3. Iterative Methoden für lineare Gleichungssysteme

(1) **Problemstellung:** $A\underline{x} = \underline{b}$

$A \in \mathbb{R}^{n \times n}$, regulär

$\underline{b} \in \mathbb{R}^n$

Speziell: n sehr gross, z.B. $n = 10'000$

A sparse (schwach besetzt)

Gesucht: Lösung $\underline{x} = A^{-1}\underline{b}$, ev. approximativ unter Ausnutzung der Sparsity: Matrix A (ein Unding!) nicht aufschreiben, nicht speichern. Nur die Abbildung $\underline{x} \mapsto A\underline{x}$ brauchen!

3.1. Die Jacobi-Iteration

Sei $A = C - B$ eine additive Zerlegung von A

C so, dass Gl'systeme mit C einfach lösbar

(2) Aus (1) folgt $C\underline{x} = \underline{b} + B\underline{x}$

eine Vektorgleichung in Fixpunktform; lösen durch Iteration:

Startvektor $\underline{x}_0$, z.B. $\underline{x}_0 = 0$

(3) $\underline{x}_{k+1}$ aus $C\underline{x}_{k+1} = \underline{b} + B\underline{x}_k$, $k = 0, 1, \ldots$

einfach lösbar!

Fehlertheorie

Subtrahiere (2) von (3):

$$C(\underline{x}_{k+1} - \underline{x}) = B(\underline{x}_k - \underline{x})$$

Def: Fehlervektor $\underline{e}_k := \underline{x}_k - \underline{x}$

Fehlergesetz: $C\,\underline{e}_{k+1} = B\,\underline{e}_k$

oder $\underline{e}_{k+1} = (C^{-1}B)\,\underline{e}_k$

Normen:

$$\boxed{\|e_{k+1}\| \leq \|C^{-1}B\| \cdot \|e_k\|}$$

SATZ: Hinreichend für die Konvergenz des Iterationsverfahrens $C\,\underline{x}_{k+1} = \underline{b} + B\,\underline{x}_k$ zur Lösung von $(C-B)\,\underline{x} = \underline{b}$ ist $\|C^{-1}B\| < 1$.

Beispiele:

(i) Jacobi-Iteration. $A = (a_{jk})$

Sei A "diagonal dominant", d.h.

(4) $$|a_{jj}| > \sum_{k \neq j} |a_{jk}|, \quad j = 1, 2, \cdots, n.$$

Wähle $C := \mathrm{diag}(a_{jj})$

Aus (4) folgt sofort $\|C^{-1}B\|_\infty < 1$

(siehe S. 21)

(ii) Gauss-Seidel-Iteration

$A = \left(\begin{array}{|cc} C & -B \\ & \diagdown \end{array} \right)$

häufig leicht besser als Jacobi

3.2. Das Verfahren der konjugierten Gradienten

cg ; M. Hestenes, E. Stiefel 1951/52

Sei $A \in \mathbb{R}^{n \times n}$ symmetrisch, $A^T = A$, und positiv definit, d.h. die zu A gehörige quadratische Form erfüllt

$$Q(\underline{x}) := \underline{x}^T A \underline{x} > 0 \quad \forall\, \underline{x} \neq 0$$

Bemerkungen.

(i) A hat lauter positive Eigenwerte

Beweis. Sei λ Ew. von A, $\underline{v}$ zugehöriger Ev.

$$A\underline{v} = \underline{v} \cdot \lambda \quad \Rightarrow \quad 0 < \underline{v}^T A \underline{v} = \underbrace{\underline{v}^T \underline{v}}_{\|\underline{v}\|^2 > 0} \cdot \lambda$$

$$\Rightarrow \lambda > 0 \text{ , q.e.d.}$$

(ii) Graphische Darstellungen von $Q(\underline{x}) = \underline{x}^T A \underline{x}$ für $\underline{x} = (x_1, x_2)^T \in \mathbb{R}^2$ siehe S. 50.

