervorschlag für den Konverter erster Ordnung

In diesem Abschnitt geben wir einen Filtervorschlag für einen Modulator erster Ordnung an. Ausgehend von der Differentialgleichung und Resultaten, die zu ihrer Herleitung führten, erhalten wir den doppelten Fehler wie in Folgerung 3.1.2. Jedoch unter Benutzung der expliziten Lösung (2.21) ergibt sich die gleiche Abschätzung wie in der Folgerung 3.1.2. Dies ist auch nicht überraschend, interessant dagegen ist die Herleitung des Filtervorschlages. Diese werden wir später auch beim Modulator zweiter Ordnung nutzen.

Bemerkung 4.2.1 In allen weiteren Überlegungen sei Δt konstant. Ziel ist eine geeignete digitale Rekonstruktion von $U_0(t)$ aus der Treppenfunktion $X(t)$. Ausgangspunkt ist die Gleichung in Teil b) im Satz 4.1.10. Zwei Aspekte von (4.4) können wir praktisch nicht erfüllen. Zum einen liegt keine punktweise sondern nur fast überall punktweise Gleichheit vor. Dies ist eine integrale Eigenschaft, da punktweise Gleichheit erst nach einer Integration gegeben ist. Da wir U_0 approximieren wollen, dividieren wir durch die entsprechende Intervallänge. Somit entsteht der erste Fehler, wir rekonstruieren nicht U_0 sondern den integralen Mittelwert

$$U_0\left(t_n + \frac{m\Delta t}{2}\right) \approx \frac{1}{m\Delta t} \int\limits_{t_n}^{t_{n+m}} U_0(\tau)\mathrm{d}\tau.$$

Der zweite Aspekt ist, daß $X^0(t)$ schwacher Grenzwert der Folge von Treppenfunktionen $(X^k)_{k\in\mathbb{N}}$ ist. Für $\Delta t > 0$ haben wir also einen zweiten Fehler, den wir mit Hilfe des Lemmas 4.1.9

abschätzen.

Lemma 4.2.2 *Es sei $m \in \mathbb{N}$ und $|U_0(t)| \leq \alpha U_{ref}$ für alle Zeiten $t \geq t_0$. Weiterhin sei $U_1(t_0) = 0$. Dann gilt*

$$\forall n \in \mathbb{N}_0 : \left| \frac{1}{m\Delta t} \int\limits_{t_n}^{t_{n+m}} U_0(\tau) \mathrm{d}\tau - \left[\frac{\alpha U_{ref}}{m} \sum_{s=1}^{m} X_{n+s-1} \right] \right| \leq \frac{4\alpha U_{ref}}{m}. \tag{4.58}$$

Dabei bezeichne $\frac{\alpha U_{ref}}{m} \sum_{s=1}^{m} X_{n+s-1}$ *unseren Filtervorschlag.*

Beweis: Es sei $n \in \mathbb{N}$. Nach (2.1), (2.2) und Teil b) in Lemma 4.1.9 gilt dann für alle $t \geq t_0$

$$|U_1(t)| = \left| \int\limits_{t_0}^{t} U_0(\tau) \mathrm{d}\tau - \alpha U_{ref} \int\limits_{t_0}^{t} X(\tau) \mathrm{d}\tau \right| \leq \frac{2\alpha \Delta t U_{ref}}{T_1}. \tag{4.59}$$

Da $X(\tau)$ stückweise konstant ist, folgt aus (4.59) insbesondere

$$\begin{aligned}
&\left| \frac{1}{m\Delta t} \int\limits_{t_n}^{t_{n+m}} U_0(\tau) \mathrm{d}\tau - \frac{\alpha U_{ref}}{m\Delta t} \int\limits_{t_n}^{t_{n+m}} X(\tau) \mathrm{d}\tau \right| \\
&= \left| \frac{1}{m\Delta t} \int\limits_{t_0}^{t_{n+m}} U_0(\tau) \mathrm{d}\tau - \frac{1}{m\Delta t} \int\limits_{t_0}^{t_n} U_0(\tau) \mathrm{d}\tau - \frac{\alpha U_{ref}}{m\Delta t} \int\limits_{t_0}^{t_{n+m}} X(\tau) \mathrm{d}\tau + \frac{\alpha U_{ref}}{m\Delta t} \int\limits_{t_0}^{t_n} X(\tau) \mathrm{d}\tau \right| \\
&\leq \left| \frac{1}{m\Delta t} \int\limits_{t_0}^{t_{n+m}} U_0(\tau) \mathrm{d}\tau - \frac{\alpha U_{ref}}{m\Delta t} \int\limits_{t_0}^{t_{n+m}} X(\tau) \mathrm{d}\tau \right| + \left| \frac{1}{m\Delta t} \int\limits_{t_0}^{t_n} U_0(\tau) \mathrm{d}\tau - \frac{\alpha U_{ref}}{m\Delta t} \int\limits_{t_0}^{t_n} X(\tau) \mathrm{d}\tau \right| \\
&= \frac{T_1}{m\Delta t} \left[|U_1(t_{n+m})| + |U_1(t_n)| \right] \\
&\leq \frac{4\alpha U_{ref}}{m}.
\end{aligned}$$

□

Folgerung 4.2.3 Es sei $m \in \mathbb{N}$ und $|U_0(t)| \leq \alpha U_{ref}$ für alle Zeiten $t \geq t_0$. Weiterhin sei $U_1(t_0) = 0$. Unter Benutzung des Satzes 2.1.6 gilt sogar

$$\forall n \in \mathbb{N}_0 : \left| \frac{1}{m\Delta t} \int\limits_{t_n}^{t_{n+m}} U_0(\tau) \mathrm{d}\tau - \left[\frac{\alpha U_{ref}}{m} \sum_{s=1}^{m} X_{n+s} \right] \right| \leq \frac{2\alpha U_{ref}}{m}. \tag{4.60}$$

Dies stimmt mit dem Resultat aus Folgerung 3.1.2 überein.

Beweis: Nach der Definition der ersten Differenzen und (2.21) sowie der Annahme, daß dort für den Schwellenwert $s = 0$ gilt, ist

$$\sum_{s=1}^{m} X_{n+s} \quad = \quad Y_{n+m+1} - Y_{n+1}$$

$$= \quad 2\left\lfloor \frac{1}{2\alpha\Delta t U_{ref}} \int\limits_{t_0}^{t_{n+m}} U_0(\tau)\mathrm{d}\tau + \frac{m+n}{2} \right\rfloor - n - m + 1$$
$$- \left(2\left\lfloor \frac{1}{2\alpha\Delta t U_{ref}} \int\limits_{t_0}^{t_n} U_0(\tau)\mathrm{d}\tau + \frac{n}{2} \right\rfloor - n + 1 \right).$$

Hiermit folgt dann

$$\frac{1}{m\Delta t} \int\limits_{t_n}^{t_{n+m}} U_0(\tau)\mathrm{d}\tau - \left[\frac{\alpha U_{ref}}{m} \sum_{s=1}^{m} X_{n+s} \right]$$
$$= \frac{2\alpha U_{ref}}{m} \left(\frac{1}{2\alpha\Delta t U_{ref}} \int\limits_{t_0}^{t_{n+m}} U_0(\tau)\mathrm{d}\tau + \frac{n+m}{2} - \left\lfloor \frac{1}{2\alpha\Delta t U_{ref}} \int\limits_{t_0}^{t_{n+m}} U_0(\tau)\mathrm{d}\tau + \frac{n+m}{2} \right\rfloor \right.$$
$$\left. - \frac{1}{2\alpha\Delta t U_{ref}} \int\limits_{t_0}^{t_n} U_0(\tau)\mathrm{d}\tau - \frac{n}{2} + \left\lfloor \frac{1}{2\alpha\Delta t U_{ref}} \int\limits_{t_0}^{t_n} U_0(\tau)\mathrm{d}\tau + \frac{n}{2} \right\rfloor \right). \tag{4.61}$$

Wegen $x - \lfloor x \rfloor \in [0, 1)$ für jedes $x \in \mathbb{R}$, folgt aus (4.61) dann die Behauptung. □

Bemerkung 4.2.4 Um einen möglichst kleinen Fehler bei gleitender Mittelung zu machen, suggeriert das Ergebnis der Folgerung 4.2.3, den Wert m möglichst groß zu wählen. Dies ist nicht richtig, da wir dort nicht U_0 approximieren sondern deren integralen Mittelwert. Wir wollen hier den eigentlichen Fehler für Polynome kleiner Ordnung angeben. Dabei gilt

$$\left| U_0\left(t_n + \frac{m\Delta t}{2}\right) - \frac{\alpha U_{ref}}{m} \sum_{s=1}^{m} X_{n+s} \right|$$
$$\leq \left| U_0\left(t_n + \frac{m\Delta t}{2}\right) - \frac{1}{m\Delta t} \int\limits_{t_n}^{t_{n+m}} U_0(\tau)\mathrm{d}\tau \right| + \left| \frac{1}{m\Delta t} \int\limits_{t_n}^{t_{n+m}} U_0(\tau)\mathrm{d}\tau - \frac{\alpha U_{ref}}{m} \sum_{s=1}^{m} X_{n+s} \right|$$
$$\leq |f(U_0)| + \frac{2\alpha U_{ref}}{m}$$

mit

$$|f(U_0)| \leq \begin{cases} 0, & \text{falls } U_0(t) = a, \\ 0, & \text{falls } U_0(t) = at + b, \\ \frac{a}{12}\Delta t^2 m^2, & \text{falls } U_0(t) = at^2 + bt + c, \\ \left(\frac{at_n}{4} + \frac{a}{8}\Delta t m + \frac{b}{12}\right)\Delta t^2 m^2, & \text{falls } U_0(t) = at^3 + bt^2 + ct + d. \end{cases}$$

Somit ist der Fehler $f(U_0)$ von der Ordnung $O((\Delta t m)^k)$ für $k = \lfloor \frac{n}{2} \rfloor$ mit $n \in \mathbb{N}_0$ bei polynomialen Verhalten von $U_0(t)$ vom Grade n. Verhält sich U_0 näherungsweise kubisch, so hängt der Fehler zusätzlich auch von t_n ab.
Sei nun beispielsweise $U_0(t) = at^2 + bt + c$ mit $a > 0$. Dann ist

$$\left| U_0\left(t_n + \frac{m\Delta t}{2}\right) - \frac{\alpha U_{ref}}{m} \sum_{s=1}^{m} X_{n+s} \right| \leq \frac{am^2\Delta t^2}{12} + \frac{2\alpha U_{ref}}{m}.$$

Danach ist der Fehler für $m_{\min} = \frac{1}{a\Delta t}\sqrt[3]{12\alpha\Delta t U_{ref} a^2}$ minimal. Er hat den Wert $\frac{1}{4}\sqrt[3]{(12\alpha\Delta t U_{ref})^2 a}$.

4.2.2 Filtervorschlag für den Konverter zweiter Ordnung

In diesem Kapitel leiten wir einen Filtervorschlag für einen Σ-Δ-Modulator zweiter Ordnung her, der bez. der Wortlänge m den zweiten integralen Mittelwert von U_0 mit der Ordnung $O\left(\frac{1}{m^2}\right)$ approximiert. Sei nun in allen Überlegungen Δt konstant.

Satz 4.2.5 *Es sei $m \in \mathbb{N}$, $U_2(t_0) = 0$ und $|U_1(t)| \le \beta U_{ref}$ für alle Zeiten $t \ge t_0$. Dann gilt für jedes $n \in \mathbb{N}$ mit $n \ge m$*

$$\begin{aligned}
&\left| \frac{1}{(m\Delta t)^2}\left(\int_{t_n}^{t_{n+m}}\int_{\tau-m\Delta t}^{\tau} U_0(r)\mathrm{d}r\mathrm{d}\tau\right)\right. \\
&\quad \left. - \left[\frac{T_1\beta U_{ref}}{m^2\Delta t}\sum_{s=1}^{m}(X_{n+s}-X_{n-m+s}) + \frac{\alpha U_{ref}}{m^2}\sum_{s=1}^{m}\left(s-\frac{1}{2}\right)(X_{n-m+s-1}+X_{n+m-s})\right]\right| \\
&\quad \le \frac{4T_1\beta U_{ref}}{m^2\Delta t}.
\end{aligned} \tag{4.62}$$

Dabei bezeichne der erste Summand aus (4.62) den integralen Mittelwert von $U_0(t_n)$ und der zweite den gewichteten Filtervorschlag für einen Modulator zweiter Ordnung.

Beweis: Es sei $m, n \in \mathbb{N}$ mit $n \ge m$. Aus Teil c) des Lemmas 4.1.12 folgt

$$\begin{aligned}
\left|\frac{T_1 T_2}{U_{ref}}U_2(t)\right| &= \left|\int_{t_0}^{t}\int_{t_0}^{\tau}\frac{U_0(r)}{U_{ref}}\mathrm{d}r\mathrm{d}\tau - \int_{t_0}^{t}\left[\int_{t_0}^{\tau}\alpha X^k(r)\mathrm{d}r + T_1\beta X^k(\tau) - \frac{U_1(t_0)}{U_{ref}}\right]\mathrm{d}\tau\right| \\
&\le 2\beta T_1\Delta t_k.
\end{aligned} \tag{4.63}$$

Aus (4.63) ist ersichtlich, daß sich nach zweimaliger Differentiation von U_2 eine Approximation von U_0 durch X ergibt. Diese Differentiation erfolgt numerisch durch Bildung zweiter Differenzen von U_2 nach der Formel

$$U_2''(t) \approx \frac{U_2(t_{n-m}) - 2U_2(t_n) + U_2(t_{n+m})}{m^2\Delta t^2}. \tag{4.64}$$

Nach Voraussetzung sind alle Annahmen der Folgerung 4.2.3 für U_1 erfüllt. Somit sind nach (2.31)

$$\left|\int_{t_n}^{t_{n+m}} U_1(\tau)\mathrm{d}\tau - \beta U_{ref}\Delta t\sum_{s=1}^{m} X_{n+s}\right| \le 2\beta\Delta t U_{ref}, \tag{4.65}$$

$$\left|\int_{t_{n-m}}^{t_n} U_1(\tau)\mathrm{d}\tau - \beta U_{ref}\Delta t\sum_{s=1}^{m} X_{n-m+s}\right| \le 2\beta\Delta t U_{ref} \tag{4.66}$$

die ersten Differenzen von U_2. Da $X(t)$ stückweise konstant ist, gilt

$$\begin{aligned}\int_{t_n}^{t_{n+m}}\int_{t_n}^{\tau} X(r)\mathrm{d}r\mathrm{d}\tau - \int_{t_{n-m}}^{t_n}\int_{t_n}^{\tau} X(r)\mathrm{d}r\mathrm{d}\tau &= \int_{t_n}^{t_{n+m}}\int_{\tau-m\Delta t}^{\tau} X(r)\mathrm{d}r\mathrm{d}\tau \\ &= \Delta t^2 \sum_{s=1}^{m}(X_{n+s}-X_{n-m+s}). \qquad (4.67)\end{aligned}$$

Also folgt aus (4.65), (4.66) und (4.67) bei der Bildung zweiter Differenzen gemäß des Zählers von (4.64)

$$\begin{aligned}&\left|\left(\int_{t_n}^{t_{n+m}} U_1(\tau)\mathrm{d}\tau - \int_{t_{n-m}}^{t_n} U_1(\tau)\mathrm{d}\tau\right) - \beta U_{ref}\Delta t\sum_{s=1}^{m}(X_{n+s}-X_{n-m+s})\right| \\ &= \left|\frac{1}{T_1}\int_{t_n}^{t_{n+m}}\int_{t_n}^{\tau}[U_0(r)-\alpha U_{ref}X(r)]\mathrm{d}r\mathrm{d}\tau + \int_{t_n}^{t_{n+m}} U_1(t_n)\mathrm{d}\tau\right. \\ &\quad -\frac{1}{T_1}\int_{t_{n-m}}^{t_n}\int_{t_n}^{\tau}[U_0(r)-\alpha U_{ref}X(r)]\mathrm{d}r\mathrm{d}\tau - \int_{t_{n-m}}^{t_n} U_1(t_n)\mathrm{d}\tau \\ &\quad \left.-\beta U_{ref}\Delta t\sum_{s=1}^{m}(X_{n+s}-X_{n-m+s})\right| \\ &= \left|\frac{1}{T_1}\int_{t_n}^{t_{n+m}}\int_{\tau-m\Delta t}^{\tau}[U_0(r)-\alpha U_{ref}X(r)]\mathrm{d}r\mathrm{d}\tau - \beta U_{ref}\Delta t\sum_{s=1}^{m}(X_{n+s}-X_{n-m+s})\right| \\ &= \left|\frac{1}{T_1}\int_{t_n}^{t_{n+m}}\int_{\tau-m\Delta t}^{\tau} U_0(r)\mathrm{d}r\mathrm{d}\tau - \frac{\alpha\Delta t^2 U_{ref}}{T_1}\sum_{s=1}^{m}\left(s-\frac{1}{2}\right)(X_{n-m+s-1}+X_{n+m-s})\right. \\ &\quad \left.-\beta U_{ref}\Delta t\sum_{s=1}^{m}(X_{n+s}-X_{n-m+s})\right| \\ &\leq 4\beta\Delta t U_{ref}. \qquad (4.68)\end{aligned}$$

Wegen

$$U_0(t_n) \approx \frac{1}{m^2\Delta t^2}\int_{t_n}^{t_{n+m}}\int_{\tau-m\Delta t}^{\tau} U_0(r)\mathrm{d}r\mathrm{d}\tau$$

ergibt sich die Behauptung durch Multiplikation von (4.68) mit $\frac{T_1}{m^2\Delta t^2}$. □

Folgerung 4.2.6 Es seien die Voraussetzungen des Satzes 4.2.5 erfüllt. Weiterhin seien die Funktionen $U_0(t), U_1(t)$ und $X(t)$ identisch Null auf dem Intervall $[t_0 - m\Delta t, t_0)$. Dann gilt die Aussage (4.62) für jedes $n \in \mathbb{N}_0$

Bemerkung 4.2.7 Nach Satz 4.2.5 ist der Fehler zwischen dem zweiten integralen Mittelwert und dem gewichteten Filtervorschlag von der Ordnung $O\left(\frac{1}{m^2}\right)$. Legt man nur die Aussage c) des Lemmas 4.1.12 zur Gewinnung einer Approximation zugrunde, dann wäre der Fehler von der Ordnung $O\left(\frac{1}{m}\right)$. Dies läßt sich gerade durch die numerische Differentiation der linken Seite von (4.63) nach der Vorschrift (4.64) verbessern. Dadurch ist es möglich, den $O\left(\frac{1}{m}\right)$-Ausdruck $\int\limits_{t_0}^{t_{n+m}} U_1(t_n)\mathrm{d}\tau$ in (4.68) zu eliminieren.

Somit läßt sich nach den gemachten Überlegungen sagen:

> Analytische Differentiation des Komparatoreingangs liefert im Grenzfall $\Delta t \to 0$ ein Anfangswertproblem.
>
> Numerische Differentiation des Komparatoreingangs liefert für $\Delta t > 0$ einen gewichteten Filtervorschlag.

Bemerkung 4.2.8 Um einen möglichst kleinen Fehler bei der oben vorgeschlagenen gewichteten Mittelung zu machen, suggeriert das Ergebnis des Satzes 4.2.5, den Wert m möglichst groß zu wählen. Dies ist nicht richtig, da wir dort nicht U_0 approximieren sondern deren zweiten integralen Mittelwert. Wir wollen hier den eigentlichen Fehler für Polynome kleiner Ordnung angeben. Dabei gilt

$$
\begin{aligned}
&\left|U_0(t_n) - \left[\frac{T_1\beta U_{ref}}{m^2\Delta t}\sum_{s=1}^{m}(X_{n+s} - X_{n-m+s}) + \frac{\alpha U_{ref}}{m^2}\sum_{s=1}^{m}\left(s-\frac{1}{2}\right)(X_{n-m+s-1} + X_{n+m-s})\right]\right| \\
&\leq \left|U_0(t_n) - \frac{1}{(m\Delta t)^2}\int\limits_{t_n}^{t_{n+m}}\int\limits_{\tau-m\Delta t}^{\tau} U_0(r)\mathrm{d}r\mathrm{d}\tau\right| \\
&\quad + \left|\frac{1}{(m\Delta t)^2}\int\limits_{t_n}^{t_{n+m}}\int\limits_{\tau-m\Delta t}^{\tau} U_0(r)\mathrm{d}r\mathrm{d}\tau\right. \\
&\quad - \left.\left[\frac{T_1\beta U_{ref}}{m^2\Delta t}\sum_{s=1}^{m}(X_{n+s} - X_{n-m+s}) + \frac{\alpha U_{ref}}{m^2}\sum_{s=1}^{m}\left(s-\frac{1}{2}\right)(X_{n-m+s-1} + X_{n+m-s})\right]\right| \\
&\leq |f(U_0)| + \frac{4T_1\beta U_{ref}}{m^2\Delta t}
\end{aligned}
$$

mit

$$
|f(U_0)| \leq \begin{cases} 0, & \text{falls } U_0(t) = a, \\ 0, & \text{falls } U_0(t) = at + b, \\ \frac{a}{6}\Delta t^2 m^2, & \text{falls } U_0(t) = at^2 + bt + c, \\ \left(\frac{at_n}{2} + \frac{b}{6}\right)\Delta t^2 m^2, & \text{falls } U_0(t) = at^3 + bt^2 + ct + d. \end{cases}
$$

Somit ist der Fehler $f(U_0)$ von der Ordnung $O((\Delta tm)^k)$ für $k = \lfloor \frac{n}{2} \rfloor$ mit $n \in \mathbb{N}_0$ bei polynomialen Verhalten von $U_0(t)$ vom Grade n. Verhält sich U_0 näherungsweise kubisch, so hängt der Fehler zusätzlich auch von t_n ab.

Sei nun beispielsweise $U_0(t) = at^2 + bt + c$ mit $a > 0$. Dann ist

$$
\left|U_0(t_n) - \left[\frac{T_1\beta U_{ref}}{m^2\Delta t}\sum_{s=1}^{m}(X_{n+s} - X_{n-m+s}) + \frac{\alpha U_{ref}}{m^2}\sum_{s=1}^{m}\left(s-\frac{1}{2}\right)(X_{n-m+s-1} + X_{n+m-s})\right]\right|
$$

$$\leq \frac{am^2\Delta t^2}{6} + \frac{4T_1\beta U_{ref}}{m^2\Delta t}.$$

Danach ist der Fehler für $m_{\min} = \frac{1}{a\Delta t}\sqrt[4]{24\beta\Delta t U_{ref} T_1 a^3}$ minimal. Er hat den Wert $\frac{2}{3}\sqrt{6\beta\Delta t U_{ref} T_1 a}$.

4.2.3 Vergleich und Ausblick

Beispiel 4.2.9 Wir wollen nun die theoretischen Ergebnisse für die beiden Modulatoren an einem Beispiel illustrieren. Dazu sei

$$U_0(t) = \frac{1}{2}\sin\left(\frac{\pi}{250}t\right).$$

Im Beispiel seien die Parameter der Modulatoren durch $T_1 = 10$, $T_2 = 1$, $\Delta t = 1$, $\alpha = 1$, $\beta = 2$, $U_{ref} = 1$ und $m = 50$ gegeben. Nach (4.60) und (4.62) gelten dann für einen Modulator erster bzw. zweiter Ordnung die oberen Schranken $\frac{2\alpha U_{ref}}{m} = 0.04$ bzw. $\frac{4T_1\beta U_{ref}}{m^2\Delta t} = 0.032$ für den Fehler zwischen intergalem Mittelwert und dem Filterausgang. Abbildung 12 stellt die Wirkungsweise unseres Filtervorschlags für einen Modulator erster Ordnung dar. Das linke Bild in Abbildung 12 zeigt die beiden Summanden unter den Betragsstrichen in (4.60). Das rechte dagegen zeigt deren Differenz.

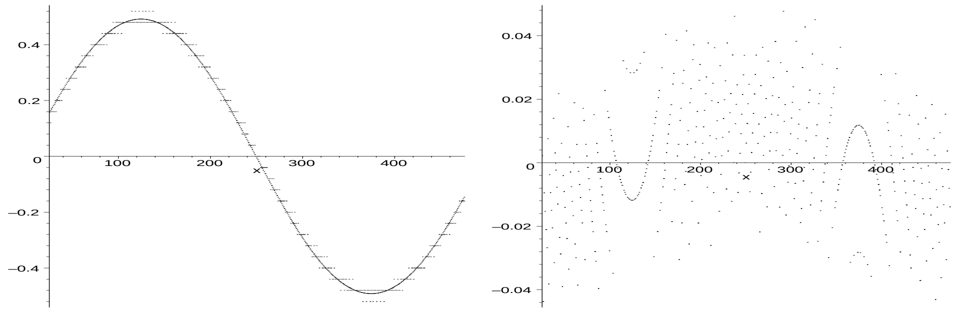


Abbildung 12: Filterung bez. Modulators erster Ordnung

Durch die gleiche Darstellung verdeutlicht Abbildung 13 die Wirkungsweise eines Filters eines Modulators zweiter Ordnung (vgl. 4.62).

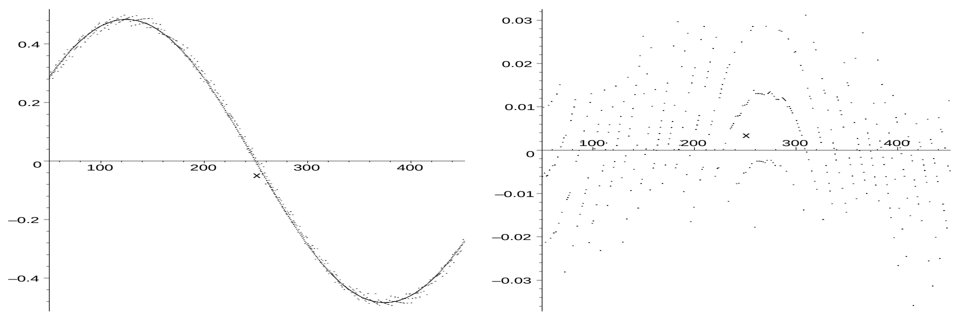


Abbildung 13: Filterung bez. Modulators zweiter Ordnung

Nun wollen wir einen weiteren interessanten Zusammenhang beim Modulator zweiter Ordnung graphisch darstellen. Sei nun $\Delta t = 0.01$ und $m = 70$. Das linke Bild in Abbildung

14 zeigt die Funktion $U_0(t) = \frac{1}{2}\sin t$ und die Treppenfunktion $\tilde{\varphi}(t) = \frac{\alpha U_{ref}}{m}\sum_{s=1}^{m} X_{n+s-1}$ für $t \in [t_{n+m}, t_{n+m+1})$. Die rechte dagegen zeigt neben U_0 auch die Lösung des Anfangswertproblems (4.29), welche gerade durch die Funktion $\tilde{\varphi}$ angenähert wird.

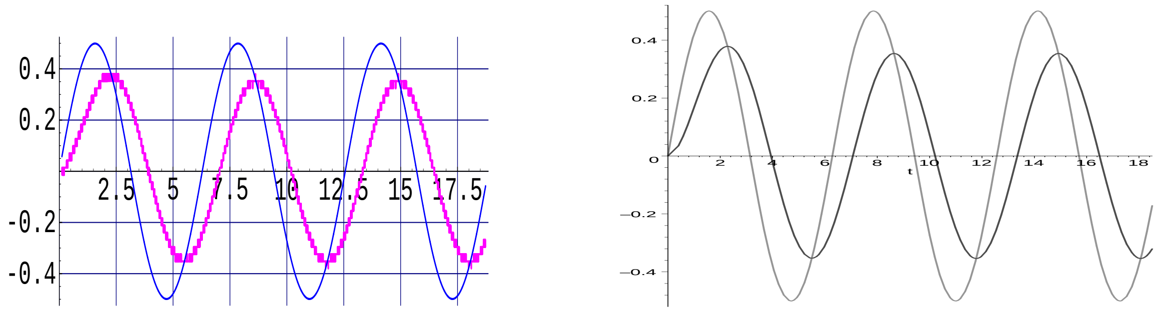


Abbildung 14: Zusammenhang zwischen einfacher Mittelung (links) und der Lösung des Anfangswertproblems (rechts) beim Modulator zweiter Ordnung

Die beiden Abbildungen legen die Vermutung nahe, daß ein einfacher Mittelwertbildner als Filter eines Konverters zweiter Ordnung die Lösung des entsprechenden AWP annähert. Die Berechnung der Phasenverschiebung und des Amplitudenfaktors, die wir in den Bildern erkennen, werden wir in Kapitel 8.5 aus den Differentialgleichungen berechnen. Dies ist auch ein Aspekt, der die Modellierung durch Differentialgleichungen bestätigt.

Bemerkung 4.2.10 Die Abschnitte 4.2.1 und 4.2.2 zusammenfassend, geben wir nun die Vorgehensweise zur Herleitung eines Filters eines Modulators n-ter Ordnung an.

a) Unter der Anahme, daß U_{n-1} beschränkt ist, kann man dann wie in den Lemmas 4.1.9 und 4.1.12 den Komparatoreingang U_n abschätzen. $U_n(t)$ entspricht aber gerade einer Summe von Mehrfachintegralen, wenn man iterativ die Gleichungen 4.34 ineinander einsetzt. Dabei ergibt sich ein n-faches Integral von U_0 und je ein r-faches gewichtetes Integral von X mit $r \in \{1, \dots, n-1\}$.

b) Um Terme der Ordnung $O\left(\frac{1}{m^r}\right)$ mit $r \in \{1, \dots, n-1\}$ zu eliminieren, benutzt man die numerische Differentiation. Diese entspricht gerade einer Differenzenbildung n-ter Ordnung.

c) Da X stückweise konstant ist, wird die Mehrfachintegration von X zur Mehrfachsummation der X_k, $k \in \mathbb{N}_0$.

d) Diese n-te Differenz wird durch das Integrationsgebiet des n-fachen Integrals von U_0 dividiert, um somit U_0 integral zu mitteln. Damit approximiert eine gewichtete Summe von X_k diesen integralen Mittelwert. Der Fehler berechnet sich durch die Abschätzung des Komparatoreingangs in a).

Durch diese Vorgehensweise gewinnen wir einen Filtervorschlag, welcher einen Fehler der Ordnung $O\left(\frac{1}{m^n}\right)$ besitzt.

Fazit 4.2.11 *Bei der digitalen Rekonstruktion des Eingangssignals ist entscheidend, wie die Folge* $(X_n)_{n\in\mathbb{N}_0}$ *entstanden ist. Wir müssen also wissen, welche Art von* Σ*-*Δ*-Konverter sich hinter der 1-Bit-Quantisierung von* U_0 *verbirgt. Aus einer Folge allein, welche nur zwei Werte besitzt, ist* U_0 *nicht rekonstruierbar, wie das die einfache gleitende Mittelung im Beispiel* 4.2.9 *in Abbildung* 14 *zeigt.*

5 Näherungslösung der Differenzengleichung eines Konverters zweiter Ordnung und abgeleitete Filterung

Im Gegensatz zu Kapitel 4, in dem eine Differentialgleichung die Grundlage zur Herleitung eines Filtervorschlages bildete, ist es hier das ursprüngliche DAWP (2.42) selbst. Alle Betrachtungen in diesem Kapitel setzen die Annahme

$$|U_1(t)| \leq \beta U_{ref} \tag{5.1}$$

voraus. Hinreichende Bedingungen für das Erfülltsein von (5.1) bei einem Konverter zweiter Ordnung sind im Satz 4.1.19 dargestellt. Wir sind leider nicht in der Lage, das DAWP (2.42) explizit zu lösen, d. h. eine Darstellung der Art

$$H_k = G(U_0(t), k) \tag{5.2}$$

anzugeben. Aus diesem Grund leiten wir eine Näherungslösung von (2.42) her und schätzen deren Fehler zur wahren Lösung ab.
Zunächst wollen wir jedoch eine weitere Vereinbarung treffen, welche den qualitativen Inhalt der Aussagen dieses Kapitels nicht ändert. Im weiteren gelte

$$\Delta t := 1 \quad \text{und} \quad t_0 := 0.$$

Somit ist dann immer $t_k = k$.

5.1 Aufgabenstellung

Wir untersuchen das DAWP

$$\begin{aligned} H_{k+2} - 2H_{k+1} + H_k &= \sigma\left(\psi_k - \left(2\beta + \frac{\alpha}{T_1}\right) H_{k+1} + \left(2\beta - \frac{\alpha}{T_1}\right) H_k\right), \\ H_0 &= H_1 = 0 \end{aligned} \tag{5.3}$$

mit

$$\psi_k = \frac{1}{T_1} \int\limits_0^k (k-\tau) \frac{U_0(\tau) + \alpha U_{ref}}{U_{ref}} \mathrm{d}\tau + \beta k, \quad k \in \mathbb{N}_0.$$

Aus unserer Annahme $|U_1(t)| \leq \beta U_{ref}$ für alle Zeiten ergibt sich für den konstanten Fall

$$U_0(t) \equiv (v - \alpha) U_{ref},$$

welchen wir im folgenden untersuchen wollen, dann mit Satz 4.1.19 die Bedingung

$$\frac{U_0(t) + \alpha U_{ref}}{U_{ref}} \equiv v \in [-\gamma + \alpha, \gamma + \alpha] \subset [0, 2\alpha]$$

sowie

$$\psi_k = \frac{1}{2T_1} k^2 v + \beta k.$$

Bemerkung 5.1.1 Im folgenden bezeichne Δ den Differenzenoperator mit

$$\Delta a_k := a_{k+1} - a_k$$

und Δ^2 den mit

$$\Delta^2 a_k := a_{k+2} - 2a_{k+1} + a_k.$$

Dann gilt

$$\Delta\left(a_k b_k\right) = \left(\Delta a_k\right) b_k + a_{k+1}\Delta b_k = a_{k+1}b_{k+1} - a_k b_k.$$

Weiterhin sei

$$V(t) := \frac{U_0(t)}{U_{ref}} + \alpha \in [-\gamma + \alpha, \gamma + \alpha].$$

Nun wollen wir ein Resultat beweisen, auf das wir später zurückgreifen.

Lemma 5.1.2 *Es sei $V(t) \in [-\gamma + \alpha, \gamma + \alpha]$ für alle Zeiten $t \geq t_0$. Dann gelten folgende Aussagen:*

$$a) \quad -\gamma + \alpha \leq \frac{2}{k^2}\int\limits_0^k (k-\tau)V(\tau)\mathrm{d}\tau \leq \gamma + \alpha,$$

$$b) \quad (-\gamma + \alpha)\left(k + \frac{1}{2}\right) \leq \Delta\left(\int\limits_0^k (k-\tau)V(\tau)\mathrm{d}\tau\right) \leq (\gamma + \alpha)\left(k + \frac{1}{2}\right),$$

$$c) \quad (-\gamma + \alpha) \leq \Delta^2\left(\int\limits_0^k (k-\tau)V(\tau)\mathrm{d}\tau\right) \leq (\gamma + \alpha).$$

Beweis: a) Dies folgt sofort aus

$$\frac{2}{k^2}\int\limits_0^k (k-\tau)\mathrm{d}\tau = 1.$$

b) Es gilt wegen (2.35)

$$\begin{aligned}
\Delta\left(\int\limits_0^k (k-\tau)V(\tau)\mathrm{d}\tau\right) &= \Delta\left(\int\limits_0^k \int\limits_0^\tau V(s)\mathrm{d}s\mathrm{d}\tau\right) \\
&= \int\limits_0^{k+1} (k+1-\tau)V(\tau)\mathrm{d}\tau - \int\limits_0^k (k-\tau)V(\tau)\mathrm{d}\tau \\
&= \int\limits_0^k (k+1-\tau)V(\tau)\mathrm{d}\tau - \int\limits_0^k (k-\tau)V(\tau)\mathrm{d}\tau + \int\limits_k^{k+1} (k+1-\tau)V(\tau)\mathrm{d}\tau \\
&= \int\limits_0^k V(\tau)\mathrm{d}\tau + \int\limits_k^{k+1} (k+1-\tau)V(\tau)\mathrm{d}\tau. \qquad (5.4)
\end{aligned}$$

Hieraus folgt wegen

$$\int\limits_0^k \mathrm{d}\tau + \int\limits_k^{k+1} (k+1-\tau)\mathrm{d}\tau = k + \frac{1}{2}$$

sofort die Behauptung in b).
c) Wegen (5.4) gilt

$$\begin{aligned}
\Delta^2 \left(\int_0^k (k-\tau)V(\tau)\mathrm{d}\tau \right) &= \Delta^2 \left(\int_0^k \int_0^\tau V(s)\mathrm{d}s\mathrm{d}\tau \right) \\
&= \int_{t_0}^{k+1} V(\tau)\mathrm{d}\tau - \int_0^k V(\tau)\mathrm{d}\tau + \int_{k+1}^{k+2} (k+2-\tau)V(\tau)\mathrm{d}\tau - \int_k^{k+1} (k+1-\tau)V(\tau)\mathrm{d}\tau \\
&= \int_k^{k+1} V(\tau)\mathrm{d}\tau - \int_k^{k+1} (k+1-\tau)V(\tau)\mathrm{d}\tau + \int_k^{k+1} [(k+2-(\tau+1))V(\tau+1)]\mathrm{d}\tau \\
&= \int_k^{k+1} (1+\tau-k-1)V(\tau)\mathrm{d}\tau + \int_k^{k+1} [(k+1-\tau)V(\tau+1)]\mathrm{d}\tau.
\end{aligned}$$

Hieraus folgt wegen

$$\int_k^{k+1} (1+\tau-k-1)\mathrm{d}\tau + \int_k^{k+1} (k+1-\tau)\mathrm{d}\tau = 1$$

sofort die Behauptung in c). □

Den weiteren Überlegungen liegt folgende Idee zugrunde:
Ein AD-Konverter zweiter Ordnung funktioniert „ordentlich“, wenn die Rückstellung durch die Rückkopplung immer gewährleistet ist. Dies bedeutet, daß der Ausdruck

$$\psi_k - \left(2\beta + \frac{\alpha}{T_1}\right) H_{k+1} + \left(2\beta - \frac{\alpha}{T_1}\right) H_k$$

für alle $k \in \mathbb{N}_0$ in einem fixen abgeschlossenen beschränkten Intervall um den Nullpunkt liegt. Je kleiner dieses Intervall ist, desto genauer wird durch den AD-Wandler das durch $t \longmapsto U_0(t)$ initiierte Signal beschrieben. Deshalb ist es sinnvoll, die Lösung $(w_k)_{k\in\mathbb{N}}$ des DAWP

$$\begin{aligned}
\left(2\beta + \frac{\alpha}{T_1}\right) w_{k+1} - \left(2\beta - \frac{\alpha}{T_1}\right) w_k &= \psi_k, \\
w_1 &= 0
\end{aligned} \tag{5.5}$$

zu berechnen und die Parameter α, β, T_1 so zu bestimmen, daß der Abstand

$$\sup_{k=1,2,\ldots} |H_k - w_k|$$

möglichst klein wird, wobei $(H_k)_{k\in\mathbb{N}}$ die Lösung des DAWP (5.3) ist.

5.2 Innere Differenzengleichung

Zunächst werden wir (5.5) für den Fall

$$V(t) \equiv v = \mathit{const.} \in [-\gamma + \alpha, \gamma + \alpha]$$

untersuchen, um dafür eine Lösung zu gewinnen. Diese Lösung besitzt dann den Charakter einer Näherungslösung von (5.5) für den allgemeinen Fall. Eine solche Näherungslösung erhalten wir durch Ersetzungregeln beim Übergang vom konstanten zum allgemeinen Fall.

5.2.1 Konstanter Fall

Wir wollen nun die Differenzengleichung (5.5) für den Fall $V(t) \equiv v = const.$ untersuchen.

Lemma 5.2.1 *Es sei*

$$2\beta T_1 > \alpha > 0.$$

Dann besitzt das DAWP

$$\begin{aligned} \left(2\beta + \frac{\alpha}{T_1}\right) w_{k+1} - \left(2\beta - \frac{\alpha}{T_1}\right) w_k &= \frac{k^2}{2T_1} v + \beta k \\ w_1 &= 0 \end{aligned} \tag{5.6}$$

die Lösung

$$\begin{aligned} w_k &= \frac{v}{4\alpha} k^2 + \frac{2\alpha\beta T_1 - (2\beta T_1 + \alpha)v}{4\alpha^2} k \\ &\quad + \frac{\beta T_1}{4\alpha^3}(2\beta T_1 + \alpha)(v - \alpha)\left[1 - \left(\frac{2\beta T_1 - \alpha}{2\beta T_1 + \alpha}\right)^k\right]. \end{aligned} \tag{5.7}$$

Insbesondere gilt $w_0 = H_0 = 0$.

