§6.5, Verfahren mit festen Schrittweiten zur Minimierung von konvexen quadratischen Funktionen

(Nach einer gemeinsamen Arbeit mit Melinda Hagedorn [2])

Florian Jarre, Math. Insititut, HHU

7.1.2026

1. Ansatz

Sei $f : \mathbb{R}^n \to \mathbb{R}$ streng konvex und quadratisch. Zur Minimierung von f ausgehend von $x^0 \in \mathbb{R}^n$ setze man $m^0 := -\nabla f(x^0)$ und betrachte folgende "Momentum Methode" für $k \geq 0$:

$$(MM) \qquad x^{k+1} = x^k + \alpha m^k, \qquad m^{k+1} = \beta m^k - \nabla f(x^{k+1}).$$

Hier ist $\alpha > 0$ eine kurze Schrittweite und $\beta \in [0, 1)$ ein Parameter, der wie beim cg-Verfahren festlegt wieviel der alten Suchrichtung zur Neuen addiert wird. Man versteht m^k als "Moment" oder "Schwung" beim Abstieg. Anders als beim cg-Verfahren wird keine line-search ausgeführt.

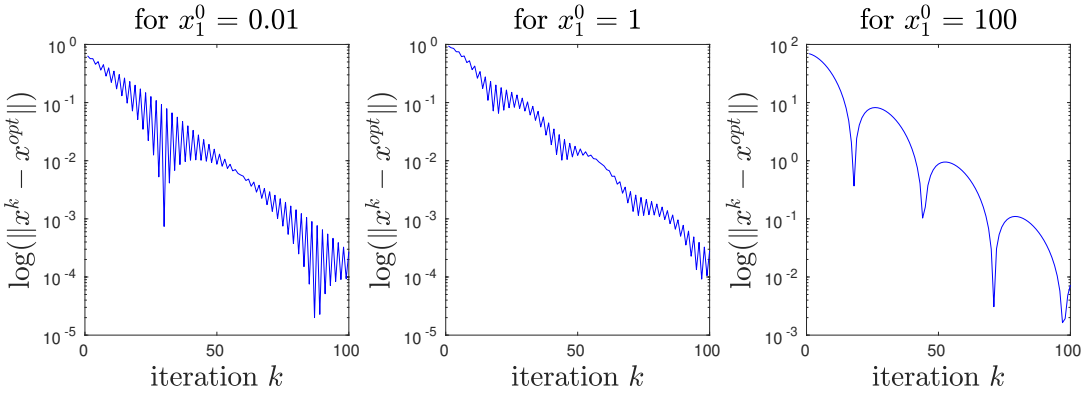
Äquivalent zur Momentum Methode (MM) ist die sogenannte "Heavy Ball Methode", die auf Polyak [5] zurück geht, der schon vor knapp 60 Jahren bewiesen hat, dass diese Methode bei passender Wahl der Parameter α und β "best möglich" ist: $x^{-1} := x^0$ und für $k \geq 0$ setze

$$(HBM) \qquad x^{k+1} = x^k - \alpha \nabla f(x^k) + \beta(x^k - x^{k-1}).$$

Diese Methode hat sich auch in der numerischen Implementierung von stochastischen Gradientenverfahren bewährt, [1, 6]. Modifizierungen von (MM) sind heute weit verbreitet, insbesondere in Bibliotheken zum Maschinellen Lernen.

Allerdings ist das Konvergenzverhalten von (MM) selbst bei Verwendung exakter Gradienten "etwas unregelmäßig". **Abbildung 1** zeigt ein typisches nicht-monotones Konvergenzverhalten von (MM) für das einfache 2-d-Beispiel $f(x) \equiv \frac{1}{2}(x_1^2 + 100x_2^2)$ mit $\beta = 0.85$ und $\alpha = \frac{1.9}{100}$ während der ersten 100 Iterationen für die drei Startpunkte $x^0 = (\frac{1}{100}, 1)^{\mathrm{T}}$, $x^0 = (1, 1)^{\mathrm{T}}$ und $x^0 = (100, 1)^{\mathrm{T}}$.

Abbildung 1: Konvergenz von (MM) für ein 2-d-Beispiel mit $\beta = 0.85$ und $\alpha = 1.9/100$.



