Nesterov's Accelerated Gradient Descent on L-smooth convex function

Proof approach 1

Andersen Ang

ECS, Uni. Southampton, UK

andersen.ang@soton.ac.uk

Homepage angms.science

Version: November 4, 2023

First draft: August 2, 2017

Content

Problem setup: smooth unconstrained convex optimisation

$$(\mathcal{P}) \ : \ \underset{\boldsymbol{x}}{\operatorname{argmin}} \ f(\boldsymbol{x}).$$

- We consider Euclidean space
- $f : \mathbb{R}^n \to \mathbb{R}$
- f is $L-$smooth
  - f is continuously differentiable
    $f \in \mathcal{C}^1$, i.e., $\nabla f(\boldsymbol{x})$ exists for all $\boldsymbol{x} \in \operatorname{dom} f$
  - ∇f is L-Lipschitz
    $L > 0$ is the least upper bound in $\frac{\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|}{\|\boldsymbol{x}-\boldsymbol{y}\|} \leq L$

$$\forall \boldsymbol{a}, \boldsymbol{b} \in \operatorname{dom} f \ : \ f(\boldsymbol{a}) - f(\boldsymbol{b}) \leq \langle \nabla f(\boldsymbol{b}), \boldsymbol{a} - \boldsymbol{b} \rangle + \frac{L}{2} \|\boldsymbol{a} - \boldsymbol{b}\|_2^2 .$$

- f is convex all local minima of $\mathcal{P}$ are global minima

$$(\forall \boldsymbol{x} \in \operatorname{dom} f)(\forall \boldsymbol{y} \in \operatorname{dom} f)\Big\{ f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle \Big\}$$

- Details of convexity, L-smoothness, see here

Gradient Descent (GD)

- Notation

$$f_k := f(\boldsymbol{x}_k)$$
$$f^* := f(\boldsymbol{x}^*)$$

- GD: start with initial point $\boldsymbol{x}_0 \in \mathbb{R}^n$, iterates

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha_k \nabla f(\boldsymbol{x}_k).$$

  For sufficiently small stepsize ($\alpha_k < \frac{2}{L}$), the sequence $\{\boldsymbol{x}_k\}_{k \in \mathbb{N}}$ converges to a stationary point of f.

  As f is convex, the sequence converges to the global minimizer $\boldsymbol{x}^*$ (if exists).

- GD convergence as $f_k - f^* \leq \mathcal{O}\left(\frac{1}{k}\right)$

Nesterov's Accelerated Gradient Descent (NAGD)

$$\boxed{(\mathcal{P}) \ : \ \min_{\boldsymbol{x}} f(\boldsymbol{x})}$$

- Start with initial point $\boldsymbol{y}_0 = \boldsymbol{x}_0 \in \mathbb{R}^n$ and $\lambda_0 = 0$, iterates

$$\begin{aligned}
\text{Gradient update} \quad \boldsymbol{y}_{k+1} &= \boldsymbol{x}_k - \frac{1}{L}\nabla f(\boldsymbol{x}_k) && (1)\\
\text{Extrapolation} \quad \boldsymbol{x}_{k+1} &= (1-\gamma_k)\boldsymbol{y}_{k+1} + \gamma_k \boldsymbol{y}_k && (2)\\
\text{Extrapolation weight} \quad \gamma_k &= \frac{1-\lambda_k}{\lambda_{k+1}} && (3)\\
\text{Extrapolation weight} \quad \lambda_k &= \frac{1+\sqrt{1+4\lambda_{k-1}^2}}{2} && (4)
\end{aligned}$$

  Note that here fix stepsize is used: $\alpha_k = \dfrac{1}{L}\forall k$.

- **Theorem**. If $f : \mathbb{R}^n \to \mathbb{R}$ is L-smooth and convex, the sequences $\big\{f(\boldsymbol{y}_k)\big\}_k$ produced by NAGD converges to the optimal value f^* at the rate $\mathcal{O}\left(\frac{1}{k^2}\right)$ as

$$f(\boldsymbol{y}_k) - f^* \ \leq \ \frac{2L\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2}{k^2}.$$

  - The convergence rate $\mathcal{O}\Big(\dfrac{1}{k^2}\Big)$ is optimal. I.e., no 1st-order algo. can perform better than NAGD in terms of convergence rate. All 1st-order algorithm can only be at most as good as NAGD. Proof here.
  - If f is nonconvex, the sequence $\big\{f(\boldsymbol{y}_k)\big\}_k$ produced by NAGD converges to the closest stationary point with the same convergence rate.

