Heavy ball method on convex quadratic problem $\min_{\mathbf{x}} \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} - \mathbf{b}^\top \mathbf{x}$

A case study

Andersen Ang

Mathématique et recherche opérationnelle
UMONS, Belgium

manshun.ang@umons.ac.be Homepage: angms.science

First draft: June 26, 2018
Last update : July 28, 2019

Overview

An inverse problem / unconstrained optimization problem

Given $\mathbf{A} \in \mathbf{R}^{n\times n}$, $\mathbf{b} \in \mathbf{R}^{n\times 1}$, find $\mathbf{x} \in \mathbf{R}^{n\times 1}$ by

$$(\mathcal{P}_0): \quad \min_{\mathbf{x}} \frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2$$

$(\mathcal{P}_0)$ is equivalent to the quadratic problem $\min_{\mathbf{x}} f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{A}'\mathbf{x} - \mathbf{b}'^\top \mathbf{x}$.

$$\begin{aligned}
\frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2 &= \frac{1}{2}(\mathbf{A}\mathbf{x}-\mathbf{b})^\top(\mathbf{A}\mathbf{x}-\mathbf{b}) \\
(\text{expand}) &= \frac{1}{2}\Big(\mathbf{x}^\top\mathbf{A}^\top\mathbf{A}\mathbf{x} - \mathbf{x}^\top\mathbf{A}^\top\mathbf{b} - \mathbf{b}^\top\mathbf{A}\mathbf{x} + \mathbf{b}^\top\mathbf{b}\Big) \\
(\mathbf{a}^\top\mathbf{b} = \mathbf{b}^\top\mathbf{a}) &= \frac{1}{2}\Big(\mathbf{x}^\top\mathbf{A}^\top\mathbf{A}\mathbf{x} - 2\mathbf{b}^\top\mathbf{A}\mathbf{x} + \|\mathbf{b}\|_2^2\Big) \\
(\mathbf{A}' = \mathbf{A}^\top\mathbf{A},\ \mathbf{b}' = \mathbf{A}^\top\mathbf{b}) &= \frac{1}{2}\mathbf{x}^\top\mathbf{A}'\mathbf{x} - \mathbf{b}'^\top\mathbf{x} + \frac{1}{2}\|\mathbf{b}\|_2^2
\end{aligned}$$

Ignoring constant $\frac{1}{2}\|\mathbf{b}\|_2^2$, denote $\mathbf{A}'$ as $\mathbf{A}$ and $\mathbf{b}'$ as $\mathbf{b}$, we now focus on the equivalent problem

$$(\mathcal{P}): \quad \min_{\mathbf{x}} f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top\mathbf{A}\mathbf{x} - \mathbf{b}^\top\mathbf{x}.$$

The convex quadratic problem

$$(\mathcal{P}): \quad \min_{\mathbf{x}} f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} - \mathbf{b}^\top \mathbf{x}.$$

- Properties of f :
  - f is convex with respect to (w.r.t) $\mathbf{x}$
  - f is differentiable w.r.t. $\mathbf{x}$
    - First order derivative (gradient) : $\nabla_{\mathbf{x}} f(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b}$
    - Second order derivative (Hessian) : $\nabla^2_{\mathbf{x}} f(\mathbf{x}) = \mathbf{A}$
- Assumption 1 : $\mathbf{A}$ is positive definite and symmetric

  Consequence of the assumption :
  - all eigenvalues of $\mathbf{A}$ are positive
  - $\mathbf{A}$ is nonsingular $\implies$ optimal solution $\mathbf{x}^*$ exists, which is $\mathbf{x}^* = \mathbf{A}^{-1}\mathbf{b}$
- We can further assume $l\mathbf{I} \preceq A \preceq L\mathbf{I}$

Gradient descent

GD solves $(\mathcal{P}): \min_{\mathbf{x}} f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} - \mathbf{b}^\top \mathbf{x}$ by generating the sequence

$$\{\mathbf{x}_k\}_{k\in\mathbb{N}} \ : \ \mathbf{x}_{k+1} = \mathbf{x}_k - t_k \nabla_{\mathbf{x}} f(\mathbf{x}_k)$$

where k is iteration $(k = 1, 2, ...)$ and t_k is step size

The sequence $\mathbf{x}_k$ converges to $\mathbf{x}^*$ at a linear rate (in optimization). The convergence is illustrated by showing the distance function $\|\mathbf{x}_k - \mathbf{x}^*\|$ is monotonically decreasing as k increases, under a suitable step size t_k