Eng verwandt mit (MM) ist Nesterov's beschleunigtes Gradientenverfahren [3] (Nesterov Acclelerated Gradient "(NAG)"). Für den Spezialfall, dass die obigen Voraussetzungen erfüllt sind, kann es wie folgt geschrieben werden: Gegeben $x^0 \in \mathbb{R}^n$ setze $y^0 := x^0$ und für $k \geq 0$:

$$(NAG) \qquad x^{k+1} := y^{k+1} + \beta(y^{k+1} - y^k) \qquad \text{wobei} \qquad y^{k+1} := x^k - \alpha \nabla f(x^k)$$

mit passenden Parametern $\alpha, \beta > 0$. (Für das NAG-Verfahren gibt es zahlreiche Modifikationen, die unter schwächeren Voraussetzungen angewendet werden können oder die mit einer Schrittweitenkontrolle genutzt werden können.) Eliminiert man die Variable y^k und wählt eine passende Initialisierung, so kann (NAG) auch in der folgenden kompakten Form geschrieben werden:

$$x^{k+1} := x^k - \alpha \nabla f(x^k) + \beta \left(x^k - x^{k-1} - \alpha(\nabla f(x^k) - \nabla f(x^{k-1})) \right), \tag{1}$$

wobei der "Momentum-Term" nicht nur die früheren Iterierten x^k und x^{k-1} umfasst sondern auch die zugehörigen Abstiegsschritte "$-\alpha \nabla f(x^k)$" und "$-\alpha \nabla f(x^{k-1})$".

Wir betrachten zunächst wieder (HBM).

Theorem 1 *Sei $f : \mathbb{R}^n \to \mathbb{R}$ streng konvex und quadratisch mit Minimalstelle x^*. Setze $H := \nabla^2 f(x^*)$. Sei $\overline{m}$ eine obere Schranke für die Eigenwerte von H (Gerschgorin) und sei $0 < \underline{m}$ eine untere Schranke für die Eigenwerte von H. (Schwieriger zu bestimmen.) Dann ist $\overline{\text{cond}}(H) := \overline{m}/\underline{m}$ eine obere Schranke für die Kondition von H. (Technische Voraussetzung: $\overline{\text{cond}}(H) \geq 28$).*

In (HBM) sei $\alpha = 2/\overline{m}$, $\beta = \left(1 - \sqrt{2 \,/\, \overline{\text{cond}}(H)}\right)^2$ und sei $\epsilon \leq 1 \,/\, \overline{\text{cond}}(H)$ gegeben. Dann wird ausgehend von $x^0 \in \mathbb{R}^n$ nach

$$1 + \lceil \sqrt{2\ \overline{\text{cond}}(H)}\ \ln(\tfrac{2}{\epsilon}) \rceil$$

Schritten von (HBM), eine Näherungslösung $\bar{x}^k := \frac{1}{2}(x^{k-1} + x^k)$ erzeugt mit

$$\|\bar{x}^k - x^*\|_2 \leq \epsilon \|x^0 - x^*\|_2.$$

Bemerkung:
Man könnte vielleicht annehmen, dass bei einer großen Wahl von $\beta \in [0, 1)$ die Schrittweite α in (MM) vorsichtiger gewählt werden sollte. Überraschenderweise folgt aus der Analyse von (MM) jedoch das Gegenteil: Wenn der Schritt selbst verlängert wird, imdem ein größeres Vielfaches des letzten Suchschritts hinzuaddiert wird, d.h. $\beta \in [0, 1)$ größer gewählt wird, dann kann auch die Schrittweite α vergrößert werden, ohne die globale Konvergenz zu verlieren.

Wenn $\overline{m}$ und $\underline{m}$ bekannt sind, dann kann die Analyse für (MM) auch auf (NAG) angewendet werden.

2. Analyse von (HBM)

Zur Analyse der (HBM) und damit auch von (MM) werden zwei Schritte betrachtet (Weitere Details in [2]):

1. In Abschnitt 2.1 wird eine lineare Rekursion für die Iterierten x^k hergeleitet und der Spektralradius der zugehörigen Systemmatrix analysiert.

2. Darauf aufbauend werden in Abschnitt 2.2 passende Parameter α, β für (HBM) bestimmt.

Sei zunächst die Funktion f gegeben durch die Vorschrift $f(z) \equiv \frac{1}{2} z^T A z - b^T z + \gamma$ mit $A \succ 0$. Sei $A = UDU^T$ mit einer orthogonalen Matrix U und einer Diagonalmatrix D eine Eigenwertzerlegung von A und sei $z^* := A^{-1}b$ sowie $x := U^T(z - z^*)$. Dann folgt

$$\frac{1}{2} x^T D x = \frac{1}{2}(z - z^*)^T U D U^T (z - z^*) = \frac{1}{2}(z - z^*)^T A (z - z^*) =$$

$$\frac{1}{2} z^T A z - \frac{2}{2} z^T A z^* + \frac{1}{2}(z^*)^T A z^* = \frac{1}{2} z^T A z - z^T A A^{-1} b + \frac{1}{2}(z^*)^T A z^* = f(z) + \frac{1}{2}(z^*)^T A z^* - \gamma.$$

