Einführung in die Grundlagen der Numerik

Vorlesungsmitschriften im Wintersemester 2018/19

INHALTSVERZEICHNIS

VORWORT

Diese Vorlesungsmitschriften werden in der Vorlesung *Einführung in die Grundlagen der Numerik* von Prof. Ira Neitzel im Wintersemester 2018/19 an der Universität Bonn angefertigt.

Wir versuchen, diese immer unter `https://pankratius.github.io` zu aktualisieren.

Teile, die von der Vorlesung abweichen, sind in violett markiert.

Betrachte $A \in \mathrm{GL}_n(\mathbb{R})$, wobei A schlecht konditioniert sein kann. Wir wollen ein Gleichungssystem der Form $Ax = b$, mit $b \in \mathbb{R}^n$ gegeben, lösen. Dazu suchen wir eine Orthogonalmatrix $Q \in \mathrm{O}_n(\mathbb{R})$ und eine obere Dreiecksmatrix $R \in \mathrm{M}_n(\mathbb{R})$ mit $A = QR$. Diese Zerlegung von A nennt man **Orthogonalzerlegung**. Dann erhalten wir das äquivalente Problem

$$Ax = b \Longleftrightarrow QRx = b \Longleftrightarrow Rx = Q^T b.$$

1.1 Eigenschaften orthogonaler Matrizen

Lemma 1.1.1. *Sei $Q \in \mathrm{O}_m(\mathbb{R})$ orthogonal. Dann ist auch Q^T orthogonal und es gilt*

$$||Qx|| = \left|\left|Q^T x\right|\right| = ||x||$$

Beweis. Es gilt

$$||Qx||^2 = x^T Q^T Q x = x^T x = ||x||.$$

Genauso für Q^T. □

Lemma 1.1.2. *Sei $A \in \mathrm{GL}_n(\mathbb{R})$ regulär und $Q \in \mathrm{O}_n(\mathbb{R})$ orthogonal. Dann gilt*

$$\kappa_2(QA) = \kappa_2(A)$$

Beweis. Die Matrixnorm $||A||$ ist durch die euklidsche Norm induziert, i.e.

$$||A|| = \max_{x \neq 0} \frac{||Ax||}{||x||}.$$

Also folgt aus lemma 1.1.1, dass $||(||\,QA) = ||(||\,A)$ gilt.
Betrachte jetzt

$$\left|\left|A^{-1}Q^T\right|\right| = \max_{x \neq 0} \frac{\left|\left|A^{-1}Q^T x\right|\right|}{||x||} \qquad = \max_{x \neq 0} \frac{\left|\left|A^{-1}Q^T x\right|\right|}{||Q^T x||} \overset{y:=Q^T x}{=} \max_{y \neq 0} \frac{\left|\left|A^{-1}\right|\right|}{||y||} = \left|\left|A^{-1}\right|\right|$$

□

Also ist für das LGS $Rx = Q^T b : \kappa_2(R) = \kappa_2(A)$. Also hat sich die Kondition des Problems nicht verschlechtert.

1.2 Anwendung: Lineare Ausgleichsgeraden

Betrachte für gegebenes $b \in \mathbb{R}^n$ und $A \in \mathrm{M}_n(\mathbb{R})$ das Optimierungsproblem

$$\min_{x \in \mathbb{R}^n} ||Ax - b||. \tag{O}$$

Dieses Problem ist äquivalent zur Optimierung von $||Ax - b||^2$.

Seien nun m Tupel $(y_i, f_i) \in \mathbb{R}^2$ $(1 \leq i \leq m)$ gegeben. Gesucht ist diejenige affine Gerade $c + dy$ in $\mathbb{R}^2$, so dass die Summe der Quadrate der Punkte von der Gerade minimal ist. Wir erhalten also das Optimierungsproblem

$$\min_{(c,d)\in\mathbb{R}^2} \left(\sum_{i=1}^{m} (c + dy_i - f_i)^2 \right) = \min_{(c,d)\in\mathbb{R}^2} \left\| \begin{pmatrix} 1 & y_1 \\ \vdots & \vdots \\ 1 & y_m \end{pmatrix} \cdot \begin{pmatrix} c \\ d \end{pmatrix} - \begin{pmatrix} f_1 \\ \vdots \\ f_m \end{pmatrix} \right\|.$$

