

Modellierung, Simulation und Optimierungsverfahren II

Dr. Colin W. Glass

Institut für Höchstleistungsrechnen

28.05.2014

Distanz

Das Distanzmaß hängt stark von der Repräsentation ab.

Die Distanz legt den Abstand zwischen zwei Lösungen fest. Sie soll also quantifizieren, wie ähnlich zwei Lösungen sind.

Folglich haben wir es hier mit einem Modell zu tun, welches "Ähnlichkeit" quantifizieren soll.

Richtung

Die Distanz quantifiziert den absoluten Abstand. Manchmal lässt sich auch eine sinnvolle Richtung bestimmen.

Beispielsweise ein Parametrisierungsproblem (Maschine, Verfahren, ...) hat meist eine sinnvolle Richtung: wird Parameter x_i um $+1$ verändert, so ist dies eine andere Richtung als $x_j + 1$ oder auch $x_i - 1$.

Wird eine Lösung $\mathbf{x}$ verändert und dadurch besser, kann die Richtung beibehalten werden oder von einer Lösung $\mathbf{y}$ kann in die selbe Richtung gesucht werden.

Nachbarschaft

Gegeben sei ein Suchraum $\mathcal{S}$. Die Nachbarschaft $\mathcal{N}$ einer Lösung $\mathbf{x}$ ist definiert als alle Lösungen, die sich innerhalb einer gewissen Distanz (ϵ) von $\mathbf{x}$ befinden.

$$\mathcal{N}(\mathbf{x}) = \{\mathbf{y} \in \mathcal{S} : \text{dist}(\mathbf{x}, \mathbf{y}) \leq \epsilon\}$$

Lokales Optimum

Eine Lösung ist dann ein lokales Optimum, wenn sie zulässig ist und sich innerhalb der Nachbarschaft keine bessere zulässige Lösung befindet.

Das heisst $\mathbf{x} \in \mathcal{F}$, so dass gilt:

$$F(\mathbf{x}) \leq F(\mathbf{y}) \quad \forall \mathbf{y} \in \mathcal{N}(\mathbf{x}) \cap \mathcal{F}$$

Ausblick von letzter Vorlesung

Es geht weiter mit (lokaler) Optimierung basierend auf Ableitungen:

- Ableitungen von $F(\mathbf{x})$
- Newtonverfahren
- Gradientenverfahren
- Konjugierte Gradienten

Ableitung

$F(\mathbf{x})$ kann differenzierbar sein (nach den Variablen der Repräsentation).

Hier greifen die üblichen Methoden aus der Analysis.

Ableitung der Evaluierungsfunktion

Für den Fall, dass eine Ableitung der Evaluierungsfunktion verfügbar ist, gibt es Verfahren, die ihre Suche darauf aufbauen.

Dies sind lokale Verfahren: der in der Nachbarschaft zu wählende nächste Punkt ($\mathbf{y} \in \mathcal{N}((\mathbf{x}))$) ist durch die Ableitung (und einer Schrittweite) gegeben.

- Newtonverfahren
- Gradientenverfahren
- Konjugierte Gradienten
- ...

Newton-Verfahren

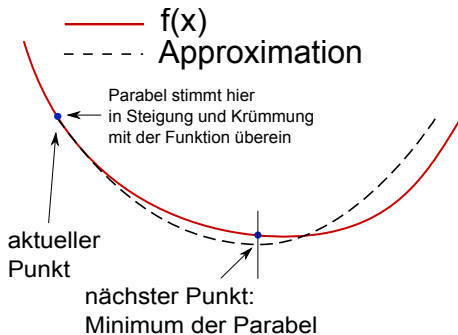
Unter der Voraussetzung, dass die 1. und 2. Ableitung von $F(x)$ berechnet werden kann, lässt sich das Newton-Verfahren einsetzen.