In den Übungen wird daraus hergeleitet, dass (HBM) invariant ist unter Verschiebungen und orthogonalen Transformationen, sodass für die Analyse o.B.d.A. vorausgesetzt wird, dass

$$f(x) \equiv \tfrac{1}{2}\, x^T D x \tag{2}$$

mit einer Diagonalmatrix D. Dabei sei bekannt, dass D positiv definit ist, die Diagonaleinträge $D_{i,i}$ seien aber nicht bekannt. (Falls z.B. $f(x) \equiv \frac{1}{2} \|Bx\|_2^2 + b^T x$ mit einer Matrix B mit vollem Zeilenrang, so ist bekannt, dass $B^T B$ nur positive Eigenwerte hat, aber die Eigenwerte selbst sind nicht bekannt. Bei der obigen Umformung wurde benutzt, dass es eine

Orthogonalbasis von Eigenvektoren gibt, die in der Matrix U zusammengefasst sind, aber U kennen wir auch nicht.) Ferner wird im Folgenden vorausgesetzt, dass

$$\beta \in [0,1) \qquad \text{und} \qquad \alpha \in (0, \bar{\alpha}]$$

wobei $\bar{\alpha} := 2/\max_{1\le i\le n}\{D_{i,i}\}$.

Die (HBM) kann mit der folgenden Rekursionsformel ausgedrückt werden

$$\begin{pmatrix} x^k \\ x^{k+1} \end{pmatrix} = \begin{pmatrix} 0 & I \\ -\beta I & (1+\beta)I - \alpha D \end{pmatrix} \begin{pmatrix} x^{k-1} \\ x^k \end{pmatrix}, \tag{3}$$

oder kurz $\hat{z}^{k+1} = \hat{M}\hat{z}^k$ mit der Variablen

$$\hat{z}^k := \begin{pmatrix} x^{k-1} \\ x^k \end{pmatrix} \qquad \text{und} \qquad \hat{M} := \begin{pmatrix} 0 & I \\ -\beta I & (1+\beta)I - \alpha D \end{pmatrix}.$$

Die Rekursion (3) ist "Block-separabel", d.h. für $i \neq j$ hängen die Variablen $x_i^k,\ x_i^{k+1}$ nicht von x_j^ℓ für irgendein $\ell \le k+1$ ab. Setzt man daher

$$z_{(i)}^k := \begin{pmatrix} x_i^{k-1} \\ x_i^k \end{pmatrix}$$

für $k \ge 0$ und $1 \le i \le n$, so lässt sich die Rekursion (3) schreiben als

$$z_{(i)}^{k+1} = M^{(i)} z_{(i)}^k \qquad \text{für } 1 \le i \le n$$

mit

$$M^{(i)} := \begin{pmatrix} 0 & 1 \\ -\beta & 1+\beta-\alpha D_{i,i} \end{pmatrix}. \tag{4}$$

(Anders ausgedrückt, können die Zeilen und Spalten von $\hat{M}$ so permutiert werden dass eine Block-Diagonalmatrix M entsteht mit 2×2 Matrizen $M^{(i)}$ auf der Diagonalen.)

2.1 Der Spektralradius von $M^{(i)}$ in Abhängigkeit von α und β

Die Konvergenz der Iterierten $\hat{z}^k$ hängt von den Normen $\|(M^{(i)})^k\|_2$ für große k ab, d.h. von dem Spektralradius von $\hat{M}$ der mit dem maximalen Spektralradius $\rho(M^{(i)})$ von $M^{(i)}$ für alle i übereinstimmt.

(Falls die Jordansche Normalform von $M^{(i)}$ gegeben ist durch $M^{(i)} = SJS^{-1}$ so ist $(M^{(i)})^k = SJ^kS^{-1}$. Für J sind zwei Fälle möglich: Wenn J eine Diagonalmatrix ist so gilt offensichtlich $\lim_{k\to\infty} J^k = 0$ genau dann wenn der betragsgrößere Eigenwert den Betrag < 1 hat. Falls J ein Jordanblock der Größe 2 zum Eigenwert λ ist, so verifiziert man durch Induktion, dass

$$\begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix}^k = \begin{pmatrix} \lambda^k & k\lambda^{k-1} \\ 0 & \lambda^k \end{pmatrix}.$$

Hier gilt dann wieder $\lim_{k\to\infty} J^k = 0$ genau dann wenn $|\lambda| < 1$ ist, aber die Konvergenz ist etwas langsamer.)

