Das CG-Verfahren

Sven Wetterauer

06.07.2010

Inhaltsverzeichnis

1 Einführung

In dieser Ausarbeitung wird ein iteratives Verfahren zur Lösung eines linearen Gleichungssystems vorgestellt, die Methode der Konjugierten Gradienten, kurz das CG-Verfahren. Dabei wird nicht direkt nach der Lösung des LGS gesucht, sondern nur indirekt. Die Lösung kann mit dem Minimum einer quadratischen Form identifiziert werden und dieses Minimum wird dann mit dem CG-Verfahren bestimmt. Zuerst wird allerdings graphisch (nicht mathematisch) die Methode des steilsten Abstiegs erläutert. Anschließend wird als Hinführung zum CG-Verfahren, die Methode der konjugierten Richtungen erklärt. Das CG-Verfahren ist schießlich nur noch ein Spezialfall der Methode der konjugierten Richtungen.

2 Die quadratische Form

Zur Lösung des linearen Gleichungssystems

$$Ax = b$$

wird die quadratische Form eingeführt:

$$f(x) = \frac{1}{2}x^T Ax + b^T x + c \tag{1}$$

wobei hier A eine symmetrische, positiv definite Matrix, x und b Vektoren und c ein Skalar ist.
Die Behauptung ist nun, dass das Minimum der quadratischen Form die Lösung des linearen Gleichungssystems ist.
Für das Minimum müssen 2 Bedigungen erfüllt sein:
1. Der Gradient der Funktion muss verschwinden.

$$f'(x) = \nabla f = \begin{pmatrix} \frac{\partial}{\partial x_1} f(x) \\ \frac{\partial}{\partial x_2} f(x) \\ \vdots \\ \frac{\partial}{\partial x_n} f(x) \end{pmatrix} = 0 \tag{2}$$

2. Die Hessematrix der Funktion muss positiv definit sein.

$$H(f) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \ddots & & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix} \tag{3}$$

Der Gradient unserer quadratischen Form (1) ist gegeben durch:

$$f'(x) = \nabla f = \frac{1}{2}A^T x + \frac{1}{2}Ax - b = Ax - b \tag{4}$$

Wobei im zweiten Schritt benutzt wurde, dass die Matrix A symmetrisch ist.
Die Hesse-Matrix entspricht offensichtlich gerade der Matrix A, welche nach Voraussetzung positiv definit ist. Damit ist also gezeigt, dass die Lösung des linearen Gleichungssystems

$$Ax = b$$

gleich dem Minimum der quadratischen Form (1) ist. Im folgenden werden iterative Lösungen zum Finden des Minimums vorgestellt.

3 Methode des steilsten Abstiegs

In diesem Kapitel soll in Grundzügen die Methode des steilsten Abstiegs vorgestellt werden. Ziel ist das Minimum der quadratischen Form zu finden. Wir starten an einem beliebigen Punkt. Da wir das Minimum suchen, muss man von diesem Punkt aus offensichtlich bergab gehen. Intuitiv ist klar, dass wir uns in Richtung des steilsten Abstiegs bewegen. Um möglichst nahe an das Minimum zu gelangen, suchen wir in dieser Richtung das Minimum. Von diesem Minimum aus gehen wir wieder in Richtung des steilsten Abstiegs und kommen dadurch dem Minimum der quadratischen Funktion immer näher. Anschaulich wird dieser Vorgehen an einem Beispiel graphisch dargstellt.
Beispiel:

$$A = \begin{pmatrix} 3 & 2 \\ 2 & 6 \end{pmatrix}, \quad b = \begin{pmatrix} 2 \\ -8 \end{pmatrix}, \quad c = 0 \tag{5}$$

Das oben genannte Verfahren wird in Graph 1 dargestellt. In (a) ist die quadratische Form als Äquipotentiallinien dargestellt. $x_{(0)}$ ist unser beliebiger Startpunkt. Die durchgezogene Linie zeigt die Richtung des steilsten Abstiegs. In Graph (c) wird die quadratische Funktion in diese Richtung dargestellt. Das Minimum der Funktion in Graph (c) entspricht unserem nächsten Startpunkt $x_{(1)}$. Wiederum nähern wir uns dem Minimum der quadratischen Funktion, indem wir in Richtung des steilsten Abstiegs gehen. Die neue Bewegungsrichtung steht orthogonal auf die vorherige. Sie entspricht der negativen Richtung des Pfeils, der in (d) dargestellt ist. Zusammengefasst ist diese Methode in Graph 2 dargestellt.