Idee: $F(x)$ ist zwei Mal differenzierbar und lässt sich in der Nähe eines lokalen Minimums (durch Taylor-Entwicklung) gut durch eine Parabel approximieren:

$$F(x) \approx F(x_k) + F'(x_k)(x - x_k) + \frac{1}{2}F''(x_k)(x - x_k)^2 \quad (1)$$

Newton-Verfahren

Dann kann das lokale Optimum iterativ aufgefunden werden, indem in jedem Schritt zum Minimum der Näherungsparabel des letzten Schritts gesprungen wird.



Newton-Verfahren

Algorithmus:

Wähle einen Startpunkt x_0 . Iteriere bis ein Abbruchkriterium erfüllt ist (z.B. $|x_k - x| < \epsilon$, wobei ϵ ein zu bestimmender Parameter ist.).

- Berechne $F(x_k)$, $F'(x_k)$, $F''(x_k)$,
wobei k den aktuellen Iterationsschritt bezeichnet.
- Berechne den neuen x -Wert x_{k+1} als Minimum der Approximationsparabel

$$x_{k+1} = x_k - \frac{F'(x_k)}{F''(x_k)} \quad (2)$$

Newton-Verfahren

Anmerkungen:

- Das Newton-Verfahren zur Minimierung entspricht dem Newton-Verfahren für das Auffinden von Nullstellen, angewendet auf die 1. Ableitung der Funktion ($F'(x)$)
- Verfahren, die Ableitungen verwenden, haben in der Regel höhere Anforderungen an die Funktion. Z.B. wird das Newtonverfahren instabil für sehr kleine Werte von $F''(x)$

Gradientenverfahren und konjugierte Gradienten

In erster Linie geht es hier um numerische Methode zur Lösung von großen, symmetrischen, positiv definiten Gleichungssystemen der Form $A\mathbf{x} = \mathbf{b}$.

Dieselben Verfahren können jedoch näherungsweise zur lokalen Optimierung von nichtlinearen Systemen verwendet werden. Wir werden also versuchen das Gradientenverfahren und die konjugierte Gradienten auf Basis des linearen Falls zu verstehen.

Wir betrachten nun die Verfahren auf Basis von:

"An Introduction to the Conjugate Gradient Method Without the Agonizing Pain"

Jonathan Richard Shewchuck

Lineare Systeme: PDEs

Der vielleicht prominenteste Fall linearer Gleichungssysteme sind numerische Approximationen für partielle Differentialgleichungen:

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \right) f(x, y, z) = g(x, y, z)$$

Wobei

$$\frac{\partial}{\partial x} \hat{=} \text{partielle Ableitung nach } x$$

P.s: Dies ist die Poisson-Gleichung (z.B. Diffusion).

Finite Differenzen Methode

Grundüberlegung: Wir approximieren eine Ableitung mit der durchschnittlichen Änderung über eine kleine Differenz (hier: zentrale Differenzenquotienten, 1D Problem).

$$f'(x) = \frac{f(x + h/2) - f(x - h/2)}{h}$$

Wenn wir nun die zweite Ableitung haben wollen, machen wir das Gleiche nochmal :

$$f''(x) = \frac{f'(x + h/2) - f'(x - h/2)}{h}$$

$$f''(x) = \frac{\frac{f(x+h)-f(x)}{h} - \frac{f(x)-f(x-h)}{h}}{h} = \frac{f(x+h) - 2 * f(x) + f(x-h)}{h^2}$$

Diskretisieren

$$f''(x) = \frac{f(x+h) - 2 * f(x) + f(x-h)}{h^2}$$

Für ein 1D Problem haben wir den Gitterindex j und wählen eine konstante Differenz h . Folglich:

$$f''(x) = \alpha_{j-1} f_{j-1} + \alpha_j f_j + \alpha_{j+1} f_{j+1}$$

Mit:

$$\alpha_{j-1} = \frac{1}{h^2} = 1 * \frac{1}{h^2}$$

$$\alpha_j = -\frac{2}{h^2} = -2 * \frac{1}{h^2}$$

$$\alpha_{j+1} = \frac{1}{h^2} = 1 * \frac{1}{h^2}$$

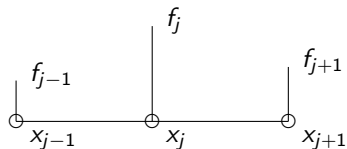


Figure: 3 Punkte in Dimension x mit konstantem Abstand h und Werten $f_x = f(x)$ (1D Problem)