Die nachfolgenden Betrachtungen beziehen sich auf einen fest (aber bieliebig) fixierten Index i. Setze

$$\alpha_i := \alpha D_{i,i} \in (0,2], \quad \beta_i := 1 + \beta - \alpha_i \in [-1,2), \quad \text{und } \gamma_i := \sqrt{\beta_i^2 - 4\beta}.$$

Hier ist γ_i entweder eine nicht-negative reelle Zahl oder eine (rein imaginäre) Zahl mit positivem Imaginärteil. In beiden Fällen wird die Beziehung $\gamma_i^2 = \beta_i^2 - 4\beta$ nachfolgend benutzt. Mit diesen Abkürzungen ist $M^{(i)}$ gegeben durch $M^{(i)} = \begin{pmatrix} 0 & 1 \\ -\beta & \beta_i \end{pmatrix}$. Sei ferner

$$\lambda_+ := \tfrac{1}{2}(\beta_i + \gamma_i), \quad \lambda_- := \tfrac{1}{2}(\beta_i - \gamma_i), \quad v_+ := \begin{pmatrix} 1 \\ \lambda_+ \end{pmatrix}, \text{ und } v_- := \begin{pmatrix} 1 \\ \lambda_- \end{pmatrix}. \tag{5}$$

Unter Beachtung von

$$\lambda_\pm^2 \;=\; \tfrac{1}{4}\beta_i^2 \pm \tfrac{1}{2}\beta_i\gamma_i + \tfrac{1}{4}\gamma_i^2 \;=\; \tfrac{1}{4}\beta_i^2 \pm \tfrac{1}{2}\beta_i\gamma_i + \tfrac{1}{4}\beta_i^2 - \beta \;=\; -\beta + \tfrac{1}{2}\beta_i^2 \pm \tfrac{1}{2}\beta_i\gamma_i \;=\; -\beta + \beta_i\lambda_\pm$$

folgt dass $M^{(i)}v_\pm = \lambda_\pm v_\pm$. Also sind die – möglicherweise komplexen – Eigenwerte von $M^{(i)}$ gegeben durch λ_+ und λ_- und der Spektralradius $\rho(M^{(i)})$ von $M^{(i)}$ ist durch den größeren der beiden Werte

$$|\lambda_\pm| \;=\; \tfrac{1}{2}|\beta_i \pm \gamma_i| \;=\; \frac{1}{2}\left|\beta_i \pm \sqrt{\beta_i^2 - 4\beta}\right|$$

gegeben. Wenn die Wurzel rein imaginär ist, d.h. wenn $\beta_i^2 < 4\beta$ dann ist das Quadrat des Betrags durch die Summe der Quadrate von Real- und Imaginärteil gegeben, d.h.

$$\rho(M^{(i)})^2 = \frac{1}{4}\left(\beta_i^2 + |\gamma_i|^2)\right) = \frac{1}{4}\left(\beta_i^2 + (4\beta - \beta_i^2)\right) = \beta.$$

Somit vereinfacht sich der Ausdruck für $\rho(M^{(i)})$ auf

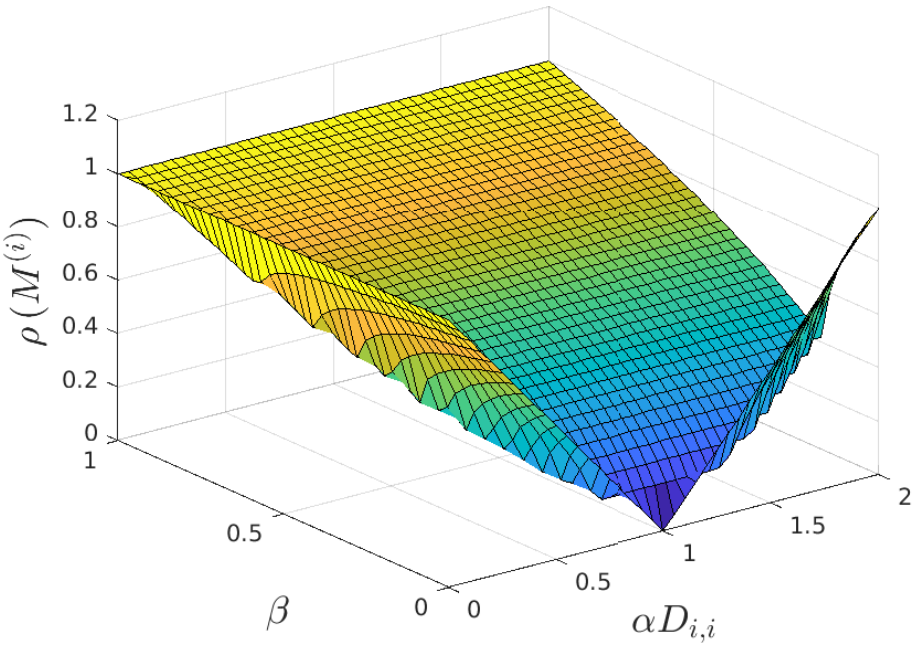
$$\rho := \rho(M^{(i)}) = \begin{cases} \sqrt{\beta} & \text{falls } \beta_i^2 < 4\beta \\ \frac{1}{2}\left(|\beta_i| + \gamma_i\right) & \text{sonst.} \end{cases} \tag{6}$$

