Nesterov's estimate sequence: 2. General scheme of optimal method

Andersen Ang

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

Homepage angms.science

Version: August 8, 2023
First draft: July 31, 2023

Content

Initialize $\boldsymbol{x}_0 \in \mathbb{R}^n$, $\gamma_0 > 0$, $\boldsymbol{v}_0 = \boldsymbol{x}_0$
for $k = 1, 2, \ldots$ **do**
  Compute $\alpha_k \in]\,0, 1\,[$ from $L\alpha_k^2 \;=\; (1-\alpha_k)\gamma_k + \alpha_k\mu$
  $\gamma_{k+1} = (1-\alpha_k)\gamma_k + \alpha_k\mu$
  $\boldsymbol{y}_k = \frac{\alpha_k\gamma_k\boldsymbol{v}_k + \gamma_{k+1}\boldsymbol{x}_k}{\gamma_k + \alpha_k\mu}$
  Find $\boldsymbol{x}_{k+1}$ s.t. $f(\boldsymbol{x}_{k+1}) \;\leq\; f(\boldsymbol{y}_k) - \frac{1}{2L}\|\nabla f(\boldsymbol{y}_k)\|_2^2$. E.g., a gradient step
  $\boldsymbol{v}_k = \frac{(1-\alpha_k)\gamma_k\boldsymbol{v}_k + \alpha_k\mu\boldsymbol{y}_k - \alpha_k\nabla f(\boldsymbol{y}_k)}{\gamma_{k+1}}$

$$f(\boldsymbol{x}_k) - f^* \;\leq\; \lambda_k\Big[f(\boldsymbol{x}_0) - f^* + \frac{\gamma_0}{2}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2\Big],\quad \lambda_0 = 1,\; \lambda_k = \prod_{i=0}^{k-1}(1-\alpha_i)$$

Convergence of $\{\lambda_k\}_{k\in\mathbb{N}}$: $\quad \lambda_k \;\leq\; \frac{\mu}{(\gamma_0-\mu)}\,\frac{1}{\left[\sinh\left((k+1)\sqrt{\frac{\mu}{L}}\right)\right]^2}$

$$f(\boldsymbol{x}_k) - f^* \;\leq\; \frac{f(\boldsymbol{x}_0) - f^* + \frac{\gamma_0}{2}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2}{4(\gamma_0 - \mu)}\,\frac{L}{(k+1)^2}$$

Reference Yurii Nesterov, Lectures on convex optimization. Springer, 2018.

Problem setup: unconstrained convex smooth optimization

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

- $f : \mathbb{R}^n \to \mathbb{R}$ is μ-strongly convex and L-smooth. — $f \in \mathcal{C}_L^{1,1}$
  - f is convex — $\operatorname{dom} f$ is a convex set and $\operatorname{epi} f$ is a convex set
  - f is μ-strongly convex, $\mu \geq 0$ — $f - \frac{\mu}{2}\|\boldsymbol{x}\|_2^2$ is convex
    - The assumption subsume the case for f is convex ($\mu = 0$)
    - If a function ψ is convex, we have $\psi(\boldsymbol{a}) \geq \psi(\boldsymbol{b}) + \langle \nabla\psi(\boldsymbol{b}), \boldsymbol{a} - \boldsymbol{b}\rangle$
  - f is continuous — no jump
  - f is continuously differentiable — $\nabla f(\boldsymbol{x})$ exists for all $\boldsymbol{x} \in \operatorname{dom} f$
  - ∇f is globally L-Lipschitz, $L > 0$ — $\left(\forall \boldsymbol{x} \forall \boldsymbol{y} \neq \boldsymbol{x}\right)\left(\dfrac{\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|}{\|\boldsymbol{x} - \boldsymbol{y}\|} \leq L\right)$

  For the details of convexity, epigraph, smoothness, see here.

- We also assume a solution $\boldsymbol{x}^* \in \mathcal{X}^*$ exists.
  - $\mathcal{X}^* := \operatorname*{argmin}_{\boldsymbol{x}} f(\boldsymbol{x})$ — solution set, assumed nonempty
  - $\boldsymbol{x}^* \in \mathcal{X}^*$ — minimizer
  - $f^* := f(\boldsymbol{x}^*)$ — optimal function value

Things from part 1

- **[Definition 2.2.1 ("what is" estimate sequence)]** A sequences pair $\left\{\phi_k(\boldsymbol{x}),\, \lambda_k\right\}_{k=0}^{\infty}$ is estimate sequence of $f(\cdot)$ if

(Def0)	λ_k	$\geq$	0	$(\forall k)$	$\{\lambda_k\}_{k\in\mathbb{N}}$ is nonnegative
(Def1)	λ_k	$\xrightarrow{k\to\infty}$	0	$(\forall k)$	$\{\lambda_k\}_{k\in\mathbb{N}}$ converges to 0
(Def2)	$\phi_k(\boldsymbol{x})$	$\leq$	$(1-\lambda_k)f(\boldsymbol{x}) + \lambda_k\phi_0(\boldsymbol{x})$	$(\forall k)(\forall \boldsymbol{x}\in\mathbb{R}^n)$	$\{\phi_k\}_{k\in\mathbb{N}} \leq$ "convex combination" of f, ϕ_0

- **[Lemma 2.2.1 ("why of" estimate sequence)]** Assume $\boldsymbol{x}^*$ exists. For a sequence $\{\boldsymbol{x}_k\}_{k\in\mathbb{N}}$:

$$\textbf{IF} \quad f(\boldsymbol{x}_k) \;\leq\; \phi_k^* \;:=\; \min_{\boldsymbol{x}\in\mathbb{R}^n} \phi_k(\boldsymbol{x}) \qquad \textbf{THEN} \quad f(\boldsymbol{x}_k) - f^* \;\leq\; \lambda_k\left(\phi_0(\boldsymbol{x}^*) - f^*\right) \xrightarrow{\text{Def1}} 0.$$