Die Methode des steilsten Abstiegs wird hier nicht weitergeführt. Sie sollte nur anschaulich ein mögliches Vorgehen darstellen. Wie in Graph 2 ersichtlich ist diese Methode nicht besonders effektiv, da wir uns öfters in die selbe Richtung bewegen müssen. Es wäre deutlich besser, wenn wir in einem Schritt die beste Lösung in dieser Richtung finden und uns damit nie wieder in diese Richtung bewegen müssen. Dazu wird als nächstes die Methode der konjugierten Richtungen dargestellt.

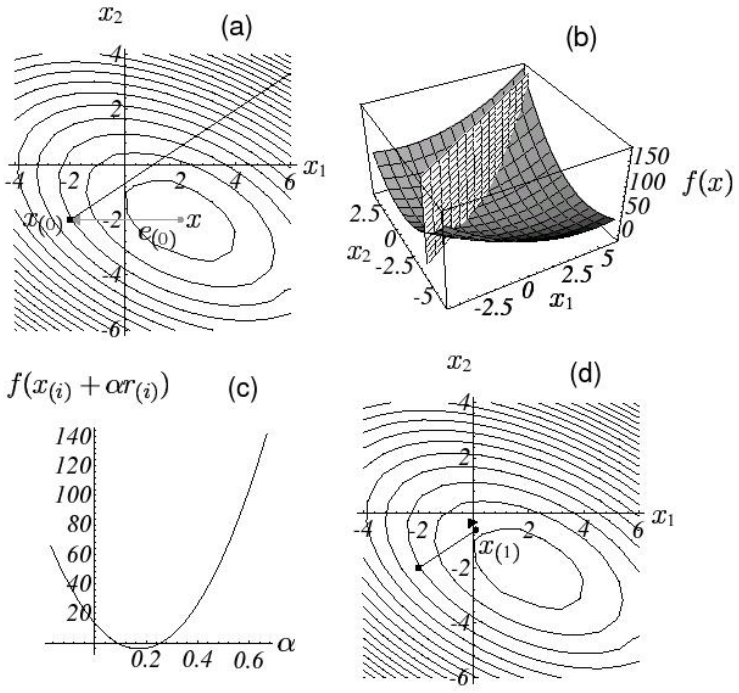

Abbildung 1: Vorgehen der Methode des steilsten Abstiegs

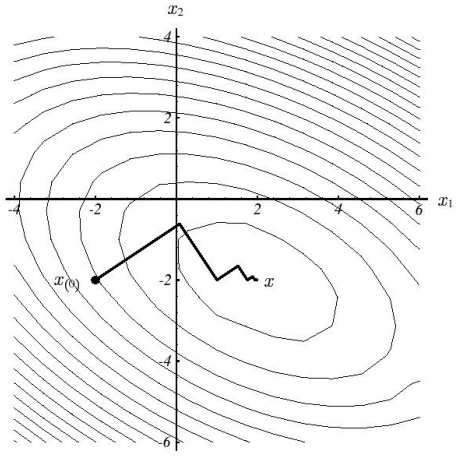

Abbildung 2: mehrmaliges Anwenden des steilsten Abstiegs

4 Methode der Konjugierten Richtungen

Diese Vorgehensweise wird nun mathematisch exakt hergeleitet und dargstellt. Bevor wir dies tun, werden allerdings noch mehrere Größen definiert:

$$e_{(i)} = x_{(i)} - x, \quad r_{(i)} = b - Ax_{(i)} = -Ae_{(i)} = -f'(x_{(i)}) \tag{6}$$

Das bedeutet, dass der Fehlervektor $e_{(i)}$ gerade unsere Entfernung zu der exakten Lösung x beschreibt, also dem Fehler unserer momentanen Lösung entspricht. Das Residuum $r_{(i)}$ entpricht dem Fehler, der von A in denselben Raum wie b transformiert wird. Aufgrund des Zusammenhanges $r_{(i)} = -f'(x_{(i)})$ entspricht das Residuum auch gerade der Richtung des steilsten Abstiegs.

Zunächst wählen wir uns eine Menge von orthogonalen Suchrichtungen $d_{(0)}$, $d_{(1)}, ..., d_{(n-1)}$. Von unserem Startpunkt aus gehen wir nacheinander in alle Suchrichtungen, und zwar gerade soweit, dass wir in dieser Richtung auf einer Ebene mit der exakten Lösung x liegen, so dass wir in jede Richtung nur einmal suchen müssen. Dies bedeutet auch automatisch, dass wir nach n Schritten unser Minimum exkat bestimmt hätten. In Graph 3 ist dieses Vorgehen graphisch aufgezeigt. Es ist offensichtlich, dass die erste Suchrichtung $d_{(0)}$ orthogonal zu $e_{(1)}$ steht. Diese Tatsache kann man auf alle weiteren Suchrichtungen verallgemeinern $d_{(i)}^T e_{(j)} = 0$ für i<j.