Die zweite Ableitung im 1D-Raum als Beispiel

- Es soll auf dem Intervall $[a, b]$ die Gleichung $f''(x) = \rho(x)$ gelöst werden.
- Die Funktion $f(x)$ wird durch den Vektor seiner Werte an den Diskretisierungspunkten x_j ersetzt, ebenso die Funktion der rechten Seite $\rho(x)$.
- das Intervall wird in gleichgroße Teilintervalle zerlegt (hier 8).

Die zweite Ableitung im 1D-Raum als Beispiel

- Wir erhalten also das folgende Gleichungssystem

$$\frac{1}{h^2} \begin{bmatrix} -2 & 1 & & & & & & \\ 1 & -2 & 1 & & & & & \\ & 1 & -2 & 1 & & & & \\ & & 1 & -2 & 1 & & & \\ & & & 1 & -2 & 1 & & \\ & & & & 1 & -2 & 1 & \\ & & & & & 1 & -2 & 1 \\ & & & & & & 1 & -2 \end{bmatrix} \begin{bmatrix} f_2 \\ f_3 \\ f_4 \\ f_5 \\ f_6 \\ f_7 \\ f_8 \end{bmatrix} = \begin{bmatrix} \rho_2 - f_a \\ \rho_3 \\ \rho_4 \\ \rho_5 \\ \rho_6 \\ \rho_7 \\ \rho_8 - f_b \end{bmatrix}$$

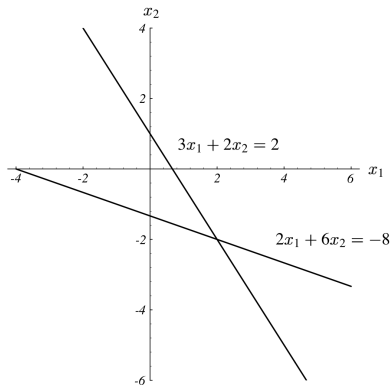
- Am ersten und letzten Punkt werden die Werte von $f_1 = f_a$ und $f_9 = f_b$ als Randbedingungen vorgegeben.
- Wir haben ein lineares Gleichungssystem der Form $Ax = b$

Lineare Systeme

Wir betrachten das einfache System

$$3x_1 + 2x_2 = 2$$

$$2x_1 + 6x_2 = -8$$



[Jonathan Richard Shewchuck]

Lineare Systeme

Zuerst formulieren wir das Problem $Ax = b$ um in die quadratische Form:

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

Der Gradient von $F(x)$ lautet:

$$\nabla F(x) = \frac{1}{2}A^T x + \frac{1}{2}Ax - b$$

Folglich für symmetrische A :

$$\nabla F(x) = Ax - b$$

Das heisst unsere gesuchte Lösung $Ax = b$ führt zu $\nabla F(x) = 0$.

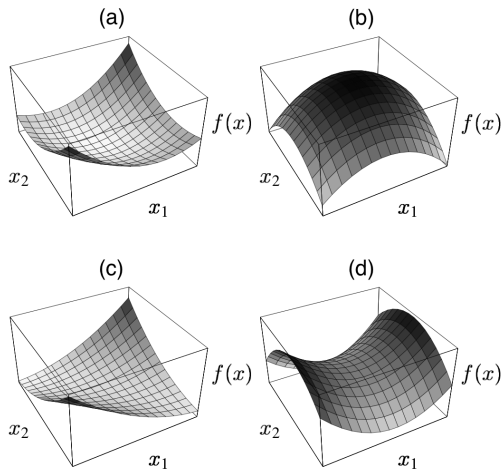
Da A symmetrisch und positiv definit ist, ist das ein Minimum.