Fälle: A positiv definit, pos. semidefinit, indefinit

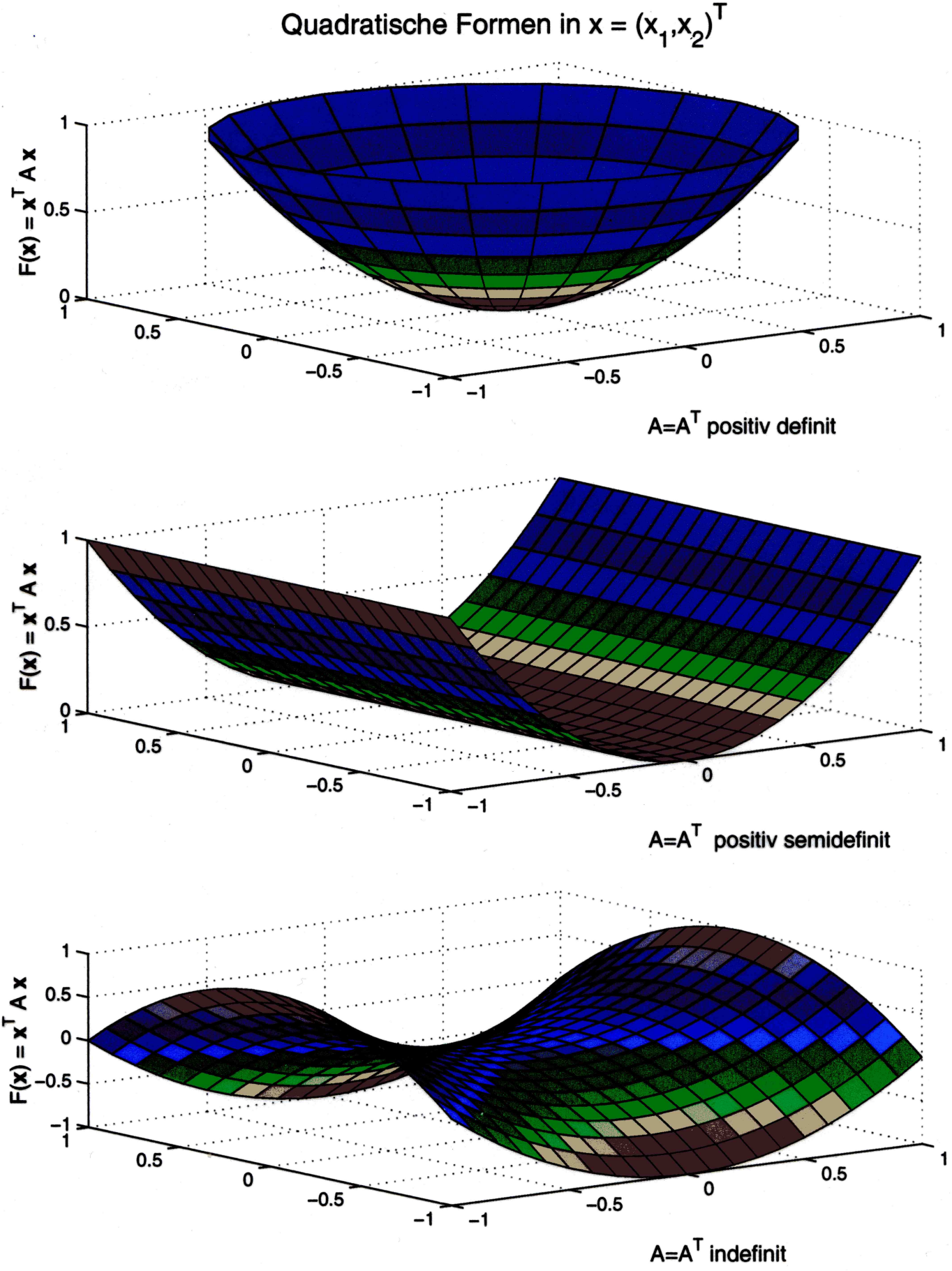
Problemstellung

Löse lineares Gleichungssystem

$$A\underline{x} + \underline{b} = 0,$$

$A^T = A$, pos. def., sparse, $\underline{x}, \underline{b} \in \mathbb{R}^n$

Bem: Differenzenverfahren für elliptische part. DGl. führen häufig auf solche Gl'systeme.


