Nesterov’s Accelerated Gradient Descent on $\beta$-smooth convex function

Proving NAGD converges at $O\left(\frac{1}{k^2}\right)$

Andersen Ang

Mathématique et recherche opérationnelle
UMONS, Belgium

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

First draft: August 2, 2017
Last update: July 28, 2019
1. The Nesterov’s Accelerated Gradient Descent (NAGD)

2. Proving NAGD converges at $O\left(\frac{1}{k^2}\right)$

3. Summary
Gradient Descent

Given a function $f : \mathbb{R}^n \to \mathbb{R}$ that is convex and $\beta$-smooth\(^1\), the unconstrained minimization problem

$$\min_{x \in \mathbb{R}^n} f(x),$$

can be solved by Gradient Descent (GD) algorithm. Starting with an initial point $x_0 \in \mathbb{R}^n$, GD iterates the following update:

$$x_{k+1} = x_k - t_k \nabla f(x_k).$$

If step size is small enough ($t_k \leq \frac{2}{\beta}$), then $\{x_k\}_{k \in \mathbb{N}}$ converges to a stationary point of $f$. As $f$ is convex, the sequence converges to the global minimizer $x^*$ (if it exists).

GD algorithm on such $f$ has convergence rate $\mathcal{O}\left(\frac{1}{k}\right)$

---

\(^1\) $f$ is $\beta$ smooth if $\nabla f$ is $\beta$-Lipschitz. If $f$ is not differentiable, $\nabla f$ is replaced by sub-gradient $g \in \partial f$
Nesterov’s Accelerated Gradient Descent

On the same problem \( \min_{x \in \mathbb{R}^n} f(x) \), \( f \) is convex and \( \beta \)-smooth), the **Nesterov’s Accelerated Gradient Descent** (NAGD) iterates the following update scheme:

- **Gradient update** \( y_{k+1} = x_k - \frac{1}{\beta} \nabla f(x_k) \) \( \tag{1} \)

- **Extrapolation** \( x_{k+1} = (1 - \gamma_k)y_{k+1} + \gamma_k y_k \) \( \tag{2} \)

- **Extrapolation weight** \( \gamma_k = \frac{1 - \lambda_k}{\lambda_{k+1}} \) \( \tag{3} \)

- **Extrapolation weight** \( \lambda_k = \frac{1 + \sqrt{1 + 4\lambda_{k-1}^2}}{2} \) \( \tag{4} \)

with initial point \( y_0 = x_0 \in \mathbb{R}^n \) and \( \lambda_0 = 0 \).

Note. In NAGD, fixed step size is used: \( t_k = \frac{1}{\beta} \).
Theorem. If $f : \mathbb{R}^n \to \mathbb{R}$ is $\beta$-smooth and convex, the sequences $f(y_k)$ produced by the NAGD algorithm converges to the optimal value $f^*$ at the rate $O\left(\frac{1}{k^2}\right)$ as

$$f(y_k) - f^* \leq \frac{2\beta\|x_0 - x^*\|_2^2}{k^2}$$

Note.

- It can be proved that the convergence rate $O\left(\frac{1}{k^2}\right)$ is optimal. i.e., no algorithm can perform better than NAGD in terms of convergence rate. A ”good” algorithm can only be at most as good as NAGD.

- If $f$ is not convex, the sequence $f(y_k)$ produced by the NAGD algorithm will converge to the closest stationary point with the same convergence rate.
Tool 1. $f$ is convex $\iff \forall x, y \in \text{dom } f$ we have

$$f(y) \geq f(x) + \nabla f(x)^T (y - x)$$

Tool 2. $f$ is $\beta$-smooth $\iff \forall x, y \in \text{dom } f$ we have

$$f(a) - f(b) \leq \nabla f(b)^T (a - b) + \frac{\beta}{2} \|a - b\|^2_2$$

From tool 1,

$$-f(y) \leq -f(x) + \nabla f(x)^T (x - y)$$

So

$$f\left(x - \frac{1}{\beta} \nabla f(x)\right) - f(y) \leq f\left(x - \frac{1}{\beta} \nabla f(x)\right) - f(x) + \nabla f(x)^T (x - y) \tag{5}$$
Consider $a = x - \frac{1}{\beta} \nabla f(x)$ and $b = x$, we have

$$a - b = -\frac{1}{\beta} \nabla f(x)$$  
$$\|a - b\|_2^2 = \frac{1}{\beta^2} \|\nabla f(x)\|_2^2$$

