Discover acceleration of gradient descent

Daniil Merkulov

Optimization methods. MIPT

Coordinate shift

Consider the following quadratic optimization problem:

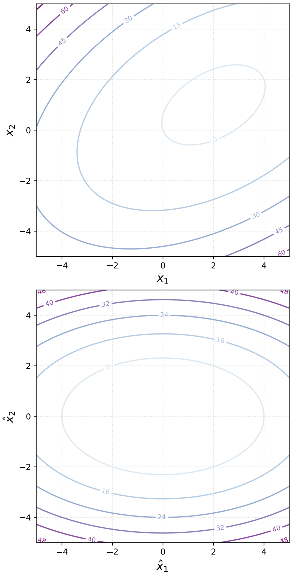
$$
\min_{x \in \mathbb{R}^d} f(x) = \min_{x \in \mathbb{R}^d} \frac{1}{2} x^\top A x - b^\top x + c, \text{ where } A \in \mathbb{S}^d_{++}.
$$

Coordinate shift

Consider the following quadratic optimization problem:

$$\min_{x \in \mathbb{R}^d} f(x) = \min_{x \in \mathbb{R}^d} \frac{1}{2} x^\top A x - b^\top x + c, \text{ where } A \in \mathbb{S}^d_{++}.$$

- Firstly, without loss of generality we can set $c = 0$, which will or affect optimization process.



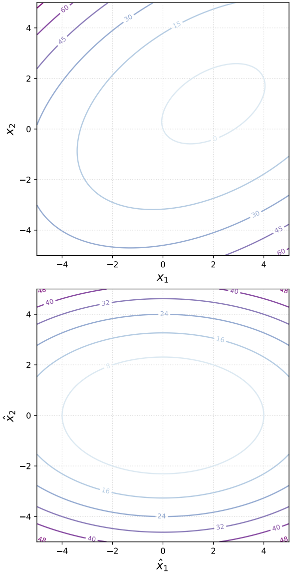
Coordinate shift

Consider the following quadratic optimization problem:

$$\min_{x \in \mathbb{R}^d} f(x) = \min_{x \in \mathbb{R}^d} \frac{1}{2} x^\top A x - b^\top x + c, \text{ where } A \in \mathbb{S}^d_{++}.$$

- Firstly, without loss of generality we can set $c = 0$, which will or affect optimization process.
- Secondly, we have a spectral decomposition of the matrix A:

$$A = Q \Lambda Q^T$$



Coordinate shift

Consider the following quadratic optimization problem:

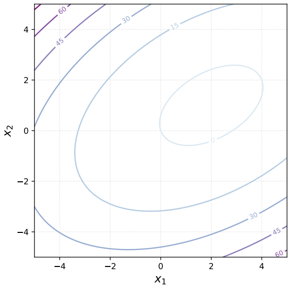
$$\min_{x \in \mathbb{R}^d} f(x) = \min_{x \in \mathbb{R}^d} \frac{1}{2} x^\top A x - b^\top x + c, \text{ where } A \in \mathbb{S}^d_{++}.$$

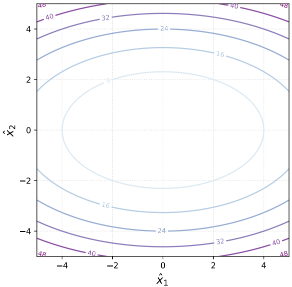
- Firstly, without loss of generality we can set $c = 0$, which will or affect optimization process.
- Secondly, we have a spectral decomposition of the matrix A:

$$A = Q \Lambda Q^T$$

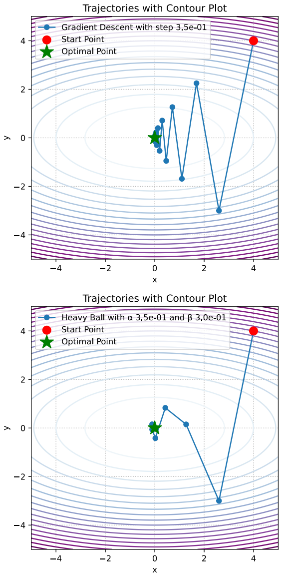
- Let's show, that we can switch coordinates in order to make an analysis a little bit easier. Let $\hat{x} = Q^T(x - x^*)$, where x^* is the minimum point of initial function, defined by $Ax^* = b$. At the same time $x = Q\hat{x} + x^*$.

