Nesterov's accelerated gradient method on m-strongly convex L-smooth functionconverges at $\mathcal{O}(\exp \frac{-k}{\sqrt{Q}})$

Andersen Ang

ECS, Uni. Southampton, UK
andersen.ang@soton.ac.uk
Homepage angms.science

Version: July 21, 2023
First draft: August 2, 2017

Content

Nesterov's estimate sequence

$$\Phi_0(\boldsymbol{x}) \coloneqq f(\boldsymbol{x}_0) + \frac{m}{2}\|\boldsymbol{x} - \boldsymbol{x}_0\|_2^2$$

$$\Phi_{k+1}(\boldsymbol{x}) \coloneqq \left(1 - \frac{1}{\sqrt{Q}}\right)\Phi_k(\boldsymbol{x}) + \frac{1}{\sqrt{Q}}\left(f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x} - \boldsymbol{x}_k\rangle + \frac{m}{2}\|\boldsymbol{x} - \boldsymbol{x}_k\|_2^2\right)$$

Lemma 1 $\Phi_k(\boldsymbol{x}) \leq f(\boldsymbol{x}) + \left(1 - \frac{1}{\sqrt{Q}}\right)^k \left(\Phi_0(\boldsymbol{x}) - f(\boldsymbol{x})\right)$

Lemma 2 $\nabla^2 \Phi_k(\boldsymbol{x}) = m\boldsymbol{I}_n$

Lemma 3 $f(\boldsymbol{y}_k) \leq \Phi_k^* \coloneqq \min_{x \in \mathbb{R}^n} \Phi_k(\boldsymbol{x})$

Lemma 4 $\boldsymbol{\nu}_k - \boldsymbol{x}_k = \sqrt{Q}(\boldsymbol{x}_k - \boldsymbol{y}_K)$

NAG convergence rate $f(\boldsymbol{y}_k) - f^* \leq \left(\frac{m+L}{2}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2\right) \exp \frac{-k}{\sqrt{Q}}$

Problem setup: unconstrained strongly convex smooth optimisation

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

- We consider Euclidean space
- $f : \mathbb{R}^n \to \mathbb{R}$ 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 globally L-Lipschitz — $L > 0$ is the least upper bound in $\dfrac{\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|}{\|\boldsymbol{x} - \boldsymbol{y}\|} \leq L$
  - $(\forall \boldsymbol{a} \in \operatorname{dom} f)(\forall \boldsymbol{b} \in \operatorname{dom} f)\Big\{ f(\boldsymbol{a}) - f(\boldsymbol{b}) \leq \langle \nabla f(\boldsymbol{b}), \boldsymbol{a} - \boldsymbol{b} \rangle + \dfrac{L}{2}\|\boldsymbol{a} - \boldsymbol{b}\|_2^2 \Big\}$
- $f : \mathbb{R}^n \to \mathbb{R}$ is m-strongly convex — the global minima of $\mathcal{P}$ is unique
  - f is convex — all local minima of $\mathcal{P}$ are global minima
    - $(\forall \boldsymbol{x} \in \operatorname{dom} f)(\forall \boldsymbol{y} \in \operatorname{dom} f)\Big\{ f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle \Big\}$
  - f is m-strongly convex
    - $f(\boldsymbol{x}) - \dfrac{m}{2}\|\boldsymbol{x}\|_2^2$ is convex
- Details of L-smoothness, convexity, strong convexity, see here

Gradient Descent (GD)

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

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

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

If stepsize is sufficiently small ($m_k < \frac{2}{L}$), then $\{\boldsymbol{x}_k\}_{k\in\mathbb{N}}$ converges to a stationary point of f.

- f convex $\implies$ {all local minimizers are global} — $\{\boldsymbol{x}_k\}_{k\in\mathbb{N}} \to$ a global minimizer $\boldsymbol{x}^*$ (if it exists)
- f strongly convex $\implies$ {unique minimizer} — global minimizer $\boldsymbol{x}^*$ is unique (if it exists)

- Notation $f^* \coloneqq f(\boldsymbol{x}^*)$ and $Q = \dfrac{L}{m}$.

- **If f is L-smooth and convex**, $f_k - f^* \leq \mathcal{O}\left(\frac{1}{k}\right)$ — convergence rate on $\{f_k\}_{k\in\mathbb{N}}$ is $\mathcal{O}\left(\frac{1}{k}\right)$
  Details

- **If f is L-smooth and m-strongly convex**, $f_k - f^* \leq \mathcal{O}\left(\exp \frac{-k}{Q}\right)$ — convergence rate on $\{f_k\}_{k\in\mathbb{N}}$ is $\mathcal{O}\left(\exp \frac{-k}{Q}\right)$
  Details

Nesterov's accelerated gradient (NAG) method

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

If f is L-smooth and convex

Algorithm 1: NAG (for convex smooth f)

1 Initialize $\boldsymbol{x}_0 \in \mathbb{R}^n$, $\lambda_1 = 1$
2 **while** not converge **do**
3

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

Theorem The sequence $\left\{f(\boldsymbol{x}_k)\right\}_{k\in\mathbb{N}}$ produced by NAG on convex L-smooth function satisfies

$$f(\boldsymbol{y}_k) - f^* \ \leq \ \left(\frac{1}{k^2}\right).$$

Details

If f is L-smooth and m-strongly convex
Fix $\gamma_k = \dfrac{\sqrt{Q}-1}{\sqrt{Q}+1}$ where $Q = \dfrac{L}{m}$

Algorithm 2: NAG (for strongly convex smooth f)

1 Initialize $\boldsymbol{x}_0 \in \mathbb{R}^n$
2 **while** not converge **do**
3

$$\begin{aligned} \boldsymbol{y}_{k+1} &= \boldsymbol{x}_k - \frac{1}{L}\nabla f(\boldsymbol{x}_k) \\ \boldsymbol{x}_{k+1} &= \left(1 - \frac{\sqrt{Q}-1}{\sqrt{Q}+1}\right)\boldsymbol{y}_{k+1} + \frac{\sqrt{Q}-1}{\sqrt{Q}+1}\boldsymbol{y}_k \end{aligned}$$

Theorem The sequence $\{f(\boldsymbol{x}_k)\}_{k\in\mathbb{N}}$ produced by NAG on m-strongly convex L-smooth function satisfies

$$f(\boldsymbol{y}_k) - f^* \ \leq \ \frac{m+L}{2}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2 \exp\left(\frac{-k}{\sqrt{Q}}\right).$$

This pdf: prove this.