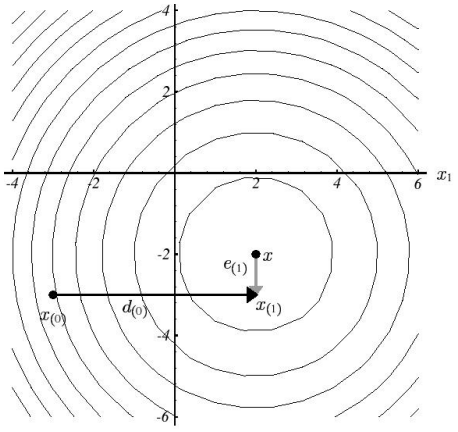

Abbildung 3: Methode der konjugierten Richtungen

Um unseren neuen Punkt $x_{(i+1)}$ zu finden, gehen wird von $x_{(i)}$ aus in Richtung von unserer Suchrichtung $d_{(i)}$

$$x_{(i+1)} = x_{(i)} + \alpha_{(i)} d_{(i)} \tag{7}$$

Um den neuen Punkt genau zu bestimmen, benötigen wir allerdings noch $\alpha_{(i)}$. Diese Gleichung ist äquivalent zu der Gleichung:

$$e_{(i+1)} = e_{(i)} + \alpha_{(i)} d_{(i)} = e_{(0)} + \sum_{j=0}^{i} \alpha_{(j)} d_{(j)} \tag{8}$$

Um den genauen Wert von $\alpha_{(i)}$ zu bestimmen, nutzen wir oben genannte Tatsache.

$$d_{(i)}^T e_{(i+1)} = 0$$

$$d_{(i)}^T (e_{(i)} + \alpha_{(i)} d_{(i)}) = 0$$

$$\alpha_{(i)} = -\frac{d_{(i)}^T e_{(i)}}{d_{(i)}^T d_{(i)}} \tag{9}$$

Allerdings haben wir damit nicht neues erreicht. Um $\alpha_{(i)}$ zu berechnen benötigen wir $e_{(i)}$. Allerdings würden wir die exakte Lösung schon kennen, wenn $e_{(i)}$ bekannt wäre.

Wir versuchen eine etwas andere Herangehensweise an das Problem. Anstatt die Suchrichtungen orthogonal zu wählen, wählen wir diese A-orthogonal. Was A-Orthogonalität bedeutet wird in Graph 4 deutlich. In (a) sind einige A-orthogonale Vektoren (als Pfeile) dargestellt. Wenn wir uns vorstellen, das Papier wäre aus Gummi und wir könnten den Graphen so dehnen, dass die Äquipotentiallinien konzentrisch erscheinen, würde der Graph wie (b) aussehen. Es ist offensichtlich, dass die Vektoren hier orthogonal stehen.

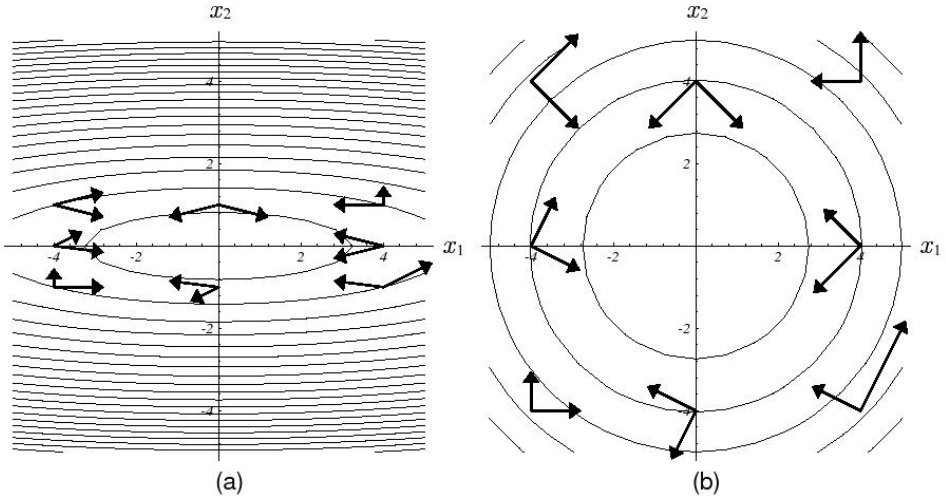

Abbildung 4: A-Orthogonalität

Mathematisch gesehen wird A-Orthogonalität so ausgedrückt:
Zwei Vektoren $d_{(i)}$ und $d_{(j)}$ heißen A-orthogonal, oder Konjugiert, wenn sie die Gleichung