Theorem (GD converge at linear rate) Consider the problem

$$\min_{\mathbf{x}} \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} - \mathbf{b}^\top \mathbf{x}$$

with $\mathbf{A}$ being pd and $l\mathbf{I} \preceq A \preceq L\mathbf{I}$, we have

$$\|\mathbf{x}_k - \mathbf{x}^*\|_2 \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^k \|\mathbf{x}_0 - \mathbf{x}^*\|_2$$

where κ is the conditional number of $\mathbf{A}$: i.e. $\kappa = \frac{L}{l}$

Useful material before the proof

As $\mathbf{x}^* = \mathbf{A}^{-1}\mathbf{b}$, we have

$$\mathbf{b} = \mathbf{A}\mathbf{x}^* \tag{1}$$

As $\nabla_{\mathbf{x}} f(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b}$, we have

$$\mathbf{x}_k - t_k \nabla_{\mathbf{x}} f(\mathbf{x}_k) = \mathbf{x}_k - t_k(\mathbf{A}\mathbf{x}_k - \mathbf{b}) \tag{2}$$

Put (1) into (2) we have

$$\begin{aligned} \mathbf{x}_k - t_k \nabla_{\mathbf{x}} f(\mathbf{x}_k) &= \mathbf{x}_k - t_k(\mathbf{A}\mathbf{x}_k - \mathbf{A}\mathbf{x}^*) \\ &= (\mathbf{I} - t_k\mathbf{A})\mathbf{x}_k + t_k\mathbf{A}\mathbf{x}^* \end{aligned} \tag{3}$$

With these we can now prove the convergence, starting with the distance function $\|\mathbf{x}_{k+1} - \mathbf{x}^*\|_2$

Convergence rate of Gradient Descent in 1 slide

$$
\begin{aligned}
\|\mathbf{x}_{k+1} - \mathbf{x}^*\|_2 &= \|\mathbf{x}_k - t_k \nabla_{\mathbf{x}} f(\mathbf{x}_k) - \mathbf{x}^*\|_2 \\
\text{by } (3) &= \|(\mathbf{I} - t_k \mathbf{A})\mathbf{x}_k + t_k \mathbf{A}\mathbf{x}^* - \mathbf{x}^*\|_2 \\
&= \|(\mathbf{I} - t_k \mathbf{A})(\mathbf{x}_k - \mathbf{x}^*)\|_2 \\
&\leq \|\mathbf{I} - t_k \mathbf{A}\|_2 \|\mathbf{x}_k - \mathbf{x}^*\|_2 \\
&\leq (1 - t_k l)\|\mathbf{x}_k - \mathbf{x}^*\|_2 \\
&\leq (1 - tl)^k \|\mathbf{x}_0 - \mathbf{x}^*\|_2 \\
&= \left(\frac{L - l}{L + l}\right)^k \|\mathbf{x}_0 - \mathbf{x}^*\|_2 \\
&= \left(\frac{\kappa - 1}{\kappa + 1}\right)^k \|\mathbf{x}_0 - \mathbf{x}^*\|_2
\end{aligned}
$$

- 1st line : by definition of GD $\mathbf{x}_{k+1} = \mathbf{x}_k - t_k \nabla_{\mathbf{x}} f(\mathbf{x}_k)$
- 4th line : by operator norm inequality $\|\mathbf{A}\mathbf{x}\|_2 \leq \|\mathbf{A}\|_2 \|\mathbf{x}\|_2$
- 5th line : by $l\mathbf{I} \preceq A \preceq L\mathbf{I} \implies (1 - t_k L)\mathbf{I} \preceq \mathbf{I} - t_k \mathbf{A} \preceq (1 - t_k l)\mathbf{I}$
- 6th line : if constant step size is used $t_k = t$
- 7th line : pick $t = \frac{2}{L+l}$, $1 - tl = \frac{L-l}{L+l}$
- 8th line : $\kappa = \frac{L}{l} \geq 1$ is the conditional number of $\mathbf{A}$.