NAGD converges rate $\mathcal{O}\left(\frac{1}{k^2}\right)$ proof 1/6 **Stage 1: make use of convexity & smoothness**

- f cvx: $(\forall \boldsymbol{x} \forall \boldsymbol{y})\Big\{ f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \big\langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \big\rangle \Big\}$ gives

$$-f(\boldsymbol{y}) \leq -f(\boldsymbol{x}) + \big\langle \nabla f(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{y} \big\rangle \tag{5}$$

- f L-smooth $(\forall \boldsymbol{a} \forall \boldsymbol{b})\Big\{ f(\boldsymbol{a}) - f(\boldsymbol{b}) \leq \big\langle \nabla f(\boldsymbol{b}), \boldsymbol{a} - \boldsymbol{b} \big\rangle + \frac{L}{2}\|\boldsymbol{a} - \boldsymbol{b}\|_2^2 \Big\}$, with $\boldsymbol{a} = \boldsymbol{x} - \frac{1}{L}\nabla f(\boldsymbol{x}), \boldsymbol{b} = \boldsymbol{x}$,

$$f\Big(\boldsymbol{x} - \frac{1}{L}\nabla f(\boldsymbol{x})\Big) - f(\boldsymbol{x}) \;\leq\; -\frac{1}{L}\|\nabla f(\boldsymbol{x})\|_2^2 + \frac{1}{2L}\|\nabla f(\boldsymbol{x})\|_2^2 = \frac{-1}{2L}\|\nabla f(\boldsymbol{x})\|_2^2. \tag{6}$$

- (5) + (6) will cancel $-f(\boldsymbol{x})$ and give

$$f\Big(\boldsymbol{x} - \frac{1}{L}\nabla f(\boldsymbol{x})\Big) - f(\boldsymbol{y}) \;\leq\; \frac{-1}{2L}\|\nabla f(\boldsymbol{x})\|_2^2 + \big\langle \nabla f(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{y} \big\rangle. \tag{7}$$

- Put $\boldsymbol{x} = \boldsymbol{x}_k$, $\boldsymbol{y} = \boldsymbol{x}^*$ in (7)

$$f\Big(\boldsymbol{x}_k - \frac{1}{L}\nabla f(\boldsymbol{x}_k)\Big) - f^* \;\leq\; \frac{-1}{2L}\|\nabla f(\boldsymbol{x}_k)\|_2^2 + \big\langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x}_k - \boldsymbol{x}^* \big\rangle. \tag{8}$$

- put $\boldsymbol{x} = \boldsymbol{x}_k$, $\boldsymbol{y} = \boldsymbol{y}_k$ in (7)

$$f\Big(\boldsymbol{x}_k - \frac{1}{L}\nabla f(\boldsymbol{x}_k)\Big) - f(\boldsymbol{y}_k) \;\leq\; \frac{-1}{2L}\|\nabla f(\boldsymbol{x}_k)\|_2^2 + \big\langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x}_k - \boldsymbol{y}_k \big\rangle. \tag{9}$$

- Proof overview: (8), (9) link $f(\boldsymbol{y}_{k+1})$, $f(\boldsymbol{y}_k)$ and f^*. We see $\nabla f(\boldsymbol{x}_k)$ appear in (8), (9) but not in the convergence result, so we eliminate $\nabla f(\boldsymbol{x}_k)$ in (8), (9).