Beweis: Die homogene Differenzengleichung

$$\left(2\beta + \frac{\alpha}{T_1}\right) w_{k+1} - \left(2\beta - \frac{\alpha}{T_1}\right) w_k = 0$$

hat mit dem Ansatz

$$w_k = \mu^k$$

die allgemeine Lösung

$$w_k = C\left(\frac{2\beta T_1 - \alpha}{2\beta T_1 + \alpha}\right)^k, C \in \mathbb{R}. \tag{5.8}$$

Über den Ansatz für die partikuläre Lösung

$$w_k^p = ak^2 + bk + c$$

erhalten wir mit

$$w_{k+1}^p = a(k+1)^2 + b(k+1) + c$$

die Bestimmungsgleichung

$$\left(2\beta + \frac{\alpha}{T_1}\right)\left(a(k+1)^2 + b(k+1) + c\right) - \left(2\beta - \frac{\alpha}{T_1}\right)\left(ak^2 + bk + c\right) = \frac{k^2}{2T_1} v + \beta k.$$

Der Koeffizientenvergleich

$$\begin{aligned} 2\frac{\alpha}{T_1} a &= \frac{v}{2T_1} \\ 4\beta a + 2\frac{\alpha}{T_1}(a + b) &= \beta \\ \frac{\alpha}{T_1}(a + b + 2c) + 2\beta(a + b) &= 0 \end{aligned}$$

liefert

$$\begin{aligned}
a &= \frac{v}{4\alpha}, \\
b &= \frac{2\alpha\beta T_1 - (2\beta T_1 + \alpha)v}{4\alpha^2}, \\
c &= \frac{\beta T_1}{4\alpha^3}(2\beta T_1 + \alpha)(v - \alpha).
\end{aligned}$$

Damit haben wir die allgemeine Lösung

$$w_k = \frac{v}{4\alpha}k^2 + \frac{2\alpha\beta T_1 - (2\beta T_1 + \alpha)v}{4\alpha^2}k + \frac{\beta T_1}{4\alpha^3}(2\beta T_1 + \alpha)(v - \alpha) + C\left(\frac{2\beta T_1 - \alpha}{2\beta T_1 + \alpha}\right)^k$$

für $k \in \mathbb{N}$ und $C \in \mathbb{R}$. Die Berücksichtigung der Anfangsbedingung

$$w_1 = 0$$

ergibt die Bestimmungsgleichung für C

$$\frac{v}{4\alpha} + \frac{2\alpha\beta T_1 - (2\beta T_1 + \alpha)v}{4\alpha^2} + \frac{\beta T_1}{4\alpha^3}(2\beta T_1 + \alpha)(v - \alpha) + C\left(\frac{2\beta T_1 - \alpha}{2\beta T_1 + \alpha}\right) = 0.$$

Deren Lösung ist

$$C = \frac{\beta T_1}{4\alpha^3}(2\beta T_1 + \alpha)(\alpha - v).$$

Somit hat das DAWP (5.6) die Lösung

$$w_k = \frac{v}{4\alpha}k^2 + \frac{2\alpha\beta T_1 - (2\beta T_1 + \alpha)v}{4\alpha^2}k + \frac{\beta T_1}{4\alpha^3}(2\beta T_1 + \alpha)(v - \alpha)\left[1 - \left(\frac{2\beta T_1 - \alpha}{2\beta T_1 + \alpha}\right)^k\right].$$

Hieraus ergibt sich die Behauptung. □

Bemerkung 5.2.2 Numerische Simulationen zeigen, daß der Ausdruck

$$\sup_{k=1,2,\ldots} |H_k - w_k|$$

beschränkt ist, falls

$$0 < \alpha < 2\beta T_1$$

gilt. Instabilitäten treten aber auch für Werte von α nahe 0 und $2\beta T_1$, wenn die Werte von v nahe bei $-\gamma + \alpha$ und $\gamma + \alpha$ liegen.

5.2.2 Allgemeiner Fall

Durch (5.7) aus Lemma 5.2.1 haben wir erkannt, wie die Struktur einer Näherungslösung von (5.5) aussehen muß. Mit dieser untersuchen wir nun den allgemeinen Fall.

Bemerkung 5.2.3 Es liegt nahe, für den allgemeinen Fall folgende Ersetzung durchzuführen

$$\begin{aligned} \frac{k^2}{2}v &\mapsto \int\limits_0^k (k-\tau)V(\tau)\mathrm{d}\tau, \\ kv &\mapsto \int\limits_0^k V(\tau)\mathrm{d}\tau, \\ v &\mapsto \int\limits_k^{k+1} V(\tau)\mathrm{d}\tau, \end{aligned} \tag{5.9}$$

da für konstantes $V(t) \equiv v$ genau die linken Seiten entstehen. Dann ergibt sich für eine Näherungslösung von (5.5) die Vermutung

$$\begin{aligned} \tilde{w}_k &= \frac{1}{2\alpha}\int\limits_0^k (k-\tau)V(\tau)\mathrm{d}\tau - \frac{1}{4\alpha^2}\left[2\alpha\beta T_1 k - (2\beta T_1+\alpha)\int\limits_0^k V(\tau)\mathrm{d}\tau\right] \\ &\quad + \frac{\beta T_1}{4\alpha^3}(2\beta T_1+\alpha)\left(\int\limits_k^{k+1} V(\tau)\mathrm{d}\tau - \alpha\right)\left(1-\left(\frac{2\beta T_1-\alpha}{2\beta T_1+\alpha}\right)^k\right), k \in \mathbb{N}. \end{aligned} \tag{5.10}$$

Bemerkung 5.2.4 Wie bereits festgestellt, sind wir nicht in der Lage, eine Lösung von (2.42) in der Form (5.2) anzugeben. Dies liegt im wesentlichen an der Unbestimmtheit von p_k, welches gerade die Unstetigkeit der Heavisidefunktion auf der rechten Seite von (2.42) ausmacht. Deshalb sind wir daran interessiert, wenigstens eine Näherungslösung von (2.42) anzugeben. Beim DAWP (5.5) liegt die Sache etwas anders. Hier können wir die Lösung explizit angeben, sie lautet

$$w_k = \frac{T_1}{2\beta T_1+\alpha}\sum_{j=2}^{k}\left(\frac{2\beta T_1-\alpha}{2\beta T_1+\alpha}\right)^{k-j}\psi_{j-1} \tag{5.11}$$

$$= \frac{T_1}{2\beta T_1+\alpha}\sum_{j=2}^{k}\left(\frac{2\beta T_1-\alpha}{2\beta T_1+\alpha}\right)^{k-j}\left(\frac{1}{T_1}\int\limits_0^{j-1}(j-1-\tau)\left(\frac{U_0(\tau)}{U_{ref}}+\alpha\right)\mathrm{d}\tau + \beta(j-1)\right).$$

Die in der Lösung enthaltene Summe ist kein geeigneter Term, um die Grundlage für eine Rekonstruktion von U_0 in einem Filter zu bilden. Deshalb interessiert uns die Näherungslösung (5.10) des DAWP (5.5), von der wir wissen möchten, wie weit sie von der wahren Lösung entfernt ist. Um diese Abschätzung zu bestimmen, haben wir zwei Möglichkeiten:

I Löse das DAWP (5.5), wie durch (5.11) geschehen. Nun bilde die Differenz zwischen (5.11) und (5.10) und schätze diese nach oben ab.

II Setze die Näherung (5.10) in die linke Seite von (5.5) ein. Die dabei entstehende rechte Seite bezeichnen wir mit φ_k. Schätze nun die Differenz der rechten Seite von (5.5) ψ_k und φ_k ab. Nun schätze

$$w_k^* := w_k - \tilde{w}_k, \quad k \in \mathbb{N}.$$

als Lösung des DAWP

$$\begin{aligned} \left(2\beta+\frac{\alpha}{T_1}\right)w_{k+1}^* - \left(2\beta-\frac{\alpha}{T_1}\right)w_k^* &= \psi_k - \varphi_k, \\ w_1^* &= 0 \end{aligned} \tag{5.12}$$

ab. Diese Vorgehensweise ist korrekt, da die Differenzengleichung linear ist. Aus gleichem Grund existiert auch eine Lösung w^* von (5.12), welche gerade die Differenz $w_k - \tilde{w}_k$ ist.

Bemerkung 5.2.5 Bevor wir $\tilde{w}_k$ in die Gleichung (5.5) einsetzen, untersuchen wir einige Ausdrücke auf Ersetzbarkeit bei Erhöhung des Indexes. Dabei gilt

$$\begin{aligned}
\int_0^{k+1} (k+1-\tau)V(\tau)\mathrm{d}\tau &= \int_0^{k+1} (k-\tau)V(\tau)\mathrm{d}\tau + \int_0^{k+1} V(\tau)\mathrm{d}\tau \\
&= \int_0^{k} (k-\tau)V(\tau)\mathrm{d}\tau + \int_0^{k} V(\tau)\mathrm{d}\tau + \int_k^{k+1} (k-\tau)V(\tau)\mathrm{d}\tau + \int_k^{k+1} V(\tau)\mathrm{d}\tau \\
&=: A + B + C + D
\end{aligned}$$

$$\int_0^{k+1} V(\tau)\mathrm{d}\tau = \int_0^{k} V(\tau)\mathrm{d}\tau + \int_k^{k+1} V(\tau)\mathrm{d}\tau = B + D$$

$$\int_{k+1}^{k+2} V(\tau)\mathrm{d}\tau = \int_{t_k}^{t_{k+1}} V(\tau)\mathrm{d}\tau + \int_k^{k+1} (V(\tau+1) - V(\tau))\mathrm{d}\tau =: D + R$$

$$\psi_k = \frac{A}{T_1} + \beta k.$$

Nun sind wir in der Lage, den Fehler der rechten Seite von (5.5) zu berechnen, wenn wir die Näherunglösung (5.10) zugrunde legen.

Lemma 5.2.6 *Es gelte für alle Zeiten* $t \geq 0$

$$|V\,(t+1) - V(t)| \leq L < +\infty.$$

Dann gilt

$$|F_k| := |\varphi_k - \psi_k| \leq \left(\frac{2\beta T_1 + \alpha}{8\alpha T_1} + \frac{\beta(2\beta T_1 + \alpha)^2}{4\alpha^3} \right) L, \quad \forall k \in \mathbb{N}.$$

Beweis: Aus der Bemerkung 5.2.5 folgt

$$\begin{aligned}
\tilde{w}_k &= \frac{A}{2\alpha} + \frac{1}{4\alpha^2}\left[2\alpha\beta T_1 k - (2\beta T_1 + \alpha)B\right] \\
&\quad + \frac{\beta T_1}{4\alpha^3}(2\beta T_1 + \alpha)(D - \alpha)\left(1 - \left(\frac{2\beta T_1 - \alpha}{2\beta T_1 + \alpha}\right)^k\right),
\end{aligned}$$

$$\begin{aligned}
\tilde{w}_{k+1} &= \frac{A+B+C+D}{2\alpha} + \frac{1}{4\alpha^2}\left[2\alpha\beta T_1(k+1) - (2\beta T_1 + \alpha)(B+D)\right] \\
&\quad + \frac{\beta T_1}{4\alpha^3}(2\beta T_1 + \alpha)(D + R - \alpha)\left(1 - \left(\frac{2\beta T_1 - \alpha}{2\beta T_1 + \alpha}\right)^k\right).
\end{aligned}$$

Wir berechnen damit nun den Fehler

$$\left(2\beta + \frac{\alpha}{T_1}\right)\tilde{w}_{k+1} - \left(2\beta - \frac{\alpha}{T_1}\right)\tilde{w}_k - \left(\frac{A}{T_1} + \beta k\right)$$
$$= \frac{1}{T_1}\left[(2\beta T_1 + \alpha)\tilde{w}_{k+1} - (2\beta T_1 - \alpha)\tilde{w}_k - A - k\beta T_1\right]$$
$$= \frac{1}{T_1}\Bigg[\frac{2\beta T_1 + \alpha}{2\alpha}(A+B+C+D) + \frac{2\beta T_1 + \alpha}{4\alpha^2}\left[2\alpha\beta T_1(k+1) - (2\beta T_1 + \alpha)(B+D)\right]$$
$$+\frac{(2\beta T_1 + \alpha)^2\beta T_1}{4\alpha}(D+R-\alpha)\left(1 - \left(\frac{2\beta T_1 - \alpha}{2\beta T_1 + \alpha}\right)^{k+1}\right)$$
$$-\frac{2\beta T_1 - \alpha}{2\alpha}A - \frac{2\beta T_1 + \alpha}{4\alpha^2}\left[2\alpha\beta T_1 k - (2\beta T_1 + \alpha)B\right]$$
$$-\frac{(2\beta T_1 + \alpha)^2\beta T_1}{4\alpha}(D-\alpha)\left(1 - \left(\frac{2\beta T_1 - \alpha}{2\beta T_1 + \alpha}\right)^{k}\right) - A - k\beta T_1\Bigg]$$
$$= \frac{2\beta T_1 + \alpha}{4\alpha T_1}(2C+D) + \frac{\beta R(2\beta T_1 + \alpha)^2}{4\alpha^3}\left(1 - \left(\frac{2\beta T_1 - \alpha}{2\beta T_1 + \alpha}\right)^{k+1}\right).$$

Wegen

$$|V(t+1) - V(t)| \le L$$

folgen die Abschätzungen

$$\begin{aligned}
2C + D &= \int_k^{k+1} \left[2(k-\tau)V(\tau) + V(\tau)\right] \mathrm{d}\tau \\
&= \int_0^1 \left[2(k-(\tau+k))V(\tau+k) + V(\tau+k)\right] \mathrm{d}\tau \\
&= \int_0^1 (1-2\tau)V(\tau+k)\,\mathrm{d}\tau \\
&= \int_0^1 (1-2\tau)V(k)\mathrm{d}\tau + \int_0^1 (1-2\tau)\left(V(\tau+k) - V(k)\right)\mathrm{d}\tau \\
&= \int_0^1 (1-2\tau)\left(V(\tau+k) - V(k)\right)\mathrm{d}\tau, \\
|2C + D| &\le \int_0^1 |1-2\tau|\,|V(\tau+k) - V(k)|\,\mathrm{d}\tau \\
&\le L\int_0^1 |1-2\tau|\,\mathrm{d}\tau = \frac{L}{2}, \\
|R| &= \left|\int_k^{k+1} \left(V(t+1) - V(t)\right)\mathrm{d}\tau\right| \le L.
\end{aligned}$$

Also gilt für den Term

$$F_k = \frac{2\beta T_1 + \alpha}{4\alpha T_1}(2C + D) + \frac{\beta R(2\beta T_1 + \alpha)^2}{4\alpha^3}\left(1 - \left(\frac{2\beta T_1 - \alpha}{2\beta T_1 + \alpha}\right)^{k+1}\right) \tag{5.13}$$

unter der Bedingung

$$2\beta T_1 > \alpha > 0$$

für alle $k \in \mathbb{N}_0$ die Abschätzung

$$|F_k| \leq \left(\frac{2\beta T_1 + \alpha}{8\alpha T_1} + \frac{\beta(2\beta T_1 + \alpha)^2}{4\alpha^3}\right) L.$$

Damit ist alles bewiesen. □

Beispiel 5.2.7 Eine volle Sinusschwingung werde 1000-fach abgetastet. Es sei $U_{ref} = 1$. Dann haben wir wegen

$$U_0(t) = \sin\left(\frac{2\pi}{1000}t\right)$$

eine Lipschitzkonstante

$$L = \frac{2\pi}{1000} \leq 0.0063.$$

Wir erhalten wegen

$$\Sigma\left(\alpha, \beta, T_1\right) := 0.0063\left(\frac{2\beta T_1 + \alpha}{8\alpha T_1} + \frac{\beta(2\beta T_1 + \alpha)^2}{4\alpha^3}\right)$$

dann den folgenden Fehler, der in Abbildung 15 dargestellt ist.

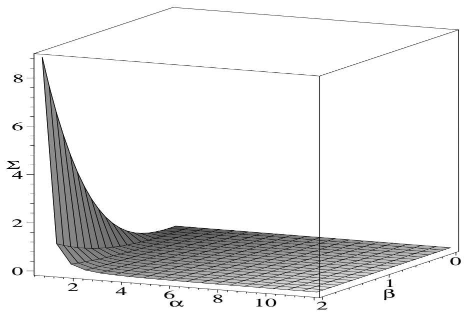


Abbildung 15: Plot der Fläche $\Sigma\left(\alpha, \beta, T_1\right)$ für $T_1 = 6$, $\alpha \in [0.6, 12]$ und $\beta \in [0, 2]$

5.2.3 Fehler der approximativen Lösung der inneren Differenzengleichung

In Lemma 5.2.6 hatten wir für den Fehler $\varphi_k - \psi_k$ eine gleichmäßige obere Schranke gefunden. Nun wollen wir entsprechend des Teils **II** der Bemerkung 5.2.4 untersuchen, wie groß der Fehler $\tilde{w}_k - w_k$ der Näherungslösung (5.10) zur wahren Lösung des Anfangswertproblems (5.5) ist. Dabei gilt folgendes Resultat.

Satz 5.2.8 *Für*

$$2\beta T_1 > \alpha > 0$$

$$-\gamma + \alpha \leq V(t) \leq \gamma + \alpha$$

$$|V(t+1) - V(t)| \leq L, \quad t \geq 0$$

hat die Differenzengleichung

$$\begin{aligned}\left(2\beta + \frac{\alpha}{T_1}\right) w_{k+1} - \left(2\beta - \frac{\alpha}{T_1}\right) w_k &= \frac{1}{T_1}\int_0^k (k-\tau)V(\tau)\mathrm{d}\tau + \beta k \\ w_0 &= 0\end{aligned}$$

die approximative Lösung

$$\begin{aligned}\tilde{w}_k &= \frac{1}{2\alpha}\int_0^k (k-\tau)V(\tau)\mathrm{d}\tau - \frac{1}{4\alpha^2}\left[2\alpha\beta T_1 k - (2\beta T_1 + \alpha)\int_0^k V(\tau)\mathrm{d}\tau\right] \\ &\quad + \frac{\beta T_1}{4\alpha^3}(2\beta T_1 + \alpha)\left(\int_k^{k+1} V(\tau)\mathrm{d}\tau - \alpha\right)\left(1 - \left(\frac{2\beta T_1 - \alpha}{2\beta T_1 + \alpha}\right)^k\right).\end{aligned}$$

Zur wahren Lösung w_k mit

$$w_k = \tilde{w}_k + w_k^*$$

gilt die Fehlerabschätzung

$$|w_k^*| \leq \frac{T_1 F}{2\alpha}$$

mit

$$\begin{aligned}F &= \sup_{k=0,1,2,\ldots}\left(\frac{2\beta T_1 + \alpha}{4\alpha T_1}(2C + D) + \frac{\beta R(2\beta T_1 + \alpha)^2}{4\alpha^3}\left(1 - \left(\frac{2\beta T_1 - \alpha}{2\beta T_1 + \alpha}\right)^{k+1}\right)\right) \\ &\leq \left(\frac{2\beta T_1 + \alpha}{8\alpha T_1} + \frac{\beta(2\beta T_1 + \alpha)^2}{4\alpha^3}\right) L\end{aligned}$$

wobei gilt

$$\begin{aligned}|2C + D| &\leq \sup_{k=0,1,2,\ldots}\left|\int_k^{k+1} \left[2(k-\tau)V(\tau) + V(\tau)\right]\mathrm{d}\tau\right| \\ &\leq L\int_0^1 |1 - 2\tau|\,\mathrm{d}\tau = \frac{L}{2}, \\ |R| &= \sup_{k=0,1,2,\ldots}\left|\int_k^{k+1} (V(t+1) - V(t))\,dt\right| \leq L.\end{aligned}$$

Beweis: Für den Beweis betrachten wir jetzt die Differenzengleichung

$$\left(2\beta + \frac{\alpha}{T_1}\right) w_{k+1}^* - \left(2\beta - \frac{\alpha}{T_1}\right) w_k^* = F_k \tag{5.14}$$

unter der Bedingung

$$|F_k| \leq F < \infty, \ k \in \mathbb{N}_0,$$

wobei F_k wie in Lemma 5.2.6 definiert seien. Nach Voraussetzung ist dann auch die Existenz eines endlichen F gesichert. Wir bestimmen die Lösung durch Variation der Konstanten

$$\begin{aligned} w_k^* &= C_k \left(\frac{2\beta T_1 - \alpha}{2\beta T_1 + \alpha}\right)^k, \\ w_{k+1}^* &= C_{k+1} \left(\frac{2\beta T_1 - \alpha}{2\beta T_1 + \alpha}\right)^{k+1}. \end{aligned}$$

Aus

$$\left(2\beta + \frac{\alpha}{T_1}\right) C_{k+1} \left(\frac{2\beta T_1 - \alpha}{2\beta T_1 + \alpha}\right)^{k+1} - \left(2\beta - \frac{\alpha}{T_1}\right) C_k \left(\frac{2\beta T_1 - \alpha}{2\beta T_1 + \alpha}\right)^k = F_k$$

folgt dann

$$C_{k+1} - C_k = \frac{T_1 \left(2\beta T_1 + \alpha\right)^k}{\left(2\beta T_1 - \alpha\right)^{k+1}} F_k.$$

Dadurch ergibt sich

$$\begin{aligned} \sum_{k=0}^{n-1} \left(C_{k+1} - C_k\right) &= \sum_{k=0}^{n-1} \frac{\left(2\beta T_1 + \alpha\right)^k}{\left(2\beta T_1 - \alpha\right)^{k+1}} T_1 F_k \\ C_n &= C_0 + \sum_{k=0}^{n-1} \frac{\left(2\beta T_1 + \alpha\right)^k}{\left(2\beta T_1 - \alpha\right)^{k+1}} T_1 F_k \end{aligned}$$

Dann folgt nach Einsetzen für $n = 1, 2, 3, ...$

$$\begin{aligned} w_n^* &= C_n \left(\frac{2\beta T_1 - \alpha}{2\beta T_1 + \alpha}\right)^n \\ &= \left(C_0 + \sum_{k=0}^{n-1} \frac{\left(2\beta T_1 + \alpha\right)^k}{\left(2\beta T_1 - \alpha\right)^{k+1}} T_1 F_k\right) \left(\frac{2\beta T_1 - \alpha}{2\beta T_1 + \alpha}\right)^n \\ &= C_0 \left(\frac{2\beta T_1 - \alpha}{2\beta T_1 + \alpha}\right)^n + \frac{1}{2\beta T_1 + \alpha} \sum_{k=0}^{n-1} \frac{\left(2\beta T_1 - \alpha\right)^{n-1-k}}{\left(2\beta T_1 + \alpha\right)^{n-1-k}} T_1 F_k. \end{aligned}$$

Wir benötigen für die Konstante C_0 nur noch eine spezielle Lösung. Mit $w_0 = 0$ folgt

$$0 = C_0.$$

Also erhalten wir durch die Abschätzung einer geometrischen Reihe

$$|w_n^*| \leq \frac{T_1 F}{2\beta T_1 + \alpha} \sum_{k=0}^{n-1} \frac{\left(2\beta T_1 - \alpha\right)^{n-1-k}}{\left(2\beta T_1 + \alpha\right)^{n-1-k}}$$

$$
\begin{aligned}
&= \frac{T_1 F}{2\beta T_1 + \alpha} \sum_{k=0}^{n-1} \left(\frac{2\beta T_1 - \alpha}{2\beta T_1 + \alpha} \right)^k \\
&= \frac{T_1 F}{2\beta T_1 + \alpha} \frac{1 - \left(\frac{2\beta T_1 - \alpha}{2\beta T_1 + \alpha} \right)^n}{1 - \left(\frac{2\beta T_1 - \alpha}{2\beta T_1 + \alpha} \right)} \\
&\leq \frac{T_1 F}{2\alpha}.
\end{aligned}
$$

Damit ist alles gezeigt. □

Bemerkung 5.2.9 Den Beweis für die Richtigkeit von (5.11) haben wir somit erst jetzt bei der Berechnung von w_k^* in (5.14) anstelle von w_k in (5.5) - dort mit anderer rechter Seite - geführt. Ein wichtiges Merkmal unserer Näherungslösung ist die Tatsache, daß der Fehler mit zunehmendem Index k nicht größer wird. Somit bleibt gewissermaßen $\tilde{w}_k$ für alle Zeiten in der „Nähe“ von w_k. Mit ähnlichen Überlegungen hoffen wir, eine Näherungslösung von (2.42) zu bestimmen, die auch diese gute Eigenschaft besitzt.

5.3 Zur Differenzengleichung eines Konverters zweiter Ordnung

Nun wollen wir uns dem DAWP (2.42) zuwenden. Da gerade dadurch das Verhalten eines Konverters zweiter Ordnung beschrieben wird, ist klar, daß deren Lösung H_k ganzzahlig ist. Wir setzen wiederum das Erfülltsein von (5.1) voraus. Denn in diesem Fall können wir die unstetige rechte Seite der Differenzengleichung in einen stetigen Anteil und eine beschränkte Unbestimmtheit zerlegen. Mit den gleichen Berechnungen wie im Abschnitt 5.2 geben wir eine Näherung von H_k an und schätzen deren Fehler ab.

5.3.1 Vorbetrachtungen

Wir wollen nun das DAWP (2.42) untersuchen. Wir nehmen wiederum an, daß für den Schwellenwert des Komparators $s = 0$ gilt. Weiterhin seien die Anfangswerte $U_1(t_0) = 0$ sowie $U_2(t_0) = 0$. Wegen $|U_1(t)| \leq \beta U_{ref}$ gilt nach Teil b) des Lemmas 4.1.12

$$
|U_2(t)| \leq \frac{2\beta \Delta t U_{ref}}{T_2}.
$$

Hieraus, aus Bemerkung 2.2.1 und $s = 0$ ergibt sich für $\Delta t = 1$ und $t_0 = 0$ mit $t_k = k$ dann

$$
\begin{aligned}
\Delta^2 H_k &= \sigma(U_2(k)) = \sigma\left(\frac{T_2}{2\beta U_{ref}} U_2(k) \right) = \frac{T_2}{2\beta \Delta t U_{ref}} U_2(k) + p_k \\
&= \frac{1}{2\beta T_1 U_{ref}} \int_0^k (k-\tau) U_0(\tau) \mathrm{d}\tau - \frac{\alpha}{2\beta T_1}(H_{k+1} + H_k) - (H_{k+1} - H_k) \\
&\quad + \frac{\alpha}{4\beta T_1} k^2 + \frac{k}{2} + p_k \\
&= \frac{1}{2\beta T_1} \int_0^k (k-\tau) V(\tau) \mathrm{d}\tau - \frac{\alpha}{2\beta T_1}(H_{k+1} + H_k) - (H_{k+1} - H_k) \\
&\quad + \frac{k}{2} + p_k \quad \text{mit} \quad p_k \in [0,1].
\end{aligned}
$$

In der ingenieurtechnischen Literatur wird ein Σ-Δ-Konverter linearisiert, indem der Komparator durch einen Summierer ersetzt wird. Die Unbestimmtheit p_k für $k \in \mathbb{N}_0$ entspricht genau dem Quantisierungsfehler, welcher in diesem Summierer mit $U_2(k)$ addiert wird, um $X_k \in \{0,1\}$ zu bilden. Weiteres dazu in Kapitel 8.
Wir untersuchen im folgenden das DAWP

$$\begin{aligned} H_{k+2} - \left(1 - \frac{\alpha}{2\beta T_1}\right) H_{k+1} + \frac{\alpha}{2\beta T_1} H_k &= \frac{1}{2\beta T_1} \int\limits_0^k (k-\tau) V(\tau) \mathrm{d}\tau + \frac{k}{2} + p_k, \\ H_0 &= H_1 = 0, \qquad k \in \mathbb{N}_0, \end{aligned} \tag{5.15}$$

welches aus (2.30), (2.31) und (2.42) folgt. Wir benutzen hier wiederum die Möglichkeit **II** aus der Bemerkung 5.2.4, da die wahre Lösung H_k von (5.15) von ähnlicher Gestalt wie (5.11) ist. Diese ist wegen ihrer Komplexität und des Enthaltenseins der Unbestimmtheit p_k nicht für unsere Diskussion geeignet. Wir lösen zunächst die homogene Gleichung

$$H_{k+2} - \left(1 - \frac{\alpha}{2\beta T_1}\right) H_{k+1} + \frac{\alpha}{2\beta T_1} H_k = 0.$$

Mit dem Ansatz

$$H_k = \mu^k$$

folgen aus

$$\mu^2 - \left(1 - \frac{\alpha}{2\beta T_1}\right) \mu + \frac{\alpha}{2\beta T_1} = 0$$

die Lösungen

$$\mu_{\pm} = \frac{(2\beta T_1 - \alpha) \pm \sqrt{(2\beta T_1 - \alpha)^2 - 8\alpha\beta T_1}}{4\beta T_1}.$$

Damit lautet die Lösung der homogenen Differenzengleichung

$$H_k = C\mu_+^k + D\mu_-^k. \tag{5.16}$$

Es können somit drei Fälle auftreten:
komplexe Lösungen für

$$\frac{\alpha}{\beta T_1} \in \left(6 - 4\sqrt{2}, 6 + 4\sqrt{2}\right),$$

Doppellösungen für

$$\frac{\alpha}{\beta T_1} \in \left\{6 - 4\sqrt{2}, 6 + 4\sqrt{2}\right\}$$

und zwei verschiedene relle Lösungen für

$$\frac{\alpha}{\beta T_1} \in \left(0, 6 - 4\sqrt{2}\right) \cup \left(6 + 4\sqrt{2}, \infty\right).$$

Wegen $H_k \in \mathbb{N}_0$ und $\dfrac{\alpha}{\beta T_1} > 0$ interessieren wir uns nur für die Intervalle

$$\frac{\alpha}{\beta T_1} \in (0, 6 - 4\sqrt{2}] \cup [6 + 4\sqrt{2}, +\infty).$$

5.3.2 Konstanter Fall

Wir gehen ähnlich wie in Kapitel 5.2 vor und betrachten zunächst den Fall, daß $V(t) \equiv v = const.$ sei. Wir wollen das DAWP

$$\begin{aligned} H_{k+2} - \left(1 - \frac{\alpha}{2\beta T_1}\right) H_{k+1} + \frac{\alpha}{2\beta T_1} H_k &= \frac{k^2 v}{4\beta T_1} + \frac{k}{2} + p_k, \\ H_0 &= H_1 = 0, \qquad k \in \mathbb{N}_0 \end{aligned} \tag{5.17}$$

untersuchen. Dazu zerlegen wir es in die zwei Probleme

$$\begin{aligned} H^I_{k+2} - \left(1 - \frac{\alpha}{2\beta T_1}\right) H^I_{k+1} + \frac{\alpha}{2\beta T_1} H^I_k &= \frac{k^2 v}{4\beta T_1} + \frac{k}{2}, \\ H^I_0 &= H^I_1 = 0, \qquad k \in \mathbb{N}_0 \end{aligned} \tag{5.18}$$

und

$$\begin{aligned} H^{II}_{k+2} - \left(1 - \frac{\alpha}{2\beta T_1}\right) H^{II}_{k+1} + \frac{\alpha}{2\beta T_1} H^{II}_k &= p_k, \\ H^{II}_0 &= H^{II}_1 = 0, \qquad k \in \mathbb{N}_0 . \end{aligned} \tag{5.19}$$

Dann stellt sich die allgemeine Lösung von (5.17) wie folgt dar

$$H_k = H^I_k + H^{II}_k. \tag{5.20}$$

Der Ansatz

$$\begin{aligned} H^I_k &= ak^2 + bk + c \\ H^I_{k+1} &= a(k+1)^2 + b(k+1) + c \\ H^I_{k+2} &= a(k+2)^2 + b(k+2) + c \end{aligned}$$

ergibt für die Differenzengleichung in (5.18) dann

$$2\alpha a k^2 + (4a\beta T_1 + 2\alpha(a+b))k + \alpha(2c + a + b) + 2\beta T_1(3a + b) = \frac{v}{2}k^2 + \beta T_1 k.$$

Der Koeffizientenvergleich liefert

$$\begin{aligned} 2\alpha a &= \frac{v}{2}, \\ 4a\beta T_1 + 2\alpha(a+b) &= \beta T_1, \\ \alpha(2c + a + b) + 2\beta T_1(3a + b) &= 0, \end{aligned}$$

und damit ist

$$\begin{aligned} a &= \frac{v}{4\alpha}, \\ b &= \frac{1}{4\alpha^2}(2\alpha\beta T_1 - v(2\beta T_1 + \alpha)), \\ c &= \frac{\beta T_1}{4\alpha^3}((2\beta T_1 - \alpha)v - (2\beta T_1 + \alpha)\alpha). \end{aligned}$$

Also folgt aus dem Ansatz für H^I_k

$$H^I_k = \frac{v}{4\alpha}k^2 + \frac{1}{4\alpha^2}(2\alpha\beta T_1 - v(2\beta T_1 + \alpha))k + \frac{\beta T_1}{4\alpha^3}((2\beta T_1 - \alpha)v - (2\beta T_1 + \alpha)\alpha). \tag{5.21}$$

Diese Lösung besitzt große Ähnlichkeit mit w_k aus Lemma 5.2.1. Aus (5.7) und (5.21) folgt dann

$$H^I_k = w_k + \frac{\beta T_1}{2\alpha} + \frac{\beta T_1}{4\alpha^3}(2\beta T_1 - \alpha)(v - \alpha)\left(\frac{2\beta T_1 - \alpha}{2\beta T_1 + \alpha}\right)^k .$$

Für $2\beta T_1 \geq \alpha > 0$ gilt

$$\left(\frac{2\beta T_1 - \alpha}{2\beta T_1 + \alpha}\right)^k \to 0 \text{ für } k \to +\infty.$$

Also ist

$$H_k^I \approx w_k + \frac{\beta T_1}{2\alpha}$$

für große $k \in \mathbb{N}_0$. Wir berechnen jetzt die Konstanten C und D der homogenen Lösung (5.16) gemäß den Anfangsbedingungen und der speziellen Lösung (5.21). Die allgemeine Lösung von (5.18) lautet

$$\begin{aligned} H_k^I &= C\mu_+^k + D\mu_-^k + \frac{v}{4\alpha}k^2 + \frac{1}{4\alpha^2}(2\alpha\beta T_1 - v(2\beta T_1 + \alpha))k \\ &\quad + \frac{\beta T_1}{4\alpha^3}((2\beta T_1 - \alpha)v - (2\beta T_1 + \alpha)\alpha) \end{aligned} \tag{5.22}$$

mit den Anfangsbedingungen

$$\begin{aligned} H_0^I &= 0 = C + D + \frac{\beta T_1}{4\alpha^3}((2\beta T_1 - \alpha)v - (2\beta T_1 + \alpha)\alpha) \\ H_1^I &= 0 = C\mu_+ + D\mu_- + \frac{v}{4\alpha} + \frac{1}{4\alpha^2}(2\alpha\beta T_1 - v(2\beta T_1 + \alpha)) \\ &\quad + \frac{\beta T_1}{4\alpha^3}((2\beta T_1 - \alpha)v - (2\beta T_1 + \alpha)\alpha). \end{aligned}$$

Somit gilt für die beiden Konstanten

$$\begin{aligned} C &= \frac{\beta T_1}{8\alpha^3}\left[\alpha(\alpha + v) + 2\beta T_1(\alpha - v)\right] \\ &\quad - \frac{\beta T_1}{8\alpha^3\sqrt{(2\beta T_1 - \alpha)^2 - 8\alpha\beta T_1}}\left[4\beta^2 T_1^2(v - \alpha) + 4\alpha\beta T_1(\alpha - 2v) - \alpha^2(\alpha + v)\right], \end{aligned} \tag{5.23}$$

$$\begin{aligned} D &= \frac{\beta T_1}{8\alpha^3}\left[\alpha(\alpha + v) + 2\beta T_1(\alpha - v)\right] \\ &\quad + \frac{\beta T_1}{8\alpha^3\sqrt{(2\beta T_1 - \alpha)^2 - 8\alpha\beta T_1}}\left[4\beta^2 T_1^2(v - \alpha) + 4\alpha\beta T_1(\alpha - 2v) - \alpha^2(\alpha + v)\right]. \end{aligned} \tag{5.24}$$

Wir zeichnen C

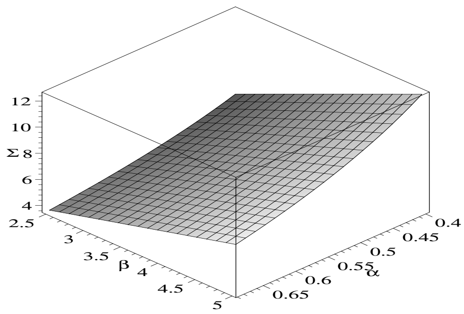


Abbildung 16: C für $T_1 = v = 1$, $\alpha \in (2.5, 5)$ und $\beta \in (0.05, 03)$

sowie D

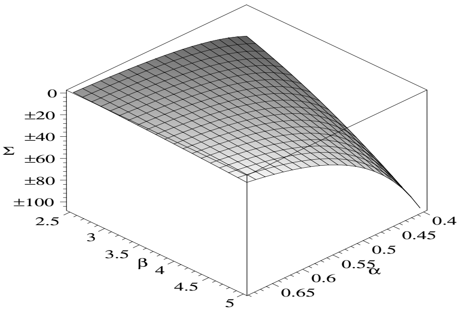


Abbildung 17: D für $T_1 = v = 1$, $\alpha \in (2.5, 5)$ und $\beta \in (0.05, 0.3)$

Nun wollen wir die Lösung H_k^{II} von (5.19) berechnen und abschätzen, um den Fehler zu bestimmen, den wir machen, wenn wir H_k^I als Näherung von H_k annehmen. Unter Benutzung der z-Transformation [1, Ch.2] ergibt sich für die Lösung H_k^{II} von (5.19) dann

$$\begin{aligned} H_k^{II} &= \frac{1}{\sqrt{(1-\lambda)^2 - 4\lambda}} \sum_{n=2}^{k} \left[\left(\frac{2\lambda}{1-\lambda-\sqrt{(1-\lambda)^2-4\lambda}} \right)^{k-n+1} \right. \\ &\qquad \left. - \left(\frac{2\lambda}{1-\lambda+\sqrt{(1-\lambda)^2-4\lambda}} \right)^{k-n+1} \right] p_{k-2} \\ &= \frac{1}{\sqrt{(1-\lambda)^2 - 4\lambda}} \sum_{n=0}^{k-1} \left[\left(\frac{2\lambda}{1-\lambda-\sqrt{(1-\lambda)^2-4\lambda}} \right)^{n} \right. \\ &\qquad \left. - \left(\frac{2\lambda}{1-\lambda+\sqrt{(1-\lambda)^2-4\lambda}} \right)^{n} \right] p_{k-n-1} \end{aligned} \tag{5.25}$$

mit

$$\lambda := \frac{\alpha}{2\beta T_1}.$$

Für $\lambda \in (0, 3-2\sqrt{2}] \cup [3+2\sqrt{2}, +\infty)$ ist der Term $\sqrt{(1-\lambda)^2 - 4\lambda}$ reell. Definieren wir

$$f^+(\lambda) := \frac{2\lambda}{1-\lambda+\sqrt{(1-\lambda)^2-4\lambda}} \quad \text{und} \quad f^-(\lambda) := \frac{2\lambda}{1-\lambda-\sqrt{(1-\lambda)^2-4\lambda}},$$

dann gilt

$$\begin{aligned} f^+(\lambda) &\in \left(0, \frac{3-2\sqrt{2}}{\sqrt{2}-1}\right) \quad \text{für} \quad \lambda \in (0, 3-2\sqrt{2}), \\ f^-(\lambda) &\in \left(\frac{3-2\sqrt{2}}{\sqrt{2}-1}, 1\right) \quad \text{für} \quad \lambda \in (0, 3-2\sqrt{2}), \\ f^+(\lambda) &\in \left(-\infty, \frac{-3-2\sqrt{2}}{\sqrt{2}+1}\right) \quad \text{für} \quad \lambda \in (3+2\sqrt{2}, +\infty), \\ f^-(\lambda) &\in \left(\frac{-3-2\sqrt{2}}{\sqrt{2}+1}, -1\right) \quad \text{für} \quad \lambda \in (3+2\sqrt{2}, +\infty). \end{aligned}$$

Somit ist

$$\frac{1}{\sqrt{(1-\lambda)^2-4\lambda}}\sum_{n=0}^{k-1}\left[\left(f^-(\lambda)\right)^n-\left(f^+(\lambda)\right)^n\right]$$

alternierend divergent für $\lambda \in (3+2\sqrt{2},+\infty)$ und wegen

$$1-\lambda+\sqrt{(1-\lambda)^2-4\lambda} > 1-\lambda-\sqrt{(1-\lambda)^2-4\lambda} > 0$$

nichtnegativ, monoton wachsend und konvergent für $\lambda \in (0, 3-2\sqrt{2})$. Es gilt somit für $\lambda \in (0, 3-2\sqrt{2})$ wegen (5.25)

$$\begin{aligned}
H_k^{II} &\leq \frac{1}{\sqrt{(1-\lambda)^2-4\lambda}}\sum_{n=0}^{\infty}\left[\left(f^-(\lambda)\right)^n-\left(f^+(\lambda)\right)^n\right] \\
&= \frac{1}{\sqrt{(1-\lambda)^2-4\lambda}}\left[\frac{1}{1-\frac{2\lambda}{1-\lambda-\sqrt{(1-\lambda)^2-4\lambda}}}-\frac{1}{1-\frac{2\lambda}{1-\lambda+\sqrt{(1-\lambda)^2-4\lambda}}}\right] \\
&= \frac{4\lambda}{\left(3\lambda-1+\sqrt{(1-\lambda)^2-4\lambda}\right)\left(3\lambda-1-\sqrt{(1-\lambda)^2-4\lambda}\right)} \\
&= \frac{1}{2\lambda}=\frac{\beta T_1}{\alpha}.
\end{aligned} \tag{5.26}$$

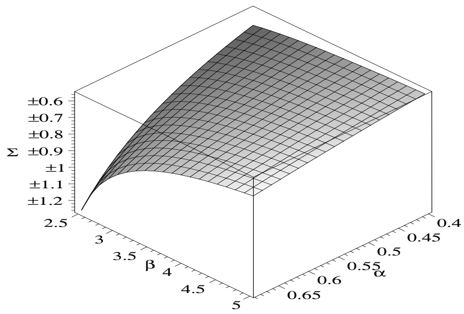


Abbildung 18: Plot von $\dfrac{\beta T_1}{\alpha}$ für $T_1 = 1$, $\alpha \in (0.4, 12-8\sqrt{2})$ und $\beta \in (2.5, 5)$

5.3.3 Allgemeiner Fall und Fehlerabschätzung

Wir betrachten nun das DAWP (5.15). Durch die Ersetzungsregeln (5.9) versuchen wir, eine Näherungslösung für H_k aus (5.15) herzuleiten und ihren Fehler abzuschätzen. Dabei erhalten wir folgende Aussage.