Polyak's Heavy Ball Method (HBM)

HBM adds a momentum term in GD

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \nabla_{\mathbf{x}} f(\mathbf{x}_k) + \underbrace{\beta_k(\mathbf{x}_k - \mathbf{x}_{k-1})}_{\text{HBM momentum}}$$

- i.e. gradient descent with momentum $\beta_k(\mathbf{x}_k - \mathbf{x}_{k-1})$
- $\beta_k \geq 0$ is the momentum parameter / extrapolation parameter
- α_k acts as the step size t_k in GD
- When $\beta_k = 0$, HBM reduces to GD
- As update direction is perturbed by the momentum, HBM is not monotone : objective function value may increase
- However, overall speaking HBM converges faster than GD (Will prove it soon)

Comparing GD, HBM and Nesterov's acceleration

Compared to HBM, Nesterov's accelerated gradient compute the gradient after applying the momentum

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \nabla_{\mathbf{x}} f\big(\mathbf{x}_k + \beta_k(\mathbf{x}_k - \mathbf{x}_{k-1})\big) + \beta_k(\mathbf{x}_k - \mathbf{x}_{k-1})$$

Consider the following notations :

$$\begin{aligned} \mathbf{x}^+ &= \mathbf{x} - t\nabla f(\mathbf{x}) \\ \mathbf{a}_k &= \beta_k(\mathbf{x}_k - \mathbf{x}_{k-1}) \end{aligned}$$

Then

$$\begin{aligned} \text{Cauchy Gradient Descent} \quad & \mathbf{x}_{k+1} = \mathbf{x}_k^+ \\ \text{Polyak HBM} \quad & \mathbf{x}_{k+1} = \mathbf{x}_k^+ + \mathbf{a}_k \\ \text{Nesterov acceleration} \quad & \mathbf{x}_{k+1} = (\mathbf{x}_k + \mathbf{a}_k)^+ \end{aligned}$$

Open question : may be ?

$$\mathbf{x}_{k+1} = (\mathbf{x}_k + \mathbf{a}_k)^+ + \mathbf{b}_k, \quad \mathbf{x}_{k+1} = \big((\mathbf{x}_k + \mathbf{a}_k)^+ + \mathbf{b}_k\big)^+, \quad \dots$$

Convergence of Heavy Ball Method

Consider $\mathbf{x}_{k+1} - \mathbf{x}^*$. By definition of HBM update :

$$\mathbf{x}_{k+1} - \mathbf{x}^* = \mathbf{x}_k - \alpha_k \nabla_{\mathbf{x}} f(\mathbf{x}_k) + \beta_k(\mathbf{x}_k - \mathbf{x}_{k-1}) - \mathbf{x}^*$$

As $\nabla_{\mathbf{x}} f(\mathbf{x}_k) = \mathbf{A}\mathbf{x}_k - \mathbf{b}$ and $\mathbf{b} = \mathbf{A}\mathbf{x}^*$, we have $\nabla_{\mathbf{x}} f(\mathbf{x}_k) = \mathbf{A}\mathbf{x}_k - \mathbf{A}\mathbf{x}^*$

$$\begin{aligned}
\mathbf{x}_{k+1} - \mathbf{x}^* &= \mathbf{x}_k - \alpha_k(\mathbf{A}\mathbf{x}_k - \mathbf{A}\mathbf{x}^*) + \beta_k(\mathbf{x}_k - \mathbf{x}_{k-1}) - \mathbf{x}^* \\
&= \mathbf{x}_k - \mathbf{x}^* - \alpha_k \mathbf{A}(\mathbf{x}_k - \mathbf{x}^*) + \beta_k(\mathbf{x}_k - \mathbf{x}_{k-1}) \\
&= (\mathbf{I} - \alpha_k \mathbf{A})(\mathbf{x}_k - \mathbf{x}^*) + \beta_k(\mathbf{x}_k - \mathbf{x}_{k-1} - \mathbf{x}^* + \mathbf{x}^*) \\
&= (\mathbf{I} - \alpha_k \mathbf{A})(\mathbf{x}_k - \mathbf{x}^*) - \beta_k(\mathbf{x}_{k-1} - \mathbf{x}^*) + \beta(\mathbf{x}_k - \mathbf{x}^*) \\
&= \Big((1+\beta_k)\mathbf{I} - \alpha_k \mathbf{A}\Big)(\mathbf{x}_k - \mathbf{x}^*) - \beta_k(\mathbf{x}_{k-1} - \mathbf{x}^*)
\end{aligned}$$