- **[Lemma 2.2.2 ("how to" estimate sequence)]**

	A1	f L-smooth μ-strongly cvx			
	A2	$\phi_0(\cdot)$ a cvx function			
	A3	$\{\boldsymbol{y}_k\}_{k=0}^{\infty}$ is a sequence	**THEN**	$\{\phi_k(\boldsymbol{x}),\, \lambda_k\}_{k=0}^{\infty}$ defined by A6 A7 is an estimate sequence of f
IF	A4a	$\alpha_k \in \left]\, 0,\, 1\,\right[\;\; \forall k$	A6	$\lambda_{k+1} = (1-\alpha_k)\lambda_k$
	A4b	$\sum_{k=0}^{\infty} \alpha_k = \infty$	A7	$\phi_{k+1}(\boldsymbol{x}) = (1-\alpha_k)\phi_k(\boldsymbol{x}) + \alpha_k\left(f(\boldsymbol{y}_k) + \langle \nabla f(\boldsymbol{y}_k), \boldsymbol{x} - \boldsymbol{y}_k\rangle + \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{y}_k\|_2^2\right)$
	A5	$\lambda_0 := 1$		

- **[Lemma 2.2.3 (a quadratic ϕ_0)]** **IF** $\phi_0(\boldsymbol{x}) := \phi_0^* + \frac{\gamma_0}{2}\|\boldsymbol{x} - \boldsymbol{v}_0\|_2^2$ **THEN** $\{\phi_k(\boldsymbol{x})\}_{k=0}^{\infty}$ defined as A7 in Lemma 2.2.2 preserves the canonical form of $\{\phi_k(\boldsymbol{x})\}_{k\in\mathbb{N}}$

$$\phi_k(\boldsymbol{x}) \;=\; \phi_k^* + \frac{\gamma_0}{2}\|\boldsymbol{x} - \boldsymbol{v}_k\|_2^2.$$

where

$$\gamma_{k+1} = (1-\alpha_k)\gamma_k + \alpha_k\mu \qquad (i)$$

$$\boldsymbol{v}_{k+1} = \frac{(1-\alpha_k)\gamma_k\boldsymbol{v}_k + \alpha_k\mu\boldsymbol{y}_k - \alpha_k\nabla f(\boldsymbol{y}_k)}{\gamma_{k+1}} \qquad (ii)$$

$$\phi_{k+1}^* = (1-\alpha_k)\phi_k^* + \alpha_k f(\boldsymbol{y}_k) - \frac{\alpha_k^2}{2\gamma_{k+1}}\|\nabla f(\boldsymbol{y}_k)\|_2^2 + \frac{\alpha_k(1-\alpha_k)\gamma_k}{\gamma_{k+1}}\left(\frac{\mu}{2}\|\boldsymbol{y}_k - \boldsymbol{v}_k\|_2^2 + \langle \nabla f(\boldsymbol{y}_k), \boldsymbol{v}_k - \boldsymbol{y}_k\rangle\right) \qquad (iii)$$

Things to focus

- **[Definition 2.2.1 ("what is" estimate sequence)]** A sequences pair $\left\{\phi_k(\boldsymbol{x}), \lambda_k\right\}_{k=0}^{\infty}$ is estimate sequence of $f(\cdot)$ if

(Def0)	λ_k	$\geq$	0	$(\forall k)$	$\{\lambda_k\}_{k\in\mathbb{N}}$ is nonnegative
(Def1)	λ_k	$\xrightarrow[k\to\infty]{}$	0	$(\forall k)$	$\{\lambda_k\}_{k\in\mathbb{N}}$ converges to 0
(Def2)	$\phi_k(\boldsymbol{x})$	$\leq$	$(1-\lambda_k)f(\boldsymbol{x}) + \lambda_k\phi_0(\boldsymbol{x})$	$(\forall k)(\forall \boldsymbol{x}\in\mathbb{R}^n)$	$\{\phi_k\}_{k\in\mathbb{N}} \leq$ "convex combination" of f, ϕ_0

- **[Lemma 2.2.1 ("why of" estimate sequence)]** Assume $\boldsymbol{x}^*$ exists. For a sequence $\{\boldsymbol{x}_k\}_{k\in\mathbb{N}}$:

IF $f(\boldsymbol{x}_k) \leq \phi_k^* := \min_{\boldsymbol{x}\in\mathbb{R}^n} \phi_k(\boldsymbol{x})$ **THEN** $f(\boldsymbol{x}_k) - f^* \leq \lambda_k\left(\phi_0(\boldsymbol{x}^*) - f^*\right) \xrightarrow{\text{Def1}} 0.$

- **[Lemma 2.2.2 ("how to" estimate sequence)]**

	A1	f L-smooth μ-strongly cvx		
	A2	$\phi_0(\cdot)$ a cvx function		
	A3	$\{\boldsymbol{y}_k\}_{k=0}^{\infty}$ is a sequence	**THEN**	$\{\phi_k(\boldsymbol{x}), \lambda_k\}_{k=0}^{\infty}$ defined by A6 A7 is an estimate sequence of f
IF	A4a	$\alpha_k \in]\,0, 1\,[\ \ \forall k$	A6	$\lambda_{k+1} = (1-\alpha_k)\lambda_k$
	A4b	$\sum_{k=0}^{\infty} \alpha_k = \infty$	A7	$\phi_{k+1}(\boldsymbol{x}) = (1-\alpha_k)\phi_k(\boldsymbol{x}) + \alpha_k\left(f(\boldsymbol{y}_k) + \langle\nabla f(\boldsymbol{y}_k), \boldsymbol{x}-\boldsymbol{y}_k\rangle + \frac{\mu}{2}\|\boldsymbol{x}-\boldsymbol{y}_k\|_2^2\right)$
	A5	$\lambda_0 := 1$		

- **[Lemma 2.2.3 (a quadratic ϕ_0)]** **IF** $\phi_0(\boldsymbol{x}) := \phi_0^* + \frac{\gamma_0}{2}\|\boldsymbol{x}-\boldsymbol{v}_0\|_2^2$ **THEN** $\{\phi_k(\boldsymbol{x})\}_{k=0}^{\infty}$ defined as A7 in Lemma 2.2.2 preserves the canonical form of $\{\phi_k(\boldsymbol{x})\}_{k\in\mathbb{N}}$

$$\phi_k(\boldsymbol{x}) = \phi_k^* + \frac{\gamma_0}{2}\|\boldsymbol{x}-\boldsymbol{v}_k\|_2^2.$$

where

$$\gamma_{k+1} = (1-\alpha_k)\gamma_k + \alpha_k\mu \qquad (i)$$

$$\boldsymbol{v}_{k+1} = \frac{(1-\alpha_k)\gamma_k\boldsymbol{v}_k + \alpha_k\mu\boldsymbol{y}_k - \alpha_k\nabla f(\boldsymbol{y}_k)}{\gamma_{k+1}} \qquad (ii)$$

$$\phi_{k+1}^* = (1-\alpha_k)\phi_k^* + \alpha_k f(\boldsymbol{y}_k) - \frac{\alpha_k^2}{2\gamma_{k+1}}\|\nabla f(\boldsymbol{y}_k)\|_2^2 + \frac{\alpha_k(1-\alpha_k)\gamma_k}{\gamma_{k+1}}\left(\frac{\mu}{2}\|\boldsymbol{y}_k-\boldsymbol{v}_k\|_2^2 + \langle\nabla f(\boldsymbol{y}_k), \boldsymbol{v}_k-\boldsymbol{y}_k\rangle\right) \qquad (iii)$$