In **Abbildung 2** ist der Wert von $\alpha D_{i,i} \in (0,2]$ von der Mitte unten aus nach rechts hin abgetragen und der Wert von β von der Mitte unten aus nach links hinten. Die zugehörigen Werte von ρ sind auf der vertikalen Achse abgetragen.

Nach Voraussetzung ist α kleiner oder gleich $2/\overline{m}$ gewählt, wobei $\overline{m} \geq \max_i D_{i,i}$, so dass die möglichen Werte von $\alpha D_{i,i}$ in einem (oft recht großen) Bereich im Interval (0,2] liegen. Dieser Bereich hängt vom Problem ab und kann durch die Wahl von β nicht beeinflusst werden.

Eine schnelle Konvergence von HBM (oder MM) ergibt sich wenn der Wert von β so gewählt ist, dass ρ für eine breite Wahl von möglichen $\alpha D_{i,i}$ klein ist. **Abbildung 2** mag

Abbildung 2: Spektralradius von $M^{(i)}$ als Funktion von $\alpha D_{i,i} \in (0, 2]$ und $\beta \in [0, 1)$.



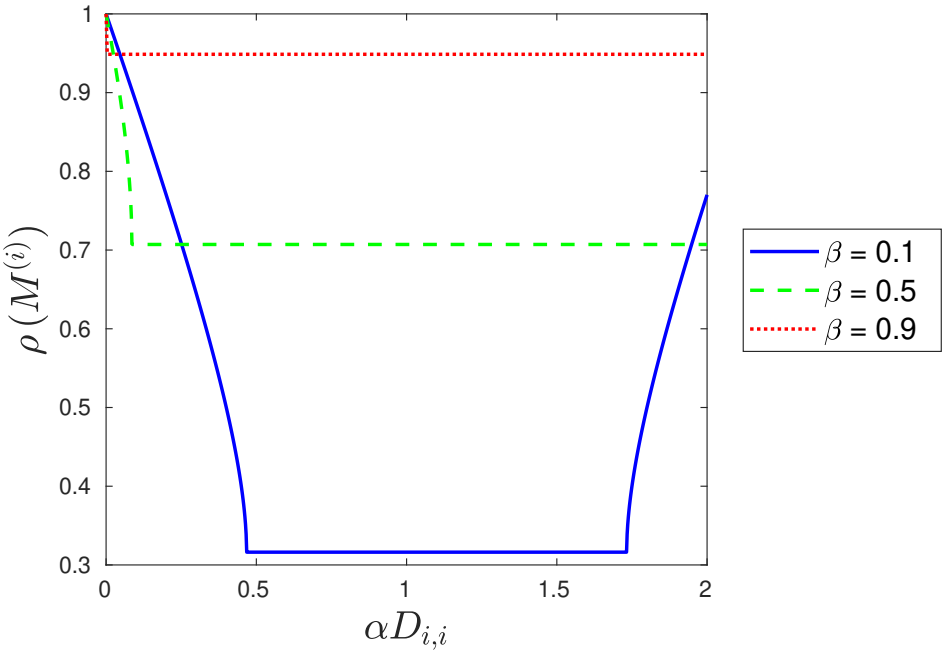
den Eindruck vermitteln, dass dies erreicht wird, wenn $\beta = 0$ gewählt wird. In der Tat, wenn die Werte von $\alpha D_{i,i}$ alle in der Nähe von "1", liegen, dann ist nicht nur das Problem f zu minimieren ziemlich einfach (weil die Konditionszahl D nahe bei 1 liegt) aber auch die Wahl $\beta = 0$ ist dann nahezu optimal.