$$d_{(i)}^T A d_{(j)} = 0, \quad i \neq j \tag{10}$$

erfüllen. Probieren wir aus, ob wir damit unser Problem lösen können. Wir wollen $\alpha_{(i)}$ in Gleichung (7) bestimmen. Aufgrund der A-Orthogonalität fordern wir jetzt, dass

$$d_{(i)}^T A e_{(i+1)} = 0 \tag{11}$$

Analog zu Gleichung (9) ergibt sich damit:

$$\alpha_{(i)} = -\frac{d_{(i)}^T A e_{(i)}}{d_{(i)}^T A d_{(i)}} = \frac{d_{(i)}^T r_{(i)}}{d_{(i)}^T A d_{(i)}} \tag{12}$$

Da uns der exakte Wert von b in unserer quadratischen Form, und damit auch das Residuum, bekannt ist, können wir diese Gleichung direkt lösen.

Die Frage ist jetzt allerdings, ob A-orthogonale Suchrichtungen unser Problem auch in n Schritten lösen, wie es orthogonale Suchrichtungen tun würden.
Es ist sehr einfach zu beweisen, dass die n A-orthogonalen Suchrichtungen linear unabhängig sind. Daher bilden die Suchrichtungen eine Basis unseres n-dimensionalen Problems.
Also können wir unseren Fehlervektor $e_{(0)}$ als Linearkombination von den Suchrichtungen $d_{(i)}$ darstellen.

$$e_{(0)} = \sum_{j=0}^{n-1} \delta_{(j)} d_{(j)} \tag{13}$$

Die genauen Werte von $\delta_{(j)}$ können mit einem mathematischen Trick bestimmt werden. Wir multiplizieren beide Seiten der Gleichung (13) mit $d_{(k)}^T A$ und nutzen die A-Orthogonalität der d-Vektoren aus. Damit ergibt sich:

$$d_{(k)}^T A e_{(0)} = \sum_{j=0}^{n-1} \delta_{(j)} d_{(k)}^T A d_{(j)}$$

$$d_{(k)}^T A e_{(0)} = \delta_{(k)} d_{(k)}^T A d_{(k)}$$

$$\delta_{(k)} = \frac{d_{(k)}^T A e_{(0)}}{d_{(k)}^T A d_{(k)}} = \frac{d_{(k)}^T A (e_{(0)} + \sum_{i=0}^{k-1} \alpha_{(i)} d_{(i)})}{d_{(k)}^T A d_{(k)}} = \frac{d_{(k)}^T A e_{(k)}}{d_{(k)}^T A d_{(k)}} \tag{14}$$

In Gleichung (14) wurde zunächst die A-Orthogonalität ausgenutzt, um mit 0 zu addieren. In der zweiten Umformung wurde Gleichung (8) benutzt.
Im Vergleich mit der Gleichung (12) fällt auf, dass der Zusammenhang $\alpha_{(i)} = -\delta_{(i)}$ gilt. Damit können wir den Fehler e auf eine neue Art betrachten. Die Tatsache, dass

wir die exakte Lösung x Komponente um Komponente ausgehend von einem beliebigen Startvektor $x_{(0)}$ aufbauen, ist gleichbedeutend damit, dass wir den Fehlervektor $e_{(0)}$ Komponente um Komponente abbauen.

$$e_{(i)} = e_{(0)} + \sum_{j=0}^{i-1} \alpha_{(j)} d_{(j)} = \sum_{j=0}^{n-1} \delta_{(j)} d_{(j)} - \sum_{j=0}^{i-1} \delta_{(j)} d_{(j)} = \sum_{j=i}^{n-1} \delta_{(j)} d_{(j)} \tag{15}$$

Das bedeutet, dass nach n Schritten der Fehler $e_{(n)} = 0$ ist. Dies wiederum bedeutet, dass wir den exakten Wert unseres Minimums der quadratischen Form gefunden haben.

4.1 Gram-Schmidt-Konjugation

Das einzige, was uns jetzt noch fehlt, sind die A-orthogonalen Suchvektoren $d_{(0)},\ d_{(1)}, ...,\ d_{(n-1)}$. Um diese Vektoren zu finden, gibt es ein einfaches Verfahren, die sogenannte Gram-Schmidt-Konjugation. Dazu nehmen wir uns zunächst eine Menge von n beliebigen linear unabhängigen Vektoren $u_0,\ u_1, ...,\ u_{n-1}$. Die Koordinatenachsen erfüllen diese Bedingung, wobei es geschicktere Wahlen gibt. Bei der Gram-Schmidt-Konjugation nehmen wir uns einfach u_i und subtrahieren alle Komponenten, die nicht A-orthogonal zu den Vektoren $d_{(0)}, ..., d_{(i-1)}$ sind.
Wir beginnen damit, dass wir $d_{(0)} = u_0$ wählen.
Für i=1,...,n-1 wählen wir