Proof 2/6 **Stage 2: eliminate gradient**

$$\boldsymbol{y}_{k+1} = \boldsymbol{x}_k - \frac{1}{L}\nabla f(\boldsymbol{x}_k) \tag{1}$$

$$f\Big(\boldsymbol{x}_k - \frac{1}{L}\nabla f(\boldsymbol{x}_k)\Big) - f^* \le \frac{-1}{2L}\|\nabla f(\boldsymbol{x}_k)\|_2^2 + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x}_k - \boldsymbol{x}^*\rangle. \tag{8}$$

$$f\Big(\boldsymbol{x}_k - \frac{1}{L}\nabla f(\boldsymbol{x}_k)\Big) - f(\boldsymbol{y}_k) \le \frac{-1}{2L}\|\nabla f(\boldsymbol{x}_k)\|_2^2 + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x}_k - \boldsymbol{y}_k\rangle. \tag{9}$$

- Simplify notation, let $\delta_k := f(\boldsymbol{y}_k) - f^*$, then

$$f\Big(\boldsymbol{x}_k - \frac{1}{L}\nabla f(\boldsymbol{x}_k)\Big) \overset{(1)}{=} f(\boldsymbol{y}_{k+1}) \tag{10}$$

$$f\Big(\boldsymbol{x}_k - \frac{1}{L}\nabla f(\boldsymbol{x}_k)\Big) - f^* \overset{(10),\delta_k}{=} \delta_{k+1} \tag{11}$$

$$\begin{aligned} f\Big(\boldsymbol{x}_k - \frac{1}{L}\nabla f(\boldsymbol{x}_k)\Big) - f(\boldsymbol{y}_k) &= f\Big(\boldsymbol{x}_k - \frac{1}{L}\nabla f(\boldsymbol{x}_k)\Big) - f^* - \Big(f(\boldsymbol{y}_k) - f^*\Big) \\ &= \delta_{k+1} - \delta_k \end{aligned} \tag{12}$$

$$\nabla f(\boldsymbol{x}_k) \overset{(1)}{=} -L(\boldsymbol{y}_{k+1} - \boldsymbol{x}_k) \tag{13}$$

$$\|\nabla f(\boldsymbol{x}_k)\|_2^2 \overset{(13)}{=} L^2\|\boldsymbol{y}_{k+1} - \boldsymbol{x}_k\|_2^2 \tag{14}$$

- Put (11,13,14) into (8)

$$\delta_{k+1} \le -\frac{L}{2}\|\boldsymbol{y}_{k+1} - \boldsymbol{x}_k\|_2^2 - L\big\langle \boldsymbol{y}_{k+1} - \boldsymbol{x}_k,\, \boldsymbol{x}_k - \boldsymbol{x}^*\big\rangle. \tag{15}$$

- Put (12,13,14) into (9)

$$\delta_{k+1} - \delta_k \le -\frac{L}{2}\|\boldsymbol{y}_{k+1} - \boldsymbol{x}_k\|_2^2 - L\big\langle \boldsymbol{y}_{k+1} - \boldsymbol{x}_k,\, \boldsymbol{x}_k - \boldsymbol{y}_k\big\rangle. \tag{16}$$

Proof 3/6 **Stage 3: form telescoping sum**

$$\begin{array}{rcll} \lambda_k &=& \frac{1}{2}\left(1+\sqrt{1+4\lambda_{k-1}^2}\right) & (4) \\ \delta_{k+1} &\leq& -\frac{L}{2}\|\boldsymbol{y}_{k+1}-\boldsymbol{x}_k\|_2^2 - L\langle \boldsymbol{y}_{k+1}-\boldsymbol{x}_k,\, \boldsymbol{x}_k-\boldsymbol{x}^* \rangle & (15) \\ \delta_{k+1}-\delta_k &\leq& -\frac{L}{2}\|\boldsymbol{y}_{k+1}-\boldsymbol{x}_k\|_2^2 - L\langle \boldsymbol{y}_{k+1}-\boldsymbol{x}_k,\, \boldsymbol{x}_k-\boldsymbol{y}_k \rangle & (16) \end{array}$$

- **Tricky step**: consider (15) + $(\lambda_k - 1)$(16).

  Left-hand size of (15) + (λ_k-1)(16) $= \delta_{k+1} + (\lambda_k-1)(\delta_{k+1}-\delta_k) = \lambda_k\delta_{k+1} - (\lambda_k-1)\delta_k$.