Wenn die Konditionszahl von D aber schlecht ist, so treten auch sehr kleine Werte von $\alpha D_{i,i} > 0$ auf. In diesem Fall eignet sich Abbildung 2 nicht, um die best mögliche Wahl von β zu verstehen. Statt dessen zeigt **Abbildung 3** die Werte von $\rho(M^{(i)})$ als Funktion von $\alpha D_{i,i} \in (0, 2]$ für $\beta \equiv 0.9$ in rot, $\beta \equiv 0.5$ in grün, und $\beta \equiv 0.1$ in blau.

In **Abbildung 3** führt die Wahl $\beta = 0.1$ in blau zu kleinen Werten von $\rho(M^{(i)})$ wenn $\alpha D_{i,i} \in [0.5, 1.5]$ (und etwas uaßerhalb dieses Intervalls). Für Werte $\alpha D_{i,i} \approx 0.2$ resultiert die wahl $\beta = 0.5$ in grün in einem viele kleineren Wert won $\rho(M^{(i)})$ als $\beta = 0.1$, und wie in **Abbildung 4** dargestellt, wenn schlecht konditionierte Probleme betrachtet werden mit Werten von $\alpha D_{i,i} \approx 0.004$ dann resultiert die Wahl $\beta = 0.9$ in rot in einem kleineren Wert von $\rho(M^{(i)})$ asl $\beta = 0.5$ oder $\beta = 0.1$. (Abbildung 4 ist ein Zoom von Abbildung 3)

In den Übungen wird gezeigt, dass eine Schrittweite $\alpha D_{i,i} > 2$ bei dem Verfahren des steilsten Abstiegs ($\beta = 0$) "zu lang" ist in dem Sinn, dass es bei diesen Schrittweiten nicht mehr konvergent ist. Aber für größere Werte $\beta \geq 0.5$, setzen sich die grüne und die rote Linie noch etwas rechts von $\alpha D_{i,i} = 2$ fort, sodass, wie eingangs erwähnt, für $\beta \geq 0.5$ sogar Schrittweiten gewählt werden können, die für das reine steepest-descent-Verfahren zu lang

Abbildung 3: $\rho(M^{(i)})$ als Funktion von $\alpha D_{i,i} \in (0, 2]$ für verschiedene Werte von β.



sind.

2.2 Zur Wahl von α und β

Nachfolgend wird vorausgesetzt, dass eine untere Schranke

$$0 < \underline{m} \leq \min_{1 \leq i \leq n} D_{i,i} \qquad \text{und eine obere Schranke} \qquad \overline{m} \geq \max_{1 \leq i \leq n} D_{i,i}$$

bekannt sind. (Gerade die untere Schranke ist für f in der Form $f(x) \equiv \frac{1}{2}x^T Ax - bx$ oft schwer zu bestimmen.) Die Schrittweite α wird dann fixiert auf

$$\alpha := 2/\overline{m}. \tag{7}$$

Dann ist

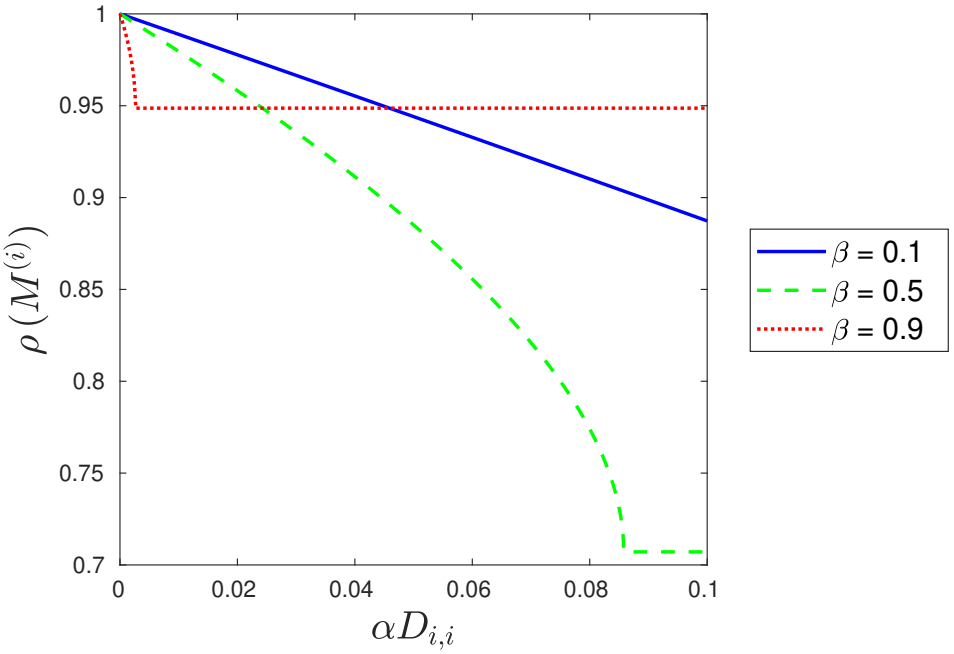
$$\overline{\text{cond}}(D) \quad := \quad \overline{m}/\underline{m} \quad \geq \quad \text{cond}(D)$$