Showing $f(\boldsymbol{x}_k) \leq \phi_k^* \ \ldots\ 1/3$

f is L-smooth and μ-strongly convex $\qquad (A1)$

IF $\quad f(\boldsymbol{x}_k) \ \leq\ \phi_k^* := \min_{\boldsymbol{x}\in\mathbb{R}^n} \phi_k(\boldsymbol{x}) \quad$ **THEN** $\quad f(\boldsymbol{x}_k) - f^* \ \leq\ \lambda_k\big(\phi_0(\boldsymbol{x}^*) - f^*\big) \xrightarrow{\text{Def1}} 0.$ (Lemma 2.2.1)

$$\phi_{k+1}^* = (1-\alpha_k)\phi_k^* + \alpha_k f(\boldsymbol{y}_k) - \frac{\alpha_k^2}{2\gamma_{k+1}}\|\nabla f(\boldsymbol{y}_k)\|_2^2 + \frac{\alpha_k(1-\alpha_k)\gamma_k}{\gamma_{k+1}}\Big(\frac{\mu}{2}\|\boldsymbol{y}_k - \boldsymbol{v}_k\|_2^2 + \langle \nabla f(\boldsymbol{y}_k), \boldsymbol{v}_k - \boldsymbol{y}_k\rangle\Big) \qquad (iii)$$

- Idea: if ϕ_{k+1} defined by (iii) satisfies the **IF** condition in Lemma 2.2.1 then we have convergence result of $f(\boldsymbol{x}_k)$ in the **THEN** statement

- By [ ], assume $\phi_k^* \geq f(\boldsymbol{x}_k)$, then (iii) becomes

$$\phi_{k+1}^* \ \geq\ (1-\alpha_k)\, f(\boldsymbol{x}_k) + \alpha_k f(\boldsymbol{y}_k) - \frac{\alpha_k^2}{2\gamma_{k+1}}\|\nabla f(\boldsymbol{y}_k)\|_2^2 + \frac{\alpha_k(1-\alpha_k)\gamma_k}{\gamma_{k+1}}\Big(\frac{\mu}{2}\|\boldsymbol{y}_k - \boldsymbol{v}_k\|_2^2 + \langle \nabla f(\boldsymbol{y}_k), \boldsymbol{v}_k - \boldsymbol{y}_k\rangle\Big) \qquad (iii)$$

- Since $\dfrac{\alpha_k(1-\alpha_k)\gamma_k}{\gamma_{k+1}}\dfrac{\mu}{2}\|\boldsymbol{y}_k - \boldsymbol{v}_k\|_2^2 \geq 0$, then from (iii) we have

$$\phi_{k+1}^* \ \geq\ (1-\alpha_k) f(\boldsymbol{x}_k) + \alpha_k f(\boldsymbol{y}_k) - \frac{\alpha_k^2}{2\gamma_{k+1}}\|\nabla f(\boldsymbol{y}_k)\|_2^2 + \frac{\alpha_k(1-\alpha_k)\gamma_k}{\gamma_{k+1}}\langle \nabla f(\boldsymbol{y}_k), \boldsymbol{v}_k - \boldsymbol{y}_k\rangle$$

- By (A1) f is convex: $f(\boldsymbol{x}_k) \ \geq\ f(\boldsymbol{y}_k) + \langle \nabla f(\boldsymbol{y}_k),\ \boldsymbol{x}_k - \boldsymbol{y}_k\rangle$, so

$$\begin{aligned}
\phi_{k+1}^* &\geq (1-\alpha_k)\Big(f(\boldsymbol{y}_k) + \langle \nabla f(\boldsymbol{y}_k),\ \boldsymbol{x}_k - \boldsymbol{y}_k\rangle\Big) + \alpha_k f(\boldsymbol{y}_k) - \frac{\alpha_k^2}{2\gamma_{k+1}}\|\nabla f(\boldsymbol{y}_k)\|_2^2 + \frac{\alpha_k(1-\alpha_k)\gamma_k}{\gamma_{k+1}}\langle \nabla f(\boldsymbol{y}_k), \boldsymbol{v}_k - \boldsymbol{y}_k\rangle \\
&= f(\boldsymbol{y}_k) + (1-\alpha_k)\,\langle\ \nabla f(\boldsymbol{y}_k)\ ,\ \boldsymbol{x}_k - \boldsymbol{y}_k\rangle - \frac{\alpha_k^2}{2\gamma_{k+1}}\|\nabla f(\boldsymbol{y}_k)\|_2^2 + \frac{\alpha_k\ (1-\alpha_k)\ \gamma_k}{\gamma_{k+1}}\langle\ \nabla f(\boldsymbol{y}_k)\ , \boldsymbol{v}_k - \boldsymbol{y}_k\rangle \\
&= f(\boldsymbol{y}_k) - \frac{\alpha_k^2}{2\gamma_{k+1}}\|\nabla f(\boldsymbol{y}_k)\|_2^2 + (1-\alpha_k)\Big\langle \nabla f(\boldsymbol{y}_k), \boldsymbol{x}_k - \boldsymbol{y}_k + \frac{\alpha_k\gamma_k}{\gamma_{k+1}}(\boldsymbol{v}_k - \boldsymbol{y}_k)\Big\rangle \qquad (\dagger)
\end{aligned}$$