Put these and $a, b$ in tool 2

$$f(x - \frac{1}{\beta} \nabla f(x)) - f(x) \leq -\frac{1}{\beta} \|\nabla f(x)\|_2^2 + \frac{1}{2\beta} \|\nabla f(x)\|_2^2$$

$$= -\frac{1}{2\beta} \|\nabla f(x)\|_2^2$$ (6)

Recall inequality (5)

$$f(x - \frac{1}{\beta} \nabla f(x)) - f(y) \leq f(x - \frac{1}{\beta} \nabla f(x)) - f(x) + \nabla f(x)^T (x - y)$$

(5) + (6) :

$$f(x - \frac{1}{\beta} \nabla f(x)) - f(y) \leq -\frac{1}{2\beta} \|\nabla f(x)\|_2^2 + \nabla f(x)^T (x - y)$$ (7)
The proof ...

We have (7)

\[ f\left(x - \frac{1}{\beta} \nabla f(x)\right) - f(y) \leq \frac{-1}{2\beta} \|\nabla f(x)\|_2^2 + \nabla f(x)^T (x - y) \]

Put \( x = x_k, \ y = x^* \) in (7) get (8), put \( x = x_k, \ y = y_k \) in (7) get (9)

\[ f\left(x_k - \frac{1}{\beta} \nabla f(x_k)\right) - f(x^*) \leq \frac{-1}{2\beta} \|\nabla f(x_k)\|_2^2 + \nabla f(x_k)^T (x_k - x^*) \quad (8) \]

\[ f\left(x_k - \frac{1}{\beta} \nabla f(x_k)\right) - f(y_k) \leq \frac{-1}{2\beta} \|\nabla f(x_k)\|_2^2 + \nabla f(x_k)^T (x_k - y_k) \quad (9) \]
The proof ... 4/12

Define \( \delta_k := f(y_k) - f^* \)

From (1) we have \( y_{k+1} = x_k - \beta^{-1} \nabla f(x_k) \) and

\[
f\left(x_k - \frac{1}{\beta} \nabla f(x_k)\right) = f(y_{k+1}) \tag{10}
\]

\[
f\left(x_k - \frac{1}{\beta} \nabla f(x_k)\right) - f^* = \delta_{k+1} \tag{11}
\]

\[
f\left(x_k - \frac{1}{\beta} \nabla f(x_k)\right) - f(y_k) = f\left(x_k - \frac{1}{\beta} \nabla f(x_k)\right) - f^* - \left(f(y_k) - f^*\right)
\]

\[
= \delta_{k+1} - \delta_k \tag{12}
\]

\[
\nabla f(x_k) = -\beta (y_{k+1} - x_k) \tag{13}
\]

\[
\|\nabla f(x_k)\|_2^2 = \beta^2 \|y_{k+1} - x_k\|_2^2 \tag{14}
\]

Put (11,13,14) into (8), put (12,13,14) into (9)

\[
\delta_{k+1} \leq -\frac{\beta}{2} \|y_{k+1} - x_k\|_2^2 - \beta (y_{k+1} - x_k)^T (x_k - x^*) \tag{15}
\]

\[
\delta_{k+1} - \delta_k \leq -\frac{\beta}{2} \|y_{k+1} - x_k\|_2^2 - \beta (y_{k+1} - x_k)^T (x_k - y_k) \tag{16}
\]
(Things go crazy here) Consider \((15) + (\lambda_k - 1)(16)\)