Wir können folglich $Ax = b$ lösen, in dem wir das Minimum von $F(x)$ finden.

Lineare Systeme

Fitnesslandschaft bei
verschiedenen
Matrizen:

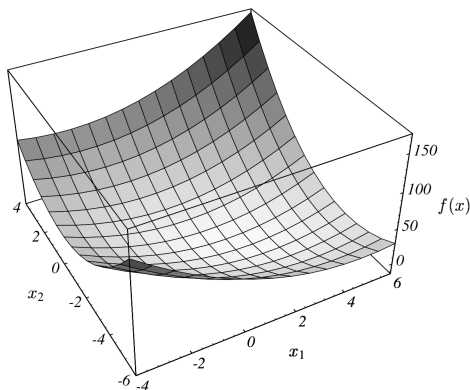
- a) positiv-definit
- b) negativ-definit
- c) singulär,
positiv-definit
- d) indefinit



[Jonathan Richard Shewchuck]

Lineare Systeme

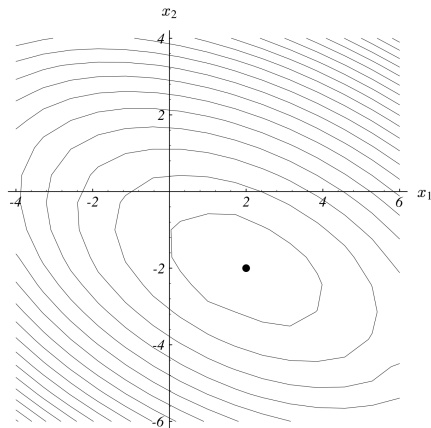
Unser System sieht in quadratischer Form so aus ($F(x)$):



[Jonathan Richard Shewchuck]

Lineare Systeme

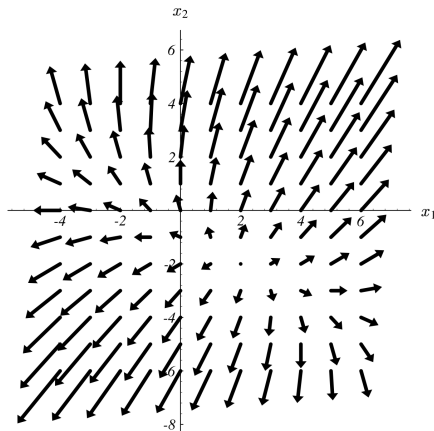
Unser System sieht in quadratischer Form so aus (Höhenlinien von $F(x)$):



[Jonathan Richard Shewchuck]

Lineare Systeme

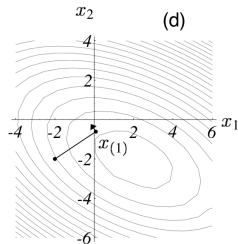
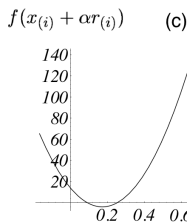
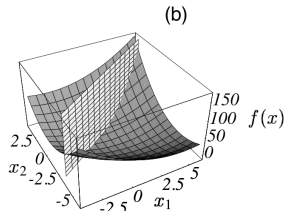
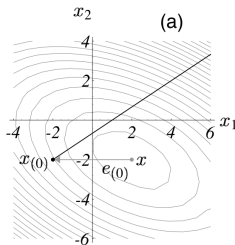
Unser System sieht in quadratischer Form so aus $(\nabla F(x))$:



[Jonathan Richard Shewchuck]

Gradientenverfahren ("Steepest Descent")

Wir wählen die Richtung $-\nabla F(x)$ und machen entlang dieser eine Liniensuche für den besten Wert. Iterieren.