Proving $f(\boldsymbol{x}_k) \leq \phi_k^*$ ... 2/3

f is L-smooth and μ-strongly convex $\qquad (A1)$

IF $f(\boldsymbol{x}_k) \leq \phi_k^* := \min_{\boldsymbol{x}\in\mathbb{R}^n} \phi_k(\boldsymbol{x})$ **THEN** $f(\boldsymbol{x}_k) - f^* \leq \lambda_k\left(\phi_0(\boldsymbol{x}^*) - f^*\right) \xrightarrow{\text{Def1}} 0.$ (Lemma 2.2.1)

$$\phi_{k+1}^* \geq f(\boldsymbol{y}_k) - \frac{\alpha_k^2}{2\gamma_{k+1}}\|\nabla f(\boldsymbol{y}_k)\|_2^2 + (1-\alpha_k)\left\langle \nabla f(\boldsymbol{y}_k), \boldsymbol{x}_k - \boldsymbol{y}_k + \frac{\alpha_k\gamma_k}{\gamma_{k+1}}(\boldsymbol{v}_k - \boldsymbol{y}_k)\right\rangle \qquad (\dagger)$$

$$\gamma_{k+1} = (1-\alpha_k)\gamma_k + \alpha_k\mu \qquad (i)$$

- What we want: make $(\dagger)$ fulfill the **IF** in Lemma 2.2.1, i.e., $\phi_{k+1}^* \geq f(\boldsymbol{x}_{k+1})$
  - Note that $f(\boldsymbol{y}_k)$ in $(\dagger)$ is on $\boldsymbol{y}$ while in the **IF** we need $f(\boldsymbol{x})$.
  - Hence we need to get $\boldsymbol{x}$ from $\boldsymbol{y}$, a way is gradient descent with constant stepsize

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

(GD) on $\underbrace{\text{smooth function } f}_{\text{A1}}$ guarantees the sufficient descent lemma (proof here)

$$f(\boldsymbol{y}_k) - \frac{1}{2L}\|\nabla f(\boldsymbol{y}_k)\|_2^2 \geq f(\boldsymbol{x}_{k+1}). \qquad \text{(SDL)}$$

  - To make use of (SDL) we need to make $\frac{1}{2L} = \frac{\alpha_k^2}{2\gamma_{k+1}}$. This can be done by defining α_k as a positive root of

$$L\alpha_k^2 = (1-\alpha_k)\gamma_k + \alpha_k\mu \overset{(i)}{=:} \gamma_{k+1} \iff L\alpha_k^2 - \gamma_{k+1} = 0 \implies \alpha_k = \frac{-0 \pm \sqrt{0^2 - 4(L)(-\gamma_{k+1})}}{2L} \iff \alpha_k^2 = \frac{\gamma_{k+1}}{L}$$

- Apply (SDL) on $(\dagger)$ gives

$$\phi_{k+1}^* \geq f(\boldsymbol{x}_{k+1}) + (1-\alpha_k)\left\langle \nabla f(\boldsymbol{y}_k), \boldsymbol{x}_k - \boldsymbol{y}_k + \frac{\alpha_k\gamma_k}{\gamma_{k+1}}(\boldsymbol{v}_k - \boldsymbol{y}_k)\right\rangle \qquad (\dagger')$$

Proving $f(\boldsymbol{x}_k) \leq \phi_k^*$ ... 3/3

f is L-smooth and μ-strongly convex $(A1)$

$\{\boldsymbol{y}_k\}_{k=0}^{\infty}$ is a sequence $(A3)$

IF $f(\boldsymbol{x}_k) \;\leq\; \phi_k^* \;:=\; \min_{\boldsymbol{x}\in\mathbb{R}^n} \phi_k(\boldsymbol{x})$ **THEN** $f(\boldsymbol{x}_k) - f^* \;\leq\; \lambda_k\Big(\phi_0(\boldsymbol{x}^*) - f^*\Big) \xrightarrow{\text{Def1}} 0.$ (Lemma 2.2.1)

$$\phi_{k+1}^* \;\geq\; f(\boldsymbol{y}_k) - \frac{\alpha_k^2}{2\gamma_{k+1}}\|\nabla f(\boldsymbol{y}_k)\|_2^2 + (1-\alpha_k)\Big\langle \nabla f(\boldsymbol{y}_k), \boldsymbol{x}_k - \boldsymbol{y}_k + \tfrac{\alpha_k\gamma_k}{\gamma_{k+1}}(\boldsymbol{v}_k - \boldsymbol{y}_k)\Big\rangle \qquad (\dagger)$$

$$\phi_{k+1}^* \;\geq\; f(\boldsymbol{x}_{k+1}) + (1-\alpha_k)\Big\langle \nabla f(\boldsymbol{y}_k), \boldsymbol{x}_k - \boldsymbol{y}_k + \tfrac{\alpha_k\gamma_k}{\gamma_{k+1}}(\boldsymbol{v}_k - \boldsymbol{y}_k)\Big\rangle \qquad (\dagger')$$

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

$$f(\boldsymbol{x}_{k+1}) \;\geq\; f(\boldsymbol{y}_k) - \frac{1}{2L}\|\nabla f(\boldsymbol{y}_k)\|_2^2 \qquad \text{(SDL)}$$

$$\gamma_{k+1} = (1-\alpha_k)\gamma_k + \alpha_k\mu \qquad (i)$$

---

1. We have defined $\boldsymbol{x}$ from $\boldsymbol{y}$
   - (GD) $\overset{A1}{\Longrightarrow}$ (SDL) so we have $(\dagger) \overset{(SDL)}{\Longrightarrow} (\dagger')$
   - (GD) is a way to define $\boldsymbol{x}$ from $\boldsymbol{y}$
2. We still haven't specify how we get $\boldsymbol{y}$ from $\boldsymbol{x}$
   - By (A3) $\boldsymbol{y}$ is free
3. Our goal is to make $(\dagger')$ satisfy , i.e., $\phi_{k+1}^* \geq f(\boldsymbol{x}_{k+1})$
   - We can consider $(1-\alpha_k)\Big\langle \nabla f(\boldsymbol{y}_k), \boldsymbol{x}_k - \boldsymbol{y}_k + \tfrac{\alpha_k\gamma_k}{\gamma_{k+1}}(\boldsymbol{v}_k - \boldsymbol{y}_k)\Big\rangle \;\geq\; 0$
   - The borderline case is $\boldsymbol{x}_k - \boldsymbol{y}_k + \frac{\alpha_k\gamma_k}{\gamma_{k+1}}(\boldsymbol{v}_k - \boldsymbol{y}_k) = \boldsymbol{0}$. Now using (i) we have

$$\boldsymbol{y}_k \overset{(i)}{=} \frac{\alpha_k\gamma_k\boldsymbol{v}_k + \gamma_{k+1}\boldsymbol{x}_k}{\gamma_k + \alpha_k\mu}$$

