Inexakte Krylov-Raum-Verfahren

Jens-Peter M. Zemke
zemke@tu-harburg.de

Institut für Numerische Simulation
Technische Universität Hamburg-Harburg

24.06.2010

TUHH
Technische Universität Hamburg-Harburg

Übersicht

Übersicht

Übersicht

Übersicht

Motivation

Häufig ist die Matrix-Vektor-Multiplikation eine der teuersten Operationen in einem Krylov-Raum-Verfahren und man versucht billigere Approximationen zu verwenden, ohne aber die Konvergenzeigenschaften negativ zu beeinflussen.

Motivation

Häufig ist die Matrix-Vektor-Multiplikation eine der teuersten Operationen in einem Krylov-Raum-Verfahren und man versucht billigere Approximationen zu verwenden, ohne aber die Konvergenzeigenschaften negativ zu beeinflussen.

Ein Beispiel hierfür ist durch sogenannte „inner–outer“-Iterationen gegeben, siehe [18, 15, 19, 20], in denen die Matrix A implizit präkonditioniert wird, heutzutage oft wieder durch ein Krylov-Raum-Verfahren. Diese Verfahren sind auch unter dem Schlagwort „zwei-stufiges Iterationsverfahren“ zu finden, siehe [22, 26, 24, 25, 14].

Motivation

Häufig ist die Matrix-Vektor-Multiplikation eine der teuersten Operationen in einem Krylov-Raum-Verfahren und man versucht billigere Approximationen zu verwenden, ohne aber die Konvergenzeigenschaften negativ zu beeinflussen.

Ein Beispiel hierfür ist durch sogenannte „inner–outer“-Iterationen gegeben, siehe [18, 15, 19, 20], in denen die Matrix A implizit präkonditioniert wird, heutzutage oft wieder durch ein Krylov-Raum-Verfahren. Diese Verfahren sind auch unter dem Schlagwort „zwei-stufiges Iterationsverfahren“ zu finden, siehe [22, 26, 24, 25, 14].

Ein anderes Beispiel ist es, wenn die Matrix A selber gar nicht gegeben ist, sondern das Produkt von A mit einem Vektor jedesmal teuer aus Messungen oder als Funktion einer Matrix berechnet werden muss.

Motivation

Häufig ist die Matrix-Vektor-Multiplikation eine der teuersten Operationen in einem Krylov-Raum-Verfahren und man versucht billigere Approximationen zu verwenden, ohne aber die Konvergenzeigenschaften negativ zu beeinflussen.

Ein Beispiel hierfür ist durch sogenannte „inner–outer“-Iterationen gegeben, siehe [18, 15, 19, 20], in denen die Matrix A implizit präkonditioniert wird, heutzutage oft wieder durch ein Krylov-Raum-Verfahren. Diese Verfahren sind auch unter dem Schlagwort „zwei-stufiges Iterationsverfahren“ zu finden, siehe [22, 26, 24, 25, 14].

Ein anderes Beispiel ist es, wenn die Matrix A selber gar nicht gegeben ist, sondern das Produkt von A mit einem Vektor jedesmal teuer aus Messungen oder als Funktion einer Matrix berechnet werden muss.

Der Fall der Funktion einer Matrix ist z. B. beim sog. Dirac Overlap-Operator in der Quantenchromodynamik (QCD) gegeben, siehe [32, 1, 10, 9].

Motivation

Die Auswirkungen einer Approximation der Matrix-Vektor-Multiplikation (MVM) lassen sich mathematisch mit der Ersetzung der exakten MVM

$$w = Aq \tag{1}$$

Motivation

Die Auswirkungen einer Approximation der Matrix-Vektor-Multiplikation (MVM) lassen sich mathematisch mit der Ersetzung der exakten MVM

$$w = Aq \tag{1}$$

durch eine inexakte MVM

$$\tilde{w} = (A + \Delta A)q = Aq + g \tag{2}$$

beschreiben,

Motivation

Die Auswirkungen einer Approximation der Matrix-Vektor-Multiplikation (MVM) lassen sich mathematisch mit der Ersetzung der exakten MVM

$$w = Aq \tag{1}$$

durch eine inexakte MVM

$$\tilde{w} = (A + \Delta A)q = Aq + g \tag{2}$$

beschreiben, wobei $g = \Delta A q$ die Abweichung beschreibt.

Motivation

Die Auswirkungen einer Approximation der Matrix-Vektor-Multiplikation (MVM) lassen sich mathematisch mit der Ersetzung der exakten MVM

$$w = Aq \tag{1}$$

durch eine inexakte MVM

$$\tilde{w} = (A + \Delta A)q = Aq + g \tag{2}$$

beschreiben, wobei $g = \Delta A q$ die Abweichung beschreibt.

Die naheliegende Frage ist die nach der maximal erlaubten Norm der Störung ΔA (oder äquivalent dazu, g), welche die Konvergenzeigenschaften nicht oder nur marginal beeinflusst.

Motivation

Die Auswirkungen einer Approximation der Matrix-Vektor-Multiplikation (MVM) lassen sich mathematisch mit der Ersetzung der exakten MVM

$$w = Aq \tag{1}$$

durch eine inexakte MVM

$$\tilde{w} = (A + \Delta A)q = Aq + g \tag{2}$$

beschreiben, wobei $g = \Delta A q$ die Abweichung beschreibt.

Die naheliegende Frage ist die nach der maximal erlaubten Norm der Störung ΔA (oder äquivalent dazu, g), welche die Konvergenzeigenschaften nicht oder nur marginal beeinflusst.

Eine Verfeinerung dieser Frage ist die, ob die Störung in der Norm konstant bleiben oder im Verlaufe der Iteration angepasst werden kann/sollte.

Historisches

Da die Matrix-Vektor-Multiplikation nicht exakt ausgeführt wird, werden diese Verfahren unter dem Schlagwort „inexakte Verfahren“ zusammengefaßt, in unserem Falle handelt es sich also um inexakte Krylov-Raum-Verfahren.

Historisches

Da die Matrix-Vektor-Multiplikation nicht exakt ausgeführt wird, werden diese Verfahren unter dem Schlagwort „inexakte Verfahren“ zusammengefaßt, in unserem Falle handelt es sich also um inexakte Krylov-Raum-Verfahren.

Untersuchungen zu inexakten Verfahren gab es bereits vorher im Bereich der inexakten Newton-Verfahren [11, 12, 23, 8] und im Bereich der inexakten klassischen Iterationsverfahren [16, 17].

Historisches

Da die Matrix-Vektor-Multiplikation nicht exakt ausgeführt wird, werden diese Verfahren unter dem Schlagwort „inexakte Verfahren“ zusammengefaßt, in unserem Falle handelt es sich also um inexakte Krylov-Raum-Verfahren.

Untersuchungen zu inexakten Verfahren gab es bereits vorher im Bereich der inexakten Newton-Verfahren [11, 12, 23, 8] und im Bereich der inexakten klassischen Iterationsverfahren [16, 17].

Im Newton-Verfahren muss in jedem Schritt ein Gleichungssystem gelöst werden. Da diese Gleichungssysteme sehr groß werden können, liegt es nahe, die Lösung nur approximativ zu berechnen.

Historisches

Da die Matrix-Vektor-Multiplikation nicht exakt ausgeführt wird, werden diese Verfahren unter dem Schlagwort „inexakte Verfahren“ zusammengefaßt, in unserem Falle handelt es sich also um inexakte Krylov-Raum-Verfahren.

Untersuchungen zu inexakten Verfahren gab es bereits vorher im Bereich der inexakten Newton-Verfahren [11, 12, 23, 8] und im Bereich der inexakten klassischen Iterationsverfahren [16, 17].

Im Newton-Verfahren muss in jedem Schritt ein Gleichungssystem gelöst werden. Da diese Gleichungssysteme sehr groß werden können, liegt es nahe, die Lösung nur approximativ zu berechnen.

Im Falle der inexakten Newton-Verfahren muss die Genauigkeit beim Lösen des Gleichungssystemes nahe bei der gesuchten Lösung (Nullstelle, stationärer Punkt) erhöht werden, große Anfangsfehler stören eventuell lokal die Konvergenz und verändern lokal die Konvergenzgeschwindigkeit, verändern aber nicht die erreichbare Genauigkeit und stören nicht das lokal-quadratische Konvergenzverhalten.

Historisches

Bei den inexakten klassischen Iterationsverfahren wird, genau wie im Fall der Krylov-Raum-Verfahren, die Matrix-Vektor-Multiplikation ungenau ausgeführt; der Grund liegt auch hier in der nur approximativen Lösung der durch Präkonditionierung entstandenen Gleichungssysteme (Splitting).

Historisches

Bei den inexakten klassischen Iterationsverfahren wird, genau wie im Fall der Krylov-Raum-Verfahren, die Matrix-Vektor-Multiplikation ungenau ausgeführt; der Grund liegt auch hier in der nur approximativen Lösung der durch Präkonditionierung entstandenen Gleichungssysteme (Splitting).

Im Bereich der inexakten klassischen Iterationsverfahren sieht die Sache ähnlich aus wie bei den inexakten Newton-Verfahren; allerdings konvergieren diese Verfahren durchweg linear, also sollte im Allgemeinen mit Hinblick auf eine Minimierung des Gesamtaufwandes die Auswertungsgenauigkeit zumindest konstant gehalten werden.

Historisches

Bei den inexakten klassischen Iterationsverfahren wird, genau wie im Fall der Krylov-Raum-Verfahren, die Matrix-Vektor-Multiplikation ungenau ausgeführt; der Grund liegt auch hier in der nur approximativen Lösung der durch Präkonditionierung entstandenen Gleichungssysteme (Splitting).

Im Bereich der inexakten klassischen Iterationsverfahren sieht die Sache ähnlich aus wie bei den inexakten Newton-Verfahren; allerdings konvergieren diese Verfahren durchweg linear, also sollte im Allgemeinen mit Hinblick auf eine Minimierung des Gesamtaufwandes die Auswertungsgenauigkeit zumindest konstant gehalten werden.

Erste theoretische Untersuchungen zum Richardson-Verfahren zweiter Ordnung und der Chebyshev-Beschleunigung wurden 1982 ([16], nur die Richardson-Iteration) und 1988 ([17], für beide Varianten) von Golub und Overton durchgeführt. Es wurden auch bereits Tests an einem inexaktem CG-Verfahren ausgeführt, allerdings mit fester Auswertungsgenauigkeit.

Historisches; inexakte Krylov-Raum-Verfahren

Bouras und Frayssé (und später Giraud) untersuchten im Jahre 2000 die Konvergenz verschiedener inexakter Krylov-Raum-Verfahren [4, 5, 7], siehe auch [3] und bald stellte sich heraus, dass diese aus dem Rahmen fallen.

Historisches; inexakte Krylov-Raum-Verfahren

Bouras und Frayssé (und später Giraud) untersuchten im Jahre 2000 die Konvergenz verschiedener inexakter Krylov-Raum-Verfahren [4, 5, 7], siehe auch [3] und bald stellte sich heraus, dass diese aus dem Rahmen fallen.

In den Krylov-Raum-Verfahren muss anfangs äußerst akkurat gerechnet werden, und in der Phase der Konvergenz kann die Genauigkeitsanforderung relaxiert werden, bis eigentlich nur noch mit Fehlern gerechnet wird.

Historisches; inexakte Krylov-Raum-Verfahren

Bouras und Frayssé (und später Giraud) untersuchten im Jahre 2000 die Konvergenz verschiedener inexakter Krylov-Raum-Verfahren [4, 5, 7], siehe auch [3] und bald stellte sich heraus, dass diese aus dem Rahmen fallen.

In den Krylov-Raum-Verfahren muss anfangs äußerst akkurat gerechnet werden, und in der Phase der Konvergenz kann die Genauigkeitsanforderung relaxiert werden, bis eigentlich nur noch mit Fehlern gerechnet wird.

In vielen Fällen konvergiert das inexakte Verfahren dennoch bis auf die gewünschte Genauigkeit gegen die Lösung.

Historisches; inexakte Krylov-Raum-Verfahren

Bouras und Frayssé (und später Giraud) untersuchten im Jahre 2000 die Konvergenz verschiedener inexakter Krylov-Raum-Verfahren [4, 5, 7], siehe auch [3] und bald stellte sich heraus, dass diese aus dem Rahmen fallen.

In den Krylov-Raum-Verfahren muss anfangs äußerst akkurat gerechnet werden, und in der Phase der Konvergenz kann die Genauigkeitsanforderung relaxiert werden, bis eigentlich nur noch mit Fehlern gerechnet wird.

In vielen Fällen konvergiert das inexakte Verfahren dennoch bis auf die gewünschte Genauigkeit gegen die Lösung.

Der „erste“ in einem Journal erschienene Artikel [6] von Bouras und Frayssé zu dem Thema der inexakten Krylov-Raum-Verfahren benötigte allerdings so lange bis zum Erscheinen, dass die beiden ersten theoretischen Arbeiten zu inexakten Krylov-Raum-Verfahren von van den Eshof & Sleijpen [34] und Szyld & Simoncini [30] vor diesem erschienen.

Historisches; inexakte Krylov-Raum-Verfahren

Es gibt zwei Arbeiten, beide im Bereich inexakter Verfahren zur Approximation von Eigenpaaren, die gewissermaßen als Vorarbeiten zu den experimentellen Untersuchungen von Bouras und Frayssé verstanden werden können.

Historisches; inexakte Krylov-Raum-Verfahren

Es gibt zwei Arbeiten, beide im Bereich inexakter Verfahren zur Approximation von Eigenpaaren, die gewissermaßen als Vorarbeiten zu den experimentellen Untersuchungen von Bouras und Frayssé verstanden werden können.

Eine erste Anmerkung zu dem erwarteten Verhalten im Falles des inexakten Lanczos-Verfahrens zur Berechnung eines Eigenwertes findet sich in einem eher philosophisch gehaltenen Artikel von Golub, Zhang und Zha von 2000.

Historisches; inexakte Krylov-Raum-Verfahren

Es gibt zwei Arbeiten, beide im Bereich inexakter Verfahren zur Approximation von Eigenpaaren, die gewissermaßen als Vorarbeiten zu den experimentellen Untersuchungen von Bouras und Frayssé verstanden werden können.

Eine erste Anmerkung zu dem erwarteten Verhalten im Falles des inexakten Lanczos-Verfahrens zur Berechnung eines Eigenwertes findet sich in einem eher philosophisch gehaltenen Artikel von Golub, Zhang und Zha von 2000. Dort wird auf Basis des exakten symmetrischen Lanczos-Verfahrens untersucht, für welche Störungen sich der gewünschte Eigenwert der konstruierten Tridiagonalmatrix robust verhält.

Historisches; inexakte Krylov-Raum-Verfahren

Es gibt zwei Arbeiten, beide im Bereich inexakter Verfahren zur Approximation von Eigenpaaren, die gewissermaßen als Vorarbeiten zu den experimentellen Untersuchungen von Bouras und Frayssé verstanden werden können.

Eine erste Anmerkung zu dem erwarteten Verhalten im Falles des inexakten Lanczos-Verfahrens zur Berechnung eines Eigenwertes findet sich in einem eher philosophisch gehaltenen Artikel von Golub, Zhang und Zha von 2000. Dort wird auf Basis des exakten symmetrischen Lanczos-Verfahrens untersucht, für welche Störungen sich der gewünschte Eigenwert der konstruierten Tridiagonalmatrix robust verhält. Quintessenz: bei schneller Konvergenz der exakten Version kann gegen Ende stärker gestört werden.

Historisches; inexakte Krylov-Raum-Verfahren

Es gibt zwei Arbeiten, beide im Bereich inexakter Verfahren zur Approximation von Eigenpaaren, die gewissermaßen als Vorarbeiten zu den experimentellen Untersuchungen von Bouras und Frayssé verstanden werden können.

Eine erste Anmerkung zu dem erwarteten Verhalten im Falles des inexakten Lanczos-Verfahrens zur Berechnung eines Eigenwertes findet sich in einem eher philosophisch gehaltenen Artikel von Golub, Zhang und Zha von 2000. Dort wird auf Basis des exakten symmetrischen Lanczos-Verfahrens untersucht, für welche Störungen sich der gewünschte Eigenwert der konstruierten Tridiagonalmatrix robust verhält. Quintessenz: bei schneller Konvergenz der exakten Version kann gegen Ende stärker gestört werden.

Ungefähr zeitgleich erschien eine Arbeit von Smit & Paardekooper, in der die inexakte inverse Iteration und eine inexakte Rayleigh-Quotienten-Iteration (inexakte RQI) für reelles symmetrisches A verglichen wurden.

Historisches; inexakte Krylov-Raum-Verfahren

Es gibt zwei Arbeiten, beide im Bereich inexakter Verfahren zur Approximation von Eigenpaaren, die gewissermaßen als Vorarbeiten zu den experimentellen Untersuchungen von Bouras und Frayssé verstanden werden können.

Eine erste Anmerkung zu dem erwarteten Verhalten im Falles des inexakten Lanczos-Verfahrens zur Berechnung eines Eigenwertes findet sich in einem eher philosophisch gehaltenen Artikel von Golub, Zhang und Zha von 2000. Dort wird auf Basis des exakten symmetrischen Lanczos-Verfahrens untersucht, für welche Störungen sich der gewünschte Eigenwert der konstruierten Tridiagonalmatrix robust verhält. Quintessenz: bei schneller Konvergenz der exakten Version kann gegen Ende stärker gestört werden.

Ungefähr zeitgleich erschien eine Arbeit von Smit & Paardekooper, in der die inexakte inverse Iteration und eine inexakte Rayleigh-Quotienten-Iteration (inexakte RQI) für reelles symmetrisches A verglichen wurden. Die Quintessenz dieser Arbeit ist, dass die Genauigkeit bei der inexakten inversen Iteration erhöht und bei inexakter RQI die Genauigkeit konstant gewählt werden sollte, wobei Genauigkeit die Größe des jeweiligen Residuums meint.

Übersicht

Problemstellung

In den genannten Fällen hat man Kontrolle über die Größen der auftretenden Fehler, also die Normen der Matrizen ΔA_k, oder, vergleichbar, die Normen der Vektoren $g_k = \Delta A_k q_k$ im Produkt $\tilde{w}_k = (A + \Delta A_k)q_k = Aq_k + g_k$.