$$\begin{aligned}
f(\hat{x}) &= \frac{1}{2}(Q\hat{x} + x^*)^\top A(Q\hat{x} + x^*) - b^\top (Q\hat{x} + x^*) \\
&= \frac{1}{2}\hat{x}^T Q^T A Q\hat{x} + (x^*)^T A Q\hat{x} + \frac{1}{2}(x^*)^T A (x^*)^T - b^T Q\hat{x} - b^T x^* \\
&= \frac{1}{2}\hat{x}^T \Lambda \hat{x}
\end{aligned}$$





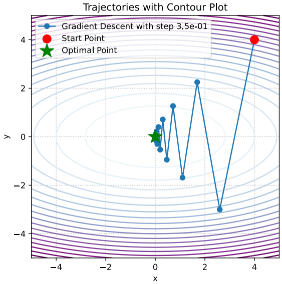
Polyak Heavy ball method

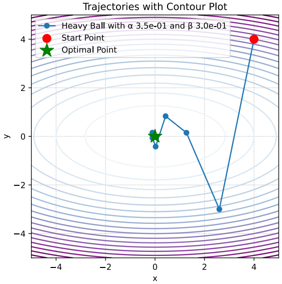


Let's introduce the idea of momentum, proposed by Polyak in 1964. Recall that the momentum update is

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

Polyak Heavy ball method





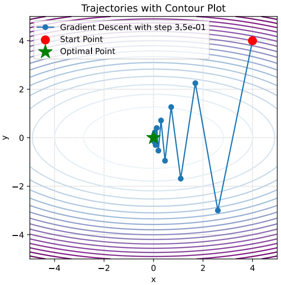
Let's introduce the idea of momentum, proposed by Polyak in 1964. Recall that the momentum update is

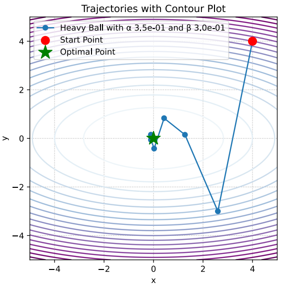
$$x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta (x_k - x_{k-1}).$$

Which is in our case is

$$\hat{x}_{k+1} = \hat{x}_k - \alpha \Lambda \hat{x}_k + \beta (\hat{x}_k - \hat{x}_{k-1}) = (I - \alpha \Lambda + \beta I)\hat{x}_k - \beta \hat{x}_{k-1}$$

Polyak Heavy ball method





Let's introduce the idea of momentum, proposed by Polyak in 1964. Recall that the momentum update is

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

Which is in our case is

$$\hat{x}_{k+1} = \hat{x}_k - \alpha \Lambda \hat{x}_k + \beta(\hat{x}_k - \hat{x}_{k-1}) = (I - \alpha\Lambda + \beta I)\hat{x}_k - \beta \hat{x}_{k-1}$$

This can be rewritten as a follows

$$\begin{aligned} \hat{x}_{k+1} &= (I - \alpha\Lambda + \beta I)\hat{x}_k - \beta \hat{x}_{k-1}, \\ \hat{x}_k &= \hat{x}_k. \end{aligned}$$

Let's use the following notation $\hat{z}_k = \begin{bmatrix} \hat{x}_{k+1} \\ \hat{x}_k \end{bmatrix}$. Therefore $\hat{z}_{k+1} = M\hat{z}_k$, where the iteration matrix M is:

$$M = \begin{bmatrix} I - \alpha\Lambda + \beta I & -\beta I \\ I & 0_d \end{bmatrix}.$$

Reduction to a scalar case

Note, that M is $2d \times 2d$ matrix with 4 block-diagonal matrices of size $d \times d$ inside. It means, that we can rearrange the order of coordinates to make M block-diagonal in the following form. Note that in the equation below, the matrix M denotes the same as in the notation above, except for the described permutation of rows and columns. We use this slight abuse of notation for the sake of clarity.

Reduction to a scalar case

Note, that M is $2d \times 2d$ matrix with 4 block-diagonal matrices of size $d \times d$ inside. It means, that we can rearrange the order of coordinates to make M block-diagonal in the following form. Note that in the equation below, the matrix M denotes the same as in the notation above, except for the described permutation of rows and columns. We use this slight abuse of notation for the sake of clarity.