Satz 5.3.1 *Es seien für alle* $k \in \mathbb{N}_0$

$$\left|\int_k^{k+1}\left(V(\tau+1)-V(\tau)\right)\mathrm{d}\tau\right| \leq L \tag{5.27}$$

sowie

$$\int_k^{k+1} V(\tau)\mathrm{d}\tau \in [-\gamma+\alpha, \gamma+\alpha] \tag{5.28}$$

und

$$\frac{\alpha}{2\beta T_1} \in (0,1).$$

Weiterhin bezeichne

$$z_k := z_k^I + z_k^{II}$$

mit

$$\begin{aligned} z_k^I \quad := \quad & \frac{1}{2\alpha}\int\limits_0^k (k-\tau)V(\tau)\mathrm{d}\tau + \frac{1}{4\alpha^2}\left[2\alpha\beta T_1 k - (2\beta T_1+\alpha)\int\limits_0^k V(\tau)\mathrm{d}\tau\right] \\ & + \frac{\beta T_1}{4\alpha^3}(2\beta T_1 - \alpha)\left(\int\limits_0^k V(\tau)\mathrm{d}\tau - \alpha\right) - \frac{\beta T_1}{2\alpha} + C_k\cdot\mu_+^k + D_k\cdot\mu_-^k \end{aligned} \tag{5.29}$$

und

$$\begin{aligned} C_k := & \frac{\beta T_1}{8\alpha^3}\left(\left(\int\limits_k^{k+1} V(\tau)\mathrm{d}\tau - \alpha\right)\left[(\alpha - 2\beta T_1) + \frac{\alpha^2 - 4\beta^2T_1^2 + 8\alpha\beta T_1}{\sqrt{(2\beta T_1-\alpha)^2 - 8\alpha\beta T_1}}\right]\right. \\ & \left. +2\alpha^2 + \frac{2\alpha^2(2\beta T_1+\alpha)}{\sqrt{(2\beta T_1-\alpha)^2 - 8\alpha\beta T_1}}\right), \\ D_k := & \frac{\beta T_1}{8\alpha^3}\left(\left(\int\limits_k^{k+1} V(\tau)\mathrm{d}\tau - \alpha\right)\left[(\alpha - 2\beta T_1) - \frac{\alpha^2 - 4\beta^2T_1^2 + 8\alpha\beta T_1}{\sqrt{(2\beta T_1-\alpha)^2 - 8\alpha\beta T_1}}\right]\right. \\ & \left. +2\alpha^2 - \frac{2\alpha^2(2\beta T_1+\alpha)}{\sqrt{(2\beta T_1-\alpha)^2 - 8\alpha\beta T_1}}\right) \end{aligned} \tag{5.30}$$

eine Näherungslösung des DAWP (5.15)

$$\begin{aligned} H_{k+2} - \left(1 - \frac{\alpha}{2\beta T_1}\right)H_{k+1} + \frac{\alpha}{2\beta T_1}H_k \quad &= \quad \frac{1}{2\beta T_1}\int\limits_0^k (k-\tau)V(\tau)\mathrm{d}\tau + \frac{k}{2} + p_k, \quad k\in\mathbb{N}_0, \\ H_0 \quad &= \quad H_1 = 0. \end{aligned}$$

Dann gilt

$$\begin{aligned} & \left| z_{k+2}^I - \left(1 - \frac{\alpha}{2\beta T_1}\right)z_{k+1}^I + \frac{\alpha}{2\beta T_1}z_k^I - \frac{1}{2\beta T_1}\int\limits_0^k (k-\tau)V(\tau)\mathrm{d}\tau - \frac{k}{2}\right| \\ & \leq L\left\{\frac{\alpha^3 + 12\alpha^2\beta T_1 + 8\beta^3T_1^3}{16\alpha^3\beta T_1} + \frac{1}{L}\left(\frac{\gamma}{\alpha} + \frac{3}{2}\right)\right. \\ & \quad \left. + \left|2(Q+P)\mu_+^{k+2} + 2(Q-P)\mu_-^{k+2} - \left(1 - \frac{\alpha}{2\beta T_1}\right)(Q+P)\mu_+^{k+1} + (Q-P)\mu_-^{k+1}\right|\right\} \end{aligned} \tag{5.31}$$

mit

$$\begin{aligned} Q(\alpha,\beta,T_1) \quad &:= \quad \frac{\beta T_1}{8\alpha^3}(\alpha - 2\beta T_1), \\ P(\alpha,\beta,T_1) \quad &:= \quad \frac{\beta T_1}{8\alpha^3}\frac{\alpha^2 - 4\beta^2T_1^2 + 8\alpha\beta T_1}{\sqrt{(2\beta T_1-\alpha)^2 - 8\alpha\beta T_1}}. \end{aligned}$$

Beweis: Ausgangspunkt unserer Überlegungen sind die Formeln (5.22) für die Lösung der Differenzengleichung im konstanten Fall ohne Unbestimmtheit p_k und (5.23) sowie (5.24) für die Terme C und D. Wenden wir auf diese Ausdrücke die Ersetzungsregeln (5.9) an, so erhalten wir (5.30) und

$$\begin{aligned} z_k^I &= \frac{1}{2\alpha}\int_0^k (k-\tau)V(\tau)\mathrm{d}\tau + \frac{1}{4\alpha^2}\left[2\alpha\beta T_1 k - (2\beta T_1+\alpha)\int_0^k V(\tau)\mathrm{d}\tau\right] \\ &\quad + \frac{\beta T_1}{4\alpha^3}(2\beta T_1-\alpha)\left(\int_0^k V(\tau)\mathrm{d}\tau - \alpha\right) - \frac{\beta T_1}{2\alpha} + C_k\cdot\mu_+^k + D_k\cdot\mu_-^k. \end{aligned}$$

Wir wollen jetzt den „homogenen“ Anteil

$$z_k^h := C_k\mu_+^k + D_k\mu_-^k$$

von z_k^I untersuchen. Es sei nun

$$\begin{aligned} R(\alpha,\beta,T_1) &:= \frac{\beta T_1}{8\alpha^3}2\alpha^2, \\ S(\alpha,\beta,T_1) &:= \frac{\beta T_1}{8\alpha^3}\frac{2\alpha^2(2\beta T_1+\alpha)}{\sqrt{(2\beta T_1-\alpha)^2-8\alpha\beta T_1}}. \end{aligned}$$

Dann ergibt sich

$$\begin{aligned} z_k^h &= \left[\left(\int_k^{k+1} V(\tau)\mathrm{d}\tau - \alpha\right)(Q+P)+(R+S)\right]\mu_+^k \\ &\quad + \left[\left(\int_k^{k+1} V(\tau)\mathrm{d}\tau - \alpha\right)(Q-P)+(R-S)\right]\mu_-^k \end{aligned}$$

und

$$\begin{aligned} z_{k+1}^h &= \left[\left(\int_{k+1}^{k+2} V(\tau)\mathrm{d}\tau - \alpha\right)(Q+P)+(R+S)\right]\mu_+^{k+1} \\ &\quad + \left[\left(\int_{k+1}^{k+2} V(\tau)\mathrm{d}\tau - \alpha\right)(Q-P)+(R-S)\right]\mu_-^{k+1} \\ &= \left[\left(\int_k^{k+1} V(\tau+1)\mathrm{d}\tau - \alpha\right)(Q+P)+(R+S)\right]\mu_+^{k+1} \\ &\quad + \left[\left(\int_k^{k+1} V(\tau+1)\mathrm{d}\tau - \alpha\right)(Q-P)+(R-S)\right]\mu_-^{k+1} \\ &= \left(\int_k^{k+1} (V(\tau+1)-V(\tau))\mathrm{d}\tau\right)\left\{\left[(Q+P)\mu_+^{k+1}\right]+\left[(Q-P)\mu_-^{k+1}\right]\right\} \\ &\quad + C_k\mu_+^{k+1} + D_k\mu_-^{k+1} \end{aligned}$$

und anaolog dazu

$$
\begin{aligned}
z^h_{k+2} &= \left[\left(\int\limits_{k+2}^{k+3} V(\tau)\mathrm{d}\tau - \alpha\right)(Q+P) + (R+S)\right]\mu_+^{k+2} \\
&\quad + \left[\left(\int\limits_{k+2}^{k+3} V(\tau)\mathrm{d}\tau - \alpha\right)(Q-P) + (R-S)\right]\mu_-^{k+2} \\
&= \left(\int\limits_{k}^{k+1} (V(\tau+2) - V(\tau))\mathrm{d}\tau\right)\left\{\left[(Q+P)\mu_+^{k+2}\right] + \left[(Q-P)\mu_-^{k+2}\right]\right\} \\
&\quad + C_k\mu_+^{k+2} + D_k\mu_-^{k+2}.
\end{aligned}
$$

Wegen

$$
\mu_\pm = \frac{(2\beta T_1 - \alpha) \pm \sqrt{(2\beta T_1 - \alpha)^2 - 8\alpha\beta T_1}}{4\beta T_1}
$$

ist

$$
\mu_\pm^2 - \left(1 - \frac{\alpha}{2\beta T_1}\right)\mu_\pm + \frac{\alpha}{2\beta T_1} = 0.
$$

Damit ergibt sich ein Fehler für den homogenen Anteil, wenn man auf z^h_k den Differenzenoperator entsprechend der linken Seite von (5.15) anwendet

$$
\begin{aligned}
& z^h_{k+2} - \left(1 - \frac{\alpha}{2\beta T_1}\right) z^h_{k+1} + \frac{\alpha}{2\beta T_1} z^h_k \\
&\quad = \left(\int\limits_{k}^{k+1} (V(\tau+2) - V(\tau))\mathrm{d}\tau\right)\left[(Q+P)\mu_+^{k+2} + (Q-P)\mu_-^{k+2}\right] \\
&\qquad - \left(1 - \frac{\alpha}{2\beta T_1}\right)\left(\int\limits_{k}^{k+1} (V(\tau+1) - V(\tau))\mathrm{d}\tau\right)\left[(Q+P)\mu_+^{k+1} + (Q-P)\mu_-^{k+1}\right].
\end{aligned}
$$

Aus den oben getroffenen Voraussetzungen (5.27) folgt dann

$$
\begin{aligned}
& \left| z^h_{k+2} - \left(1 - \frac{\alpha}{2\beta T_1}\right) z^h_{k+1} + \frac{\alpha}{2\beta T_1} z^h_k \right| \qquad (5.32) \\
&\quad \le L \left| 2(Q+P)\mu_+^{k+2} + 2(Q-P)\mu_-^{k+2} - \left(1 - \frac{\alpha}{2\beta T_1}\right)(Q+P)\mu_+^{k+1} + (Q-P)\mu_-^{k+1} \right|.
\end{aligned}
$$

Nun untersuchen wir das Verhalten des „partikulären“ Anteils der Näherungslösung

$$
\begin{aligned}
z^p_k &:= \frac{1}{2\alpha}\int\limits_0^k (k-\tau)V(\tau)\mathrm{d}\tau + \frac{1}{4\alpha^2}\left[2\alpha\beta T_1 k - (2\beta T_1 + \alpha)\int\limits_0^k V(\tau)\mathrm{d}\tau\right] \\
&\quad + \frac{\beta T_1}{4\alpha^3}(2\beta T_1 - \alpha)\left(\int\limits_0^k V(\tau)\mathrm{d}\tau - \alpha\right) - \frac{\beta T_1}{2\alpha}
\end{aligned}
$$

bez. der Differenzengleichung. Hierfür benötigen wir Ersetzbarkeitsformeln zur Abschätzung des Residuums

$$
\begin{aligned}
\int_0^{k+1} (k+1-\tau)\, V(\tau)\, \mathrm{d}\tau &= \int_0^{k+1} (k-\tau)\, V(\tau)\, \mathrm{d}\tau + \int_0^{k+1} V(\tau)\, \mathrm{d}\tau \\
&= \int_0^{k} (k-\tau)\, V(\tau)\, \mathrm{d}\tau + \int_0^{k} V(\tau)\, \mathrm{d}\tau + \int_k^{k+1} (k-\tau)\, V(\tau)\, \mathrm{d}\tau + \int_k^{k+1} V(\tau)\, \mathrm{d}\tau \\
&= A + B + C + E,
\end{aligned}
$$

$$
\begin{aligned}
\int_0^{k+1} V(\tau)\, \mathrm{d}\tau &= \int_0^{k} V(\tau)\, \mathrm{d}\tau + \int_k^{k+1} V(\tau)\, \mathrm{d}\tau \\
&= B + E,
\end{aligned}
$$

$$
\begin{aligned}
\int_{k+1}^{k+2} V(\tau)\, \mathrm{d}\tau &= \int_{t_k}^{t_{k+1}} V(\tau)\, \mathrm{d}\tau + \int_k^{k+1} (V(\tau+1) - V(\tau))\, \mathrm{d}\tau \\
&= E + R,
\end{aligned}
$$

$$
\begin{aligned}
\int_0^{k+2} (k+2-\tau)\, V(\tau)\, \mathrm{d}\tau &= \int_0^{k+2} (k-\tau)\, V(\tau)\, \mathrm{d}\tau + 2\int_0^{k+2} V(\tau)\, \mathrm{d}\tau \\
&= \int_0^{k} (k-\tau)\, V(\tau)\, \mathrm{d}\tau + 2\int_0^{k} V(\tau)\, \mathrm{d}\tau + \int_k^{k+2} (k-\tau)\, V(\tau)\, \mathrm{d}\tau + 2\int_k^{k+2} V(\tau)\, \mathrm{d}\tau \\
&= A + 2B + G + H,
\end{aligned}
$$

$$
\begin{aligned}
G &= \int_k^{k+2} (k-\tau)\, V(\tau)\, \mathrm{d}\tau = \int_k^{k+1} (k-\tau)\, V(\tau)\, \mathrm{d}\tau + \int_{k+1}^{k+2} (k-\tau)\, V(\tau)\, \mathrm{d}\tau \\
&= C + \int_k^{k+1} (k-\tau-1)\, V(\tau+1)\, \mathrm{d}\tau \\
&= C + \int_k^{k+1} (k-\tau)\, V(\tau+1)\, \mathrm{d}\tau - \int_k^{k+1} V(\tau+1) - V(\tau)\, \mathrm{d}\tau - \int_k^{k+1} V(\tau)\, \mathrm{d}\tau \\
&= C - E - R + \int_k^{k+1} (k-\tau)\, V(\tau+1)\, \mathrm{d}\tau \\
&= 2C - E - R + \int_k^{k+1} (k-\tau)(V(\tau+1) - V(\tau))\mathrm{d}\tau \\
&= 2C - E - R + U,
\end{aligned}
$$

$$\begin{aligned}\int_0^{k+2} V(\tau)\,\mathrm{d}\tau &= \int_0^{k} V(\tau)\,\mathrm{d}\tau + \int_k^{k+2} V(\tau)\,\mathrm{d}\tau \\ &= B + H,\end{aligned}$$

$$\begin{aligned}H &= \int_k^{k+2} V(\tau)\,\mathrm{d}\tau = 2\int_k^{k+1} V(\tau)\,\mathrm{d}\tau + \int_k^{k+1} V(\tau+1) - V(\tau)\,\mathrm{d}\tau \\ &= 2E + R,\end{aligned}$$

$$\begin{aligned}\int_{k+2}^{k+3} V(\tau)\,\mathrm{d}\tau &= \int_{t_k}^{t_{k+1}} V(\tau)\,\mathrm{d}\tau + \int_k^{k+1} (V(\tau+2) - V(\tau))\,\mathrm{d}\tau \\ &= E + S,\end{aligned}$$

$$\begin{aligned}S &= \int_k^{k+1} (V(\tau+2) - V(\tau))\,\mathrm{d}\tau \\ &= \int_k^{k+1} (V(\tau+2) - V(\tau+1))\mathrm{d}\tau + \int_k^{k+1} (V(\tau+1) - V(\tau))\,\mathrm{d}\tau \\ &= T + R.\end{aligned}$$

Mit diesen Termen erhalten wir dann

$$\begin{aligned}z_k^p &= \frac{1}{2\alpha}\int_0^k (k-\tau)V(\tau)\mathrm{d}\tau + \frac{1}{4\alpha^2}\left[2\alpha\beta T_1 k - (2\beta T_1 + \alpha)\int_0^k V(\tau)\mathrm{d}\tau\right] \\ &\quad + \frac{\beta T_1}{4\alpha^3}(2\beta T_1 - \alpha)\left(\int_k^{k+1} V(\tau)\mathrm{d}\tau - \alpha\right) - \frac{\beta T_1}{2\alpha} \\ &= \frac{1}{2\alpha}A + \frac{1}{4\alpha^2}\left[2\alpha\beta T_1 k - (2\beta T_1 + \alpha)B\right] + \frac{\beta T_1}{4\alpha^3}(2\beta T_1 - \alpha)(E - \alpha) - \frac{\beta T_1}{2\alpha}, \\ z_{k+1}^p &= \frac{1}{2\alpha}\int_0^{k+1} (k+1-\tau)V(\tau)\mathrm{d}\tau + \frac{1}{4\alpha^2}\left[2\alpha\beta T_1 (k+1) - (2\beta T_1 + \alpha)\int_0^{k+1} V(\tau)\mathrm{d}\tau\right] \\ &\quad + \frac{\beta T_1}{4\alpha^3}(2\beta T_1 - \alpha)\left(\int_{k+1}^{k+2} V(\tau)\mathrm{d}\tau - \alpha\right) - \frac{\beta T_1}{2\alpha} \\ &= \frac{1}{2\alpha}(A + B + C + E) + \frac{1}{4\alpha^2}\left[2\alpha\beta T_1 k - (2\beta T_1 + \alpha)(B + E)\right] \\ &\quad + \frac{\beta T_1}{4\alpha^3}(2\beta T_1 - \alpha)(E + R - \alpha) - \frac{\beta T_1}{2\alpha},\end{aligned}$$

$$
\begin{aligned}
z^p_{k+2} &= \frac{1}{2\alpha}\int\limits_0^{k+2}(k+2-\tau)V(\tau)\mathrm{d}\tau + \frac{1}{4\alpha^2}\left[2\alpha\beta T_1(k+2) - (2\beta T_1+\alpha)\int\limits_0^{k+2}V(\tau)\mathrm{d}\tau\right] \\
&\quad + \frac{\beta T_1}{4\alpha^3}(2\beta T_1-\alpha)\left(\int\limits_{k+2}^{k+3}V(\tau)\mathrm{d}\tau - \alpha\right) - \frac{\beta T_1}{2\alpha} \\
&= \frac{1}{2\alpha}(A+2B+2C+E+U) \\
&\quad + \frac{1}{4\alpha^2}\left[2\alpha\beta T_1(k+2) - (2\beta T_1+\alpha)(B+2E+R)\right] \\
&\quad + \frac{\beta T_1}{4\alpha^3}(2\beta T_1-\alpha)\left(E+T+R-\alpha\right) - \frac{\beta T_1}{2\alpha}.
\end{aligned}
$$

Setzen wir nun diese partikulären Lösungen in die Differenzengleichung (5.15) - ohne p_k auf der rechten Seite, da dieser Term in die Betrachtungen von H_k^{II} resp. hier in die von z_k^{II} eingeht - ein und subtrahieren deren rechte Seite, so gilt

$$
\begin{aligned}
& z^p_{k+2} - \left(1-\frac{\alpha}{2\beta T_1}\right)z^p_{k+1} + \frac{\alpha}{2\beta T_1}z^p_k - \frac{1}{2\beta T_1}A - \frac{k}{2} \\
&= -\frac{2\beta T_1+3\alpha}{8\alpha^2}R + \frac{2\beta T_1+\alpha}{8\alpha\beta T_1}(2C+E) - \frac{E}{\alpha} + \frac{U}{2\alpha} + \frac{\beta T_1}{4\alpha^3}(2\beta T_1-\alpha)T - \frac{1}{2}. \qquad (5.33)
\end{aligned}
$$

Dabei gelten wegen (5.27) folgende Abschätzungen

$$
\begin{aligned}
|2C+E| &\leq \sup_{k=0,1,2,\ldots}\int\limits_0^1 |1-2\tau|\,|V(\tau+k)-V(k)|\,\mathrm{d}\tau \\
&\leq L\int\limits_0^1 |1-2\tau|\,\mathrm{d}\tau = \frac{L}{2} \\
|R| &= \sup_{k=0,1,2,\ldots}\left|\int\limits_k^{k+1}(V(\tau+1)-V(\tau))\,\mathrm{d}\tau\right| \leq L \\
|U| &\leq \sup_{k=0,1,2,\ldots}\left|\int\limits_k^{k+1}(\tau-k)\,(V(\tau+1)-V(\tau))\mathrm{d}\tau\right| \\
&\leq L\int\limits_k^{k+1}(\tau-k)\,\mathrm{d}\tau = \frac{1}{2}L \\
|T| &\leq L.
\end{aligned}
$$

Der einzige Term aus (5.33), welcher auf diese Art nicht durch L abschätzbar ist, lautet

$$
E_k := \int\limits_k^{k+1}V(\tau)\mathrm{d}\tau.
$$

Nach Voraussetzungen gilt aber

$$\left|-E_k - \frac{1}{2}\right| \le \frac{\gamma}{\alpha} + \frac{3}{2}.$$

Mit den obigen Abschätzungen und dem partikulären Anteil der Näherungslösung gilt dann

$$\begin{aligned}
&\left|z^p_{k+2} - \left(1 - \frac{\alpha}{2\beta T_1}\right) z^p_{k+1} + \frac{\alpha}{2\beta T_1} z^p_k - \frac{1}{2\beta T_1} A - \frac{k}{2}\right| \\
&\quad = \left|-\frac{2\beta T_1 + 3\alpha}{8\alpha^2} R + \frac{2\beta T_1 + \alpha}{8\alpha\beta T_1}(2C + E) + \frac{U}{2\alpha} + \frac{\beta T_1}{4\alpha^3}(2\beta T_1 - \alpha)T - \frac{E}{2} - \frac{1}{2}\right| \\
&\quad \le L\left(\frac{2\beta T_1 + 3\alpha}{8\alpha^2} + \frac{2\beta T_1 + \alpha}{8\alpha\beta T_1}\frac{1}{4} + \frac{1}{4\alpha} + \frac{\beta T_1(2\beta T_1 - \alpha)}{4\alpha^3} + \frac{1}{L}\left(\frac{\gamma}{\alpha} + \frac{3}{2}\right)\right) \\
&\quad = L\left(\frac{\alpha^3 + 12\alpha^2\beta T_1 + 8\beta^3 T_1^3}{16\alpha^3\beta T_1} + \frac{1}{L}\left(\frac{\gamma}{\alpha} + \frac{3}{2}\right)\right). \qquad (5.34)
\end{aligned}$$

Nach Superposition setzt sich der Gesamtfehler aus dem homogene Fehleranteil (5.32) und dem partikulären (5.34) zusammen und lautet

$$\begin{aligned}
&\left|z^I_{k+2} - \left(1 - \frac{\alpha}{2\beta T_1}\right) z^I_{k+1} + \frac{\alpha}{2\beta T_1} z^I_k - \frac{1}{2\beta T_1}\int_0^k (k-\tau)V(\tau)\mathrm{d}\tau - \frac{k}{2}\right| \\
&\le L\left\{\frac{\alpha^3 + 12\alpha^2\beta T_1 + 8\beta^3 T_1^3}{16\alpha^3\beta T_1} + \frac{1}{L}\left(\frac{\gamma}{\alpha} + \frac{3}{2}\right)\right. \\
&\qquad \left. + \left|2(Q+P)\mu_+^{k+2} + 2(Q-P)\mu_-^{k+2} - \left(1 - \frac{\alpha}{2\beta T_1}\right)(Q+P)\mu_+^{k+1} + (Q-P)\mu_-^{k+1}\right|\right\}.
\end{aligned}$$

Damit ist alles gezeigt. □

Bemerkung 5.3.2 Es sei

$$\lambda := \frac{2\alpha}{\beta T_1}.$$

Dann ist

$$\mu_\pm = \mu_\pm(\lambda) = \frac{4 - \lambda \pm \sqrt{(4-\lambda)^2 - 16\lambda}}{8}.$$

Dabei ist $\mu_+(\lambda)$ streng monoton fallend und $\mu_-(\lambda)$ streng monoton wachsend für $\lambda \in (0, 12 - 8\sqrt{2})$ mit

$$\begin{aligned}
\mu_+(\lambda) &\in (\sqrt{2} - 1, 1), \\
\mu_-(\lambda) &\in (0, \sqrt{2} - 1).
\end{aligned}$$

Somit erhalten wir wegen

$$\mu_\pm^k \to 0 \quad \text{für} \quad k \to +\infty$$

dann

$$z_k^h = C_k\mu_+^k + D_k\mu_-^k \to 0 \quad \text{für} \quad k \to +\infty$$

und damit

$$L\left\{\frac{\alpha^3+12\alpha^2\beta T_1+8\beta^3T_1^3}{16\alpha^3\beta T_1}+\frac{1}{L}\left(\frac{\gamma}{\alpha}+\frac{3}{2}\right)\right.$$
$$\left.+\left|2(Q+P)\mu_+^{k+2}+2(Q-P)\mu_-^{k+2}-\left(1-\frac{\alpha}{2\beta T_1}\right)(Q+P)\mu_+^{k+1}+(Q-P)\mu_-^{k+1}\right|\right\}$$
$$\to L\left\{\frac{\alpha^3+12\alpha^2\beta T_1+8\beta^3T_1^3}{16\alpha^3\beta T_1}+\frac{1}{L}\left(\frac{\gamma}{\alpha}+\frac{3}{2}\right)\right\}\quad \text{für}\quad k\to+\infty.$$

Wir wollen nun die einzelnen Fehleranteile von (5.31) graphisch darstellen. In den folgenden Abbildungen sei

$$T_1 := 1,\quad \alpha\in(0.4, 12-8\sqrt{2}),\quad \beta\in(2.5,5).$$

Zunächst die Terme $\mu_\pm(\lambda)$:

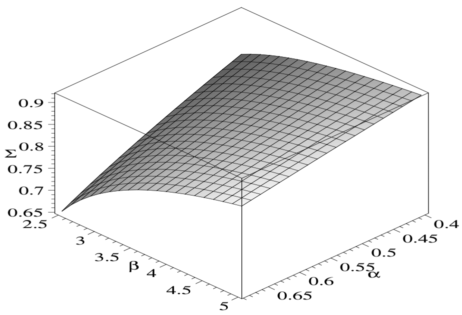


Abbildung 19: Plot von $\mu_+(\alpha,\beta,T_1)=\dfrac{2\beta T_1-\alpha+\sqrt{(2\beta T_1-\alpha)^2-8\alpha\beta T_1}}{4\beta T_1}$

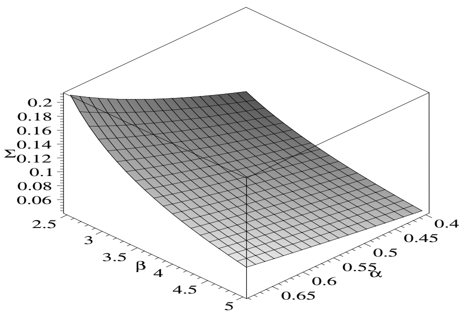


Abbildung 20: Plot von $\mu_-(\alpha,\beta,T_1)=\dfrac{2\beta T_1-\alpha-\sqrt{(2\beta T_1-\alpha)^2-8\alpha\beta T_1}}{4\beta T_1}$

Darüber hinaus gilt für den Term der speziellen Lösung

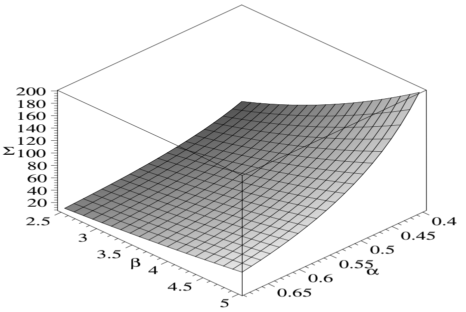


Abbildung 21: Plot von $\dfrac{\alpha^3 + 12\alpha^2\beta T_1 + 8\beta^3 T_1^3}{16\alpha^3\beta T_1}$

und für die Fehlerterme der homogenen Lösung, welche für $k = 0$ maximal sind, gilt und

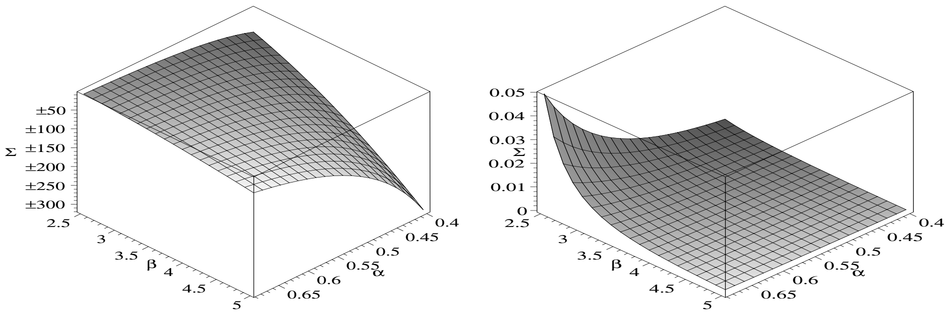


Abbildung 22: Plot von $2(Q+P)\mu_+^2$ und $2(Q-P)\mu_-^2$

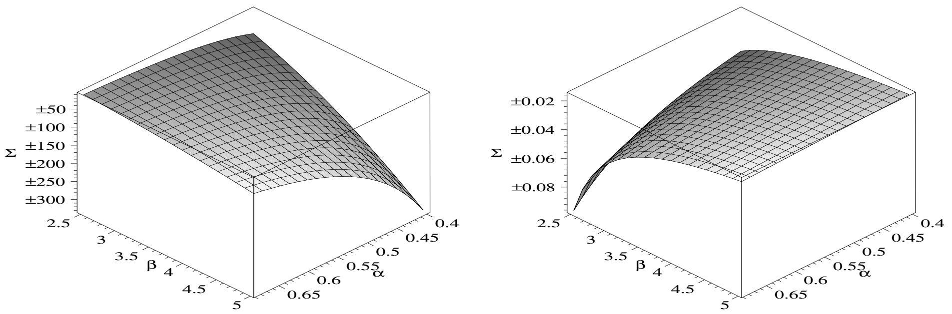


Abbildung 23: Plot von $\left(1 - \dfrac{\alpha}{2\beta T_1}\right)(Q+P)\mu_+^1$ und $\left(1 - \dfrac{\alpha}{2\beta T_1}\right)(Q-P)\mu_-^1$

Bisher haben wir den Fehler im „Wertebereich“, d. h. den Fehler der Näherungslösung zur rechten Seite von (5.15) abgeschätzt. Nun wollen wir - mit Hinblick auf einen Filtervorschlag durch diese Näherungslösung - uns dem Fehler im „Urbildbereich“, d. h. dem Fehler der Näherungslösung zur wahren Lösung von (5.15), zuwenden. Die gleichen Überlegungen wie bei der Herleitung von (5.26) und ein paar einfache Rechnungen liefern folgende Aussage.

Satz 5.3.3 *Es bezeichne $H_k \in \mathbb{Z}$ die wahre Lösung von* (5.15). *Weiterhin gelten die Voraussetzungen aus dem Satz* 5.3.1. *Ist*

$$\frac{2\alpha}{\beta T_1} \in \left(0, 12 - 8\sqrt{2}\right],$$

dann gilt

$$\left|z_k^I - H_k\right| \leq \left(L\left\{\frac{\alpha^3+12\alpha^2\beta T_1+8\beta^3T_1^3}{16\alpha^3\beta T_1} + \frac{1}{8\alpha^3}\left|-5\alpha^2+12\alpha\beta T_1-4\beta T_1\right|\right\} + \frac{\gamma}{\alpha}+\frac{3}{2}\right)\frac{\beta T_1}{\alpha}. \tag{5.35}$$

Beweis: Nach Bemerkung 5.3.2 ist der Ausdruck

$$\left|2(Q+P)\mu_+^{k+2}+2(Q-P)\mu_-^{k+2}-\left(1-\frac{\alpha}{2\beta T_1}\right)(Q+P)\mu_+^{k+1}+(Q-P)\mu_-^{k+1}\right|$$

für $k=0$ maximal. Dafür gilt gerade

$$\begin{aligned}
&2(Q+P)\mu_+^2+2(Q-P)\mu_-^2-\left(1-\frac{\alpha}{2\beta T_1}\right)(Q+P)\mu_++(Q-P)\mu_-\\
&\quad=\frac{1}{2\alpha^3}\left(-2\alpha^2+5\alpha\beta T_1-2\beta^2T_1^2\right)-\left(1-\frac{\alpha}{2\beta T_1}\right)\frac{\beta T_1}{4\alpha^3}(3\alpha-2\beta T_1)\\
&\quad=\frac{1}{8\alpha^3}\left(-5\alpha^2+12\alpha\beta T_1-4\beta^2T_1^2\right).
\end{aligned}\tag{5.36}$$

Insbesondere gilt also

$$\begin{aligned}
&\left|2(Q+P)\mu_+^2+2(Q-P)\mu_-^2-\left(1-\frac{\alpha}{2\beta T_1}\right)(Q+P)\mu_++(Q-P)\mu_-\right|\\
&\quad\leq\frac{1}{8\alpha^3}\begin{cases}5\alpha^2-12\alpha\beta T_1+4\beta^2T_1^2, & \text{für } \alpha\in\left(0,\frac{2}{5}\beta T_1\right],\\ -5\alpha^2+12\alpha\beta T_1-4\beta^2T_1^2, & \text{für } \alpha\in\left(\frac{2}{5}\beta T_1,(6-4\sqrt{2})\beta T_1\right].\end{cases}
\end{aligned}$$

Nun sei

$$\begin{aligned}
M_k := L\Bigg\{&\frac{\alpha^3+12\alpha^2\beta T_1+8\beta^3T_1^3}{16\alpha^3\beta T_1}\\
&+\left|2(Q+P)\mu_+^{k+2}+2(Q-P)\mu_-^{k+2}-\left(1-\frac{\alpha}{2\beta T_1}\right)(Q+P)\mu_+^{k+1}+(Q-P)\mu_-^{k+1}\right|\Bigg\}.
\end{aligned}$$

Dann ist nach Satz 5.3.1 der Fehler der Näherungslösung zur rechten Seite von (5.36) durch

$$\left|z_{k+2}^I-\left(1-\frac{\alpha}{2\beta T_1}\right)z_{k+1}^I+\frac{\alpha}{2\beta T_1}z_k^I-\frac{1}{2\beta T_1}\int\limits_0^k(k-\tau)V(\tau)\mathrm{d}\tau-\frac{k}{2}\right|\leq M_k+\frac{\gamma}{\alpha}+\frac{3}{2}$$

gegeben. Um den Fehler der Näherungslösung zur wahren Lösung abschätzen zu können, betrachten wir das DAWP

$$\begin{aligned}
q_{k+2}-\left(1-\frac{\alpha}{2\beta T_1}\right)q_{k+1}+\frac{\alpha}{2\beta T_1}q_k &= M_k+\frac{\gamma}{\alpha}+\frac{3}{2}, \quad k\in\mathbb{N}_0,\\
q_0 &= q_1 = 0.
\end{aligned}\tag{5.37}$$

Für die Lösung q_k von (5.37) gilt wie bei (5.26) wegen (5.25) dann

$$q_k = \frac{2\beta T_1}{\sqrt{(2\beta T_1-\alpha)^2-8\alpha\beta T_1}}\sum_{n=0}^{k-1}\left[\left(\frac{2\alpha}{2\beta T_1-\alpha-\sqrt{(2\beta T_1-\alpha)^2-8\alpha\beta T_1}}\right)^n\right.$$

$$
\begin{aligned}
& \quad - \left(\frac{2\alpha}{2\beta T_1 - \alpha + \sqrt{(2\beta T_1-\alpha)^2 - 8\alpha\beta T_1}} \right)^n \Bigg] \left(M_{k-n-1} + \frac{\gamma}{\alpha} + \frac{3}{2} \right) \\
\leq \; & \frac{2\beta T_1}{\sqrt{(2\beta T_1-\alpha)^2 - 8\alpha\beta T_1}} \sum_{n=0}^{\infty} \Bigg[\left(\frac{2\alpha}{2\beta T_1 - \alpha - \sqrt{(2\beta T_1-\alpha)^2 - 8\alpha\beta T_1}} \right)^n \\
& \quad - \left(\frac{2\alpha}{2\beta T_1 - \alpha + \sqrt{(2\beta T_1-\alpha)^2 - 8\alpha\beta T_1}} \right)^n \Bigg] \left(M_0 + \frac{\gamma}{\alpha} + \frac{3}{2} \right) \\
= \; & \left(M_0 + \frac{\gamma}{\alpha} + \frac{3}{2} \right) \frac{\beta T_1}{\alpha} \\
\leq \; & \Bigg(L \Bigg\{ \frac{\alpha^3 + 12\alpha^2\beta T_1 + 8\beta^3 T_1^3}{16\alpha^3\beta T_1} \\
& \quad + \frac{1}{8\alpha^3} \left| -5\alpha^2 + 12\alpha\beta T_1 - 4\beta T_1 \right| \Bigg\} + \frac{\gamma}{\alpha} + \frac{3}{2} \Bigg) \frac{\beta T_1}{\alpha}.
\end{aligned}
$$

Damit ist alles gezeigt. □

Beispiel 5.3.4 Kehren wir nun zum Standardbeispiel einer 1000-fach abgetasteten Sinusschwingung

$$U_0(t) = \sin\left(\frac{2\pi}{1000} t\right)$$

zurück. Dann ist hier

$$L = \frac{2\pi}{1000}.$$

Wir betrachten den Parameterbereich

$$T_1 = 1, \quad \alpha \in (0.4, 12 - 8\sqrt{2}), \quad \beta \in \left(\frac{5}{2}, 5\right).$$

Dort ist die Bedingung $\frac{2\alpha}{\beta T_1} \in (0, 12 - 8\sqrt{2})$ erfüllt. In diesem Bereich gilt nach Satz 5.3.1 für den Fehler der Näherungslösung z_k^I - eingesetzt in die Differenzengleichung - zur rechten Seite folgende grobe Abschätzung

$$\left| z_{k+2}^I - \left(1 - \frac{\alpha}{2\beta T_1}\right) z_{k+1}^I + \frac{\alpha}{2\beta T_1} z_k^I - \frac{1}{2\beta T_1} \int_0^k (k-\tau) \sin\left(\frac{2\pi}{1000}\tau\right) \mathrm{d}\tau + \frac{k}{2} \right| \in (16, 28).$$

Nach Satz 5.3.3 haben wir dann für den Fehler der Näherungslösung zur wahren Lösung die Abschätzung

$$\left| z_k^I - H_k \right| \in (5, 35).$$

Die oberen bzw. unteren Fehlergrenzen werden für

$$(\alpha, \beta) = (0.4, 5) \quad \text{bzw.} \quad (\alpha, \beta) = (2.5, 12 - 8\sqrt{2})$$

angenommen.

5.4 Filterung bei einem Konverter zweiter Ordnung und Fehlerbetrachtung

In diesem Abschnitt wollen wir auf der Grundlage der Näherungslösung des DAWP (5.15) Filtervorschläge angeben, wenn wir annehmen, daß U_0 stückweise linear bzw. quadratisch ist. Diese Filtervorschläge wollen wir mit dem aus der Differentialgleichung hergeleiteten (4.62) vergleichen und ihre Fehler angeben.

5.4.1 Linearer Ansatz

Im Kapitel 5.3 haben wir eine Näherungslösung des DAWP (5.15) hergeleitet und den Fehler zur wahren Lösung abgeschätzt. Dieses Ergebnis wollen wir nun benutzen, um einen Filtervorschlag mittels zweiter Differenzen dieser Näherungslösung anzugeben und seinen Fehler mit dem aus Kapitel 4.2.2 zu vergleichen.
In den weiteren Betrachtungen sei $U_{ref} = 1$, $\Delta t = 1$ und

$$U_0(t) = a + b(t-q).$$

Dann ist

$$V(t) = a + \alpha + b(t-q)$$

in die Näherungslösung des DAWP (5.15) einzusetzen. Es sei nun

$$\begin{aligned} z_{k,q} \quad := \quad & \frac{1}{2\alpha}\int_0^k (k-\tau)(a+\alpha+b(\tau-q))\mathrm{d}\tau \\ & +\frac{1}{4\alpha^2}\left(2\alpha\beta T_1 k - (2\beta T_1+\alpha)\int_0^k (a+\alpha+b(\tau-q))\mathrm{d}\tau\right) \\ & +\frac{\beta T_1}{4\alpha^3}(2\beta T_1-\alpha)\left(\int_k^{k+1}(a+\alpha+b(\tau-q))\mathrm{d}\tau-\alpha\right)-\frac{\beta T_1}{2\alpha}. \end{aligned} \tag{5.38}$$

Dies ist gerade der partikuläre Anteil von z_k^I. Dann erhalten wir folgende Aussage.