Problemstellung

In den genannten Fällen hat man Kontrolle über die Größen der auftretenden Fehler, also die Normen der Matrizen ΔA_k, oder, vergleichbar, die Normen der Vektoren $g_k = \Delta A_k q_k$ im Produkt $\tilde{w}_k = (A + \Delta A_k)q_k = Aq_k + g_k$.

Da ein kleinerer Fehler einen größeren Arbeitsaufwand bedeutet, liegt die Frage nahe, wieviel Ungenauigkeit man zu welcher Zeit erlauben kann, um immer noch ein schnell gegen eine approximative Lösung $\tilde{x}$ konvergentes Verfahren zu bekommen.

Problemstellung

In den genannten Fällen hat man Kontrolle über die Größen der auftretenden Fehler, also die Normen der Matrizen ΔA_k, oder, vergleichbar, die Normen der Vektoren $g_k = \Delta A_k q_k$ im Produkt $\tilde{w}_k = (A + \Delta A_k)q_k = Aq_k + g_k$.

Da ein kleinerer Fehler einen größeren Arbeitsaufwand bedeutet, liegt die Frage nahe, wieviel Ungenauigkeit man zu welcher Zeit erlauben kann, um immer noch ein schnell gegen eine approximative Lösung $\tilde{x}$ konvergentes Verfahren zu bekommen.

Die auf Beispielen aufbauenden Arbeiten von Bouras und Frayssé benutzen Termini wie den normweisen relativen Rückwärtsfehler, und arbeiten deshalb in natürlicher Weise mit den Normen der Fehlermatrizen ΔA_k. Auch in der Arbeit [30] von Simoncini und Szyld werden diese Fehlermatrizen verwendet.

Problemstellung

In den genannten Fällen hat man Kontrolle über die Größen der auftretenden Fehler, also die Normen der Matrizen ΔA_k, oder, vergleichbar, die Normen der Vektoren $g_k = \Delta A_k q_k$ im Produkt $\tilde{w}_k = (A + \Delta A_k)q_k = Aq_k + g_k$.

Da ein kleinerer Fehler einen größeren Arbeitsaufwand bedeutet, liegt die Frage nahe, wieviel Ungenauigkeit man zu welcher Zeit erlauben kann, um immer noch ein schnell gegen eine approximative Lösung $\tilde{x}$ konvergentes Verfahren zu bekommen.

Die auf Beispielen aufbauenden Arbeiten von Bouras und Frayssé benutzen Termini wie den normweisen relativen Rückwärtsfehler, und arbeiten deshalb in natürlicher Weise mit den Normen der Fehlermatrizen ΔA_k. Auch in der Arbeit [30] von Simoncini und Szyld werden diese Fehlermatrizen verwendet.

Die auf Theorie gründende Arbeit [34] von van den Eshof und Sleijpen verwendet hingegen die Normen der Vektoren g_k.

Problemstellung

In den genannten Fällen hat man Kontrolle über die Größen der auftretenden Fehler, also die Normen der Matrizen ΔA_k, oder, vergleichbar, die Normen der Vektoren $g_k = \Delta A_k q_k$ im Produkt $\tilde{w}_k = (A + \Delta A_k)q_k = Aq_k + g_k$.

Da ein kleinerer Fehler einen größeren Arbeitsaufwand bedeutet, liegt die Frage nahe, wieviel Ungenauigkeit man zu welcher Zeit erlauben kann, um immer noch ein schnell gegen eine approximative Lösung $\tilde{x}$ konvergentes Verfahren zu bekommen.

Die auf Beispielen aufbauenden Arbeiten von Bouras und Frayssé benutzen Termini wie den normweisen relativen Rückwärtsfehler, und arbeiten deshalb in natürlicher Weise mit den Normen der Fehlermatrizen ΔA_k. Auch in der Arbeit [30] von Simoncini und Szyld werden diese Fehlermatrizen verwendet.

Die auf Theorie gründende Arbeit [34] von van den Eshof und Sleijpen verwendet hingegen die Normen der Vektoren g_k. Wir gehen im Folgenden zuerst auf die Arbeit von Bouras und Frayssé, danach auf die Arbeit von van den Eshof und Sleijpen ein.

Problemstellung; Bouras & Frayssé; Hintergrund

Der normweise relative Rückwärtsfehler einer approximativen Lösung $\tilde{x}$ eines linearen Gleichungssystemes $Ax = r_0$ ist definiert als die kleinste Zahl η, so dass

$$(A + \Delta A)\tilde{x} = r_0, \quad \|\Delta A\| \leqslant \eta \|A\| \tag{3}$$

gilt.

Problemstellung; Bouras & Frayssé; Hintergrund

Der normweise relative Rückwärtsfehler einer approximativen Lösung $\tilde{x}$ eines linearen Gleichungssystemes $Ax = r_0$ ist definiert als die kleinste Zahl η, so dass

$$(A + \Delta A)\tilde{x} = r_0, \quad \|\Delta A\| \leqslant \eta \|A\| \tag{3}$$

gilt.

Der (normweise relative) Rückwärtsfehler gibt also an, um wieviel die Matrix A (in der Norm, relativ) gestört werden muss, damit die berechnete Lösung $\tilde{x}$ die tatsächliche Lösung eines „in der Nähe liegenden“ Gleichungssystemes ist.

Problemstellung; Bouras & Frayssé; Hintergrund

Der normweise relative Rückwärtsfehler einer approximativen Lösung $\tilde{x}$ eines linearen Gleichungssystemes $Ax = r_0$ ist definiert als die kleinste Zahl η, so dass

$$(A + \Delta A)\tilde{x} = r_0, \quad \|\Delta A\| \leqslant \eta \|A\| \tag{3}$$

gilt.

Der (normweise relative) Rückwärtsfehler gibt also an, um wieviel die Matrix A (in der Norm, relativ) gestört werden muss, damit die berechnete Lösung $\tilde{x}$ die tatsächliche Lösung eines „in der Nähe liegenden“ Gleichungssystemes ist.

Theorem (Rigal und Gaches; 1967)

Der normweise relative Rückwärtsfehler η aus (3) *läßt sich berechnen durch*

$$\eta = \frac{\|r_0 - A\tilde{x}\|}{\|A\|\|\tilde{x}\|}. \tag{4}$$

Problemstellung

Bouras und Frayssé streben an, eine approximative Lösung $\tilde{x}$ mit einem Rückwärtsfehler der gegebenen Größe $\eta \geqslant 10^{-16}$ zu erzielen.

Problemstellung

Bouras und Frayssé streben an, eine approximative Lösung $\tilde{x}$ mit einem Rückwärtsfehler der gegebenen Größe $\eta \geqslant 10^{-16}$ zu erzielen.

Bei gegebener Größe η des Rückwärtsfehlers suchen wir eine Folge $\{\epsilon_k\}_{k=1}^{\infty}$ von Schranken; ϵ_k ist die Schranke für den erlaubten relativen Fehler in der Matrix-Vektor-Multiplikation im $k$ten Schritt,

$$\|\Delta A_k\| \leqslant \epsilon_k \|A\|. \tag{5}$$

Problemstellung

Bouras und Frayssé streben an, eine approximative Lösung $\tilde{x}$ mit einem Rückwärtsfehler der gegebenen Größe $\eta \geqslant 10^{-16}$ zu erzielen.

Bei gegebener Größe η des Rückwärtsfehlers suchen wir eine Folge $\{\epsilon_k\}_{k=1}^{\infty}$ von Schranken; ϵ_k ist die Schranke für den erlaubten relativen Fehler in der Matrix-Vektor-Multiplikation im $k$ten Schritt,

$$\|\Delta A_k\| \leqslant \epsilon_k \|A\|. \tag{5}$$

Bouras und Frayssé haben mit der aus experimentiellen Studien gewonnenen Wahl

$$\epsilon_k = \min\left(\frac{\eta}{\min(\|r_{k-1}\|, 1)}, 1\right) \tag{6}$$

gearbeitet, wobei $r_{k-1} = r_0 - Ax_{k-1}$ das tatsächliche Residuum bezeichnet.

Problemstellung

Bouras und Frayssé streben an, eine approximative Lösung $\tilde{x}$ mit einem Rückwärtsfehler der gegebenen Größe $\eta \geqslant 10^{-16}$ zu erzielen.

Bei gegebener Größe η des Rückwärtsfehlers suchen wir eine Folge $\{\epsilon_k\}_{k=1}^{\infty}$ von Schranken; ϵ_k ist die Schranke für den erlaubten relativen Fehler in der Matrix-Vektor-Multiplikation im $k$ten Schritt,

$$\|\Delta A_k\| \leqslant \epsilon_k \|A\|. \tag{5}$$

Bouras und Frayssé haben mit der aus experimentiellen Studien gewonnenen Wahl

$$\epsilon_k = \min\left(\frac{\eta}{\min(\|r_{k-1}\|, 1)}, 1\right) \tag{6}$$

gearbeitet, wobei $r_{k-1} = r_0 - Ax_{k-1}$ das tatsächliche Residuum bezeichnet. Dieses ist selbstverständlich normalerweise nicht verfügbar.

Übersicht

GMRes

Wir betrachten die Ergebnisse von Bouras und Frayssé für GMRes. Um unsere im Folgenden verwendete Notation einzuführen, geben wir eine nochmalige kurze Herleitung des GMRes-Verfahrens von Saad und Schultz.

GMRes

Wir betrachten die Ergebnisse von Bouras und Frayssé für GMRes. Um unsere im Folgenden verwendete Notation einzuführen, geben wir eine nochmalige kurze Herleitung des GMRes-Verfahrens von Saad und Schultz.

GMRes basiert auf der Methode von Arnoldi zur Bestimmung einer Orthonormalbasis eines Krylovraumes, welche in der Krylov-Zerlegung

$$AQ_k = Q_{k+1}\underline{H}_k \tag{7}$$

resultiert, wobei $A \in \mathbb{C}^{(n,n)}$ eine gegebene Matrix eines Gleichungssystemes $Ax = r_0$ ist und die Methode von Arnoldi mit dem Startvektor $q_1 = r_0/\|r_0\|$ begonnen wurde.

GMRes

Wir betrachten die Ergebnisse von Bouras und Frayssé für GMRes. Um unsere im Folgenden verwendete Notation einzuführen, geben wir eine nochmalige kurze Herleitung des GMRes-Verfahrens von Saad und Schultz.

GMRes basiert auf der Methode von Arnoldi zur Bestimmung einer Orthonormalbasis eines Krylovraumes, welche in der Krylov-Zerlegung

$$AQ_k = Q_{k+1}\underline{H}_k \tag{7}$$

resultiert, wobei $A \in \mathbb{C}^{(n,n)}$ eine gegebene Matrix eines Gleichungssystemes $Ax = r_0$ ist und die Methode von Arnoldi mit dem Startvektor $q_1 = r_0/\|r_0\|$ begonnen wurde.

Nach Konstruktion ist $Q_{k+1} \in \mathbb{C}^{(n,k+1)}$ orthonormal, die Spalten sind gerade die orthonormalen Basisvektoren des Krylovraumes $\mathcal{K}_{k+1}$. Die rechteckige Matrix $\underline{H}_k \in \mathbb{C}^{(k+1,k)}$ ist eine um eine zusätzliche, unten angefügte Zeile erweiterte unreduzierte Hessenbergmatrix.

GMRes

Die Spalten der Matrix Q_k sind die Basisvektoren des $k$ten Krylovraumes $\mathcal{K}_k$, also läßt sich jeder Vektor $x \in \mathcal{K}_k$ mittels gewisser Koeffizienten $z \in \mathbb{C}^k$ durch die Gleichung $x = Q_k z$ eindeutig parametrisieren.

GMRes

Die Spalten der Matrix Q_k sind die Basisvektoren des $k$ten Krylovraumes $\mathcal{K}_k$, also läßt sich jeder Vektor $x \in \mathcal{K}_k$ mittels gewisser Koeffizienten $z \in \mathbb{C}^k$ durch die Gleichung $x = Q_k z$ eindeutig parametrisieren.

GMRes ist dadurch definiert, dass im $k$ten Schritt diejenige approximative Lösung des Gleichungssystemes mit minimalem Residuum berechnet wird. Diese Approximation aus dem $k$ten Krylovraum wird im Folgenden mit $\underline{x}_k$, der zugehörige Parametervektor mit $\underline{z}_k$ bezeichnet.

GMRes

Die Spalten der Matrix Q_k sind die Basisvektoren des $k$ten Krylovraumes $\mathcal{K}_k$, also läßt sich jeder Vektor $x \in \mathcal{K}_k$ mittels gewisser Koeffizienten $z \in \mathbb{C}^k$ durch die Gleichung $x = Q_k z$ eindeutig parametrisieren.

GMRes ist dadurch definiert, dass im $k$ten Schritt diejenige approximative Lösung des Gleichungssystemes mit minimalem Residuum berechnet wird. Diese Approximation aus dem $k$ten Krylovraum wird im Folgenden mit $\underline{x}_k$, der zugehörige Parametervektor mit $\underline{z}_k$ bezeichnet.

Da Q_{k+1} orthonormal ist und der Vektor $\underline{x}_k \in \mathcal{K}_k$ das Residuum $\underline{r}_k = r_0 - A\underline{x}_k$ minimieren soll, läßt er sich mittels

$$\|r_0 - A\underline{x}_k\| = \|q_1\|r_0\| - AQ_k\underline{z}_k\| = \|Q_{k+1}(\underline{e}_1\|r_0\| - \underline{H}_k\underline{z}_k)\| \tag{8}$$

$$= \|\underline{e}_1\|r_0\| - \underline{H}_k\underline{z}_k\| = \min \tag{9}$$

beschreiben, existiert also immer, und ist somit eindeutig festgelegt durch

$$\underline{z}_k = \underline{H}_k^{\dagger}\underline{e}_1\|r_0\|. \tag{10}$$

GMRes

In der inexakten Variante wird gemäß der Wahl von Bouras und Frayssé im $k$ten Schritt eine Fehlermatrix ΔA_k der relativen Norm ϵ_k konstruiert, wobei

$$\epsilon_k = \frac{\|\Delta A_k\|}{\|A\|} = \min\left(\frac{\eta}{\min(\|\underline{r}_{k-1}\|, 1)}, 1\right). \tag{11}$$

GMRes

In der inexakten Variante wird gemäß der Wahl von Bouras und Frayssé im $k$ten Schritt eine Fehlermatrix ΔA_k der relativen Norm ϵ_k konstruiert, wobei

$$\epsilon_k = \frac{\|\Delta A_k\|}{\|A\|} = \min\left(\frac{\eta}{\min(\|\underline{r}_{k-1}\|, 1)}, 1\right). \tag{11}$$

Damit ist der relative Fehler von A garantiert im Intervall $[\eta, 1]$. In Matlab kann man eine Fehlermatrix mit derselben Besetztheitsstruktur wie die Matrix A mittels `sprand(A)` erzeugen.

GMRes

In der inexakten Variante wird gemäß der Wahl von Bouras und Frayssé im $k$ten Schritt eine Fehlermatrix ΔA_k der relativen Norm ϵ_k konstruiert, wobei

$$\epsilon_k = \frac{\|\Delta A_k\|}{\|A\|} = \min\left(\frac{\eta}{\min(\|\underline{r}_{k-1}\|, 1)}, 1\right). \tag{11}$$

Damit ist der relative Fehler von A garantiert im Intervall $[\eta, 1]$. In Matlab kann man eine Fehlermatrix mit derselben Besetztheitsstruktur wie die Matrix A mittels `sprand(A)` erzeugen.