- Right-hand side of (15) + (λ_k-1)(16)

$$\begin{aligned} & -\frac{L}{2}\|\boldsymbol{y}_{k+1}-\boldsymbol{x}_k\|_2^2 - L\Big\langle \boldsymbol{y}_{k+1}-\boldsymbol{x}_k,\, \boldsymbol{x}_k-\boldsymbol{x}^*\Big\rangle + (\lambda_k-1)\left(\frac{-L}{2}\|\boldsymbol{y}_{k+1}-\boldsymbol{x}_k\|_2^2 - L\Big\langle \boldsymbol{y}_{k+1}-\boldsymbol{x}_k,\, \boldsymbol{x}_k-\boldsymbol{y}_k\Big\rangle\right) \\ = \; & -\frac{\lambda_k L}{2}\|\boldsymbol{y}_{k+1}-\boldsymbol{x}_k\|_2^2 - L\Big\langle \boldsymbol{y}_{k+1}-\boldsymbol{x}_k,\, \boldsymbol{x}_k-\boldsymbol{x}^* + (\lambda_k-1)(\boldsymbol{x}_k-\boldsymbol{y}_k)\Big\rangle \\ = \; & -\frac{\lambda_k L}{2}\|\boldsymbol{y}_{k+1}-\boldsymbol{x}_k\|_2^2 - L\Big\langle \boldsymbol{y}_{k+1}-\boldsymbol{x}_k,\, \lambda_k\boldsymbol{x}_k - (\lambda_k-1)\boldsymbol{y}_k - \boldsymbol{x}^*\Big\rangle \end{aligned}$$

- By LHS = RHS $\lambda_k\delta_{k+1} - (\lambda_k-1)\delta_k \leq -\frac{\lambda_k L}{2}\|\boldsymbol{y}_{k+1}-\boldsymbol{x}_k\|_2^2 - L\Big\langle \boldsymbol{y}_{k+1}-\boldsymbol{x}_k,\, \lambda_k\boldsymbol{x}_k-(\lambda_k-1)\boldsymbol{y}_k-\boldsymbol{x}^*\Big\rangle$.

  Multiply the inequality with λ_k:

$$\begin{aligned} \lambda_k^2\delta_{k+1} - \lambda_k(\lambda_k-1)\delta_k \;&\leq\; -\frac{\lambda_k^2 L}{2}\|\boldsymbol{y}_{k+1}-\boldsymbol{x}_k\|_2^2 - \lambda_k L\Big\langle \boldsymbol{y}_{k+1}-\boldsymbol{x}_k,\, \lambda_k\boldsymbol{x}_k-(\lambda_k-1)\boldsymbol{y}_k-\boldsymbol{x}^*\Big\rangle \\ &=\; -\frac{L}{2}\left(\lambda_k^2\|\boldsymbol{y}_{k+1}-\boldsymbol{x}_k\|_2^2 + 2\lambda_k\Big\langle \boldsymbol{y}_{k+1}-\boldsymbol{x}_k,\, \lambda_k\boldsymbol{x}_k-(\lambda_k-1)\boldsymbol{y}_k-\boldsymbol{x}^*\Big\rangle\right). \qquad (\#) \end{aligned}$$

- (4) gives $(2\lambda_k-1)^2 = 1+4\lambda_{k-1}^2 \iff 4\lambda_k^2 - 4\lambda_k + 1 = 1 + 4\lambda_{k-1}^2 \iff \lambda_{k-1}^2 = \lambda_k(\lambda_k-1)$, put this into (#) gives

$$\lambda_k^2\delta_{k+1} - \lambda_{k-1}^2\delta_k \leq -\frac{L}{2}\left(\lambda_k^2\|\boldsymbol{y}_{k+1}-\boldsymbol{x}_k\|_2^2 + 2\lambda_k\Big\langle \boldsymbol{y}_{k+1}-\boldsymbol{x}_k,\, \lambda_k\boldsymbol{x}_k-(\lambda_k-1)\boldsymbol{y}_k-\boldsymbol{x}^*\Big\rangle\right) \qquad (17)$$

Proof 4/6

$$\lambda_k = \tfrac{1}{2}\Big(1+\sqrt{1+4\lambda_{k-1}^2}\Big) \tag{4}$$

$$\lambda_k^2\delta_{k+1} - \lambda_k(\lambda_k-1)\delta_k \leq -\frac{L}{2}\Big(\lambda_k^2\|\boldsymbol{y}_{k+1}-\boldsymbol{x}_k\|_2^2 + 2\lambda_k\big\langle \boldsymbol{y}_{k+1}-\boldsymbol{x}_k,\ \lambda_k\boldsymbol{x}_k-(\lambda_k-1)\boldsymbol{y}_k-\boldsymbol{x}^*\big\rangle\Big) \tag{17}$$