Lemma 5.4.1 *Es sei $m, q \in \mathbb{N}$, und $U_0(t)$ sei linear im Intervall $[q, q+2m)$ mit $U_0(t) = a + b(t-q)$. Weiterhin sei*

$$A_{m,q} := \frac{2\alpha}{m^2}(z_{q+2m,q} - 2z_{q+m,q} + z_{q,q})$$

sowie

$$B_{m,q} := \frac{2\alpha}{(m-1)^2}(z_{q+2m-2,q} - 2z_{q+m-1,q} + z_{q,q}).$$

Dann gilt

$$\frac{1}{2}\left(\frac{2\beta T_1+3\alpha}{\alpha}A_{m,q} - \frac{2\beta T_1+\alpha}{\alpha}B_{m,q}\right) = a + bm = \frac{1}{2m}\int_q^{q+2m}(a+b(\tau-q))\mathrm{d}\tau. \tag{5.39}$$

Somit stimmt die rechte Seite von (5.39) mit $U_0(m+q)$ überein.

Beweis: Setzen wir (5.38) in $A_{m,q}$ und $B_{m,q}$ ein, so ergibt sich

$$A_{m,q} = a + bm - \frac{\beta T_1}{\alpha} b - \frac{b}{2} = a + bm - b\frac{2\beta T_1 + \alpha}{2\alpha}$$

und

$$B_{m,q} = a + b(m-1) - \frac{\beta T_1}{\alpha} b - \frac{b}{2} = a + b(m-1) - b\frac{2\beta T_1 + \alpha}{2\alpha}.$$

Damit ist

$$A_{m,q} - B_{m,q} = b,$$

und es gilt

$$A_{m,q} = a + bm - (A_{m,q} - B_{m,q})\frac{2\beta T_1 + \alpha}{2\alpha}.$$

Wegen

$$\begin{aligned} a + bm &= A_{m,q} + (A_{m,q} - B_{m,q})\frac{2\beta T_1 + \alpha}{2\alpha} \\ &= \frac{1}{2}\left(\frac{2\beta T_1 + 3\alpha}{\alpha} A_{m,q} - \frac{2\beta T_1 + \alpha}{\alpha} B_{m,q}\right) \end{aligned}$$

folgt unmittelbar die Behauptung. □

Bemerkung 5.4.2 Bezeichne nun (5.39) unseren Filtervorschlag, welcher aus den zweiten Differenzen - siehe $A_{m,q}$ und $B_{m,q}$ - der obigen Näherungslösung besteht. Dann ist diese zwar identisch mit dem Wert $U_0(q+m)$. Das liegt aber daran, daß U_0 bereits in $z_{k,q}$ enthalten ist. Im realen Filter stehen uns diese Informationen nicht zur Verfügung, da $H_k \in \mathbb{Z}$. Beachte, daß $H_k \in \mathbb{Z}$ die wahre Lösung von (5.15) ist.

Wir wollen nun den Fehler abschätzen, wenn wir im Filtervorschlag (5.39) die Näherungslösung $z_{k,q}$ durch die wahre Lösung H_k ersetzen.

Satz 5.4.3 *Es seien die Voraussetzungen von Satz* 5.3.1 *und Lemma* 5.4.1 *gegeben. Weiterhin bezeichne $H_k \in \mathbb{Z}$ die wahre Lösung von* (5.15). *Dann gilt*

$$\begin{aligned} &\left| a + bm - \frac{1}{2}\left(\frac{2\beta T_1 + 3\alpha}{\alpha} A_{m,q}(H) - \frac{2\beta T_1 + \alpha}{\alpha} B_{m,q}(H)\right)\right| \\ &\quad \le \left|\frac{2\beta T_1 + 3\alpha}{m^2} - \frac{2\beta T_1 + \alpha}{(m-1)^2}\right| \left(L\frac{\alpha^3 + 12\alpha^2\beta T_1 + 8\beta^3 T_1^3}{16\alpha^3\beta T_1} + \frac{\gamma}{\alpha} + \frac{3}{2}\right)\frac{4\beta T_1}{\alpha}. \end{aligned} \tag{5.40}$$

Beweis: Aus der Formel (5.34) im Beweis des Satzes 5.3.1 und der Tatsache, daß $z_{k,q}$ dem partikulären Teil von z_k^I entspricht, folgt dann

$$\begin{aligned} &\left| z_{k+2,q} - \left(1 - \frac{\alpha}{2\beta T_1}\right) z_{k+1,q} + \frac{\alpha}{2\beta T_1} z_{k,q} - \frac{1}{2\beta T_1}\int_0^k (a + b(\tau - q))\mathrm{d}\tau - \frac{k}{2}\right| \\ &\quad \le L\frac{\alpha^3 + 12\alpha^2\beta T_1 + 8\beta^3 T_1^3}{16\alpha^3\beta T_1} + \frac{\gamma}{\alpha} + \frac{3}{2}. \end{aligned}$$

Hieraus und gleichen Überlegungen wie beim Satz 5.3.3 folgt dann

$$|z_{k,q} - H_k| \leq \left(L \frac{\alpha^3 + 12\alpha^2\beta T_1 + 8\beta^3 T_1^3}{16\alpha^3\beta T_1} + \frac{\gamma}{\alpha} + \frac{3}{2} \right) \frac{\beta T_1}{\alpha}.$$

Hieraus, aus der Definition der Größen $A_{k,q}$ und $B_{k,q}$, hier als zweite Differenzen von H_k benutzt, ergibt sich aus Lemma 5.4.1 und wegen $L = |b|$ dann

$$\begin{aligned}
&\left| a + bm - \frac{1}{2} \left(\frac{2\beta T_1 + 3\alpha}{\alpha} A_{m,q}(H) - \frac{2\beta T_1 + \alpha}{\alpha} B_{m,q}(H) \right) \right| \\
&\quad \leq \left| \frac{2\beta T_1 + 3\alpha}{m^2} - \frac{2\beta T_1 + \alpha}{(m-1)^2} \right| \left(L \frac{\alpha^3 + 12\alpha^2\beta T_1 + 8\beta^3 T_1^3}{16\alpha^3\beta T_1} + \frac{\gamma}{\alpha} + \frac{3}{2} \right) \frac{4\beta T_1}{\alpha}.
\end{aligned}$$

Damit ist alles gezeigt. □

Die Aussage in Satz 5.4.3 ist von lokaler Natur. Man kann sie in folgender Weise verallgemeinern.

Folgerung 5.4.4 Es sei $m \in \mathbb{N}$ und $U_0(t)$ eine auf jedem Intervall doppelter Länge durch eine lineare Funktion mit der Lipschitzkonstante L approximierbar. D. h., Terme höherer Ordnung seien bei einer polynomialen Approximation von U_0 vernachlässigbar. Dann haben wir für unseren Filtervorschlag

$$\frac{1}{2} \left(\frac{2\beta T_1 + 3\alpha}{\alpha} A_{m,q}(H) - \frac{2\beta T_1 + \alpha}{\alpha} B_{m,q}(H) \right) \tag{5.41}$$

die Fehlerabschätzung (5.40).

Bemerkung 5.4.5 Es sei $m \in \mathbb{N}$. Dann ist der erste Faktor

$$\left| \frac{2\beta T_1 + 3\alpha}{m^2} - \frac{2\beta T_1 + \alpha}{(m-1)^2} \right|$$

von (5.40) identisch null für die beiden Werte

$$m_{1,2} = \frac{3\alpha + 2\beta T_1 \pm \sqrt{3\alpha^2 + 8\alpha\beta T_1 + 4\beta^2 T_1^2}}{2\alpha}.$$

Dies ist insgesamt recht überraschend. Es sei aber daran erinnert, daß wir für dieses Resultat annahmen, daß U_0 stückweise in jedem Intervall linear sei. Nur konstante Funktionen erfüllen diese Annahme.

Zurück zu unserem Standardbeispiel.

Beispiel 5.4.6 Wir betrachten den gleichen Parameterbereich wie in Beispiel 5.3.4. Fassen wir den 1000-fach abgetasteten Sinus für $m = 25$ als eine in jedem Intervall der Länge $2m$ lineare Funktion auf, dann liegt der Filterfehler bez. 5.41, welcher $\sin \frac{2\pi}{1000} \left(k + \frac{m}{2} \right)$ approximiert, im Intervall $(0, 0.04)$. Dies entspricht einer oberen Fehlerschranke von maximal zwei Prozent.

Folgerung 5.4.7 Der Filtervorschlag (5.41) hat bez. der Ausgangsfunktion $X(t) \in \{-1, +1\}$ des Komparators die Gestalt

$$\begin{aligned}
&\frac{(m-1)^2(2\beta T_1 + 3\alpha) - m^2(2\beta T_1 + \alpha)}{m^2(m-1)^2} \sum_{n=1}^{m} \sum_{j=0}^{m-1} X_{q+n+j} \\
&\quad + \frac{(\alpha + 2\beta T_1)}{(m-1)^2} \left(\sum_{j=0}^{m-2} 2X_{q+m+j} + X_{q+2m-1} \right).
\end{aligned}$$

Beweis: Wegen

$$X_k = \Delta^2 H_k$$

gilt

$$\begin{aligned} H_k &= \sum_{n=0}^{k-1}\sum_{j=0}^{n-1} X_j, \quad k \in \mathbb{N}\setminus\{0,1\}, \\ H_0 &= H_1 = 0. \end{aligned}$$

Hieraus folgt dann

$$\begin{aligned} A_{m,q}(H) &= \frac{2\alpha}{m^2}(H_{q+2m} - 2H_{q+m} + H_q) \\ &= \frac{2\alpha}{m^2}\left[\sum_{n=0}^{q+2m-1}\sum_{j=0}^{n-1} X_j - 2\sum_{n=0}^{q+m-1}\sum_{j=0}^{n-1} X_j + \sum_{n=0}^{q-1}\sum_{j=0}^{n-1} X_j\right] \\ &= \frac{2\alpha}{m^2}\sum_{n=1}^{m}\sum_{j=0}^{m-1} X_{q+n+j} \\ &= \frac{2\alpha}{m^2}\left(\sum_{j=1}^{m-1} j(X_{q+j} + X_{q+2m-j}) + mX_{q+m}\right), \end{aligned} \tag{5.42}$$

und analog dazu ist

$$\begin{aligned} B_{m,q}(H) &= \frac{2\alpha}{(m-1)^2}(H_{q+2m-2} - 2H_{q+m-1} + H_q) \\ &= \frac{2\alpha}{(m-1)^2}\sum_{n=1}^{m-1}\sum_{j=1}^{m-1} X_{q+n+j} \\ &= \frac{2\alpha}{(m-1)^2}\left(\sum_{j=1}^{m-1} j(X_{q+j} + X_{q+2m-j}) + mX_{q+m}\right), \end{aligned} \tag{5.43}$$

Aus (5.42), (5.43) sowie (5.42) ergibt sich dann

$$\begin{aligned} &\frac{1}{2}\left(\frac{2\beta T_1 + 3\alpha}{\alpha}A_{m,q}(H) - \frac{2\beta T_1 + \alpha}{\alpha}B_{m,q}(H)\right) \\ &\quad= \frac{3\alpha + 2\beta T_1}{m^2}\sum_{n=1}^{m}\sum_{j=0}^{m-1} X_{q+n+j} - \frac{\alpha + 2\beta T_1}{(m-1)^2}\sum_{n=1}^{m-1}\sum_{j=0}^{m-2} X_{q+n+j} \\ &\quad= \frac{3\alpha + 2\beta T_1}{m^2}\sum_{n=1}^{m}\sum_{j=0}^{m-1} X_{q+n+j} - \frac{\alpha + 2\beta T_1}{(m-1)^2}\sum_{n=1}^{m}\sum_{j=0}^{m-1} X_{q+n+j} \\ &\qquad+ \frac{\alpha + 2\beta T_1}{(m-1)^2}\left[\sum_{j=0}^{m-2} X_{q+m+j} + \sum_{j=0}^{m-2} X_{q+m+j} + X_{q+2m-1}\right] \\ &\quad= \frac{(m-1)^2(2\beta T_1 + 3\alpha) - m^2(2\beta T_1 + \alpha)}{m^2(m-1)^2}\sum_{n=1}^{m}\sum_{j=0}^{m-1} X_{q+n+j} \\ &\qquad+ \frac{(\alpha + 2\beta T_1)}{(m-1)^2}\left(\sum_{j=0}^{m-2} 2X_{q+m+j} + X_{q+2m-1}\right). \end{aligned}$$

Damit ist alles gezeigt. $\square$

Bemerkung 5.4.8 Mit Hinblick auf den Filtervorschlag in (4.62), den wir aus der Differentialgleichung gewonnen haben, und dem aus Folgerung 5.4.7 unterscheiden sich beide wesentlich voneinander. Worin beide jedoch fast übereinstimmen, sind die Darstellung der Doppelsummen. Dies sieht man leicht, wenn man die aus der Folgerung 5.4.7 als Einfachsumme darstellt

$$\sum_{n=1}^{m}\sum_{j=0}^{m-1} X_{q+n+j} = \sum_{j=1}^{m-1} j(X_{q+j} + X_{q+2m-j}) + mX_{q+m}.$$

Natürlich differieren auch die Faktoren vor den Summen wesentlich.

5.4.2 Quadratischer Ansatz

In diesem Kapitel wollen wir nur aufzeigen, wie man einen Filtervorschlag herleiten kann, wenn man eine beste Approximation von Polynomen zweiter Ordnung zugrunde legt. Diese Überlegungen beenden wir mit der Darstellung einer solchen speziellen Funktion U_0 als Summe zweiter Differenzen des partikulären Anteils der Näherungslösung von (5.15), also analog zu den Betrachtungen im vorangegangenen Abschnitt.

Satz 5.4.9 *Es sei $m, q \in \mathbb{N}$, und $U_0(t)$ sei quadratisch im Intervall $[q, q+2m)$ mit $U_0(t) = a + b(t-q) + c(t-q)^2$. Der Term $z_{k,q}$ sei wie in* (5.38) *definiert. Weiterhin sei*

$$A_{m,q} := \frac{2\alpha}{m^2}(z_{q+2m,q} - 2z_{q+m,q} + z_{q,q})$$

und

$$B_{m,q} := \frac{2\alpha}{(m-1)^2}(z_{q+2m-2,q} - 2z_{q+m-1,q} + z_{q,q})$$

sowie

$$C_{m,q} := \frac{2\alpha}{(m-2)^2}(z_{q+2m-4,q} - 2z_{q+m-2,q} + z_{q,q}).$$

Dann gilt

$$\begin{aligned}
\frac{1}{2m}\int_{q}^{2m+q} &\left(a + b\,(\tau - q) + c\,(\tau - q)^2\right) \mathrm{d}\tau \\
&= a + bm + \frac{4}{3}m^2 c \\
&= A_{m,q} + \frac{A_{m,q}}{28\alpha}\left[m(2\alpha m - 2(\alpha + 2\beta T_1)) + 78\beta T_1 + 27\alpha\right] \\
&\qquad + \frac{B_{m,q}}{28\alpha}\left[m(2\alpha m - 2(\alpha + 2\beta T_1)) + 64\beta T_1 + 20\alpha\right] \\
&\qquad + \frac{C_{m,q}}{28\alpha}\left[m(2\alpha m - 2(\alpha + 2\beta T_1)) + 50\beta T_1 + 13\alpha\right]. \qquad (5.44)
\end{aligned}$$

Somit stimmt die rechte Seite von (5.44) *mit dem integralen Mittelwert von $U_0(m+q)$ im Intervall $[q, q+2m)$ überein.*

Beweis: Setzen wir (5.38) in die zweiten Differenzen von $A_{m,q}$, $B_{m,q}$ und $C_{m,q}$ ein, so erhalten wir

$$\begin{aligned}
A_{m,q} &= a + bm + \frac{4}{3}m^2 c - \frac{b}{2\alpha}b\left(\alpha + 2\beta T_1\right) \\
&\quad - \frac{c}{6\alpha^2}\left(7\alpha^2 m^2 - 6\alpha^2 m + 12\beta^2 T_1^2 - 6\alpha\beta T_1 - 12\alpha\beta T_1 m\right), \qquad (5.45)\\
B_{m,q} &= a + (m-1)\,b + \frac{4}{3}(m-1)^2 c - \frac{b}{2\alpha}\left(\alpha + 2\beta T_1\right) \\
&\quad - \frac{c}{6\alpha^2}\left(\alpha^2 m^2 - 6\alpha\beta T_1 + 4\alpha^2 m + 12\alpha\beta T_1 m - 12\beta^2 T_1^2 - 5\alpha^2\right), \\
C_{m,q} &= a + (m-2)\,b + \frac{4}{3}(m-2)^2 c - \frac{b}{2\alpha}\left(\alpha + 2\beta T_1\right) \\
&\quad - \frac{c}{6\alpha^2}\left(12\alpha\beta T_1 m + \alpha^2 m^2 + 2\alpha^2 m - 12\beta^2 T_1^2 - 8\alpha^2 - 18\alpha\beta T_1\right).
\end{aligned}$$

Löst man dieses lineare Gleichungssystem nach a, b und c auf, so folgt unmittelbar der zweite Teil der Gleichung. Der erste ist elementar. Damit ist alles bewiesen. □

Fazit 5.4.10 *Es gilt im Grunde der gleiche Schluß, wie der am Ende des Kapitels* 4. *Um das Eingangssignal* U_0 *möglichst gut digital rekonstruieren zu können, müssen die Ordnung des Konverters und die Dimensionierung der Systemparameter bekannt sein. Eine einfache gleitende Mittellung wie beim Konverter erster Ordnung liefert eine schlechte Approximation.*

6 Betrachtungen zum Phasenporträt eines Konverters zweiter Ordnung

In diesem Kapitel wollen wir uns mit dem Verhalten eines Konverters zweiter Ordnung beschäftigen, welches sich am besten geometrisch anhand des zugrunde liegenden Phasenporträts darstellen läßt. Wir werden Aussagen bezüglich der Beschränktheit der Funktionen $U_1(t)$ und $U_2(t)$ sowie deren periodisches respektive unperiodisches Verhalten unter speziellen Voraussetzungen treffen. Aus physikalischen Gründen bedeutet Stabilität in der Technik, daß beschränkte Eingangssignale auf beschränkte Ausgangssignale abgebildet werden. Interessante Arbeiten zu diesem Thema in der ingenieurtechnischen Literatur sind [55], [24], [25], [59], [12] und [66]. Untersuchungen zur Beschränktheit stehen in engem Zusammenhang zur Periodizität. Arbeiten hierzu sind beispielsweise [22] für Konverter erster und [13] sowie [55] für Konverter zweiter Ordnung. Die Betrachtungen für Konverter zweiter Ordnung basieren auf der Modellbeschreibung im Zeitbereich, wie auch die, die wir in diesem Kapitel vorstellen wollen. Die Resultate sind im wesentlichen geometrisch motiviert, sowie analytisch und durch Simulation bestätigt. Betrachtungen zur Stabilität - im technischen Sinne - im Frequenzbereich werden später in den Kapiteln 7.1 und 8.5 geführt. Zunächst einmal wollen wir einige allgemeine Resultate herleiten.

6.1 Einleitende Resultate

In diesem Kapitel werden wir das zeitliche Verhalten eines Konverters zweiter Ordnung durch eine Folge von Anfangswertproblemen darstellen. Diese Folge von Anfangswertproblemen ist rekursiv, und deren Lösungen beschreiben - intervallweise zusammengesetzt - das zeitliche Verhalten des Konverters zweiter Ordnung. Zur Vereinfachung betrachten wir den einfachsten Fall, d. h., das Eingangssignal $U_0(t)$ sei konstant, analog zu den Betrachtungen von Gray [22]. Im weiteren sei $k \in \mathbb{N}_0$ und $t \in [t_k, t_{k+1})$. Ausgehend von Abbildung 2 und den Formeln (2.30) sowie (2.31) gilt dann

$$
\begin{aligned}
U_1(t) &= \frac{1}{T_1}\int_{t_k}^{t} (U_0(\tau) - \alpha U_{ref} X(\tau))\mathrm{d}\tau + U_1(t_k), \\
U_2(t) &= \frac{1}{T_2}\int_{t_k}^{t} (U_1(\tau) - \beta U_{ref} X(\tau))\mathrm{d}\tau + U_2(t_k).
\end{aligned}
\tag{6.1}
$$

Somit ist $U_0(t) - \alpha U_{ref} X(t)$ wegen (2.33) in $[t_k, t_{k+1})$ stetig. Nach Differentiation von (6.1) ergeben sich für $k \in \mathbb{N}_0$ die Anfangswertprobleme

$$
\begin{aligned}
\dot{U}_1(t) &= \frac{1}{T_1}(U_0(t) - \alpha U_{ref} X(t)), \quad U_1(t_k) = x_k, \\
\dot{U}_2(t) &= \frac{1}{T_2}(U_1(t) - \beta U_{ref} X(t)), \quad U_2(t_k) = y_k.
\end{aligned}
$$

Dabei ist jedoch nur Anfangslage $P_0 = (x_0, y_0)$ gegeben. Die nachfolgenden Anfangswerte errechnen sich für $k \in \mathbb{N}$ durch

$$
\lim_{t \to t_k - 0} U_1(t) =: x_k \quad \text{und} \quad \lim_{t \to t_k - 0} U_2(t) =: y_k
$$

aus den Lösungen der AWA im vorangegangenen Intervall $[t_{k-1}, t_k)$. Durch dieses Zusammenheften der Lösungen entsteht eine stetige Lösung von (6.1) beginnend bei $k = 0$.
Um zum Wesen der Betrachtungen zu gelangen, vereinfachen wir wie folgt. Es sei $T_1 = T_2 =$

$U_{ref} = 1$ und $U_1(t) = x(t)$ sowie $U_2(t) = y(t)$. Darüber hinaus sei im folgenden $U_0(t) \equiv K = const$. Somit lauten unsere Anfangswertprobleme in $t \in [t_k, t_{k+1})$ mit $k \in \mathbb{N}_0$

$$\begin{aligned} \dot{x}(t) &= K - \alpha X(t), \quad x(t_k) = x_k, \\ \dot{y}(t) &= x(t) - \beta X(t), \quad y(t_k) = y_k. \end{aligned} \tag{6.2}$$

Mit dem Vergleichswert des Komparators $s = 0$ und

$$X(t) = 2\sigma(y_k) - 1 \tag{6.3}$$

erhalten wir für $t \in [t_k, t_{k+1})$ zwei autonome Systeme linearer Differentialgleichungen

$$\begin{array}{lclclcl} \dot{x}(t) = K - \alpha & \text{falls} & y_k \geq 0 & \text{und} & \dot{x}(t) = K + \alpha & \text{falls} & y_k < 0 \\ \dot{y}(t) = x(t) - \beta & & & & \dot{y}(t) = x(t) + \beta & & \end{array} \tag{6.4}$$

mit den Anfangswerten

$$x(t_k) = x_k \quad \text{und} \quad y(t_k) = y_k.$$

Das Anfangswertproblem beginnend mit (x_0, y_0) für beliebige Zeiten $t \geq t_0$ ist nicht autonom. Seine Lösung ist rekursiv zusammengesetzt, da die Werte

$$\lim_{t \to t_k - 0} x(t) = x_k \quad \text{und} \quad \lim_{t \to t_k - 0} y(t) = y_k \tag{6.5}$$

die Anfangslage für das Problem im Intervall $[t_k, t_{k+1})$ bilden. Darüber hinaus entscheidet der Wert y_k, nach welchem der beiden Differentialgleichungen von (6.4) sich die Lösung im Intervall $[t_k, t_{k+1})$ berechnet.
Durch das Zusammenheften via (6.5) und durch (6.3) erhalten wir die AWA, deren Lösung eindeutig durch die Anfangslage (x_0, y_0) bestimmt ist

$$\begin{aligned} \dot{x}(t) &= K - \alpha X(t), \quad x(t_0) = x_0, \\ \dot{y}(t) &= x(t) - \beta X(t), \quad y(t_0) = y_0. \end{aligned} \tag{6.6}$$

Ohne großen Rechenaufwand ergeben sich die beiden Lösungen von (6.4). Für $k \in \mathbb{N}_0$ und $y_k \geq 0$ gilt

$$\begin{aligned} x(t) &= (K - \alpha)(t - t_k) + x_k, \\ y(t) &= (K - \alpha)\frac{(t - t_k)^2}{2} + (x_k - \beta)(t - t_k) + y_k \end{aligned} \tag{6.7}$$

bzw.

$$y(x) = \frac{1}{K - \alpha}\left(\frac{(x - x_k)^2}{2} + (x_k - \beta)(x - x_k)\right) + y_k. \tag{6.8}$$

Für $y_k < 0$ gilt dagegen

$$\begin{aligned} x(t) &= (K + \alpha)(t - t_k) + x_k, \\ y(t) &= (K + \alpha)\frac{(t - t_k)^2}{2} + (x_k + \beta)(t - t_k) + y_k \end{aligned} \tag{6.9}$$

bzw.

$$y(x) = \frac{1}{K + \alpha}\left(\frac{(x - x_k)^2}{2} + (x_k + \beta)(x - x_k)\right) + y_k. \tag{6.10}$$

Bemerkung 6.1.1 Aus (6.8) und (6.10) ist ersichtlich, daß die Lösung von (6.6) aus Parabelstücken in der x-y-Ebene besteht. Die Parabeln aus (6.8) haben ihre Scheitelpunkte auf der Geraden $x = \beta$ und die aus (6.10) auf $x = -\beta$. Aus diesem Grund werden wir im folgenden auch von „oberen" bzw. von „unteren" Bahnen sprechen.
Wenn wir zur geometrischen Motivation in unseren Beweisen das Phasenporträt von (6.6) benutzen, so besteht dies wegen (6.4) aus zwei zusammengesetzten Phasenportäts. Diese „durchdringen" sich gerade soweit, wie die Lösung die x-Achse in einem Takt Δt „überspringen" kann. Hierin spiegelt sich die Diskretheit des Modells wider.
Weiterhin folgt aus (6.7) bzw. (6.9), daß die oberen bzw. unteren Lösungskurven in x-Richtung in einem Takt Δt um $K - \alpha$ bzw. $K + \alpha$ durchlaufen werden. Dies ist ein Aspekt, den wir später bei den allgemeinen Betrachtungen und denen zur Periodizität berücksichtigen werden.

Definition 6.1.2 *Die Lösung von (6.6) heißt beschränkt, falls ein $M \in (0, +\infty)$ existiert, so daß $|x(t)| + |y(t)| \leq M$ für alle Zeiten $t \geq t_0$ erfüllt ist. Existiert kein solches M, so heißt die Lösung unbeschränkt.*

Nun einige Aussagen, die leicht beweisbar sind und unmittelbar aus der Definition 6.1.2 folgen.

Folgerung 6.1.3 Es sei $|K| > \alpha$. Dann ist die Lösung von (6.6) für $t \to +\infty$ unbeschränkt.

Beweis: Es sei $K > \alpha$. Dann sind die Parabeln wegen (6.8) und (6.10) nach oben offen. Dabei werden die Parabeln wegen (6.7) und (6.9) von links nach rechts in der x-y-Ebene durchlaufen. Das Phasenporträt von (6.6) stellt sich somit wie in Abbildung 24 dar. Wegen

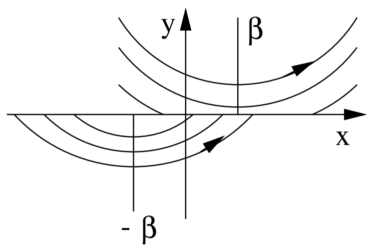


Abbildung 24: Phasenporträt für $K > \alpha$

(6.7) und (6.9) gilt nach Annahme, daß $(x_k)_{k\in\mathbb{N}_0}$ eine streng monoton wachsende Folge mit $x_{k+1} - x_k \geq (K - \alpha)\Delta t > 0$ ist. Damit ist $x(t) \to +\infty$ für $t \to +\infty$. Somit existiert ein $k_0 \in \mathbb{N}_0$ mit $x_k \geq \beta$ für alle $k \geq k_0$. Wegen (6.7) und (6.9) gilt dann, daß $(y_k)_{k\geq k_0}$ eine streng monoton wachsende Folge ist. Für sie ist $y_{k+1} - y_k \geq (K - \alpha)\frac{\Delta t^2}{2} + (x_{k_0} - \beta)\Delta t > 0$ für alle $k \geq k_0$. Somit gilt $y(t) \to +\infty$ für $t \to +\infty$. Daraus folgt die Behauptung für $K > \alpha$.
Der Fall $K < -\alpha$ ist analog beweisbar. $\square$

Folgerung 6.1.4 Es gelten folgende Aussagen:
a) Es sei $K = \alpha$. Weiterhin existiere ein $k \in \mathbb{N}_0$ so, daß gleichzeitig $x_k = \beta$ und $y_k \geq 0$ erfüllt sind. Dann gilt $x(t) = \beta$ und $y(t) = y_k$ für alle Zeiten $t \geq t_k$.
b) Es sei $K = -\alpha$. Weiterhin existiere ein $k \in \mathbb{N}_0$ so, daß gleichzeitig $x_k = -\beta$ und $y_k < 0$ erfüllt sind. Dann gilt $x(t) = -\beta$ und $y(t) = y_k$ für alle Zeiten $t \geq t_k$.
Somit ergeben sich in diesen beiden Fällen stationäre Lösungen.

Beweis: Auch hier ist der Beweis elementar und folgt unmittelbar aus (6.4). $\square$

Folgerung 6.1.5 Es gelten folgende Aussagen:
a) Es sei $K = \alpha$. Weiterhin existiere kein $k \in \mathbb{N}_0$ so, daß gleichzeitig $x_k = \beta$ und $y_k \geq 0$ erfüllt sind. Dann ist die Lösung von (6.6) für $t \to +\infty$ unbeschränkt. In diesem Fall gilt $x(t) \to x_n > \beta$ für ein $n \in \mathbb{N}_0$ und $y(t) \to +\infty$ jeweils für $t \to +\infty$.
b) Es sei $K = -\alpha$. Weiterhin existiere kein $k \in \mathbb{N}_0$ so, daß gleichzeitig $x_k = -\beta$ und $y_k < 0$ erfüllt sind. Dann ist die Lösung von (6.6) für $t \to +\infty$ unbeschränkt. In diesem Fall gilt $x(t) \to x_n < -\beta$ für ein $n \in \mathbb{N}_0$ und $y(t) \to -\infty$ jeweils für $t \to +\infty$.

Beweis: a) Aus (6.7) und (6.10) erhalten wir das Phasenporträt, wie in Abbildung 25 dargestellt.

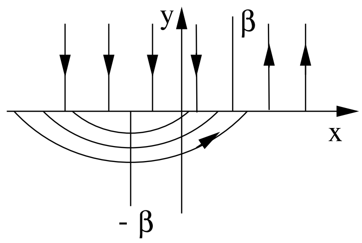


Abbildung 25: Phasenporträt für $K = \alpha$

Es sei $x_0 > \beta$ und $y_0 \geq 0$. Aus (6.6) folgt dann

$$\begin{aligned} \dot{x}(t) &= 0, \\ \dot{y}(t) &= x_0 - \beta > 0. \end{aligned} \tag{6.11}$$

Wegen $y_0 \geq 0$ folgt aus (6.7) dann $y_1 > 0$ und $x_1 = x_0$. Damit sind wir wieder bei (6.11) und wegen (6.7) gilt schließlich $x(t) \equiv x_0$ und $y(t) \to +\infty$ für $t \to +\infty$.
Sei nun $x_0 < \beta$ und o. B. d. A. $y_0 \geq 0$. Wegen (6.6) gilt dann

$$\begin{aligned} \dot{x}(t) &= 0, \\ \dot{y}(t) &= x_0 - \beta < 0. \end{aligned} \tag{6.12}$$

Wegen $y_0 \geq 0$ folgt aus (6.7) dann $y_1 < y_0$ und $x_1 = x_0$. Dann ist zu entscheiden, ob y_1 kleiner als Null ist oder nicht. Ist dies nicht der Fall, so gilt wiederum (6.12) und zwar so lange, bis $y_m < 0$ ist. In diesem Fall folgt aus (6.6) dann

$$\begin{aligned} \dot{x}(t) &= 2K, \\ \dot{y}(t) &= x_0 + \beta > 0. \end{aligned} \tag{6.13}$$

Nach (6.9) ist $x(t)$ für $t \geq t_m$ streng monoton wachsend mit der Geschwindigkeit $\dot{x} = 2K$. Aus diesem Grund existiert eine Zeit t_s mit $x_s \geq -\beta$ und $y_s < 0$. Wiederum aus (6.9) folgt, daß $y(t)$ für $t \geq t_s$ streng monoton wachsend ist. Hieraus folgt die Existenz eines $y_r > 0$ mit $y_{r-1} \leq 0$. Gilt darüber hinaus $x_r = x_n > \beta$ - Gleichheit ist nach Voraussetzung ausgeschlossen -, so sind wir bei der Diskussion von (6.11) und erhalten die Behauptung. Gilt aber immer noch $x_r < \beta$, so setzen wir mit unseren Überlegungen wieder bei (6.12) ein. Dieser Prozeß bricht ab, wenn $x_n > \beta$ erfüllt ist. Damit ist die Behauptung von a) bewiesen.
b) Die Behauptung folgt aus analogen Überlegungen wie in a). □

Bemerkung 6.1.6 In allen weiteren Untersuchungen werden wir stillschweigend $|K| < \alpha$ voraussetzen. Ist diese Voraussetzung nicht erfüllt, so ist nach den Folgerungen 6.1.3 bis 6.1.5 jede Diskussion über Beschränktheit von Lösungen und Periodizität dieser, was ja Beschränktheit voraussetzt, sinnlos.

6.2 Aussagen zur Beschränktheit

Es sei $|K| < \alpha$. Ziel der folgenden Überlegungen ist es, ein Gebiet G anzugeben, so daß die Lösung $(x(t), y(t))$ der in Abschnitt 6.1 beschriebenen AWA mit der Anfangslage $(x_0, y_0) \in \mathbb{R}^2$ das Gebiet G für $t \geq \tilde{t}$ nie mehr verläßt, falls es einen Zeitpunkt $\tilde{t} \geq t_0$ mit $(x(\tilde{t}), y(\tilde{t})) \in G$ gibt. Eine geometrische Motivation gibt die Abbildung 26. An ihr wollen wir unsere Ideen erläutern.

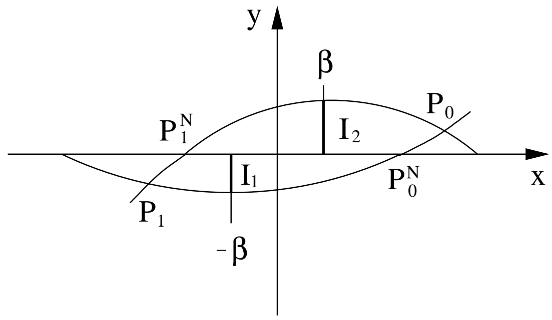


Abbildung 26: Geometrische Motivation für die Beschränktheit

Bemerkung 6.2.1 Da sich die Lösung von (6.6) wegen (6.8) und (6.10) auf Parabeln bewegt, ist das Gebiet G, das wir suchen, durch eine obere und eine untere Parabel berandet. In G liegen somit alle oberen Parabeln mit einem Scheitelpunkt in I_2. Andererseits enthält G alle unteren Parabeln mit einem Scheitelpunkt in I_1. Wenn wir fordern, daß die Lösung von (6.6) G nicht verlassen soll, so bedeutet das: Mit einem Takt Δt kann die Lösung, welche entlang einer zulässigen unteren Parabel läuft, nicht auf eine obere Bahn gelangen, die oberhalb von der liegt, die durch P_0 eindeutig festgelegt ist. Genauso kann sie nicht in einem Takt auf eine untere Bahn gelangen, die unterhalb der liegt, die ebenfalls durch P_0 eindeutig bestimmt ist.
Es seien $A, B \in \mathbb{R}^2$ zwei Punkte, die auf einer Parabel liegen. Bezeichne

$$T(A, B)$$

die Zeit, die die Lösung von A nach B entlang einer der beiden Parabeltypen benötigt - Umlaufrichtung entgegen des Uhrzeigersinns -, so liegt es nahe, ein Gebiet G zu beschreiben, welches folgende Eigenschaften besitzt:

a) Es ist $T(P_0^N, P_0) = T(P_1^N, P_1) = \Delta t$.

b) Es bezeichne P_i die linke Nullstelle einer zulässigen oberen Parabel und P_a den linken Schnittpunkt dieser mit der unteren Bahn. Dann gilt $T(P_i, P_a) \geq \Delta t$. In Abbildung 26 sind die Punkte für die oberste zulässige obere Bahn speziell durch P_1^N und P_1 gekennzeichnet.

c) Es bezeichne P_i die rechte Nullstelle einer zulässigen unteren Parabel und P_a den rechten Schnittpunkt dieser mit der oberen Bahn. Dann gilt $T(P_i, P_a) \geq \Delta t$. In Abbildung 26 sind die Punkte für die unterste zulässige untere Bahn speziell durch P_0^N und P_0 gekennzeichnet.

Wegen a) sind also beide Parabeln, die G beranden, allein durch den Punkt P_0 eindeutig festgelegt. Als erstes wollen wir P_0 in Abhängigkeit von Δt beschreiben. Hierbei muß P_0 nicht identisch mit der Anfangslage unserer AWA sein.

Lemma 6.2.2 *Es sei $|K| < \alpha$ und $P_0 = (x_0, y_0)$. Dann wird die Lage des rechten Schnittpunktes einer unteren und einer oberen Parabel, für die $T(P_0^N, P_0) = T(P_1^N, P_1) = \Delta t$ gilt, parametrisch* *durch*

$$\begin{aligned} x_0(\Delta t) &= \frac{-2K\beta^2 + 2K\alpha\beta\Delta t - 2\alpha\beta\Delta t + \alpha^3\Delta t^2}{2\alpha(\alpha\Delta t - \beta)}, \\ y_0(\Delta t) &= -\frac{\Delta t(\alpha^2 K\Delta t^2 - 3\alpha\beta K\Delta t + 2\beta^2 K - \alpha^2\beta\Delta t + 2\alpha\beta^2)}{2\alpha(\alpha\Delta t - \beta)} \end{aligned} \tag{6.14}$$

beschrieben.