$$d_{(i)} = u_i + \sum_{k=0}^{i-1} \beta_{ik} d_{(k)}, \tag{16}$$

wobei β_{ik} nur für $i > k$ definiert sind. Die Koeffizienten β_{ij} werden durch die Forderung gewonnen, dass die $d_{(i)}$ A-orthogonal sein sollen. Wir benutzen den gleichen Trick, den wir schon bei der Berechnung von $\delta_{(j)}$ benutzt haben, wir multiplizieren beide Seiten der Gleichung (16) mit $Ad_{(j)}$, wobei $i > j$ gesetzt wird:

$$d_{(i)}^T A d_{(j)} = u_i^T A d_{(j)} + \sum_{k=0}^{i-1} \beta_{ik} d_{(k)}^T A d_{(j)}$$

$$0 = u_i^T A d_{(j)} + \beta_{ij} d_{(j)}^T A d_{(j)}$$

Nach β_{ij} auflösen:

$$\beta_{ij} = -\frac{u_i^T A d_j}{d_{(j)}^T A d_{(j)}} \tag{17}$$

Dieses Verfahren hat allerdings den großen Nachteil, dass wir alle $u_{(i)}$ speichern müssen, da wir diese bis zum Schlußzum berechnen der β_{ij} benötigen. Mit diesen Vorraussetzungen können wir nun die Methode der konjugierten Gradienten herleiten.

5 Methode des konjugierten Gradienten

Um die Methode des konjugierten Gradienten aus den konjugierten Richtungen zu erhalten, wählen wir $u_i = r_{(i)}$.
An dieser Stelle wollen wir noch 2 Eigenschaften der Residuen festhalten. Dazu nehmen wir uns noch einmal Gleichung (15) vor und multiplizieren sie mit $-d_{(i)}^T A$.

$$-d_{(i)}^T A e_{(j)} = -\sum_{k=j}^{n-1} \delta_{(k)} d_{(i)}^T A d_{(k)} \tag{18}$$

Für den Fall, dass i < j und nach ausnutzen der A-Orthogonalität ist diese Gleichung äquivalent zu

$$d_{(i)}^T r_{(j)} = 0 \tag{19}$$

Das bedeutet, dass das Residuum r_j orthogonal zu allen vorherigen Suchvektoren $d_{(0)}, d_{(1)}, ...d_{(j-1)}$ ist.
Da die Suchvektoren $d_{(i)}$ aus den Residuen zusammengebaut werden, Gleichung(16) mit $u_{(i)} = r_{(i)}$, muss gelten: $span\{r_{(0)}, r_{(1)}, ..., r_{(i-1)}\} = span\{d_{(0)}, d_{(1)}, ..., d_{(i-1)}\}$. Allerdings ist auch jedes Residuum orthogonal zu den vorherigen Suchvektoren, daher müssen die Residuen orthogonal aufeinander stehen:

$$r_{(i)}^T r_{(j)} = 0, \quad i \neq j \tag{20}$$

In Gleichnung (17) haben wir den Zusammenhang $\beta_{ij} = -\frac{r_{(i)}^T A d_{(j)}}{d_{(j)}^T A d_{(j)}}$ gefunden, wobei $i > j$ gilt. Durch unsere Wahl der u_i können wir diesen Zusammenhang etwas vereinfachen.

$$r_{(i)}^T r_{(j+1)} = r_{(i)}^T r_{(j)} - \alpha_{(j)} r_{(i)}^T A d_{(j)}$$

$$\alpha_{(j)} r_{(i)}^T A d_{(j)} = r_{(i)}^T r_{(j)} - r_{(i)}^T r_{(j+1)}$$

$$r_{(i)}^T A d_{(j)} = \begin{cases} \frac{1}{\alpha_{(i)}} r_{(i)}^T r_{(i)}, & i = j, \\ -\frac{1}{\alpha_{(i-1)}} r_{(i)}^T r_{(i)}; & i = j+1, \\ 0, & sonst \end{cases}$$

Damit vereinfacht sich dann β_{ij}:

$$\beta_{ij} = \begin{cases} \frac{1}{\alpha_{(i-1)}} \frac{r_{(i)}^T r_{(i)}}{d_{(i-1)}^T A d_{(i-1)}}, & i = j+1, \\ 0, & i > j+1 \end{cases} \tag{21}$$