On the left hand side we have

\[
\delta_{k+1} + (\lambda_k - 1)(\delta_{k+1} - \delta_k) = \lambda_k \delta_{k+1} - (\lambda_k - 1)\delta_k
\]

On the right hand side we have

\[
- \frac{\beta}{2} \|y_{k+1} - x_k\|_2^2 - \beta (y_{k+1} - x_k)^T (x_k - x^*)
\]

\[
- (\lambda_k - 1) \left( \frac{\beta}{2} \|y_{k+1} - x_k\|_2^2 - \beta (y_{k+1} - x_k)^T (x_k - y_k) \right)
\]

\[
= - \frac{\lambda_k \beta}{2} \|y_{k+1} - x_k\|_2^2 - \beta (y_{k+1} - x_k)^T \left( x_k - x^* + (\lambda_k - 1)(x_k - y_k) \right)
\]

\[
= - \frac{\lambda_k \beta}{2} \|y_{k+1} - x_k\|_2^2 - \beta (y_{k+1} - x_k)^T \left( \lambda_k x_k - (\lambda_k - 1)y_k - x^* \right)
\]

Hence we have

\[
\lambda_k \delta_{k+1} - (\lambda_k - 1)\delta_k \leq - \frac{\lambda_k \beta}{2} \|y_{k+1} - x_k\|_2^2 - \beta (y_{k+1} - x_k)^T \left( \lambda_k x_k - (\lambda_k - 1)y_k - x^* \right)
\]
Multiply (17) inequality with $\lambda_k$:

$$
\lambda_k^2 \delta_{k+1} - \lambda_k (\lambda_k - 1) \delta_k \leq - \frac{\lambda_k^2 \beta}{2} \|y_{k+1} - x_k\|^2_2
$$

$$
- \lambda_k \beta (y_{k+1} - x_k)^T \left( \lambda_k x_k - (\lambda_k - 1) y_k - x^* \right)
$$

From (4) $\lambda_k = \frac{1}{2} \left( 1 + \sqrt{1 + 4 \lambda_{k-1}^2} \right)$, we get

$$(2 \lambda_k - 1)^2 = 1 + 4 \lambda_{k-1}^2 \iff \lambda_{k-1}^2 = \lambda_k (\lambda_k - 1)$$

Put $\lambda_{k-1}^2 = \lambda_k (\lambda_k - 1)$ in the equation above

$$
\lambda_k^2 \delta_{k+1} - \lambda_{k-1}^2 \delta_k \leq - \frac{\lambda_k^2 \beta}{2} \|y_{k+1} - x_k\|^2_2
$$

$$
- \lambda_k \beta (y_{k+1} - x_k)^T \left( \lambda_k x_k - (\lambda_k - 1) y_k - x^* \right)
$$

$$
- \frac{\beta}{2} \left( \lambda_k^2 \|y_{k+1} - x_k\|^2_2
$$

$$
+ 2 \lambda_k (y_{k+1} - x_k)^T (\lambda_k x_k - (\lambda_k - 1) y_k - x^*) \right)
$$

[re-arrange RHS] $$
\Rightarrow (18)$$
Super tricky: the expression

\[ \lambda_k^2 \| y_{k+1} - x_k \|_2^2 + 2\lambda_k (y_{k+1} - x_k)^T (\lambda_k x_k - (\lambda_k - 1)y_k - x^*) \]

is equivalent to

\[ \| \lambda_k y_{k+1} - (\lambda_k - 1)y_k - x^* \|_2^2 + \| \lambda_k x_k - (\lambda_k - 1)y_k - x^* \|_2^2 \]

Why:

\[ \lambda^2 (y - x)^2 + 2\lambda(y - x)(\lambda x - (\lambda - 1)z - w) = (\lambda y - (\lambda - 1)z - w)^2 + (\lambda x - (\lambda - 1)z - w)^2 \]