Betrachte allgemeiner das Polynom

$$p(y) = \sum_{k=0}^{n-1} a_k y^k.$$

Gesucht sind jetzt die Koeffizienten $a_0, ..., a_{n-1}$ mit

$$\sum_{j=1}^{m} (p(y_j) - f_j)^2$$

ist minimal. Schreibe dies ebenfalls als Optimierungsproblem:

$$\min_{a_0,\ldots,a_{n-1}} \left\| \begin{pmatrix} y_1^0 & \ldots & y_1^{n-1} \\ \vdots & \ddots & \vdots \\ y_m^0 & \ldots & y_m \end{pmatrix} \cdot \begin{pmatrix} a_0 \\ \vdots \\ a_{n-1} \end{pmatrix} - \begin{pmatrix} f_1 \\ \vdots \\ f_m \end{pmatrix} \right\|^2.$$

End of Lecture 1

Die Existenz der Lösung des Optimierungsproblems folgt aus

$$\lim_{x\to\infty} ||Ax - b|| \to \infty,$$

und einer anschließenden Anwendung des Satzes von Weierstraß auf die kompakten Niveaumengen der Abbildung.

Theorem 1.2.1. *(Weierstraß) Sei X ein kompakter metrischer Raum und $f : X \to \mathbb{R}$ eine stetige Abbildung. Dann nimmt f auf X sowohl ein Maximum als auch ein Minimum an.*

Sei $f : X \to \mathbb{R}$, mit $X \subseteq \mathbb{R}$. Dann heißt $x_0 \in X$ ein **lokales Maximum** bzw. **lokales Minimum**, falls es eine Umgebung $x \in V$ gibt, so dass $f(x) \leq f(x_0)$ bzw. $f(x) \geq f(x_0)$ für alle $x \in V$ gilt.

Lemma 1.2.2. *Sei $U \subset \mathbb{R}^n$ offen, und $f : U \to \mathbb{R}$. Angenommen, f hat in $x_0 \in U$ eine Extremstelle und ist in x_0 partiell differenzierbar. Dann gilt*

$$\nabla f(a) = 0.$$

Beweis. Betrachte für ein hinreichend kleines $\varepsilon > 0$ und $1 \leq j \leq n$ die Funktion

$$F : (-\varepsilon, \varepsilon) \to \mathbb{R}, \; t \mapsto F(x_0 + te_j).$$

Dann hat F in $t = 0$ eine Extremstelle, und es gilt

$$0 = F'(0) = \partial_j f(x_0)$$

$\square$

Betrachte zur Lösung des Optimierungsproblems nun immer die Funktion

$$f : \mathbb{R}^n \to \mathbb{R},\ x \mapsto \frac{1}{2} \left|\left|Ax - b\right|\right|^2 .$$

Dann gilt für beliebiges $\overline{x} \in \mathbb{R}^n$

$$\begin{aligned} \langle \nabla f(\overline{x}), h \rangle &= \langle A\overline{x} - b, Ah \rangle \\ &= \langle A^T(A\overline{x} - b), h \rangle \end{aligned}$$

Also gilt für eine Extremstelle $\overline{x}$

$$\nabla f(\overline{x}) = A^T(A\overline{x} - b) \overset{!}{=} 0 \iff \boxed{A^T A\overline{x} = A^T b}. \tag{NE}$$

Zur Lösung des Optimierungsproblems müssen wir also ebenfalls ein Gleichungssystem lösen. Die Gleichung (NE) heißt **Normalengleichung**.

Für $A \in \mathbb{R}^{m,n}$ ist $B := AA^T$ symmetrisch. Weiterhin ist B positiv semi-definit, denn es gilt

$$\langle x, Bx \rangle = x^T Bx = x^T AA^T x = \langle A^T x, A^T x \rangle = \left|\left|A^T x\right|\right| \geq 0.$$

Weiterhin impliziert dies, dass alle Eigenwerte von x größer gleich null sind. Sei dazu λ ein Eigenwert und x ein korrespondierender Eigenvektor von B,

$$0 \leq \langle Bx, x \rangle = \langle \lambda x, x \rangle = \lambda \left|\left|x\right|\right|^2 .$$

Also ist B positiv semi-definit. Angenommen, A hat nun vollen Rang. Dann ist A^T injektiv. Sei x ein Eigenvektor von B zum Eigenwert 0. Dann gilt

$$0 = \langle x, Bx \rangle = \left|\left|A^T x\right|\right|^2 \implies A^T x = 0 \implies x = 0.$$

Also hat B nur positive Eigenwerte. Nach dem euklidschen Spektralsatz [Sch18, S. 20.25] ist B aber diagonalisierbar. Also ist B sogar invertierbar, und die Normalengleichung hat hier jeweils eine eindeutig bestimmte Lösung. Im folgende habe A also immer maximalen Rang. Weil AA^T symmetrisch positiv-definit ist, kann man (NE) mit der Choleskyzerlegung lösen. Es gilt aber $\kappa_2(AA^T) = \kappa_2(A)^2$,

Wie ist die Kondition von A im nicht-quadratischen Fall definiert?

weshalb weitere Lösungsverfahren betrachtet werden müssen.