Quadratische Formen in x = (x1,x2)T
F(x) = xT A x
1
0.5
0
−0.5
−1
A=AT positiv definit
A=AT positiv semidefinit
A=AT indefinit

SATZ: $A\underline{x} + \underline{b} = 0$, A sym. pos. def.

ist äquivalent mit

$$F(\underline{u}) := \frac{1}{2}\underline{u}^T A \underline{u} + \underline{u}^T \cdot \underline{b}$$

hat in $\underline{x}$ ein Minimum.

Beweis.

(i) Differentiation ergibt

$$\text{grad}\, F = \nabla F = A\underline{u} + \underline{b} =: \underline{r} = \text{Residuum}$$

Damit folgt "$\Leftarrow$" sowie "$\underline{x}$ ist stationärer Punkt".

(ii) Minimaleigenschaft:

Betrachte $\underline{v} \in \mathbb{R}^n$, $\underline{v} \neq 0$:

$$F(\underline{x} + \underline{v}) = \frac{1}{2}(\underline{x}+\underline{v})^T A(\underline{x}+\underline{v}) + (\underline{x}+\underline{v})^T \underline{b}$$

$$= F(\underline{x}) + \underbrace{\frac{1}{2}\underline{x}^T A \underline{v}}_{= \frac{1}{2}\underline{v}^T A \underline{x} \text{ wegen } A^T = A} + \frac{1}{2}\underline{v}^T A \underline{x} + \frac{1}{2}\underline{v}^T A \underline{v} + \underline{v}^T \underline{b}$$

$$= F(\underline{x}) + \underline{v}^T \underbrace{(A\underline{x} + \underline{b})}_{0} + \frac{1}{2}\underline{v}^T A \underline{v} > F(\underline{x})$$

q.e.d.

Grundidee: Minimiere F! $\rightarrow$

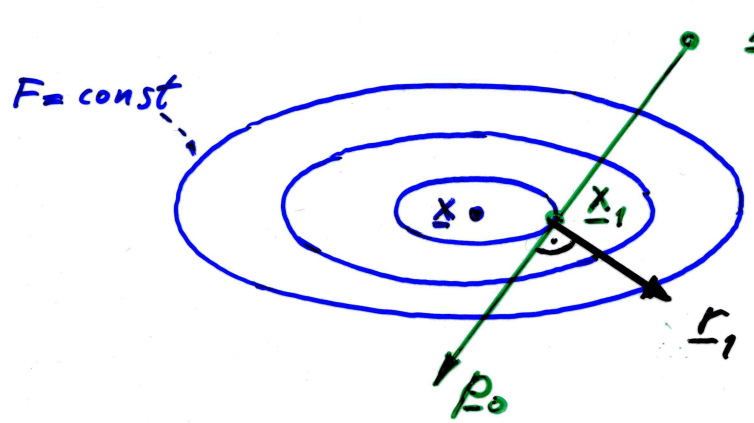
Dies löst automatisch $A\underline{x} + \underline{b} = 0$

Vorbereitung: 1-dimensionale Minimierung

Gegeben: Startpunkt $\underline{x}_0$, Suchrichtung $\underline{p}_0$

Gesucht: $\rho = \rho_0 \in \mathbb{R}$ so, dass

$$F(\underline{x}_0 + \rho\, \underline{p}_0) \overset{!}{=} \min$$



$$x_1 = x_0 + \rho_0 p_0$$

$$r_1 = A x_1 + b$$

Notwendige Bedingung:

$$\frac{d}{d\rho}\left[\frac{1}{2}(x_0^T + \rho p_0^T) A (x_0 + \rho p_0) + (x_0^T + \rho p_0^T) b\right] = 0$$

$$(*) \quad \Rightarrow \quad \rho (p_0^T A p_0) + p_0^T \underbrace{(A x_0 + b)}_{r_0} = 0$$

$$\Rightarrow \quad \rho_0 = -\frac{p_0^T r_0}{p_0^T A p_0}, \qquad x_1 = x_0 + \rho_0 p_0$$