Offensichtlich sind die meisten Terme β_{ij} verschwunden. Die einzigen, die uns noch einen Beitrag bringen, sind diejenige, für die i=j+1 gilt. Daher werden wir ab sofort zur Vereinfachung die Notation $\beta_{(i)} = \beta_{i,i-1}$ verwenden und vereinfachen noch weiter:

$$\beta_{(i)} = \frac{r_{(i)}^T r_{(i)}}{d_{(i-1)}^T r_{(i-1)}} \qquad aus\ Gleichung\ (12)$$

Wir haben oben schon den Zusammenhang $d_{(i)}^T r_{(j)} = 0$ für i<j hergeleitet. Was passiert mit dieser Gleichung allerdings im Fall j=i?

$$d_{(i)}^T r_{(i)} = u_{(i)}^T r_{(i)} + \sum_{k=0}^{i-1} \beta_{ik} d_{(k)}^T r_{(i)} \quad \textit{aus Gleichung (16)}$$

Durch Ausnutzen der Gleichung (19) verschwindet die hintere Summe und es bleibt übrig:

$$d_{(i)}^T r_{(i)} = u_{(i)}^T r_{(i)} \tag{22}$$

Diese Gleichung können wir verwenden, um unser $\beta_{(i)}$ weiter zu vereinfachen:

$$\beta_{(i)} = \frac{r_{(i)}^T r_{(i)}}{r_{(i-1)}^T r_{(i-1)}} \tag{23}$$

Das Problem, das oben erwähnt wurde (Speicherung der $u_{(i)}$), ist hier auch gelöst. Im i-ten Schritt benötigen wir nur noch $r_{(i-1)}, r_{(i)}$ zur Berechnung von $\beta_{(i)}$.
Damit haben wir genügend Vorarbeit geleistet. Wenn wir alles Zusammenfassen erhalten wir die Methode des konjugierten Gradienten:

$$d_{(0)} = r_{(0)} = b - Ax_{(0)} \tag{24}$$

$$\alpha_{(i)} = \frac{r_{(i)}^T r_{(i)}}{d_{(i)}^T A d_{(i)}} \tag{25}$$

$$x_{(i+1)} = x_{(i)} + \alpha_{(i)} d_{(i)} \tag{26}$$

$$r_{(i+1)} = r_{(i)} - \alpha_{(i)} A d_{(i)} \tag{27}$$

$$\beta_{(i+1)} = \frac{r_{(i+1)}^T r_{(i+1)}}{r_{(i)}^T r_{(i)}} \tag{28}$$

$$d_{(i+1)} = r_{(i+1)} + \beta_{(i+1)} d_{(i)} \tag{29}$$

5.1 Konvergenzanalyse

Wir haben weiter oben bereits gezeigt, dass das CG-Verfahren nach n Schritten die exakte Lösung x berechnet. Also kann man sich ersteinmal die Frage stellen, warum hier überhaupt eine Konvergenzanalyse durchgeführt wird.
Wir sind in der kompletten Herleitung des CG-Verfahrens immer von exakter Artihmetik ausgegangen. Durch Rundungsfehler kann es passieren, dass unsere Suchvektoren $d_{(0)}, ..., d_{(n-1)}$ die A-Orthogonalität verlieren, welche aber essentiell für die Konvergenz in n Schritten ist. Dadurch kann es passieren, dass das CG-Verfahren nicht nach dem n-ten Schritt mit der exakten Lösung abbricht, sondern sich dieser immer nur mehr annähert. Daher wollen wir im folgenden von gerundeter Artihmetik ausgehen, und damit das Konvergenzverhalten des CG-Verfahrens untersuchen.
Dazu wird zunächt ein Hilfssatz eingeführt und bewiesen:

Hilfssatz
Für ein Plynom $p \in P_i$ mit $p(0) = 1$ gelte auf einer Menge $S \subset \mathbb{R}$, welche alle Eigenwerte von A enthält,

$$\sup_{\mu \in S} |p(\mu)| \leq M$$

Dann gilt:

$$||x_{(i)} - x||_A \leq M||x_{(0)} - x||_A \tag{30}$$

Beweis:
Nach Konstruktion des CG-Verfahrens gilt offensichtlich:

$$D_i := span\{d_{(0)}, ..., d_{(i-1)}\} = span\{A^0 r_{(0)}, A^1 r_{(0)}, ..., A^{i-1} r_{(0)}\}$$

Außerdem gilt (ohne Beweis):

$$||x_{(i)} - x||_A = \min_{y \in x_{(0)} + D_i} ||y - x||_A$$

Das heißt y kann geschrieben werden, als:

$$y = x_{(0)} + \sum_{k=0}^{i-1} \eta_k d_{(k)} = x_{(0)} + \sum_{k=0}^{i-1} \eta'_k A^k r_{(0)}$$