- Inspecting the inner product in (17) we see that it is completing squares (Thanks to Tony Silveti-Falls for figuring it out, 2023 Nov 3).

$$\|\lambda\boldsymbol{a}+\boldsymbol{b}\|_2^2 = \lambda^2\|\boldsymbol{a}\|_2^2 + 2\lambda\langle\boldsymbol{a},\boldsymbol{b}\rangle + \|\boldsymbol{b}\|_2^2 \iff \lambda^2\|\boldsymbol{a}\|_2^2 + 2\lambda\langle\boldsymbol{a},\boldsymbol{b}\rangle = \|\lambda\boldsymbol{a}+\boldsymbol{b}\|_2^2 - \|\boldsymbol{b}\|_2^2.$$

$$\begin{aligned}
&\lambda_k^2\|\boldsymbol{y}_{k+1}-\boldsymbol{x}_k\|_2^2 + 2\lambda_k\big\langle \boldsymbol{y}_{k+1}-\boldsymbol{x}_k,\ \lambda_k\boldsymbol{x}_k-(\lambda_k-1)\boldsymbol{y}_k-\boldsymbol{x}^*\big\rangle\\
=\ &\big\|\lambda(\boldsymbol{y}_{k+1}-\boldsymbol{x}_k)+\lambda_k\boldsymbol{x}_k-(\lambda_k-1)\boldsymbol{y}_k-\boldsymbol{x}^*\big\|_2^2 - \big\|\lambda_k\boldsymbol{x}_k-(\lambda_k-1)\boldsymbol{y}_k-\boldsymbol{x}^*\big\|_2^2\\
=\ &\big\|\lambda_k\boldsymbol{y}_{k+1}-(\lambda_k-1)\boldsymbol{y}_k-\boldsymbol{x}^*\big\|_2^2 - \big\|\lambda_k\boldsymbol{x}_k-(\lambda_k-1)\boldsymbol{y}_k-\boldsymbol{x}^*\big\|_2^2.
\end{aligned}$$

- Using this (17) becomes

$$\lambda_k^2\delta_{k+1} - \lambda_{k-1}^2\delta_k \leq -\frac{L}{2}\Big(\big\|\lambda_k\boldsymbol{y}_{k+1}-(\lambda_k-1)\boldsymbol{y}_k-\boldsymbol{x}^*\big\|_2^2 - \big\|\lambda_k\boldsymbol{x}_k-(\lambda_k-1)\boldsymbol{y}_k-\boldsymbol{x}^*\big\|_2^2\Big). \tag{18}$$

- We have $\lambda_k\boldsymbol{x}_k-(\lambda_k-1)\boldsymbol{y}_k = (1-\lambda_{k-1})\boldsymbol{y}_{k-1}+\lambda_{k-1}\boldsymbol{y}_k$.

Proof: $\gamma_k \overset{(3)}{=} \frac{1-\lambda_k}{\lambda_{k+1}} \iff \gamma_k\lambda_{k+1} = 1-\lambda_k$.
By (2) $\boldsymbol{x}_{k+1} = (1-\gamma_k)\boldsymbol{y}_{k+1}+\gamma_k\boldsymbol{y}_k$ gives $x_{k+1} = \boldsymbol{y}_{k+1}+\gamma_k(\boldsymbol{y}_k-\boldsymbol{y}_{k+1})$, multiply with λ_{k+1} gives
$\lambda_{k+1}x_{k+1} = \lambda_{k+1}\boldsymbol{y}_{k+1}+\lambda_{k+1}\gamma_k(\boldsymbol{y}_k-\boldsymbol{y}_{k+1}) = \lambda_{k+1}\boldsymbol{y}_{k+1}+(1-\lambda_k)(\boldsymbol{y}_k-\boldsymbol{y}_{k+1})$, rearrange gives
$\lambda_{k+1}x_{k+1}-\lambda_{k+1}\boldsymbol{y}_{k+1} = (1-\lambda_k)(\boldsymbol{y}_k-\boldsymbol{y}_{k+1})$, add $\boldsymbol{y}_{k+1}$ on both side gives
$\lambda_{k+1}x_{k+1}-(\lambda_{k+1}-1)\boldsymbol{y}_{k+1} = (1-\lambda_k)\boldsymbol{y}_k+\lambda_k\boldsymbol{y}_{k+1}$. Move counter k by -1 gives the result.