Beweis: Dem Beweis liegt folgende Idee zugrunde. Es sei $P_0 = (x_0, y_0) \in \mathbb{R}^2$ beliebig. Nun berechnen wir die Punkte P_1^N, P_1 und P_0^N in Abhängigkeit von (x_0, y_0). Danach bestimmen wir $T(P_1^N, P_1)$ und $T(P_0^N, P_0)$. Beide Zeiten setzen wir gleich der Zeit Δt und lösen nach x_0 und y_0 auf.
Wir nehmen o. B. d. A. an, daß $y_0 \geq 0$ ist. Dann muß $x_0 > -\beta$ sein, da der Scheitelpunkt der unteren Parabeln $-\beta$ ist (vgl Abb. 26). Mit (6.8) und (6.10) ergibt sich unmittelbar

$$\begin{aligned} P_1^N &= \left(\beta - \sqrt{(\beta - x_0)^2 + 2y_0(\alpha - K)}, 0\right), \\ P_1 &= \left(-x_0 + \frac{2\beta K}{\alpha}, y_0 + \frac{2\beta}{\alpha}\left(\frac{\beta K}{\alpha} - x_0\right)\right), \\ P_0^N &= \left(-\beta + \sqrt{(\beta + x_0)^2 - 2y_0(\alpha + K)}, 0\right). \end{aligned} \tag{6.15}$$

Mit den Formeln (6.7) und (6.9) folgt aus (6.15) dann

$$\begin{aligned} T(P_1^N, P_1) &= \frac{1}{\alpha - K}\left(x_0 + \beta - \sqrt{(\beta - x_0)^2 + 2y_0(\alpha - K)} - \frac{2\beta K}{\alpha(\alpha - K)}\right), \\ T(P_0^N, P_0) &= \frac{1}{\alpha + K}\left(x_0 + \beta - \sqrt{(\beta + x_0)^2 - 2y_0(\alpha + K)}\right). \end{aligned} \tag{6.16}$$

Nun lösen wir das Gleichungssystem

$$\begin{aligned} T(P_1^N, P_1) &= \Delta t, \\ T(P_0^N, P_0) &= \Delta t \end{aligned}$$

nach (x_0, y_0) auf. Mit (6.16) ergibt sich dann

$$\begin{aligned} x_0(\Delta t) &= \frac{-2K\beta^2 + 2K\alpha\beta\Delta t - 2\alpha\beta\Delta t + \alpha^3\Delta t^2}{2\alpha(\alpha\Delta t - \beta)}, \\ y_0(\Delta t) &= -\frac{\Delta t(\alpha^2 K\Delta t^2 - 3\alpha\beta K\Delta t + 2\beta^2 K - \alpha^2\beta\Delta t + 2\alpha\beta^2)}{2\alpha(\alpha\Delta t - \beta)}. \end{aligned} \tag{6.17}$$

Mit diesem Punkt sind nun die beiden Parabeln bestimmt, auf denen die Lösung von der x-Achse bis zu einem der Punkte P_0 bzw. P_1 die gleiche Zeit braucht (vgl. Abb. 26). □

Folgerung 6.2.3 Es sei $|K| < \alpha$. Dann gelten folgende Aussagen:
a) Die Lage des rechten Schnittpunktes einer unteren und einer oberen Parabel, für welche $T(P_0^N, P_0) = T(P_1^N, P_1) = \Delta t$ gilt, ist durch

$$\begin{aligned} y_0(x_0) = {} & \frac{1}{2\alpha^4(x_0\alpha - K\beta - M)}(-x_0\alpha + K\beta - \alpha\beta + M) \times \\ & \times \big(2Kx_0^2\alpha^2 - 4x_0\alpha K^2\beta - Kx_0\alpha^2\beta - 2Kx_0\alpha M + 2K^3\beta^2 + K^2\beta^2\alpha \\ & + 2K^2\beta M + 2K\beta^2\alpha^2 + K\alpha\beta M - \beta\alpha^3 x_0 + \beta^2\alpha^3 + \beta\alpha^2 M\big) \end{aligned} \tag{6.18}$$

mit

$$M = \sqrt{(x_0\alpha - K\beta)^2 + \alpha^2\beta^2}$$

explizit gegeben.
b) Es gilt

$$(x_0(\Delta t), y_0(\Delta t)) \to \left(\frac{\beta K}{\alpha}, 0\right) \text{ für } \Delta t \to 0. \tag{6.19}$$

c) Es ist

$$\begin{aligned}
\Delta t(x_0) &= \frac{2x_0\alpha - 2K\beta + 2\alpha\beta - 2\sqrt{x_0^2\alpha^2 - 2x_0\alpha K\beta + K^2\beta^2 + \alpha^2\beta^2}}{2\alpha^2}\\
\Delta t(y_0) &= \frac{1}{3\alpha K}\left((-9\beta\alpha^3 K y_0 - 9\beta^3\alpha K^2 + \beta^3\alpha^3 + 3QK)^{\frac{1}{3}}\right)\\
&\quad + \frac{-6Ky_0\alpha^2 + 3K^2\beta^2 + \beta^2\alpha^2}{3\alpha K\left((-9\beta\alpha^3 K y_0 - 9\beta^3\alpha K^2 + \beta^3\alpha^3 + 3QK)^{\frac{1}{3}}\right)} + \frac{\beta(3K+\alpha)}{3\alpha K}.
\end{aligned}$$

mit

$$\begin{aligned}
Q &= \left(24Ky_0^3\alpha^6 - 36K^2y_0^2\alpha^4\beta^2 - 3y_0^2\alpha^6\beta^2 + 18K^3y_0\alpha^2\beta^4 + 30Ky_0\alpha^4\beta^4 - 3K^4\beta^6\right.\\
&\quad \left. + 6K^2\beta^6\alpha^2 - 3\beta^6\alpha^4\right)^{\frac{1}{2}}.
\end{aligned}$$

d) Es gilt

$$\begin{aligned}
\Delta t(x_0) &\to \frac{\beta}{\alpha} \text{ für } x_0 \to +\infty,\\
\Delta t(y_0) &\to \frac{\beta}{\alpha} \text{ für } y_0 \to +\infty.
\end{aligned} \tag{6.20}$$

e) Asymptotisches Verhalten ist charakterisiert durch

$$\frac{x_0(\Delta t)}{y_0(\Delta t)} \to \frac{\alpha}{\beta} \text{ für } \Delta t \to \frac{\beta}{\alpha}.$$

Beweis: Alle Behauptungen folgen unmittelbar aus dem Lemma 6.2.2 bzw. nach Auflösung von $x_0(\Delta t)$ nach Δt und Einsetzen von $\Delta t(x_0)$ in $y_0(\Delta t)$. □

Bemerkung 6.2.4 Bevor wir zum Hauptresultat dieses Kapitels kommen, wollen wir eine Fallunterscheidung angeben, die sich nach den obigen Ideen in natürlicher Weise ergibt. Entscheidendes Merkmal dabei ist $P_i = P_1^N$ bzw. $P_i \neq P_1^N$, wie in den Abbildungen 27 und 28 dargestellt ist. Wir werden die Aussagen für alle vier Fälle angeben, aber nur die der Abbildung 27 ausführlich beweisen. Bezeichne x_{st} die rechte Nullstelle der oberen Parabel in Bild 27. Dann gilt

$$\begin{aligned}
&P_1^N \text{ liegt links von } P_i &&\text{für } x_{st} > 3\beta,\\
&P_1^N = P_i &&\text{für } 2\beta - \frac{\beta K}{\alpha} < x_{st} \leq 3\beta.
\end{aligned}$$

Dabei darf x_{st} nicht kleiner als $2\beta - \dfrac{\beta K}{\alpha}$ sein, da es sonst keinen zugehörigen Punkt P_1, wie in Lemma 6.2.2 erklärt, gibt.
In den folgenden vier Aussagen gibt es jeweils einen Aspekt, der analytisch schwer zu beschreiben ist. Aus diesem Grund benutzen wir numerische Resultate, die unsere Aussagen stützen. Wir

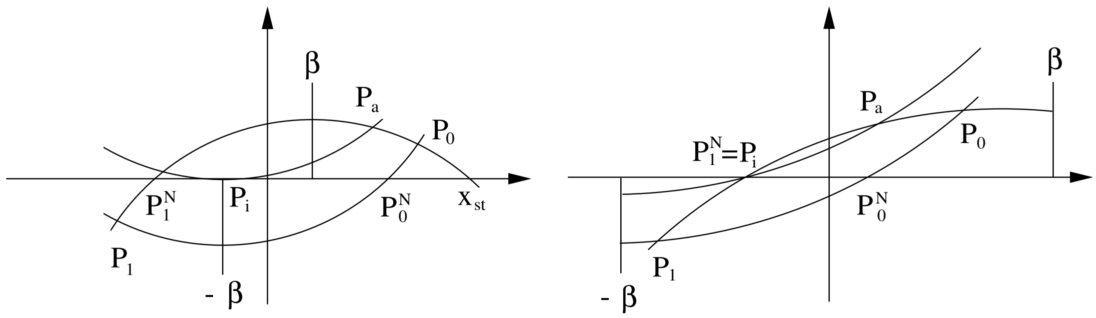


Abbildung 27: Die beiden Fälle $y_0 \geq 0$ mit $x_{st} \geq 3\beta$ und $x_{st} < 3\beta$

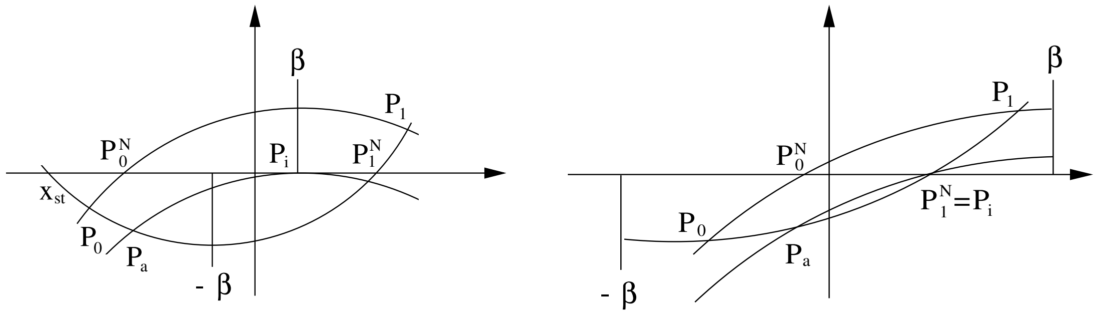


Abbildung 28: Die beiden Fälle $y_0 < 0$ mit $x_{st} \leq -3\beta$ und $x_{st} > -3\beta$

werden jedoch an den entsprechenden Stellen noch einmal darauf aufmerksam machen.
Betrachten wir die erste Darstellung in Abbildung 27, so ist wegen Lemma 6.2.2 und Folgerung 6.2.3 leicht ersichtlich, daß P_0, x_{st} und Δt identifiziert werden können. In diesem Sinne werden wir die folgenden Resultate nun darstellen.

Lemma 6.2.5 *Es sei $x_{st} \geq 3\beta$. Die Bezeichnungen seien entsprechend der ersten Darstellung der Abbildung 27 gewählt. Dann gilt*

$$T(P_i, P_a) \geq \Delta t$$

für alle unteren Parabeln mit Scheitelpunkt in I_1.

Beweis: Zur Veranschaulichung diene die Abbildung 29.
Bezeichne Σ_1 die Menge aller rechten Nullstellen der unteren Parabeln, welche mit der oberen einen Schnittpunkt haben, der links von P_0 liegt. Zusätzlich sollen sich diese Schnittpunkte nicht unterhalb von y_0 befinden. In Abbildung 29 sind dies alle Punkte der oberen Parabel oberhalb der gestrichelten Linie. Betrachten wir nun eine untere Parabel Π, deren rechte Nullstelle im Intervall Σ_1 liegt. Da alle unteren Parabeln zueinander parallel sind, ist somit der Anstieg von Π rechts ihrer Nullstelle kleiner oder gleich zu der der unteren Bahn zwischen P_0^N und P_0. Also ist der Weg in x-Richtung, um von der Nullstelle zur gestrichelten Linie zu gelangen, auf Π größer oder gleich als auf der unteren Bahn, die durch P_0 geht. Nach (6.9) gilt dann für Π

$$T(P_i, P_a) \geq \Delta t = T(P_0^N, P_0). \tag{6.21}$$

Da Π eine beliebige untere Bahn ist, die in Σ_1 eine rechte Nullstelle hat, ist die Behauptung für dieses Intervall gezeigt.

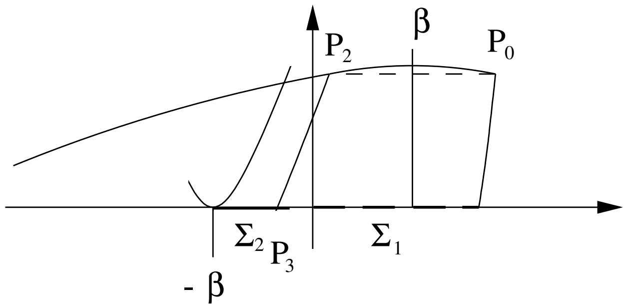


Abbildung 29: Geometrische Motivation für ersten Fall

Dieser geometrische Beweis ist nur für das Intervall Σ_1 möglich. Bezeichne nun Σ_2 die Menge aller rechten Nullstellen unterer Parabeln, deren rechter Schnittpunkt mit der oberen Parabel links von P_0 und kleiner als y_0 ist. Wir nehmen an, daß Σ_2 nicht leer sei. Dies ist äquivalent dazu, daß für $P_3 = (x_3, y_3)$

$$\begin{aligned} &x_0 \geq \beta, \\ &x_3 \geq -\beta \end{aligned} \tag{6.22}$$

erfüllt sind, wobei x_3 die rechte Nullstelle der unteren Parabel ist, die die obere im Punkt $P_2 = (-x_0 + 2\beta, y_0)$ schneidet. Wir wollen alle folgenden Abhängigkeiten als solche von x_{st}, d. h. von Lage der rechten Nullstelle der oberen Parabel, darstellen. Die rechte Nullstelle der oberen Bahn lautet

$$x_{st} = \beta - \sqrt{(\beta - x_0)^2 + 2y_0(\alpha - K)}.$$

Setzen wir $y_0(x_0)$ nach Formel (6.18) in x_{st} ein und lösen dies nach $x_0(x_{st})$ auf, so gilt für die erste Bedingung aus (6.22) dann

$$x_0(x_{st}) = \frac{-\alpha K x_{st} + 2\alpha\beta K - 2\beta K^2 + R}{\alpha(\alpha - 2K)} \geq \beta \tag{6.23}$$

mit

$$\begin{aligned} R \;=\; &\big(\alpha^2 K^2 x_{st}^2 - 2\alpha^2 \beta K^2 x_{st} - 4\alpha^2\beta^2 K^2 + 2\alpha\beta^2 K^3 + 4\alpha^3\beta K x_{st} - 2\alpha^3 K x_{st}^2 \\ &-2\alpha^4\beta x_{st} + \alpha^4 x_{st}^2 + 2\alpha^3\beta^2 K\big)^{\frac{1}{2}}. \end{aligned}$$

Das Umstellen von (6.23) nach K liefert

$$\alpha\left(\frac{S}{6\beta} + \frac{2(x_{st} - 4\beta)(x_{st} - \beta)}{3\beta V} - \frac{x_{st} - 4\beta}{3\beta}\right) \geq K \tag{6.24}$$

mit

$$\begin{aligned} S \;=\; &\big(6\beta x_{st}^2 + 12\beta^2 x_{st} - 10\beta^3 - 8x_{st}^3 \\ &+6\sqrt{3\beta}\sqrt{148\beta^2 x_{st}^3 - 53\beta x_{st}^4 + 8x_{st}^5 - 206\beta^3 x_{st}^2 + 140\beta^4 x_{st} - 37\beta^5}\bigg)^{\frac{1}{3}}. \end{aligned}$$

Setzen wir (6.23) in (6.10) ein, so erhalten wir

$$\begin{aligned} y_0(x_{st}) &= \frac{1}{2(K-\alpha)}\left(\frac{-\alpha K x_{st}+2\alpha\beta K-2\beta K^2+R}{\alpha(\alpha-2K)}-x_{st}\right)^2 \\ &\quad +(x_{st}-\beta)\left(\frac{-\alpha K x_{st}+2\alpha\beta K-2\beta K^2+R}{\alpha(\alpha-2K)}-x_{st}\right). \end{aligned} \tag{6.25}$$

Mit Hilfe von (6.23) und (6.25) können wir nun die rechte Nullstelle x_3 der unteren Parabel berechnen, welche die obere in P_2 schneidet. Dabei soll

$$\begin{aligned} x_3(x_{st}) &= \frac{1}{(-\alpha K+\alpha^2)(-2K+\alpha)}\left[3\alpha^2K\beta-\alpha^3\beta-2\alpha K^2\beta\right. \\ &\quad +\left(9\alpha^6\beta^2+5\alpha^4K^2x_{st}^2+4\alpha^2K^4x_{st}^2-8\alpha^3K^3x_{st}^2+128\alpha^2K^4\beta^2\right. \\ &\quad -232\alpha^3K^3\beta^2-66\alpha^5K\beta^2+185\alpha^4K^2\beta^2+4\alpha^2K^2Tx_{st}+36\alpha^3KT\beta \\ &\quad +16\alpha K^3T\beta-24\alpha^2K^4x_{st}\beta-42\alpha^4K^2x_{st}\beta+56\alpha^3K^3x_{st}\beta-8\alpha^4T\beta \\ &\quad -2\alpha^5Kx_{st}^2-24\alpha K^5\beta^2-4\alpha^3TKx_{st}+\alpha^6x_{st}^2+12\alpha^5Kx_{st}\beta \\ &\quad \left.\left.-44\alpha^2K^2T\beta-2\alpha^6x_{st}\beta\right)^{\frac{1}{2}}\right] \geq -\beta \end{aligned} \tag{6.26}$$

mit

$$\begin{aligned} T &= \left(\alpha^2K^2x_{st}^2-2\alpha^2K^2x_{st}\beta-4K^2\beta^2\alpha^2+2\alpha K^3\beta^2+4K\alpha^3x_{st}\beta-2K\alpha^3x_{st}^2\right. \\ &\quad \left.-2\alpha^4x_{st}\beta+\alpha^4x_{st}^2+2\alpha^3\beta^2K\right)^{\frac{1}{2}} \end{aligned}$$

erfüllt sein. Maple ist nun nicht mehr in der Lage, (6.26) nach K aufzulösen. Somit müssen wir es mit der Bedingung (6.26) bewenden lassen. Im folgenden seien die Bedingungen (6.23) und (6.26) erfüllt. Dann ist $\Sigma_2 \neq \emptyset$. Weiterhin sei $s \in \Sigma_2$. Dann ist der rechte Schnittpunkt der unteren Parabel mit rechter Nullstelle in s und der oberen Parabel mit rechter Nullstelle in $x_{st} \geq 3\beta$ in

$$x(s) = \frac{1}{2\alpha}\left(2\beta K + U(s)\right)$$

mit

$$\begin{aligned} U(s) &= \left(4\beta^2K^2+4s\beta\alpha^2-2\alpha s^2K+2s^2\alpha^2-4\alpha s\beta K-4x_{st}\beta\alpha^2+2\alpha x_{st}^2K\right. \\ &\quad \left.+2x_{st}^2\alpha^2-4\alpha x_{st}\beta K\right)^{\frac{1}{2}}. \end{aligned}$$

Für die Zeit, die die Lösung zwischen s und $x(s)$ benötigt, gilt dann

$$t(s) = \frac{1}{2\alpha(\alpha+K)}\left(2\beta K + U(s) - 2\alpha s\right).$$

Eine gleichmäßige numerische Abtastung des zulässigen Parameterbereiches ergab, daß

$$t(s) \geq \Delta t \quad \forall s \in \Sigma_2 \tag{6.27}$$

erfüllt ist. Aus (6.21) und (6.27) ergibt sich dann die Behauptung. □

Nun wollen wir den zweiten Fall betrachten, bei dem nun $x_{st} \in \left(2\beta - \frac{\beta K}{\alpha}, 3\beta\right]$ gilt.

Lemma 6.2.6 *Es sei* $x_{st} \in \left(2\beta - \frac{\beta K}{\alpha}, 3\beta\right]$. *Weiterhin sei* $\Delta t \geq \frac{\beta(K-\alpha)}{\alpha(K-2\alpha)}$. *Die Bezeichnungen seien entsprechend der zweiten Darstellung der Abbildung 27 gewählt. Dann gilt*

$$T(P_i, P_a) \geq \Delta t$$

für alle unteren Parabeln mit Scheitelpunkt in I_1.

Beweis: Es sei $x_{st} \leq 3\beta$ die rechte Nullstelle der oberen Parabel. Dann liegt die linke bei

$$-x_{st} + 2\beta.$$

Dieser Wert ist somit gleichzeitig die rechte Nullstelle der obersten unteren Parabel Π. Der rechte Schnittpunkt von Π und der oberen Parabel ist dann in

$$x_{st} - 2\beta + \frac{2\beta K}{\alpha}.$$

Somit gilt für Π

$$T(P_i, P_a) = 2\frac{\alpha x_{st} - 2\alpha\beta + \beta K}{\alpha(\alpha + K)}. \tag{6.28}$$

Nun stellt sich die Frage, ob $T(P_i, P_a) \geq \Delta t$ gilt, wenn x_{st} das Intervall $\left(2\beta - \frac{\beta K}{\alpha}, 3\beta\right]$ durchläuft. Dafür müssen wir x_{st} in Abhängigkeit von Δt darstellen. Benutzen wir die Formeln in (6.17) und berechnen mit ihnen die rechte Nullstelle x_{st} der oberen Parabel, so erhalten wir

$$x_{st}(\Delta t) = -\frac{-2\beta^2 K + 4\beta\Delta t K\alpha + 4\beta^2\alpha - 2\alpha^2\Delta t^2 K - 4\alpha^2\beta\Delta t + \Delta t^2\alpha^3}{2\alpha(\Delta t\alpha - \beta)}. \tag{6.29}$$

Setzen wir $x_{st}(\Delta t)$ in (6.28) ein, so ergibt sich

$$T(P_i, P_a) = f(\Delta t) = \frac{\Delta t(-2\beta K + 2\alpha\Delta t K - \Delta t\alpha^2)}{(\Delta t\alpha + \beta)(\alpha K)}. \tag{6.30}$$

Die Funktion $f(\Delta t)$ besitzt folgende Eigenschaften. Für

$$\Delta t \in \left\{0, \frac{\beta(K-\alpha)}{\alpha(K-2\alpha)}\right\} \subset [0, +\infty)$$

ist

$$f(\Delta t) = \Delta t.$$

Die Funktion $f(\Delta t)$ ist streng konvex. Für ihre zweite Ableitung gilt

$$f(\Delta t) = -\frac{2\alpha^2\beta^2}{(\alpha\Delta t - \beta)^3(K+\alpha)} > 0,$$

da dies nach Voraussetzung äquivalent mit $\Delta t \in \left[0, \frac{\beta}{\alpha}\right)$ ist. Wo liegt aber nun der Wert $\frac{\beta(K-\alpha)}{\alpha(K-2\alpha)}$? Entspricht er einem $x_{st} \in \left(2\beta - \frac{\beta K}{\alpha}, 3\beta\right]$ oder nicht? Lösen wir (6.29) nach $\Delta t(x_{st})$ auf und berechnen den Wert an der Stelle $x_{st} = 3\beta$, dann gilt

$$\Delta t_{3\beta} = \frac{(-\alpha - 2K + \sqrt{3\alpha^2 + 2K\alpha})\beta}{(-2K + \alpha)\alpha}. \tag{6.31}$$

Somit ist $x_{st} \in \left(2\beta - \frac{\beta K}{\alpha}, 3\beta\right]$ gleichbedeutend mit $\Delta t \in (0, \Delta t_{3\beta}]$. Wegen

$$f(\Delta t_{3\beta}) = 2\frac{\beta}{\alpha} > \Delta t$$

und der strengen Konvexität von $f(\Delta t)$, gilt dann

$$\frac{\beta(K-\alpha)}{\alpha(K-2\alpha)} \in (0, \Delta t_{3\beta}].$$

Hieraus folgt

$$f(\Delta t) \geq \Delta t \quad \text{für} \quad \Delta t \in \left(\frac{\beta(K-\alpha)}{\alpha(K-2\alpha)}, \Delta t_{3\beta}\right].$$

Damit erfüllt die oberste untere Parabel, welche ja von Δt abhängt, die Behauptung

$$T(P_i, P_a) \geq \Delta t \quad \text{für} \quad \Delta t \in \left(\frac{\beta(K-\alpha)}{\alpha(K-2\alpha)}, \Delta t_{3\beta}\right].$$

Im weiteren sei $\Delta t \in \left(\frac{\beta(K-\alpha)}{\alpha(K-2\alpha)}, \Delta t_{3\beta}\right]$. Nun wissen wir, daß nach den Voraussetzungen $T(P_i, P_a) \geq \Delta t$ für die oberste und unterste untere Parabel gilt. Gilt dies auch für die, die zwischen diesen beiden Bahnen liegen? Nach gleichen Rechnungen wie im Beweis von Lemma 6.2.5 gilt für die Zeit zwischen rechter Nullstelle einer unteren Parabel und ihrem Schnittpunkt mit der oberen Bahn

$$t(s) = \frac{1}{2\alpha(\alpha+K)}\left(2\beta K + U(s) - 2\alpha s\right)$$

mit

$$\begin{aligned} U(s) \quad = \quad & \left(4\beta^2K^2 + 4s\beta\alpha^2 - 2\alpha s^2 K + 2s^2\alpha^2 - 4\alpha s\beta K - 4x_{st}\beta\alpha^2\right. \\ & \left. + 2\alpha x_{st}^2 K + 2x_{st}^2\alpha^2 - 4\alpha x_{st}\beta K\right)^{\frac{1}{2}}. \end{aligned}$$

Eine gleichmäßige numerische Abtastung des zulässigen Parameterbereiches ergab, daß

$$T(P_i, P_a) \geq \Delta t$$

für alle unteren Bahnen, wenn zusätzlich $\Delta t \in \left(\frac{\beta(K-\alpha)}{\alpha(K-2\alpha)}, \Delta t_{3\beta}\right]$ erfüllt ist. Damit ist alles gezeigt. □

Nun wollen wir die beiden anderen Fälle betrachten. Die Rechnungen sind analog zu denen der Lemmata 6.2.5 und 6.2.6. Aus diesem Grund formulieren wir hier die beiden Aussagen und geben nur die Unterschiede in den Beweisen an.

Lemma 6.2.7 *Es sei $x_{st} \leq -3\beta$. Die Bezeichnungen seien entsprechend der ersten Darstellung der Abbildung 28 gewählt. Dann gilt*

$$T(P_i, P_a) \geq \Delta t$$

für alle oberen Parabeln mit Scheitelpunkt in I_2.

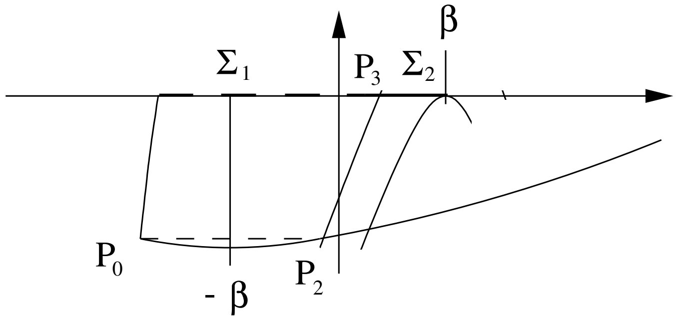


Abbildung 30: Geometrische Motivation für dritten Fall

Beweis: Zur Veranschaulichung diene die Abbildung 30. Es sei nun $x_{st} \leq -3\beta$. Der geometrische Nachweis, daß

$$T(P_i, P_a) \geq \Delta t$$

erfüllt ist, falls die linke Nullstelle der oberen Parabel Element von Σ_1 ist, folgt aus gleichen Überlegungen wie im Beweis von Lemma 6.2.5. Die beiden Bedingungen, damit das Intervall Σ_2 nicht leer ist, lauten

$$x_0 \leq -\beta \quad \text{und} \quad x_3 \leq \beta.$$

Dies ist äquivalent zu

$$\alpha\left(\frac{V}{6\beta} + \frac{2(x_{st}+4\beta)(x_{st}+\beta)}{3\beta V} - \frac{x_{st}+4\beta}{3\beta}\right) \geq K$$

mit

$$\begin{aligned} V &= \Big(-6\beta x_{st}^2 + 12\beta^2 x_{st} + 10\beta^3 - 8x_{st}^3 \\ &\quad +6\sqrt{3\beta}\sqrt{-\left(148\beta^2 x_{st}^3 + 53\beta x_{st}^4 + 8x_{st}^5 + 206\beta^3 x_{st}^2 + 140\beta^4 x_{st} + 37\beta^5\right)}\Big)^{\frac{1}{3}}. \end{aligned}$$

und

$$\begin{aligned} x_3(x_{st}) &= \frac{1}{\alpha(\alpha+2K)}\Big(-2K^2\beta - 3Kx_{st}\alpha + W - 8\alpha K\beta - \alpha^2 x_{st} - 3\alpha^2\beta \\ &\quad + \big(2K\beta^2\alpha^3 - 14\alpha K^3\beta^2 + 6K\alpha^3 x_{st}^2 + 6K^2 x_{st}^2\alpha^2 - 8K^2\beta^2\alpha^2 + 6\alpha^2 x_{st}\beta K^2 \\ &\quad +\alpha^4 x_{st}^2 + 2\alpha^4 x_{st}\beta + 10K\alpha^3 x_{st}\beta - 4K^4\beta^2 - 4K^3 x_{st}\alpha\beta + \alpha^4\beta^2 + 2W\alpha^2\beta \\ &\quad +2Kx_{st}\alpha W + 8\alpha K\beta W + 4K^2\beta W\big)^{\frac{1}{2}}\Big) \leq \beta \end{aligned}$$

mit

$$\begin{aligned} W &= \big(K^2 x_{st}^2\alpha^2 + 2\alpha^2 x_{st}\beta K^2 - 4K^2\beta^2\alpha^2 - 2\alpha K^3\beta^2 - 2K\beta^2\alpha^3 + \alpha^4 x_{st}^2 \\ &\quad +2\alpha^4 x_{st}\beta + 2K\alpha^3 x_{st}^2 + 4K\alpha^3 x_{st}\beta\big)^{\frac{1}{2}}. \end{aligned}$$

Der Nachweis, daß

$$T(P_i, P_a) \geq \Delta t$$

erfüllt ist, falls die linke Nullstelle der oberen Parabel Element von Σ_2 ist, erfolgt wiederum durch numerische Abtastung. $\square$

Nun wollen wir den zweiten Fall betrachten, bei dem nun $x_{st} \in \left[-3\beta, -2\beta + \frac{\beta K}{\alpha}\right)$ gilt.

Lemma 6.2.8 *Es sei* $x_{st} \in \left[-3\beta, -2\beta + \dfrac{\beta K}{\alpha}\right)$. *Weiterhin sei* $\Delta t \geq \dfrac{\beta(K+\alpha)}{\alpha(K+2\alpha)}$. *Die Bezeichnungen seien entsprechend der zweiten Darstellung der Abbildung* 28 *gewählt. Dann gilt*

$$T(P_i, P_a) \geq \Delta t$$

für alle oberen Parabeln mit Scheitelpunkt in I_2.

Beweis: Der Beweis wird auf die gleiche Weise geführt, wie der des Lemmas 6.2.6. Bezeichne Π die obere Parabel, deren linke Nullstelle mit der rechten der unteren Bahn übereinstimmt. Dann gilt für Π

$$T(P_i, P_a) = g(\Delta t) = \frac{\Delta t(\Delta t\alpha^2 - 2\beta K + 2\Delta t\alpha K)}{(\Delta t\alpha - \beta)(K-\alpha)}).$$

Die Funktion $g(\Delta t)$ besitzt folgende Eigenschaften. Es ist $g(\Delta t) = \Delta t$ für

$$\Delta t \in \left\{0, \frac{\beta(K+\alpha)}{\alpha(K+2\alpha)}\right\} \subset [0, +\infty).$$

Die Bedingung $x_{st} = -3\beta$ ist gleichbedeutend mit

$$\Delta t_{-3\beta} = \frac{(-\alpha + 2K + \sqrt{3\alpha^2 - 2K\alpha})\beta}{(2K+\alpha)\alpha}.$$

Es gilt

$$g(\Delta t_{-3\beta}) = \frac{2\beta}{\alpha} > \Delta t.$$

Hieraus und aus der Tatsache, daß g streng konvex im Intervall $(0, \Delta t_{-3\beta}]$ ist, folgt

$$\frac{\beta(K+\alpha)}{\alpha(K+2\alpha)} \leq \Delta t_{-3\beta}.$$

Damit ist die Behauptung für die unterste der oberen Parabeln Π bewiesen. Der Nachweis, daß

$$T(P_i, P_a) \geq \Delta t.$$

für alle anderen oberen Parabeln unter der Bedingung $\Delta t \in \left[\frac{\beta(K+\alpha)}{\alpha(K+2\alpha)}, \Delta t_{3\beta}\right]$ erfüllt ist, erfolgt wiederum durch numerische Abtastung. $\square$

Nun wollen wir die letzten vier Resultate zusammenfassen und den Punkt $P_0 = (x_0, y_0)$ im ersten Quadranten angeben, durch den die beiden Parabeln eindeutig bestimmt sind, die das Gebiet G beranden.

Satz 6.2.9 *Es sei* $|K| < \alpha$ *und* $\Delta t \in \left[\max\left\{\frac{\beta(K-\alpha)}{\alpha(K-2\alpha)}, \frac{\beta(K+\alpha)}{\alpha(K+2\alpha)}\right\}, \frac{\beta}{\alpha}\right)$. *Dann gilt für* P_0

$$x_0(\Delta t) = \frac{-2K\beta^2 + 2K\beta\Delta t\alpha - 2\beta\Delta t\alpha^2 + \Delta t^2\alpha^3}{2\alpha(\Delta t\alpha - \beta)}$$

und

$$y_0(\Delta t) = -\frac{\Delta t(\Delta t^2\alpha^2 K - 3K\beta\Delta t\alpha + 2K\beta^2 - \beta\Delta t\alpha^2 + 2\beta^2\alpha)}{2\alpha(\Delta t\alpha - \beta)}.$$

Beweis: Die Aussage ist unmittelbare Konsequenz der Lemmata 6.2.5 bis 6.2.8. Die Tatsache, daß für das linke Intervallende des zulässigen Bereiches für Δt

$$\max\left\{\frac{\beta(K-\alpha)}{\alpha(K-2\alpha)}, \frac{\beta(K+\alpha)}{\alpha(K+2\alpha)}\right\}$$

gilt, folgt daraus, daß die Aussagen von Lemma 6.2.6 sowie von Lemma 6.2.8 gleichzeitig erfüllt sein müssen. $\square$

Bemerkung 6.2.10 Welche qualitativen Auswirkungen hat Satz 6.2.9? Wie verhält sich das Beschränktheitsgebiet G, wenn man am konstanten Eingang K „wackelt“? Die erste Darstellung in Abbildung 31 zeigt die Beschränktheitsgebiete für die Wahl $\alpha = 2, \beta = 1, \Delta t = \frac{1}{3}$ und $K \in \{0.9, 1, 1.1\}$. Offensichtlich sind sie nicht echt ineinander enthalten. Deshalb ist die Be-

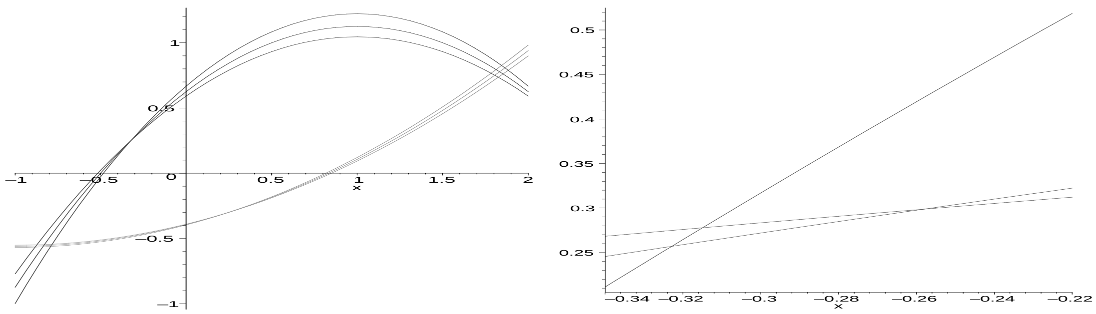


Abbildung 31: Beschränktheitsgebiete für verschiedene K

randung des Beschränktheitsgebietes im nichtkonstanten Fall nicht einfach durch den Rand der Vereinigung aller Beschränktheitsgebiete beschreibbar. Bisher haben wir keine Argumentation zur Darstellung von G in diesem Fall.
Weiterhin suggeriert die erste Darstellung von Abbildung 31, daß sich alle oberen bzw. unteren Parabeln, die sich mit $P_0 = (x_0, y_0)$ nach Satz 6.2.9 berechnen lassen, in je zwei Punkten schneiden. Vergrößern wir aber z.B. den linken „Schnittpunkt“ der oberen Bahnen für $K \in \left\{-\frac{3}{2}, 0, \frac{3}{2}\right\}$, wie im zweiten Teil der Abbildung 31 geschehen, so erkennt man, daß dies nicht der Fall ist.

Fazit 6.2.11 *Der Satz* 6.2.9 *gibt also darüber Auskunft, wie sich das Beschränktheitsgebiet G in Abhängigkeit von Δt und den anderen Parametern darstellt. Ist die Lösung von* (6.6) *einmal in G, so verläßt sie es unter den gegebenen Voraussetzungen nicht mehr.*
Ein Beschränktheitsgebiet ist für uns besonders interessant, da wir dies zur Grundlage einer ganzen Reihe von Betrachtungen gemacht haben. Häufig setzten wir $|U_1(t)| \leq \beta U_{ref}$ voraus. Dies ist mit den obigen Überlegungen gleichbedeutend zu $|x(t)| \leq \beta$. Für welches G dies immer erfüllt ist, können wir nun mit Hilfe des Satzes 6.2.9 *berechnen.*
Offensichtlich ist das in Satz 6.2.9 *bestimmte G das kleinste in der Menge aller Beschränktheitsgebiete in folgendem Sinn. Es sei P_1 ein Punkt, der auf dem Graphen der Funktion $y_0(x_0)$ wie in* (6.18) *liegt, derart, daß das Gebiet G_1, welches durch die beiden Parabeln, die durch P_1 gehen, berandet ist, G echt enthält. Dann ist auch G_1 ein Beschränktheitsgebiet der Lösung von* (6.6), *welches nie mehr verlassen werden kann. Die Argumentation ist die gleiche, die zum Satz* 6.2.9 *führte. Daß dagegen nicht jede Obermenge von G auch Beschränktheitsmenge ist, folgt aus der Charakteristik des Phasenporträts.*
Nun stellt sich aber folgende Frage. Von welchen Startpunkten $Q \in \mathbb{R}^2$ erreicht die Lösung nach

endlicher Zeit dieses G? Oder umgekehrt, gibt es Anfangslagen $Q \in \mathbb{R}^2$, die nicht zu G laufen? Diese Fragen sind vom mathematischen Standpunkt aus sehr interessant und noch offen. Vom physikalischen Gesichtspunkt spielt dies eine untergeordnete Rolle. Beim Beginn eines Meßvorgangs wird der Modulator so initialisiert, daß die Anfangslage ein fest definierter Punkt in $\mathbb{R}^2$ ist. Dieser sollte bereits in G enthalten sein.

6.3 Aussagen zur Periodizität

In diesem Kapitel widmen wir uns der Frage, ob periodische Lösungen der AWA (6.6) auftreten können, und wie sie dann charakterisiert werden. Auch hier untersuchen wir den einfachsten Fall $U_0(t) = const.$ Zunächst erklären wir, was wir unter einer periodischen Lösung von (6.6) verstehen wollen.

Definition 6.3.1 *Die Indizierung (x_m, y_m) bezeichne den Ort der Lösung von (6.6) zum Zeitpunkt $t_0 + m\Delta t$ mit $m \in \mathbb{N}_0$. Dagegen beschreibe (x_m^i, y_m^i) den Ort des i-ten Wechsels der Lösung von einer oberen auf eine untere Bahn bzw. umgekehrt, wobei (x_m, y_m) auch ein solcher sei, d. h., bei dieser Indizierung findet nach genau m Takten ein solcher Wechsel statt. Die Bedeutung der Bezeichnung geht aus dem Kontext hervor, so daß es zu keinen Verwechslungen kommen sollte. Es sei $n \in \mathbb{N}$. Dann heißt eine Lösung von (6.6) n-periodisch, falls es ein $m \in \mathbb{N}$ gibt, so daß*

$$(x_m, y_m) = (x_m^{2n}, y_m^{2n}) \quad \text{und} \quad (x_m, y_m) \notin \{(x_m^i, y_m^i) \in \mathbb{R}^2 : i \in \{1, \ldots, 2n-1\}\}$$

erfüllt ist.

Bemerkung 6.3.2 Zwei Aspekte aus Definition 6.1.2 sind interessant. Zum einen kann sich die zu untersuchende Periodizität erst nach einer gewissen Zeit einstellen. Zum anderen muß demnach aus $(x_m, y_m) = (x_m^{2n}, y_m^{2n})$ dann $(x_m, y_m) = (x_m^{2kn}, y_m^{2kn})$ für alle $k \in \mathbb{N}$ folgen. In der Abbildung 32 sind eine ein- und eine zweiperiodische Lösung dargestellt. Wir sind nun daran interessiert, welche Bedingungen notwendig und hinreichend für das Auftreten periodischer Lösungen sind (vgl. [22] für Konverter erster Ordnung).

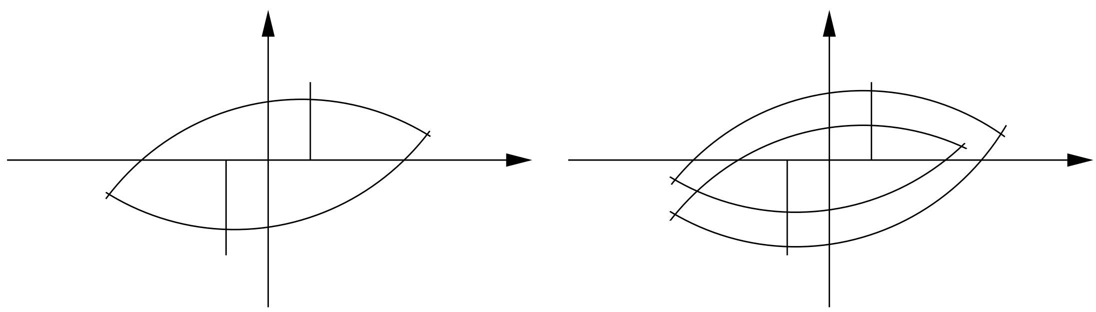


Abbildung 32: Ein- und zweiperiodische Lösung

Lemma 6.3.3 *Es sei $|K| < \alpha$. Besitzt das Anfangswertproblem (6.6) eine n-periodische Lösung, dann gilt*

$$\frac{\alpha - K}{\alpha + K} \in \mathbb{Q}.$$

Beweis: Nach Definition 6.3.1 existiert ein $m \in \mathbb{N}_0$ mit

$$x_m = x_m^{2n}. \tag{6.32}$$

O. B. d. A. sei $y_m \geq 0$. Dann läuft die Lösung die ersten a_1 Takte Δt entlang einer oberen Bahn bis $(x_{m+a_1}, y_{m+a_1}) = (x_m^1, y_m^1)$ mit $y_m^1 < 0$, danach b_1 Takte entlang einer unteren Bahn bis $(x_{m+a_1+b_1}, y_{m+a_1+b_1}) = (x_m^2, y_m^2)$ mit $y_m^2 \geq 0$. Diese Prozeß geht bis $(x_m^{2n}, y_m^{2n}) = (x_m, y_m)$. Somit ist (6.32) äquivalent dazu, daß der Weg in $\sum\limits_{i=1}^{n} a_i$ Takten Δt in x-Richtung entlang oberer Parabeln dem Betrag nach gleich groß wie der in $\sum\limits_{i=1}^{n} b_i$ Takten entlang der entsprechenden unteren Parabeln. Wegen (6.7) und (6.9) gilt dann

$$-\sum_{i=1}^{n} a_i \Delta t (K - \alpha) = \sum_{i=1}^{n} b_i \Delta t (K + \alpha).$$

Dies ist äquivalent zu

$$\frac{K-\alpha}{K+\alpha} = -\frac{\sum\limits_{i=1}^{n} b_i}{\sum\limits_{i=1}^{n} a_i}.$$

Hieraus und wegen $a_i, b_i \in \mathbb{N}$ folgt dann die Behauptung. $\square$

Wir wollen nun notwendige und hinreichende Bedingungen für die Existenz ein- bzw. zweiperiodischer Lösungen angeben.