Offensichtlich wird nur über die Koeffizienten η_k und die Matrix A summiert. Diese Summe kann auch als ein Polynom vom Grad i-1 aufgefasst werden, wobei das Argument eine Matrix ist.

$$p(A) = \sum_{k=0}^{i-1} \eta'_k A^k$$

Wenden wir dies an:

$$||x_{(i)} - x||_A = \min_{p \in P_{i-1}} ||x_{(0)} - x + p(A) r_{(0)}||_A = \min_{p \in P_{i-1}} ||x_{(0)} - x + p(A)A(x_{(0)} - x)||_A$$

$$= \min_{p \in P_{i-1}} ||[I + p(A)A](x_{(0)} - x)||_A$$

$I + p(A) \cdot A$ kann wiederum als neues Polynom vom Grad i aufgefasst werden, wobei gelten muss, dass $p(0) = 1$. (Falls das Argument eine Matrix ist, dann wird 1 zu der Einheitsmatrix I.)

$$||x_{(i)} - x||_A = \min_{p \in P_i, p(0)=1} ||p(A)(x_{(0)} - x)||_A \leq \min_{p \in P_i, p(0)=1} ||p(A)||_A ||x_{(0)} - x||_A \tag{31}$$

Da A symmetrisch positiv definit ist, existiert eine Orthonormalbasis aus Eigenvektoren $\{v_{(0)}, ...v_{(n-1)}\}$. Das heißt wir können jeden Vektor y als Linearkombination der Eigenvektoren darstellen.

$$y = \sum_{k=0}^{n-1} \xi_k v_{(k)}$$

Für Eigenvektoren hat das Matrixpolynom von oben auch eine sehr intressante Eigenschaft:

$$p(A)v_{(j)} = \sum_{k=0}^{i-1} \eta_k A^k v_{(j)} = \sum_{k=0}^{i-1} \eta_k \lambda_j^k v_{(j)} = p(\lambda_j)v_{(j)}$$

Hier ist λ_j der Eigenwert zum Eigenvektor $v_{(j)}$.
Damit gilt:

$$p(A)y = p(A)\sum_{k=0}^{n-1} \xi_k v_{(k)} = \sum_{k=0}^{n-1} \xi_k p(\lambda_k) v_{(k)}$$

und damit für die Norm:

$$||p(A)y||_A^2 = ||\sum_{k=0}^{n-1} \xi_k p(\lambda_k) v_{(k)}||_A^2 = \sum_{k=0}^{n-1} \xi_k^2 p(\lambda_k)^2 v_{(k)}^T A v_{(k)} = \sum_{k=0}^{n-1} \xi_k^2 p(\lambda_k)^2 \lambda_k$$

$$\leq M^2 \sum_{k=0}^{n-1} \lambda_k \xi_k^2 = M^2 ||y||_A^2$$

Damit erhalten wir den Zusammenhang:

$$||p(A)||_A = \sup_{y \in \mathbb{R}^n, y \neq 0} \frac{||p(A)y||_A}{||y||_A} \leq M$$

Wenn wir diese Gleichung in (31) einsetzen erhalten wir die Behauptung.

Mit diesem Hilfssatz kommen wir jetzt zur Konvergenz des CG-Verfahrens:

Satz
Für das CG-Verfahren gilt die Fehlerabschätzung:

$$||x_{(i)} - x||_A \leq 2(\frac{1 - 1/\sqrt{\kappa}}{1 + 1/\sqrt{\kappa}})^i ||x_{(0)} - x||_A,$$

wobei $i \in \mathbb{N}$ und $\kappa = cond_2(A) = \frac{\Lambda}{\lambda}$ die Spektralkonditionszahl von A ist. Λ beschreibt dabei den größten Eigenwert, λ den kleinsten. Zur Reduzierung des Anfangsfehlers um den Faktor ε sind höchstens

$$i(\varepsilon) \leq \frac{1}{2}\sqrt{\kappa} ln(\frac{2}{\varepsilon}) + 1$$

Iterationsschritte nötig.
Beweis:
Wir setzen in unserem Hilfssatz $S = [\lambda, \Lambda]$, damit folgt:

$$||x_{(i)} - x||_A \leq \min_{p \in P_i, p(0)=1} \{ \sup_{\lambda \leq \mu \leq \Lambda} |p(\mu)| \} ||x_{(0)} - x||_A$$

Daraus erhalten wir direkt die Behauptung wenn wir zeigen können, dass

$$\min_{p\in P_i, p(0)=1}\{\sup_{\lambda\leq\mu\leq\Lambda}|p(\mu)|\}\leq 2(\frac{1-1/\kappa}{1+1/\kappa})^i$$