Die Krylov-Zerlegung wird zur perturbierten Krylov-Zerlegung

$$AQ_k + F_k = Q_{k+1}\underline{H}_k, \tag{12}$$

wobei die Perturbation $F_k \in \mathbb{C}^{(n,k)}$ die folgende Gestalt hat,

$$F_k = G_k = \left(\Delta A_1 q_1, \ldots, \Delta A_k q_k\right). \tag{13}$$

GMRes; Algorithmus

Ein Algorithmus für GMRes sieht folgendermaßen aus:

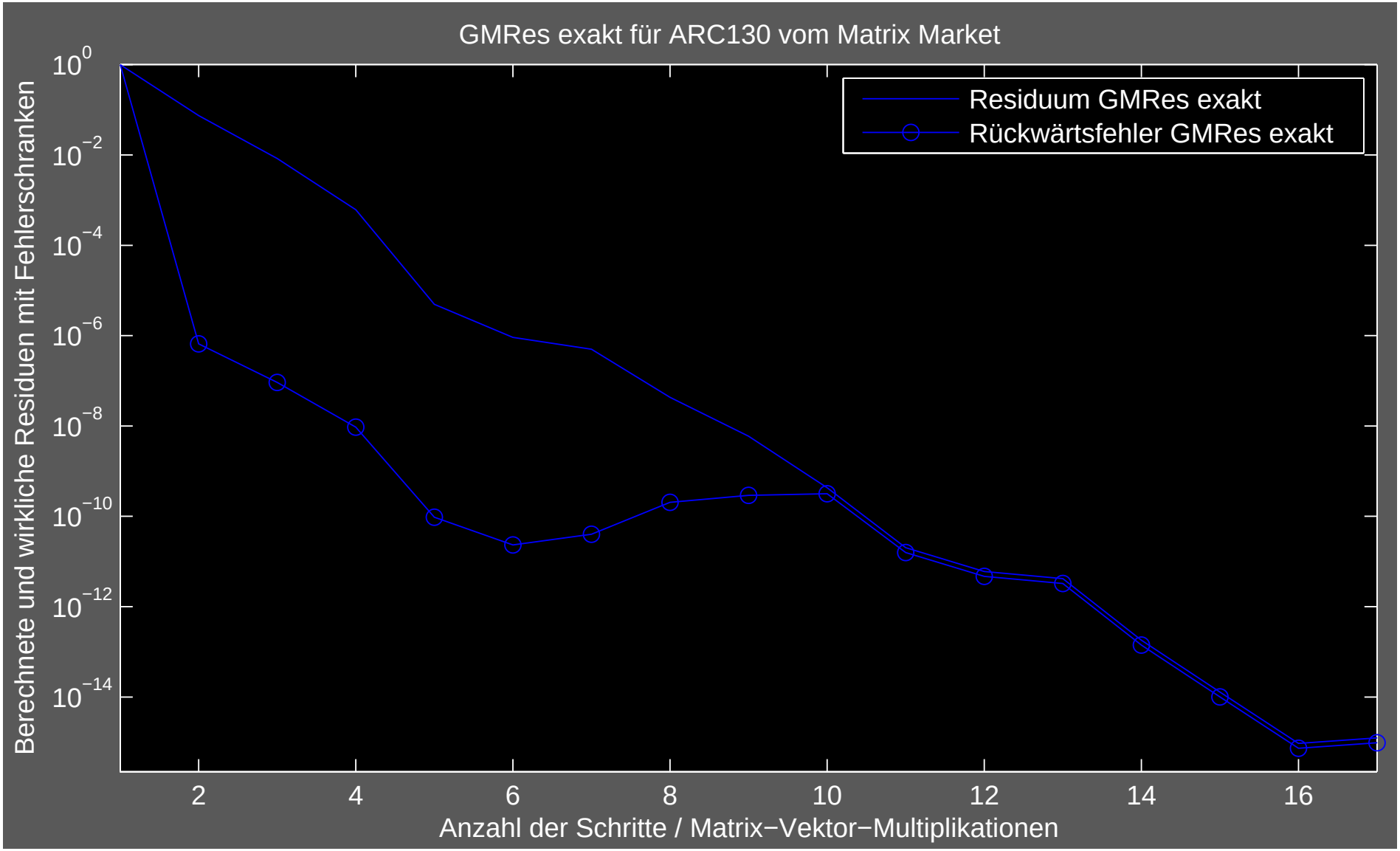
```
Gegeben A,r0
Setze x0 = o, beta = norm(r0);
q(:,1) = r0/beta;
for j = 1,2,...

  w =  A*q(:,j);
  for i = 1:j % Modifiziertes Gram-Schmidt-Verfahren
    h(i,j) = q(:,i)'*w;
    w = w - q(:,i)*h(i,j);
  end
  h(j+1,j) = norm(w);
  q(:,j+1) = w/h(j+1,j);
  z(1:j,j) = h(1:j+1,1:j)\eye(j+1,1)*beta;
  x(:,j) = q(:,1:j)*z(1:j,j);

  Teste auf Konvergenz
end
```

GMRes; Algorithmus

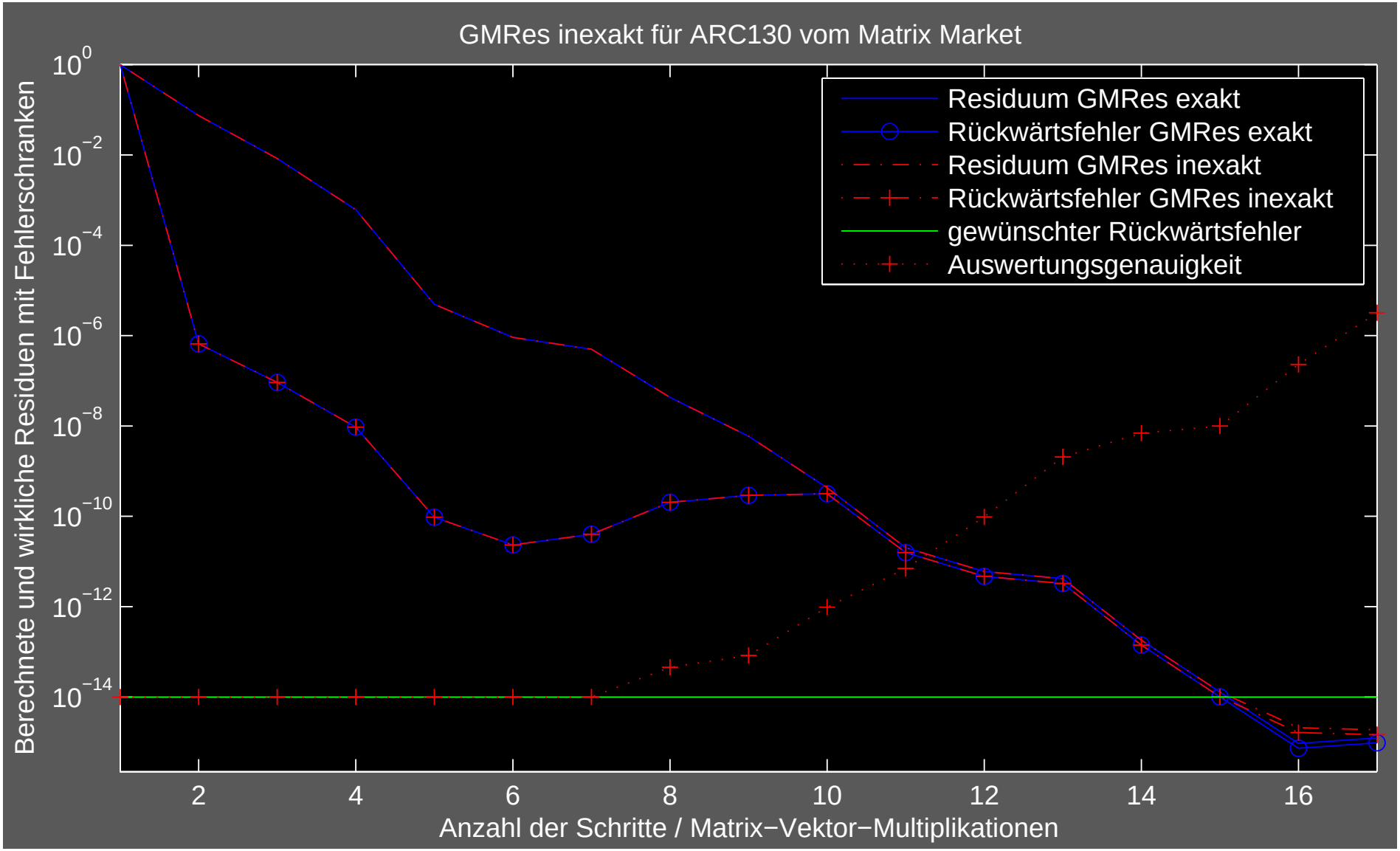
Ein Algorithmus für inexaktes GMRes sieht folgendermaßen aus:

```
Gegeben A,r0,eta
Setze x0 = o, beta = norm(r0);
q(:,1) = r0/beta;
for j = 1,2,...
  Berechne den Fehlerterm DeltaAj
  w = (A+DeltaAj)*q(:,j);
  for i = 1:j % Modifiziertes Gram-Schmidt-Verfahren
    h(i,j) = q(:,i)'*w;
    w = w - q(:,i)*h(i,j);
  end
  h(j+1,j) = norm(w);
  q(:,j+1) = w/h(j+1,j);
  z(1:j,j) = h(1:j+1,1:j)\eye(j+1,1)*beta;
  x(:,j) = q(:,1:j)*z(1:j,j);
  r(:,j) = r0-A*x(:,j);
  Teste auf Konvergenz
end
```

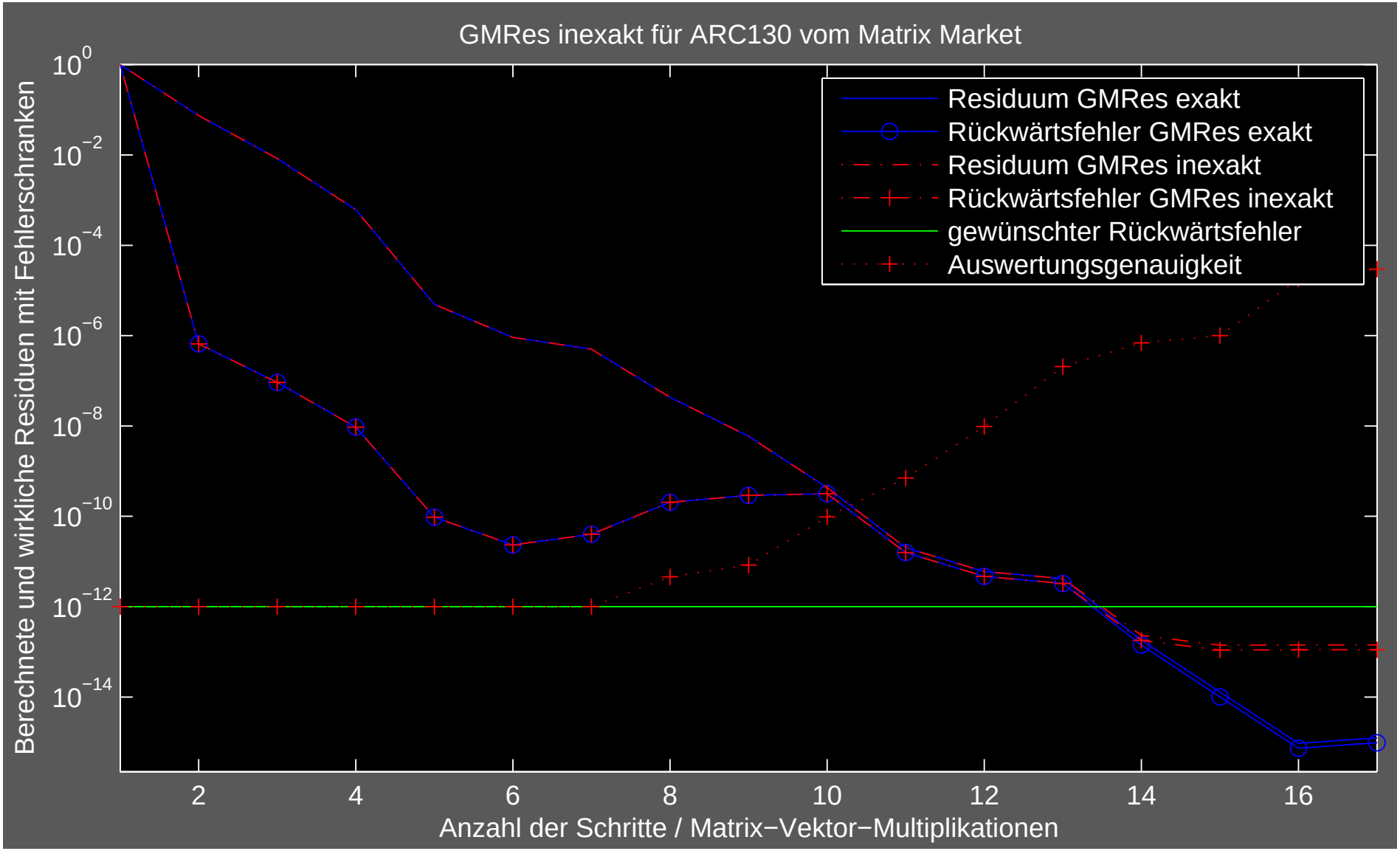
GMRes; exakt


GMRes exakt für ARC130 vom Matrix Market
Residuum GMRes exakt
Rückwärtsfehler GMRes exakt
Berechnete und wirkliche Residuen mit Fehlerschranken
10^{0}
10^{-2}
10^{-4}
10^{-6}
10^{-8}
10^{-10}
10^{-12}
10^{-14}
2
4
6
8
10
12
14
16
Anzahl der Schritte / Matrix−Vektor−Multiplikationen