Strategien

(a) Steilster Abstieg

Wähle $p_k := -r_k = -\nabla F(x_k)$, $k \geq 0$

Algorithmus: Gegeben $x_0 \in \mathbb{R}^n$

for $k = 0, 1, \ldots$

$$r_k = A x_k + b$$

$$\rho_k = \frac{r_k^T r_k}{r_k^T A r_k}$$

$$x_{k+1} = x_k - \rho_k r_k$$

Leider: In $n > 2$ Dimensionen viel zu langsame Konvergenz!

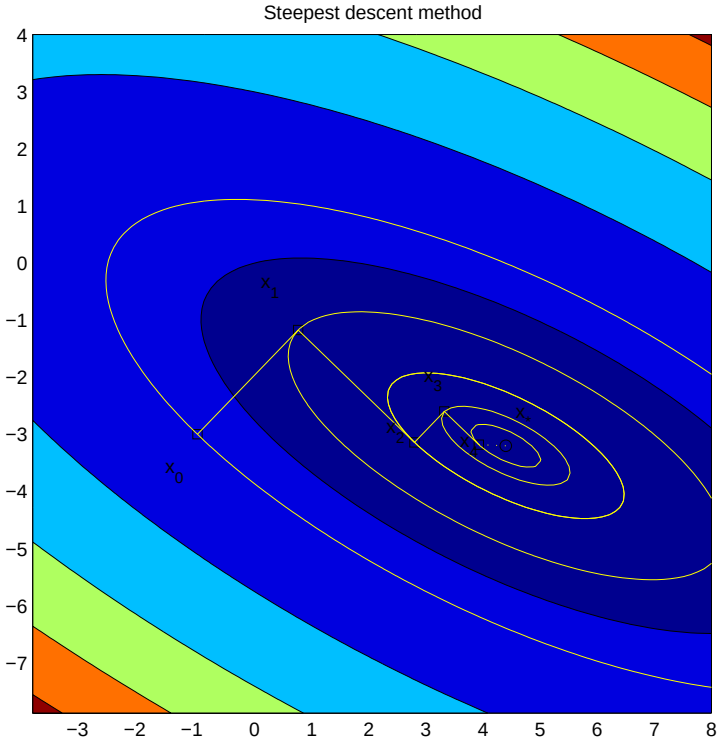


Figure 1: The steepest descent method for $N = 2$.

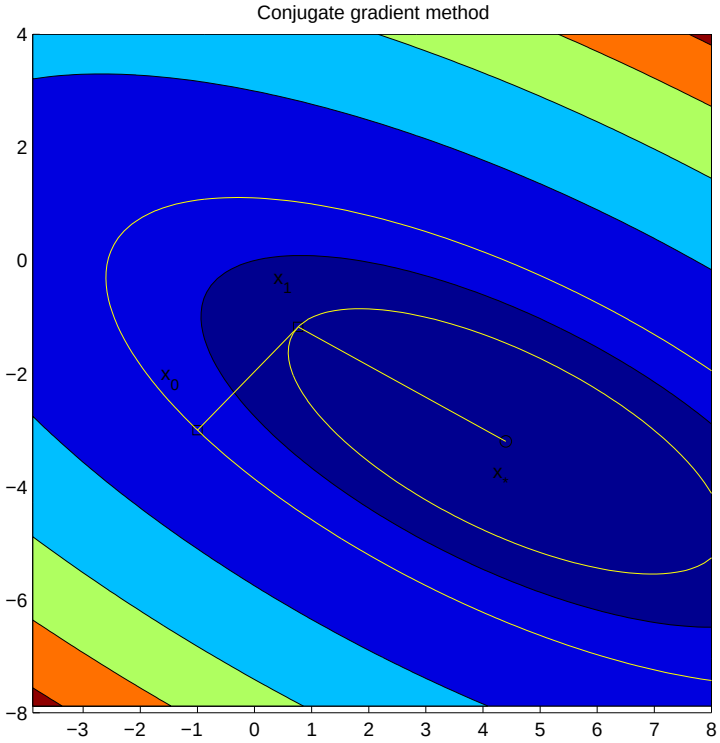


Figure 2: Conjugate directions — the CG method for $N = 2$.