Solche Probleme werden von einem Tschebyscheff-Polynom gelöst.(vgl. Numerik 0-Vorlesung, Rannacherskript).
Ein Tschebyscheff-Polynom $T_i(x)$ löst dieses Problem allerdings nur auf dem Intervall $[-1,1]$ Daher müssen wir das Polynom an unser Intervall $[\lambda,\Lambda]$ anpassen:

$$T_i(\mu)\rightsquigarrow T_i(\frac{\Lambda+\lambda-2\mu}{\Lambda-\lambda})$$

Allerdings fordern wir an unser Polynom $\overline{p}(\mu)$, dass gilt: $\overline{p}(0)=1$. Damit erhalten wir also:

$$\overline{p}(\mu)=T_i(\frac{\Lambda+\lambda-2\mu}{\Lambda-\lambda})(T_i(\frac{\Lambda+\lambda}{\Lambda-\lambda}))^{-1}$$

Der erste Faktor in diesem Produkt kann nur zwischen ±1 oszillieren. Damit gilt also für das Supremum:

$$\sup_{\lambda\leq\mu\Lambda}\overline{p}(\mu)=(T_i(\frac{\Lambda+\lambda}{\Lambda-\lambda}))^{-1}$$

Aus der allgemeinen Darstellung des Tschebyscheff-Polynoms

$$T_i(\mu)=\frac{1}{2}[(\mu+\sqrt{\mu^2-1})^i+(\mu-\sqrt{\mu^2-1})^i]$$

mit den Identitäten

$$\frac{\kappa+1}{\kappa-1}+\sqrt{(\frac{\kappa+1}{\kappa-1})^2-1}=\frac{\kappa+1}{\kappa-1}+\frac{2\sqrt{\kappa}}{\kappa-1}=\frac{(\sqrt{\kappa}+1)^2}{\kappa-1}=\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}$$

$$\frac{\kappa+1}{\kappa-1}-\sqrt{(\frac{\kappa+1}{\kappa-1})^2-1}=\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$$

folgt:

$$T_i(\frac{\Lambda+\lambda}{\Lambda-\lambda})=T_i(\frac{\kappa+1}{\kappa-1})=\frac{1}{2}[(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1})^i+(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1})^i]\geq\frac{1}{2}(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1})^i$$

Also gilt:

$$\sup_{\lambda\leq\mu\leq\Lambda}\overline{p}(\mu)\leq 2(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1})^i$$

Woraus die erste Behauptung folgt.
Für die zweite Behauptung fordern wir, dass

$$2(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1})^i=\varepsilon$$

und formen dies nach i um.

$$i(\varepsilon) = ln(\frac{2}{\varepsilon})(ln(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}))^{-1}$$

Für den Logarithmus gilt die Summendarstellung:

$$ln(x) = \sum_{k=0}^{\infty} \frac{2}{2k+1}(\frac{x-1}{x+1})^{2k+1}$$

und somit:

$$ln(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}) = 2\sum_{k=0}^{\infty} \frac{1}{2k+1}(\frac{1}{x})^{2k+1}$$

$$= 2[\frac{1}{x} + \frac{1}{3x^3} + \frac{1}{5x^5} + ...] \geq \frac{2}{x}$$

Damit erhalten wir:

$$i(\varepsilon) \leq \frac{1}{2}\sqrt{\kappa}ln(\frac{2}{\varepsilon})$$

Womit die zweite Behauptung bewiesen wäre.

5.2 Vergleich mit Richardsoniteration

Die Richardsonitaration ist eine Iteration gemäßder Vorschrift:

$$x_(i+1) = (I_\omega A)x_{(i)} + \omega b$$

Für das KOnvergenzverhalten gilt damit im symmetrisch positiv definiten Fall:

$$||x_{(i)} - x| \leq (\frac{\kappa-1}{\kappa+1})^i ||x_{(0)} - x||.$$

Um den Fehler um den Faktor ε zu vermindern werden

$$i(\varepsilon) \geq \frac{1}{2}\kappa ln(\frac{1}{\varepsilon})$$

benötigt. Das heißt wir hier $O(\kappa)$ Schritte, wohingegen beim CG-Verfahren $O(\sqrt{\kappa})$ nötig sind.

5.3 Vorteile des CG-Verfahrens

- eigent sich gut für dünn besetzte Matrizen
- hauptsächlicher Aufwand besteht aus Matrix-Vektor-Mulitplikation
- Aufwand O(m), wobei m die Anzahl der nichtnegativen Einträge von A sind
- liefert für sehr große Systeme schon in weniger als n Schritten gute Näherungen der exakten Lösung