Dazu definieren wir die **erweiterte Orthogonalzerlegung** von $A \in \mathbb{R}^{m,n}$, mit $m \geq n$ durch

$$A = QR,\ \text{mit}\ Q \in \mathrm{O}_m\,\mathbb{R}\ \text{und}\ R = \left(\frac{\hat{R}}{0}\right),$$

wobei $\hat{R} \in \mathbb{R}^{n,n}$ eine obere Dreiecksmatrix ist. R heißt in diesem Fall **erweiterte obere Dreiecksmatrix**
Angenommen, eine solche Zerlegung existiert. Dann gilt

$$\begin{aligned} ||Ax-b||^2 &= ||QRx-b||^2 \\ &= \left|\left|QRX - QQ^Tb\right|\right|^2 \\ &= \left|\left|Q(Rx+Q^Tb)\right|\right| \\ &= \left|\left|\left(\frac{\hat{R}}{0}\right)x - \left(\frac{y_1}{y_2}\right)\right|\right| \\ &= \left|\left|\hat{R}x' - y_1\right|\right|^2 + ||y_2||^2, \end{aligned}$$

für passend gewählte y_1, y_2. Weil y_2 aber fix ist, reicht es,

$$\left|\left|\hat{R}x - y_1\right|\right|^2$$

zu minimieren.

Theorem 1.2.3. *Sei $1 \leq n \leq m$ und $A \in \mathbb{R}^{m,n}$, mit $\deg A = n$. Angenommen, A hat eine erweiterte Orthogonalzerlegung. Sei*

$$Q^T =: \left(\frac{y_1}{y_2}\right),$$

mit $y_1 \in \mathbb{R}^n$ und $y_2 \in \mathbb{R}^{m-n}$. Dann sind äquivalent

i) $\overline{x} \in \mathbb{R}^n$ löst das Optimierungsproblem (O).

ii) $\hat{R}x = y_1$.

1.3 Gram-Schmidt-Verfahren

Wir wollen die Existenz einer Orthogonalzerlegung von A zeigen. Dazu verwenden wir [Sch18, S. 11.32]

Proposition 1.3.1. *(Gram-Schmidtsches Orthogonalisierungsverfahren) Sei V ein n-dimensionaler euklidischer Vektorraum und $(b_1, ..., b_n)$ eine geordnete Basis von V. Für $1 \leq i \leq n$ sei*

$$V_i := \operatorname{Lin}(b_1, ..., b_i),$$

die also eine Flagge

$$0 = V_0 \subset V_1 \subset ... \subset V_n = V$$

bilden. Dann gibt es eine geordnete Orthogonalbasis $(\hat{b}_n, ..., \hat{b}_n)$ von V, so dass

$$V_i = \operatorname{Lin}(\hat{b}_1, ..., \hat{b}_n) \qquad (*)$$

gilt.

Theorem 1.3.2. *Für jede reguläre Matrix $A \in \operatorname{GL}_n(\mathbb{R})$ existiert eine orthogonale Matrix $Q \in \operatorname{O}_n(\mathbb{R})$ und eine obere Dreiecksmatrix $R \in \operatorname{M}_n \mathbb{R}$, so dass $A = QR$ gilt.*

Beweis. Aus ($*$) folgt, dass für die Abbildung

$$g : V \to V,\ b_i \mapsto \hat{b}_i$$

oben die Koordinatenmatrix bezüglich B in oberer Dreiecksform sein muss. Weiterhin ist g per Definition ein Isomorphismus.
Betrachte nun den konkreten Fall für $A \in \mathrm{GL}_n(\mathbb{R})$. Dann bilden die Spalten von A^{-1} eine geordnete Basis von $\mathbb{R}^n$. Also gibt es eine obere Dreiecksmatrix g, so dass

$$gA^{-1} = Q,$$

wobei $Q \in \mathrm{O}_n(\mathbb{R})$ orthogonal ist. Dabei haben wir benutzt, dass eine Matrix Q genau dann orthogonal ist, wenn ihre Spalten eine Orthonormalbasis von $\mathbb{R}^n$ bilden [**schroer**]. Damit erhalten wir aber

$$A = Q^{-1}g,$$

und $Q^{-1} \in \mathrm{O}_n(\mathbb{R})$, weil die orthogonalen Matrizen eine Gruppe bilden. □

1.4 Householder-Transformationen

Betrachte für ein fixes $w \in \mathbb{R}^s$ mit $w^T w = 1$ die Abbildung

$$H : \mathbb{R}^s \to \mathbb{R}^s, x \mapsto x - 2ww^T x.$$

Die Abbildung H heißt **Householder-Spiegelung**.