Convergence rate of NAG - proof idea: Nesterov's estimate sequence

- There are a few ways to prove the convergence of NAG.
- A way is to use a non-trivial technique known as the Nesterov's estimate sequence.
- Consider a sequence of function $\left\{\Phi_k(\boldsymbol{x})\right\}_{k\in\mathbb{N}}$ that
  - $\Phi_k(\boldsymbol{x})$ has a general structure with "parameters" varies with iteration k.
  - $\Phi_k(\boldsymbol{x})$ is based on f
  - $\Phi_k(\boldsymbol{x})$ is m-strongly convex
- $\Phi_k(\boldsymbol{x})$ can be defined as

$$\begin{aligned}\Phi_0(\boldsymbol{x}) &:= f(\boldsymbol{x}_0) + \frac{m}{2}\|\boldsymbol{x}-\boldsymbol{x}_0\|_2^2 \\ \Phi_{k+1}(\boldsymbol{x}) &:= \left(1-\frac{1}{\sqrt{Q}}\right)\Phi_k(\boldsymbol{x}) + \frac{1}{\sqrt{Q}}\left(f(\boldsymbol{x}_k) + \langle\nabla f(\boldsymbol{x}_k), \boldsymbol{x}-\boldsymbol{x}_k\rangle + \frac{m}{2}\|\boldsymbol{x}-\boldsymbol{x}_k\|_2^2\right)\end{aligned}$$

Details of the theory of Nesterov's estimating sequence.

Understanding Nesterov's estimate sequence

$$\Phi_0(\boldsymbol{x}) \quad := \quad f(\boldsymbol{x}_0) + \frac{m}{2}\|\boldsymbol{x} - \boldsymbol{x}_0\|_2^2$$

$$\Phi_{k+1}(\boldsymbol{x}) \quad := \quad \left(1 - \frac{1}{\sqrt{Q}}\right)\Phi_k(\boldsymbol{x}) + \frac{1}{\sqrt{Q}}\left(f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x} - \boldsymbol{x}_k \rangle + \frac{m}{2}\|\boldsymbol{x} - \boldsymbol{x}_k\|_2^2\right)$$

- $\Phi_k(\boldsymbol{x})$ is based on f
- $\Phi_k(\boldsymbol{x})$ is m-strongly convex
- $\Phi_k(\boldsymbol{x})$ varies with iteration k.
- We can see Φ_{k+1} is in the form $\Phi_{k+1} = (1-\lambda)a + \lambda b$.
  - Φ_{k+1} is a convex combination of Φ_k and the 2nd-order Taylor approximation of f at $\boldsymbol{x}_k$.
  - $Q = 1 \iff$ the level sets of f is circular: Φ_{k+1} is more like the Taylor approximation

    In fact by definition of NAG, if $Q = 1$, there is no acceleration and NAG reduces to GD.

    In this case GD should solve the optimization problem in 1 step. Details.
  - $Q \gg 1 \iff$ the level sets of f is elliptic: Φ_{k+1} is more like previous Φ_k

The derivatives of Nesterov's estimating sequence

$$\begin{aligned}\Phi_0(\boldsymbol{x}) &:= f(\boldsymbol{x}_0) + \frac{m}{2}\|\boldsymbol{x}-\boldsymbol{x}_0\|_2^2 \\ \Phi_{k+1}(\boldsymbol{x}) &:= \left(1-\frac{1}{\sqrt{Q}}\right)\Phi_k(\boldsymbol{x}) + \frac{1}{\sqrt{Q}}\left(f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x}-\boldsymbol{x}_k\rangle + \frac{m}{2}\|\boldsymbol{x}-\boldsymbol{x}_k\|_2^2\right)\end{aligned}$$

- With respected to $\boldsymbol{x}$, the gradient and Hessian are

$$\begin{aligned}\nabla\Phi_0(\boldsymbol{x}) &= m(\boldsymbol{x}-\boldsymbol{x}_0) && (1)\\ \nabla^2\Phi_0(\boldsymbol{x}) &= m\boldsymbol{I}_n && (2)\\ \nabla\Phi_{k+1}(\boldsymbol{x}) &= \left(1-\frac{1}{\sqrt{Q}}\right)\nabla\Phi_k(\boldsymbol{x}) + \frac{1}{\sqrt{Q}}\Big(\nabla f(\boldsymbol{x}_k) + m(\boldsymbol{x}-\boldsymbol{x}_k)\Big) && (3)\\ \nabla^2\Phi_{k+1}(\boldsymbol{x}) &= \left(1-\frac{1}{\sqrt{Q}}\right)\nabla^2\Phi_k(\boldsymbol{x}) + \frac{1}{\sqrt{Q}}m\boldsymbol{I}_n && (4)\end{aligned}$$

- In other words,
  - Φ_{k+1} is a convex combination of Φ_k and the 2nd-order Taylor approximation of f at $\boldsymbol{x}_k$.
  - $\nabla\Phi_{k+1}(\boldsymbol{x})$ is a convex combination of $\nabla\Phi_k(\boldsymbol{x})$ and $\nabla f(\boldsymbol{x}_k) + m(\boldsymbol{x}-\boldsymbol{x}_k)$.
  - $\nabla^2\Phi_{k+1}(\boldsymbol{x})$ is a convex combination of $\nabla^2\Phi_k(\boldsymbol{x})$ and $m\boldsymbol{I}_n$.
    In fact we are going to show $\nabla^2\Phi_{k+1}(\boldsymbol{x}) = m\boldsymbol{I}_n$ in Lemma 2.
  - In fact the derivatives of Φ_k plays an important role in the whole proof.

$\Phi_k(\boldsymbol{x})$ with $k = 0, 1$

$$\Phi_{k+1}(\boldsymbol{x}) := \left(1 - \frac{1}{\sqrt{Q}}\right)\Phi_k(\boldsymbol{x}) + \frac{1}{\sqrt{Q}}\left(f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x} - \boldsymbol{x}_k\rangle + \frac{m}{2}\|\boldsymbol{x} - \boldsymbol{x}_k\|_2^2\right)$$

$$f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x} - \boldsymbol{x}_k\rangle + \frac{m}{2}\|\boldsymbol{x} - \boldsymbol{x}_k\|_2^2 \;\leq\; f(\boldsymbol{x}) \;\forall \boldsymbol{x}, \boldsymbol{x}_k$$