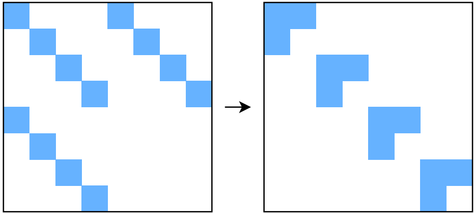


Figure 1: Illustration of matrix M rearrangement

$$
\begin{bmatrix} \hat{x}_k^{(1)} \\ \vdots \\ \hat{x}_k^{(d)} \\ \hat{x}_{k-1}^{(1)} \\ \vdots \\ \hat{x}_{k-1}^{(d)} \end{bmatrix} \to \begin{bmatrix} \hat{x}_k^{(1)} \\ \hat{x}_{k-1}^{(1)} \\ \vdots \\ \hat{x}_k^{(d)} \\ \hat{x}_{k-1}^{(d)} \end{bmatrix} \quad M = \begin{bmatrix} M_1 & & & \\ & M_2 & & \\ & & \ldots & \\ & & & M_d \end{bmatrix}
$$

where $\hat{x}_k^{(i)}$ is i-th coordinate of vector $\hat{x}_k \in \mathbb{R}^d$ and M_i stands for 2×2 matrix. This rearrangement allows us to study the dynamics of the method independently for each dimension. One may observe, that the asymptotic convergence rate of the $2d$-dimensional vector sequence of $\hat{z}_k$ is defined by the worst convergence rate among its block of coordinates. Thus, it is enough to study the optimization in a one-dimensional case.

Reduction to a scalar case

For i-th coordinate with λ_i as an i-th eigenvalue of matrix W we have:

$$M_i = \begin{bmatrix} 1 - \alpha\lambda_i + \beta & -\beta \\ 1 & 0 \end{bmatrix}.$$

The method will be convergent if $\rho(M) < 1$, and the optimal parameters can be computed by optimizing the spectral radius

$$\alpha^*, \beta^* = \arg\min_{\alpha,\beta} \max_{\lambda \in [\mu, L]} \rho(M) \quad \alpha^* = \frac{4}{(\sqrt{L} + \sqrt{\mu})^2}; \quad \beta^* = \left(\frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}} \right)^2.$$

It can be shown, that for such parameters the matrix M has complex eigenvalues, which forms a conjugate pair, so the distance to the optimum (in this case, $\|z_k\|$), generally, will not go to zero monotonically.

Heavy ball quadratic convergence

We can explicitly calculate the eigenvalues of M_i:

$$\lambda^M_1, \lambda^M_2 = \lambda \left(\begin{bmatrix} 1 - \alpha\lambda_i + \beta & -\beta \\ 1 & 0 \end{bmatrix} \right) = \frac{1 + \beta - \alpha\lambda_i \pm \sqrt{(1 + \beta - \alpha\lambda_i)^2 - 4\beta}}{2}.$$

When α and β are optimal (α^*, β^*), the eigenvalues are complex-conjugated pair $(1 + \beta - \alpha\lambda_i)^2 - 4\beta \leq 0$, i.e. $\beta \geq (1 - \sqrt{\alpha\lambda_i})^2$.

$$\text{Re}(\lambda^M_1) = \frac{L + \mu - 2\lambda_i}{(\sqrt{L} + \sqrt{\mu})^2}; \quad \text{Im}(\lambda^M_1) = \frac{\pm 2\sqrt{(L - \lambda_i)(\lambda_i - \mu)}}{(\sqrt{L} + \sqrt{\mu})^2}; \quad |\lambda^M_1| = \frac{L - \mu}{(\sqrt{L} + \sqrt{\mu})^2}.$$

And the convergence rate does not depend on the stepsize and equals to $\sqrt{\beta^*}$.

Heavy ball method summary

- Ensures accelerated convergence for strongly convex quadratic problems

Heavy ball method summary

- Ensures accelerated convergence for strongly convex quadratic problems
- Local accelerated convergence was proved in the original paper.

Heavy ball method summary

- Ensures accelerated convergence for strongly convex quadratic problems
- Local accelerated convergence was proved in the original paper.
- Recently was proved, that there is no global accelerated convergence for the method.

Heavy ball method summary

- Ensures accelerated convergence for strongly convex quadratic problems
- Local accelerated convergence was proved in the original paper.
- Recently was proved, that there is no global accelerated convergence for the method.
- Method was not extremely popular until the ML boom

Heavy ball method summary

- Ensures accelerated convergence for strongly convex quadratic problems
- Local accelerated convergence was proved in the original paper.
- Recently was proved, that there is no global accelerated convergence for the method.
- Method was not extremely popular until the ML boom
- Nowadays, it is de-facto standard for practical acceleration of gradient methods, even for the non-convex problems (neural network training)

Nesterov accelerated gradient

$$x_{k+1} = x_k - \alpha \nabla f(x_k) \tag{GD}$$

$$x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta (x_k - x_{k-1}) \tag{HB}$$

$$\begin{cases} y_{k+1} = x_k + \beta (x_k - x_{k-1}) \\ x_{k+1} = y_{k+1} - \alpha \nabla f(y_{k+1}) \end{cases} \tag{NAG}$$