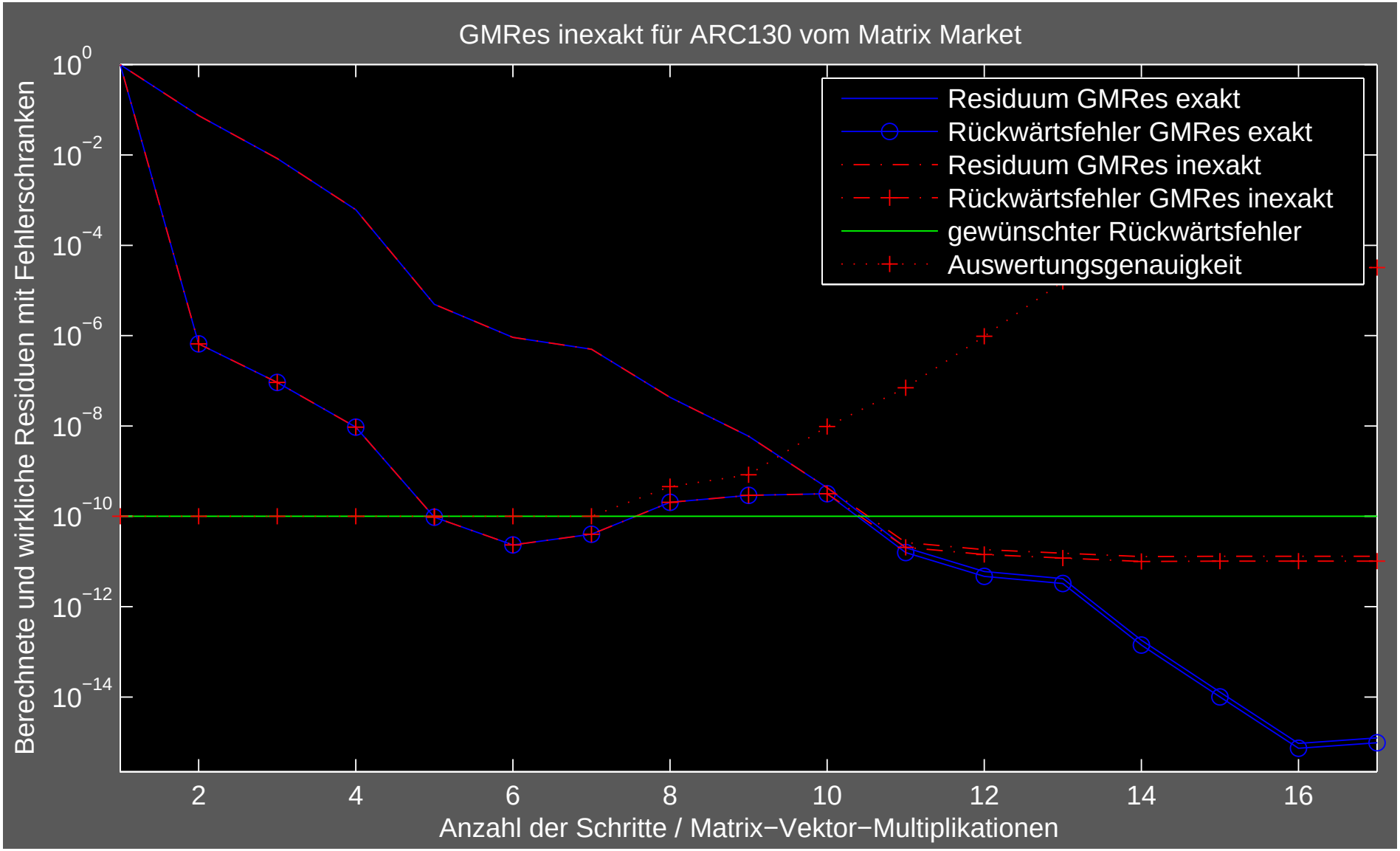
GMRes; inexakt, $\eta = 10^{-14}$



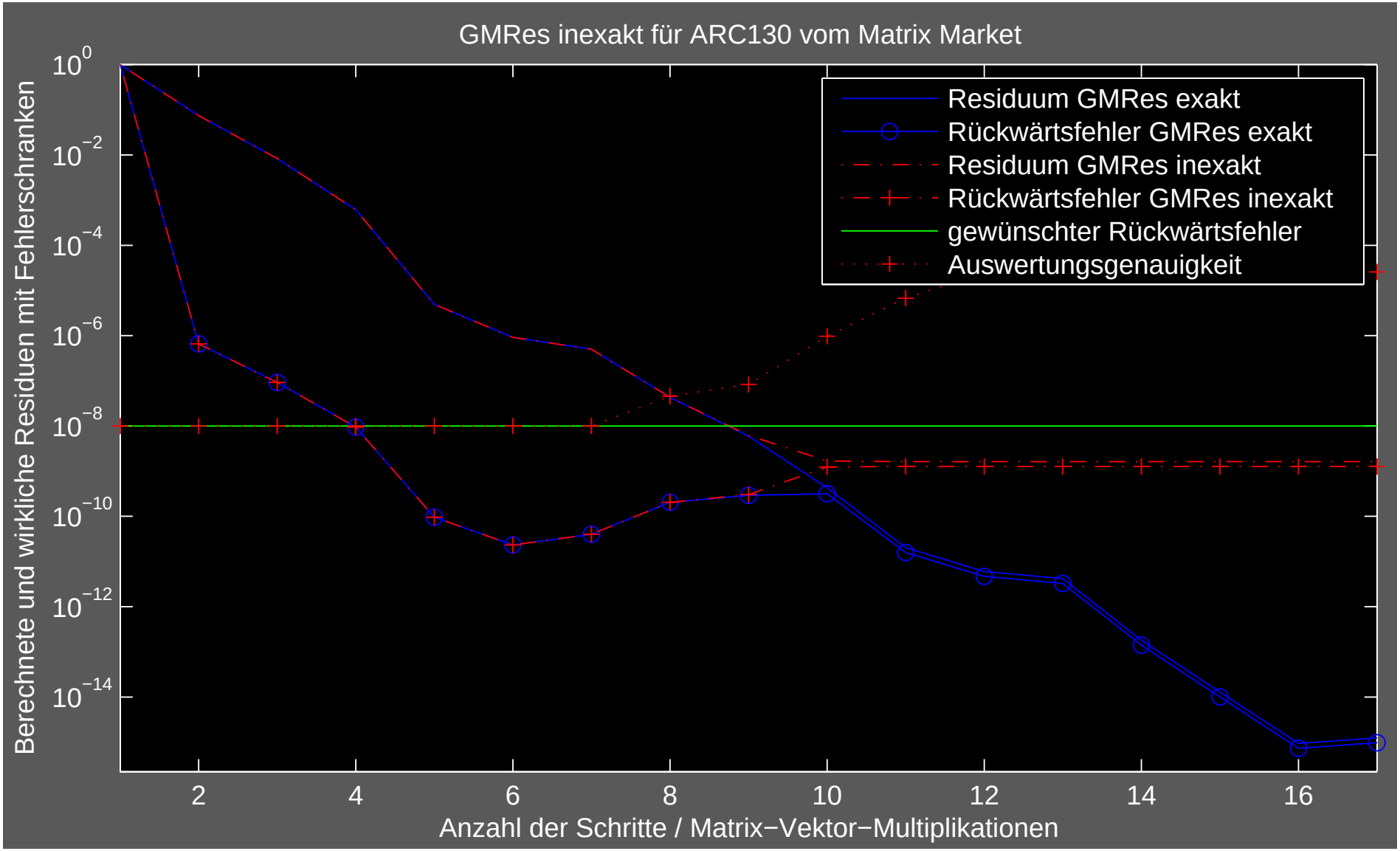
GMRes; inexakt, $\eta = 10^{-12}$



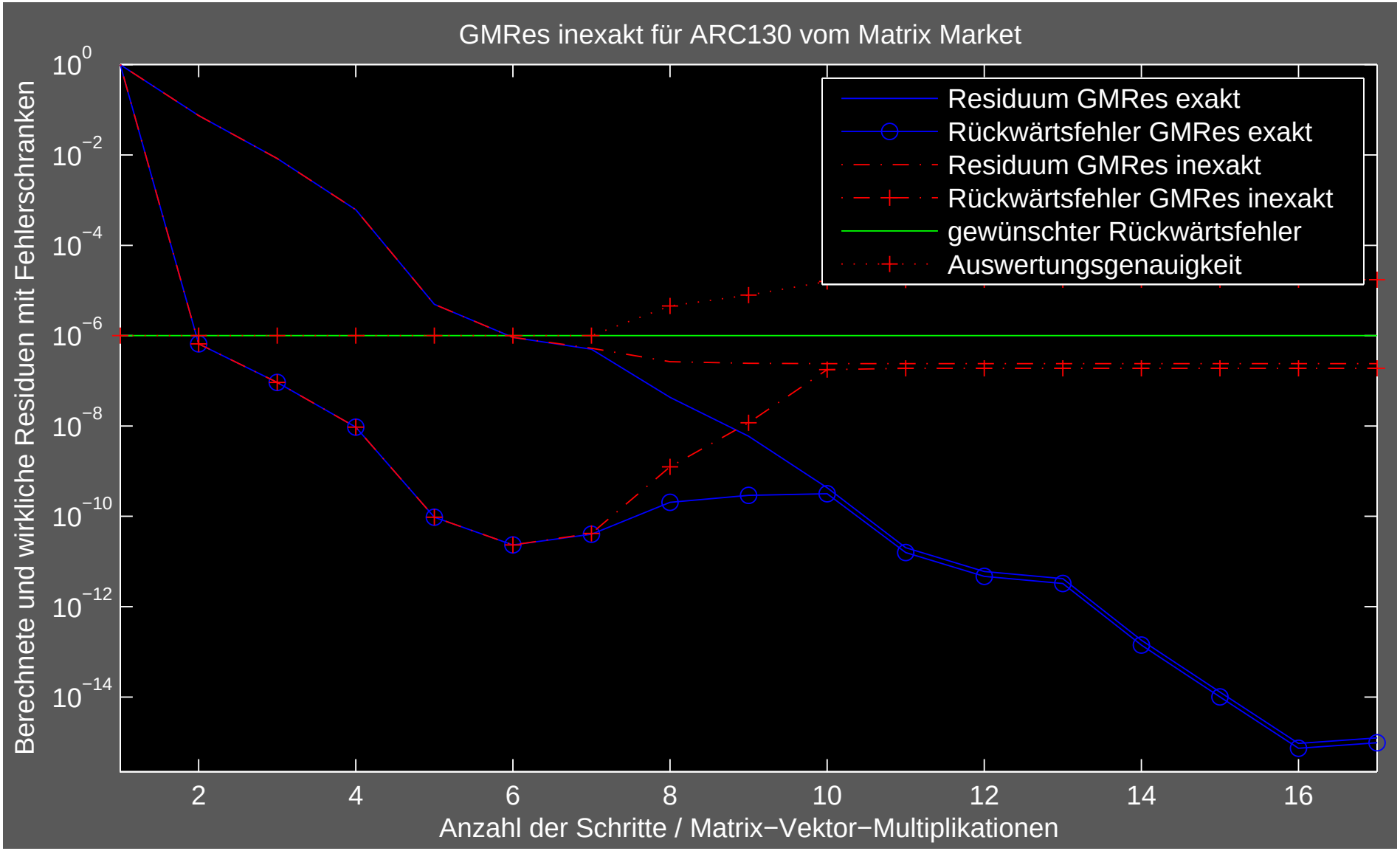
GMRes; inexakt, $\eta = 10^{-10}$



GMRes; inexakt, $\eta = 10^{-8}$



GMRes; inexakt, $\eta = 10^{-6}$



Übersicht

FOM

Auch FOM von Saad und Schultz basiert auf der Krylov-Zerlegung aus der Methode von Arnoldi,

$$AQ_k = Q_{k+1}\underline{H}_k = Q_k H_k + q_{k+1} h_{k+1,k} e_k^T . \tag{14}$$

FOM

Auch FOM von Saad und Schultz basiert auf der Krylov-Zerlegung aus der Methode von Arnoldi,

$$AQ_k = Q_{k+1}\underline{H}_k = Q_k H_k + q_{k+1} h_{k+1,k} e_k^T. \tag{14}$$

Bei FOM wird diejenige approximative Lösung $x_k = Q_k z_k$ aus dem $k$ten Krylovraum $\mathcal{K}_k$ gesucht, deren Residuum $r_k = r_0 - Ax_k$ auf $\mathcal{K}_k$ senkrecht steht,

$$o = Q_k^H (r_0 - Ax_k) = e_1 \|r_0\| - Q_k^H A Q_k z_k \tag{15}$$

$$= e_1 \|r_0\| - Q_k^H Q_k H_k z_k + Q_k^H q_{k+1} h_{k+1,k} e_k^T z_k \tag{16}$$

$$= e_1 \|r_0\| - H_k z_k. \tag{17}$$

FOM

Auch FOM von Saad und Schultz basiert auf der Krylov-Zerlegung aus der Methode von Arnoldi,

$$AQ_k = Q_{k+1}\underline{H}_k = Q_k H_k + q_{k+1} h_{k+1,k} e_k^T. \tag{14}$$

Bei FOM wird diejenige approximative Lösung $x_k = Q_k z_k$ aus dem $k$ten Krylovraum $\mathcal{K}_k$ gesucht, deren Residuum $r_k = r_0 - Ax_k$ auf $\mathcal{K}_k$ senkrecht steht,

$$o = Q_k^H(r_0 - Ax_k) = e_1\|r_0\| - Q_k^H A Q_k z_k \tag{15}$$

$$= e_1\|r_0\| - Q_k^H Q_k H_k z_k + Q_k^H q_{k+1} h_{k+1,k} e_k^T z_k \tag{16}$$

$$= e_1\|r_0\| - H_k z_k. \tag{17}$$

Also existiert eine approximative Lösung nur, wenn H_k regulär ist, und ist dann gegeben durch den Parametervektor

$$z_k = H_k^{-1} e_1 \|r_0\|. \tag{18}$$

FOM; Algorithmus

Ein Algorithmus für FOM sieht folgendermaßen aus:

```
Gegeben A,r0
Setze x0 = o, beta = norm(r0);
q(:,1) = r0/beta;
for j = 1,2,...

  w =  A*q(:,j);
  for i = 1:j % Modifiziertes Gram-Schmidt-Verfahren
    h(i,j) = q(:,i)'*w;
    w = w - q(:,i)*h(i,j);
  end
  h(j+1,j) = norm(w);
  q(:,j+1) = w/h(j+1,j);
  z(1:j,j) = h(1:j,1:j)\eye(j,1)*beta;
  x(:,j) = q(:,1:j)*z(1:j,j);

  Teste auf Konvergenz
end
```

FOM; Algorithmus

Ein Algorithmus für inexaktes FOM sieht folgendermaßen aus:

```
Gegeben A,r0,eta
Setze x0 = o, beta = norm(r0);
q(:,1) = r0/beta;
for j = 1,2,...
  Berechne den Fehlerterm DeltaAj
  w = (A+DeltaAj)*q(:,j);
  for i = 1:j % Modifiziertes Gram-Schmidt-Verfahren
    h(i,j) = q(:,i)'*w;
    w = w - q(:,i)*h(i,j);
  end
  h(j+1,j) = norm(w);
  q(:,j+1) = w/h(j+1,j);
  z(1:j,j) = h(1:j,1:j)\eye(j,1)*beta;
  x(:,j) = q(:,1:j)*z(1:j,j);
  r(:,j) = r0-A*x(:,j);
  Teste auf Konvergenz
end
```

FOM; Relaxationsstrategien

Bouras und Frayssé schlagen wieder die Relaxationsstrategie

$$\epsilon_k = \frac{\|\Delta A_k\|}{\|A\|} = \min\left(\frac{\eta}{\min(\|r_{k-1}\|, 1)}, 1\right) \tag{19}$$

vor, wobei diese Mal aber die r_{k-1} die tatsächlichen Residuen von FOM sind.

FOM; Relaxationsstrategien

Bouras und Frayssé schlagen wieder die Relaxationsstrategie

$$\epsilon_k = \frac{\|\Delta A_k\|}{\|A\|} = \min\left(\frac{\eta}{\min(\|r_{k-1}\|, 1)}, 1\right) \tag{19}$$

vor, wobei diese Mal aber die r_{k-1} die tatsächlichen Residuen von FOM sind.

Selbstverständlich sind auch die tatsächlichen Residuen von FOM in einem realen Umfeld nicht gegeben, analog zu GMRes.

FOM; Relaxationsstrategien

Bouras und Frayssé schlagen wieder die Relaxationsstrategie

$$\epsilon_k = \frac{\|\Delta A_k\|}{\|A\|} = \min\left(\frac{\eta}{\min(\|r_{k-1}\|, 1)}, 1\right) \tag{19}$$

vor, wobei diese Mal aber die r_{k-1} die tatsächlichen Residuen von FOM sind.

Selbstverständlich sind auch die tatsächlichen Residuen von FOM in einem realen Umfeld nicht gegeben, analog zu GMRes.

Da sowohl GMRes als auch FOM auf der orthogonalen Krylov-Zerlegung nach Arnoldi basieren, liegt die Frage nach einem Vergleich und Gemeinsamkeiten der beiden Relaxationsstrategien nahe.

FOM; Relaxationsstrategien

Bouras und Frayssé schlagen wieder die Relaxationsstrategie

$$\epsilon_k = \frac{\|\Delta A_k\|}{\|A\|} = \min\left(\frac{\eta}{\min(\|r_{k-1}\|, 1)}, 1\right) \tag{19}$$

vor, wobei diese Mal aber die r_{k-1} die tatsächlichen Residuen von FOM sind.

Selbstverständlich sind auch die tatsächlichen Residuen von FOM in einem realen Umfeld nicht gegeben, analog zu GMRes.

Da sowohl GMRes als auch FOM auf der orthogonalen Krylov-Zerlegung nach Arnoldi basieren, liegt die Frage nach einem Vergleich und Gemeinsamkeiten der beiden Relaxationsstrategien nahe.

Solch ein Vergleich wurde von van den Eshof und Sleijpen 2004 verwendet, um eine alternative, stärkere Relaxationsstrategie vorzuschlagen; diese Strategie wurde zusätzlich mit Theorie untermauert.

FOM; Relaxationsstrategien

Es gibt bekannte Relationen zwischen Paaren von Methoden, von denen eine auf minimalen Residuen $\underline{r}_k$ (MR) und die andere auf orthogonalen Residuen r_k (OR) basiert. Eine dieser Relationen hat die Form

$$\|\underline{r}_k\| = \frac{1}{\sqrt{\sum_{j=0}^{k} 1/\|r_j\|^2}} \tag{20}$$

FOM; Relaxationsstrategien

Es gibt bekannte Relationen zwischen Paaren von Methoden, von denen eine auf minimalen Residuen $\underline{r}_k$ (MR) und die andere auf orthogonalen Residuen r_k (OR) basiert. Eine dieser Relationen hat die Form

$$\|\underline{r}_k\| = \frac{1}{\sqrt{\sum_{j=0}^{k} 1/\|r_j\|^2}} \equiv \rho_k \leqslant \|r_k\|. \tag{20}$$

Da diese Relation auch im inexakten Fall gilt, und somit wahrscheinlich die Abweichung der Basis des konstruierten Raumes vom wirklichen Krylovraum der Hauptgrund der eventuellen Nichtkonvergenz ist, liegt die folgende Relaxationsstrategie nahe,

$$\epsilon_k = \frac{\|\Delta A_k\|}{\|A\|} = \min\left(\frac{\eta}{\min(\rho_{k-1}, 1)}, 1\right). \tag{21}$$

FOM; Relaxationsstrategien

Es gibt bekannte Relationen zwischen Paaren von Methoden, von denen eine auf minimalen Residuen $\underline{r}_k$ (MR) und die andere auf orthogonalen Residuen r_k (OR) basiert. Eine dieser Relationen hat die Form

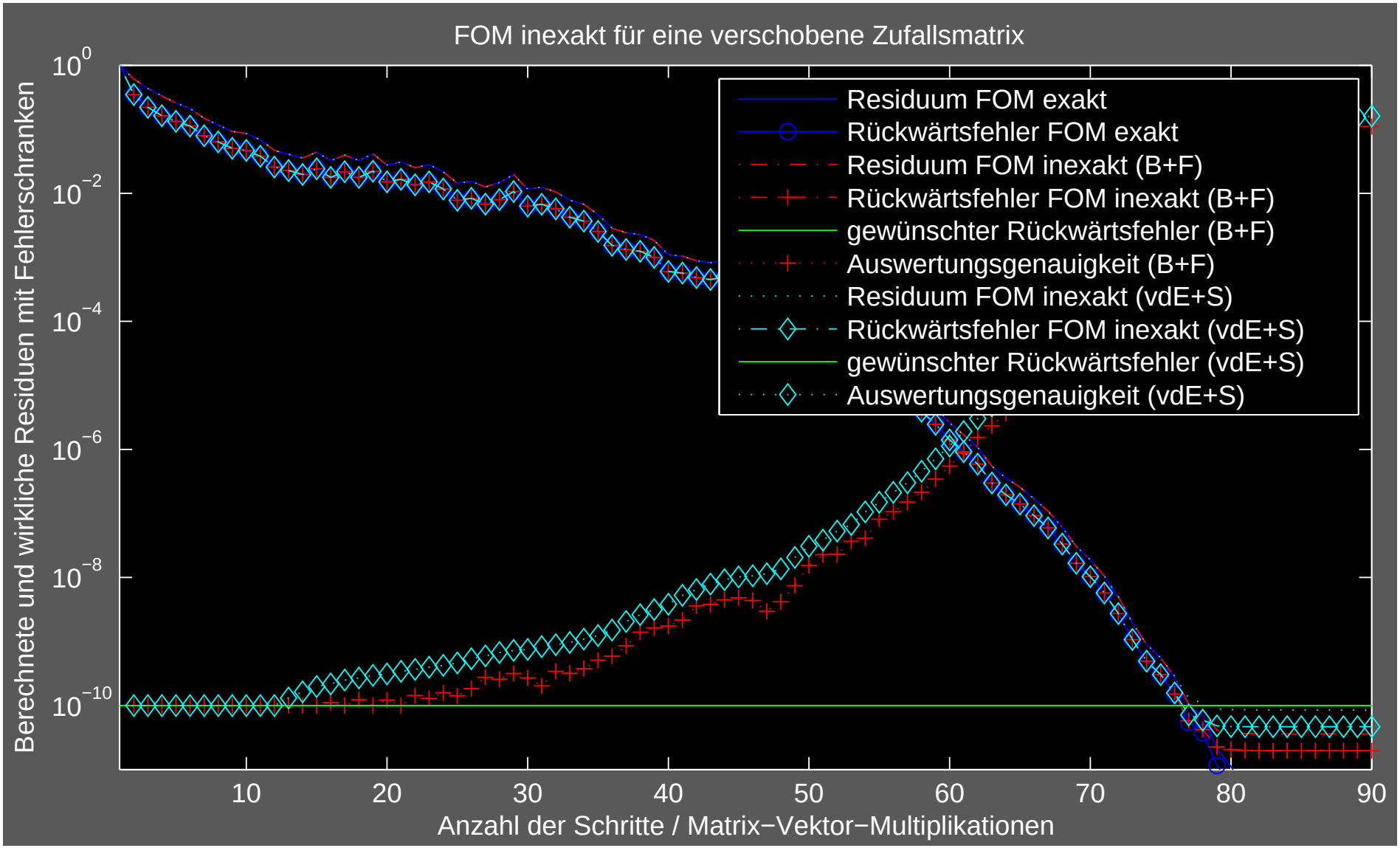
$$\|\underline{r}_k\| = \frac{1}{\sqrt{\sum_{j=0}^{k} 1/\|r_j\|^2}} \equiv \rho_k \leqslant \|r_k\|. \tag{20}$$

Da diese Relation auch im inexakten Fall gilt, und somit wahrscheinlich die Abweichung der Basis des konstruierten Raumes vom wirklichen Krylovraum der Hauptgrund der eventuellen Nichtkonvergenz ist, liegt die folgende Relaxationsstrategie nahe,

$$\epsilon_k = \frac{\|\Delta A_k\|}{\|A\|} = \min\left(\frac{\eta}{\min(\rho_{k-1}, 1)}, 1\right). \tag{21}$$

Das folgende Bild zeigt die beiden Relaxationsstrategien für
`A = randn(100)+sqrt(100)*eye(100).`

FOM; inexakt, $\eta = 10^{-10}$



Übersicht

Richardson-Iteration

Ein einfaches (leichter zu untersuchendes) klassisches Iterationsverfahren zur approximativen Lösung eines linearen Gleichungssystemes ist die Richardson-Iteration.