Lemma 6.3.4 *Es sei $|K| < \alpha$. Dann gelten folgende Aussagen:*
a) Es sei $a_1, b_1 \in \mathbb{N}$. Folgende Aussagen sind äquivalent:

(i) Die Lösung von (6.6) *ist einperiodisch mit*

$$(x_k, y_k) = (x_{k+a_1+b_1}, y_{k+a_1+b_1})$$

für ein $k \in \mathbb{N}_0$.

(ii) Es existiert ein $m \in \mathbb{N}_0$, so daß gilt

$$y_m \geq 0,$$
$$\forall n \in \mathbb{N}_0 : y_{m+(a_1+b_1)n+a_1-1} \geq 0 \quad \textit{und} \quad y_{m+(a_1+b_1)n+a_1} < 0, \tag{6.33}$$
$$\forall n \in \mathbb{N} : y_{m+(a_1+b_1)n-1} < 0 \quad \textit{und} \quad y_{m+(a_1+b_1)n} \geq 0, \tag{6.34}$$
$$-a_1(K-\alpha) = b_1(K+\alpha). \tag{6.35}$$

b) Es sei $a_1, a_2, b_1, b_2 \in \mathbb{N}$. Folgende Aussagen sind äquivalent:

(iii) Die Lösung von (6.6) *ist zweiperiodisch mit*

$$(x_k, y_k) = (x_{k+a_1+b_1+a_2+b_2}, y_{k+a_1+b_1+a_2+b_2})$$

für ein $k \in \mathbb{N}_0$.

(iv) Es existiert ein $m \in \mathbb{N}_0$, so daß gilt

$$\begin{aligned}
&y_m \geq 0, \\
&\forall n \in \mathbb{N}_0 : y_{m+(a_1+b_1+a_2+b_2)n+a_1-1} \geq 0 \quad \textit{und} \quad y_{m+(a_1+b_1+a_2+b_2)n+a_1} < 0, \\
&\forall n \in \mathbb{N}_0 : y_{m+(a_1+b_1+a_2+b_2)n+a_1+b_1-1} < 0 \quad \textit{und} \quad y_{m+(a_1+b_1+a_2+b_2)n+a_1+b_1} \geq 0, \\
&\forall n \in \mathbb{N}_0 : y_{m+(a_1+b_1+a_2+b_2)n+a_1+b_1+a_2-1} \geq 0 \quad \textit{und} \quad y_{m+(a_1+b_1+a_2+b_2)n+a_1+b_1+a_2} < 0, \\
&\forall n \in \mathbb{N} : y_{m+(a_1+b_1+a_2+b_2)n-1} < 0 \quad \textit{und} \quad y_{m+(a_1+b_1+a_2+b_2)n} \geq 0, \\
&-(a_1+a_2)(K-\alpha) = (b_1+b_2)(K+\alpha).
\end{aligned}$$

Beweis: a) $(i) \Rightarrow (ii)$ Der Beweis ist elementar. Er folgt unmittelbar aus der Definition 6.3.1, Lemma 6.3.3 und der Darstellung des Phasenporträts nach (6.7) und (6.9).
$(ii) \Rightarrow (i)$ Aus (6.33) und (6.34) folgt, daß die Lösung von (6.6) nach genau a_1 Takten - zum Zeitpunkt t_{m+a_1} - von einer oberen auf eine untere Parabel springt und nach genau b_1 Takten wieder zurück. Dieser Prozeß setzt sich unendlich fort. Hieraus, aus (6.35) und (6.7) sowie (6.9) ergibt sich

$$\begin{aligned}
x_{m+a_1+b_1} &= (K+\alpha)b_1\Delta t + x_{m+a_1} \\
&= \Delta t[(K-\alpha)a_1 + (K+\alpha)b_1] + x_m \\
&= x_m.
\end{aligned}$$

Induktiv läßt sich diese Aussage fortsetzen zu

$$x_{m+(a_1+b_1)n} = x_m \quad \forall n \in \mathbb{N}. \tag{6.36}$$

Wegen (6.36) bleibt zu zeigen, daß $y_{m+(a_1+b_1)n} = y_m$ erfüllt ist. Nun sei $y_{m+a_1+b_1} > y_m$. Wegen (6.7) und (6.9) gilt

$$\begin{aligned}
x_{m+a_1} &= (K-\alpha)a_1\Delta t + x_m, \\
y_{m+a_1} &= (K-\alpha)\frac{a_1^2\Delta t^2}{2} + (x_m-\beta)a_1\Delta t + y_m, \\
y_{m+a_1+b_1} &= (K+\alpha)\frac{b_1^2\Delta t^2}{2} + (x_{m+a_1}-\beta)b_1\Delta t + y_{m+a_1} \\
&= [(K+\alpha)b_1^2 + (K-\alpha)a_1^2]\frac{\Delta t^2}{2} + [(K-\alpha)a_1b_1\Delta t^2 + x_m - \beta]b_1\Delta t \\
&\quad +(x_m-\beta)a_1\Delta t + y_m. \qquad (6.37)
\end{aligned}$$

Damit ist unsere Annahme äquivalent zu

$$[(K+\alpha)b_1^2 + (K-\alpha)a_1^2]\frac{\Delta t^2}{2} + (K-\alpha)a_1b_1\Delta t^2 + (x_m-\beta)(a_1+b_1)\Delta t =: \varepsilon > 0. \tag{6.38}$$

Analog zu (6.37) gilt wegen (6.36) und (6.38) dann

$$\begin{aligned}
y_{m+(a_1+b_1)(n+1)} &= [(K+\alpha)b_1^2 + (K-\alpha)a_1^2]\frac{\Delta t^2}{2} + (K-\alpha)a_1\Delta t \\
&\quad +(x_{m+(a_1+b_1)n} - \beta)(a_1+b_1)\Delta t + y_{m+(a_1+b_1)n} \\
&= [(K+\alpha)b_1^2 + (K-\alpha)a_1^2]\frac{\Delta t^2}{2} + (K-\alpha)a_1b_1\Delta t^2 \\
&\quad +(x_m-\beta)(a_1+b_1)\Delta t + y_{m+(a_1+b_1)n} \\
&= \varepsilon + y_{m+(a_1+b_1)n} \\
&> y_{m+(a_1+b_1)n}.
\end{aligned}$$

Somit ist $(y_{m+(a_1+b_1)n})_{n\in\mathbb{N}_0}$ eine unbeschränkte, streng monoton wachsende Folge. Andererseits folgt aus (6.34) und (6.9)

$$\begin{aligned} y_{m+(a_1+b_1)n} - y_{m+(a_1+b_1)n-1} &= (K+\alpha)\frac{\Delta t^2}{2} + (x_m - (K+\alpha)\Delta t + \beta)\Delta t \\ &= -(K+\alpha)\frac{\Delta t^2}{2} + (x_m + \beta)\Delta t \in (0, +\infty). \end{aligned}$$

Also kann nicht gleichzeitig $y_{m+(a_1+b_1)n-1}$ negativ sein und $y_{m+(a_1+b_1)n}$ beliebig groß. Die Annahme $y_{m+a_1+b_1} > y_m$ ist falsch.
Analog führt man $y_{m+a_1+b_1} < y_m$ zu einem Widerspruch. Es gilt dann

$$y_{m+(a_1+b_1)n} = y_m \quad \forall n \in \mathbb{N}. \tag{6.39}$$

Wegen (6.36) und (6.39) ist

$$(x_m, y_m) = (x_{m+(a_1+b_1)n}, y_{m+(a_1+b_1)n}) \quad \forall n \in \mathbb{N}.$$

was nach Definition 6.3.1 äquivalent zu (i) ist. Damit ist die Aussage in a) bewiesen.
b) Der Nachweis erfolgt durch die gleichen Überlegungen wie in a). □

Bemerkung 6.3.5 Bis jetzt haben wir gesehen, daß es beschränkte nichtperiodische Lösungen gibt, wenn das Eingangssignal konstant und einer Beschränkung unterliegt. Ein- bzw. zweiperiodische Lösungen existieren, wenn man die Bedingungen des Lemmas 6.3.4 berücksichtigt. Diese sind jedoch recht unhandlich. Wir sind bisher nicht in der Lage folgende zwei Fragen zu beantworten.

(i) Ab wann ist eine Lösung von (6.6) periodisch, und welcher Art ist diese Periodizität?

(ii) Wo liegen beispielsweise für alle einperiodischen Lösungen die Wechselpunkte von unterer zu oberer Bahn? Bilden diese Punkte eine besondere Struktur im $\mathbb{R}^2$?

Bezüglich der Frage (ii) läßt sich diese auf alle n-periodischen Lösungen erweitern. Darüber hinaus wollen wir diese Frage schon zu einem kleinen Teil beantworten, jedoch nicht allgemein.

Lemma 6.3.6 *Es sei $P_0 = (x_0, y_0)$ der Wechselpunkt von unterer zu oberer Bahn einer einperiodischen Lösung, d. h., es ist $(x_0, y_0) = (x_0^{2n}, y_0^{2n})$ für alle $n \in \mathbb{N}$. Dann gelten folgende Aussagen.*
a) Es existieren $\varepsilon_1, \varepsilon_2 \geq 0$, wobei mindestens einer der beiden Werte größer als Null ist, derart, daß alle Punkte

$$Q \in \{(x,y) \in \mathbb{R}^2 : x = x_0, y \in [y_0 - \varepsilon_2, y_0 + \varepsilon_1]\}$$

auch solche Wechselpunkte einperiodischer Lösungen sind.
b) Es sei $\varepsilon > 0$. Dann sind

$$Q \in \{(x,y) \in \mathbb{R}^2 : x \in (x_0 - \varepsilon, x_0 + \varepsilon) \setminus \{x_0\}, y = y_0\}$$

keine solche Wechselpunkte einperiodischer Lösungen.

Beweis: a) Es sei $Q = (x_0, y_0 + \mu)$ mit $\mu \in (-\varepsilon_2, \varepsilon_1)$. Somit ist Q eine vertikale Verschiebung von P_0. Die Lösung von (6.6) ist wegen der Charakteristik des Phasenporträts, aus zueinander vertikal verschobenen Parabeln zu bestehen, auch vertikal verschoben. Die Differenz μ in der y-Koordinate darf jedoch nicht zu groß sein, damit die Umschaltbedingungen zwischen den beiden Parabeltypen noch gewährleistet sind, wie sie in Teil a) des Lemmas 6.3.4 beschrieben sind.

Dies geht in die Größen von ε_i ein, wobei beide nicht gleich sein müssen (vgl. (6.33) und (6.34)). Daraus folgt die Behauptung in a).
b) Es sei $Q = (x_0 + \mu, y_0)$ mit $\mu \in (-\varepsilon, \varepsilon) \setminus \{0\}$. Weiterhin gelte die Notation wie in Lemma 6.3.4. Dann folgt aus analogen Rechnungen wie im Beweis des Lemmas 6.3.4

$$\begin{aligned} x_{0+a_1+b_1} &= x_0 + \mu, \\ y_{0+a_1+b_1} &= [(K+\alpha)b_1^2 + (K-\alpha)a_1^2]\frac{\Delta t^2}{2} + (K-\alpha)a_1 b_1 \Delta t^2 + (x_0 - \beta)(a_1 + b_1)\Delta t \\ &\quad + \mu \Delta t (a_1 + b_1) + y_0, \end{aligned}$$

wobei hier $(x_{0+a_1+b_1}, y_{0+a_1+b_1})$ die Lage der Lösung nach $a_1 + b_1$ Takten beginnend in Q beschreibe. Da nach Voraussetzung P_0 Wechselpunkt einer einperiodischen Lösung ist, gilt (vgl. den Beweis von Lemma 6.3.4)

$$[(K+\alpha)b_1^2 + (K-\alpha)a_1^2]\frac{\Delta t^2}{2} + (K-\alpha)a_1 b_1 \Delta t^2 + (x_0 - \beta)(a_1 + b_1)\Delta t = 0.$$

Hiermit ergibt sich wegen $\mu \neq 0$ sofort

$$y_{0+a_1+b_1} = \mu \Delta t (a_1 + b_1) + y_0 \neq y_0.$$

Also ist die Lösung von (6.6) beginnend in Q nach $a_1 + b_1$ Takten von Q verschieden. Damit ist alles gezeigt. $\square$

Bemerkung 6.3.7 Allgemein läßt sich wegen Lemma 6.3.6 feststellen, daß in jeder Umgebung von P_0 ein- und nichteinperiodische Lösungen existieren. Ob dagegen die Lösung der in x-Richtung gestörten AWA später periodisch wird oder nicht, ist zur Zeit noch nicht entscheidbar.

7 Stabilität und Robustheit

In den beiden folgenden Abschnitten wollen wir einige Aussagen bezüglich der Stabilität und der Robustheit der Konverter treffen. Dabei unterscheiden wir die Modellbildung, die durch eine Differentialgleichung beschrieben wird, und die der Differenzengleichung. Die Begriffe und Definitionen haben wir im kontinuierlichen Fall aus [67], [57] und [65] entnommen. Im diskreten Fall stützen wir uns auf die Darstellung in [69], [40] sowie [12].

7.1 Stabilität und Robustheit der Differentialgleichungen

Die Stabilitätsanalyse linearer Regelkreise geht von der Tatsache aus, daß jeder Regelvorgang bezüglich einer Anfangsstörung asymptotisch stabil im Ganzen ist, wenn alle Nullstellen des charakteristischen Polynoms negative Realteile haben. Die Systemparameter, die diese Bedingung erfüllen, bilden im Parameterraum den Stabilitätsbereich. Ist der Parameterraum mit einer geeigneten Norm versehen, so betrachtet man nun alle Vektoren, die vom ungestörten System zum Rand des Stabilitätsbereiches führen. Die minimale Länge aller dieser Vektoren nennt man Stabiltätsradius. Somit ist der Stabilitätsradius ein Maß dafür, wie stark das homogene System gestört werden kann, ohne daß das System die Eigenschaft, asymptotisch stabil zu sein, verliert, falls dies bereits für das ungestörte System galt.
Ausgangspunkt unserer Überlegungen sind die Differentialgleichungen, die wir im Kapitel 4 für die Konverter erster bis dritter Ordnung hergeleitet haben

$$\varphi(t) = \frac{U_0(t)}{\alpha U_{ref}}, \tag{7.1}$$

$$\dot{\varphi}(t) + \frac{\alpha}{\beta T_1}\varphi(t) = \frac{U_0(t)}{\beta T_1 U_{ref}}, \tag{7.2}$$

$$\ddot{\varphi}(t) + \frac{\beta}{\gamma T_2}\dot{\varphi}(t) + \frac{\alpha}{\gamma T_1 T_2}\varphi(t) = \frac{U_0(t)}{\gamma T_1 T_2 U_{ref}}. \tag{7.3}$$

Zunächst wollen wir die entsprechenden Begriffsbildungen, die wir oben verbal benannt haben, bereitstellen.

Definition 7.1.1 *a) Es sei $n \in \mathbb{N}$, $A \in \mathbb{R}^{n\times n}$ eine reelle Matrix und x eine n-dimensionale reelle Funktion. Dann heißt die Differentialgleichung*

$$x'(t) = Ax(t)$$

asymptotisch stabil, falls $\mathrm{Re}\lambda < 0$ für alle Eigenwerte von A gilt, bzw. stabil, falls $\mathrm{Re}\lambda \leq 0$ für alle Eigenwerte von A gilt, wobei λ mit $\mathrm{Re}\lambda = 0$ halbeinfach ist. Ist die Differentialgleichung weder asymptotisch stabil noch stabil, so nennt man sie instabil.
b) Es sei $\mathbb{C}_g$ und $\mathbb{C}_b$ eine Zerlegung von $\mathbb{C}$ in zwei einfach zusammenhängende Gebiete, wobei $\mathbb{C}_g$ offen sei. Dabei bezeichne $\mathbb{C}_g$ den Stabilitätsbereich. Weiterhin sei $n \in \mathbb{N}$ und $\mathbb{R}^n$ versehen mit einer Norm $||.||$. Es bezeichne $\mathcal{P}_n$ die Menge aller reellen Polynome n-ten Grades. Dann ist eine Abbildung von $\mathbb{R} \times \mathbb{R}^n$ nach $\mathcal{P}_n$ durch

$$p(s,k) = s^n + (s^{n-1}, s^{n-2}, \dots, 1)(Fk + g)$$

definiert, wobei $F \in \mathbb{R}^{n\times n}$ und $g \in \mathbb{R}^n$ ist. Ein Polynom in $\mathcal{P}_n$ heißt stabil, falls alle seine Nullstellen in $\mathbb{C}_g$ liegen, im anderen Fall instabil. Durch

$$\varrho := \inf\{||k|| : k \in \mathbb{R}^n, p(s,k) \text{ ist instabil}\}$$

ist der Stabilitätsradius bezüglich der Norm $||.||$ definiert, wobei F, g und $\mathbb{C}_g$ gegeben sind. (Hierbei sei $\inf\emptyset := +\infty$.)

Bemerkung 7.1.2 Anhand dieser Begriffsbildungen wollen wir nun einige vorbereitende Aussagen treffen, um später die Konverter zweiter und dritter Ordnung zu charakterisieren. Dabei wird in der Literatur oft eine andere Definition für die asymptotische Stabilität bzw. Instabilität einer linearen Differentialgleichung angeführt. Teil a) der Definition 7.1.1 ist dort dann ein notwendiges und hinreichendes Kriterium. Wir haben dieses Kriterium als Definition aufgeführt, da es unserer Situation besser angepaßt ist.
Was aber hat Teil b) der Definition 7.1.1 mit den Differentialgleichungen zu tun? Hierfür betrachten wir die charakteristischen Polynome der homogenen Gleichungen von (7.2) bzw. (7.3) und die entsprechenden Polynome $p(s,k)$ in Teil b) mit $k = 0 \in \mathbb{R}^n$, $n = 1, 2$, wobei g der Vektor ist, welcher aus den Koeffizienten von φ bzw. $\dot{\varphi}$ und φ besteht.
Eine Diskussion für den Konverter erster Ordnung erübrigt sich, da es sich hierbei nicht um eine klassische Differentialgleichung handelt, auf die wir die Begriffe der Definition anwenden können.

Nun wollen wir je ein Kriterium für asymptotische Stabilität und den Stabilitätsradius angeben.

Lemma 7.1.3 *a) (Hurwitz-Kriterium) Es sei $n \in \mathbb{N}$. Eine notwendige und hinreichende Bedingung, daß die Nullstellen einer algebraischen Gleichung*

$$f(s) = a_0 s^n + a_1 s^{n-1} + \cdots + a_n = 0$$

mit reellen Koeffizienten und $a_0 > 0$ sämtlich negative Realteile haben, lautet: Die n Hurwitzschen Determinanten

$$D_\nu = \begin{vmatrix} a_1 & a_0 & 0 & 0 & \dots & 0 \\ a_3 & a_2 & a_1 & a_0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ a_{2\nu-1} & a_{2\nu-2} & a_{2\nu-3} & a_{2\nu-4} & \dots & a_\nu \end{vmatrix}, \quad \nu \in \{1, \dots, n\}$$

mit $a_i = 0$ für $i > n$ sind alle positiv.
b) Es sei $i \in [1, +\infty]$ und $n \in \mathbb{N}$. Weiterhin bezeichne $||.||_i^\star$ die Norm des zu $(\mathbb{R}^n, ||.||_i)$ dualen normierten Raumes. Dabei ist

$$\forall i \in [1, +\infty) : ||x||_i = \left(\sum_{j=1}^{n} |x_j|^i \right)^{\frac{1}{i}} \qquad \textit{und} \qquad ||x||_\infty = \max_{j \in \{1,\dots,n\}} |x_j|.$$

Dann gilt

$$\begin{aligned} \varrho_i &= \inf_{s \in \partial \mathbb{C}_g} \{ \inf\{ ||k||_i : k \in \mathbb{R}^n, u^\top k = 1, v^\top k = 0 \} \} \\ &= \begin{cases} +\infty, & \textit{falls} \quad rang(u \;\; v) \neq rang \begin{pmatrix} u & v \\ 1 & 0 \end{pmatrix}, \\ \inf\limits_{s \in \partial \mathbb{C}_g} \dfrac{1}{||u||_i^\star}, & \textit{falls} \quad rang(u \;\; v) = 1 = rang \begin{pmatrix} u & v \\ 1 & 0 \end{pmatrix}, \\ \inf\limits_{s \in \partial \mathbb{C}_g} \left\{ \sup\limits_{\alpha \in \mathbb{R}} \dfrac{1}{||u + \alpha v||_i^\star} \right\}, & \textit{falls} \quad rang(u \;\; v) = 2, \end{cases} \end{aligned}$$

wobei $u := \mathrm{Re} w$ und $v := \mathrm{Im} w$ für

$$w := \frac{F^\top (s^{n-1} s^{n-2} \dots 1)^\top}{-s^n - (s^{n-1} s^{n-2} \dots 1) g}$$

ist.

Beweis: a) Diese Aussage ist in [38] bewiesen.
b) Den Beweis findet man in [57]. Nur im Fall $rang(u\ \ v) = 2$ ist die Berechnung von ϱ_i nichttrivial. Für den ersten Fall haben wir entweder lineare Abhängigkeit von u und v oder beide sind Nullvektoren. Dann sind aber $u^\top k = 1$ und $v^\top k = 0$ nicht gleichzeitig erfüllbar. Der zweite Fall folgt aus dem dritten, da

$$rang(u\ \ v) = 1 = rang \begin{pmatrix} u & v \\ 1 & 0 \end{pmatrix}$$

äquivalent zu $u \neq 0$ und $v = 0$ ist. □

Nun wollen wir ϱ_i im nichttrivialen Fall $rang(u\ \ v) = 2$ berechnen.

Satz 7.1.4 *Es sei $n \in \mathbb{N}$ und $u, v \in \mathbb{R}^n$. Für $rang(u\ \ v) = 2$ gelten dann folgende Aussagen:*
a)

$$\begin{aligned} \inf\{||k||_\infty : k \in \mathbb{R}^n, u^\top k = 1, v^\top k = 0\} &= \sup_{h\in\mathbb{R}} \frac{1}{||u + hv||_1} \\ &= \max\left\{\frac{1}{||u+hv||_1} : h \in \left\{-\frac{u_i}{v_i} : i = 1, \ldots, n, v_i \neq 0\right\}\right\}, \end{aligned}$$

b)

$$\inf\{||k||_2 : k \in \mathbb{R}^n, u^\top k = 1, v^\top k = 0\} = \sup_{h\in\mathbb{R}} \frac{1}{||u+hv||_2} = \frac{||v||_2}{\sqrt{||u||_2^2||v||_2^2 - (u^\top v)^2}},$$

c)

$$\inf\{||k||_1 : k \in \mathbb{R}^n, u^\top k = 1, v^\top k = 0\} = \sup_{h\in\mathbb{R}} \frac{1}{||u+hv||_\infty} = \max\left\{\frac{1}{||u+hv||_\infty} : h \in \Lambda\right\}$$

mit

$$\begin{aligned} \Lambda &= \{h : u_i + hv_i = u_j + hv_j, 1 \leq i < j \leq n\} \\ &\quad \cup \{h : u_i + hv_i = -u_j - hv_j, 1 \leq i < j \leq n\} \\ &= \left\{-\frac{u_i - u_j}{v_i - v_j} : 1 \leq i < j \leq n, v_i - v_j \neq 0\right\} \\ &\quad \cup \left\{-\frac{u_i + u_j}{v_i + v_j} : 1 \leq i < j \leq n, v_i + v_j \neq 0\right\}. \end{aligned}$$

Beweis: Den Beweis findet man in [57]. □

In den beiden folgenden Aussagen wollen wir die Stabilitätsradien für (7.2) und (7.3) berechnen und diese exemplarisch für die Einheitsmatrizen $F = I_n$, $n = 1, 2$ angeben. Weiterhin geben wir Bedingungen an, unter denen (7.2) und (7.3) - also für das ungestörte System ($k = 0 \in \mathbb{R}^n$ für $n = 1, 2$) - asymptotisch stabil sind.

Folgerung 7.1.5 a) Ein Konverter zweiter Ordnung, der durch (7.2) beschrieben wird, ist genau dann asymptotisch stabil, wenn $\frac{\alpha}{\beta T_1} > 0$ gilt.
b) Der zugehörige Stabilitätsradius für $\mathbb{C}_g := \{z \in \mathbb{C} : \mathrm{Re} z < 0\}$ ist

$$\varrho_i = \left|\frac{\alpha}{F\beta T_1}\right| \quad \text{für} \quad i \in \{1, 2, +\infty\}.$$

c) Für $F = I_1 = 1$ gilt $\varrho_i = \frac{\alpha}{\beta T_1}$ für $i \in \{1, 2, +\infty\}$.

Beweis: a) Das charakteristische Polynom der homogenen Gleichung von (7.2) lautet

$$p(s,0) = s + \frac{\alpha}{\beta T_1}. \tag{7.4}$$

Dies ist nach dem Hurwitzschen Kriterium genau dann asymptotisch stabil, wenn $\frac{\alpha}{\beta T_1} > 0$ gilt.
b) Wir stören nun das charakteristische Polynom (7.4) wie in Teil b) der Definition 7.1.1 beschrieben

$$p(s,k) = s + Fk + \frac{\alpha}{\beta T_1}. \tag{7.5}$$

Beweis 1: Um den Stabilitätsradius zu bestimmen, muß $\sup\{||Fk||_i, k \in \mathbb{R}\}$ für $i \in \{1, 2, +\infty\}$ angegeben werden, so daß (7.5) noch stabil ist. Aus dem Hurwitzschen Kriterium und aus der Tatsache, daß die Normen $||.||_i$ im $\mathbb{R}$ gleich sind, folgt unmittelbar

$$\varrho_i = \left| \frac{\alpha}{F\beta T_1} \right|.$$

Wir wollen aber noch eine weitere Berechnungsmöglichkeit von ϱ_i angeben, wie sie in [57] beschrieben ist.
Beweis 2: Es sei nun $s = s_1 + is_2 \in \mathbb{C}$. Dann gilt für w aus Lemma 7.1.3 mit $g = \frac{\alpha}{\beta T_1}$

$$w = \frac{F}{-s - \frac{\alpha}{\beta T_1}},$$

und somit ist

$$\begin{aligned} u = \mathrm{Re} w &= -F \frac{s_1 + \frac{\alpha}{\beta T_1}}{\left(s_1 + \frac{\alpha}{\beta T_1}\right)^2 + s_2^2} \\ v = \mathrm{Im} w &= F \frac{s_2}{\left(s_1 + \frac{\alpha}{\beta T_1}\right)^2 + s_2^2}. \end{aligned}$$

Aus $uk = 1$ und $vk = 0$ folgt $s_1 \neq -\frac{\alpha}{\beta T_1}$ und $s_2 = 0$. Hieraus ergibt sich

$$u = -\frac{F}{s_1 + \frac{\alpha}{\beta T_1}} \quad \text{und} \quad v = 0.$$

Nach Lemma 7.1.3 gilt wegen

$$rang(u \;\; v) = 1 = rang \begin{pmatrix} u & v \\ 1 & 0 \end{pmatrix}$$

und wegen $||.||_1 = ||.||_2 = ||.||_\infty$ dann

$$\varrho_i = \left. \frac{1}{||u||_i^\star} \right|_{s \in \partial \mathbb{C}_g} = \left| \frac{\alpha}{F\beta T_1} \right|.$$

Hieraus folgt die Behauptung in b).
c) trivial. □

Nun wollen wir analoge Aussagen für das System, welches durch (7.3) beschrieben wird, treffen. Dabei kann der Stabilitätsradius nur in Abhängigkeit von $F \in \mathbb{R}^{2\times 2}$ dargestellt und erst im konkreten Fall auch berechnet werden.

Folgerung 7.1.6 a) Ein Konverter dritter Ordnung, der durch (7.3) beschrieben wird, ist genau dann asymptotisch stabil, wenn $\frac{\beta}{\gamma T_2} > 0$ und $\frac{\alpha}{\gamma T_1 T_2} > 0$ gilt.
b) Der Stabilitätsradius von (7.3) lautet für $\mathbb{C}_g = \{z \in \mathbb{C} : \mathrm{Re} z < 0\}$

$$\varrho_i = \inf_{s \in \partial \mathbb{C}_g} \left\{ \sup_{h \in \mathbb{R}} \frac{1}{||u - hv||_i^\star} \right\} \quad \text{für} \quad i \in \{1, 2, +\infty\}.$$

c) Für $F = I_2$ gilt $\varrho_i = \frac{\beta}{\gamma T_2}$ für $i \in \{1, 2, +\infty\}$.

Beweis: Im weiteren sei

$$a := \frac{\beta}{\gamma T_2} \quad \text{und} \quad b := \frac{\alpha}{\gamma T_1 T_2}.$$

a) Das charakteristische Polynom der homogenen Gleichung von (7.3) lautet

$$p(s, 0) = s^2 + as + b. \tag{7.6}$$

Dies ist nach dem Hurwitzschen Kriterium genau dann asymptotisch stabil, wenn $a > 0$ und $b > 0$ gilt.
b) Das gestörte charakteristische Polynom von (7.3) lautet

$$p(s, k) = s^2 + a_1(k)s + a_2(k)$$

mit

$$\begin{pmatrix} a_1(k) \\ a_2(k) \end{pmatrix} = \begin{pmatrix} F_1 & F_2 \\ F_3 & F_4 \end{pmatrix} \begin{pmatrix} k_1 \\ k_2 \end{pmatrix} + \begin{pmatrix} a \\ b \end{pmatrix} = Fk + g.$$

Hieraus folgt für w wie in Teil b) von Lemma 7.1.3 dann

$$w = -\frac{1}{s^2 + as + b} \begin{pmatrix} F_1 s + F_2 s \\ F_3 + F_4 \end{pmatrix}.$$

Für $s = s_1 + is_2$ sowie

$$A := s_1^2 - s_2 + as_1 + b \quad \text{und} \quad B := 2s_1 s_2 + as_2$$

gilt für $u, v \in \mathbb{R}^2$ wie in Teil b) von Lemma 7.1.3 schließlich

$$u = -\frac{1}{A^2 + B^2} \begin{pmatrix} (As_1 + Bs_2)(F_1 + F_2) \\ A(F_3 + F_4) \end{pmatrix}$$

und

$$v = -\frac{1}{A^2 + B^2} \begin{pmatrix} (As_2 - Bs_1)(F_1 + F_2) \\ -B(F_3 + F_4) \end{pmatrix}.$$

Diese sind für $s \in \partial \mathbb{C}_g$, also $s = is_2$, linear unabhängig. Es gilt also $rang(u \;\; v) = 2$. Nach Lemma 7.1.3 berechnet sich der Stabiltätsradius dann durch

$$\varrho_i = \inf_{s \in \partial \mathbb{C}_g} \left\{ \sup_{h \in \mathbb{R}} \frac{1}{||u - hv||_i^\star} \right\} \quad \text{für} \quad i \in \{1, 2, +\infty\}.$$

c) Aus den Forderungen $F = I_2$ und an $\mathbb{C}_g$ in der Voraussetzung folgt

$$u = -\frac{1}{(b - s_2^2)^2 + s_2^2 a^2} \begin{pmatrix} as_2^2 \\ b - s_2^2 \end{pmatrix} \quad \text{und} \quad v = -\frac{1}{(b - s_2^2)^2 + s_2^2 a^2} \begin{pmatrix} (b - s_2^2)s_2 \\ -as_2 \end{pmatrix}.$$

Sei nun $i = +\infty$. Nach Teil a) des Satzes 7.1.4 haben wir

$$\begin{aligned}
&\inf\{||k||_\infty : k \in \mathbb{R}^2, u^\top k = 1, v^\top k = 0\} \\
&\quad = \max\left\{\frac{1}{||u+hv||_1} : h \in \left\{-\frac{u_i}{v_i} : i = 1,2, v_1, v_2 \neq 0\right\}\right\} \\
&\quad = \frac{1}{\min\left\{\left|u_1 - \frac{u_1}{v_1}v_1\right| + \left|u_2 - \frac{u_1}{v_1}v_2\right|, \left|u_1 - \frac{u_2}{v_2}v_1\right| + \left|u_2 - \frac{u_2}{v_2}v_2\right|\right\}} \\
&\quad = \frac{1}{\min\left\{\left|u_2 - \frac{u_1}{v_1}v_2\right|, \left|u_1 - \frac{u_2}{v_2}v_1\right|\right\}}.
\end{aligned}$$

Wegen

$$\begin{aligned}
&\min\left\{\left|u_2 - \frac{u_1}{v_1}v_2\right|, \left|u_1 - \frac{u_2}{v_2}v_1\right|\right\} \\
&\quad = \frac{1}{(b-s_2^2)^2 + s_2^2a^2}\min\left\{\left|\frac{(b-s_2^2)^2 + s_2^2a^2}{b-s_2^2}\right|, \left|\frac{(b-s_2^2)^2 + s_2^2a^2}{a}\right|\right\} \\
&\quad = \min\left\{\frac{1}{|b-s_2^2|}, \frac{1}{|a|}\right\}
\end{aligned}$$

haben wir dann

$$\begin{aligned}
\varrho_\infty &= \inf_{s_2\in\mathbb{R}}\left\{\frac{1}{\min\left\{\left|u_1 - \frac{u_2}{v_2}v_1\right|, \left|u_2 - \frac{u_1}{v_1}v_2\right|\right\}}\right\} \\
&= \inf_{s_2\in\mathbb{R}}\left\{\max\{|b-s_2^2|, |a|\}\right\} \\
&= |a| = \frac{\beta}{\gamma T_2},
\end{aligned}$$

da alle Parameter positiv sind.
Sei nun $i = 2$. Dann folgt aus den Teilen b) des Lemmas 7.1.3 und des Satzes 7.1.4

$$\begin{aligned}
\varrho_2 &= \inf_{s_2\in\mathbb{R}}\left\{\frac{||v||_2}{\sqrt{||u||_2^2||v||_2^2 - (u^\top v)^2}}\right\} \\
&= \inf_{s_2\in\mathbb{R}}\left\{\frac{(s_2^2-b)^2 + s_2^2a^2}{\sqrt{s_2^4a^2 + s_2^4 - 2s_2^2b + b^2 - \frac{a^2s_2^2(-s_2^4+s_2^2b+s_2^2-b)^2}{s_2^6-2bs_2^4+b^2s_2^2+a^2s_2^2}}}\right\} \\
&= \inf_{s_2\in\mathbb{R}}\left\{\sqrt{(s_2^2-b)^2 + a^2}\right\} \\
&= |a| = \frac{\beta}{\gamma T_2}.
\end{aligned}$$

Es sei schließlich $i = 1$. Dann folgt aus Teil b) des Lemmas 7.1.3 und aus Teil c) des Satzes 7.1.4

$$\begin{aligned}
\varrho_1 &= \inf_{s_2\in\mathbb{R}}\left\{\max\left\{\frac{1}{||u+hv||_\infty} : h \in \left\{-\frac{u_1-u_2}{v_1-v_2}, -\frac{u_1+u_2}{v_1+v_2}\right\}\right\}\right\} \\
&= \inf_{s_2\in\mathbb{R}}\left\{\max\left\{\left|\frac{1}{u_1 - \frac{u_1-u_2}{v_1-v_2}v_1}\right|, \left|\frac{1}{u_2 - \frac{u_1-u_2}{v_1-v_2}v_2}\right|, \left|\frac{1}{u_1 - \frac{u_1-u_2}{v_1-v_2}v_1}\right|, \left|\frac{1}{u_2 - \frac{u_1+u_2}{v_1+v_2}v_2}\right|\right\}\right\}
\end{aligned}$$

$$
\begin{aligned}
&= \inf_{s_2\in\mathbb{R}} \left\{ |u_1v_2 - u_2v_1| \min\left\{ \frac{1}{|v_1-v_2|}, \frac{1}{|v_1+v_2|} \right\} \right\} \\
&= \inf_{s_2\in\mathbb{R}} \left\{ \max\{|-a|, |b-s_2^2|\} \right\} \\
&= |-a| = \frac{\beta}{\gamma T_2}.
\end{aligned}
$$

Damit ist alles bewiesen. □

Bemerkung 7.1.7 Betrachten wir anstelle der homogenen Differentialgleichung von (7.3) nun

$$\ddot{\varphi}(t) + (a + k_1)\dot{\varphi}(t) + (b + k_2)\varphi(t) = 0 \tag{7.7}$$

mit $a = \frac{\beta}{\gamma T_2}$ und $b = \frac{\alpha}{\gamma T_1 T_2}$, so lautet das zugehörige charakteristische Polynom

$$p(s,k) = s^2 + a_1(k)s + a_2(k)$$

mit

$$\begin{pmatrix} a_1(k) \\ a_2(k) \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} k_1 \\ k_2 \end{pmatrix} + \begin{pmatrix} a \\ b \end{pmatrix}.$$

Dies bedeutet gerade $F = I_2$. Da der allgemeine Fall $F \in \mathbb{R}^{2\times 2}$ gleichbedeutend mit der homogenen Differentialgleichung

$$\ddot{\varphi}(t) + (a + F_1k_1 + F_2k_2)\dot{\varphi}(t) + (b + F_3k_1 + F_4k_2)\varphi(t) = 0$$

ist, welche sich qualitativ nicht von (7.7) unterscheidet und sich auf diese zurückführen läßt, ist es sinnvoll, den Stabilitätsradius ϱ_i für $F = I_2$ in Teil c) der Folgerung 7.6 zu bestimmen.

7.2 Stabilität und Robustheit der Differenzengleichungen

In diesem Abschnitt untersuchen wir nun die beiden Anfangswertaufgaben

$$
\begin{aligned}
W_{k+1} &= W_k + \sigma\left(\varphi(k) - W_k\right), \\
W_0 &= a_0
\end{aligned}
\tag{7.8}
$$

und

$$
\begin{aligned}
H_{k+2} - 2H_{k+1} + H_k &= \sigma\left(\psi(k) + \mu H_{k+1} + \nu H_k\right), \\
H_0 &= a_0, \quad H_1 = a_1,
\end{aligned}
\tag{7.9}
$$

welche die Konverter erster und zweiter Ordnung charakterisieren. Es bezeichne $(W_k)_{k\in\mathbb{N}_0}$ bzw. $(H_k)_{k\in\mathbb{N}_0}$ die Lösungen von (7.8) bzw. von (7.9). Diese AWA können wir in dreierlei Hinsicht stören, in den Anfangswerten, in der Funktion U_0, welche in φ bzw. ψ integriert wird, sowie in den Systemparametern α, β, T_1 und T_2 in φ bzw. in ψ. Was haben dann noch die Lösungen $(\tilde{W}_k)_{k\in\mathbb{N}_0}$ bzw. $(\tilde{H}_k)_{k\in\mathbb{N}_0}$ der gestörten AWA mit $(W_k)_{k\in\mathbb{N}_0}$ bzw. mit $(H_k)_{k\in\mathbb{N}_0}$ zu tun? Zunächst untersuchen wir (7.8) unter zusätzlichen Annahmen wie in Satz 2.1.3.

Lemma 7.2.1 *Die Funktion* $\varphi : \mathbb{N}_0 \longrightarrow \mathbb{R}$ *erfülle die Bedingungen*

$$
\begin{aligned}
&\varphi(0) - a_0 \in [-1, 1), \\
&\varphi(k) - \varphi(k-1) \in [0, 1], \quad \forall k \in \mathbb{N}.
\end{aligned}
\tag{7.10}
$$

Dann besitzt das Anfangswertproblem (7.8) *genau eine Lösung der Form*

$$W_k = \begin{cases} a_0, & \text{für } k = 0 \\ \lfloor \varphi(k-1) - a_0 \rfloor + a_0 + 1, & \text{für } k \in \mathbb{N}, \end{cases} \tag{7.11}$$

Beweis: Der Beweis ergibt sich aus der Transformation $G_k := W_k - a_0$ und aus Satz 2.1.3, da die neue AWA gerade die Voraussetzungen des Satzes erfüllt. $\square$

Folgerung 7.2.2 Es bezeichne $\mathcal{L}$ die Menge aller Lösungen $(W_k)_{k\in\mathbb{N}_0}$ von (7.8) unter den Nebenbedingungen (7.10). Dann ist $\mathcal{L}$ versehen mit der Metrik

$$\varrho(W,\tilde{W}) := \sum_{k=0}^{\infty} \frac{|W_k - \tilde{W}_k|}{2^k}$$

ein metrischer, aber kein normierter Raum.