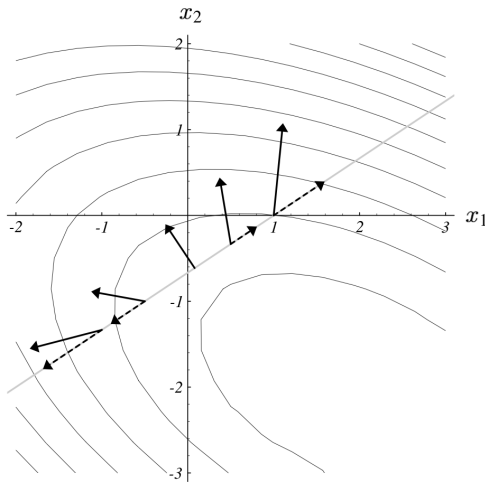


[Jonathan Richard Shewchuck]

Gradientenverfahren ("Steepest Descent")

Wir wählen die Richtung $-\nabla F(x)$ und machen entlang dieser eine Liniensuche für den besten Wert. Iterieren.

Der beste Punkt ist natürlich der, dessen Gradient orthogonal zur Suchrichtung ist.



[Jonathan Richard Shewchuck]

Gradientenverfahren ("Steepest Descent")

Wir bestimmen nun die Schrittweite:

$$-\nabla F(x) = b - Ax = r \quad \text{dies nennen wir "Residuum"}$$

$$x_{i+1} = x_i + \alpha r_i \quad \alpha \text{ ist die Schrittweite}$$

$$r_{i+1}^T r_i = 0 \quad \text{Orthogonal!}$$

$$(b - Ax_{i+1})^T r_i = 0$$

$$(b - A(x_i + \alpha_i r_i))^T r_i = 0$$

$$(b - Ax_i)^T r_i - \alpha_i (Ar_i)^T r_i = 0$$

$$(b - Ax_i)^T r_i = \alpha_i (Ar_i)^T r_i$$

$$r_i^T r_i = \alpha_i r_i^T (Ar_i)$$

$$\alpha_i = \frac{r_i^T r_i}{r_i^T (Ar_i)}$$

Gradientenverfahren ("Steepest Descent")

Das Gradientenverfahren funktioniert folgendermassen:

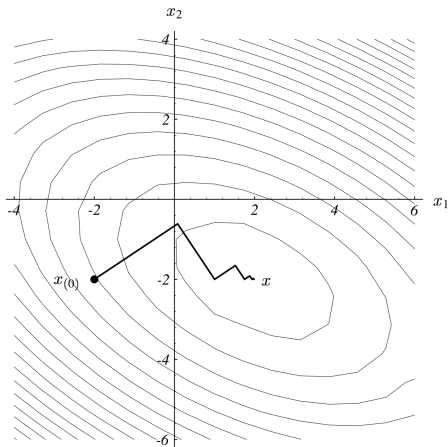
$$\begin{aligned}r_i &= b - Ax_i \\ \alpha_i &= \frac{r_i^T r_i}{r_i^T (Ar_i)} \\ x_{i+1} &= x_i + \alpha_i r_i\end{aligned}$$

P.s: Ein Matrix-Vektor Produkt können wir los werden. Die letzte Gleichung mit $-A$ multiplizieren und b addieren führt zu:

$$r_{i+1} = r_i - \alpha_i Ar_i$$

Gradientenverfahren ("Steepest Descent")

Konvergenzverhalten: je elliptischer die Fitnesslandschaft desto schlechter!



[Jonathan Richard Shewchuck]

Konjugierte Gradienten

Das Konjugierte Gradienten Verfahren ist eine effiziente numerische Methode zur Lösung von großen, symmetrischen, positiv definiten Gleichungssystemen der Form $A\mathbf{x} = \mathbf{b}$.

Das Verfahren konvergiert nach spätestens m Schritten zur exakten Lösung, wobei m die Größe der quadratischen Matrix A ist.

Konjugierte Gradienten

Das Konjugierte Gradienten Verfahren ist eine effiziente numerische Methode zur Lösung von großen, symmetrischen, positiv definiten Gleichungssystemen der Form $A\mathbf{x} = \mathbf{b}$.