Richardson-Iteration

Ein einfaches (leichter zu untersuchendes) klassisches Iterationsverfahren zur approximativen Lösung eines linearen Gleichungssystemes ist die Richardson-Iteration. Man denke sich ein Gleichungssystem

$$Ax = r_0 \quad \Leftrightarrow \quad o = r_0 - Ax, \qquad A \in \mathbb{C}^{(n,n)}, r_0 \in \mathbb{C}^n \tag{22}$$

gegeben, präkonditioniere es von links mit einer Präkonditionsmatrix P,

$$o = Pr_0 - PAx \tag{23}$$

und addiere auf beiden Seiten die (unbekannte) Lösung dazu,

$$x = (I - PA)x + Pr_0. \tag{24}$$

Richardson-Iteration

Ein einfaches (leichter zu untersuchendes) klassisches Iterationsverfahren zur approximativen Lösung eines linearen Gleichungssystemes ist die Richardson-Iteration. Man denke sich ein Gleichungssystem

$$Ax = r_0 \quad \Leftrightarrow \quad o = r_0 - Ax, \qquad A \in \mathbb{C}^{(n,n)}, r_0 \in \mathbb{C}^n \tag{22}$$

gegeben, präkonditioniere es von links mit einer Präkonditionsmatrix P,

$$o = Pr_0 - PAx \tag{23}$$

und addiere auf beiden Seiten die (unbekannte) Lösung dazu,

$$x = (I - PA)x + Pr_0. \tag{24}$$

Nun konstruiert man ausgehend von einer gegebenen Approximation x_1 eine Folge von Näherungslösungen gemäß

$$x_{k+1} = (I - PA)x_k + Pr_0. \tag{25}$$

Richardson-Iteration

Man sieht leicht, dass die Iterierten

$$x_{k+1} = (I - PA)x_k + Pr_0. \tag{26}$$

genau dann gegen die Lösung konvergieren, wenn die Matrizen $(I - PA)^k$ gegen die Nullmatrix O konvergieren.

Richardson-Iteration

Man sieht leicht, dass die Iterierten

$$x_{k+1} = (I - PA)x_k + Pr_0. \tag{26}$$

genau dann gegen die Lösung konvergieren, wenn die Matrizen $(I - PA)^k$ gegen die Nullmatrix O konvergieren.

Man spricht dann auch von einer konvergenten Matrix $I - PA$ und kann zeigen, dass $I - PA$ genau dann konvergent ist, wenn der Spektralradius von $I - PA$ kleiner als 1 ist,

$$\lim_{k \to \infty} (I - PA)^k = O \quad \Leftrightarrow \quad \rho(I - PA) < 1. \tag{27}$$

Richardson-Iteration

Man sieht leicht, dass die Iterierten

$$x_{k+1} = (I - PA)x_k + Pr_0. \tag{26}$$

genau dann gegen die Lösung konvergieren, wenn die Matrizen $(I - PA)^k$ gegen die Nullmatrix O konvergieren.

Man spricht dann auch von einer konvergenten Matrix $I - PA$ und kann zeigen, dass $I - PA$ genau dann konvergent ist, wenn der Spektralradius von $I - PA$ kleiner als 1 ist,

$$\lim_{k \to \infty} (I - PA)^k = O \quad \Leftrightarrow \quad \rho(I - PA) < 1. \tag{27}$$

Wir betrachten den einfachsten Fall einer positiv definiten Matrix A und eines skalaren Präkonditinierers $P = \omega I$. Aus der Bedingung (27) folgert man, dass eine optimale Wahl für ω gegeben ist durch

$$\omega = \frac{2}{\lambda_{\max} + \lambda_{\min}}. \tag{28}$$

Richardson-Iteration

Wir formen die Iterationsvorschrift der Richardson-Iteration noch leicht um:

$$x_{k+1} = (I - \omega A)x_k + \omega r_0 = x_k + \omega(r_0 - Ax_k) \tag{29}$$

$$= x_k + \omega r_k. \tag{30}$$

Die Residuen werden mittels eine Updates rekursiv berechnet; pro Schritt muss wieder nur eine Matrix-Vektor-Multiplikation erfolgen:

$$r_{k+1} = r_k - \omega A r_k. \tag{31}$$

Richardson-Iteration

Wir formen die Iterationsvorschrift der Richardson-Iteration noch leicht um:

$$x_{k+1} = (I - \omega A)x_k + \omega r_0 = x_k + \omega(r_0 - Ax_k) \tag{29}$$

$$= x_k + \omega r_k. \tag{30}$$

Die Residuen werden mittels eine Updates rekursiv berechnet; pro Schritt muss wieder nur eine Matrix-Vektor-Multiplikation erfolgen:

$$r_{k+1} = r_k - \omega A r_k. \tag{31}$$

Damit lautet der vollständige Algorithmus:

```
Gegeben A,r0,omega
x(:,1) = o;
r(:,1) = r0;
for j = 2,3,...
  r(:,j) = r(:,j-1)-omega*(A*r(:,j-1));
  x(:,j) = x(:,j-1)+omega*r(:,j-1);
  Teste auf Konvergenz
end
```

Richardson-Iteration

Wir formen die Iterationsvorschrift der Richardson-Iteration noch leicht um:

$$x_{k+1} = (I - \omega A)x_k + \omega r_0 = x_k + \omega(r_0 - Ax_k) \tag{29}$$

$$= x_k + \omega r_k. \tag{30}$$

Die Residuen werden mittels eine Updates rekursiv berechnet; pro Schritt muss wieder nur eine inexakte Matrix-Vektor-Multiplikation erfolgen:

$$r_{k+1} = r_k - \omega(Ar_k + g_k). \tag{31}$$

Damit lautet der vollständige Algorithmus:

```
Gegeben A,r0,omega
x(:,1) = o;
r(:,1) = r0;
for j = 2,3,...
  r(:,j) = r(:,j-1)-omega*(A*r(:,j-1));
  x(:,j) = x(:,j-1)+omega*r(:,j-1);
  Teste auf Konvergenz
end
```

Richardson-Iteration

Wir formen die Iterationsvorschrift der Richardson-Iteration noch leicht um:

$$x_{k+1} = (I - \omega A)x_k + \omega r_0 = x_k + \omega(r_0 - Ax_k) \tag{29}$$

$$= x_k + \omega r_k. \tag{30}$$

Die Residuen werden mittels eine Updates rekursiv berechnet; pro Schritt muss wieder nur eine inexakte Matrix-Vektor-Multiplikation erfolgen:

$$r_{k+1} = r_k - \omega(Ar_k + g_k). \tag{31}$$

Damit lautet der vollständige inexakte Algorithmus:

```
Gegeben A,r0,omega,epsilon
x(:,1) = o;
r(:,1) = r0;
for j = 2,3,...
  r(:,j) = r(:,j-1)-omega*(A*r(:,j-1)+g(:,j-1));
  x(:,j) = x(:,j-1)+omega*r(:,j-1);
  Teste auf Konvergenz
end
```

Richardson-Iteration; inexakt

In der inexakten Variante wird jede Matrix-Vektor-Multiplikation mit einer vorgegebenen Genauigkeit ausgeführt. Wir geben die theoretischen Ergebnisse von van den Eshof und Sleijpen für die inexakte Variante wieder.

Richardson-Iteration; inexakt

In der inexakten Variante wird jede Matrix-Vektor-Multiplikation mit einer vorgegebenen Genauigkeit ausgeführt. Wir geben die theoretischen Ergebnisse von van den Eshof und Sleijpen für die inexakte Variante wieder.

Um ein inexaktes Verfahren zu simulieren, wird ausgenutzt, dass es immer einen Vektor g_k zu jedem ΔA_k gibt, so dass

$$(A + \Delta A_k)r_k = Ar_k + g_k, \qquad \|g_k\| \leqslant \|\Delta A_k\|\|r_k\|. \tag{32}$$

Richardson-Iteration; inexakt

In der inexakten Variante wird jede Matrix-Vektor-Multiplikation mit einer vorgegebenen Genauigkeit ausgeführt. Wir geben die theoretischen Ergebnisse von van den Eshof und Sleijpen für die inexakte Variante wieder.

Um ein inexaktes Verfahren zu simulieren, wird ausgenutzt, dass es immer einen Vektor g_k zu jedem ΔA_k gibt, so dass

$$(A + \Delta A_k)r_k = Ar_k + g_k, \qquad \|g_k\| \leqslant \|\Delta A_k\|\|r_k\|. \tag{32}$$

Als allgemeine Relaxationsstrategie wird

$$\|g_k\| \leqslant \epsilon_k \|A\|\|r_k\| \tag{33}$$

mit wählbarem ϵ_k und den berechneten Residuen r_k vorgeschlagen.

Richardson-Iteration; inexakt

Die zur Strategie von Bouras und Frayssé analoge Wahl

$$\epsilon_k = \frac{\epsilon}{\|r_k\|} \approx \frac{\|\Delta A_k\|}{\|A\|} ? \tag{34}$$

wird im Falle der Richardson-Iteration verwendet, also eine absolute konstante Auswertungsgenauigkeit

$$\|g_k\| \leqslant \epsilon \|A\| \quad \forall\, k. \tag{35}$$

Richardson-Iteration; inexakt

Die zur Strategie von Bouras und Frayssé analoge Wahl

$$\epsilon_k = \frac{\epsilon}{\|r_k\|} \approx \frac{\|\Delta A_k\|}{\|A\|} ? \tag{34}$$

wird im Falle der Richardson-Iteration verwendet, also eine absolute konstante Auswertungsgenauigkeit

$$\|g_k\| \leqslant \epsilon \|A\| \quad \forall\, k. \tag{35}$$

Im Gegensatz zu Bouras und Frayssé zeigen van den Eshof und Sleijpen durch theoretische Betrachtungen, dass diese Wahl zum Erfolg führt.

Richardson-Iteration; inexakt

Die zur Strategie von Bouras und Frayssé analoge Wahl

$$\epsilon_k = \frac{\epsilon}{\|r_k\|} \approx \frac{\|\Delta A_k\|}{\|A\|} ? \tag{34}$$

wird im Falle der Richardson-Iteration verwendet, also eine absolute konstante Auswertungsgenauigkeit

$$\|g_k\| \leqslant \epsilon \|A\| \quad \forall\, k. \tag{35}$$

Im Gegensatz zu Bouras und Frayssé zeigen van den Eshof und Sleijpen durch theoretische Betrachtungen, dass diese Wahl zum Erfolg führt.

Um die Ersetzung der tatsächlichen Residuen durch die berechneten Residuen zu begründen, zeigen van den Eshof und Sleijpen, dass die berechneten Residuen hinreichend nah bei tatsächlichen Residuen liegen, solange das Verfahren noch nicht auf gewünschte Genauigkeit konvergiert ist.

Richardson-Iteration; inexakt

Um Schranken für die Abweichung des wirklichen Residuums

$$r_k^{\text{exakt}} = r_0 - Ax_k \tag{36}$$

vom berechneten Residuum r_k anzugeben, nutzen van den Eshof und Sleijpen eine speziell skalierte Krylov-Zerlegung, welche auch im Falle der Richardson-Iteration aufgestellt werden kann.

Richardson-Iteration; inexakt

Um Schranken für die Abweichung des wirklichen Residuums

$$r_k^{\text{exakt}} = r_0 - Ax_k \tag{36}$$

vom berechneten Residuum r_k anzugeben, nutzen van den Eshof und Sleijpen eine speziell skalierte Krylov-Zerlegung, welche auch im Falle der Richardson-Iteration aufgestellt werden kann.

Diese Skalierung erzwingt, dass die Basisvektoren die berechneten Residuen sind. Die Matrix der Basisvektoren wird deshalb im Folgenden mit R_k statt Q_k bezeichnet.

Richardson-Iteration; inexakt

Um Schranken für die Abweichung des wirklichen Residuums

$$r_k^{\text{exakt}} = r_0 - Ax_k \tag{36}$$

vom berechneten Residuum r_k anzugeben, nutzen van den Eshof und Sleijpen eine speziell skalierte Krylov-Zerlegung, welche auch im Falle der Richardson-Iteration aufgestellt werden kann.

Diese Skalierung erzwingt, dass die Basisvektoren die berechneten Residuen sind. Die Matrix der Basisvektoren wird deshalb im Folgenden mit R_k statt Q_k bezeichnet.

In Matrixform erfüllen aufgrund der Gleichungen $r_{k+1} = r_k - \omega A r_k$ die Residuen $R_k = \left(r_0, \ldots, r_{k-1}\right)$ der Richardson-Iteration die Krylov-Zerlegung

$$AR_k = R_{k+1}\underline{B}_k, \quad \underline{B}_k = \underline{L}_k\omega^{-1}, \quad \underline{L}_k = \text{bidiag}(-1, 1). \tag{37}$$

Richardson-Iteration; inexakt

Aus Betrachtung der Krylov-Zerlegung

$$AR_k = R_{k+1}\underline{B}_k, \quad \underline{B}_k = \underline{L}_k\omega^{-1}, \quad \underline{L}_k = \text{bidiag}(-1,1) \tag{38}$$

der Richardson-Iteration folgt sofort, dass der Vektor $\underline{e} \in \mathbb{C}^{k+1}$ aus lauter Einsen den Linksnullraum von $\underline{B}_k$ aufspannt,

$$\underline{e}^H\underline{B}_k = e^H B_k - \omega^{-1}e_k^H = \underline{o}^H \quad \Rightarrow \quad \omega^{-1}e_k^H B_k^{-1} = e^H. \tag{39}$$

Richardson-Iteration; inexakt

Aus Betrachtung der Krylov-Zerlegung

$$AR_k = R_{k+1}\underline{B}_k, \quad \underline{B}_k = \underline{L}_k\omega^{-1}, \quad \underline{L}_k = \text{bidiag}(-1,1) \tag{38}$$

der Richardson-Iteration folgt sofort, dass der Vektor $\underline{e} \in \mathbb{C}^{k+1}$ aus lauter Einsen den Linksnullraum von $\underline{B}_k$ aufspannt,

$$\underline{e}^H\underline{B}_k = e^H B_k - \omega^{-1}e_k^H = \underline{o}^H \quad \Rightarrow \quad \omega^{-1}e_k^H B_k^{-1} = e^H. \tag{39}$$

Daraus folgt, dass

$$\underline{B}_k B_k^{-1} e_1 = e_1 - e_{k+1}. \tag{40}$$

Richardson-Iteration; inexakt

Aus Betrachtung der Krylov-Zerlegung

$$AR_k = R_{k+1}\underline{B}_k, \quad \underline{B}_k = \underline{L}_k\omega^{-1}, \quad \underline{L}_k = \text{bidiag}(-1,1) \tag{38}$$

der Richardson-Iteration folgt sofort, dass der Vektor $\underline{e} \in \mathbb{C}^{k+1}$ aus lauter Einsen den Linksnullraum von $\underline{B}_k$ aufspannt,

$$\underline{e}^H\underline{B}_k = e^H B_k - \omega^{-1}e_k^H = \underline{o}^H \quad \Rightarrow \quad \omega^{-1}e_k^H B_k^{-1} = e^H. \tag{39}$$

Daraus folgt, dass

$$\underline{B}_k B_k^{-1} e_1 = e_1 - e_{k+1}. \tag{40}$$

Interessanterweise sind die Iterierten charakterisiert durch $x_k = R_k B_k^{-1} e_1$.

Richardson-Iteration; inexakt

Aus Betrachtung der Krylov-Zerlegung

$$AR_k = R_{k+1}\underline{B}_k, \quad \underline{B}_k = \underline{L}_k\omega^{-1}, \quad \underline{L}_k = \operatorname{bidiag}(-1,1) \tag{38}$$

der Richardson-Iteration folgt sofort, dass der Vektor $\underline{e} \in \mathbb{C}^{k+1}$ aus lauter Einsen den Linksnullraum von $\underline{B}_k$ aufspannt,

$$\underline{e}^H\underline{B}_k = e^H B_k - \omega^{-1}e_k^H = \underline{o}^H \quad \Rightarrow \quad \omega^{-1}e_k^H B_k^{-1} = e^H. \tag{39}$$

Daraus folgt, dass

$$\underline{B}_k B_k^{-1} e_1 = e_1 - e_{k+1}. \tag{40}$$

Interessanterweise sind die Iterierten charakterisiert durch $x_k = R_k B_k^{-1} e_1$.

Die Matrix $X_k = \left(o, x_1, \ldots, x_{k-1}\right)$ erfüllt somit die Gleichungen

$$-R_k = X_{k+1}\underline{B}_k, \quad X_k e_1 = o. \tag{41}$$

Richardson-Iteration; inexakt

Wir nehmen jetzt an, dass die Krylov-Zerlegung perturbiert wurde zu

$$AR_k + F_k = R_{k+1}\underline{B}_k, \qquad \underline{e}^H\underline{B}_k = o^H. \tag{42}$$

Richardson-Iteration; inexakt

Wir nehmen jetzt an, dass die Krylov-Zerlegung perturbiert wurde zu

$$AR_k + F_k = R_{k+1}\underline{B}_k, \qquad \underline{e}^H\underline{B}_k = o^H. \tag{42}$$

Der Abstand zwischen tatsächlichem und berechnetem Residuum läßt sich dann beschreiben durch

$$r_k - (r_0 - Ax_k) = (r_k - r_0) + AR_kB_k^{-1}e_1 = (r_k - r_0) + (R_{k+1}\underline{B}_k - F_k)B_k^{-1}e_1 \tag{43}$$

$$= (r_k - r_0) + (r_0 - r_k) - F_kB_k^{-1}e_1 = -\sum_{j=1}^{k} f_j e_j^H B_k^{-1} e_1. \tag{44}$$

Richardson-Iteration; inexakt

Wir nehmen jetzt an, dass die Krylov-Zerlegung perturbiert wurde zu

$$AR_k + F_k = R_{k+1}\underline{B}_k, \qquad \underline{e}^H\underline{B}_k = o^H. \tag{42}$$

Der Abstand zwischen tatsächlichem und berechnetem Residuum läßt sich dann beschreiben durch

$$r_k - (r_0 - Ax_k) = (r_k - r_0) + AR_kB_k^{-1}e_1 = (r_k - r_0) + (R_{k+1}\underline{B}_k - F_k)B_k^{-1}e_1 \tag{43}$$

$$= (r_k - r_0) + (r_0 - r_k) - F_kB_k^{-1}e_1 = -\sum_{j=1}^{k} f_j e_j^H B_k^{-1} e_1. \tag{44}$$

Der Abstand läßt sich also durch eine Linearkombination der Fehler beschreiben; die Koeffizienten der Linearkombination sind die Einträge des Parametervektors, der auch die Linearkombination der Approximation in der Basis der Residuen beschreibt, vergleiche mit [35].