(b) Konjugierte Gradienten

Die (zündende!) Idee:

Simultane Minimierung von F mit zwei (lin. unabh.) Suchrichtungen, $\underline{p}_0, \underline{p}_1$

Gesucht: ρ_0, ρ_1 so, dass

$$F(\underline{x}_0 + \rho_0 \underline{p}_0 + \rho_1 \underline{p}_1) \stackrel{!}{=} \min$$

Analog zu (*) ergibt sich:

$$(**) \quad \begin{aligned} \rho_0 (\underline{p}_0^T A \underline{p}_0) + \rho_1 (\underline{p}_1^T A \underline{p}_0) + \underline{p}_0^T \underline{r}_0 &= 0 \\ \rho_0 (\underline{p}_0^T A \underline{p}_1) + \rho_1 (\underline{p}_1^T A \underline{p}_1) + \underline{p}_1^T \underline{r}_0 &= 0 \end{aligned}$$

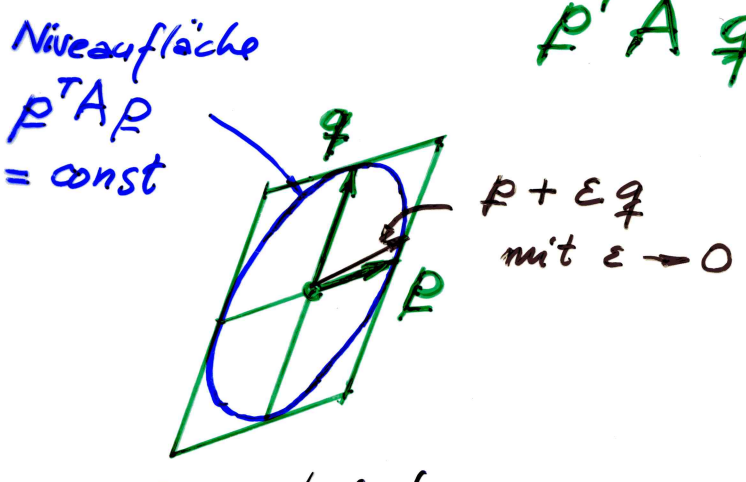
Dieses System von 2 linearen Gleichungen ist diagonal (entkoppelt), falls

$$\underline{p}_0^T A \underline{p}_1 = \underline{p}_1^T A \underline{p}_0 = 0$$

Definition

Die beiden Richtungen $\underline{p}, \underline{q} \in \mathbb{R}^n$ heissen konjugiert bezüglich $A = A^T$, falls

$$\underline{p}^T A \underline{q} = 0$$



geometrisch

Algebraisch:
Geg $\underline{p}$; finde eine Tangentenrichtg $\underline{q}$ an Niveaufläche im Endpunkt von $\underline{p}$.

Bedingung an q, im $\lim_{\varepsilon \to 0}$:

$$(p + \varepsilon q)^T A (p + \varepsilon q) = 1$$

subtrahiere $p^T A p = 1$

$$2\varepsilon \cdot \underbrace{p^T A q}_{=0} + \mathcal{O}(\varepsilon^2) = 0$$

Das cg-Verfahren:

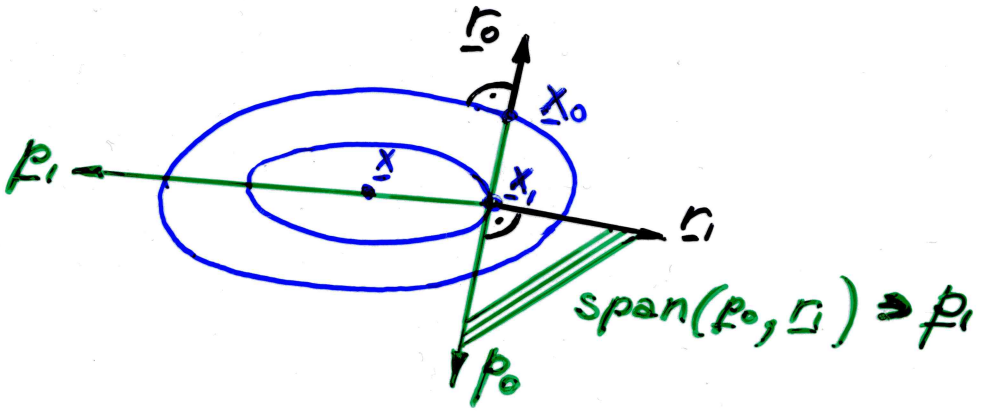
Wähle Suchrichtungen $p_0, p_1, p_2, \dots$

so, dass p_1 konj. zu p_0, $p_1^T A p_0 = 0$

p_2 konj. zu p_1, $p_2^T A p_1 = 0, \dots$

Konsequenzen:

- Algebraisch: Gleichungen für p_0, p_1 entkoppelt
- Geometrisch, z.B. für Dimension $n = 2$:



p_1 zeigt direkt auf das Minimum x

Berechnung (Wahl!) von p_1:

Wähle $p_1 \in \text{span}(p_0, r_1)$

Ansatz: $p_1 = \sigma_1 p_0 - r_1$

$$p_0, p_1 \text{ konj} \Rightarrow 0 = p_0^T A p_1 = \sigma_1 p_0^T A p_0 - p_0^T A r_1$$

$$\Rightarrow \quad \sigma_1 = \frac{\underline{p}_0^T A \underline{r}_1}{\underline{p}_0^T A \underline{p}_0}$$

Da $\underline{p}_0, \underline{p}_1$ konjugiert, folgt aus (**, S. 54):

$$\rho_1 = -\frac{\underline{p}_1^T \underline{r}_1}{\underline{p}_1^T A \underline{p}_1}$$

und weiter

$$\underline{x}_2 = \underline{x}_1 + \rho_1 \underline{p}_1$$

Zusammengefasst:

cg-Algorithmus zur Lösung von

$$A\underline{x} + \underline{b} = 0, \quad A \text{ sym. pos. def.}$$

Initialisierung: $\underline{x}_0 = 0, \quad \underline{r}_0 = \underline{b}, \quad \underline{p}_{-1} = 0$

for $k = 0, 1, \ldots$

(1) $$\sigma_k = \begin{cases} 0, & (k=0) \\ \dfrac{\underline{p}_{k-1}^T A \underline{r}_k}{\underline{p}_{k-1}^T A \underline{p}_{k-1}} = \dfrac{\underline{r}_k^T \underline{r}_k}{\underline{r}_{k-1}^T \underline{r}_{k-1}} & (k>0) \end{cases}$$

(2) $$\underline{p}_k = \sigma_k \underline{p}_{k-1} - \underline{r}_k$$

(3) $$\rho_k = -\frac{\underline{p}_k^T \underline{r}_k}{\underline{p}_k^T A \underline{p}_k} = \frac{\underline{r}_k^T \underline{r}_k}{\underline{p}_k^T A \underline{p}_k}$$

(4) $$\underline{x}_{k+1} = \underline{x}_k + \rho_k \underline{p}_k$$

(5) $$\underline{r}_{k+1} = A\underline{x}_{k+1} + \underline{b} = \underline{r}_k + \rho_k A \underline{p}_k$$