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

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

- 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

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

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

- **Theorem**. If $f : \mathbb{R}^n \to \mathbb{R}$ is L-smooth and convex, the sequences $\{f(\boldsymbol{y}_k)\}_k$ 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 $\mathcal{O}\left(\frac{1}{k^2}\right)$ 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 $\{f(\boldsymbol{y}_k)\}_k$ 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 $(\forall \boldsymbol{a} \forall \boldsymbol{b})\Big\{ f(\boldsymbol{a}) - f(\boldsymbol{b}) \leq \big\langle \nabla f(\boldsymbol{b}), \boldsymbol{a} - \boldsymbol{b} \big\rangle + \dfrac{L}{2}\|\boldsymbol{a} - \boldsymbol{b}\|_2^2 \Big\}$, with $\boldsymbol{a} = \boldsymbol{x} - \dfrac{1}{L}\nabla f(\boldsymbol{x}), \boldsymbol{b} = \boldsymbol{x}$,

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

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

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

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

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

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

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

- 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**

$$
\begin{aligned}
\boldsymbol{y}_{k+1} &= \boldsymbol{x}_k - \frac{1}{L}\nabla f(\boldsymbol{x}_k) && (1)\\
f\Big(\boldsymbol{x}_k - \frac{1}{L}\nabla f(\boldsymbol{x}_k)\Big) - f^* &\le \frac{-1}{2L}\|\nabla f(\boldsymbol{x}_k)\|_2^2 + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x}_k - \boldsymbol{x}^*\rangle. && (8)\\
f\Big(\boldsymbol{x}_k - \frac{1}{L}\nabla f(\boldsymbol{x}_k)\Big) - f(\boldsymbol{y}_k) &\le \frac{-1}{2L}\|\nabla f(\boldsymbol{x}_k)\|_2^2 + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x}_k - \boldsymbol{y}_k\rangle. && (9)
\end{aligned}
$$

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

$$
\begin{aligned}
f\Big(\boldsymbol{x}_k - \frac{1}{L}\nabla f(\boldsymbol{x}_k)\Big) &\overset{(1)}{=} f(\boldsymbol{y}_{k+1}) && (10)\\
f\Big(\boldsymbol{x}_k - \frac{1}{L}\nabla f(\boldsymbol{x}_k)\Big) - f^* &\overset{(10),\delta_k}{=} \delta_{k+1} && (11)\\
f\Big(\boldsymbol{x}_k - \frac{1}{L}\nabla f(\boldsymbol{x}_k)\Big) - f(\boldsymbol{y}_k) &= f\Big(\boldsymbol{x}_k - \frac{1}{L}\nabla f(\boldsymbol{x}_k)\Big) - f^* - \Big(f(\boldsymbol{y}_k) - f^*\Big)\\
&= \delta_{k+1} - \delta_k && (12)\\
\nabla f(\boldsymbol{x}_k) &\overset{(1)}{=} -L(\boldsymbol{y}_{k+1} - \boldsymbol{x}_k) && (13)\\
\|\nabla f(\boldsymbol{x}_k)\|_2^2 &\overset{(13)}{=} L^2\|\boldsymbol{y}_{k+1} - \boldsymbol{x}_k\|_2^2 && (14)
\end{aligned}
$$

- 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) + (λ_k-1)(16).

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

- 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 $\quad \lambda_k\delta_{k+1}-(\lambda_k-1)\delta_k \;\leq\; -\dfrac{\lambda_k L}{2}\|\boldsymbol{y}_{k+1}-\boldsymbol{x}_k\|_2^2 - L\Big\langle \boldsymbol{y}_{k+1}-\boldsymbol{x}_k,\; \lambda_k\boldsymbol{x}_k-(\lambda_k-1)\boldsymbol{y}_k-\boldsymbol{x}^*\Big\rangle.$
  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

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

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

- 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

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

- 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

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

Proof ... 5/6

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

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

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

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

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

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:

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

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

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

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

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

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 $f(\boldsymbol{y}_k) \leq \Phi_k^* := \min\limits_{x \in \mathbb{R}^n} \Phi_k(\boldsymbol{x})$