eine obere Schranke für $\text{cond}(D)$. Ferner wird angenommen, dass

$$\overline{\text{cond}}(D) > 2. \tag{8}$$

(Die Voraussetzung in Theorem 1 ist noch etwas strenger.)

Abbildung 4: $\rho(M^{(i)})$ als Funktion von $\alpha D_{i,i} \in (0, 0.1)$ für verschiedene Werte von β.



Der Spektralradius von $M^{(i)}$ stimmt laut (6) unabhängig von α_i immer mit $\sqrt{\beta}$ überein, sofern nur $\beta_i^2 \leq 4\beta$ erfüllt ist, wobei $\beta_i = 1+\beta-\alpha_i$ und $\alpha_i \in (0, 2]$. Die Gleichung $\beta_i^2 - 4\beta = 0$ lässt sich unter Ausnutzen der Definition von β_i nach β auflösen:

$$\beta_i^2 - 4\beta = 0 \iff ((1-\alpha_i)+\beta)^2 - 4\beta = 0 \iff \beta^2 + 2(1-\alpha_i)\beta + (1-\alpha_i)^2 - 4\beta = 0$$

$$\iff \beta^2 - 2(1+\alpha_i)\beta + (1-\alpha_i)^2 = 0 \iff \beta = 1+\alpha_i \pm \sqrt{(1+\alpha_i)^2 - (1-\alpha_i)^2} \Longleftarrow \beta = (1-\sqrt{\alpha_i})^2,$$

wobei im letzten Schritt der Wert $\beta < 1$ gewählt wurde, (sodass der Spektralradius $\sqrt{\beta}$ auch kleiner als 1 ist). Wenn also $\beta \in [0, 1]$ größer oder gleich dem größten Wert aller $(1-\sqrt{\alpha_i})^2$ gewählt wird, so ist die Konvergenzrate gleich $\sqrt{\beta}$.

Sei $\underline{\alpha} = \alpha \underline{m}$ eine untere Schranke für mögliche Werte von $\alpha D_{i,i}$. Voraussetzung (8) impliziert $\underline{\alpha} < 1$. Mit $\beta := (1-\sqrt{\underline{\alpha}})^2$ ist der Spektralradius aller $M^{(i)}$ (und damit $\rho(M)$)) von oben beschränkt durch $\sqrt{\beta} = 1 - \sqrt{\underline{\alpha}}$, d.h.

$$\rho(M) \leq \sqrt{\beta} = 1 - \sqrt{\underline{\alpha}} = 1 - \frac{\sqrt{2}}{\sqrt{\overline{\text{cond}}(D)}}.$$

Nun lässt sich die Gleichung

$$\left(1 - \sqrt{\frac{2}{\overline{\text{cond}}(D)}}\right)^k \leq \epsilon$$

nachfolgend umformen,

$$\begin{aligned}
&\iff k \ln\left(1 - \sqrt{\frac{2}{\overline{\text{cond}}(D)}}\right) \leq \ln \epsilon \\
&\impliedby k \left(-\sqrt{\frac{2}{\overline{\text{cond}}(D)}}\right) \leq \ln \epsilon \\
&\iff k \geq \ln(\frac{1}{\epsilon}) \sqrt{\frac{\overline{\text{cond}}(D)}{2}}.
\end{aligned}$$

Der Spektralradius liefert aber nur asymptotische Aussagen für große Iterationszahlen k und berücksichtigt nicht, dass die Transformtaionsmatrizen für die Eigenwertzerlegung schlecht konditioniert sein können oder dass ein Jordan-Block der Größe 2 auftreten kann. Der vollständige Beweis von Theorem 1 in [2] ist etwas technisch und sprengt den Rahmen dieser Vorlesung.