So (18) becomes

$$\lambda_k^2\delta_{k+1} - \lambda_{k-1}^2\delta_k \leq -\frac{L}{2}\Big(\big\|\lambda_k\boldsymbol{y}_{k+1}-(\lambda_k-1)\boldsymbol{y}_k-\boldsymbol{x}^*\big\|_2^2 - \big\|(1-\lambda_{k-1})\boldsymbol{y}_{k-1}+\lambda_{k-1}\boldsymbol{y}_k-\boldsymbol{x}^*\big\|_2^2\Big).$$

Proof ... 5/6

We have $\lambda_k^2\delta_{k+1} - \lambda_{k-1}^2\delta_k \quad \leq \quad -\dfrac{L}{2}\left(\left\| \lambda_k \boldsymbol{y}_{k+1} - (\lambda_k - 1)\boldsymbol{y}_k - \boldsymbol{x}^* \right\|_2^2 - \left\| (1-\lambda_{k-1})\boldsymbol{y}_{k-1} + \lambda_{k-1}\boldsymbol{y}_k - \boldsymbol{x}^* \right\|_2^2 \right).$

Rearrange the second term to make the terms in right-hand side have similar form

$$\lambda_k^2\delta_{k+1} - \lambda_{k-1}^2\delta_k \quad \leq \quad -\frac{L}{2}\left(\left\| \lambda_k \boldsymbol{y}_{k+1} - (\lambda_k - 1)\boldsymbol{y}_k - \boldsymbol{x}^* \right\|_2^2 - \left\| \lambda_{k-1}\boldsymbol{y}_k - (\lambda_{k-1} - 1)\boldsymbol{y}_{k-1} - \boldsymbol{x}^* \right\|_2^2 \right). \tag{19}$$

Let $\boldsymbol{u}_k = \lambda_k \boldsymbol{y}_{k+1} - (\lambda_k - 1)\boldsymbol{y}_k - \boldsymbol{x}^*$ so $\lambda_{k-1}\boldsymbol{y}_k - (\lambda_{k-1} - 1)\boldsymbol{y}_{k-1} - \boldsymbol{x}^* = \boldsymbol{u}_{k-1}$ and (19) becomes

$$\begin{array}{rcll}
\lambda_k^2\delta_{k+1} - \lambda_{k-1}^2\delta_k & \leq & -\frac{L}{2}\left(\|\boldsymbol{u}_k\|_2^2 - \|\boldsymbol{u}_{k-1}\|_2^2\right) & \\
\lambda_1^2\delta_2 - \lambda_0^2\delta_1 & \leq & -\frac{L}{2}\left(\|\boldsymbol{u}_1\|_2^2 - \|\boldsymbol{u}_0\|_2^2\right) & \text{case } k=1 \\
\lambda_2^2\delta_3 - \lambda_1^2\delta_2 & \leq & -\frac{L}{2}\left(\|\boldsymbol{u}_2\|_2^2 - \|\boldsymbol{u}_1\|_2^2\right) & \text{case } k=2 \\
 & \vdots & & \\
\lambda_{K-1}^2\delta_K - \lambda_{K-2}^2\delta_{K-1} & \leq & -\frac{L}{2}\left(\|\boldsymbol{u}_{K-1}\|_2^2 - \|\boldsymbol{u}_{K-2}\|_2^2\right) & \text{case } k=K-1 \\
\lambda_{K-1}^2\delta_K - \lambda_0^2\delta_1 & \leq & -\frac{L}{2}\left(\|\boldsymbol{u}_{K-1}\|_2^2 - \|\boldsymbol{u}_0\|_2^2\right) & \text{sum } k=1 \text{ to } k=K-1 \\
 & = & \frac{L}{2}\left(\|\boldsymbol{u}_0\|_2^2 - \|\boldsymbol{u}_{K-1}\|_2^2\right) & \\
 & \leq & \frac{L}{2}\|\boldsymbol{u}_0\|_2^2 & \|\boldsymbol{u}_{K-1}\|_2^2 \geq 0
\end{array}$$