Apply to (18)

\[ \lambda_k^2 \delta_{k+1} - \lambda_{k-1}^2 \delta_k \leq -\frac{\beta}{2} \left( \| \lambda_k y_{k+1} - (\lambda_k - 1)y_k - x^* \|_2^2 + \| \lambda_k x_k - (\lambda_k - 1)y_k - x^* \|_2^2 \right) \]
We have the following equality:

\[ \lambda_k x_k - (\lambda_k - 1)y_k = (1 - \lambda_{k-1})y_{k-1} + \lambda_{k-1}y_k \]

Proof: By (3) \( \gamma_k = \frac{1-\lambda_k}{\lambda_{k+1}} \), \( \gamma_k \lambda_{k+1} = 1 - \lambda_k \).

By (2) \( x_{k+1} = (1 - \gamma_k)y_{k+1} + \gamma_k y_k \) get \( x_{k+1} = y_{k+1} + \gamma_k (y_k - y_{k+1}) \),
multiply with \( \lambda_{k+1} \) gives \( \lambda_{k+1} x_{k+1} = \lambda_{k+1} y_{k+1} + \lambda_{k+1} \gamma_k (y_k - y_{k+1}) = \lambda_{k+1} y_{k+1} + (1 - \lambda_k)(y_k - y_{k+1}) \), rearrange this get
\[ \lambda_{k+1} x_{k+1} - \lambda_{k+1} y_{k+1} = (1 - \lambda_k)(y_k - y_{k+1}) \], add \( y_{k+1} \) on both side get
\[ \lambda_{k+1} x_{k+1} - (\lambda_{k+1} - 1)y_{k+1} = (1 - \lambda_k)y_k + \lambda_k y_{k+1} \].

hence (19) becomes

\[ \lambda_k^2 \delta_{k+1} - \lambda_{k-1}^2 \delta_k \leq -\frac{\beta}{2} \left( \| \lambda_k y_{k+1} - (\lambda_k - 1)y_k - x^* \|_2^2 \right) \]

\[ + \| (1 - \lambda_{k-1})y_{k-1} + \lambda_{k-1}y_k - x^* \|_2^2 \]
Re-arrange (20) that the two terms in right hand side have similar form

$$\lambda_k^2 \delta_{k+1} - \lambda_{k-1}^2 \delta_k \leq -\frac{\beta}{2} \left( \| \lambda_k y_{k+1} - (\lambda_k - 1)y_k - x^* \|_2^2 \right)$$

Let $u_k = \lambda_k y_{k+1} - (\lambda_k - 1)y_k - x^*$ so

$$\lambda_{k-1} y_k - (\lambda_{k-1} - 1)y_{k-1} - x^* = u_{k-1}$$

and (21) becomes

$$\lambda_k^2 \delta_{k+1} - \lambda_{k-1}^2 \delta_k \leq -\frac{\beta}{2} \left( \| u_k \|_2^2 - \| u_{k-1} \|_2^2 \right)$$

which forms a telescoping series!
The proof ...

\[ k = 1 \quad \lambda_1^2 \delta_2 - \lambda_0^2 \delta_1 \leq -\frac{\beta}{2} \left( \|u_1\|_2^2 - \|u_0\|_2^2 \right) \]

\[ k = 2 \quad \lambda_2^2 \delta_3 - \lambda_1^2 \delta_2 \leq -\frac{\beta}{2} \left( \|u_2\|_2^2 - \|u_1\|_2^2 \right) \]

\[ \vdots \]

\[ k = K - 1 \quad \lambda_{K-1}^2 \delta_K - \lambda_{K-2}^2 \delta_{K-1} \leq -\frac{\beta}{2} \left( \|u_{K-1}\|_2^2 - \|u_{K-2}\|_2^2 \right) \]

**Sum all**

\[ \lambda_{K-1}^2 \delta_K - \lambda_0^2 \delta_1 \leq -\frac{\beta}{2} \left( \|u_{K-1}\|_2^2 - \|u_0\|_2^2 \right) \]