Algorithm: Nesterov's general scheme of optimal 1st-order method

Initialize $\boldsymbol{x}_0 \in \mathbb{R}^n$, $\gamma_0 > 0$, $\boldsymbol{v}_0 = \boldsymbol{x}_0$

for $k = 1, 2, \ldots$ **do**

(a) Compute $\alpha_k \in]\,0, 1\,[$ from the equation

$$L\alpha_k^2 \;=\; (1 - \alpha_k)\gamma_k + \alpha_k \mu$$

Update $\gamma_{k+1} = (1 - \alpha_k)\gamma_k + \alpha_k \mu$

(b) $\boldsymbol{y}_k = \dfrac{\alpha_k \gamma_k \boldsymbol{v}_k + \gamma_{k+1} \boldsymbol{x}_k}{\gamma_k + \alpha_k \mu}$

Compute $f(\boldsymbol{y}_k)$ and $\nabla f(\boldsymbol{y}_k)$

(c) Find $\boldsymbol{x}_{k+1}$ such that

$$f(\boldsymbol{x}_{k+1}) \;\leq\; f(\boldsymbol{y}_k) - \frac{1}{2L}\|\nabla f(\boldsymbol{y}_k)\|_2^2$$

(For example, a gradient descent step)

(d) $\boldsymbol{v}_k = \dfrac{(1 - \alpha_k)\gamma_k \boldsymbol{v}_k + \alpha_k \mu \boldsymbol{y}_k - \alpha_k \nabla f(\boldsymbol{y}_k)}{\gamma_{k+1}}$

Algorithm: Nesterov's general scheme of optimal 1st-order method (compact form)

Initialize $\boldsymbol{x}_0 \in \mathbb{R}^n$, $\gamma_0 > 0$, $\boldsymbol{v}_0 = \boldsymbol{x}_0$
for $k = 1, 2, \ldots$ **do**
  Compute $\alpha_k \in]\,0, 1\,[$ from $L\alpha_k^2 = (1-\alpha_k)\gamma_k + \alpha_k\mu$
  $\gamma_{k+1} = (1-\alpha_k)\gamma_k + \alpha_k\mu$
  $\boldsymbol{y}_k = \frac{\alpha_k\gamma_k\boldsymbol{v}_k + \gamma_{k+1}\boldsymbol{x}_k}{\gamma_k + \alpha_k\mu}$
  Find $\boldsymbol{x}_{k+1}$ such that $f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{y}_k) - \frac{1}{2L}\|\nabla f(\boldsymbol{y}_k)\|_2^2$. E.g., a gradient descent step
  $\boldsymbol{v}_k = \frac{(1-\alpha_k)\gamma_k\boldsymbol{v}_k + \alpha_k\mu\boldsymbol{y}_k - \alpha_k\nabla f(\boldsymbol{y}_k)}{\gamma_{k+1}}$

- **Theorem** For the sequence $\{\boldsymbol{x}_k\}_{k\in\mathbb{N}}$ generated by optimal 1st-order method,we have

$$f(\boldsymbol{x}_k) - f^* \;\leq\; \lambda_k\Big[f(\boldsymbol{x}_0) - f^* + \frac{\gamma_0}{2}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2\Big],$$

where $\lambda_0 = 1$ and $\lambda_k = \prod_{i=0}^{k-1}(1-\alpha_i)$.

- **Proof**
  - Let $\phi_0(\boldsymbol{x}) = f(\boldsymbol{x}_0) + \frac{\gamma_0}{2}\|\boldsymbol{x} - \boldsymbol{v}_0\|_2^2$, then $f(\boldsymbol{x}_0) = \phi_0^*$ satisfies $f(\boldsymbol{x}_k) \leq \phi_k^*$ at $k = 0$
  - The scheme satisfies $f(\boldsymbol{x}_k) \leq \phi_k^*$ at all k (this is exactly what we spent on 3 slides to show)
  - Then

$$\begin{aligned}
f(\boldsymbol{x}_k) - f^* &\leq \lambda_k\Big[\phi_0(\boldsymbol{x}^*) - f^*\Big] && \textbf{Lemma 2.2.1}\\
&\leq \lambda_k\Big[\Big(f(\boldsymbol{x}_0) + \frac{\gamma_0}{2}\|\boldsymbol{x} - \boldsymbol{v}_0\|_2^2\Big)\Big|_{\boldsymbol{x}=\boldsymbol{x}^*} - f^*\Big] && \phi_0(\boldsymbol{x}) = f(\boldsymbol{x}_0) + \frac{\gamma_0}{2}\|\boldsymbol{x} - \boldsymbol{v}_0\|_2^2\\
&\leq \lambda_k\Big[f(\boldsymbol{x}_0) - f^* + \frac{\gamma_0}{2}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2\Big] && \boldsymbol{v}_0 = \boldsymbol{x}_0 \quad \square
\end{aligned}$$

Convergence rate of method follows the convergence of $\{\lambda_k\}_{k\in\mathbb{N}}$

Algorithm: Nesterov's general scheme of optimal 1st-order method (compact form)

Initialize $\boldsymbol{x}_0 \in \mathbb{R}^n$, $\gamma_0 > 0$, $\boldsymbol{v}_0 = \boldsymbol{x}_0$
for $k = 1, 2, \ldots$ **do**
  Compute $\alpha_k \in]\,0, 1\,[$ from $L\alpha_k^2 = (1-\alpha_k)\gamma_k + \alpha_k\mu$
  $\gamma_{k+1} = (1-\alpha_k)\gamma_k + \alpha_k\mu$
  $\boldsymbol{y}_k = \frac{\alpha_k\gamma_k\boldsymbol{v}_k + \gamma_{k+1}\boldsymbol{x}_k}{\gamma_k + \alpha_k\mu}$
  Find $\boldsymbol{x}_{k+1}$ such that $f(\boldsymbol{x}_{k+1}) \le f(\boldsymbol{y}_k) - \frac{1}{2L}\|\nabla f(\boldsymbol{y}_k)\|_2^2$. E.g., a gradient descent step
  $\boldsymbol{v}_k = \frac{(1-\alpha_k)\gamma_k\boldsymbol{v}_k + \alpha_k\mu\boldsymbol{y}_k - \alpha_k\nabla f(\boldsymbol{y}_k)}{\gamma_{k+1}}$

- **Theorem** For $\lambda_0 = 1$, $\lambda_k = \prod_{i=0}^{k-1}(1-\alpha_i)$, we have

$$f(\boldsymbol{x}_k) - f^* \;\le\; \lambda_k\Big[f(\boldsymbol{x}_0) - f^* + \frac{\gamma_0}{2}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2\Big].$$