f is m-strongly cvx

$$\begin{aligned}
\Phi_0(\boldsymbol{x}) &:= f(\boldsymbol{x}_0) + \frac{m}{2}\|\boldsymbol{x} - \boldsymbol{x}_0\|_2^2 \\
\Phi_1(\boldsymbol{x}) &:= \left(1 - \frac{1}{\sqrt{Q}}\right)\Phi_0(\boldsymbol{x}) + \frac{1}{\sqrt{Q}}\left(f(\boldsymbol{x}_0) + \langle \nabla f(\boldsymbol{x}_0), \boldsymbol{x} - \boldsymbol{x}_0\rangle + \frac{m}{2}\|\boldsymbol{x} - \boldsymbol{x}_0\|_2^2\right) \\
&= f(\boldsymbol{x}) + \left(1 - \frac{1}{\sqrt{Q}}\right)\Phi_0(\boldsymbol{x}) + \frac{1}{\sqrt{Q}}\left(f(\boldsymbol{x}_0) + \langle \nabla f(\boldsymbol{x}_0), \boldsymbol{x} - \boldsymbol{x}_0\rangle + \frac{m}{2}\|\boldsymbol{x} - \boldsymbol{x}_0\|_2^2\right) - f(\boldsymbol{x}) \\
&\leq f(\boldsymbol{x}) + \left(1 - \frac{1}{\sqrt{Q}}\right)\Phi_0(\boldsymbol{x}) + \frac{1}{\sqrt{Q}} f(\boldsymbol{x}) - f(\boldsymbol{x}) \\
&= f(\boldsymbol{x}) + \left(1 - \frac{1}{\sqrt{Q}}\right)\Phi_0(\boldsymbol{x}) - \left(1 - \frac{1}{\sqrt{Q}}\right) f(\boldsymbol{x}) \\
&= f(\boldsymbol{x}) + \left(1 - \frac{1}{\sqrt{Q}}\right)\left(\Phi_0(\boldsymbol{x}) - f(\boldsymbol{x})\right)
\end{aligned}$$

$\Phi_k(\boldsymbol{x})$ with $k = 2$

$$\Phi_{k+1}(\boldsymbol{x}) \coloneqq \left(1 - \frac{1}{\sqrt{Q}}\right)\Phi_k(\boldsymbol{x}) + \frac{1}{\sqrt{Q}}\left(f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x} - \boldsymbol{x}_k\rangle + \frac{m}{2}\|\boldsymbol{x} - \boldsymbol{x}_k\|_2^2\right)$$

$$f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x} - \boldsymbol{x}_k\rangle + \frac{m}{2}\|\boldsymbol{x} - \boldsymbol{x}_k\|_2^2 \;\leq\; f(\boldsymbol{x}) \;\forall \boldsymbol{x}, \boldsymbol{x}_k$$

f is m-strongly cvx

$$\begin{aligned}
\Phi_0(\boldsymbol{x}) &\coloneqq f(\boldsymbol{x}_0) + \frac{m}{2}\|\boldsymbol{x} - \boldsymbol{x}_0\|_2^2 \\
\Phi_1(\boldsymbol{x}) &\leq f(\boldsymbol{x}) + \left(1 - \frac{1}{\sqrt{Q}}\right)\left(\Phi_0(\boldsymbol{x}) - f(\boldsymbol{x})\right) \\
\Phi_2(\boldsymbol{x}) &\coloneqq \left(1 - \frac{1}{\sqrt{Q}}\right)\Phi_1(\boldsymbol{x}) + \frac{1}{\sqrt{Q}}\left(f(\boldsymbol{x}_1) + \langle \nabla f(\boldsymbol{x}_1), \boldsymbol{x} - \boldsymbol{x}_1\rangle + \frac{m}{2}\|\boldsymbol{x} - \boldsymbol{x}_1\|_2^2\right) \\
&= f(\boldsymbol{x}) + \left(1 - \frac{1}{\sqrt{Q}}\right)\Phi_1(\boldsymbol{x}) + \frac{1}{\sqrt{Q}}\left(f(\boldsymbol{x}_1) + \langle \nabla f(\boldsymbol{x}_1), \boldsymbol{x} - \boldsymbol{x}_1\rangle + \frac{m}{2}\|\boldsymbol{x} - \boldsymbol{x}_1\|_2^2\right) - f(\boldsymbol{x}) \\
&\leq f(\boldsymbol{x}) + \left(1 - \frac{1}{\sqrt{Q}}\right)\Phi_1(\boldsymbol{x}) + \frac{1}{\sqrt{Q}} f(\boldsymbol{x}) - f(\boldsymbol{x}) \\
&= f(\boldsymbol{x}) + \left(1 - \frac{1}{\sqrt{Q}}\right)\Phi_1(\boldsymbol{x}) - \left(1 - \frac{1}{\sqrt{Q}}\right) f(\boldsymbol{x}) \\
&= f(\boldsymbol{x}) + \left(1 - \frac{1}{\sqrt{Q}}\right)\left(\Phi_1(\boldsymbol{x}) - f(\boldsymbol{x})\right) \\
&\leq f(\boldsymbol{x}) + \left(1 - \frac{1}{\sqrt{Q}}\right)\left(1 - \frac{1}{\sqrt{Q}}\right)\left(\Phi_0(\boldsymbol{x}) - f(\boldsymbol{x})\right) \;=\; f(\boldsymbol{x}) + \left(1 - \frac{1}{\sqrt{Q}}\right)^2\left(\Phi_0(\boldsymbol{x}) - f(\boldsymbol{x})\right)
\end{aligned}$$

Lemma 1

$$\Phi_{k+1}(\boldsymbol{x}) := \left(1 - \frac{1}{\sqrt{Q}}\right)\Phi_k(\boldsymbol{x}) + \frac{1}{\sqrt{Q}}\left(f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x} - \boldsymbol{x}_k\rangle + \frac{m}{2}\|\boldsymbol{x} - \boldsymbol{x}_k\|_2^2\right)$$

$f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x} - \boldsymbol{x}_k\rangle + \frac{m}{2}\|\boldsymbol{x} - \boldsymbol{x}_k\|_2^2 \;\leq\; f(\boldsymbol{x}) \;\forall \boldsymbol{x}, \boldsymbol{x}_k$ $\quad$ f is m-strongly cvx

$$\begin{aligned}
\Phi_0(\boldsymbol{x}) &:= f(\boldsymbol{x}_0) + \frac{m}{2}\|\boldsymbol{x} - \boldsymbol{x}_0\|_2^2 \\
\Phi_1(\boldsymbol{x}) &\leq f(\boldsymbol{x}) + \left(1 - \frac{1}{\sqrt{Q}}\right)\left(\Phi_0(\boldsymbol{x}) - f(\boldsymbol{x})\right) \\
\Phi_2(\boldsymbol{x}) &\leq f(\boldsymbol{x}) + \left(1 - \frac{1}{\sqrt{Q}}\right)^2\left(\Phi_0(\boldsymbol{x}) - f(\boldsymbol{x})\right)
\end{aligned}$$

Lemma 1 For all $k \in \mathbb{N} = \{1, 2, \ldots\}$,

$$\Phi_k(\boldsymbol{x}) \leq f(\boldsymbol{x}) + \left(1 - \frac{1}{\sqrt{Q}}\right)^k\left(\Phi_0(\boldsymbol{x}) - f(\boldsymbol{x})\right).$$

Proof by induction

- Based case is already proved.
- For case $k+1$, repeat the procedure on deriving Φ_2 and make use of the induction hypothesis.