Proposition 1.4.1. *Für ein solches H gilt*

i) $H^T = H$

ii) $H^2 = E_s$,

also ist H insbesondere eine orthogonale Matrix, $H \in \mathrm{O}_s(\mathbb{R})$.

Beweis. i) $H^T = (E_s - 2ww^T)^T = E_s^T - 2(w^T)(w^T) = E_s - 2ww^T = H$

ii) Es gilt $\mathbb{R}^s = \mathrm{Lin}(w) \perp (\mathrm{Lin}(w))^\perp$. Weil H linear ist, genügen die folgenden beiden Ergebnisse

$$H(w) = w - 2(ww^T)w = w - 2w(w^T w) = w - 2w = -w,$$

und für $v \perp w$

$$H(v) = v - 2(ww^T)v = v - 2w(w^T v) = v.$$

□

End of Lecture 2

Die Idee ist jetzt, eine gegebene Matrix A durch Householder-Transformationen in eine verallgemeinerte obere Dreiecksform zu bringen. Dazu wollen wir zuerst einen Vektor $w \in \mathbb{R}^2$ finden, so dass die Spiegelung der ersten Spalte von A an w ein skalares Vielfaches des ersten Einheitsvektors ist. Induktiv fährt man dann mit der $m-1 \times n-1$-Teilmatrix fort:

$$\begin{pmatrix} a_{11} & \times & \times & \times \\ \times & a_{22} & \times & \times \\ \times & \times & \times & \times \\ \times & \times & \times & \times \\ \times & \times & \times & \times \end{pmatrix} \rightsquigarrow \begin{pmatrix} \times & \times & \times & \times \\ 0 & a_{22} & \times & \times \\ 0 & \times & \times & \times \\ 0 & \times & \times & \times \\ 0 & \times & \times & \times \end{pmatrix} \rightsquigarrow \begin{pmatrix} \times & \times & \times & \times \\ 0 & \times & \times & \times \\ 0 & 0 & \times & \times \\ 0 & 0 & \times & \times \\ 0 & 0 & \times & \times \end{pmatrix}$$

Lemma 1.4.2. *Sei $0 \neq x \in \mathbb{R}^s$, so dass $x \ni \operatorname{Lin} e_1$. Für*

$$w := \frac{x + \sigma e_1}{||x + \sigma e_1||} \text{ mit } \sigma = \pm\, ||x||$$

gilt:

i) $||w|| = 1$,

ii) $(E_s - 2ww^T)x = -\sigma e_1$

Beweis. Es gilt $||x + \sigma e_1||$, da x und e_1 nach Voraussetzung linear unabhängig sind. i) folgt dann sofort. ii) folgt durch rechnen:

$$\begin{aligned} ||x + \sigma e_1||^2 = \langle x + \sigma e_1, x + \sigma e_1 \rangle &= \langle x, x \rangle + 2\langle x, e_1 \rangle + \sigma^2 \langle e_1, e_1 \rangle \\ &= 2\langle x, x \rangle + 2\sigma \langle e_1, x \rangle \\ &= 2\langle x + \sigma e_1, x \rangle \end{aligned}$$

Mit der Definition von w folgt

$$\begin{aligned} 2w^T x &= \frac{2(x + \sigma e_1)^T x}{||x + \sigma e_1||} \\ &= \frac{2\langle x + \sigma e_1, x \rangle}{||x + \sigma e_1||} \\ &= 2\, ||x + \sigma e_1||\,, \end{aligned}$$

so dass wir schlussendlich

$$2ww^T x = \frac{x + \sigma e_1}{||x + \sigma e_1||}\, ||x + \sigma e_1|| \implies (E_s - 2ww^T)x = x - (x + \sigma e_1) = -\sigma e_1$$

erhalten. $\square$

Sei nun $a \in \mathbb{R}^{m,n}$ mit vollem Rang. Setze

$$A^{(1)} := A, \text{ und}$$

matrix $A^{(k)}$ hinzufügen.