Beweis: $\mathcal{L}$ ist kein linearer Raum, da Vielfache bzw. die Summe von Lösungen wegen der Unstetigkeit der rechten Seite von (7.11) nicht in $\mathcal{L}$ liegen müssen. Somit ist $\mathcal{L}$ auch kein normierter Raum.
Aus der Darstellung der Lösung der AWA (7.8) in der Form (7.11) und den beiden Bedingungen in (7.10) folgt unmittelbar $W_k \in [a_0, k + a_0]$. Hieraus ergibt sich dann

$$\varrho(W,0) = \sum_{k=0}^{\infty} \frac{|W_k|}{2^k} < +\infty.$$

Damit ist der Abstandsbegriff überhaupt sinnvoll. Die Eigenschaften der Metrik - Identität, Symmetrie und Dreiecksungleichung - sind elementar beweisbar. $\square$

Nun wollen wir stetige Abhängigkeiten der Lösung W von der Anfangslage und von $(\varphi(k))_{k\in\mathbb{N}_0}$ untersuchen.

Definition 7.2.3 *Es sei $(\varphi(k))_{k\in\mathbb{N}_0}$ eine Folge aus $\mathbb{R}$, welche (7.10) erfüllt, und W bezeichne die zugehörige Lösung zu einem festen Anfangswert a_0. Existiert nun zu jedem $\varepsilon > 0$ ein $\delta > 0$, so daß für jede Folge $(\tilde{\varphi}(k))_{k\in\mathbb{N}_0}$ mit*

$$\sup_{k\in\mathbb{N}_0} |\varphi(k) - \tilde{\varphi}(k)| < \varepsilon,$$

welche zudem (7.10) erfüllt, gilt

$$\varrho(W,\tilde{W}) < \delta,$$

dann hängt die Lösung von (7.8) stetig von $(\varphi(k))_{k\in\mathbb{N}_0}$ ab.

Die folgende Aussage zeigt, daß Konverter erster Ordnung den Begriffen der Definition 7.2.3 nicht genügen.

Lemma 7.2.4 *Es sei $(\varphi(k))_{k\in\mathbb{N}_0}$ eine beliebige aber feste Folge aus $\mathbb{R}$ und $a_0 \in \mathbb{R}$ so gewählt, daß (7.10) erfüllt ist. Dann gelten folgende Aussagen.*
a) Die Lösung $(W_k)_{k\in\mathbb{N}_0}$ von (7.8) hängt nicht stetig vom Anfangswert a_0 ab.
b) Die Lösung hängt auch nicht stetig von $(\varphi(k))_{k\in\mathbb{N}_0}$ ab.

Beweis: Beide Aussagen wollen wir durch ein Gegenbeispiel zeigen.
a) Da $(\varphi(k))_{k\in\mathbb{N}_0}$ beliebig ist, wählen wir

$$\varphi(k) := a_0 \quad \forall k \in \mathbb{N}.$$

Dann sind beide Bedingungen aus (7.10) erfüllt. Nun sei o. B. d. A.

$$0 < b_0 - a_0 < \delta \leq 1.$$

Es bezeichne W^{a_0} und W^{b_0} die Lösungen von (7.8) bez. der Anfangswerte a_0 und b_0. Dann gilt nach der Definition der Metrik in $\mathcal{L}$ aus Folgerung 7.2.2 und nach (7.11)

$$\begin{aligned}
\varrho(W^{a_0}, W^{b_0}) &= \sum_{k=0}^{\infty} \frac{|W_k^{a_0} - W_k^{b_0}|}{2^k} \\
&= |a_0 - b_0| + \sum_{k=1}^{\infty} \frac{|\lfloor\varphi(k-1) - a_0\rfloor + a_0 - (\lfloor\varphi(k-1) - b_0\rfloor + b_0)|}{2^k} \\
&= |a_0 - b_0| + \sum_{k=1}^{\infty} \frac{|1 + (a_0 - b_0)|}{2^k} \\
&= b_0 - a_0 + |1 + (a_0 - b_0)| \\
&= 1.
\end{aligned}$$

Also ist der Abstand zwischen W^{a_0} und W^{b_0} gleich 1, gleich wie klein auch δ gewählt wird.
b) Es sei $\delta \in (0, 1)$ klein und

$$\varphi(k) := a_0 \quad \text{und} \quad \tilde{\varphi}(k) := a_0 - \frac{\delta}{k+1}.$$

Dann erfüllen φ und $\tilde{\varphi}$ die Bedingungen in (7.10). Weiterhin gilt

$$\max_{k \in \mathbb{N}_0} |\varphi(k) - \tilde{\varphi}(k)| = \delta.$$

Es bezeichne W und $\tilde{W}$ die Lösungen von (7.8) bez. der Folgen φ und $\tilde{\varphi}$. Dann gilt

$$\begin{aligned}
\varrho(W, \tilde{W}) &= \sum_{k=1}^{\infty} \frac{|\lfloor\varphi(k-1) - a_0\rfloor + a_0 + 1 - (\lfloor\tilde{\varphi}(k-1) - a_0\rfloor + a_0 + 1)|}{2^k} \\
&= \sum_{k=1}^{\infty} \frac{1}{2^k} \\
&= 1.
\end{aligned}$$

Also ist der Abstand zwischen W und $\tilde{W}$ immer gleich 1, egal wie klein δ gewählt wird. □

Offensichtlich sind die Sprungstellen von $\lfloor\cdot\rfloor$ die kritischen Stellen, an denen Stetigkeit vom Anfangswert und von der Folge $(\varphi(k))_{k \in \mathbb{N}_0}$ nicht gegeben ist. Wir wollen nun zwei Kriterien angeben, mit denen gerade diese Stetigkeit erzwungen werden kann.

Folgerung 7.2.5 Es sei $\varphi(k) - a_0 \notin \mathbb{Z}$ für alle $k \in \mathbb{N}_0$. Weiterhin seien die Bedingungen in (7.10) erfüllt. Dann gelten folgende Aussagen.
a) Die Lösung $(W_k)_{k \in \mathbb{N}_0}$ hängt stetig vom Anfangswert ab.
b) Die Lösung $(W_k)_{k \in \mathbb{N}_0}$ hängt stetig vom $(\varphi(k))_{k \in \mathbb{N}_0}$ ab.

Beweis: a) Es bezeichne W^{a_0} und W^{b_0} die Lösungen von (7.8) bezüglich der Anfangswerte a_0 und b_0. Dann gilt für $|a_0 - b_0| < \delta$

$$\begin{aligned}
\varrho(W^{a_0}, W^{b_0}) &= |a_0 - b_0| + \sum_{k=1}^{\infty} \frac{|\lfloor\varphi(k-1) - a_0\rfloor + a_0 - (\lfloor\varphi(k-1) - b_0\rfloor + b_0)|}{2^k} \\
&< 2\delta + \sum_{\{k \in \mathbb{N} : \lfloor\varphi(k-1) - a_0\rfloor \neq \lfloor\varphi(k-1) - b_0\rfloor\}} \frac{1}{2^k}.
\end{aligned}$$

Es sei nun $\tau > 0$. Dann gibt es ein $n_0 \in \mathbb{N}$, so daß

$$\sum_{k=n_0}^{\infty} \frac{1}{2^k} < \tau$$

erfüllt ist. Nach Voraussetzung und der Tatsache, daß n_0 endlich ist, gibt es ein $\delta > 0$ mit

$$\delta < \min_{k \in \{0,\ldots,n_0-1\}} \{\varphi(k) - a_0 - \lfloor \varphi(k) - a_0 \rfloor, \lfloor \varphi(k) - a_0 + 1 \rfloor - (\varphi(k) - a_0)\}.$$

Für alle b_0 mit $|a_0 - b_0| < \delta$ gilt dann

$$\begin{aligned}
&|a_0 - b_0| + \sum_{k=1}^{\infty} \frac{|\lfloor \varphi(k-1) - a_0 \rfloor + a_0 - (\lfloor \varphi(k-1) - b_0 \rfloor + b_0)|}{2^k} \\
&\quad < 2\delta + \sum_{\{k \in \mathbb{N} : \lfloor \varphi(k-1) - a_0 \rfloor \neq \lfloor \varphi(k-1) - b_0 \rfloor\}} \frac{1}{2^k} \\
&\quad < 2\delta + \sum_{k=n_0}^{\infty} \frac{1}{2^k} \\
&\quad < 2\delta + \tau.
\end{aligned}$$

Hieraus und wegen $\tau \to 0$ für $\delta \to 0$ wird der Abstand $\varrho(W^{a_0}, W^{b_0})$ beliebig klein, falls δ beliebig klein ist.
b) Die gleichen Überlegungen wie im Beweis von a) führen zur Behauptung in b). Es sei $n_0 \in \mathbb{N}$. Dann existiert nach obigen Voraussetzungen ein $\delta > 0$ mit

$$\delta < \min_{k \in \{0,\ldots,n_0-1\}} \{\varphi(k) - a_0 - \lfloor \varphi(k) - a_0 \rfloor, \lfloor \varphi(k) - a_0 + 1 \rfloor - (\varphi(k) - a_0)\}.$$

Somit gilt für alle $(\tilde{\varphi}(k))_{k \in \mathbb{N}_0}$, für die

$$\max_{k \in \mathbb{N}_0} |\varphi(k) - \tilde{\varphi}(k)| < \delta$$

erfüllt ist, schließlich

$$\begin{aligned}
\varrho(W, \tilde{W}) &= \sum_{k=1}^{\infty} \frac{|\lfloor \varphi(k-1) - a_0 \rfloor - \lfloor \tilde{\varphi}(k-1) - a_0 \rfloor|}{2^k} \\
&= \sum_{k=n_0}^{\infty} \frac{|\lfloor \varphi(k-1) - a_0 \rfloor - \lfloor \tilde{\varphi}(k-1) - a_0 \rfloor|}{2^k} \\
&\leq \sum_{k=n_0}^{\infty} \frac{1}{2^k}.
\end{aligned}$$

Da man für beliebiges n_0 ein $\delta > 0$ finden kann, kann man die rechte Seite der letzten Ungleichung beliebig klein machen. Damit ist alles gezeigt. □

Bemerkung 7.2.6 a) Wie wir gesehen haben, gibt es einen gewaltigen Unterschied zwischen der Störung der Anfangslage und der von $(\varphi(k))_{k \in \mathbb{N}_0}$. Aus $a_0 \neq b_0$ folgt immer $\varrho(W^{a_0}, W^{b_0}) \neq 0$. Aber aus $\varphi \neq \tilde{\varphi}$ kann $\varrho(W, \tilde{W}) = 0$ folgen. Das bedeutet, unterschiedliche AWA können gleiche Lösungen besitzen.
b) Bis hierher waren die Antworten, ob im allgemeinen Fall, gestörte Lösungen noch etwas mit der ungestörten zu tun haben, bis auf wenige Spezialfälle negativ. Nun wollen wir den punktweisen Abstand zwischen Original und gestörter Lösung untersuchen. Dazu geben wir erst einige Begriffsbildungen an.

Definition 7.2.7 *Es bezeichne W^{a_0} und W^{b_0} zwei Lösungen aus $\mathcal{L}$ bezüglich der Anfangswerte a_0 und b_0. Dann heißt die Lösung W^{a_0} stabil, falls es zu jedem $\varepsilon > 0$ ein $\delta > 0$ existiert, so daß aus $|a_0 - b_0| < \delta$ für alle $k \in \mathbb{N}_0$ folgt $|W_k^{a_0} - W_k^{b_0}| < \varepsilon$.*

Lemma 7.2.8 *Die Lösung W von (7.8) ist im allgemeinen nicht stabil.*

Beweis: Das Gegenbeispiel $\varphi(k) := a_0$ und $a_0 < b_0$ mit $|a_0 - b_0| < \delta = \frac{1}{2}$ im Beweis des Teils a) des Lemmas 7.2.4 liefert für alle $k \in \mathbb{N}$

$$\begin{aligned} |W_k^{a_0} - W_k^{b_0}| &= |a_0 - b_0 + \lfloor \varphi(k) - a_0 \rfloor - \lfloor \varphi(k) - b_0 \rfloor| \\ &> \frac{1}{2}. \end{aligned}$$

$\square$

Wir schwächen nun Definition 7.2.7 in folgender Weise ab.

Definition 7.2.9 *a) Die Lösung $W^{a_0} \in \mathcal{L}$ heißt K-stabil, falls ein $K > 0$ und ein $\delta > 0$ existieren, so daß aus $|a_0 - b_0| < \delta$ für alle $k \in \mathbb{N}_0$ gilt $|W_k^{a_0} - W_k^{b_0}| < K$.*
b) Es bezeichne W^p und W^r zwei Lösungen aus $\mathcal{L}$ bezüglich der Parametervektoren $p, r \in \mathbb{R}^n$ bei gleichem Anfangswert. Dann heißt die Lösung W^p K-robust, falls ein $K > 0$ und ein $\delta > 0$ existieren, so daß aus $||p - r|| < \delta$ mit einer geeigneten Norm in $\mathbb{R}^n$ für alle $k \in \mathbb{N}_0$ gilt $|W_k^p - W_k^r| < K$.

Bemerkung 7.2.10 In Definition 7.2.9 stören wir in a) die Anfangslage der AWA (7.8) und in b) Parameter, die in φ enthalten sind. Diese sind für einen Konverter erster Ordnung α und U_{ref}. Die Taktzeit wollen wir nicht stören, da die Lösung eines solcherart gestörten Modells nichts mehr mit der Originallösung zu tun hätte.

Lemma 7.2.11 *a) Jede Lösung $W^{a_0} \in \mathcal{L}$ ist K-stabil.*
b) Eine Lösung W aus $\mathcal{L}$ ist im allgemeinen nicht K-robust.
c) Gilt dagegen

$$\forall k \in \mathbb{N} : \left| \int_{t_0}^{t_k} U_0(\tau) \mathrm{d}\tau \right| \leq M < +\infty,$$

dann ist W sogar K-robust.

Beweis: a) Es sei $\delta \in (0, 1)$ und $|a_0 - b_0| < \delta$. Dann folgt aus

$$|\lfloor \varphi(k-1) - a_0 \rfloor - \lfloor \varphi(k-1) - b_0 \rfloor| \in \{0, 1\}$$

schließlich für alle $k \in \mathbb{N}$

$$|W_k^{a_0} - W_k^{b_0}| \leq |a_0 - b_0| + |\lfloor \varphi(k-1) - a_0 \rfloor - \lfloor \varphi(k-1) - b_0 \rfloor| < \delta + 1 =: K.$$

Da die Lösung von (7.8) im allgemeinen nicht stabil ist, kann K nicht beliebig klein werden.
b) In der Darstellung von φ gemäß (2.13) seien o. B. d. A. α und U_{ref} störbare Systemparameter, und es sei $s = 0$ für den Schwellenwert des Komparators. Weiterhin sei $U_0 = const.$ von Null verschieden. Dann folgt für $\alpha U_{ref} \neq \tilde{\alpha} \tilde{U}_{ref}$ aus

$$\lim_{k \to \infty} \left| \frac{1}{2\alpha U_{ref} \Delta t} \int_{t_0}^{t_k} U_0(\tau) \mathrm{d}\tau - \frac{1}{2\tilde{\alpha} \tilde{U}_{ref} \Delta t} \int_{t_0}^{t_k} U_0(\tau) \mathrm{d}\tau \right| = +\infty$$

schließlich

$$\begin{aligned}
&\lim_{k\to\infty} |W_k - \tilde{W}_k| \\
&\quad = \lim_{k\to\infty} \Bigg| \left\lfloor \frac{1}{2\alpha U_{ref}\Delta t} \int\limits_{t_0}^{t_{k-1}} U_0(\tau)\mathrm{d}\tau + \frac{k-1}{2} - a_0 \right\rfloor \\
&\qquad\qquad - \left\lfloor \frac{1}{2\tilde{\alpha}\tilde{U}_{ref}\Delta t} \int\limits_{t_0}^{t_{k-1}} U_0(\tau)\mathrm{d}\tau + \frac{k-1}{2} - a_0 \right\rfloor \Bigg| \\
&\quad = +\infty.
\end{aligned}$$

Also wächst $|W_k - \tilde{W}_k|$ bei beliebig kleiner Störung der Parameter mit $\alpha U_{ref} \neq \tilde{\alpha}\tilde{U}_{ref}$ über jede Schranke, und somit kann die Lösung von (7.8) nicht K-robust sein.
c) Es sei $(\tilde{\alpha}, \tilde{U}_{ref})$ eine Störung von (α, U_{ref}). Dann ist

$$\left| \frac{\alpha U_{ref} - \tilde{\alpha}\tilde{U}_{ref}}{2\alpha\tilde{\alpha}U_{ref}\tilde{U}_{ref}\Delta t} \right| =: Q < +\infty.$$

Hieraus, aus $x - \lfloor x \rfloor \in [0,1)$ und der Dreiecksungleichung folgt analog zu b)

$$\begin{aligned}
&|W_k - \tilde{W}_k| \\
&= \left| \left\lfloor \frac{1}{2\alpha U_{ref}\Delta t} \int\limits_{t_0}^{t_{k-1}} U_0(\tau)\mathrm{d}\tau + \frac{k-1}{2} - a_0 \right\rfloor - \left\lfloor \frac{1}{2\tilde{\alpha}\tilde{U}_{ref}\Delta t} \int\limits_{t_0}^{t_{k-1}} U_0(\tau)\mathrm{d}\tau + \frac{k-1}{2} - a_0 \right\rfloor \right| \\
&\leq 2 + \left| \left(\frac{1}{2\alpha U_{ref}\Delta t} - \frac{1}{2\tilde{\alpha}\tilde{U}_{ref}\Delta t} \right) \int\limits_{t_0}^{t_{k-1}} U_0(\tau)\mathrm{d}\tau \right| \\
&\leq 2 + \left| \frac{\alpha U_{ref} - \tilde{\alpha}\tilde{U}_{ref}}{2\alpha\tilde{\alpha}U_{ref}\tilde{U}_{ref}\Delta t} \right| M = 2 + QM =: K < +\infty.
\end{aligned}$$

Damit ist alles bewiesen. □

Bemerkung 7.2.12 Die vorangegangenen Aussagen führen also zu folgendem Fazit. Bei Änderung des Anfangswertes bleibt die gestörte Lösung von (7.8) in einem K-Schlauch um die Originallösung. Bei beliebig kleiner Störung der Systemparameter ist auch dies für unendliche Zeiten im allgemeinen nicht mehr realisierbar.

Nun wollen wir das Verhalten der Lösung der AWA (7.9) untersuchen. Ein einfacher Weg hierbei besteht darin, anzunehmen, daß der innere Kreis des Konverters zweiter Ordnung wie einer erster Ordnung funktioniert. Unter dieser Einschränkung lassen sich die vorangegangenen Aussagen leicht auf die nun gewonnene AWA erster Ordnung anwenden.

Lemma 7.2.13 *Es sei $|U_1(t)| \leq \beta U_{ref}$ für alle Zeiten $t \geq t_0$. Weiterhin gelten für*

$$\varphi_{U_1}(k) := \begin{cases} 0, & \text{falls } k = 0, \\ \dfrac{1}{2\beta U_{ref}\Delta t} \displaystyle\int\limits_{t_0}^{t_k} U_1(\tau)\mathrm{d}\tau + \frac{k}{2}, & \text{falls } k \in \mathbb{N}, \end{cases}$$

die beiden Bedingungen

$$\varphi_{U_1}(0) - (a_1 - a_0) \in [-1, 1),$$
$$\varphi_{U_1}(k) - \varphi_{U_1}(k-1) \in [0, 1], \quad \forall k \in \mathbb{N}.$$

Dann hängt die Lösung $(H_k)_{k\in\mathbb{N}_0}$ *der AWA* (7.9) *nicht stetig vom Anfangswert* (a_0, a_1) *und auch nicht von* $(\varphi_{U_1}(k))_{k\in\mathbb{N}_0}$ *ab. Weiterhin ist* $(H_k)_{k\in\mathbb{N}_0}$ *im allgemeinen weder stabil, K-stabil noch K-robust.*

Beweis: Wegen $|U_1(t)| \leq \beta U_{ref}$ ist das Verhalten eines Konverters zweiter Ordnung durch die AWA

$$\begin{array}{lcl} Q_{k+1} & = & Q_k + \sigma\left(\varphi_{U_1}(k) - Q_k\right), \\ Q_0 & = & a_1 - a_0 \end{array} \tag{7.12}$$

beschreibbar. Dies wird durch die Transfomation

$$H_{k+1} - H_k = Q_k \quad \text{und} \quad H_0 = a_0$$

vermittelt. Wegen der Forderungen an φ_{U_1} besitzt die AWA (7.12) nach Satz 2.1.3 die eindeutig bestimmte Lösung

$$Q_k = \begin{cases} a_1 - a_0, & \text{für } k = 0, \\ \lfloor \varphi_{U_1}(k-1) - (a_1 - a_0) \rfloor + (a_1 - a_0) + 1, & \text{für } k \in \mathbb{N}. \end{cases}$$

Wegen der obigen Transformation lautet die Lösung dann für $k \in \mathbb{N}$

$$H_k = a_0 + \sum_{n=0}^{k-1} Q_n = a_0 + (k-1)(a_1 - a_0 + 1) + \sum_{n=0}^{k-1} \lfloor \varphi_{U_1}(n-1) - (a_1 - a_0) \rfloor.$$

Es sei o. B. d. A. $U_0 = const$. Dann gilt für die L-Konstante aus Satz 4.1.19 $L_U = 0$. Weiterhin seien die Parameter, die in diesem Satz eine Rolle spielen, so gewählt, daß

$$\frac{\gamma}{\alpha} + \frac{4\alpha\Delta t}{\beta T_1} \in (0, 1)$$

hinreichend klein ist. Darüber hinaus gelte für die gestörte Anfangslage von (7.9)

$$(a_1 - a_0) \neq (\tilde{a}_1 - \tilde{a}_0).$$

Ist diese Störung hinreichend klein, dann existiert nach Satz 4.1.19 ein U_1, welches

$$\varphi_{U_1}(n-1) - (a_1 - a_0) \in \begin{cases} \left(\frac{n-1}{2}, \frac{n}{2}\right), & n \text{ ungerade}, \\ \left(\frac{n}{2}, \frac{n+1}{2}\right), & n \text{ gerade}, \end{cases}$$

und

$$\varphi_{U_1}(n-1) - (\tilde{a}_1 - \tilde{a}_0) \in \begin{cases} \left(\frac{n-1}{2}, \frac{n}{2}\right), & n \text{ ungerade}, \\ \left(\frac{n}{2}, \frac{n+1}{2}\right), & n \text{ gerade} \end{cases}$$

erfüllt. Somit gilt für die beiden Lösungen von (7.9)

$$|H_k - \tilde{H}_k| = |(a_0 - \tilde{a}_0) + (k-1)[(a_1 - a_0) - (\tilde{a}_1 - \tilde{a}_0)]| \to +\infty.$$

Also ist H nicht K-stabil und kann demzufolge erst recht nicht stabil sein.
Daß H nicht K-robust ist, folgt aus der Tatsache, daß bereits die Lösung Q von (7.12) nicht K-robust ist. Die beiden anderen Behauptungen folgen ebenfalls aus Lemma 7.2.4, angewandt auf (7.12). Damit ist alles gezeigt. □

Bemerkung 7.2.14 Die Existenz des Gegenbeispiels im Beweis von Lemma 7.2.13 benötigt die Aussage des Satzes 4.1.19. Die explizite Angabe von U_0, so daß H nicht K-stabil ist, läuft wiederum auf die Lösung der Differenzengleichung, die den Konverter zweiter Ordnung beschreibt, hinaus. Dies ist aber nur näherungsweise gelöst.

8 Vergleich zur Modellbildung der Ingenieure

In diesem Kapitel wollen wir die grundlegenden Herangehensweisen, wie sie in der ingenieurtechnischen Literatur häufig dargestellt sind, erläutern. Wir werden die wichtigsten Begriffe formulieren, Vor- und Nachteile der Modellbildungen angeben und Vergleiche zu unserem Modell ziehen. Kompakte Darstellungen zu dieser Thematik sind in [28], [52], [46], [37] zu finden. Welche Bedeutung die aktuelle Forschung zum Thema Σ-Δ-Modulatoren besitzt, läßt sich unter anderem an der Vielzahl von Publikationen in der ingenieurtechnischen Literatur ablesen. Allein, wenn man online auf der Internetseite der Ingenieursvereinigung IEEE unter www.ieee.org nach Artikeln sucht, die den Begriff Σ-Δ-Converter enthalten, kann man sich in mehreren hundert Aufsätzen zu dieser Thematik informieren. Insofern ist es nicht möglich, einen kompletten Vergleich durchzuführen.
Wir wollen hier Konverter mit einem Komparator betrachten, die nur eine Quantisierungsstufe besitzen. Fassen wir den Ausgang des Konverters als Zahlenfolge $(X_k)_{k\in\mathbb{N}_0}$ auf, so besitzen die Codewörter X_k nur ein Bit. Somit können gewisse Fehler gar nicht auftreten, wie sie bei Komparatoren, die mehrere Quantisierungsstufen haben, möglich sind. Das sind z.B. Fehler aufgrund ungleicher Sprunghöhen des Vergleichers, differentielle Nichtlinearitäten oder Leerstellen in Codewörtern (missing codes). Außerdem werden durch die 1-Bit-Quantisierung die Genauigkeitsanforderungen an die Bauelemente erheblich entschärft und somit eine kostengünstige Integration erleichtert. Da am Ausgang des Konverters nur zwei Werte auftreten, kann - bei geeigneter Schaltung - der zur Rückkopplung benötigte Digital-Analog-Konverter entweder sehr einfach sein oder sogar ganz entfallen [21, p.77]. In unserem Modell haben wir ihn von Anfang an als Multiplizierer dargestellt (vgl. Abb. 1). Der geringen Wortbreite von einem Bit entspricht eine sehr einfache Arithmetik des nachfolgenden digitalen Filters. Dies sind die wesentlichen Gründe, warum Konverter, die nach dem Σ-Δ-Prinzip arbeiten, so populär sind.
Wir werden nicht alle auftretenden Begriffe aus der Signalverarbeitung erläutern können, da dies den Rahmen sprengen würde. Wir wollen aber auf diejenigen näher eingehen, die bei der Σ-Δ-Modulation eine besondere Rolle spielen.

8.1 Überblick

Die technologischen Fortschritte in der Halbleitertechnik ermöglichen es in zunehmendem Maße, Methoden der digitalen Signalverarbeitung auch zur Realisierung hochgenauer Analogfunktionen einzusetzen. Besondere Bedeutung hat in den letzten Jahren die Σ-Δ-Wandlung zur Realisierung hochauflösender und schneller AD- und DA-Wandler-Schaltkreise in VLSI-Technik erlangt.
Dieses Verfahren zur AD-Umsetzung kombiniert die Überabtastung (oversampling) eines Signals mit einer Rauschformung (noise shaping) und nachfolgender digitaler Filterung. Damit wird die Auflösung erhöht und gleichzeitig das Quantisierungsrauschen in niederen Frequenzen verringert. Die Abtastung des Eingangssignals erfolgt z. T. mit hundert- bis tausendfacher Nyquistrate [43, p.167].

- Ein Σ-Δ-Konverter wandelt das analoge Eingangssignal in eine hochfrequente serielle Folge grob quantisierter Abtastwerte. Typisch ist eine Bitfolge mit einem Bit Auflösung.
- Durch anschließende digitale Tiefpaßfilterung (finite impulse response = FIR, Mittelwertbildung) wird das Konvertersignal in hochauflösende Parallelworte mit stark verringerter Abtastrate (decimation) umgesetzt (vgl. Abb. 33).

Durch die hohe Überabtastung des Eingangssignals $U_0(t)$ mit der Frequenz $f_m = \frac{1}{\Delta t}$ wird die Energie des Quantisierungsrauschens gleichmäßig auf ein großes Frequenzintervall $\left(0, \frac{f_m}{2}\right)$ ver-

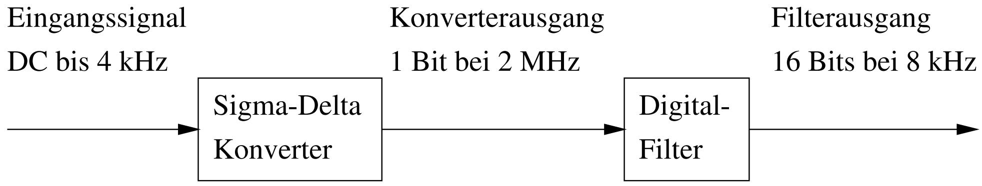


Abbildung 33: Blockstruktur einer Σ-Δ-Modulation (Beispiel: Sprachsignal)

teilt. Der in das kleinere Frequenzband des Eingangssignals fallende Anteil dieses rauschförmigen Fehlersignals wird durch zusätzliche Rauschformung und nachträgliche Filterung stark reduziert (vgl. Abschnitte 8.4.2, 8.4.3).
Ein Σ-Δ-Modulator kann als Umsetzer aufgefaßt werden, der ein Analogsignal mit einer sehr hohen Frequenz abtastet und die dabei entstehenden digitalen Daten seriell bei einer viel kleineren Frequenz wieder ausgibt. Bild 34 stellt die Relationen zwischen der Bandbreite (Frequenzintervall) des Eingangssignals $U_0(t)$ und den verschiedenen Abtastfrequenzen dar.

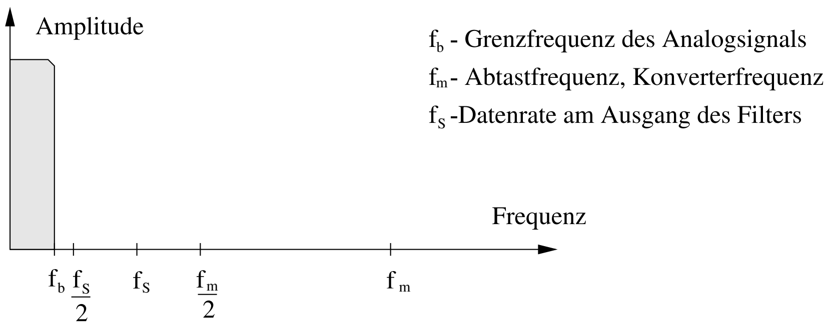


Abbildung 34: Relationen zwischen den Abtastfrequenzen

Σ-Δ-Modulatoren eignen sich hervorragend für Aufgaben der Signalerfassung, im Gegensatz zum Begriff der Datenerfassung, und haben gegenüber vielen anderen Modulatortypen (Stufenwandler, Parallelwandler) sogar spezifische Vorteile:

- komplette monolithische Wandlerbausteine, einschließlich der Filterung,
- kostengünstige Lösung bei hoher Auflösung,
- frei von differentieller Nichtlinearität,
- kein externes Sample-and-Hold-Glied nötig,
- großer Signal-Rausch-Abstand,
- Reduktion der Anforderungen an analoge Vorfilter.

Haupteinsatzgebiet von monolithischen Wandlern nach dem Σ-Δ-Verfahren ist gegenwärtig die Digitalisierung von Audiosignalen (Sprache, Musik) mit Auflösungen bis zu 18 Bit. Dies entspricht einer Unterteilung des zu untersuchenden Bereiches in $2^{18} = 262144$ Werten.
Zunehmend werden Σ-Δ-Modulatoren auch für die Signalerfassung verwendet, wenn es um hohe Auflösung bei gleichzeitig gutem Stör- und Rauschverhalten geht. Zur AD-Umsetzung niederfrequenter Signale sind heute Schaltkreise mit Auflösungen bis 24 Bit verfügbar.
Die Grundschaltung eines Σ-Δ-Modulators erster Ordnung besteht im wesentlichen aus einem Konverter mit analogem Integrator (es existieren auch Konverter mit diskreten Integratoren)

und einer digitalen Rechenschaltung (digitaler Filter), wie sie im Bild 36 dargestellt ist. Dabei kann das Element im Rückzweig als Analogschalter, als 1-Bit-DA-Umsetzer oder wie im Bild 1 als Multiplizierer aufgefaßt werden. Der Analogkonverter arbeitet im Prinzip als abgetasteter (diskreter) Regelkreis. Der daran anschließende digitale Signalprozessor wirkt als Digitalfilter und realisiert gleichzeitig die notwendige digitale Nachverarbeitung (Mittelwertbildung) der Konverterausgangsfolge $(X_k)_{k\in\mathbb{N}_0}$. Die Wirkungsweise des Konverters ist in Kapitel 2 erläutert.
Der Funktionsweise eines Konverters erster Ordnung entsprechend, vorausgesetzt seien ideale Bauelemente, entsteht für das Eingangssignal $U_0(t) = 0$ am Ausgang eine Bitfolge $(X_k)_{k\in\mathbb{N}_0}$ mit symmetrischem Tastverhältnis. Eine Unsymmetrie im Tastverhältnis repräsentiert somit alle Offset- und Schaltschwellenfehler sowie das Quantisierungsrauschen der Anordnung [37], [51].

- Die Größe von U_1 am Integratorausgang schwankt somit ständig um den Umschaltwert des Komparators und führt so das Konvertersignal $\pm\alpha U_{ref}$ dem Eingangssignal $U_0(t)$ taktsynchron nach.
- Die Pulsfolgen am Konverter- bzw. Modulatorausgang geben an, ob der aktuelle Analogwert des Eingangssignals größer oder kleiner als der vorhergehende ist.
- Die zeitlichen integralen Mittelwerte von U_0 und X sind bis auf einen beschränkten Fehler gleich (vgl. Lemma 4.2.2), falls $|U_0(t)| \leq \alpha U_{ref}$ vorausgesetzt ist.

Die Differenz zwischen den integralen Mittelwerten von U_0 und X

$$F_m^n = \frac{1}{m\Delta t} \int\limits_{t_n}^{t_{n+m}} U_0(\tau)\mathrm{d}\tau - \frac{1}{m\Delta t} \int\limits_{t_n}^{t_{n+m}} X(\tau)\mathrm{d}\tau$$

entspricht einem Fehlersignal, welches als Quantisierungsrauschen bezeichnet wird, dessen Energie nur von der Auflösung bestimmt wird. Die serielle Bitfolge $(X_k)_{k\in\mathbb{N}_0}$ am Konverterausgang enthält bereits die vollständige Information über das Analogsignal, die aber durch das Fehlersignal verdeckt ist.

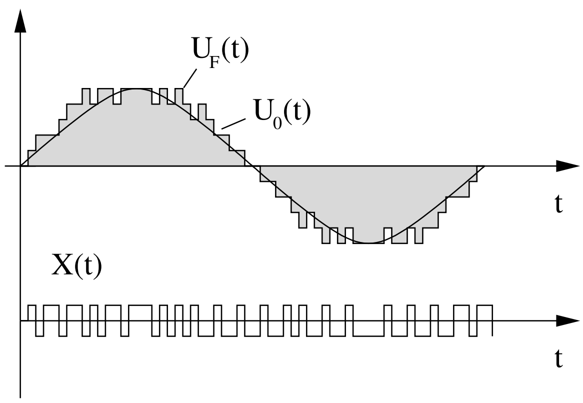


Abbildung 35: Analogsignal, Filterausgang und zugehörige Bitfolge

Um zu verstehen, wie nach dem Σ-Δ-Verfahren aus der 1-Bit-Folge am Konverterausgang hochauflösende Datenworte erzeugt werden, muß man die Arbeitsweise nicht im Detail kennen. Notwendig sind aber zumindest qualitative Kenntnisse zu den Konzepten

- Überabtastung (oversampling),
- Rauschformung (noise shaping),

- digitale Filterung,
- Abtastratenreduktion (decimation),

welche wir im Kapitel 8.4 darstellen werden.
Das Ziel der nachfolgend im einzelnen beschriebenen Konzepte ist es, eine möglichst originalgetreue Abbildung des analogen Eingangssignals U_0 aus der Bitfolge $(X_k)_{k\in\mathbb{N}_0}$ zu gewinnen. Die große Leistungsfähigkeit des Σ-Δ-Verfahrens wird erst durch die kombinierte Wirkung der einzelnen Konzepte erreicht.

8.2 Beschreibung im Zeitbereich

Die Modellierung des Σ-Δ-Verfahrens im Zeitbereich entspricht genau der Herangehensweise, wie wir sie in den Kapiteln 2.1.1 mit der Herleitung der Differenzengleichungen und den Kapiteln 3.1.1, 3.1.2 mit den Betrachtungen zum digitalen Filter dargestellt haben. Dabei definiert die Differenz zwischen letztem Integratorausgang und Komparatorausgang des abgetasteten Regelkreises z. B. erster Ordnung

$$U_1(t_k) - X(t_k) =: e_k \quad \text{zum Abtastzeitpunkt} \quad t_k = t_0 + k\Delta t$$

den Quantisierungsfehler. Die Verteilung des Quantisierungsfehlers in Abhängigkeit von der Frequenz spielt eine wichtige Rolle in der Charakterisierung des Σ-Δ-Verfahrens.
Eine analytische Auswertung zwischen Einganssignal U_0, Bitfolge $(X_k)_{k\in\mathbb{N}_0}$ und Quantsierungsfehler $(e_k)_{k\in\mathbb{N}_0}$ bereitet aufgrund der Unstetigkeit und Nichtlinearität des Regelkreises sowie der Tatsache, daß U_0 im allgemeinen nicht bekannt ist, große Probleme. An dieser Stelle werden in der Literatur häufig vereinfachende Annahmen an stochastische Charakteristika von $(e_k)_{k\in\mathbb{N}_0}$ und an die L-Stetigkeit von U_0 getroffen, um eine Abschätzung der Fehlergrößen im Vergleich zur Signalgröße zu erhalten. Dieses Verhältnis wird als Signal-Rausch-Abstand (signal-to-noise-ratio) bezeichnet. Eine andere Möglichkeit sind Simulationen, um Schlüsse daraus zu ziehen. Darauf wollen wir hier nicht näher eingehen.
Wir wollen nun die Streuung der Mittelwerte der Ausgangsworte bei verschiedenen digitalen Filtern berechnen. Dazu wird vereinfachend angenommen, daß die nichtlinearen Änderungen von U_0 im Zeitintervall $[t_j, t_{j+m})$ vernachlässigt werden können und die einzelnen Fehler e_k voneinander unabhängig bzw. unkorreliert und identisch gleichverteilt sind. Sie haben dann innerhalb eines Ausgangswortes den gleichen Erwartungswert

$$E(e_k) = e$$

und die quadratischen Effektivwerte der einzelnen Fehler e_k können addiert werden

$$\overline{R}^2 = \sum_{k=1}^{m} \overline{e}_{j+k}^2.$$

Wir nehmen weiterhin $\Delta t = T_1$ und $\alpha = U_{ref} = 1$ an. Es gelten die Bezeichnungen aus den Kapiteln 3.1.1 und 3.1.2 Für einen Σ-Δ-Modulator erster Ordnung erhalten wir dann folgende Aussagen. Die Streuung für einfache Mittellung lautet nach [28], Lemma 3.2.1

$$D(F_m^j) = \frac{\sqrt{2}e}{m}$$

und für eine gleitende Mittelung nach Lemma 3.2.2

$$D(F_{mn}^j) = \begin{cases} \dfrac{\sqrt{2}e}{\sqrt{m}(n-m+1)}, & \text{für } m \leq \dfrac{n+1}{2} \\ \dfrac{\sqrt{2}e}{m\sqrt{n-m+1}}, & \text{für } m > \dfrac{n+1}{2}. \end{cases}$$

Der Ausdruck

$$F := \frac{e^2}{D^2(.)} = \frac{\text{Rauschleistung eines quantisierten Abtastwertes}}{\text{Rauschleistung eines Ausgangswortes}}$$

wird als Rauschreduktionsfaktor bezeichnet. Wir sehen also, daß sich bei gleitender Mittelwertbildung der durch die Quantisierung bewirkte mittlere Fehler des Mittelwertes von m Abtastwerten gegenüber einfacher Mittelwertbildung reduziert. Die folgende Tabelle zeigt ausgezeichnete Werte für den Rauschreduktionsfaktor F und die Auflösung A bei verschiedenen Werten von m in Abhängigkeit von n. Dabei sei

$$F_1 = \frac{m(n-m+1)^2}{2}, \quad F_2 = \frac{m^2(n-m+1)}{2} \quad \text{und} \quad A = m(n-m+1).$$

m	F_1	F_2	A	Bemerkung
1	$\frac{n^2}{2}$	-	n	einfacher Mittelwert
$\frac{n+1}{3}$	$2\left(\frac{n+1}{3}\right)^3$	-	$2\left(\frac{n+1}{3}\right)^3$	Trapezwichtung, max. Rauschunterdrückung
$\frac{n+1}{2}$	$\frac{1}{2}\left(\frac{n+1}{2}\right)^3$	$\frac{1}{2}\left(\frac{n+1}{2}\right)^3$	$\left(\frac{n+1}{2}\right)^3$	Dreieckwichtung, max. Auflösung
$2\frac{n+1}{3}$	-	$2\left(\frac{n+1}{3}\right)^3$	$2\left(\frac{n+1}{3}\right)^3$	Trapezwichtung, max. Rauschunterdrückung
n	-	$\frac{n^2}{2}$	n	einfacher Mittelwert

Wir wollen nun noch die Ergebnisse für eine gewichtete Mittelung angeben. Dabei bezeichne $\{w_1, \ldots, w_m\} \subset (0, +\infty)$ die Wichtungsfaktoren mit

$$W = \sum_{j=1}^{m} w_j.$$

Dann gilt für den Rauschreduktionsfaktor F bei gewichtetem Mittel

$$F = \begin{cases} \dfrac{W^2}{2m} = \dfrac{m}{2}(n-m+1)^2, & \text{für } m \leq \dfrac{n+1}{2} \\[2ex] \dfrac{W^2}{2(n-m+1)} = \dfrac{m^2}{2}(n-m+1), & \text{für } m > \dfrac{n+1}{2}. \end{cases}$$

Nach [4] gibt es bei gewichteter Mittelwertbildung ein theoretisches Maximum für den Rauschreduktionsfaktor F bei konstantem n für parabolische Wichtungsfaktoren w_j

$$F_{opt} = \frac{n(n+1)(n+2)}{12} \quad \text{bei} \quad w_j = \frac{j(n-j+1)}{n}, \quad 0 \leq j \leq n+1.$$

Allerdings ist eine technische Realisierung dieser parabolischen Wichtungsfaktoren problematisch, dagegen ist eine Dreieckswichtung relativ einfach mit Zähler und Akkumulator zu erreichen [61], [63]. Darüber hinaus liegt der Rauschreduktionsfaktor F bei Dreieckswichtung nur um etwa 1.25 dB unter diesem optimalen Wert F_{opt}.