  - The method converge: $f(\boldsymbol{x}_k) \to f^*$ because $\lambda_k \to 0$
  - The convergence rate of method follows the convergence rate of $\{\lambda_k\}_{k\in\mathbb{N}}$

- **Theorem (Convergence rate of λ_k)** If $\gamma_0 \in]\,\mu, 3L+\mu\,]$, then

$$\lambda_k \;\le\; \frac{\mu}{(\gamma_0-\mu)\Big[\sinh\Big((k+1)\sqrt{\frac{\mu}{L}}\Big)\Big]^2} \quad \forall k \ge 0.$$

Proof of λ_k

$$\begin{array}{rll} \gamma_{k+1} & = (1-\alpha_k)\gamma_k + \alpha_k\mu & (i) \\ \lambda_{k+1} & = (1-\alpha_k)\lambda_k & (A6) \\ \lambda_0 & := 1 & (A7) \\ L\alpha_k^2 & = (1-\alpha_k)\gamma_k + \alpha_k\mu & \text{step 1 of algo} \end{array}$$

- First

$$\begin{array}{rcll} \gamma_{k+1} - \mu & = & (1-\alpha_k)\gamma_k + \alpha_k\mu - \mu & (i) \\ & = & (1-\alpha_k)(\gamma_k - \mu) & \\ & = & (1-\alpha_k)(1-\alpha_{k-1})(\gamma_{k-1} - \mu) & \text{calling } (i) \text{ on } k-1 \\ & = & \prod_{i=0}^{k}(1-\alpha_i)(\gamma_0 - \mu) & \text{calling } (i) \text{ multiple times} \\ & = & \lambda_{k+1}(\gamma_0 - \mu) & \text{by } (A6, A7) \end{array}$$

Hence

$$\gamma_{k+1} - \mu = \lambda_{k+1}(\gamma_0 - \mu) \quad \Longrightarrow \quad \gamma_{k+1} = \mu + \lambda_{k+1}(\gamma_0 - \mu) \qquad (\#)$$

- Next

$$L\alpha_k^2 \;=\; (1-\alpha_k)\gamma_k + \alpha_k\mu \overset{(i)}{=:} \gamma_{k+1} \iff L\alpha_k^2 - \gamma_{k+1} = 0 \implies \alpha_k = \frac{-0 \pm \sqrt{0^2 - 4(L)(-\gamma_{k+1})}}{2L} = \sqrt{\frac{\gamma_{k+1}}{L}}. \quad (\#\#)$$

Rearrange (A6)

$$1 - \frac{\lambda_{k+1}}{\lambda_k} \overset{(A6)}{=} \alpha_k \overset{(\#\#)}{=} \sqrt{\frac{\gamma_{k+1}}{L}} \overset{(\#)}{=} \sqrt{\frac{\mu + \lambda_{k+1}(\gamma_0 - \mu)}{L}} = \sqrt{\frac{\mu}{L} + \lambda_{k+1}\frac{\gamma_0 - \mu}{L}} \qquad (\#\#\#)$$

Several facts on $\{\lambda_k\}_k$

$$\begin{aligned} \alpha_k &\in]0,1[&& (A4a)\\ \lambda_{k+1} &= (1-\alpha_k)\lambda_k && (A6)\\ \lambda_0 &:= 1 && (A7)\\ 1-\frac{\lambda_{k+1}}{\lambda_k} &= \sqrt{\frac{\mu}{L}+\lambda_{k+1}\frac{\gamma_0-\mu}{L}} && (\#\#\#) \end{aligned}$$

- When we get (###) from (A6) by division of λ_k, we need to make sure $\lambda_k \neq 0$
- First $\lambda_{k+1} \overset{(A6)}{=} (1-\alpha_k)\lambda_k \overset{(A4a)}{<} \lambda_k$ (Strict Decrease)
- By (A4a), (A6), (A7) that $\lambda_k = \prod_{i=0}^{k}(1-\alpha_i) > 0$ so
  - $\{\lambda_k\}_{k\in\mathbb{N}}$ is a strictly positive sequence that is upper bounded by 1 ($\lambda_k > 0$)
  - we can take inverse of λ_k (inverse ok)
  - we can take square-root of λ_k (square-root ok)
- Let $q = \frac{\mu}{L}$. By (inverse ok) we can multiply (###) by $\frac{1}{\lambda_{k+1}}$ on both sides gives

$$\frac{1}{\lambda_{k+1}} - \frac{1}{\lambda_k} \overset{(\#\#\#)}{=} \frac{1}{\lambda_{k+1}}\sqrt{q+\lambda_{k+1}\frac{\gamma_0-\mu}{L}} \overset{\text{(square-root ok)}}{=} \frac{1}{\sqrt{\lambda_{k+1}}}\sqrt{q\frac{1}{\lambda_{k+1}}+\frac{\gamma_0-\mu}{L}} \qquad (\heartsuit)$$

- By (inverse ok) and (square-root ok) we can apply a not-so-trivial decomposition $a-b = (\sqrt{a}+\sqrt{b})(\sqrt{a}-\sqrt{b})$

$$\frac{1}{\lambda_{k+1}} - \frac{1}{\lambda_k} = \frac{1}{\sqrt{\lambda_{k+1}}^2} - \frac{1}{\sqrt{\lambda_k}^2} = \Big(\frac{1}{\sqrt{\lambda_{k+1}}}+\frac{1}{\sqrt{\lambda_k}}\Big)\Big(\frac{1}{\sqrt{\lambda_{k+1}}}-\frac{1}{\sqrt{\lambda_k}}\Big) \qquad (\heartsuit\heartsuit)$$

- By (inverse ok)

$$\lambda_{k+1} \overset{\text{(Strict Decrease)}}{<} \lambda_k \text{ AND } \{\lambda_k\}_k \overset{\lambda_k>0}{>} 0 \overset{\text{(inverse ok)}}{\Longrightarrow} \frac{1}{\lambda_k} < \frac{1}{\lambda_{k+1}} \overset{\lambda>0}{\Longrightarrow} \boxed{\frac{1}{\sqrt{\lambda_k}} < \frac{1}{\sqrt{\lambda_{k+1}}}}$$