In this sense, we have to consider $\mathbf{x}_k - \mathbf{x}^*$ and $\mathbf{x}_{k-1} - \mathbf{x}^*$ at the same time

$$\begin{bmatrix} \mathbf{x}_{k+1} - \mathbf{x}^* \\ \mathbf{x}_k - \mathbf{x}^* \end{bmatrix} = \underbrace{\begin{bmatrix} (1+\beta_k)\mathbf{I} - \alpha_k \mathbf{A} & -\beta_k \mathbf{I} \\ \mathbf{I} & 0 \end{bmatrix}}_{\mathbf{T}_k(\alpha,\beta)} \begin{bmatrix} \mathbf{x}_k - \mathbf{x}^* \\ \mathbf{x}_{k-1} - \mathbf{x}^* \end{bmatrix},$$

$\mathbf{T}_k(\alpha, \beta)$ is the transition matrix

Convergence of Heavy Ball Method - Transition matrix $\mathbf{T}$

Compact expression

$$\begin{bmatrix} \mathbf{x}_{k+1} - \mathbf{x}^* \\ \mathbf{x}_k - \mathbf{x}^* \end{bmatrix} = \mathbf{T}_k(\alpha, \beta) \begin{bmatrix} \mathbf{x}_k - \mathbf{x}^* \\ \mathbf{x}_{k-1} - \mathbf{x}^* \end{bmatrix}$$

Take constant constant α_k and β_k in $\mathbf{T}_k$, so $\mathbf{T}_k = \mathbf{T}$ and

$$\begin{bmatrix} \mathbf{x}_{k+1} - \mathbf{x}^* \\ \mathbf{x}_k - \mathbf{x}^* \end{bmatrix} = \mathbf{T}^k \begin{bmatrix} \mathbf{x}_1 - \mathbf{x}^* \\ \mathbf{x}_0 - \mathbf{x}^* \end{bmatrix}$$

Take norm

$$\left\| \begin{bmatrix} \mathbf{x}_{k+1} - \mathbf{x}^* \\ \mathbf{x}_k - \mathbf{x}^* \end{bmatrix} \right\|_2 = \left\| \mathbf{T}^k \begin{bmatrix} \mathbf{x}_1 - \mathbf{x}^* \\ \mathbf{x}_0 - \mathbf{x}^* \end{bmatrix} \right\|_2 \leq \|\mathbf{T}^k\|_2 \left\| \begin{bmatrix} \mathbf{x}_1 - \mathbf{x}^* \\ \mathbf{x}_0 - \mathbf{x}^* \end{bmatrix} \right\|_2$$

So if $\|\mathbf{T}^k\|_2$ is bounded, the series $\mathbf{x}_k$ produced by HBM converges.

Tools for bounding $\|\mathbf{T}^k\|_2$

Recall

- Spectrum (all eigenvalues) of a block diagonal matrix are the eigenvalues of the block submatrices.
- Spectrum of a matrix are the roots of characteristic equation.
- For 2-by-2 matrix, the characteristic equation is in the form $ax^2 + bx + c = 0$ with roots $x = \frac{1}{2}(-b \pm \sqrt{b^2 - 4ac})$. The roots are complex conjugate if $\Delta = b^2 - 4ac \leq 0$
- Complex roots in the form $a + ib$ share same magnitude as $\sqrt{a^2 + b^2}$

We need two lemmas

Lemma 1. For a $n \times n$ matrix $\mathbf{T}$, there exists a sequences $\varepsilon_k \geq 0$ that

$$\|\mathbf{T}^k\| \leq (\rho(\mathbf{T}) + \varepsilon_k)^k$$

Lemma 2. For $\beta > (1 - \sqrt{\alpha L})^2$, $\rho(\mathbf{T}) < \beta$.

where $\rho(\mathbf{T}) = \max\{|\lambda_1|, |\lambda_2|, ..., |\lambda_n|\}$ is the spectral radius of matrix $\mathbf{T}$, and λ_i are the eigenvalues of $\mathbf{T}$

The logic flow of bounding $\|\mathbf{T}^k\|_2$

Ultimate goal : show $\mathbf{x}_k$ produced by HBM converges to $\mathbf{x}^*$

- $\mathbf{x}_k$ produced by HBM converges to $\mathbf{x}^*$ if $\|\mathbf{T}^k\|_2$ is bounded
- We use lemma 1 to bound $\|\mathbf{T}^k\|$.
- To use lemma 1, we need $\rho(\mathbf{T})$, which we use lemma 2

We will not prove lemma 1 but lemma 2.