8.3 Beschreibung im Frequenzbereich

In diesem Abschnitt orientieren wir uns stark an [3]. Eine Analyse der Σ-Δ-Modulation im Frequenzbereich zeigt ebenfalls sehr deutlich seine grundsätzliche Wirkungsweise und Funktion für die AD-Wandlung. Σ-Δ-Modulatoren, wie sie in den Abbildungen 1 und 2 dargestellt sind, entsprechen nichtlinearen abgetasteten Regelsystemen. Die genaue deterministische Analyse ist auch dementsprechend schwierig und aufwendig. Deshalb versucht man, mit Vereinfachungen Voraussetzungen zu schaffen, die es erlauben, Analysemethoden der linearen Systeme anzuwenden. Die so erhaltenen Aussagen müssen danach mit den nichtlinearen Simulationen im Zeitbereich verglichen werden. Die wesentliche Vereinfachung wird in zwei Schritten durchgeführt.

- In einem ersten Schritt wird der Komparator durch eine Addition eines Fehlereinganges $e(t)$ (Quantisierungsfehler) zum Integratorausgangssignal (Nutzsignal) ersetzt. Damit eliminiert man die Nichtlinearität des ursprünglichen Systems noch nicht, sondern verschiebt diese auf den Fehler $e(t)$. Dieser ist nicht nur vom Eingangssignal $U_0(t)$, sondern über die Rückkopplung sogar von den vorangegangenen Werten abhängig.

- Der entscheidende Schritt zur Linearisierung besteht darin, daß man den Fehler $e(t)$ vom Nutzsignaleingang entkoppelt und als unabhängigen Eingang betrachtet.

Die erste Vereinfachung verändert das nichtlineare Modell nicht, sondern kann als Definition des Quantisierungsfehlers betrachtet werden. Die zweite dagegen wirkt stark einschränkend, erlaubt aber nun, zusätzliche Annahmen für $e(t)$ zu treffen, deterministische wie auch statistische. Die Signalflußdiagramme im z-Bereich - vereinfachend als diskreter Frequenzbereich auffaßbar - für Modulatoren erster und zweiter Ordnung sind in Bild 36 dargestellt. Die Herleitung der Übertragungsfunktionen und die besondere Darstellung in diesem Fall sind in [3], [46] zu finden.

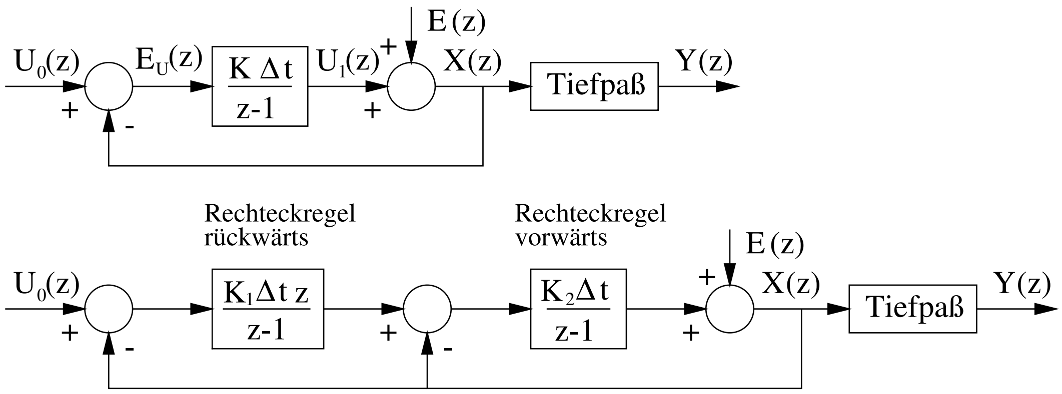


Abbildung 36: Signalflußdiagramme der Σ-Δ-Modulatoren erster und zweiter Ordnung im z-Bereich

In der graphischen Darstellung der Signalflüsse bedeuten K bzw. $K_{1,2}$ Vertärkungsfaktoren, die im engen Zusammenhang mit den Integrationszeitkonstanten stehen. Man kann deutlich das Tiefpaßverhalten der Integration ablesen. Nun ist man daran interessiert, im z-Bereich einen funktionalen Zusammenhang der Art

$$X(z) = G_U(z)U_0(z) + G_E(z)E(z)$$

herzuleiten. In dieser Form wären dann die Übertragung des Eingangssignals $U_0(z)$ und des Fehlersignals $E(z)$ voneinander getrennt, wobei $G_U(z)$ bzw. $G_E(z)$ die Übertragungsfunktionen des Eingangs und des Fehlers sind. Kennt man diese Funktionen, so ist das System vollständig beschreibbar. Leider ist dies im allgemeinen Fall nicht möglich. Man kann diese Trennung nur erreichen, falls man den Quantisierungsfehler $e(t)$ abtastet und die Integration durch Summation (numerische Integration durch Rechteckregel) ersetzt. Falls das Eingangssignal konstant ist, führt man mit obigen Operationen keinen Fehler ein. Aus diesem Grund ist diese Klasse von Signalen von großem Interesse. Dagegen ist bei sinusförmigen Signalen $U_0(t)$ mit zusätzlichen Fehlern zu rechnen. Diese sind aber bei Frequenzen, die weit unter der halben Abtastfrequenz liegen, vernachlässigbar (Abtasttheorem, vgl.[20]). Nach dieser dritten Vereinfachung bzw. Annahme erhalten wir für die Konverter erster und zweiter Ordnung folgende Zusammenhänge [3]

$$\begin{aligned}
X(z) &= \frac{K\Delta t}{z-1+K\Delta t}U_0(z) + \frac{z-1}{z-1+K\Delta t}E(z),\\
X(z) &= \frac{K_1K_2\Delta t^2 z}{(z-1)^2+K_2\Delta t(K_1\Delta t+1)z-K_2\Delta t}U_0(z)\\
&\quad + \frac{(z-1)^2}{(z-1)^2+K_2\Delta t(K_1\Delta t+1)z-K_2\Delta t}E(z).
\end{aligned}$$

Die nachfolgende Abbildung stellt die Übertragungsfunktionen der Fehler der Konverter erster (oberer Graph) und zweiter (unterer Graph) Ordnung für $K = K_1 = K_2 = \Delta t = 1$ dar. Die Tatsache, daß niederfrequente Fehleranteile beim Konverter zweiter Ordnung gedämpft, dafür höherfrequente verstärkt werden, nennt man Rauschformung. Wir werden darauf im Abschnitt 8.4.3 zurückkommen.

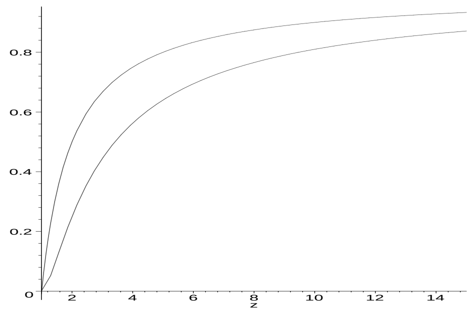


Abbildung 37: Fehlerübertragungsfunktionen bei Σ-Δ-Konverter erster und zweiter Ordnung im z-Bereich mit $z \in (0.99, 15)$

8.4 Begriffe

In den folgenden Abschnitten wollen wir uns der speziellen Charakteristik des Σ-Δ-Verfahrens widmen und die Begriffe, die wir in Kapitel 8.1 aufführten, erläutern.

8.4.1 Überabtastung

Da das analoge Eingangssignal jeden Wert innerhalb eines zulässigen Bereiches annehmen kann, das digitale Ausgangssignal aber quantisiert ist, liegt bei idealer Modulation die Differenz zwischen beiden im Intervall $\left[-\frac{LSB}{2}, \frac{LSB}{2}\right]$. Dabei hat das Least Significant Bit für eine Wortlänge 2^m den Wert

$$LSB = \frac{\text{Größe des aufzulösenden Bereiches}}{2^m}.$$

Auch dieser Fehler wird als Quantisierungsfehler bezeichnet, wie die Differenz von Integrator- und Komparatorausgang.
Wird von einem idealen Konverter ein konstantes Eingangssignal mehrfach gewandelt (vgl. [22]), so entsteht immer das gleiche Ausgangssignal, dessen Auflösung durch den Quantisierungsfehler begrenzt ist. Überlagert man das konstante Eingangssignal mit einem mittelwertfreien Signal (Störung), nimmt eine große Anzahl von Abtastwerten des Konverters und bildet daraus den Mittelwert, so kann die Eingangsgröße mit einer viel höheren Auflösung gemessen werden, als sie der Modulator selbst liefert. Dabei sei die Störung mittelwertfrei für hinreichend große Zeiten. Die Überlagerung kann ein Sinus-, Dreieck- oder Rauschsignal sein, deren Frequenz nicht mit der Abtastfrequenz des Konverters korreliert.
Σ-Δ-Modulatoren erzeugen ihr eigenes Dithersignal und benötigen keine externe Quelle. Die Art des überlagerten Signals beeinflußt die effektiv erreichbare Auflösung. Wesentlich an dieser Stelle ist, daß für eine Erhöhung der Auflösung um je ein Bit die Anzahl der Abtastungen, also die Abtastfrequenz, jeweils verdoppelt werden muß.

- Durch Überabtastung (oversampling) kann eine m-fache Auflösung der AD-Wandlung erreicht werden, wenn mit einer 2^m-fachen Abtastfrequenz gearbeitet wird.
- Überabtastung ist ein Mittel zur Erhöhung der Auflösung von AD-Modulatoren. Es besteht ein enger Zusammenhang zum Quantisierungsfehler bzw. Quantisierungsrauschen.

Ein idealer Modulator besitzt im Frequenzintervall $\left[0, \frac{f_S}{2}\right]$ und darüber hinaus ein gleichverteiltes Quantisierungsrauschen. Die Frequenzanteile des Quantisierungsfehlers haben in diesem Intervall die gleiche Amplitude. Dieser Amplitudenwert des weißen Rauschens hat den Wert

$$|u_r| = \frac{LSB}{\sqrt{12 f_s}}.$$

Der klassische Weg zur Verbesserung des Signal-Rausch-Abstandes ist der Einsatz von Modulatoren mit höherer Auflösung, d. h., LSB soll möglichst klein sein. Nach den letzten Betrachtungen läßt sich die spektrale Rauschamplitude aber auch dann drastisch reduzieren, wenn die Abtastfrequenz f_S entsprechend erhöht wird.
Ist ein gegebener Modulator nicht ideal, so ist seine Rauschamplitude größer als die theoretisch minimale. Damit ist dann auch die effektive Auflösung dieses Modulators kleiner als n Bit. Die effektive Auflösung (effective number of bits - ENOB) wird definiert durch den Ausdruck

$$\text{ENOB} = \frac{\text{Signal-Rausch-Abstand[dB]} - 1.76\text{dB}}{6.02\text{dB}}.$$

Hieraus folgt, daß sich die effektive Auflösung eines AD-Umsetzers bei Vergrößerung des Signal-Rausch-Abstandes um je 6 dB um ein Bit erhöht.

8.4.2 Digitale Filterung und Abtastratenreduktion

Wie wir in Abschnitt 8.4.1 gesehen haben, wird die spektrale Rauschamplitude in dem Maße kleiner, desto höher die Abtastfrequenz f_S gewählt wird, da sich die konstante Rauschleistung über ein größeres Frequenzintervall verteilt (wieder alles im Idealfall eines weißen Quantisierungsrauschens). Wird nun dem Modulator ein digitales Tiefpaßfilter (z. B. Mittelwertbildner) nachgeschaltet, so kann ein großer Teil der hohen Frequenzen des Quantisierungsrauschens herausgefiltert werden, ohne das niederfrequente Nutzsignal zu beeinflussen (vgl. Abbildung 38). Durch diese Vorgehensweise wird also die effektive Anzahl von Bits und somit auch die Auflösung

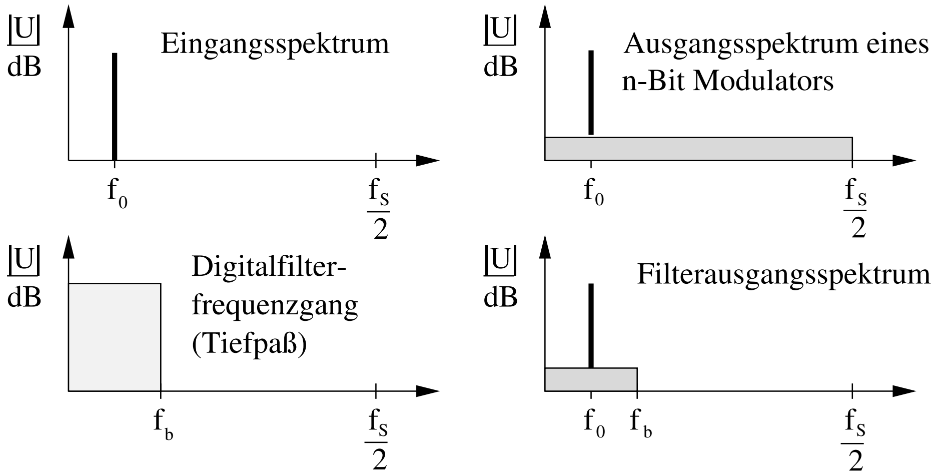


Abbildung 38: Filterung des Quantisierungsrauschens

erhöht. Man erreicht so eine hochauflösende AD-Wandlung unter Verwendung eines Konverters mit relativ geringer Auflösung.
Da durch den Einsatz dieses Tiefpaßfilters das Frequenzintervall reduziert wird, kann die Datenrate des Ausgangssignals kleiner sein als die originale Abtastrate und trotzdem das Nyquistkriterium [45, p.170], [17, Kap.3.3] erfüllen. Dazu wird nur jedes m-te Ergebnis am Ausgang ausgegeben, während man die anderen unterdrückt. Dieser Zusammenhang ist bekannt als Dezimation um den Faktor m (vgl. Abbildung 39).
Dabei kann m jeden Wert annehmen, solange die Abtastrate des Ausgangssignals größer bleibt als die zweifache Signalbandbreite (Abtasttheorem). Die Abtastratenreduktion bewirkt keinen Verlust an Information bez. des Eingangssignals.

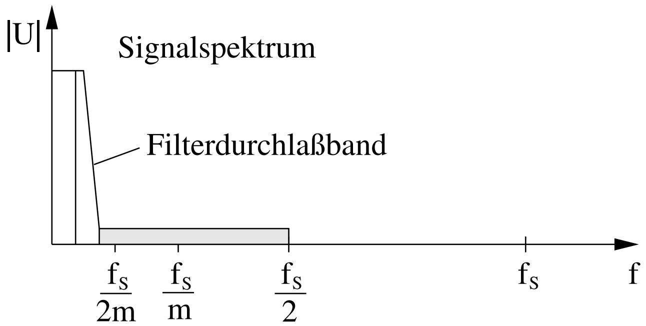


Abbildung 39: Zur Abtastratenreduktion der Ausgangsrate

8.4.3 Rauschformung

Wird zur Erhöhung der Auflösung nur die Überabtastung eingesetzt, so muß - wie wir in Abschnitt 8.4.1 kennengelernt haben - mit einer 2^n-fachen Abtastfrequenz gearbeitet werden, um eine n-fache Auflösung zu erhalten.
Durch die sogenannte Rauschformung (noise shaping) kann man mit einer z. T. wesentlich geringeren Abtastrate arbeiten, um die gleiche Auflösung zu erzielen. Oder umgekehrt erreicht man bei gegebener Abtastfrequenz eine höhere Auflösung. Dabei wird durch die Rauschformung das Quantisierungsrauschen aus dem Signalfrequenzband, d. h. aus den niederen Frequenzen herausgedrängt und kann danach wie in Abschnitt 8.4.2 durch einen Tiefpaßfilter ausgefiltert werden.

Die Rauschformung wird bei Σ-Δ-Modulatoren durch den Konverter realisiert (vgl. Bild 36). Bei Einsatz mehrerer Integratoren und Summierstufen (in Reihe geschaltet) (vgl. Abbildung 2) oder durch Kombination mehrerer Konverter erster oder zweiter Ordnung (kaskadiert - MASH-Konverter) können Konverter höherer Ordnung realisiert werden [52, Ch.6]. Mit ihnen kann die Rauschformung und damit letztlich wieder die Auflösung noch weiter erhöht werden. Bild 40 zeigt den Zusammenhang zwischen Überabtastung, Ordnung (Anzahl der in Reihe geschalteten Integratoren) der Konverter und dem erreichbaren Signal-Rausch-Abstand [61].

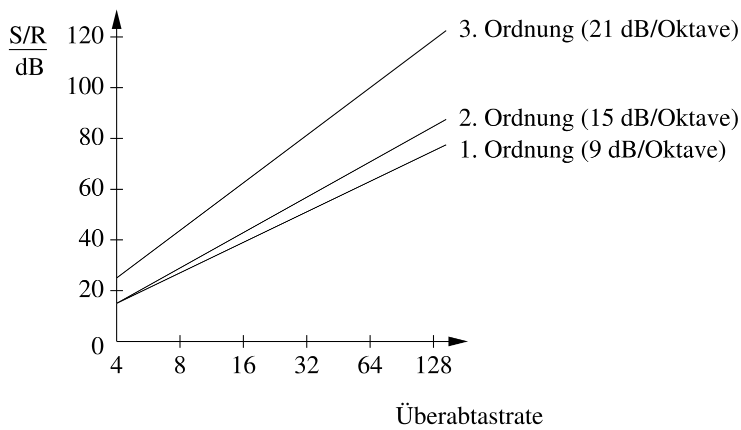


Abbildung 40: Signal-Rausch-Abstand, Überabtastung und Ordnung des Konverters

8.5 Σ-Δ-Modulatoren als analoges System

Ausgangspunkt für diesen Abschnitt sind die in den Kapiteln 4.1.2 und 4.1.3 hergeleiteten AWP für Σ-Δ-Modulatoren. Diese analogen Systeme wollen wir mit den Begriffen aus der Theorie der Signalverarbeitung untersuchen. Hierbei kann es keine Terme wie Quantisierungsfehler oder Fehlerübertragungsfunktion geben, da wegen $\Delta t \to 0$ die Auflösung unendlich groß und damit das Quantisierungsrauschen unendlich klein werden. Wenn man Σ-Δ-Modulatoren also näherungsweise als analoge Systeme betrachtet, so bleibt nur eine Übertragungsfunktion des Nutzsignals $U_0(t)$ übrig. Die entsprechenden Definitionen haben wir [42], [53] und [45] entnommen.
Die Laplacetransformation von (4.38) und (4.39) liefert

$$\begin{aligned} \Phi(s) &= \frac{1}{\alpha U_{ref}} \mathcal{U}_0(s), \\ s\Phi(s) + \frac{\alpha}{\beta T_1}\Phi(s) &= \frac{1}{\beta T_1 U_{ref}} \mathcal{U}_0(s). \end{aligned}$$

Damit lauten die Übertragungsfunktionen $H(s) := \frac{\Phi(s)}{\mathcal{U}_0(s)}$ der Σ-Δ-Modulatoren erster und zweiter Ordnung im Laplacebereich

$$\begin{aligned} H(s) &= \frac{1}{\alpha U_{ref}}, \\ H(s) &= \frac{1}{U_{ref}(\beta T_1 s + \alpha)}. \end{aligned}$$

Definition 8.5.1 *Ein lineares zeitinvariantes System (linear time invariant = LTI) heißt stabil, falls aus der Beschränktheit des Eingangssignals die des Ausgangssignals folgt.*

Satz 8.5.2 *Ein LTI-System ist genau dann stabil, wenn für die Übertragungfunktion* $H(s) = \frac{\mathcal{P}(s)}{\mathcal{Q}(s)}$ *gilt*

$$\begin{array}{ll} i) & deg\mathcal{P}(s) \leq deg\mathcal{Q}(s), \\ ii) & Re(s_{Pol}) < 0 \quad \forall Polstellen\ s_{Pol}\ von\ H(s). \end{array}$$

Folgerung 8.5.3 Die Modulatoren erster und zweiter Ordnung sind stabil.

Beweis: Die Übertragungsfunktion des Modulators erster Ordnung besitzt keine Polstellen, und die des Modulators zweiter Ordnung liegt in

$$s = -\frac{\alpha}{\beta T_1} < 0.$$

Daraus folgt unmittelbar die Behauptung. □

Bemerkung 8.5.4 Dieses Resultat deckt sich auch mit den Untersuchungen des Abschnitts 7.1

Nach [42] (dort ohne Normierungsfaktor!) lautet die Fouriertransformation und ihre Inverse

$$\begin{aligned} G(f) &= \int_{-\infty}^{\infty} g(t)\mathrm{e}^{-\mathrm{i}2\pi ft}\mathrm{d}t, \\ g(t) &= \int_{-\infty}^{\infty} G(f)\mathrm{e}^{\mathrm{i}2\pi tf}\mathrm{d}f. \end{aligned}$$

Und nach [45] besteht ein Zusammenhang zwischen zweiseitiger Laplace- und Fouriertransformation durch die Substitution

$$s = \mathrm{i}2\pi f,$$

wenn folgende zwei Bedingungen erfüllt sind:

i) das Signal ist absolut integrierbar,

ii) das LTI-System ist stabil.

Unter Annahme des Erfülltseins dieser Bedingungen lauten nun die Übertragungsfunktionen im Frequenzbereich

$$\begin{aligned} G(f) &= \frac{1}{\alpha U_{ref}}, \\ G(f) &= \frac{1}{U_{ref}(\mathrm{i}2\pi\beta T_1 f + \alpha)}. \end{aligned}$$

Wir wollen nun eine Charakterisierung der analogen Systeme im Frequenzbereich angeben.

Definition 8.5.5 *Wir betrachten die Spektraldarstellung*

$$G(f) = |G(f)| \exp\{\mathrm{i}\Psi(f)\}$$

der Fouriertransformierten der Übertragungsfunktion eines LTI-Systems. Dann bezeichnet $|G(f)|$ *die Betragscharakteristik und* $\Psi(f)$ *die Phasencharakteristik. Beide kennzeichnen das Übertragungsverhalten in Abhängigkeit zur Frequenz.* $|G(f)|$ *bezeichnet die maximale Auslenkung des*

Eingangssignals und $\Psi(f)$ die Phasenverschiebung des Ausgangs- zum Eingangssignal. Daraus abgeleitete Begriffe sind

$$\begin{array}{ll} \textit{Dämpfung:} & a(f)_{in\ dB} = -20 \lg |G(f)|, \\ \textit{Phasenlaufzeit:} & T_{ph} = -\frac{1}{2\pi f}\Psi(f), \\ \textit{Gruppenlaufzeit:} & T_{gr} = -\frac{1}{2\pi}\frac{\mathrm{d}\Psi}{\mathrm{d}f}. \end{array}$$

Die Größen Phasenlaufzeit und Gruppenlaufzeit beschreiben - verbal ausgedrückt - um welche Zeit ein Signal in einem Modulator verzögert wird. Weiteres dazu in [45]. Für die Modulatoren erster bis dritter Ordnung, als analoges System aufgefaßt, ergibt sich aus Definition 8.5.5 und den entsprechenden Differentialgleichungen leicht folgende Aussagen.

Folgerung 8.5.6 a) Für einen Σ-Δ-Modulator erster Ordnung gilt

$$\begin{aligned} |G(f)| &= \frac{1}{\alpha U_{ref}}, \\ \Psi(f) &= \frac{1}{\mathrm{i}} \ln \frac{G(f)}{|G(f)|}, \\ a(f)_{\text{in dB}} &= -20 \lg |G(f)| = const., \\ T_{ph} &= 0 = const., \\ T_{gr} &= 0 = const. \end{aligned}$$

Aufgrund der Tatsache, daß Dämpfung, Phasenlaufzeit und Gruppenlaufzeit frequenzunabhängig sind, ist ein Modulator erster Ordnung ein verzerrungsfreies System.
b) Für einen Σ-Δ-Modulator zweiter Ordnung gilt

$$\begin{aligned} |G(f)| &= \frac{1}{U_{ref}\sqrt{4\pi^2\beta^2T_1^2f^2+\alpha^2}}, \\ \Psi(f) &= \arctan\left(-\frac{2\pi\beta T_1}{\alpha}f\right), \\ T_{gr} &= \frac{\alpha\beta T_1}{\alpha^2+4\pi^2\beta^2T_1^2f^2}. \end{aligned}$$

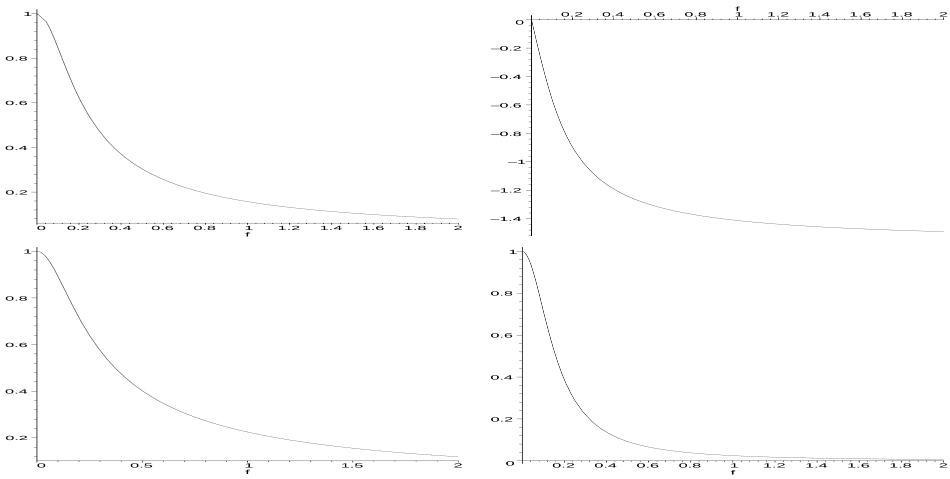


Abbildung 41: Betrags- und Phasencharakteristik (oben), Phasen- und Gruppenlaufzeit (unten) eines Modulators zweiter Ordnung

c) Für einen Σ-Δ-Modulator dritter Ordnung gilt

$$
\begin{aligned}
|G(f)| &= \frac{1}{U_{ref}\sqrt{(2\pi\beta T_1 f)^2 + (\alpha - 4\pi^2\gamma T_1 T_2 f^2)^2}}, \\
\Psi(f) &= \arctan\left(-\frac{2\pi\beta}{\alpha - 4\pi^2\gamma T_1 T_2 f^2}\right), \\
T_{gr} &= -\frac{\alpha + 4\gamma T_1 T_2 \pi^2 f^2}{\alpha^2 - 8\alpha\gamma T_1 T_2 \pi^2 f^2 + 16\gamma^2 T_1^2 T_2^2 \pi^4 f^4 + 4\pi^2 f^2}.
\end{aligned}
$$

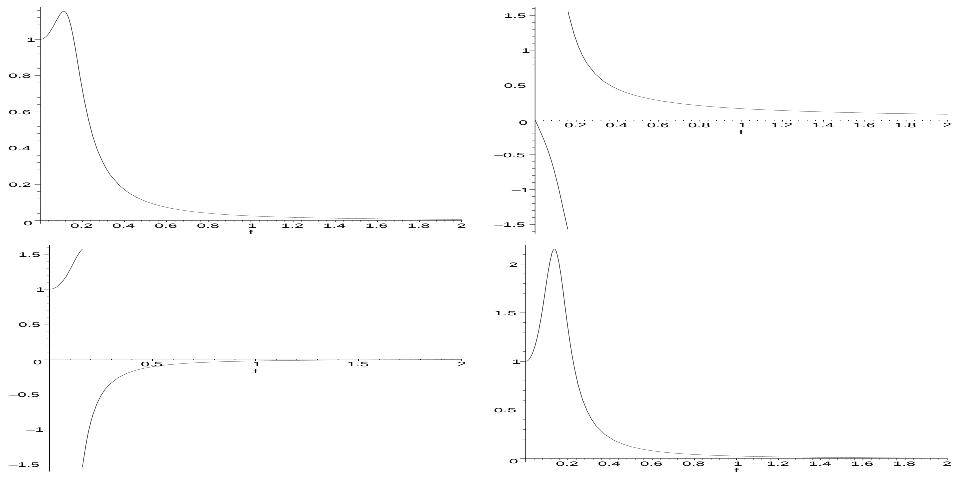


Abbildung 42: Betrags- und Phasencharakteristik (oben), Phasen- und Gruppenlaufzeit (unten) eines Modulators dritter Ordnung

Literatur

[1] Agarwal, R. P.: Difference Equations and Inequalities: Theory, Methods, and Applications, Dekker, New York, 1992.

[2] Agarwal, R. P.; Wong, P. J. Y.: Advanced Topics in Difference Equations, Kluwer Academic Publishers, Dordrecht, 1997.

[3] Barmettler, M.; Gruber, P.: Anwendung von Oversampling-Verfahren zur Erhöhung der Auflösung digital erfaßter Signale, Technisches Messen 59(1992)1,21-25;2,69-75; 6,262-268; 7/8,312-319.

[4] Candy, J. C.; Ching, Y. C.; Alexander, D. S.: Using Triangulary Weighted Interpolation to Get 13-Bit PCM from a Sigma-Delta Modulator, IEEE Transactions on Communications, vol. COM-24, pp.1268-1275, November 1976.

[5] Dunford, N.; Schwartz, J. T.: Linear Operators, Part I: General Theory, Wiley, New York, 1988.

[6] Elaydi, S. N.: An Introduction to Difference Equations, Springer, New York, 1999.

[7] Elaydi, S. N. et al. (edit.): Proceedings of the First International Conference of Differential Equations and Applications, Trinity University, San Antonio, Texas, May 25-28, 1994, Gordon and Breach Publishers, Amsterdam, 1995.

[8] Elaydi, S. N. et al. (edit.): Proceedings of the Second International Conference of Differential Equations and Applications, Veszprém, Hungary, August 7-11, 1995, Gordon and Breach Publishers, Amsterdam, 1997.

[9] Elaydi, S. N. et al. (edit.): Proceedings of the Third International Conference of Differential Equations and Applications, Taipei, Republic of China, September 1-5, 1997, Gordon and Breach Publishers, Amsterdam, 1999.

[10] Elaydi, S. N. et al. (edit.): Proceedings of the Fourth International Conference of Differential Equations and Applications, Poznan, Poland, August 27-31, 1998, Gordon and Breach Publishers, Amsterdam, 2000.

[11] Elstrodt, J.: Maß- und Integrationstheorie, Springer, Berlin, 1996.

[12] Farrell, R.; Feely, O.: Bounding the Integrator Outputs of Second-Order Sigma-Delta Modulators, IEEE Transactions on Circuits and Systems-II: Analog and digital signal processing, pp. 691-702, Vol.45, No.6, June 1998.

[13] Feely, O.: A tutorial Introduction to non-linear Dynamics and Chaos and their Application to Sigma-Delta Modulators, International Journal of Ciruit Theory and Applications, pp.347-367, Vol.25, 1997.

[14] Fenyő, S.; Stolle, H. W.: Theorie und Praxis der linearen Integralgleichungen, Band 2, Verlag der Wissenschaften, Berlin, 1983.

[15] Fisz, M.: Wahrscheinlichkeitsrechnung und mathematische Statistik (Übers. aus dem Polnischen), Verlag der Wissenschaften, Berlin, 1988.

[16] Galton, I.: Granular Quantization Noise in the First-Order Delta-Sigma Modulator, IEEE Transactions on Information Theory, pp.1944-1956, Vol.39, No.6, November 1993.

[17] Geering, H. P.: Regelungstechnik: Mathematische Grundlagen, Entwurfsmethoden, Beispiele, Springer, Berlin, 1994.

[18] Gnedenko, B. W.: Einführung in die Wahrscheinlichkeitstheorie, Akademie Verlag, Berlin, 1991.

[19] Goldberg, S.: Introduction to Difference Equations, Wiley, New York, 1958.

[20] Göldner, K.: Mathematische Grundlagen der Systemanalyse, Bde. 1,2, Fachbuchverlag, Leipzig, 1989.

[21] Goßlau, A.: Rückgekoppelte Analog-Digital-Umsetzer als nichtlineare zeitdiskrete Systeme - Analyse und Optimierung -, Dissertation, Universität der Bundeswehr München, 1989.

[22] Gray, R. M.: Oversampled Sigma-Delta modulation, IEEE Transactions on Communications, vol. COM-35, pp. 481-488, May 1987.

[23] Hackbusch, W.: Integralgleichungen: Theorie und Numerik, Teubner, Stuttgart, 1989.

[24] Hein, S.; Zakhor, A.: Stability and scaling of double loop sigma-delta modulators, Proceeding of the IEEE International Symposium on Circiuts ans Systems, pp.1312-1315, 1992.

[25] Hein, S.; Zakhor, A.: Sigma Delta Modulators: Nonlinear Decoding Algorithms and Stability Analysis, Kluwer Academic Publishers, Boston, 1993.

[26] Hess, W.: Digitale Filter, Teubner, Stuttgart, 1993.

[27] Hewitt, E.; Stromberg, K.: Real and Abstract Analysis. Springer, New York, 1965.

[28] Hirt, N.: Interfacetechnik, Vorlesungsskript, Technische Universität Ilmenau, 1996.

[29] Hoffmann, A.: On the covariance of a random sequence generated by the fractional part of multiples of a uniformly distributed random variable, pp.27-35, Stochastics and Stochastic Reports 666, 1999.

[30] Hoffmann, A.; Marx, B.: Mathematical Modelling of a First Order Sigma-Delta Modulator, Journal of Differerence Equations and Applications, pp. 533-550, Vol. 4, 1998.

[31] Hoffmann, A.; Marx, B.: Error Estimations of a Sigma-Delta Converter, Preprint No. M02/00, Technische Universität Ilmenau, 2000.

[32] Hoffmann, A.; Marx, B.: Mathematical modelling of a second order sigma-delta modulator, (in preparation).

[33] Hoffmann, A.; Marx, B.; Metzger, B.: Filtering of analog digital converters using an exact time approach, 43rd International Scientific Colloquium, Technical University of Ilmenau, Proceedings, Bd.1, pp. 491-496, 1997.

[34] Hoffmann, A.; Marx, B.; Metzger, B.: Mathematical Time Approach of a First Order Sigma-Delta-Modulator, Proceeding of the 5th Electronic Devices and Systems Conference, Brno, Czech Republik, 11.-12. Juni 1998, pp.129-132.

[35] Hoffmann, A.; Marx, B.; Metzger, B.: Mathematical Time Approach of First and Second Order Sigma-Delta-Modulator, Extended Abstracts of the 4th International Conference on Difference Equations and Applications, Poznan, Poland, 27.-31. August 1998.

[36] Hoffmann, A.; Marx, B.; Metzger, B.: Linear ODEs for describing Sigma-Delta-Modulators, Filter Design and Error Estimations, Proceedings of the 5th International Conference on Difference Equations and Applications, Temuco, Chile, 2.-7. Januar 2000, Gordon Breach Publishers, Amsterdam, (accepted 10/2000).

[37] Horbach, U.: Entwurf und Realisierung hochauflösender Sigma-Delta A/D und D/A Umsetzer, Dissertation, Erlangen, 1991.

[38] Hurwitz, A.: Über die Bedingungen, unter welchen eine Gleichung nur Wurzeln mit negativen reellen Teilen besitzt, Math. Ann. 46, 273-284, (1895)

[39] Kammeyer, K. D.; Kroschel, K.: Digitale Signalverarbeitung: Filterung und Spektralanalyse, Teubner, Stuttgart, 1996.

[40] Kocic, V. L.; Ladas, G.: Global Behavior of Nonlinear Difference Equations of Higher Order with Applications, Kluwer Academic Publishers, Dordrecht, 1993.

[41] Kolmogorov, A. N.; Fomin, S. V.: Reelle Funktionen und Funktionalanalysis, Verlag der Wissenschaften, Berlin, 1975.

[42] Kress, D.; Irmer, R.: Angewandte Systemtheorie: kontinuierliche und zeitdiskrete Signalverarbeitung, Verlag Technik, Berlin, 1989.

[43] Kronmüller, H.: Digitale Signalverarbeitung, Springer, 1991.

[44] Levy, H.; Lessman, F.: Finite Difference Equations, Macmillan, New York, 1961.

[45] Lüke, H. D.: Signalübertragung, Springer, Berlin, 1995.

[46] Medeiro, F.; Pérez-Verdú, A.; Rodríguez-Vázquez, A.: Top-Down Design oh High-Performance Sigma-Delta Modulators, Kluwer Academic Publishers, Boston, 1999.

[47] Megginson, R. E.: An Introduction to Banach Space Theory, Springer, New York, 1998.

[48] Metzger, B.; Neubert, Th.: Mathematical modelling of a second order sigma-delta modulator by difference equations, Tagungsband des Workshops der Informatik-Graduiertenkollegs im Rahmen der Jahrestagung der Gesellschaft für Informatik an der RWTH Aachen vom 22.-23. Sept. 1997, pp.169-181 (english).

[49] Mickens, R. E.: Difference Equations, Theory and Applications, Van Nostred Reinhold, New York, 1990.

[50] Natanson, I. P.: Theorie der Funktion einer reellen Veränderlichen, Harri Deutsch, Thun, Frankfurt/Main, 1981.

[51] Neubert, T.: A/D-Wandlungsverfahren: gegenwärtiger Stand, aktuelle Entwicklungen und Trends, Technische Universität Ilmenau, 1999.

[52] Norsworthy, S. T.; Schreier, R.; Temes, G. C. (Edit.): Delta-Sigma Data Converters, IEEE Press, Piscataway, 1997.

[53] Orfanidis, S. J.: Introduction to Signal Processing, Prentice Hall, London, 1996.

[54] Parks, T. W.; Burrus, C. S.: Digital Filter Design, Wiley, New York, 1987.

[55] Pinault, St. C.; Lopresti, Ph. V.: On the Behavior of the Double-Loop Sigma-Delta Modulator, IEEE Transactions on Circuits and Systems-II: Analog and digital signal processing, pp.467-479, Vol.40, No.8, August 1993.

[56] Potchinkov, A.: Der Entwurf digitaler FIR-Filter mit Methoden der konvexen semi-infiniten Optimierung, Dissertation, Technische Universität Berlin, 1994.

[57] Qiu, L.; Davison, E. J.: A simple procedure for the exact stability robustness computation of polynomials with affine coefficient pertubations, Systems and Control Letters 13, pp. 413-420, 1989.

[58] Schilder, F.: Time-domain simulation of $\Sigma\Delta$-analog-digital converter with Ptolemy, 43rd International Scientific Colloquium, Technical University of Ilmenau, Proceedings, Bd.2, pp. 60-64, 1997.

[59] Schreier, R.: Noise shaped coding, PhD Thesis, University of Toronto, 1991.

[60] Schulte, F.: Untersuchungen des Verhaltens von AD-Wandlern nach dem Σ-Δ-Prinzip mit einem oder zwei Integratoren und deren Simulation mit MATHEMATICA, Diplomarbeit, Technische Universität Ilmenau, 1997.

[61] Seifart, M.: Analoge Schaltungen, Verlag Technik, Berlin, 1987.

[62] Sobol, J. M.: Chislenye metody Monte Carlo, Izd. nauka, Moskva, 1973.

[63] Tietze, U.; Schenk, Ch.: Halbleiter-Schaltungstechnik, Springer, Berlin, 1988.

[64] van Engelen, J.; van de Plassche, R.: Stability Criteria for Design of Higher Order Sigma Delta Modulators, (aus formalen Gründen unveröffentlichter Artikel) May 1998.

[65] Walter, W.: Gewöhnliche Differentialgleichungen, Springer, Berlin, 1996.

[66] Wang, H.: A geometric view of $\Sigma\Delta$ modulations, IEEE Transactions on Circuits and Systems-II: Analog and digital signal processing, pp.402-405, Vol.39, June 1992.

[67] Weinmann, A.: Uncertain Models and Robust Control, Springer, Wien, 1991

[68] Wilms, R.: Fractional parts of random variables: Limits theorems and infinite divisibility, PhD Thesis, Technical University of Eindhoven, 1990.

[69] Wirth, F. R.: Robust Stability of Discrete-Time Systems under Time-Varying Perturbations, Dissertation, Universität Bremen, 1995.

[70] Yosida, K.: Functional Analysis, Springer, Berlin, 1995.

Hiermit versichere ich, daß ich die vorliegende Arbeit selbst angefertigt habe und alle benutzten Hilfsmittel bzw. Quellen darin angegeben sind, daß ich weder diese Arbeit noch Teile daraus für eine staatliche oder andere wissenschaftliche Prüfung eingereicht habe und daß sie weder bei einer anderen Fakultät noch Universität als Dissertation eingereicht wurde.

Jena, am 5. Juni 2001

Bob Metzger