Tricky step

$$
\begin{aligned}
\frac{1}{\lambda_{k+1}} - \frac{1}{\lambda_k} &= \frac{1}{\sqrt{\lambda_{k+1}}}\sqrt{q\frac{1}{\lambda_{k+1}} + \frac{\gamma_0 - \mu}{L}} && (\heartsuit)\\
\frac{1}{\lambda_{k+1}} - \frac{1}{\lambda_k} &= \Big(\frac{1}{\sqrt{\lambda_{k+1}}} + \frac{1}{\sqrt{\lambda_k}}\Big)\Big(\frac{1}{\sqrt{\lambda_{k+1}}} - \frac{1}{\sqrt{\lambda_k}}\Big) && (\heartsuit\heartsuit)\\
\frac{1}{\sqrt{\lambda_k}} &< \frac{1}{\sqrt{\lambda_{k+1}}} &&\\
\gamma_0 &\in \left]\, \mu, 3L + \mu \,\right[, \quad L > 0 &&
\end{aligned}
$$

- Now combine , $(\heartsuit)$ and $(\heartsuit\heartsuit)$

$$\frac{1}{\sqrt{\lambda_{k+1}}}\sqrt{q\frac{1}{\lambda_{k+1}} + \frac{\gamma_0 - \mu}{L}} \overset{(\heartsuit)}{=} \frac{1}{\lambda_{k+1}} - \frac{1}{\lambda_k} \overset{(\heartsuit\heartsuit)}{=} \Big(\frac{1}{\sqrt{\lambda_{k+1}}} + \frac{1}{\sqrt{\lambda_k}}\Big)\Big(\frac{1}{\sqrt{\lambda_{k+1}}} - \frac{1}{\sqrt{\lambda_k}}\Big) < \frac{2}{\sqrt{\lambda_{k+1}}}\Big(\frac{1}{\sqrt{\lambda_{k+1}}} - \frac{1}{\sqrt{\lambda_k}}\Big).$$

Now we see that why are we doing these: we can cancel out $\frac{1}{\sqrt{\lambda_{k+1}}}$ to simplify

$$\sqrt{q\frac{1}{\lambda_{k+1}} + \frac{\gamma_0 - \mu}{L}} < 2\Big(\frac{1}{\sqrt{\lambda_{k+1}}} - \frac{1}{\sqrt{\lambda_k}}\Big) \iff \frac{1}{\sqrt{\lambda_{k+1}}} - \frac{1}{\sqrt{\lambda_k}} > \frac{1}{2}\sqrt{q\frac{1}{\lambda_{k+1}} + \frac{\gamma_0 - \mu}{L}} \qquad (\maltese)$$

- By assumption $\gamma_0 \in]\mu, 3L + \mu]$ we have $\gamma_0 > \mu$ so multiplying $\sqrt{\frac{L}{\gamma_0 - \mu}}$ to $(\maltese)$ do not change the sign

$$\sqrt{\frac{1}{\lambda_{k+1}}\frac{L}{\gamma_0 - \mu}} - \sqrt{\frac{1}{\lambda_k}\frac{L}{\gamma_0 - \mu}} > \frac{1}{2}\sqrt{q\frac{1}{\lambda_{k+1}}\frac{L}{\gamma_0 - \mu} + 1}$$

- Denote $\xi_k := \sqrt{\frac{1}{\lambda_k}\frac{L}{\gamma_0 - \mu}}$, we have

$$\xi_{k+1} - \xi_k > \frac{1}{2}\sqrt{q\xi_{k+1}^2 + 1} \qquad (\square)$$

$$
\begin{aligned}
L &\geq \mu \geq 0 \\
\xi_k &:= \sqrt{\frac{L}{(\gamma_0 - \mu)\lambda_k}} \\
\lambda_0 &:= 1 \\
\gamma_0 &\in \,]\,\mu, 3L + \mu\,]
\end{aligned}
$$

- Let $\delta = \frac{1}{2}\sqrt{q} = \frac{1}{2}\sqrt{\frac{\mu}{L}}$, we are going to use mathematical induction to show

$$
\xi_k \;\geq\; \frac{e^{(k+1)\delta} - e^{-(k+1)\delta}}{4\delta} \quad \forall k \geq 0 \tag{$\triangle$}
$$

- **Base case ($k = 0$)**: we want to show $\sqrt{\frac{L}{\gamma_0 - \mu}} \geq \frac{e^{\delta} - e^{-\delta}}{4\delta}$

- First, by $\gamma_0 \in \,]\mu, 3L + \mu]$, we have

$$
0 < \gamma_0 - \mu \leq 3L \quad \overset{\text{strict positive, so division ok}}{\Longrightarrow} \quad \frac{1}{\gamma_0 - \mu} \geq \frac{1}{3L} > 0 \quad \overset{\text{strict positive, ok to } \sqrt{\cdot}}{\Longrightarrow} \quad \frac{1}{\sqrt{\gamma_0 - \mu}} \geq \frac{1}{\sqrt{3L}} \quad \Longrightarrow \quad \sqrt{\frac{L}{\gamma_0 - \mu}} \geq \frac{1}{\sqrt{3}}
$$

- By $L \geq \mu \geq 0$, we have

$$
0 \leq q := \frac{\mu}{L} \leq 1 \quad \Longrightarrow \quad 0 \leq \sqrt{q} \leq 1 \quad \Longrightarrow \quad 0 \leq \delta \leq \frac{1}{2}
$$

- A tricky step: the function $\frac{e^x - e^{-x}}{4x}$ for $0 \leq x \leq 0.5$ is bounded above by $\frac{e^x - e^{-x}}{4x}\Big|_{x=0.5}$. I.e., $\frac{e^{1/2} - e^{-1/2}}{2} \geq \frac{e^{\delta} - e^{-\delta}}{4\delta}$.