Lemma 1. For a $n \times n$ matrix $\mathbf{T}$, there exists a sequences $\varepsilon_k \geq 0$ that

$$\|\mathbf{T}^k\| \leq (\rho(\mathbf{T}) + \varepsilon_k)^k$$

where $\lim_{k\to\infty} \varepsilon_k = 0$.

Proof. Skipped (too long).

The logic of bounding $\|\mathbf{T}^k\|_2$

Lemma 2. For $\beta > (1 - \sqrt{\alpha L})^2$, $\rho(\mathbf{T}) < \beta$.

Flow of proving lemma 2 :

- First show $\|\mathbf{T}\|$ can be decomposed into blocks $\|\mathbf{T}_i\|$
- Then spectrum of $\mathbf{T}$ are the eigenvalues of $\mathbf{T}_i$
- As $\rho(\mathbf{T})$ is considering on the magnitude of eigenvalues, so we consider the magnitude of the eigenvalues of $\mathbf{T}_i$
- $\mathbf{T}_i$ is 2-by-2 matrix, so the eigenvalues are the root of characteristic equation in the form $ax^2 + bx + c = 0$
- Roots of $ax^2 + bx + c = 0$ are complex conjugate that share the same magnitude if $b^2 \leq 4ac$

Proving lemma 2 - eigendecomposition

Lemma 2. For $\beta > (1 - \sqrt{\alpha L})^2$, $\rho(\mathbf{T}) < \beta$.

Proof. First assume $\beta \geq (1 - \sqrt{\alpha L})^2$.
As $\mathbf{A}$ is pd, $\mathbf{A}$ has eigendecomposition as $\mathbf{V}\Lambda\mathbf{V}^\top$. Then

$$\begin{aligned}\mathbf{T} &= \begin{bmatrix}(1+\beta_k)\mathbf{I} - \alpha_k\mathbf{A} & -\beta_k\mathbf{I} \\ \mathbf{I} & 0\end{bmatrix} \\ &= \begin{bmatrix}(1+\beta_k)\mathbf{I} - \alpha_k\mathbf{V}\Lambda\mathbf{V}^\top & -\beta_k\mathbf{I} \\ \mathbf{I} & 0\end{bmatrix}\end{aligned}$$

As $\mathbf{T}$ is diagonal and $\mathbf{V}$ forms a basis, so

$$\|\mathbf{T}\| = \left\|\begin{bmatrix}(1+\beta_k)\mathbf{I} - \alpha_k\Lambda & -\beta_k\mathbf{I} \\ \mathbf{I} & 0\end{bmatrix}\right\|$$

Proving lemma 2 - block decomposition of $\mathbf{T}$

Note $\mathbf{T}$ is block diagonal, it can be decomposed into blocks.
e.g. when $n = 2$, let $x_i = 1 + \beta - \alpha\lambda_i$

$$\|\mathbf{T}\| = \left\| \begin{bmatrix} x_1 & & -\beta & \\ & x_2 & & -\beta \\ 1 & & & \\ & 1 & & \end{bmatrix} \right\|$$

As norm is invariant under permutation, we can swap :
row 3 $\iff$ row 2, column 3 $\iff$ column 2

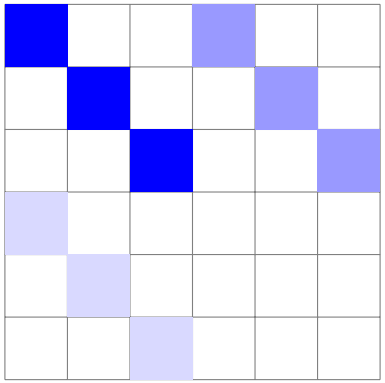
$$\rightarrow \left\| \begin{bmatrix} x_1 & & -\beta & \\ 1 & & & \\ & x_2 & & -\beta \\ & 1 & & \end{bmatrix} \right\| \rightarrow \left\| \begin{bmatrix} x_1 & -\beta & & \\ 1 & & & \\ & & x_2 & -\beta \\ & & 1 & \end{bmatrix} \right\|$$