Lemma 2 $\nabla^2 \Phi_k(\boldsymbol{x}) = m\boldsymbol{I}_n$

$$
\begin{aligned}
\Phi_0(\boldsymbol{x}) &:= f(\boldsymbol{x}_0) + \frac{m}{2}\|\boldsymbol{x} - \boldsymbol{x}_0\|_2^2 \\
\Phi_{k+1}(\boldsymbol{x}) &:= \left(1 - \frac{1}{\sqrt{Q}}\right)\Phi_k(\boldsymbol{x}) + \frac{1}{\sqrt{Q}}\left(f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x} - \boldsymbol{x}_k\rangle + \frac{m}{2}\|\boldsymbol{x} - \boldsymbol{x}_k\|_2^2\right)
\end{aligned}
$$

Proof by induction

- **Base case** $k = 0$
  $\nabla^2 \Phi_0(\boldsymbol{x}) = m\boldsymbol{I}_n$ by definition.

- **Induction Hypothesis** $\nabla^2 \Phi_k(\boldsymbol{x}) = m\boldsymbol{I}_n$

- **Case** $k + 1$

$$
\begin{aligned}
\Phi_{k+1}(\boldsymbol{x}) &= \left(1 - \frac{1}{\sqrt{Q}}\right)\Phi_k(\boldsymbol{x}) + \frac{1}{\sqrt{Q}}\left(f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x} - \boldsymbol{x}_k\rangle + \frac{m}{2}\|\boldsymbol{x} - \boldsymbol{x}_k\|_2^2\right) && \text{by definition} \\
\nabla^2 \Phi_{k+1}(\boldsymbol{x}) &= \left(1 - \frac{1}{\sqrt{Q}}\right)\nabla^2 \Phi_k(\boldsymbol{x}) + \frac{1}{\sqrt{Q}} m\boldsymbol{I}_n \\
&= \left(1 - \frac{1}{\sqrt{Q}}\right) m\boldsymbol{I}_n + \frac{1}{\sqrt{Q}} m\boldsymbol{I}_n && \text{induction hypothesis} \\
&= m\boldsymbol{I}_n \quad \square
\end{aligned}
$$

Lemma 3 $f(\boldsymbol{y}_k) \leq \Phi_k^* \coloneqq \min_{\boldsymbol{x}\in\mathbb{R}^n} \Phi_k(\boldsymbol{x})$ ... 1/7

$\Phi_0(\boldsymbol{x})$	$\coloneqq$	$f(\boldsymbol{x}_0) + \frac{m}{2}\|\boldsymbol{x} - \boldsymbol{x}_0\|_2^2$	estimate seq.
$\boldsymbol{x}_0$	$=$	$\boldsymbol{y}_0$	NAG def.
$\boldsymbol{y}_{k+1}$	$=$	$\boldsymbol{x}_k - \frac{1}{L}\nabla f(\boldsymbol{x}_k)$	NAG def.
$f(\boldsymbol{a}) - f(\boldsymbol{b})$	$\leq$	$\langle \nabla f(\boldsymbol{b}), \boldsymbol{a} - \boldsymbol{b}\rangle + \frac{L}{2}\|\boldsymbol{a} - \boldsymbol{b}\|_2^2$	L-smooth

Proof by induction

- **Base case** $k = 0$
  $\Phi_0^* = \min_{\boldsymbol{x}\in\mathbb{R}^n} \Phi_0(\boldsymbol{x}_0) = \min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}_0) + \frac{m}{2}\|\boldsymbol{x} - \boldsymbol{x}_0\|_2^2 = f(\boldsymbol{x}_0) = f(\boldsymbol{y}_0)$
- **Induction Hypothesis** $f(\boldsymbol{y}_k) \leq \Phi_k^*$
- **Case** $k+1$ Consider $f(\boldsymbol{y}_{k+1})$ and L-smoothness of f

$$
\begin{aligned}
f(\boldsymbol{y}_{k+1}) &\leq f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k),\, \boldsymbol{y}_{k+1} - \boldsymbol{x}_k\rangle + \frac{L}{2}\|\boldsymbol{y}_{k+1} - \boldsymbol{x}_k\|_2^2 \\
&= f(\boldsymbol{x}_k) + \left\langle \nabla f(\boldsymbol{x}_k),\, \frac{-\nabla f(\boldsymbol{x}_k)}{L}\right\rangle + \frac{L}{2}\left\|\frac{-\nabla f(\boldsymbol{x}_k)}{L}\right\|_2^2 \quad \text{NAG update} \\
&= f(\boldsymbol{x}_k) - \frac{1}{L}\|\nabla f(\boldsymbol{x}_k)\|_2^2 + \frac{1}{2L}\|\nabla f(\boldsymbol{x}_k)\|_2^2 \\
&= f(\boldsymbol{x}_k) - \frac{1}{2L}\|\nabla f(\boldsymbol{x}_k)\|_2^2
\end{aligned}
$$

Now for shorthand notation we will let $g \coloneqq \frac{1}{2L}\|\nabla f(\boldsymbol{x}_k)\|_2^2$, we have $f(\boldsymbol{y}_{k+1}) \leq f(\boldsymbol{x}_k) - g$.

Lemma 3 ... 2/7

$f(\boldsymbol{y}_k) \geq f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{y}_k - \boldsymbol{x}_k \rangle$ f convex
$f(\boldsymbol{y}_k) \leq \Phi_k^*$ Induction Hypothesis

- From $f(\boldsymbol{y}_{k+1}) \leq f(\boldsymbol{x}_k) - g$, two tricky steps to create $\left(1 - \frac{1}{\sqrt{Q}}\right)$