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 $\dfrac{\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|}{\|\boldsymbol{x} - \boldsymbol{y}\|} \leq L$
  - $(\forall \boldsymbol{a} \in \operatorname{dom} f)(\forall \boldsymbol{b} \in \operatorname{dom} f)\Big\{ f(\boldsymbol{a}) - f(\boldsymbol{b}) \leq \big\langle \nabla f(\boldsymbol{b}), \boldsymbol{a} - \boldsymbol{b} \big\rangle + \dfrac{L}{2}\|\boldsymbol{a} - \boldsymbol{b}\|_2^2 \Big\}$
- $f : \mathbb{R}^n \to \mathbb{R}$ is m-strongly convex — the global minima of $\mathcal{P}$ is unique
  - f is convex — all local minima of $\mathcal{P}$ are global minima
    - $(\forall \boldsymbol{x} \in \operatorname{dom} f)(\forall \boldsymbol{y} \in \operatorname{dom} f)\Big\{ f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \big\langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \big\rangle \Big\}$
  - 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)

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

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

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

- **If f is L-smooth and m-strongly convex**, $f_k - f^* \le \mathcal{O}\left(\exp \frac{-k}{Q}\right)$ 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)

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

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

Theorem The sequence $\{f(\boldsymbol{x}_k)\}_{k\in\mathbb{N}}$ 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)

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

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

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 $\{\Phi_k(\boldsymbol{x})\}_{k\in\mathbb{N}}$ 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

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

- $\Phi_k(\boldsymbol{x})$ 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

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

- 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 $k = 0, 1$

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

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

f is m-strongly cvx

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

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

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

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

f is m-strongly cvx

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

Lemma 1

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

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

f is m-strongly cvx

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

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

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

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

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

Proof by induction

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

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

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

$$\begin{array}{lll} f(\boldsymbol{y}_k) & \geq & f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{y}_k - \boldsymbol{x}_k \rangle \quad f \text{ convex} \\ f(\boldsymbol{y}_k) & \leq & \Phi_k^* \quad \text{Induction Hypothesis} \end{array}$$

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

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

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

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

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

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

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

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

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

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

Now

$$\begin{aligned}
\boldsymbol{\nu}_{k+1} &= \left(1-\frac{1}{\sqrt{Q}}\right)\boldsymbol{\nu}_k+\frac{1}{\sqrt{Q}}\left(\boldsymbol{x}_k-\frac{\nabla f(\boldsymbol{x}_k)}{m}\right) \qquad (5) \\
\iff \quad -\boldsymbol{\nu}_{k+1} &= -\left(1-\frac{1}{\sqrt{Q}}\right)\boldsymbol{\nu}_k-\frac{1}{\sqrt{Q}}\left(\boldsymbol{x}_k-\frac{\nabla f(\boldsymbol{x}_k)}{m}\right) \\
\iff \quad \boldsymbol{x}_k-\boldsymbol{\nu}_{k+1} &= \boldsymbol{x}_k-\left(1-\frac{1}{\sqrt{Q}}\right)\boldsymbol{\nu}_k-\frac{1}{\sqrt{Q}}\boldsymbol{x}_k+\frac{1}{\sqrt{Q}}\frac{\nabla f(\boldsymbol{x}_k)}{m} \\
&= \left(1-\frac{1}{\sqrt{Q}}\right)\Big(\boldsymbol{x}_k-\boldsymbol{\nu}_k\Big)+\frac{\nabla f(\boldsymbol{x}_k)}{m\sqrt{Q}} \\
\iff \quad \|\boldsymbol{x}_k-\boldsymbol{\nu}_{k+1}\|_2^2 &= \left(1-\frac{1}{\sqrt{Q}}\right)^2\|\boldsymbol{x}_k-\boldsymbol{\nu}_k\|_2^2+2\left(1-\frac{1}{\sqrt{Q}}\right)\frac{\big\langle\nabla f(\boldsymbol{x}_k),\,\boldsymbol{x}_k-\boldsymbol{\nu}_k\big\rangle}{m\sqrt{Q}}+\frac{\|\nabla f(\boldsymbol{x}_k)\|_2^2}{m^2Q}
\end{aligned}$$

Lemma 3 ... 5/7

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

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

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

  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

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

- $(*) = (**)$ 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

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

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

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

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.

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

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

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

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

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 $a = \Phi_{k+1}^* = \underbrace{\square}_{\ge 0} + \square \ge \square \overset{\text{now we have}}{\ge} f(\boldsymbol{y}_{k+1})$, 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. □

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 $f(\boldsymbol{a}) - f(\boldsymbol{b}) \leq \langle \nabla f(\boldsymbol{b}), \boldsymbol{a} - \boldsymbol{b} \rangle + \frac{L}{2} \|\boldsymbol{a} - \boldsymbol{b}\|_2^2$
$\nabla f(\boldsymbol{x}^*) = \mathbf{0}$

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

- **Proof**

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

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 $\big\{f(\boldsymbol{y}_k)\big\}_{k \in \mathbb{N}}$ 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