Richardson-Iteration; inexakt

Mit dem Ausdruck

$$r_k - (r_0 - Ax_k) = -\sum_{j=1}^{k} f_j e_j^H B_k^{-1} e_1 \tag{45}$$

Richardson-Iteration; inexakt

Mit dem Ausdruck

$$r_k - (r_0 - Ax_k) = -\sum_{j=1}^{k} f_j e_j^H B_k^{-1} e_1 \tag{45}$$

kann man im Fall der Richardson-Iteration unter Verwendung von

$$e_j^H B_k^{-1} e_1 = \omega \tag{46}$$

Richardson-Iteration; inexakt

Mit dem Ausdruck

$$r_k - (r_0 - Ax_k) = -\sum_{j=1}^{k} f_j e_j^H B_k^{-1} e_1 \tag{45}$$

kann man im Fall der Richardson-Iteration unter Verwendung von

$$e_j^H B_k^{-1} e_1 = \omega \tag{46}$$

und der noch unbestimmten Wahl

$$f_k = g_{k-1}, \quad \|g_k\| \leqslant \epsilon_k \|A\| \|r_k\| \tag{47}$$

Richardson-Iteration; inexakt

Mit dem Ausdruck

$$r_k - (r_0 - Ax_k) = -\sum_{j=1}^{k} f_j e_j^H B_k^{-1} e_1 \tag{45}$$

kann man im Fall der Richardson-Iteration unter Verwendung von

$$e_j^H B_k^{-1} e_1 = \omega \tag{46}$$

und der noch unbestimmten Wahl

$$f_k = g_{k-1}, \quad \|g_k\| \leqslant \epsilon_k \|A\| \|r_k\| \tag{47}$$

den Abstand beschreiben durch

$$\|r_k - (r_0 - Ax_k)\| = \|\sum_{j=1}^{k} f_j e_j^H B_k^{-1} e_1\| = \|\sum_{j=1}^{k} f_j \omega\| \leqslant \omega \|A\| \sum_{j=0}^{k-1} \epsilon_j \|r_j\|. \tag{48}$$

Richardson-Iteration; inexakt

Die Abstand zwischen den Residuen

$$\|r_k - (r_0 - Ax_k)\| = \|\sum_{j=0}^{k-1} g_j\omega\| \leqslant \omega\|A\| \sum_{j=0}^{k-1} \epsilon_j\|r_j\| \tag{49}$$

soll höchstens so groß wie die gewünschte Genauigkeit ϵ werden, was in der Wahl $\epsilon_j = \epsilon/\|r_j\|$ resultiert.

Richardson-Iteration; inexakt

Die Abstand zwischen den Residuen

$$\|r_k - (r_0 - Ax_k)\| = \|\sum_{j=0}^{k-1} g_j\omega\| \leqslant \omega\|A\| \sum_{j=0}^{k-1} \epsilon_j\|r_j\| \tag{49}$$

soll höchstens so groß wie die gewünschte Genauigkeit ϵ werden, was in der Wahl $\epsilon_j = \epsilon/\|r_j\|$ resultiert.

Mit dieser Wahl gilt die Schranke

$$\|r_k - (r_0 - Ax_k)\| \leqslant \omega\|A\| \sum_{j=0}^{k-1} \epsilon_j\|r_k\| = k\epsilon\omega\|A\| = 2k\epsilon\frac{\lambda_{\max}}{\lambda_{\min} + \lambda_{\max}} < 2k\epsilon. \tag{50}$$

Richardson-Iteration; inexakt

Die Abstand zwischen den Residuen

$$\|r_k - (r_0 - Ax_k)\| = \|\sum_{j=0}^{k-1} g_j\omega\| \leqslant \omega\|A\| \sum_{j=0}^{k-1} \epsilon_j\|r_j\| \tag{49}$$

soll höchstens so groß wie die gewünschte Genauigkeit ϵ werden, was in der Wahl $\epsilon_j = \epsilon/\|r_j\|$ resultiert.

Mit dieser Wahl gilt die Schranke

$$\|r_k - (r_0 - Ax_k)\| \leqslant \omega\|A\| \sum_{j=0}^{k-1} \epsilon_j\|r_k\| = k\epsilon\omega\|A\| = 2k\epsilon\frac{\lambda_{\max}}{\lambda_{\min} + \lambda_{\max}} < 2k\epsilon. \tag{50}$$

Aus der Herleitung ersieht man, dass die Schranke für Fehlerterme gewählt als Vielfache eines gegebenen Vektors maximiert wird ($\geqslant k\epsilon$).

Richardson-Iteration; inexakt

Die Abstand zwischen den Residuen

$$\|r_k - (r_0 - Ax_k)\| = \|\sum_{j=0}^{k-1} g_j\omega\| \leqslant \omega\|A\| \sum_{j=0}^{k-1} \epsilon_j\|r_j\| \tag{49}$$

soll höchstens so groß wie die gewünschte Genauigkeit ϵ werden, was in der Wahl $\epsilon_j = \epsilon/\|r_j\|$ resultiert.

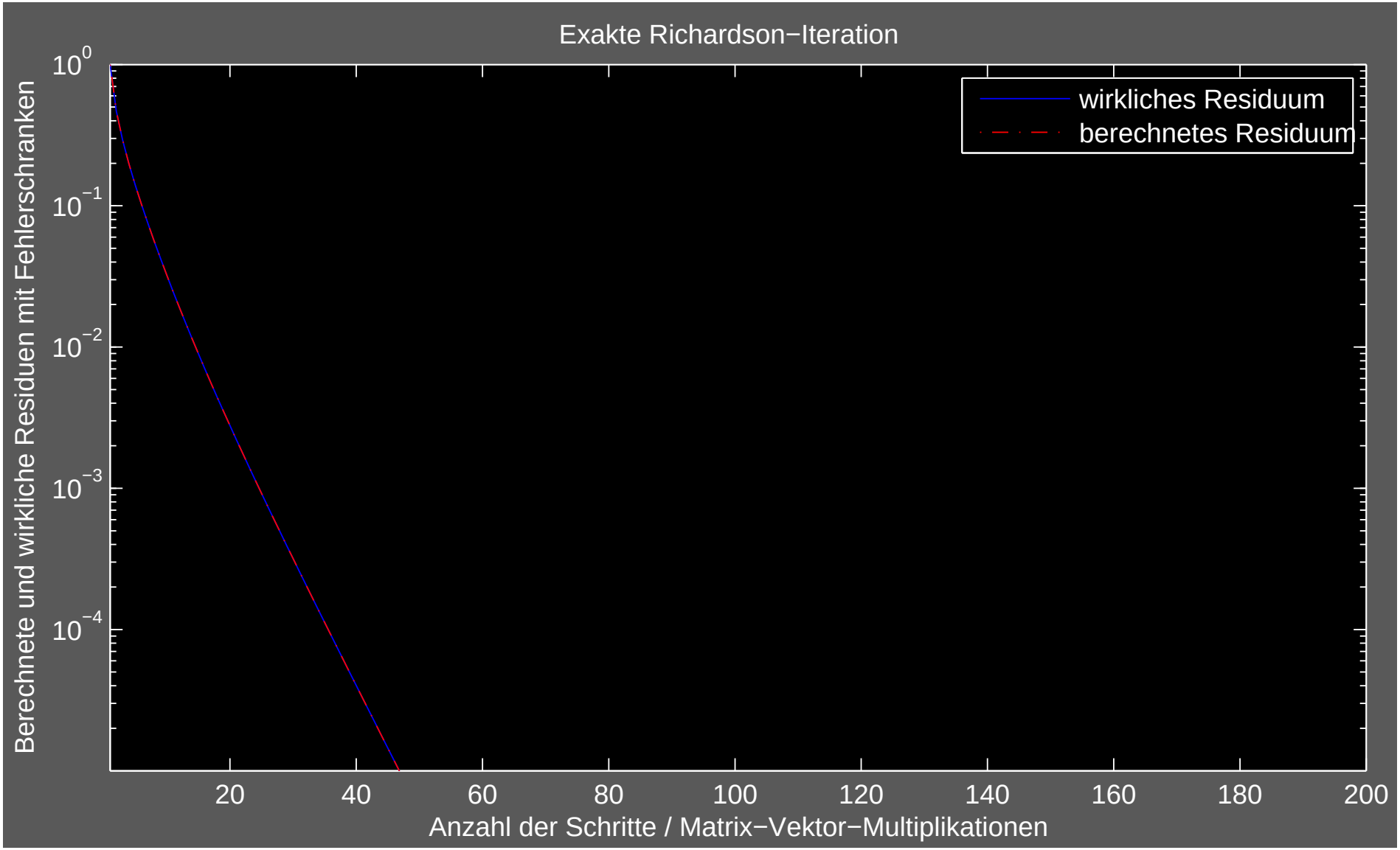
Mit dieser Wahl gilt die Schranke

$$\|r_k - (r_0 - Ax_k)\| \leqslant \omega\|A\| \sum_{j=0}^{k-1} \epsilon_j\|r_k\| = k\epsilon\omega\|A\| = 2k\epsilon\frac{\lambda_{\max}}{\lambda_{\min} + \lambda_{\max}} < 2k\epsilon. \tag{50}$$

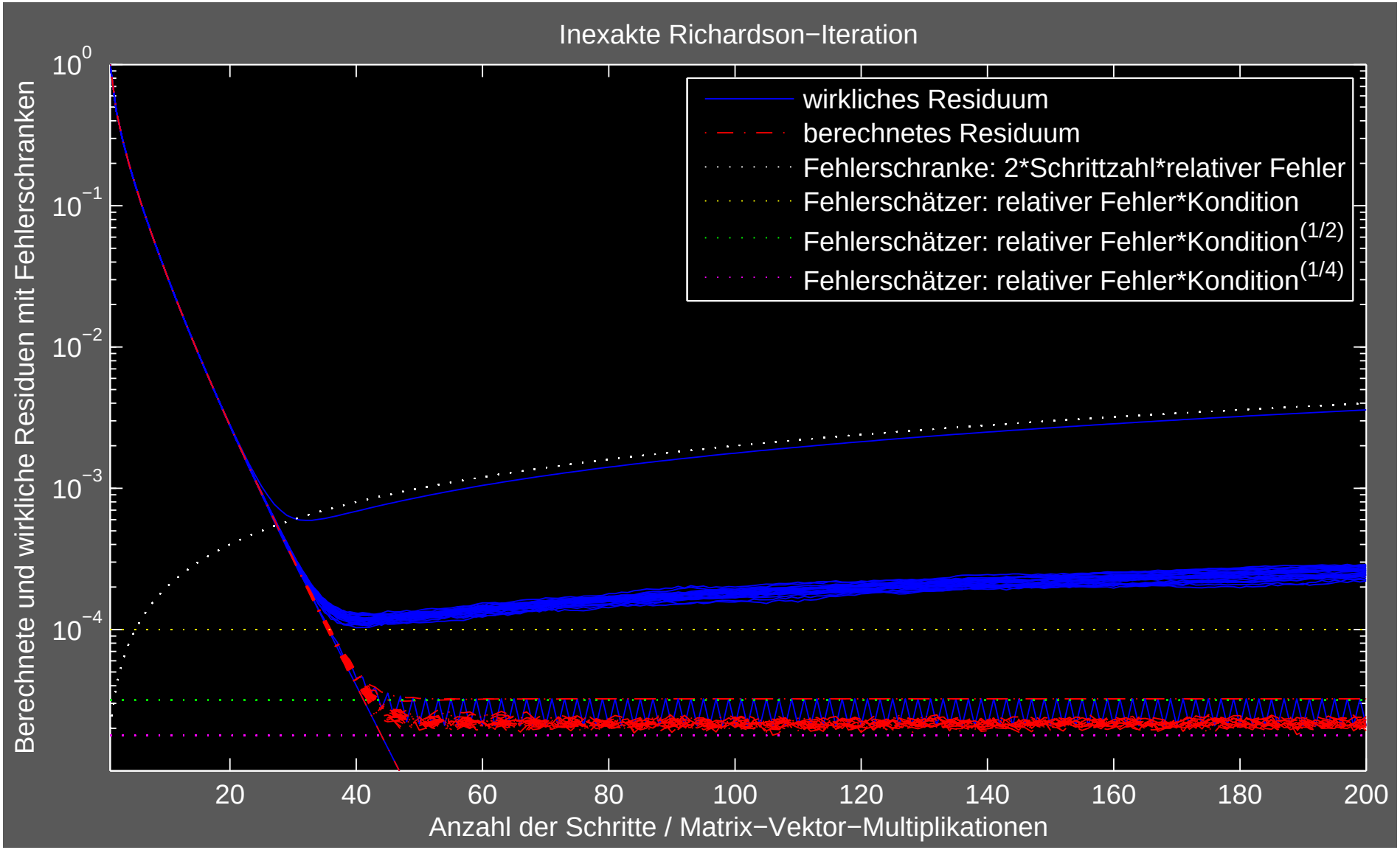
Aus der Herleitung ersieht man, dass die Schranke für Fehlerterme gewählt als Vielfache eines gegebenen Vektors maximiert wird ($\geqslant k\epsilon$).

Die folgenden Bildern zeigen die inexakte Richardson-Iteration mit $n = 100$, $\kappa(A) = 10$ und relativem Fehler $\epsilon = 10^{-5}$.

Richardson-Iteration



Richardson-Iteration; inexakt



Richardson-Iteration; inexakt

Die Konvergenz der inexakten Richardson-Iteration folgt aus der Konvergenz der exakten Richardson-Iteration, wie von van den Eshof und Sleijpen gezeigt wird.

Richardson-Iteration; inexakt

Die Konvergenz der inexakten Richardson-Iteration folgt aus der Konvergenz der exakten Richardson-Iteration, wie von van den Eshof und Sleijpen gezeigt wird.

Theorem (van den Eshof und Sleijpen; 2004)

Sei $\bar{r}_k$ das Residuum der exakten Richardson-Iteration und r_k das berechnete Residuum aus der inexakten Richardson-Iteration mit $\epsilon_k = \epsilon / \|r_k\|$.
Dann gilt

$$\|r_k - \bar{r}_k\| \leqslant \epsilon \kappa(A) = \epsilon \frac{\lambda_{\max}}{\lambda_{\min}}. \tag{51}$$

Richardson-Iteration; inexakt

Die Konvergenz der inexakten Richardson-Iteration folgt aus der Konvergenz der exakten Richardson-Iteration, wie von van den Eshof und Sleijpen gezeigt wird.

Theorem (van den Eshof und Sleijpen; 2004)

Sei $\bar{r}_k$ das Residuum der exakten Richardson-Iteration und r_k das berechnete Residuum aus der inexakten Richardson-Iteration mit $\epsilon_k = \epsilon / \|r_k\|$.
Dann gilt

$$\|r_k - \bar{r}_k\| \leqslant \epsilon\kappa(A) = \epsilon \frac{\lambda_{\max}}{\lambda_{\min}}. \tag{51}$$

Da für $k \to \infty$ das Residuum $\bar{r}_k$ gegen Null konvergiert, konvergiert auch das berechnete Residuum bis zu einem Wert nahe $\epsilon\kappa(A)$, und da der Abstand zwischen dem berechneten und dem tatsächlichen Residuum auch durch die Relaxationsstrategie um circa $\epsilon\kappa(A)$ abweicht, konvergiert auch das inexakte Verfahren.

Richardson-Iteration; inexakt

Beweis.

Die Differenz der Residuen ist gegeben durch

$$r_k - \bar{r}_k = (I-\omega A)^k r_0 + \omega \sum_{j=1}^{k} (I-\omega A)^{k-j} f_j - (I-\omega A)^k r_0 = \omega \sum_{j=1}^{k} (I-\omega A)^{k-j} f_j. \quad (52)$$

Richardson-Iteration; inexakt

Beweis.

Die Differenz der Residuen ist gegeben durch

$$r_k - \bar{r}_k = (I-\omega A)^k r_0 + \omega \sum_{j=1}^{k} (I-\omega A)^{k-j} f_j - (I-\omega A)^k r_0 = \omega \sum_{j=1}^{k} (I-\omega A)^{k-j} f_j. \quad (52)$$

Nach Wahl der Relaxationsstrategie gilt

$$\|r_k - \bar{r}_k\| \leqslant \epsilon|\omega| \| \sum_{j=1}^{k} (I-\omega A)^{k-j} \| \|A\| \quad (53)$$

$$\leqslant \epsilon|\omega| \|(\omega A)^{-1}\| \|A\| = \epsilon\kappa(A). \quad (54)$$

(Der letzte Schritt verwendet eine Neumannsche Reihe.) □

Übersicht

CG

Das CG-Verfahren, besser gesagt, das Verfahren der konjugierten Gradienten von Hestenes und Stiefel, wurde zuerst von Golub und Overton 1988 [17] im Rahmen inexakter Matrix-Vektor-Multiplikationen untersucht. Sie stellten fest, dass das CG-Verfahren nicht so leicht zu untersuchen ist wie die Iterationen nach Richardson und Chebyshev, und dass in ihren numerischen Tests die Konvergenz stark von der erlaubten Störung abhängt.

CG

Das CG-Verfahren, besser gesagt, das Verfahren der konjugierten Gradienten von Hestenes und Stiefel, wurde zuerst von Golub und Overton 1988 [17] im Rahmen inexakter Matrix-Vektor-Multiplikationen untersucht. Sie stellten fest, dass das CG-Verfahren nicht so leicht zu untersuchen ist wie die Iterationen nach Richardson und Chebyshev, und dass in ihren numerischen Tests die Konvergenz stark von der erlaubten Störung abhängt.