We can see $\mathbf{T}$ can be decomposed into a block diagonal matrix consists of 2×2 component T_i as

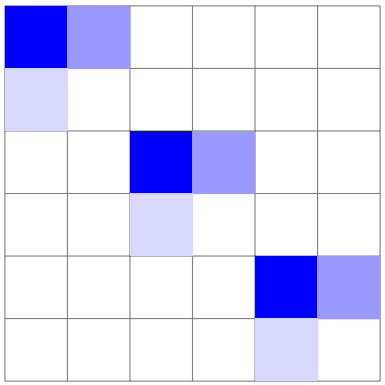
$$\mathbf{T} = \begin{bmatrix} \mathbf{T}_1 & & \\ & \mathbf{T}_2 & \\ & & \ddots \end{bmatrix}, \quad \text{where } \mathbf{T}_i = \begin{bmatrix} 1 + \beta - \alpha\lambda_i & -\beta \\ 1 & \end{bmatrix}$$

Proving lemma 2 - block decomposition of $\mathbf{T}$

Structure of $\mathbf{T}$ in general



After transformation



Hence, spectrum of $\mathbf{T}$ = all the eigenvalues of all $\mathbf{T}_i$

Proving lemma 2 - spectrum of $\mathbf{T}$ = eigenvalues of $\mathbf{T}_i$

$\mathbf{T}_i$ is 2-by-2 $\implies$ we use the roots of the characteristic equation to find its eigenvalues, which is to solve $\det(\mathbf{T}_i - u\mathbf{I}) = 0$:

$$\det \begin{pmatrix} 1+\beta-\alpha\lambda_i-u & -\beta \\ 1 & -u \end{pmatrix} = 0 \iff u^2 - (1+\beta-\alpha\lambda_i)u + \beta = 0$$

$$\iff u = \frac{1}{2}\left(1+\beta-\alpha\lambda_i \pm \sqrt{(1+\beta-\alpha\lambda_i)^2 - 4\beta}\right)$$

Magnitude of roots are the same iff the roots are complex :
Let $\Delta = (1+\beta-\alpha\lambda_i)^2 - 4\beta$. Then $\sqrt{\Delta}$ is imaginary $\iff \Delta \leq 0$.

$$\begin{aligned}
\iff \quad & (1+\beta-\alpha\lambda_i)^2 && \leq 4\beta \\
\iff \quad & 1+\beta-\alpha\lambda_i && \leq 2\sqrt{\beta} \\
\iff \quad & 1-2\sqrt{\beta}+\beta && \leq \alpha\lambda_i \\
\iff \quad & (1-\sqrt{\beta})^2 && \leq \alpha\lambda_i \\
\iff \quad & 1-\sqrt{\beta} && \leq \sqrt{\alpha\lambda_i} \\
\iff \quad & 1-\sqrt{\alpha\lambda_i} && \leq \sqrt{\beta} \\
\iff \quad & \beta && \geq (1-\sqrt{\alpha\lambda_i})^2
\end{aligned}$$

Proving lemma 2 - Complex roots $\iff \beta \geq (1-\sqrt{\alpha\lambda_i})^2$

Note : $\beta \geq (1-\sqrt{\alpha\lambda_i})^2$ is automatically satisfied due to assumptions $\beta \geq (1-\sqrt{\alpha L})^2$ and $\mathbf{A} \preceq L\mathbf{I}$

Then we have

$$\begin{aligned} 1+\beta-\alpha\lambda_i &\geq 1+(1-\sqrt{\alpha\lambda_i})^2-\alpha\lambda_i \\ &= 1+(1-2\sqrt{\alpha\lambda_i}+\alpha\lambda_i)-\alpha\lambda_i \\ &= 2(1-\sqrt{\alpha\lambda_i}) \end{aligned}$$

Hence

$$\begin{aligned} u &= \frac{1}{2}\Big(1+\beta-\alpha\lambda_i \pm \sqrt{(1+\beta-\alpha\lambda_i)^2-4\beta}\Big) \\ &\geq \frac{1}{2}\Big(2(1-\sqrt{\alpha\lambda_i}) \pm \sqrt{(2(1-\sqrt{\alpha\lambda_i}))^2-4\beta}\Big) \\ &= \Big(1-\sqrt{\alpha\lambda_i} \pm \sqrt{(1-\sqrt{\alpha\lambda_i})^2-\beta}\Big) \end{aligned}$$