$$
\begin{aligned}
f(\boldsymbol{y}_{k+1}) &\leq f(\boldsymbol{x}_k) - \frac{f(\boldsymbol{x}_k)}{\sqrt{Q}} + \frac{f(\boldsymbol{x}_k)}{\sqrt{Q}} - g \\
&= \left(1 - \frac{1}{\sqrt{Q}}\right) f(\boldsymbol{x}_k) + \frac{f(\boldsymbol{x}_k)}{\sqrt{Q}} - g \\
&= \left(1 - \frac{1}{\sqrt{Q}}\right) f(\boldsymbol{x}_k) - \left(1 - \frac{1}{\sqrt{Q}}\right) f(\boldsymbol{y}_k) + \left(1 - \frac{1}{\sqrt{Q}}\right) f(\boldsymbol{y}_k) + \frac{f(\boldsymbol{x}_k)}{\sqrt{Q}} - g \\
&= \left(1 - \frac{1}{\sqrt{Q}}\right) \left(f(\boldsymbol{x}_k) - f(\boldsymbol{y}_k)\right) + \left(1 - \frac{1}{\sqrt{Q}}\right) f(\boldsymbol{y}_k) + \frac{f(\boldsymbol{x}_k)}{\sqrt{Q}} - g \\
&\leq \left(1 - \frac{1}{\sqrt{Q}}\right) \left(f(\boldsymbol{x}_k) - f(\boldsymbol{y}_k)\right) + \left(1 - \frac{1}{\sqrt{Q}}\right) \Phi_k^* + \frac{1}{\sqrt{Q}} f(\boldsymbol{x}_k) - g && \text{induction hypothesis} \\
&\leq \left(1 - \frac{1}{\sqrt{Q}}\right) \left\langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x}_k - \boldsymbol{y}_k \right\rangle + \left(1 - \frac{1}{\sqrt{Q}}\right) \Phi_k^* + \frac{1}{\sqrt{Q}} f(\boldsymbol{x}_k) - g && f \text{ convex}
\end{aligned}
$$

$$
f(\boldsymbol{y}_{k+1}) \leq \left(1 - \frac{1}{\sqrt{Q}}\right) \left\langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x}_k - \boldsymbol{y}_k \right\rangle + \left(1 - \frac{1}{\sqrt{Q}}\right) \Phi_k^* + \frac{1}{\sqrt{Q}} f(\boldsymbol{x}_k) - g \qquad \text{(now we have)}
$$

- Recall our goal is to show $f(\boldsymbol{y}_{k+1}) \leq \Phi_{k+1}^*$, we can try to show

$$
\left(1 - \frac{1}{\sqrt{Q}}\right) \left\langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x}_k - \boldsymbol{y}_k \right\rangle + \left(1 - \frac{1}{\sqrt{Q}}\right) \Phi_k^* + \frac{1}{\sqrt{Q}} f(\boldsymbol{x}_k) - g \leq \Phi_{k+1}^*. \qquad \text{(what we want to prove)}
$$

This is what we are going to do in the next 4 - 5 slides.

Lemma 3 ... 3/7

- Now consider $\Phi_k(\boldsymbol{x})$. Lemma 2 $\nabla^2\Phi_k(\boldsymbol{x}) = m\boldsymbol{I}_n$ implies $\Phi_k(\boldsymbol{x}) = \Phi_k^* + \frac{m}{2}\|\boldsymbol{x} - \boldsymbol{\nu}_k\|_2^2$ for some $\boldsymbol{\nu}_k \in \mathbb{R}^n$ implies
  1. $\nabla\Phi_k(\boldsymbol{x}) = m(\boldsymbol{x} - \boldsymbol{\nu}_k)$
  2. Φ_k is minimized at $\boldsymbol{\nu}_k$, which implies $\nabla\Phi_k(\boldsymbol{\nu}_k) = 0$
  3. Points 1,2 work for all k, including $k+1$
  4. From $\Phi_0(\boldsymbol{x}) = f(\boldsymbol{x}_0) + \frac{m}{2}\|\boldsymbol{x} - \boldsymbol{x}_0\|_2^2$, $\boldsymbol{\nu}_0 = \boldsymbol{x}_0$

- By definition of $\Phi_{k+1}(\boldsymbol{x})$ in Nesterov's estimate sequence

$$
\begin{aligned}
\Phi_{k+1}(\boldsymbol{x}) &= \left(1 - \frac{1}{\sqrt{Q}}\right)\Phi_k(\boldsymbol{x}) + \frac{1}{\sqrt{Q}}\left(f(\boldsymbol{x}_k) + \langle\nabla f(\boldsymbol{x}_k), \boldsymbol{x} - \boldsymbol{x}_k\rangle + \frac{m}{2}\|\boldsymbol{x} - \boldsymbol{x}_k\|_2^2\right) \\
\nabla\Phi_{k+1}(\boldsymbol{x}) &= \left(1 - \frac{1}{\sqrt{Q}}\right)\nabla\Phi_k(\boldsymbol{x}) + \frac{1}{\sqrt{Q}}\left(\nabla f(\boldsymbol{x}_k) + m(\boldsymbol{x} - \boldsymbol{x}_k)\right) \\
&= \left(1 - \frac{1}{\sqrt{Q}}\right)m(\boldsymbol{x} - \boldsymbol{\nu}_k) + \frac{1}{\sqrt{Q}}\left(\nabla f(\boldsymbol{x}_k) + m(\boldsymbol{x} - \boldsymbol{x}_k)\right) \\
\nabla\Phi_{k+1}(\boldsymbol{\nu}_{k+1}) &= \left(1 - \frac{1}{\sqrt{Q}}\right)m(\boldsymbol{\nu}_{k+1} - \boldsymbol{\nu}_k) + \frac{1}{\sqrt{Q}}\left(\nabla f(\boldsymbol{x}_k) + m(\boldsymbol{\nu}_{k+1} - \boldsymbol{x}_k)\right) \\
&= \mathbf{0} \qquad \text{(2) \& (3) gives } \nabla\Phi_{k+1}(\boldsymbol{\nu}_{k+1}) = \mathbf{0}
\end{aligned}
$$

Lemma 3 ... 4/7 (just some algebra)

$$\boxed{\left(1-\frac{1}{\sqrt{Q}}\right)m(\boldsymbol{\nu}_{k+1}-\boldsymbol{\nu}_k)+\frac{1}{\sqrt{Q}}\left(\nabla f(\boldsymbol{x}_k)+m(\boldsymbol{\nu}_{k+1}-\boldsymbol{x}_k)\right)\;=\;\mathbf{0}}$$

$$\left(1-\frac{1}{\sqrt{Q}}\right)(\boldsymbol{\nu}_{k+1}-\boldsymbol{\nu}_k)+\frac{1}{\sqrt{Q}}\left(\frac{\nabla f(\boldsymbol{x}_k)}{m}+(\boldsymbol{\nu}_{k+1}-\boldsymbol{x}_k)\right) \;=\; \mathbf{0}$$

$$\iff \quad \left(1-\frac{1}{\sqrt{Q}}\right)\boldsymbol{\nu}_{k+1}-\left(1-\frac{1}{\sqrt{Q}}\right)\boldsymbol{\nu}_k+\frac{1}{\sqrt{Q}}\boldsymbol{\nu}_{k+1}+\frac{1}{\sqrt{Q}}\left(\frac{\nabla f(\boldsymbol{x}_k)}{m}-\boldsymbol{x}_k\right) \;=\; \mathbf{0}$$