Wir suchen nun eine orthogonale Matrix $\hat{H}_k \in \mathrm{O}_m(\mathbb{R})$, so dass

$$A^{(k+1)} = \hat{H}_k A^{(k)} \text{ mit } \hat{H}_k = \begin{pmatrix} I_{k-1} & 0 \\ 0 & H_k \end{pmatrix},$$

woher kommen die beiden Matrizen?

Nach lemma 1.4.2 gibt es aber eine Householder-Transformation H_k, so dass

$$H_k \begin{pmatrix} \tilde{a}_{k,k}^{(k)} \\ \vdots \\ \tilde{a}_{m,k}^{(k)} \end{pmatrix} = \begin{pmatrix} -\sigma_k \\ 0 \\ \vdots \end{pmatrix}$$

ist. Iterativ erhalten wir eine Matrix

$$R := \hat{H}_{n^*-1} \dots \hat{H}_1 A \in \mathbb{R}^{m,n}$$

die in verallgemeinerter oberer Dreiecksform ist. Also gilt, weil die $\hat{H}_k$ orthogonal sind,

$$A = \left(\hat{H}_{n^*-1} \dots \hat{H}_1\right)^T R = \hat{H}_1^T \dots \hat{H}_{n^*-1}^T R$$

Also haben wir gezeigt:

Theorem 1.4.3 (Existenz einer QR-Zerlegung)**.** *Zu jeder Matrix $A \in \mathbb{R}^{m,n}$ mit $m > n$ und maximialem Rang gibt es eine orthogonale Matrix $Q \in \mathrm{O}_n(\mathbb{R})$ sowie eine reguläre Matrix $R \in \mathbb{R}m, n$ mit*

$$R \left(\frac{\hat{R}}{0} \right) \quad \textit{und } \hat{R} \in \mathrm{M}_n(\mathbb{R}) \textit{ in oberer Dreiecksform,}$$

so dass

$$A = QR.$$

Example 1.4.4. Betrachte

$$A = \begin{pmatrix} 3 & 5 \\ 0 & 2 \\ 0 & 0 \\ 4 & 5 \end{pmatrix} \in \mathbb{R}4, 2$$

i) Spiegelung von $v_1 := (3, 0, 0, 4)^T$ auf $(-1, 0, 0, 0)^T$: Setze $\sigma_1 = 1$ und es gilt $||v_1|| = 5$. Nach lemma 1.4.2 ist für

End of Lecture 3

A.0 Sheet 0

Definition A.0.1. Die **Frechet-Ableitung** bezeichnet die gewöhnliche totale Ableitung einer Funktion $\mathbb{R}^n \to \mathbb{R}$.

Definition A.0.2. Sei $A \in \mathrm{M}_n(\mathbb{K})$. Der **Spektralradius** von A ist definiert als

$$\rho(A) := \max\{|\lambda_1|, ..., |\lambda_n|\},$$

wobei $\lambda_1, ..., \lambda_n$ die (möglicherweise komplexen) Eigenwerte von A darstellen.

Proposition A.0.3. *Sei $M \in \mathrm{M}_n(\mathbb{R})$, und*

$$\varphi : \mathbb{R}^n \to \mathbb{R}^n, x \mapsto Mx + c.$$

Dann sind äquivalent:

i) Für den Spektralradius p gilt $p(M) < 1$.

ii) Die Fixpunktiteration

$$x_{k+1} := \varphi(x_k)$$

konvergiert für ein beliebiges $x_0 \in \mathbb{R}^n$.

Solution A.0.2. Es gilt für das in der Aufgabenstellung spezifizierte ψ:

$$\{\text{Eigenwerte von } \psi\} = \lambda + (1 - \lambda) \cdot \{\text{Eigenwerte von } \varphi\},$$

wobei jeweils nur der lineare Teil betrachtet wurde. Nutze nun proposition A.0.3.

Proposition A.0.4. *Das Jacobi-Verfahren für die Matrix*

$$A = D - L - R$$

konvergiert, falls für die Matrix

$$I_{Jac.} := D^{-1}(L + R)$$

der Spektralradius größer 1 ist. Es konvergiert nicht, wenn der Spektralradius kleiner 1 ist.

Definition A.0.5. Die **Newton-Iteration** ist gegeben durch

$$x_{k+1} := x_k - \frac{f(x_k)}{f'(x_k)}$$

A.1 Sheet 1

Proposition A.1.1. *Für $A \in \mathbb{R}^{m,n}$ ist die Operatornorm bzgl. der eukldischen Norm gegeben durch*

$$||A||_2 = \rho(A^T A)$$

[Sch18] Jan Schrøer. **Lineare Algebra**. Vorlesungsskript, 2018.