Allerdings untersuchten sie Störungen einer festen Größe, was ja für andere Krylov-Raum-Verfahren nicht der Weisheit letzter Schluss scheint.

CG

Das CG-Verfahren, besser gesagt, das Verfahren der konjugierten Gradienten von Hestenes und Stiefel, wurde zuerst von Golub und Overton 1988 [17] im Rahmen inexakter Matrix-Vektor-Multiplikationen untersucht. Sie stellten fest, dass das CG-Verfahren nicht so leicht zu untersuchen ist wie die Iterationen nach Richardson und Chebyshev, und dass in ihren numerischen Tests die Konvergenz stark von der erlaubten Störung abhängt.

Allerdings untersuchten sie Störungen einer festen Größe, was ja für andere Krylov-Raum-Verfahren nicht der Weisheit letzter Schluss scheint.

Bouras und Frayssé untersuchten in ihren Studien auch das CG-Verfahren mit ihrer bekannten Relaxationsstrategie, verfeinert wurde diese Studien später durch Analysen von van den Eshof und Sleijpen.

CG

Das CG-Verfahren, besser gesagt, das Verfahren der konjugierten Gradienten von Hestenes und Stiefel, wurde zuerst von Golub und Overton 1988 [17] im Rahmen inexakter Matrix-Vektor-Multiplikationen untersucht. Sie stellten fest, dass das CG-Verfahren nicht so leicht zu untersuchen ist wie die Iterationen nach Richardson und Chebyshev, und dass in ihren numerischen Tests die Konvergenz stark von der erlaubten Störung abhängt.

Allerdings untersuchten sie Störungen einer festen Größe, was ja für andere Krylov-Raum-Verfahren nicht der Weisheit letzter Schluss scheint.

Bouras und Frayssé untersuchten in ihren Studien auch das CG-Verfahren mit ihrer bekannten Relaxationsstrategie, verfeinert wurde diese Studien später durch Analysen von van den Eshof und Sleijpen.

Erschwert wird jegliches Verstehen des inexakten CG-Verfahrens durch die Tatsache, dass die Konvergenz nahezu immer durch Ausführung in endlicher Genauigkeit beeinflusst wird.

CG; inexakt

Auch CG läßt sich klassifizieren: CG ist ein Verfahren, dass auf orthogonalen Residuen (OR) basiert. Das zugehörige Verfahren der Klasse MR nennt sich Verfahren der konjugierten Residuen, kurz CR.

CG; inexakt

Auch CG läßt sich klassifizieren: CG ist ein Verfahren, dass auf orthogonalen Residuen (OR) basiert. Das zugehörige Verfahren der Klasse MR nennt sich Verfahren der konjugierten Residuen, kurz CR.

Bouras und Frayssé schlagen wieder

$$\epsilon_k = \min\left(\frac{\epsilon}{\min(\|r_{k-1}\|, 1)}, 1\right) \tag{55}$$

als Relaxationsstrategie vor, hier verwendet mit den berechneten Residuen;

CG; inexakt

Auch CG läßt sich klassifizieren: CG ist ein Verfahren, dass auf orthogonalen Residuen (OR) basiert. Das zugehörige Verfahren der Klasse MR nennt sich Verfahren der konjugierten Residuen, kurz CR.

Bouras und Frayssé schlagen wieder

$$\epsilon_k = \min\left(\frac{\epsilon}{\min(\|r_{k-1}\|, 1)}, 1\right) \tag{55}$$

als Relaxationsstrategie vor, hier verwendet mit den berechneten Residuen; van den Eshof und Sleijpen verwenden nach Untermauerung durch handfeste Theorie die geglätteten Residuen von CR, also die Strategie

$$\epsilon_k = \min\left(\frac{\epsilon}{\min(\rho_{k-1}, 1)}, 1\right) \quad \text{mit} \quad \rho_{k-1} = \left(\sqrt{\sum_{j=0}^{k-1} \frac{1}{\|r_j\|^2}}\right)^{-1}. \tag{56}$$

CG; inexakt

Im Gegensatz zu FOM/GMRes, welche lange Rekursionen verwenden, da beide auf dem Verfahren von Arnoldi basieren, verwenden CG/CR kurze Rekursionen, da beide auf der symmetrischen Version des Verfahrens von Arnoldi und damit von Lanczos basieren.

CG; inexakt

Im Gegensatz zu FOM/GMRes, welche lange Rekursionen verwenden, da beide auf dem Verfahren von Arnoldi basieren, verwenden CG/CR kurze Rekursionen, da beide auf der symmetrischen Version des Verfahrens von Arnoldi und damit von Lanczos basieren.

Durch die Verwendung nur der Information der letzten paar Schritte läßt sich ein starker Verlust der Orthogonalität der Residuen nicht verhindern. Dieses geschieht ziemlich früh, nämlich wenn der erste Ritz-Wert konvergiert.

CG; inexakt

Im Gegensatz zu FOM/GMRes, welche lange Rekursionen verwenden, da beide auf dem Verfahren von Arnoldi basieren, verwenden CG/CR kurze Rekursionen, da beide auf der symmetrischen Version des Verfahrens von Arnoldi und damit von Lanczos basieren.

Durch die Verwendung nur der Information der letzten paar Schritte läßt sich ein starker Verlust der Orthogonalität der Residuen nicht verhindern. Dieses geschieht ziemlich früh, nämlich wenn der erste Ritz-Wert konvergiert.

Durch diese Abhängigkeit von den verstärkten Rundungsfehlern weisen mathematisch äquivalente Versionen eines Verfahrens oftmals stark unterschiedliche Verhaltensweisen auf.

CG; inexakt

Im Gegensatz zu FOM/GMRes, welche lange Rekursionen verwenden, da beide auf dem Verfahren von Arnoldi basieren, verwenden CG/CR kurze Rekursionen, da beide auf der symmetrischen Version des Verfahrens von Arnoldi und damit von Lanczos basieren.

Durch die Verwendung nur der Information der letzten paar Schritte läßt sich ein starker Verlust der Orthogonalität der Residuen nicht verhindern. Dieses geschieht ziemlich früh, nämlich wenn der erste Ritz-Wert konvergiert.

Durch diese Abhängigkeit von den verstärkten Rundungsfehlern weisen mathematisch äquivalente Versionen eines Verfahrens oftmals stark unterschiedliche Verhaltensweisen auf.

Noch interessanter ist die Abweichung solcher Verfahren, wenn nur leicht gestört wird. Dann kann es sein, dass die Konvergenzkurven zweier Durchläufe ein und der selben Variante des selben Verfahrens nicht deckungsgleich sind.

CG; inexakt

Von CG (und auch CR) gibt es drei wohlbekannte Versionen, welche meist als CG-OMin, CG-ORes und CG-ODir bezeichnet werden. Diese unterscheiden sich bzgl. der Länge der verwendeten Rekursionen für die Residuen, Iterierten (und Richtungsvektoren).

CG; inexakt

Von CG (und auch CR) gibt es drei wohlbekannte Versionen, welche meist als CG-OMin, CG-ORes und CG-ODir bezeichnet werden. Diese unterscheiden sich bzgl. der Länge der verwendeten Rekursionen für die Residuen, Iterierten (und Richtungsvektoren).

Die aus zwei gekoppelten Zwei-Term-Rekursionen für die Residuen und die Richtungsvektoren und einer Zwei-Term-Rekursion für die Iterierten bestehende Version, welche in den meisten Büchern steht und so auch von Hestenes und Stiefel 1952 beschrieben wurde, stellt CG-OMin dar.

CG; inexakt

Von CG (und auch CR) gibt es drei wohlbekannte Versionen, welche meist als CG-OMin, CG-ORes und CG-ODir bezeichnet werden. Diese unterscheiden sich bzgl. der Länge der verwendeten Rekursionen für die Residuen, Iterierten (und Richtungsvektoren).

Die aus zwei gekoppelten Zwei-Term-Rekursionen für die Residuen und die Richtungsvektoren und einer Zwei-Term-Rekursion für die Iterierten bestehende Version, welche in den meisten Büchern steht und so auch von Hestenes und Stiefel 1952 beschrieben wurde, stellt CG-OMin dar.

Die aus einer Drei-Term-Rekursion für die Residuen und einer Drei-Term-Rekursion für die Iterierten bestehende Variante CG-ORes ist näher am mathematisch äquivalenten Verfahren von Lanczos und wird in den folgenden Experimenten aus Stabilitätsgründen mitverwendet.

CG; inexakt

Von CG (und auch CR) gibt es drei wohlbekannte Versionen, welche meist als CG-OMin, CG-ORes und CG-ODir bezeichnet werden. Diese unterscheiden sich bzgl. der Länge der verwendeten Rekursionen für die Residuen, Iterierten (und Richtungsvektoren).

Die aus zwei gekoppelten Zwei-Term-Rekursionen für die Residuen und die Richtungsvektoren und einer Zwei-Term-Rekursion für die Iterierten bestehende Version, welche in den meisten Büchern steht und so auch von Hestenes und Stiefel 1952 beschrieben wurde, stellt CG-OMin dar.

Die aus einer Drei-Term-Rekursion für die Residuen und einer Drei-Term-Rekursion für die Iterierten bestehende Variante CG-ORes ist näher am mathematisch äquivalenten Verfahren von Lanczos und wird in den folgenden Experimenten aus Stabilitätsgründen mitverwendet.

In dem Artikel von van den Eshof und Sleijpen wird auch noch eine Variante von Rutishauser verwendet; diese ist im Verhalten aber vergleichbar mit CG-ORes und wird hier nicht auch noch vorgestellt.

CG; inexakt

CG ist eigentlich nur für definite Matrizen geeignet. Bei indefiniten Matrizen kann es passieren, dass zwischendurch sehr große Residuen erzeugt werden, was in endlicher Genauigkeit zu Verlust an erreichbarer Genauigkeit führen kann. CG-ORes ist vergleichsweise unempfindlich gegenüber diesen Spitzen in der Konvergenz.

CG; inexakt

CG ist eigentlich nur für definite Matrizen geeignet. Bei indefiniten Matrizen kann es passieren, dass zwischendurch sehr große Residuen erzeugt werden, was in endlicher Genauigkeit zu Verlust an erreichbarer Genauigkeit führen kann. CG-ORes ist vergleichsweise unempfindlich gegenüber diesen Spitzen in der Konvergenz.

CG-OMin ist hingegen im traditionellen Umfeld (SPD-Matrizen; keine inexakten Matrix-Vektor-Multiplikationen) der Sieger, da die gekoppelte Rekursion meist eine höhere Genauigkeit als CG-ORes erreicht.

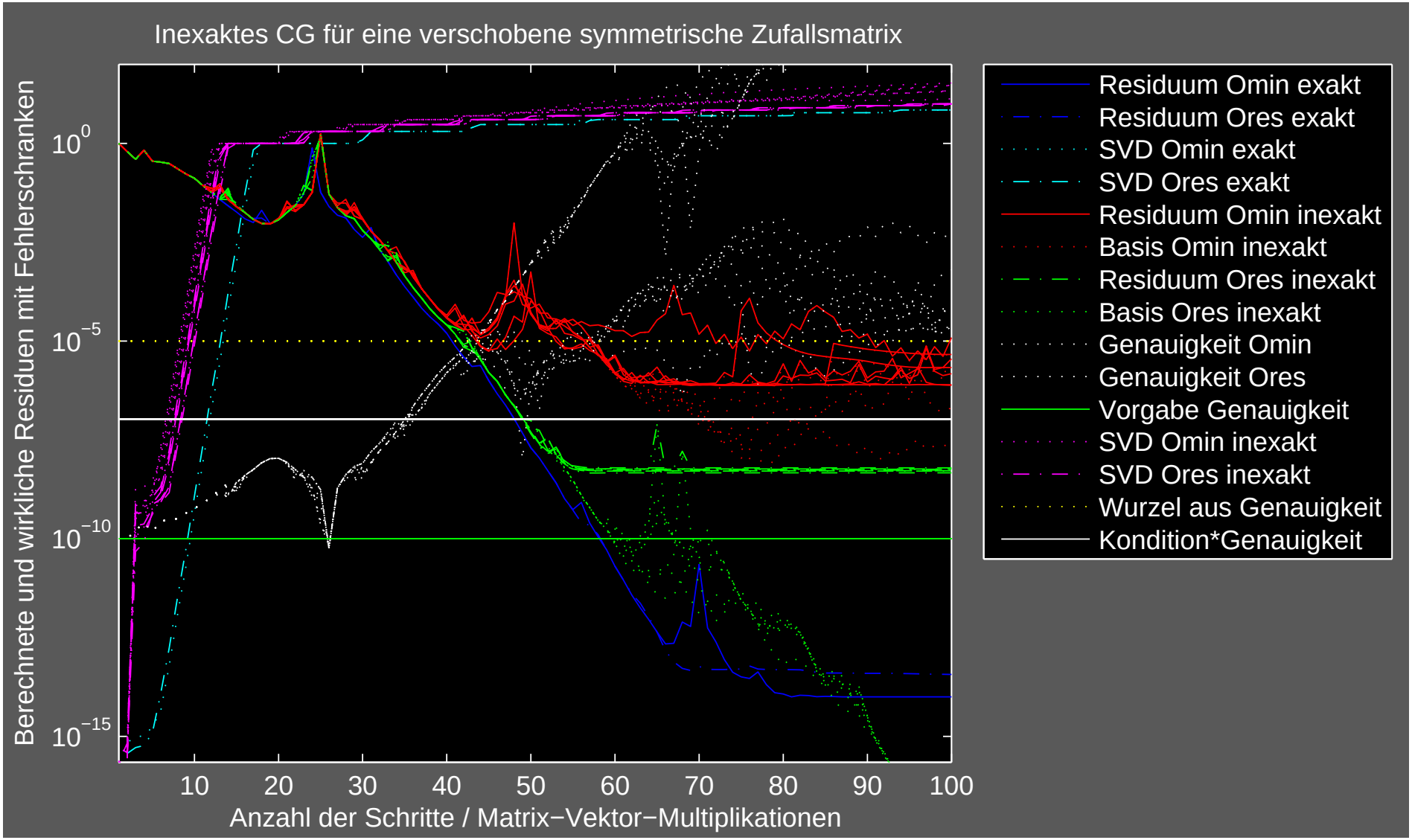
CG; inexakt

CG ist eigentlich nur für definite Matrizen geeignet. Bei indefiniten Matrizen kann es passieren, dass zwischendurch sehr große Residuen erzeugt werden, was in endlicher Genauigkeit zu Verlust an erreichbarer Genauigkeit führen kann. CG-ORes ist vergleichsweise unempfindlich gegenüber diesen Spitzen in der Konvergenz.

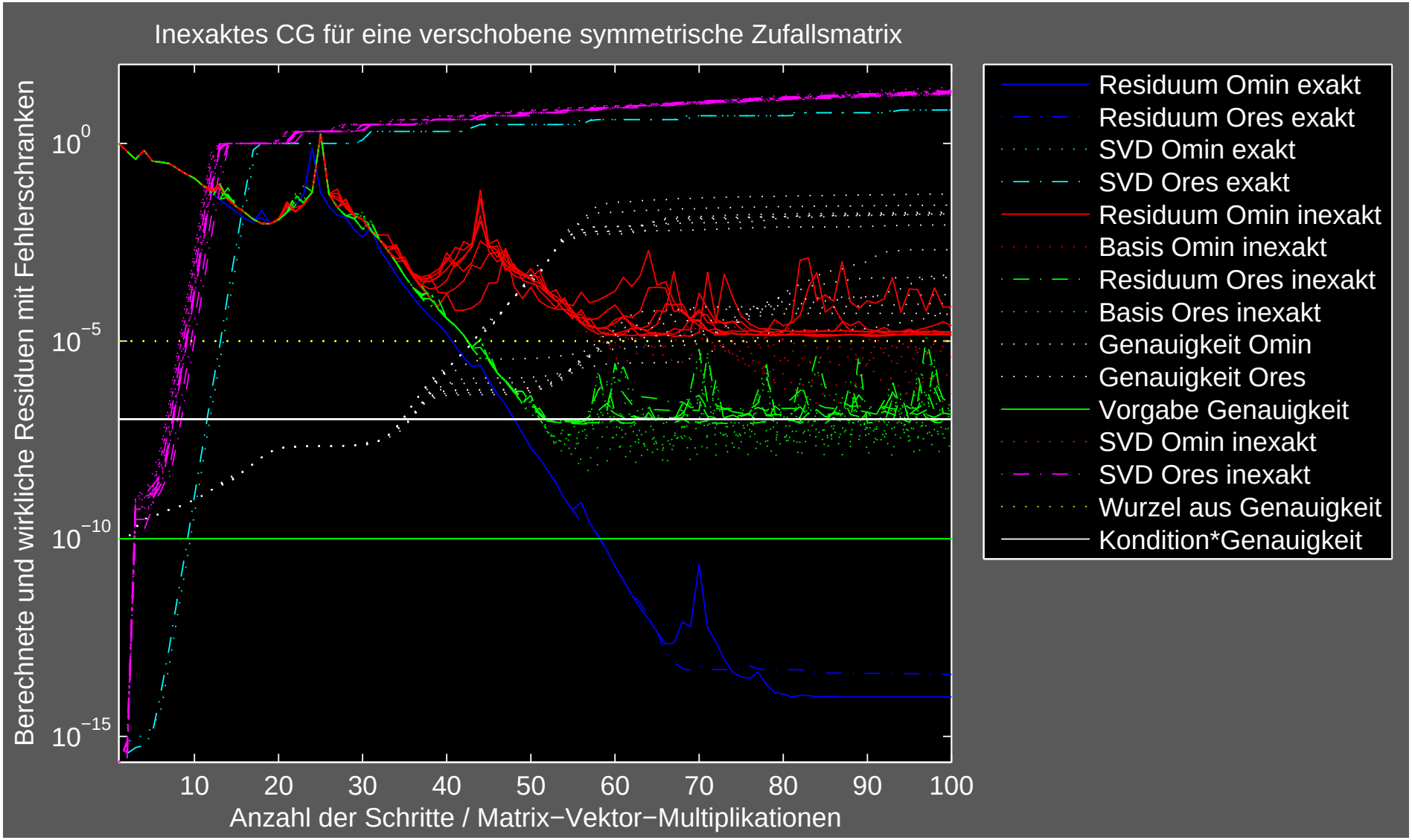
CG-OMin ist hingegen im traditionellen Umfeld (SPD-Matrizen; keine inexakten Matrix-Vektor-Multiplikationen) der Sieger, da die gekoppelte Rekursion meist eine höhere Genauigkeit als CG-ORes erreicht.