**re-arrange**

\[ \lambda_{K-1}^2 \delta_K - \lambda_0^2 \delta_1 \leq \frac{\beta}{2} \left( \|u_0\|_2^2 - \|u_{K-1}\|_2^2 \right) \]

By definition, \( \lambda_0 = 0 \), \( u_0 = \lambda_0 y_1 - (\lambda_0 - 1) y_0 - x^* = y_0 - x^* \), and \( y_0 = x_0 \):

\[ \lambda_{K-1}^2 \delta_K \leq \frac{\beta}{2} \left( \|x_0 - x^*\|_2^2 - \|u_{K-1}\|_2^2 \right) \]

As \( \|u_{K-1}\|_2^2 \geq 0 \)

\[ \delta_K \leq \frac{\beta \|x_0 - x^*\|_2^2}{2 \lambda_{K-1}^2} \]
Lemma. $\lambda_{k-1} \geq \frac{k}{2}$.

Proof by induction. For $k = 0$ it is trivial ($0 \geq 0/2$).

When $k = 1$, by definition $\lambda_k = \frac{1 + \sqrt{1 + 4\lambda^2_{k-1}}}{2}$, so $k_1 = 1 > \frac{1}{2} = \lceil \frac{k}{2} \rceil_{k=1}$ (Induction hypothesis). Assume $\lambda_{n-1} \geq \frac{n}{2}$.

When $k = n$,

\[
\lambda_n = \frac{1 + \sqrt{1 + 4\lambda^2_{n-1}}}{2} \geq \frac{1 + \sqrt{1 + 4\left(\frac{n}{2}\right)^2}}{2} = \frac{1 + \sqrt{1 + n^2}}{2} > \frac{1 + \sqrt{n^2}}{2} = \frac{1 + n}{2} \]

\[\square\]
With $\lambda_{k-1} \geq \frac{k}{2}$, so $\frac{1}{\lambda_{k-1}^2} \leq \frac{4}{k^2}$ and

$$\delta_K \leq \frac{\beta \|x_0 - x^*\|_2^2}{2\lambda_{K-1}^2}$$

becomes

$$f_{y_k} - f^* \leq \frac{2\beta \|x_0 - x^*\|_2^2}{k^2}$$

where $f_{y_k} - f^* = \delta_k$ (by definition). $\square$
NAGD vs GD

GD has convergence rate $\mathcal{O}\left(\frac{1}{k}\right)$. NAGD has convergence rate $\mathcal{O}\left(\frac{1}{k^2}\right)$. The improvement from $\mathcal{O}\left(\frac{1}{k}\right)$ to $\mathcal{O}\left(\frac{1}{k^2}\right)$ is HUGE

![Graphs showing convergence rates]

e.g. Consider at iteration $k = 10$:

For GD, $f(x_{10}) - f^*$ is bounded by $\frac{2\beta\|x_0 - x^*\|_2^2}{10}$.

For NAGD, $f(x_{10}) - f^*$ is bounded by $\frac{2\beta\|x_0 - x^*\|_2^2}{10^2}$.
For unconstrained problem $\min_{x \in \mathbb{R}^n} f(x)$, with $f : \mathbb{R}^n \to \mathbb{R}$ being $\beta$-smooth and convex, the NAGD algorithm starts with initial point $x_0 = y_0 \in \mathbb{R}^n$ and $\lambda_0 = 0$ and iterates the following:

$$y_{k+1} = x_k - \frac{1}{\beta} \nabla f(x_k) \quad x_{k+1} = (1 - \gamma_k)y_{k+1} + \gamma_k y_k$$

$$\gamma_k = \frac{1 - \lambda_k}{\lambda_{k+1}} \quad \lambda_k = \frac{1}{2} \left( 1 + \sqrt{1 + 4\lambda_{k-1}^2} \right)$$

will produce a sequences $f(y_k)$ that converges to the optimal value $f^*$ at order of $O\left(\frac{1}{k^2}\right)$ as

$$f(y_k) - f^* \leq \frac{2\beta\|x_0 - x^*\|_2^2}{k^2}$$

End of document