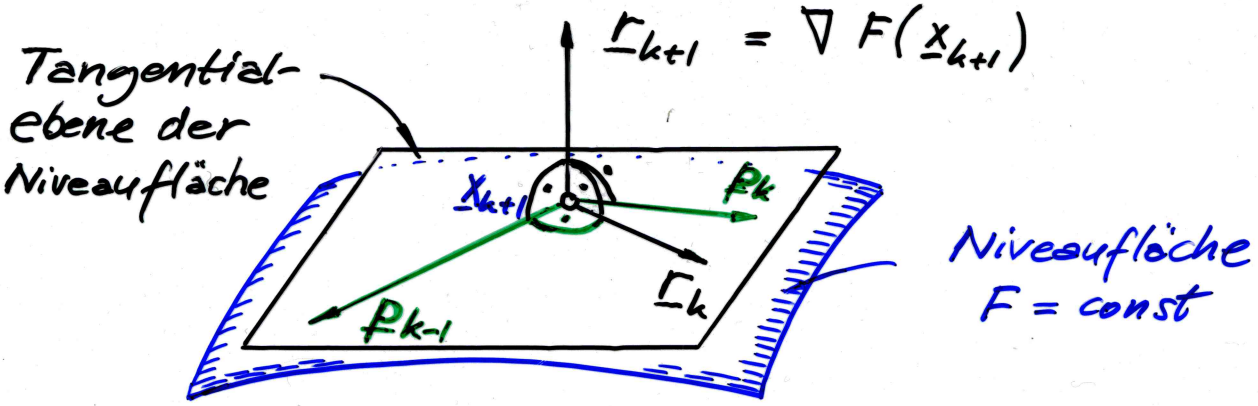
äquivalent, aber "handlicher", s. S. 57

Für den oben formulierten cg-Algorithmus gilt der

SATZ: $\underline{r}_{k+1} \perp \underline{r}_k$, $\underline{r}_{k+1} \perp \underline{p}_{k-1}$, $\underline{r}_{k+1} \perp \underline{p}_k$

für $k = 0, 1, \cdots$

Beweis: formal durch vollständige Induktion; besser geometrisch:



Damit Umrechnung der Beziehungen (3), (5), (1):

$$(3): \quad \text{Zähler} = \underline{p}_k^T \underline{r}_k \overset{(2)}{=} (\sigma_k \underline{p}_{k-1} - \underline{r}_k)^T \underline{r}_k$$

$$\overset{\text{Satz}}{=} -\underline{r}_k^T \underline{r}_k$$

$$(5): \quad \underline{r}_{k+1} \overset{(4)}{=} A(\underline{x}_k + \rho_k \underline{p}_k) + \underline{b}$$

$$= \underline{r}_k + \rho_k A \underline{p}_k, \quad \text{da } \underline{r}_k := A\underline{x}_k + \underline{b}$$

$$(1): \quad \sigma_k = \frac{(\underline{p}_{k-1} \underline{p}_{k-1}^T A)\, \underline{r}_k}{(\underline{p}_{k-1} \underline{p}_{k-1}^T A)\, \underline{p}_{k-1}}$$

$$\overset{(5)}{=} \frac{(\underline{r}_k - \underline{r}_{k-1})^T \underline{r}_k}{(\underline{r}_k - \underline{r}_{k-1})^T \underline{p}_{k-1}} \overset{\substack{\text{Satz} \\ (3)}}{=} \frac{\underline{r}_k^T \underline{r}_k}{\underline{r}_{k-1}^T \underline{r}_{k-1}}$$

Folgerungen aus dem SATZ (S.57):

- Die Suchrichtungen $p_0, p_1, \dots$ sind paarweise konjugiert
- Die Residuen (Gradienten) $r_0, r_1, \dots$ sind paarweise orthogonal.

$\Rightarrow$ $r_0 \in \mathbb{R}^n, r_1 \in \mathbb{R}^n, \dots, r_{n-1} \in \mathbb{R}^n$ bilden vollständige orthog. Basis in $\mathbb{R}^n$

$\Rightarrow$ $r_n = 0$ (da $\in \mathbb{R}^n$ und $\perp \mathbb{R}^n$)

$\Rightarrow$ $\boxed{x_n \text{ ist exakte Lösung}, \; A x_n + b = 0}$

Praktische Aspekte

(i) Wegen Rundungsfehlern bei numerischer Rechnung wird die volle Genauigkeit erst nach ein paar zusätzlichen Schritten erreicht. Einfach weiter rechnen!

(ii) Zielprobleme: $n > 10'000$

Durchführung von $\approx n$ Iterationsschritten kommt nicht in Frage!