2.3 Nesterov's beschleunigtes Gradientenverfahren

Für the Analyse of Nesterov's beschleunigtem Gradientenverfahren genügt es auch, die Funktion f in (2) mit $\nabla f(x) = Dx$ zu betrachten. Ersetzt man (HBM) durch (NAG) in der Form (1) so ist die Matrix $M^{(i)} = \begin{pmatrix} 0 & 1 \\ -\beta & \beta_i \end{pmatrix}$ in (4) zu ersetzen durch

$$M^{(i)} := \begin{pmatrix} 0 & 1 \\ -\beta(1-\alpha_i) & (1+\beta)(1-\alpha_i) \end{pmatrix}$$

wieder mit $\alpha_i := \alpha D_{i,i}$. Ersetzt man also β und β_i in (4) durch $\beta(1-\alpha_i)$ und $(1+\beta)(1-\alpha_i)$ so folgt, dass ρ in (6) ersetzt wird durch

$$\rho := \rho(M^{(i)}) = \begin{cases} \sqrt{\beta(1-\alpha_i)} & \text{falls } \ (1+\beta)^2(1-\alpha_i)^2 \leq 4\beta(1-\alpha_i) \\ \frac{1}{2}\left(|(1+\beta)(1-\alpha_i)| + \bar{\gamma}_i\right) & \text{sonst} \end{cases} \tag{9}$$

wobei $\bar{\gamma}_i := \sqrt{(1+\beta)^2(1-\alpha_i)^2 - 4\beta(1-\alpha_i)}$.

Der erste Fall von Definition (9) kann nur auftreten wenn $\alpha_i \leq 1$. Daraus ergibt sich die halbe Schrittweite $\alpha := \frac{1}{\overline{m}}$ im Vergleich zu (7) für (HBM), und daraus folgt für $\alpha_i \in [1/\overline{\text{cond}}(D), 1]$ und $\beta := \frac{\left(\sqrt{\overline{\text{cond}}(D)}-1\right)^2}{\overline{\text{cond}}(D)-1}$ dass $(1+\beta)^2(1-\alpha_i) \leq 4\beta$ und

$$\rho = \sqrt{\beta(1-\alpha_i)} \leq 1 - \frac{1}{\sqrt{\overline{\text{cond}}(D)}}.$$

Dies ist dieselbe Schranke wie für (MM) wenn dort die halbe Schrittweite α gewählt wird oder – was auf dasselbe hinausläuft – wenn $\overline{m}$ in (MM) durch $2\overline{m}$ ersetzt wird.

Die Eigenwerte von $M^{(i)}$ sind dann gegeben durch

$$\lambda_{\pm} = \tfrac{1}{2}(\ (1+\beta)(1-\alpha_i) \pm \sqrt{(1+\beta)^2(1-\alpha_i)^2 - 4\beta(1-\alpha_i)}\ ).$$

In [2] wird damit eine etwas schwächere Komplexitätsaussage hergeleitet wie für (MM); die asymptotische Konvergenzrate ist

- $1 - \sqrt{1/\overline{\text{cond}}(H)}$ für (NAG) und
- $1 - \sqrt{2/\overline{\text{cond}}(H)}$ für (HBM).

(Die Rate für (HBM) ist in [5], Theorem 9 (3) noch etwas besser; $1 - 2/\sqrt{\overline{\text{cond}}(H)}$ da dort noch etwas längere Schrittweiten α betrachtet werden.)

Auch bei Nutzung von stochastischen Gradienten hat sich die praktische Anwendung der Momentum Methode als sehr effizient und robust erwiesen.

Literatur

[1] Defazio, A. (2021): Momentum via Primal Averaging: Theoretical Insights and Learning Rate Schedules for Non-Convex Optimization. https://arxiv.org/pdf/2010.00406.pdf

[2] Hagedorn, M., Jarre, F. (2023): Iteration Complexity of Fixed-Step Methods by Nesterov and Polyak for Convex Quadratic Functions. J Optim Theory Appl. https://doi.org/10.1007/s10957-023-02261-w

[3] Nesterov, Y.E. (1983): A method for solving the convex programming problem with convergence rate $O(1/k^2)$. Dokl. akad. nauk Sssr 269, 543-547.

[4] Nesterov, Y.E. (2003): Introductory lectures on convex optimization: A basic course. Springer Science and Business Media, Vol. 87.

[5] Polyak, B.T. (1964): Some methods of speeding up the convergence of iteration methods. USSR Computational Mathematics and Mathematical Physics, 4(5):1-17.

[6] Sebbouh, O., Gower, R.M., Defazio, A. (2020): On the convergence of the Stochastic Heavy Ball Method. https://othmanesebbouh.github.io/publications/heavy_ball.pdf