Das Verfahren konvergiert nach spätestens m Schritten zur exakten Lösung, wobei m die Größe der quadratischen Matrix A ist.

Das ist ja äusserst attraktiv ...

Konjugierte Gradienten

Grundidee: Wir machen Schritte nur in Richtungen, die alle zueinander orthogonal sind. Ausserdem machen wir immer genau den richtigen Schritt in eine Richtung!

Konjugierte Gradienten

Grundidee: Wir machen Schritte nur in Richtungen, die alle zueinander orthogonal sind. Ausserdem machen wir immer genau den richtigen Schritt in eine Richtung!

Problem: Um das zu erreichen, müssen wir den Abstand zur Lösung kennen (e, von Error). Diesen kennen wir jedoch nur im Raum, der durch A verzerrt ist ($r = b - Ax$).

Konjugierte Gradienten

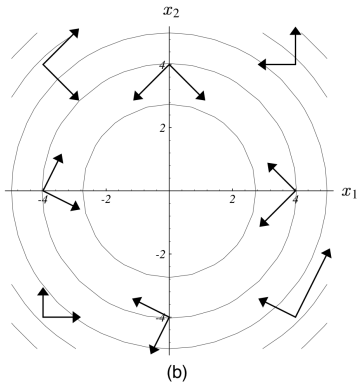
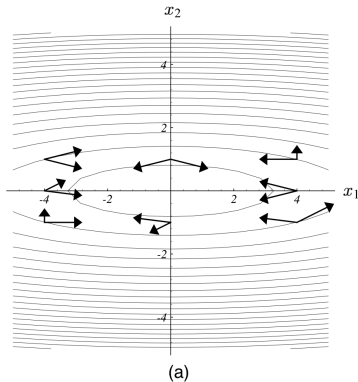
Grundidee: Wir machen Schritte nur in Richtungen, die alle zueinander orthogonal sind. Ausserdem machen wir immer genau den richtigen Schritt in eine Richtung!

Problem: Um das zu erreichen, müssen wir den Abstand zur Lösung kennen (e , von Error). Diesen kennen wir jedoch nur im Raum, der durch A verzerrt ist ($r = b - Ax$).

Lösung: Wir verlangen, dass die Schritte im verzerrten Raum orthogonal sind! Dies heisst A -orthogonal und die Richtungen (nennen wir sie d) heissen konjugiert.

Konjugierte Gradienten

A-orthogonale Vektoren sehen z.B. so aus:

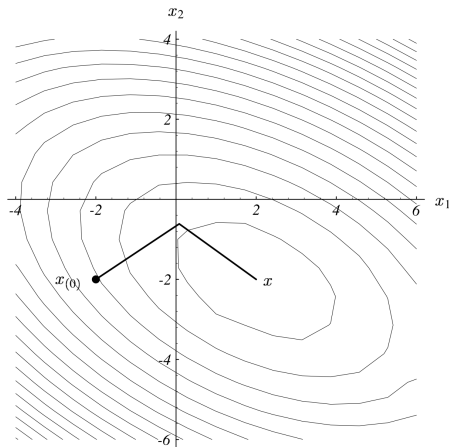


Konjugierte Gradienten

Da wir entlang A-orthogonalen Richtungen gehen, machen wir keine Schritte "zurück". Mit etwas linearer Algebra ... ergibt sich folgendes Verfahren:

$$\begin{aligned}d_0 &= r_0 = b - Ax_0 \\ \alpha_i &= \frac{r_i^T r_i}{d_i^T A d_i} \\ x_{i+1} &= x_i + \alpha_i d_i \\ r_{i+1} &= r_i - \alpha_i A d_i \\ \beta_{i+1} &= \frac{r_{i+1}^T r_{i+1}}{r_i^T r_i} \\ d_{i+1} &= r_{i+1} + \beta_{i+1} d_i\end{aligned}$$