Das Wunder:
Bei guter Kondition von A genügen schon $\sqrt{n}$ Iterationsschritte!

(sonst wäre der cg-Algorithmus längst vergessen).

(c) Vorkonditionierung

Def. Sei $\underline{x} \in \mathbb{R}^n$, $A \in \mathbb{R}^{n\times n}$, symm. pos. def

A-Norm von $\underline{x}$: $\|\underline{x}\|_A := \sqrt{\underline{x}^T A \underline{x}}$

SATZ (Konvergenz des cg-Algorithmus S.56):

$$\|\underline{x}_k - \underline{x}\|_A \leq 2\mu^k \|\underline{x}_0 - \underline{x}\|_A$$

($\underline{x}$: exakte Lösung, $A\underline{x}+\underline{b}=0$)

mit $$\mu = \frac{\sqrt{\text{cond}_2(A)} - 1}{\sqrt{\text{cond}_2(A)} + 1} \quad (<1)$$

Idee der Vorkonditionierung (preconditioning)

Betrachte neues Gleichungssystem

(1) $$\tilde{A}\,\tilde{\underline{x}} + \tilde{\underline{b}} = 0$$

mit

(2)
$$\left.\begin{aligned} \tilde{A} &= C^{-T} A C^{-1} \\ \tilde{\underline{x}} &= C\,\underline{x} \\ \tilde{\underline{b}} &= C^{-T}\underline{b} \end{aligned}\right\}$$

später: $\tilde{p} = Cp$, $\tilde{\underline{z}} = C\underline{z}$, $\tilde{\underline{r}} = C^{-T}\underline{r}$

Notation: $C^{-T} = (C^{-1})^T = (C^T)^{-1}$

Das Glsystem (1) mit den Tfsf. (2) ist äquivalent zu $A\underline{x} + \underline{b} = 0$.

Ziel: Wahl der regulären Matrix C so, dass $\text{cond}_2(\tilde{A}) \approx 1$.

Vorbereitung: cg-Algorithmus S.56 neu schreiben mit zusätzlicher Vektorfolge $\underline{z}_k := \underline{r}_k$

cg-Algorithmus mit Vorkonditionierung

zur Lösung von $A\underline{x} + \underline{b} = 0$

Initialisierung: $\underline{x}_0 = 0$, $\underline{r}_0 = \underline{b}$, $\underline{p}_{-1} = 0$

for $k = 0, 1, \ldots$

(3) ~~$\underline{z}_k = \underline{r}_k$~~ ersetzen durch **Berechne $\underline{z}_k$ aus $M\underline{z}_k = \underline{r}_k$**

$$\sigma_k = \begin{cases} 0, & k = 0 \\ \underline{z}_k^T \underline{r}_k \;/\; \underline{z}_{k-1}^T \underline{r}_{k-1}, & k > 0 \end{cases}$$

$$\underline{p}_k = \sigma_k \underline{p}_{k-1} - \underline{z}_k$$

$$\rho_k = \underline{z}_k^T \underline{r}_k \;/\; \underline{p}_k^T A \underline{p}_k$$

$$\underline{x}_{k+1} = \underline{x}_k + \rho_k \underline{p}_k$$

$$\underline{r}_{k+1} = \underline{r}_k + \rho_k \cdot A\underline{p}_k$$

Unser Vorgehen: (i) Überall ~ aufsetzen, d.h. wir wollen System (1) lösen

(ii) Trsf. (2) anwenden

Resultat: Alles bleibt invariant (bitte nachrechnen!) ausser Gleichung (3):

(3) $\xrightarrow{(i)}$ $\tilde{\underline{z}}_k = \tilde{\underline{r}}_k$ $\xrightarrow{(ii)}$ $\underbrace{(C^T C)}_{M,\ \text{"preconditioner"}} \underline{z}_k = \underline{r}_k$

Anforderungen an M:

- Gleichungssystem $M\underline{z}_k = \underline{r}_k$ einfach lösbar
- $M \approx A$ — $M = A$ würde Lösung für $k=0$ bringen!

$M = I$ ist cg ohne Vorkond