Now

$$\boldsymbol{\nu}_{k+1} = \left(1-\frac{1}{\sqrt{Q}}\right)\boldsymbol{\nu}_k+\frac{1}{\sqrt{Q}}\left(\boldsymbol{x}_k-\frac{\nabla f(\boldsymbol{x}_k)}{m}\right) \tag{5}$$

$$\iff \quad -\boldsymbol{\nu}_{k+1} = -\left(1-\frac{1}{\sqrt{Q}}\right)\boldsymbol{\nu}_k-\frac{1}{\sqrt{Q}}\left(\boldsymbol{x}_k-\frac{\nabla f(\boldsymbol{x}_k)}{m}\right)$$

$$\iff \quad \boldsymbol{x}_k-\boldsymbol{\nu}_{k+1} = \boldsymbol{x}_k-\left(1-\frac{1}{\sqrt{Q}}\right)\boldsymbol{\nu}_k-\frac{1}{\sqrt{Q}}\boldsymbol{x}_k+\frac{1}{\sqrt{Q}}\frac{\nabla f(\boldsymbol{x}_k)}{m}$$

$$= \left(1-\frac{1}{\sqrt{Q}}\right)\left(\boldsymbol{x}_k-\boldsymbol{\nu}_k\right)+\frac{\nabla f(\boldsymbol{x}_k)}{m\sqrt{Q}}$$

$$\iff \quad \|\boldsymbol{x}_k-\boldsymbol{\nu}_{k+1}\|_2^2 = \left(1-\frac{1}{\sqrt{Q}}\right)^2\|\boldsymbol{x}_k-\boldsymbol{\nu}_k\|_2^2+2\left(1-\frac{1}{\sqrt{Q}}\right)\frac{\langle\nabla f(\boldsymbol{x}_k),\,\boldsymbol{x}_k-\boldsymbol{\nu}_k\rangle}{m\sqrt{Q}}+\frac{\|\nabla f(\boldsymbol{x}_k)\|_2^2}{m^2Q}$$

Lemma 3 ... 5/7

$$\|\boldsymbol{x}_k - \boldsymbol{\nu}_{k+1}\|_2^2 = \left(1 - \frac{1}{\sqrt{Q}}\right)^2 \|\boldsymbol{x}_k - \boldsymbol{\nu}_k\|_2^2 + 2\left(1 - \frac{1}{\sqrt{Q}}\right) \frac{\langle \nabla f(\boldsymbol{x}_k),\, \boldsymbol{x}_k - \boldsymbol{\nu}_k \rangle}{m\sqrt{Q}} + \frac{\|\nabla f(\boldsymbol{x}_k)\|_2^2}{m^2 Q}$$

- Now consider $\Phi_{k+1}(\boldsymbol{x})$ evaluate at $\boldsymbol{x}_k$, from in slide 14 we have

$$\begin{aligned}
\Phi_{k+1}(\boldsymbol{x}_k) &= \Phi^*_{k+1} + \tfrac{m}{2}\|\boldsymbol{x}_k - \boldsymbol{\nu}_{k+1}\|_2^2 \\
&= \Phi^*_{k+1} + \frac{m}{2}\left(1 - \frac{1}{\sqrt{Q}}\right)^2 \|\boldsymbol{x}_k - \boldsymbol{\nu}_k\|_2^2 + \left(1 - \frac{1}{\sqrt{Q}}\right) \frac{\langle \nabla f(\boldsymbol{x}_k),\, \boldsymbol{x}_k - \boldsymbol{\nu}_k \rangle}{\sqrt{Q}} + \frac{\|\nabla f(\boldsymbol{x}_k)\|_2^2}{2mQ} \\
&= \Phi^*_{k+1} + \frac{m}{2}\left(1 - \frac{1}{\sqrt{Q}}\right)^2 \|\boldsymbol{x}_k - \boldsymbol{\nu}_k\|_2^2 + \left(1 - \frac{1}{\sqrt{Q}}\right) \frac{\langle \nabla f(\boldsymbol{x}_k),\, \boldsymbol{x}_k - \boldsymbol{\nu}_k \rangle}{\sqrt{Q}} + g \quad (*)
\end{aligned}$$

by using the fact $mQ = L$ and $g = \frac{1}{2L}\|\nabla f(\boldsymbol{x}_k)\|_2^2$.

- By definition of $\Phi_{k+1}(\boldsymbol{x})$ from page 5, $\Phi_{k+1}(\boldsymbol{x}_k)$ is

$$\begin{aligned}
\Phi_{k+1}(\boldsymbol{x}_k) &= \left(1 - \frac{1}{\sqrt{Q}}\right)\Phi_k(\boldsymbol{x}_k) + \frac{1}{\sqrt{Q}}\left(f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \underbrace{\boldsymbol{x}_k - \boldsymbol{x}_k}_{=\mathbf{0}} \rangle + \frac{m}{2}\|\underbrace{\boldsymbol{x}_k - \boldsymbol{x}_k}_{=\mathbf{0}}\|_2^2\right) \\
&= \left(1 - \frac{1}{\sqrt{Q}}\right)\Phi_k(\boldsymbol{x}_k) + \frac{1}{\sqrt{Q}} f(\boldsymbol{x}_k) \quad (**)
\end{aligned}$$

- $(*) = (**)$ gives

$$\Phi^*_{k+1} + \frac{m}{2}\left(1 - \frac{1}{\sqrt{Q}}\right)^2 \|\boldsymbol{x}_k - \boldsymbol{\nu}_k\|_2^2 + \left(1 - \frac{1}{\sqrt{Q}}\right) \frac{\langle \nabla f(\boldsymbol{x}_k),\, \boldsymbol{x}_k - \boldsymbol{\nu}_k \rangle}{\sqrt{Q}} + g = \left(1 - \frac{1}{\sqrt{Q}}\right)\Phi_k(\boldsymbol{x}_k) + \frac{1}{\sqrt{Q}} f(\boldsymbol{x}_k)$$

Lemma 3 ... 6/7

$$\Phi^*_{k+1} = -\frac{m}{2}\left(1-\frac{1}{\sqrt{Q}}\right)^2 \|\boldsymbol{x}_k - \boldsymbol{\nu}_k\|_2^2 - \left(1-\frac{1}{\sqrt{Q}}\right)\frac{\langle \nabla f(\boldsymbol{x}_k),\, \boldsymbol{x}_k - \boldsymbol{\nu}_k\rangle}{\sqrt{Q}} - g + \left(1-\frac{1}{\sqrt{Q}}\right)\Phi_k(\boldsymbol{x}_k) + \frac{1}{\sqrt{Q}} f(\boldsymbol{x}_k)$$