As $\beta \geq (1-\sqrt{\alpha\lambda_i})^2$, so $(1-\sqrt{\alpha\lambda_i})^2-\beta \leq 0$ and $\sqrt{(1-\sqrt{\alpha\lambda_i})^2-\beta}$ is imaginary

Thus the roots u will be complex number in the form $a \pm ib$, where $a = 1-\sqrt{\alpha\lambda_i}$ and $b = \left|\sqrt{(1-\sqrt{\alpha\lambda_i})^2-\beta}\right|$

Proving lemma 2 - magnitude of $u \leq \beta$

The magnitude of u in the form of $a + ib$ is $\sqrt{a^2 + b^2}$

$$\begin{aligned} |u| &= \sqrt{(1 - \sqrt{\alpha\lambda_i})^2 + (1 - \sqrt{\alpha\lambda_i})^2 - \beta} \\ &\leq \sqrt{\beta + \beta - \beta} \\ &= \sqrt{\beta} \end{aligned}$$

So the magnitude of eigenvalues of $\mathbf{T}_i$ $(\forall i)$ are less than $\sqrt{\beta}$

By assumption $\beta > (1 - \sqrt{\alpha L})^2$, we have $\sqrt{\beta} > 1 - \sqrt{\alpha L}$ and $\sqrt{\beta} \leq \beta$

Therefore, the largest eigenvalue (spectral radius) of $\mathbf{T} \leq \beta$. And we finish the proof of Lemma 2. $\square$

Convergence of HBM

Assume $\beta \geq (1 - \sqrt{\alpha L})$, by lemma 2 we have $\rho(\mathbf{T}) = \max |\lambda_i(\mathbf{T})| \leq \beta$.

By lemma 1, we have $\|\mathbf{T}^k\| \leq (\rho(\mathbf{T}) + \varepsilon_k)^k$ with $\lim_{k \to \infty} \varepsilon_k = 0$

Put lemma 2 into lemma 1 we have

$$\|\mathbf{T}^k\| \leq (\beta + \varepsilon_k)^k$$

Lastly, let $\alpha = \dfrac{4}{(\sqrt{L} + \sqrt{l})^2}$ and $\beta = \dfrac{\sqrt{L} - \sqrt{l}}{\sqrt{L} + \sqrt{l}}$ in $\mathbf{T}$ we have

$$\left\| \begin{bmatrix} \mathbf{x}_{k+1} - \mathbf{x}^* \\ \mathbf{x}_k - \mathbf{x}^* \end{bmatrix} \right\| \leq \left(\frac{\sqrt{L} - \sqrt{l}}{\sqrt{L} + \sqrt{l}} + \varepsilon \right)^k \left\| \begin{bmatrix} \mathbf{x}_1 - \mathbf{x}^* \\ \mathbf{x}_0 - \mathbf{x}^* \end{bmatrix} \right\|$$

or

$$\|\mathbf{x}_k - \mathbf{x}^*\| \leq \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} + \varepsilon \right)^k \|\mathbf{x}_0 - \mathbf{x}^*\|$$

where $\kappa = \dfrac{L}{l}$

Last page - summary

- Gradient descent $\mathbf{x}_{k+1} = \mathbf{x}_k - t_k \nabla_{\mathbf{x}} f(\mathbf{x}_k)$ has convergence

$$\|\mathbf{x}_k - \mathbf{x}^*\| \leq \left(\frac{\kappa - 1}{\kappa + 1} + \varepsilon\right)^k \|\mathbf{x}_0 - \mathbf{x}^*\|$$

- Heavy Ball Method $\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \nabla_{\mathbf{x}} f(\mathbf{x}_k) + \beta_k(\mathbf{x}_k - \mathbf{x}_{k-1})$ has convergence

$$\|\mathbf{x}_k - \mathbf{x}^*\| \leq \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} + \varepsilon\right)^k \|\mathbf{x}_0 - \mathbf{x}^*\|$$

- Improvement from $\left(\frac{\kappa - 1}{\kappa + 1} + \varepsilon\right)^k = \left(1 - \frac{2}{\kappa + 1} + \varepsilon\right)^k$ to $\left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} + \varepsilon\right)^k = \left(1 - \frac{2}{\sqrt{\kappa} + 1} + \varepsilon\right)^k$

End of document