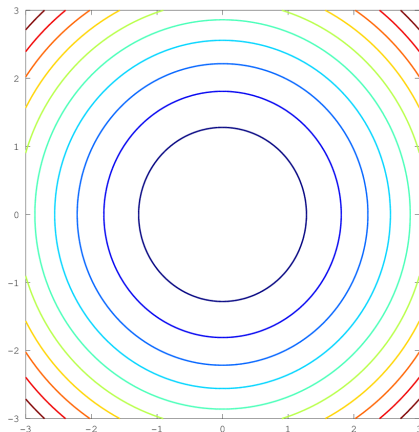
Konjugierte Gradienten



[Jonathan Richard Shewchuck]

Kondition

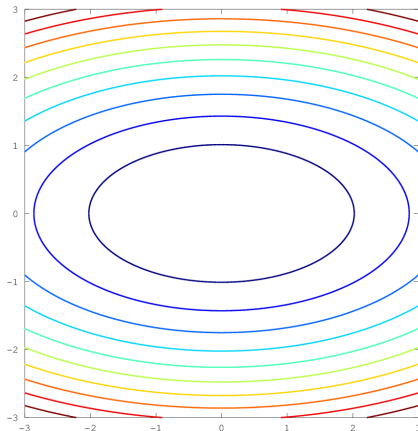
Die Konditionszahl κ gibt das Verhältnis des größten zum kleinsten Eigenwert einer Matrix. Die Landschaft hat die Kondition $\kappa = 1$.



Kondition

Hier ist die Kondition $\kappa = 2$. Interessant wird es, wenn die Kondition z.B. 10^6 beträgt.

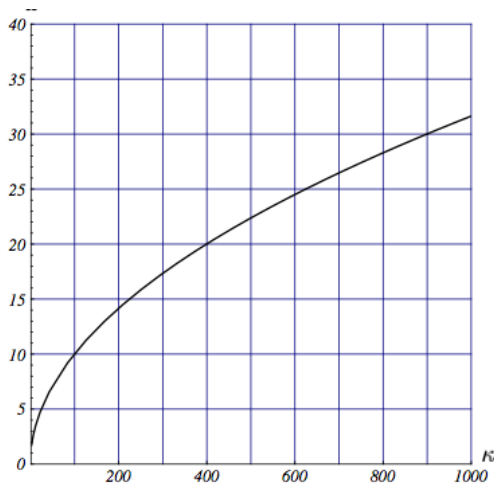
In realen Problemen können sehr grosse Konditionen auftreten, man überlege sich allein, wenn wir z.B. ein Problem in Meter vs. Mikrometer repräsentieren.



Gradientenverfahren vs. Konjugierte Gradienten

Je höher die Kondition, je besser funktionieren konjugierte Gradienten im Verhältnis zu Gradienten.

Hier: Kondition vs. Anzahl Iteration



[Jonathan Richard Shewchuck]

Nichtlineare Probleme

Für nichtlineare Optimierungsprobleme können sowohl das Gradientenverfahren als auch Konjugierte Gradienten trotzdem verwendet werden. Die Hauptunterschiede sind:

- Das Optimum wird vermutlich ein lokales Optimum sein.
- Wir wissen nicht, wann wir das Optimum erreicht haben.
- Die Schrittweite muss anderweitig bestimmt werden.

Für Konjugierte Gradienten ergeben sich noch weitere Unterschiede.

Konjugierte Gradienten für nichtlineare Probleme

Wenn wir ein paar knifflige Änderungen vornehmen, können wir das Verfahren auch für nichtlineare Problem verwenden.

Im Wesentlichen sieht es so aus:

- Wir kennen r nicht. Wir nehmen trotzdem einfach $-\nabla F$ (im linearen Fall gilt ja $r = -\nabla F$)
- Wir verwenden eine generalisierte Liniensuche, z.B. Newton-Verfahren.
- Es gibt verschiedene Ansätze, β anders zu wählen.
- Wir müssen regelmässig neu starten! (Sprich, so tun als würden wir vom momentanen Punkt anfangen.)