By definition, $\lambda_0 = 0$, $\boldsymbol{y}_0 = \boldsymbol{x}_0$, $\boldsymbol{u}_0 = \lambda_0\boldsymbol{y}_1 - (\lambda_0 - 1)\boldsymbol{y}_0 - \boldsymbol{x}^* \overset{\lambda_0=0}{=} \boldsymbol{y}_0 - \boldsymbol{x}^* \overset{\boldsymbol{y}_0=\boldsymbol{x}_0}{=} \boldsymbol{x}_0 - \boldsymbol{x}^*$, thus

$$\lambda_{K-1}^2\delta_K \leq \frac{L}{2}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2 \quad \Longrightarrow \quad \delta_K \leq \frac{L\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2}{2\lambda_{K-1}^2}.$$

Proof ... 6/6

Lemma. $\lambda_{k-1} \geq \frac{k}{2}$.
Proof (by induction)

- Case $k = 0$ and $\lambda_0 = 0$. It is trivial $0 \geq 0/2$.
- Case $k = 1$. By definition,
  $$\lambda_k = \frac{1+\sqrt{1+4\lambda_{k-1}^2}}{2} = \frac{1+\sqrt{1+4\cdot 0^2}}{2} = 1 > \frac{1}{2} = \frac{k}{2}\Big|_{k=1}$$
- Induction hypothesis: assume $\lambda_{n-1} \geq \frac{n}{2}$.
- Case $k = n$

$$\begin{aligned}
\lambda_n &= \frac{1+\sqrt{1+4\lambda_{n-1}^2}}{2} \\
&\geq \frac{1+\sqrt{1+4\left(\frac{n}{2}\right)^2}}{2} \qquad \text{[Induction hypothesis]} \\
&= \frac{1+\sqrt{1+n^2}}{2} \\
&> \frac{1+\sqrt{n^2}}{2} \\
&= \frac{1+n}{2}. \quad \square
\end{aligned}$$

With $\lambda_{k-1} \geq \dfrac{k}{2}$, so

$$\frac{1}{\lambda_{k-1}^2} \leq \frac{4}{k^2}.$$

Therefore $\delta_K \leq \dfrac{L\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2}{2\lambda_{K-1}^2}$ becomes

$$f(\boldsymbol{y}_K) - f^* \leq \frac{2L\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2}{K^2}.$$

where $f(\boldsymbol{y}_K) - f^* =: \delta_K$. $\square$

Rename K as k gives

$$f(\boldsymbol{y}_k) - f^* \leq \frac{2L\|x_0 - \boldsymbol{x}^*\|_2^2}{k^2}.$$

This $\begin{cases} \text{complicated} \\ \text{highly-involved} \\ \text{non-intuitive} \end{cases}$ proof is now completed.

Last page - summary

For unconstrained convex smooth problem

$$(\mathcal{P}) \;:\; \underset{\boldsymbol{x}}{\operatorname{argmin}}\; f(\boldsymbol{x})$$

with $f : \mathbb{R}^n \to \mathbb{R}$ being convex, L-smooth, the NAGD algorithm starts with initial point $\boldsymbol{x}_0 = \boldsymbol{y}_0 \in \mathbb{R}^n$ and $\lambda_0 = 0$ and iterates the following:

$$\begin{aligned}
\text{Gradient update} \quad \boldsymbol{y}_{k+1} &= \boldsymbol{x}_k - \frac{1}{L}\nabla f(\boldsymbol{x}_k) \\
\text{Extrapolation} \quad \boldsymbol{x}_{k+1} &= (1-\gamma_k)\boldsymbol{y}_{k+1} + \gamma_k \boldsymbol{y}_k \\
\text{Extrapolation weight} \quad \gamma_k &= \frac{1-\lambda_k}{\lambda_{k+1}} \\
\text{Extrapolation weight} \quad \lambda_k &= \frac{1+\sqrt{1+4\lambda_{k-1}^2}}{2}
\end{aligned}$$

the sequences $\{f(\boldsymbol{y}_k)\}_{k\in\mathbb{N}}$ produced will converges to the optimal f^* at order of $\mathcal{O}\left(\frac{1}{k^2}\right)$ as

$$f(\boldsymbol{y}_k) - f^* \;\leq\; \frac{2L\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2}{k^2}.$$

The proof can be used for proximal gradient descent.

End of document