In den Beispielen fangen wir mit einer leicht indefiniten symmetrischen Matrix an, welche nur einen betragsmäßig kleinen negativen Eigenwert besitzt. Danach folgt der Standardfall einer SPD-Matrix. Die verwendeten Matrizen sind durch Spektralverschiebung auseinander hervorgegangen.

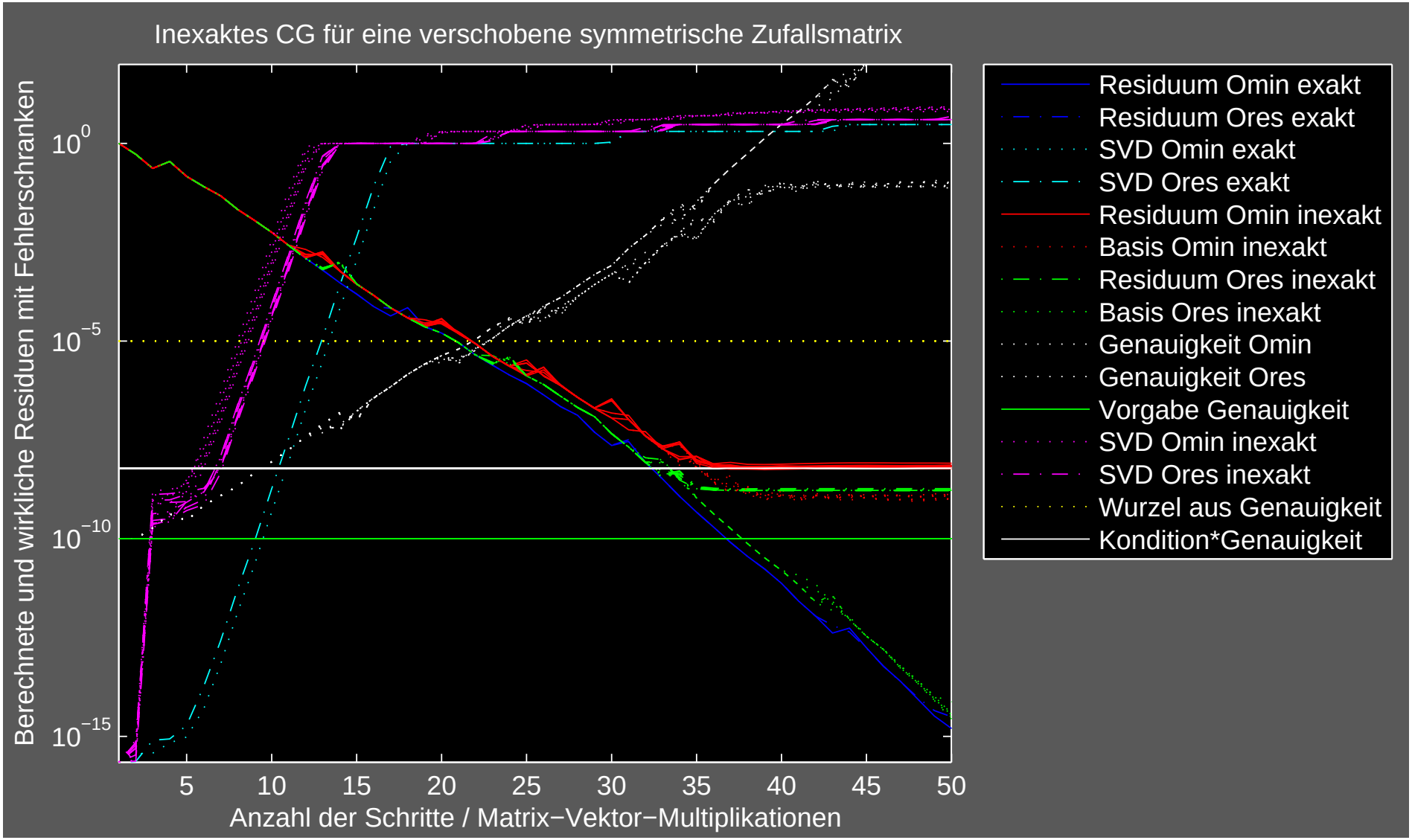
CG; inexakt, CG-basierte Relaxation, leicht indefinit



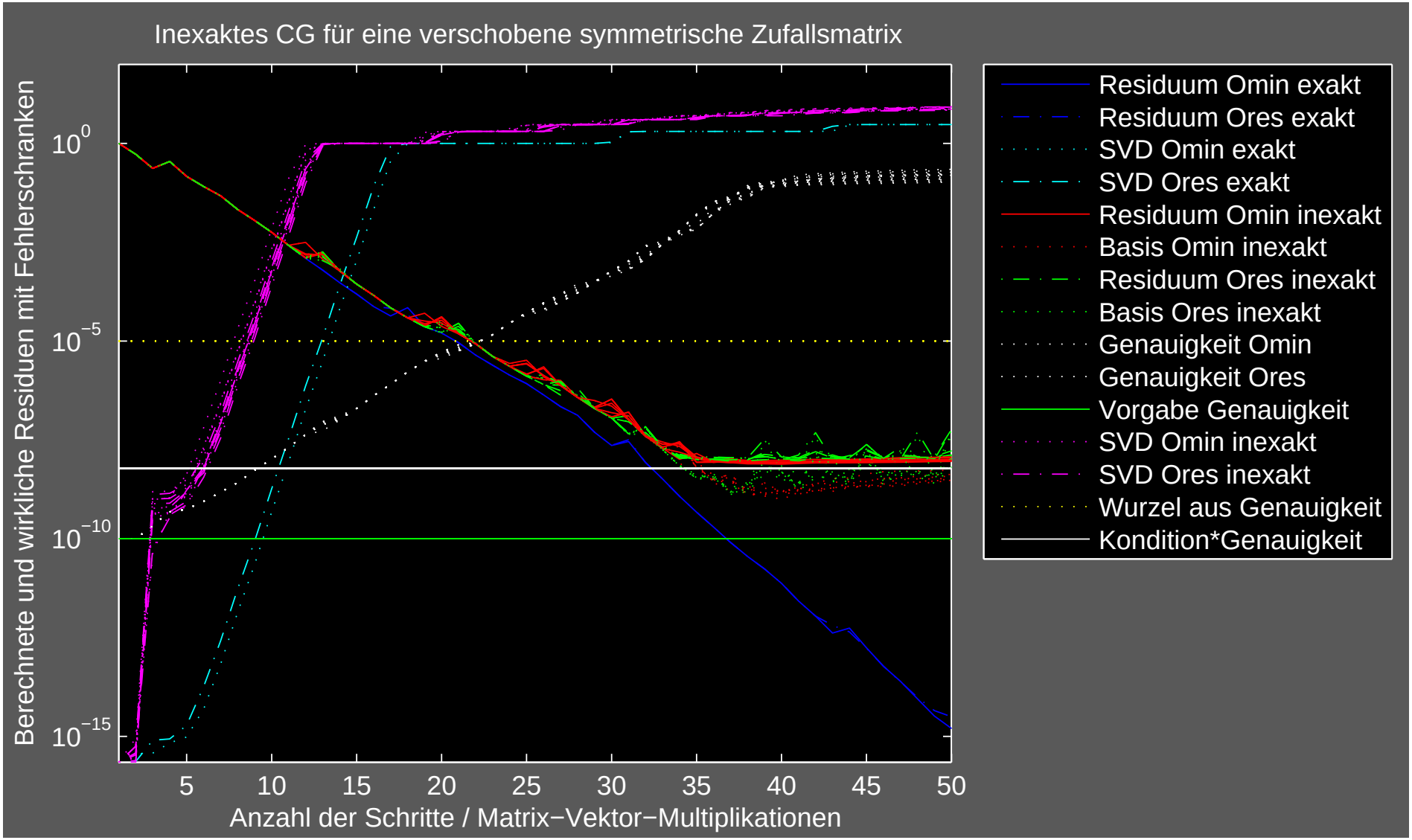
CG; inexakt, CR-basierte Relaxation, leicht indefinit



CG; inexakt, CG-basierte Relaxation, definit



CG; inexakt, CR-basierte Relaxation, definit



Inexakte Verfahren für Eigenwertaufgaben

Grundsätzlich sind Krylov-Verfahren genau wie für lineare Gleichungssysteme auch für Eigenwertaufgaben inexakt ausführbar, was ja auch im Experiment [4, 5, 7, 3, 6] und in der theoretischen Vorarbeit [21] angedeutet wurde.

Inexakte Verfahren für Eigenwertaufgaben

Grundsätzlich sind Krylov-Verfahren genau wie für lineare Gleichungssysteme auch für Eigenwertaufgaben inexakt ausführbar, was ja auch im Experiment [4, 5, 7, 3, 6] und in der theoretischen Vorarbeit [21] angedeutet wurde.

Eine Analyse von inexakter RQI (Rayleigh-Quotienten-Iteration) wurde 2002 von Valeria Simoncini und Lars Eldén veröffentlicht, siehe [29]. Inexakte Krylov-Raum-Verfahren für das Eigenwertproblem wurden 2005 von Valeria Simoncini mit theoretischer Untermauerung versehen und propagiert, siehe [28].

Inexakte Verfahren für Eigenwertaufgaben

Grundsätzlich sind Krylov-Verfahren genau wie für lineare Gleichungssysteme auch für Eigenwertaufgaben inexakt ausführbar, was ja auch im Experiment [4, 5, 7, 3, 6] und in der theoretischen Vorarbeit [21] angedeutet wurde.

Eine Analyse von inexakter RQI (Rayleigh-Quotienten-Iteration) wurde 2002 von Valeria Simoncini und Lars Eldén veröffentlicht, siehe [29]. Inexakte Krylov-Raum-Verfahren für das Eigenwertproblem wurden 2005 von Valeria Simoncini mit theoretischer Untermauerung versehen und propagiert, siehe [28].

Die 2007 erschienene Dissertation [13] von Melina A. Freitag beschäftigt sich u.a. mit der Wahl angepasster Präkonditionierer, da die zu lösenden Gleichungssysteme von besonderer Gestalt sind und es bekannt ist, dass die rechten Seiten (im Laufe des Algorithmus) bereits Näherungen für Eigenvektoren sind.

[1] Guido Arnold, Nigel Cundy, Jasper van den Eshof, Andreas Frommer, Stefan Krieg, Thomas Lippert, and Katrin Schäfer.
Numerical methods for the QCD overlap operator: II. optimal Krylov subspace methods.
In Boriçi et al. [2], pages 153–167. hep-lat/0311025.

[2] Artan Boriçi, Andreas Frommer, Bálint Joó, Anthony Kennedy, and Brian Pendleton, editors.
QCD and Numerical Analysis III, volume 47 of *Lecture Notes in Computational Science and Engineering*. Springer, 2005.

[3] Amina Bouras.
Contrôle de convergence de solveurs emboîtés pour le calcul de valeurs propres avec inversion.
Thése, Université de Toulouse I, 2000. CERFACS Technical Report TH/PA/00/77.

[4] Amina Bouras and Valérie Frayssé.
A relaxation strategy for inexact matrix-vector products for Krylov methods.
Technical Report TR/PA/00/15, Université Toulouse I and CERFACS, 2000.

[5] Amina Bouras and Valérie Frayssé.
A relaxation strategy for the Arnoldi method in eigenproblems.
Technical Report TR/PA/00/16, Université Toulouse I and CERFACS, 2000.

[6] Amina Bouras and Valérie Frayssé.
Inexact matrix-vector products in Krylov methods for solving linear systems: A relaxation strategy.
SIAM J. Matrix Anal. Appl., 26(3):660–678, 2005.

[7] Amina Bouras, Valérie Frayssé, and Luc Giraud.
A relaxation strategy for inner-outer linear solvers in domain decomposition methods.
Technical Report TR/PA/00/17, Université Toulouse I and CERFACS, 2000.

[8] Emil Cătinaş.
The inexact, inexact perturbed, and quasi-Newton methods are equivalent models.
Math. Comp., 74(249):291–301, 2005.

[9] N. Cundy, S. Krieg, G. Arnold, A. Frommer, Th. Lippert, and K. Schilling.
Numerical methods for the QCD overlap operator IV: Hybrid Monte Carlo.
arXiv.org preprint hep-lat/0502007, arXiv.org, 2005. v1.

[10] N. Cundy, J. van den Eshof, A. Frommer, S. Krieg, Th. Lippert, and K. Schäfer.
Numerical methods for the QCD overlap operator: III. Nested iterations.
Comp. Physics Comm., 165(3):221–242, 2005. hep-lat/0405003.

[11] Ron S. Dembo, Stanley C. Eisenstat, and Trond Steihaug.
Inexact Newton methods.
SIAM J. Numer. Anal., 19(2):400–408, 1982.

[12] Stanley C. Eisenstat and Homer F. Walker.
Choosing the forcing terms in an inexact Newton method.
SIAM J. Sci. Comput., 17(1):16–32, 1996.

[13] Melina Annerose Freitag.
Inner-outer Iterative Methods for Eigenvalue Problems — Convergence and Preconditioning.
PhD thesis, Department of Mathematical Sciences, University of Bath, September 2007.

[14] Andreas Frommer and Daniel B. Szyld.
H-splittings and two-stage iterative methods.
Numer. Math., 63(1):345–356, 1992.

[15] Eldar Giladi, Gene H. Golub, and Joseph B. Keller.
Inner and outer iterations for the Chebyshev algorithm.
SIAM J. Numer. Anal., 35(1):300–319, 1998.

[16] Gene H. Golub and Michael L. Overton.
Convergence of a two-stage Richardson iterative procedure for solving systems of linear equations.
In *Numerical Analysis; Proceedings of the 9th Biennial Conference Held at Dundee, Scotland, June 23–26, 1981*, volume 912 of *Lecture Notes in Mathematics*, pages 125–139. Springer, Berlin, Heidelberg, 1982.

[17] Gene H. Golub and Michael L. Overton.
The convergence of inexact Chebyshev and Richardson iterative methods for solving linear systems.
Numer. Math., 53(5):571–593, 1988.

[18] Gene H. Golub and Qiang Ye.
Inexact preconditioned conjugate gradient method with inner-outer iteration.
Technical Report SCCM97-04, Department of Computer Science, Stanford University, Stanford, California, USA, 1997.

[19] Gene H. Golub and Qiang Ye.
Inexact inverse iterations for the generalized eigenvalue problems.
Technical Report SCCM99-02, Department of Computer Science, Stanford University, Stanford, California, USA, 1999.

[20] Gene H. Golub and Qiang Ye.
Inexact preconditioned conjugate gradient method with inner-outer iteration.
SIAM J. Sci. Comput., 21(4):1305–1320, 1999.

[21] Gene H. Golub, Zhenyue Zhang, and Hongyuan Zha.
Large sparse symmetric eigenvalue problems with homogeneous linear constraints: the Lanczos process with inner–outer iterations.
Linear Algebra Appl., 309(1–3):289–306, 2000.

[22] James E. Gunn.
The numerical solution of $\nabla \cdot a\nabla u = f$ by a semi-explicit alternating-direction iterative technique.
Numer. Math., 6(1):181–184, 1964.

[23] Benedetta Morini.
Convergence behaviour of inexact Newton methods.
Math. Comp., 68(228):1605–1613, 1999.

[24] Nancy K. Nichols.
On the convergence of two-stage iterative processes for solving linear equations.
SIAM J. Numer. Anal., 10(3):460–469, 1973.

[25] R. A. Nicolaides.
On the local convergence of certain two step terative procedures.
Numer. Math., 24(2):95–101, 1975.
Title contains misprint ‘Terative’ instead of ‘Iterative’.

[26] Victor Pereyra.
Accelerating the convergence of discretization algorithms.
SIAM J. Numer. Anal., 4(4):508–533, 1967.

[27] Katrin Schäfer.
Krylov-Unterraum-Verfahren für die Matrix-Signum-Funktion.
Diplomarbeit, Fachbereich Mathematik, Bergische Universität Gesamthochschule Wuppertal, Januar 2002.

[28] V. Simoncini.
Variable accuracy of matrix-vector products in projection methods for eigencomputation.
SIAM J. Numer. Anal., 43(3):1155–1174, 2005.

[29] Valeria Simoncini and Lars Eldén.
Inexact Rayleigh quotient-type methods for eigenvalue computations.
BIT Numerical Mathematics, 42(1):159–182, March 2002.

[30] Valeria Simoncini and Daniel B. Szyld.
Theory of inexact Krylov subspace methods and applications to scientific computing.
SIAM J. Sci. Comput., 25(2):454–477, 2003.

[31] P. Smit and M. H. C Paardekooper.
The effects of inexact solvers in algorithms for symmetric eigenvalue problems.
Linear Algebra Appl., 287(1–3):337–357, 1999.

[32] J. van den Eshof, A. Frommer, Th. Lippert, K. Schilling, and H. A. van der Vorst.
Numerical methods for the QCDd overlap operator. I. Sign-function and error bounds.
Comp. Physics Comm., 146(2):203–224, 2002.
Title contains misprint ‘QCDd’ instead of ‘QCD’. hep-lat/0202025.

[33] Jasper van den Eshof.
Nested iteration methods for nonlinear matrix problems.
Proefschrift, Faculteit der Wiskunde en Informatica, Universiteit Utrecht, 2003.

[34] Jasper van den Eshof and Gerard L. G. Sleijpen.
Inexact Krylov subspace methods for linear systems.
SIAM J. Matrix Anal. Appl., 26(1):125–153, 2004.

[35] Jens-Peter M. Zemke.
Abstract perturbed Krylov methods.
Linear Algebra Appl., 424(2-3):405–434, 2007.