Konjugierte Gradienten für nichtlineare Probleme

Ausserdem haben wir keine Matrix A (hier: β gemäss Polak-Ribiere):

$$d_0 = r_0 = -\nabla F(0)$$

$\Rightarrow$ Liniensuche (entlang d_i) nach α_i , dass $F(x_i + \alpha_i d_i)$ minimal wird.

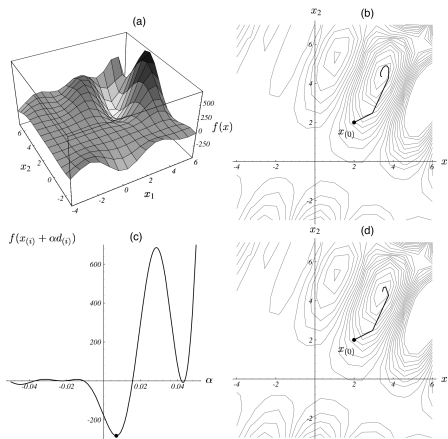
$$x_{i+1} = x_i + \alpha_i d_i$$

$$r_{i+1} = -\nabla F(x_{i+1})$$

$$\beta_{i+1} = \max \left(\frac{r_{i+1}^T (r_{i+1} - r_i)}{r_i^T r_i}, 0 \right)$$

$$d_{i+1} = r_{i+1} + \beta_{i+1} d_i$$

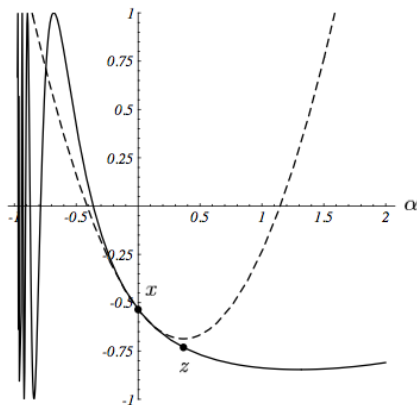
Konjugierte Gradienten für nichtlineare Probleme



[Jonathan Richard Shewchuck]

Konjugierte Gradienten für nichtlineare Probleme

Liniensuche mit Newtonverfahren



[Jonathan Richard Shewchuck]

Die Schrittweite

Für die Bestimmung der Schrittweite können wir diverse Ansätze wählen. Wie eben gesehen:

- Vereinfachte Liniensuche mit Newtonverfahren

Für das Gradientenverfahren bietet sich auch an:

- Proportional zum Gradienten (klar)

Natürlich gibt es viele weitere Möglichkeiten. Im allgemeinen ist die Liniensuche schlicht ein 1D Optimierungsproblem.

Schrittweite: Adaptiv

Die adaptive Anpassung der Schrittweite wird z.B. bei Hillclimbern gerne eingesetzt und bedeutet einfach folgendes:

- Wir machen einen Schritt.
- Ist der Schritt gut (führt zu besserer Lösung), erhöhe die Schrittweite.
- Ist der Schritt schlecht, reduziere die Schrittweite (und wirf je nach dem die neue Lösung weg).

Die Erhöhung und die Reduktion werden so getuned, dass bei einer bestimmten Erfolgsquote die Schrittweite gemittelt gleich bleibt.

Häufig wird hier die $\frac{1}{5}$ -Regel angewendet (einer aus fünf Schritten soll erfolgreich sein)

Anmerkung: Numerische Differentiation

Wie bereits besprochen, kann die Ableitung von $F(\mathbf{x})$ numerisch approximiert werden.

Diese numerisch approximierten Ableitungen können ebenfalls als Basis für die eben besprochenen Verfahren verwendet werden.

Ausblick

Die nächste Vorlesung wird sich mit einigen Beispielehaften Verfahren für globale Optimierung befassen.

- Tabu Search (Nachtrag)
- Simplex
- Random Search
- Exhaustive Search
- Branch and Bound