By $\Phi_k(\boldsymbol{x}) = \Phi^*_k + \frac{m}{2}\|\boldsymbol{x} - \boldsymbol{\nu}_k\|_2^2$ (slide 14)

$$\left(1-\frac{1}{\sqrt{Q}}\right)\Phi_k(\boldsymbol{x}_k) \;=\; \left(1-\frac{1}{\sqrt{Q}}\right)\Phi^*_k + \left(1-\frac{1}{\sqrt{Q}}\right)\frac{m}{2}\|\boldsymbol{x}_k - \boldsymbol{\nu}_k\|_2^2$$

Hence

$$\Phi^*_{k+1} \;=\; -\frac{m}{2}\left(1-\frac{1}{\sqrt{Q}}\right)^2 \|\boldsymbol{x}_k - \boldsymbol{\nu}_k\|_2^2 - \left(1-\frac{1}{\sqrt{Q}}\right)\frac{\langle \nabla f(\boldsymbol{x}_k),\, \boldsymbol{x}_k - \boldsymbol{\nu}_k\rangle}{\sqrt{Q}} - g + \left(1-\frac{1}{\sqrt{Q}}\right)\Phi^*_k + \left(1-\frac{1}{\sqrt{Q}}\right)\frac{m}{2}\|\boldsymbol{x}_k - \boldsymbol{\nu}_k\|_2^2 + \frac{f(\boldsymbol{x}_k)}{\sqrt{Q}}$$

Simplify the term

$$\Phi^*_{k+1} \;=\; \frac{m}{2\sqrt{Q}}\left(1-\frac{1}{\sqrt{Q}}\right)^2 \|\boldsymbol{x}_k - \boldsymbol{\nu}_k\|_2^2 - \left(1-\frac{1}{\sqrt{Q}}\right)\frac{\langle \nabla f(\boldsymbol{x}_k),\, \boldsymbol{x}_k - \boldsymbol{\nu}_k\rangle}{\sqrt{Q}} + \left(1-\frac{1}{\sqrt{Q}}\right)\Phi^*_k - g + \frac{f(\boldsymbol{x}_k)}{\sqrt{Q}}$$

To proceed, we need lemma 4.

Lemma 4 $\boldsymbol{\nu}_k - \boldsymbol{x}_k = \sqrt{Q}(\boldsymbol{x}_k - \boldsymbol{y}_k)$

$$
\begin{array}{rcll}
Q & = & \frac{L}{m} & \text{def of } Q \\
\boldsymbol{y}_{k+1} & = & \boldsymbol{x}_k - \frac{\nabla f(\boldsymbol{x}_k)}{L} & \text{NAG def (1)} \\
\boldsymbol{x}_{k+1} & = & \left(1 + \frac{\sqrt{Q}-1}{\sqrt{Q}+1}\right)\boldsymbol{y}_{k+1} - \frac{\sqrt{Q}-1}{\sqrt{Q}+1}\boldsymbol{y}_k & \text{NAG def (2)}
\end{array}
$$

Proof by induction

- **Base case** $k = 0$ is true by $\boldsymbol{x}_0 = \boldsymbol{y}_0$ hence $\boldsymbol{\nu}_0 = \boldsymbol{x}_0$.
- **Induction hypothesis** $\boldsymbol{\nu}_k - \boldsymbol{x}_k = \sqrt{Q}(\boldsymbol{x}_k - \boldsymbol{y}_k)$
- **Case** $k+1$

$$
\begin{aligned}
\boldsymbol{\nu}_{k+1} &\overset{(5)}{=} \left(1 - \frac{1}{\sqrt{Q}}\right)\boldsymbol{\nu}_k + \frac{1}{\sqrt{Q}}\left(\boldsymbol{x}_k - \frac{\nabla f(\boldsymbol{x}_k)}{m}\right) \\
&= \left(1 - \frac{1}{\sqrt{Q}}\right)\boldsymbol{\nu}_k + \frac{1}{\sqrt{Q}}\left(\boldsymbol{x}_k - \frac{Q\nabla f(\boldsymbol{x}_k)}{L}\right) && \text{def of } Q \\
\boldsymbol{\nu}_{k+1} - \boldsymbol{x}_{k+1} &= \left(1 - \frac{1}{\sqrt{Q}}\right)\boldsymbol{\nu}_k + \frac{1}{\sqrt{Q}}\left(\boldsymbol{x}_k - \frac{Q\nabla f(\boldsymbol{x}_k)}{L}\right) - \boldsymbol{x}_{k+1} \\
&= \left(1 - \frac{1}{\sqrt{Q}}\right)\left(\boldsymbol{x}_k + \sqrt{Q}(\boldsymbol{x}_k - \boldsymbol{y}_k)\right) + \frac{1}{\sqrt{Q}}\boldsymbol{x}_k - \sqrt{Q}\frac{\nabla f(\boldsymbol{x}_k)}{L} - \boldsymbol{x}_{k+1} && \text{induction hypothesis} \\
&= \sqrt{Q}\left(\boldsymbol{x}_k - \frac{\nabla f(\boldsymbol{x}_k)}{L}\right) - \left(\sqrt{Q} - 1\right)\boldsymbol{y}_k - \boldsymbol{x}_{k+1} \\
&= \sqrt{Q}\boldsymbol{y}_{k+1} + \left(\sqrt{Q} + 1\right)\boldsymbol{x}_{k+1} - 2\sqrt{Q}\boldsymbol{y}_{k+1} - \boldsymbol{x}_{k+1} && \text{NAG def (1) NAG def (2)} \\
&= \sqrt{Q}(\boldsymbol{x}_{k+1} - \boldsymbol{y}_{k+1}) \quad \square
\end{aligned}
$$

Lemma 3 ... 7/7

Lemma 4 $\boldsymbol{\nu}_k - \boldsymbol{x}_k = \sqrt{Q}(\boldsymbol{x}_k - \boldsymbol{y}_k)$

The proof of Lemma 3 stops at

$$\Phi^*_{k+1} = \frac{m}{2\sqrt{Q}}\left(1-\frac{1}{\sqrt{Q}}\right)^2 \|\boldsymbol{x}_k - \boldsymbol{\nu}_k\|_2^2 - \left(1-\frac{1}{\sqrt{Q}}\right)\frac{\langle \nabla f(\boldsymbol{x}_k),\, \boldsymbol{x}_k - \boldsymbol{\nu}_k\rangle}{\sqrt{Q}} + \left(1-\frac{1}{\sqrt{Q}}\right)\Phi^*_k - g + \frac{f(\boldsymbol{x}_k)}{\sqrt{Q}}$$

By lemma 4 we have

$$\begin{aligned}\Phi^*_{k+1} &= \frac{m}{2\sqrt{Q}}\left(1-\frac{1}{\sqrt{Q}}\right)^2 \|\boldsymbol{x}_k - \boldsymbol{\nu}_k\|_2^2 - \left(1-\frac{1}{\sqrt{Q}}\right)\frac{\langle \nabla f(\boldsymbol{x}_k),\, \boldsymbol{x}_k - \boldsymbol{\nu}_k\rangle}{\sqrt{Q}} + \left(1-\frac{1}{\sqrt{Q}}\right)\Phi^*_k - g + \frac{f(\boldsymbol{x}_k)}{\sqrt{Q}} \\ &= \frac{m\sqrt{Q}}{2}\left(1-\frac{1}{\sqrt{Q}}\right)^2 \|\boldsymbol{x}_k - \boldsymbol{y}_k\|_2^2 + \left(1-\frac{1}{\sqrt{Q}}\right)\langle \nabla f(\boldsymbol{x}_k),\, \boldsymbol{x}_k - \boldsymbol{y}_k\rangle + \left(1-\frac{1}{\sqrt{Q}}\right)\Phi^*_k - g + \frac{f(\boldsymbol{x}_k)}{\sqrt{Q}}\end{aligned}$$

Recall (slide 13)

$$f(\boldsymbol{y}_{k+1}) \le \left(1-\frac{1}{\sqrt{Q}}\right)\left\langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x}_k - \boldsymbol{y}_k\right\rangle + \left(1-\frac{1}{\sqrt{Q}}\right)\Phi^*_k + \frac{1}{\sqrt{Q}}f(\boldsymbol{x}_k) - g \qquad \text{(now we have)}$$

By $a = \Phi^*_{k+1} = \underbrace{\quad}_{\ge 0} + \quad \ge \quad \overset{\text{now we have}}{\ge} f(\boldsymbol{y}_{k+1})$, we have proved for the case $k+1$ that $f(\boldsymbol{y}_{k+1}) \le \Phi^*_{k+1}$.

' By induction, Lemma 3 is now proved. $\square$

Proving NAG convergence rate

Lemma 1 $\Phi_{k+1}(\boldsymbol{x}) \leq f(\boldsymbol{x}) + \left(1 - \frac{1}{\sqrt{Q}}\right)^k \left(\Phi_0(\boldsymbol{x}) - f(\boldsymbol{x})\right) \ \forall k$

Lemma 3 $f(\boldsymbol{y}_k) \leq \Phi_k^* \ \forall k$

f L-smooth $f(\boldsymbol{a}) - f(\boldsymbol{b}) \leq \langle \nabla f(\boldsymbol{b}), \boldsymbol{a} - \boldsymbol{b} \rangle + \frac{L}{2}\|\boldsymbol{a} - \boldsymbol{b}\|_2^2$

$\nabla f(\boldsymbol{x}^*) = \boldsymbol{0}$

- **Theorem** $f(\boldsymbol{y}_k) - f^* \leq \frac{m+L}{2}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2 e^{\frac{-k}{\sqrt{Q}}}$

- **Proof**

$$
\begin{aligned}
f(\boldsymbol{y}_k) - f^* &\leq \Phi_k(\boldsymbol{x}^*) - f^* && \text{lemma 3} \\
&\leq f(\boldsymbol{x}^*) + \left(1 - \tfrac{1}{\sqrt{Q}}\right)^k \left(\Phi_0(\boldsymbol{x}^*) - f(\boldsymbol{x}^*)\right) - f^* && \text{lemma 1} \\
&= \left(\Phi_0(\boldsymbol{x}^*) - f^*\right) \left(1 - \tfrac{1}{\sqrt{Q}}\right)^k && f(\boldsymbol{x}^*) = f^* \\
&= \left(f(\boldsymbol{x}_0) - f^* + \tfrac{m}{2}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2\right) \left(1 - \tfrac{1}{\sqrt{Q}}\right)^k && \text{Def. of } \Phi_0(\boldsymbol{x}) \\
&\leq \left(\langle \nabla f(\boldsymbol{x}^*), \boldsymbol{x}_0 - \boldsymbol{x}^* \rangle + \tfrac{L}{2}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2 + \tfrac{m}{2}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2\right) \left(1 - \tfrac{1}{\sqrt{Q}}\right)^k && f \text{ L-smooth} \\
&\leq \frac{m+L}{2}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2 \left(1 + (-\tfrac{1}{\sqrt{Q}})\right)^k && \nabla f(\boldsymbol{x}^*) = \boldsymbol{0} \\
&\leq \frac{m+L}{2}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2 \left(\exp(-\tfrac{1}{\sqrt{Q}})\right)^k && 1 + x \leq e^x \\
&= \frac{m+L}{2}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2 \exp\left(\frac{-k}{\sqrt{Q}}\right)
\end{aligned}
$$

Discussion

- If we stop the algorithm when ϵ-accuracy is achieved

$$\frac{m+L}{2}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2 \exp\left(\frac{-k}{\sqrt{Q}}\right) \quad \leq \quad \epsilon.$$

  Re-arrange

$$k \quad \geq \quad \sqrt{Q}\ln\frac{1}{\epsilon} + \text{constant}.$$

  I.e. it takes $\mathcal{O}\left(\sqrt{Q}\ln\frac{1}{\epsilon}\right)$ steps for NAG to converges.

- Compared to GD with rate $\mathcal{O}\left(Q\ln\frac{1}{\epsilon}\right)$, the improvement $Q \to \sqrt{Q}$ is significant as m can be viewed as regularization parameter in various machine learning model (norm regularized) and $\frac{1}{m}$ can be as large as sample size. Here the number of step reduced from sample size to $\sqrt{\text{sample size}}$.

Last page - summary

- For unconstrained smooth strongly-convex problem $\min_{x\in\mathbb{R}^n} f(\boldsymbol{x})$, with $f:\mathbb{R}^n\to\mathbb{R}$ being L-smooth and m-strongly convex, the NAG algorithm iterates the following :

$$\boldsymbol{y}_{k+1} = \boldsymbol{x}_k - \frac{1}{L}\nabla f(\boldsymbol{x}_k), \qquad \boldsymbol{x}_{k+1} = \left(1 - \frac{\sqrt{Q}-1}{\sqrt{Q}+1}\right)\boldsymbol{y}_{k+1} + \frac{\sqrt{Q}-1}{\sqrt{Q}+1}\boldsymbol{y}_k, \qquad Q = \frac{L}{m}$$

with initial point $\boldsymbol{x}_0 = \boldsymbol{y}_0 \in \mathbb{R}^n$, will produce a sequences $\left\{f(\boldsymbol{y}_k)\right\}_{k\in\mathbb{N}}$ that

$$f(\boldsymbol{y}_k) - f^* \;\leq\; \frac{m+L}{2}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2 \exp\left(\frac{-k}{\sqrt{Q}}\right).$$

End of document