- $0.5773\ldots = \frac{1}{\sqrt{3}} > \frac{e^{1/2} - e^{-1/2}}{2} = 0.521095$

- Lastly , , , $\Longrightarrow$

$$
\begin{aligned}
\psi(t) &:= \frac{e^{(t+1)\delta} - e^{-(t+1)\delta}}{4\delta} && \text{thing we want to prove, with integer } k \text{ generalized to real } t \\
\psi'(t) &= \frac{e^{(t+1)\delta} + e^{-(t+1)\delta}}{4} && \text{the derivative of } \psi\text{, note that } \delta \text{ in the denominator is gone} \\
\psi'(t) &\to \infty \text{ so } \psi \text{ is a convex function} \\
\psi(a) &\geq \psi(b) + \psi'(b)(a-b) && \psi \text{ is convex} \\
\psi'(t+1) &:= \frac{e^{(t+2)\delta} + e^{-(t+2)\delta}}{4} \\
\psi(t) &\geq \psi(t+1) + \frac{e^{(t+2)\delta} + e^{-(t+2)\delta}}{4} && \psi(a) \geq \psi(b) + \psi'(b)(a-b) \text{ with } a = t, b = t+1 \\
&= \psi(t+1) + \frac{1}{2}\sqrt{\frac{\left(e^{(t+2)\delta} + e^{-(t+2)\delta}\right)^2}{4}} && \text{a WTF step} \\
&= \psi(t+1) + \frac{1}{2}\sqrt{\frac{e^{2(t+2)\delta} + 2 + e^{-2(t+2)\delta}}{4}} \\
&= \psi(t+1) + \frac{1}{2}\sqrt{\frac{e^{2(t+2)\delta} - 2 + e^{-2(t+2)\delta}}{4} + \frac{4}{4}} \\
&= \psi(t+1) + \frac{1}{2}\sqrt{4\delta^2\left(\frac{e^{(t+2)\delta} - e^{-(t+2)\delta}}{4\delta}\right)^2 + 1} && \text{another WTF step: everything inside square-root is positive} \\
&> \psi(t+1) - \frac{1}{2}\sqrt{4\delta^2\left(\frac{e^{(t+2)\delta} - e^{-(t+2)\delta}}{4\delta}\right)^2 + 1} && \text{because everything inside square-root is positive}
\end{aligned}
$$

After so many efforts, we now have

$$
\psi(t) \geq \psi(t+1) - \frac{1}{2}\sqrt{4\delta^2\left(\frac{e^{(t+2)\delta} - e^{-(t+2)\delta}}{4\delta}\right)^2 + 1} \qquad \text{(WTF)}
$$

The last step: proving case $k+1$ by contradiction

- Now we have

$$\begin{aligned}
\psi(t) &:= \frac{e^{(t+1)\delta} - e^{-(t+1)\delta}}{4\delta} && \text{integer } k \text{ generalized to real } t \\
\xi_{k+1} - \xi_k &> \frac{1}{2}\sqrt{q\xi_{k+1}^2 + 1} && (\square) \\
\psi(t) &\geq \psi(t+1) - \frac{1}{2}\sqrt{4\delta^2\left(\frac{e^{(t+2)\delta} - e^{-(t+2)\delta}}{4\delta}\right)^2 + 1} && (WTF) \\
\xi_k &\geq \frac{e^{(k+1)\delta} - e^{-(k+1)\delta}}{4\delta} && \text{induction hypothesis}
\end{aligned}$$

- What are going to do: use these to show that at case $k+1$

$$\xi_{k+1} \geq \frac{e^{(k+2)\delta} - e^{-(k+2)\delta}}{4\delta} \overset{t=k}{=} \psi(t+1)$$

by contradiction: $(\square)$ and induction hypothesis will contradicts with (WTF) ←this is why we derive (WTF)

- A preparatory material

$$\xi_{k+1} - \frac{1}{2}\sqrt{q\xi_{k+1}^2 + 1} \overset{(\square)}{>} \xi_k \overset{\text{induction hypothesis}}{\geq} \frac{e^{(k+1)\delta} - e^{-(k+1)\delta}}{4\delta} = \psi(t)$$

which is

$$\psi(t) < \xi_{k+1} - \frac{1}{2}\sqrt{q\xi_{k+1}^2 + 1}$$

- For the purpose of contradiction, assume $\xi_{k+1} < \psi(t+1)$, now we have

$$\psi(t) < \psi(t+1) - \frac{1}{2}\sqrt{q\xi_{k+1}^2 + 1} \qquad (X)$$

Proof $\xi_{k+1} \geq \frac{e^{(k+2)\delta} - e^{-(k+2)\delta}}{4\delta} \overset{t=k}{=} \psi(t+1)$

$$
\begin{array}{rcll}
\psi(t) & < & \psi(t+1) - \frac{1}{2}\sqrt{q\xi_{k+1}^2 + 1} & (X) \\
\psi(t) & \geq & \psi(t+1) - \frac{1}{2}\sqrt{4\delta^2\left(\frac{e^{(t+2)\delta} - e^{-(t+2)\delta}}{4\delta}\right)^2 + 1} & (WTF) \\
\delta & := & \frac{1}{2}\sqrt{q} & \\
\psi(t) & := & \frac{e^{(t+1)\delta} - e^{-(t+1)\delta}}{4\delta} &
\end{array}
$$

$$
\psi(t+1) - \frac{1}{2}\sqrt{4\delta^2\left(\frac{e^{(t+2)\delta} - e^{-(t+2)\delta}}{4\delta}\right)^2 + 1} \overset{(WTF)}{\leq} \psi(t) \overset{(X)}{<} \psi(t+1) - \frac{1}{2}\sqrt{q\xi_{k+1}^2 + 1}
$$

$$
-\frac{1}{2}\sqrt{4\delta^2\left(\frac{e^{(t+2)\delta} - e^{-(t+2)\delta}}{4\delta}\right)^2 + 1} < -\frac{1}{2}\sqrt{q\xi_{k+1}^2 + 1}
$$

$$
\sqrt{4\delta^2\left(\frac{e^{(t+2)\delta} - e^{-(t+2)\delta}}{4\delta}\right)^2 + 1} > \sqrt{q\xi_{k+1}^2 + 1}
$$

$$
4\delta^2\left(\frac{e^{(t+2)\delta} - e^{-(t+2)\delta}}{4\delta}\right)^2 > q\xi_{k+1}^2
$$

$$
2\delta\left(\frac{e^{(t+2)\delta} - e^{-(t+2)\delta}}{4\delta}\right) > \sqrt{q}\xi_{k+1}
$$

$$
\frac{e^{(t+2)\delta} - e^{-(t+2)\delta}}{4\delta} > \xi_{k+1}
$$

$$
\underbrace{\psi(t+1) > \xi_{k+1}}
$$

contradicts with $\xi_{k+1} < \psi(t+1)$

What now: we just proved $\xi_k := \sqrt{\frac{L}{(\gamma_0-\mu)\lambda_k}} \geq \frac{e^{(k+1)\delta}-e^{-(k+1)\delta}}{4\delta} \quad \forall k$

$$\begin{aligned}
\lambda_k &\leq \frac{16\delta^2 L}{(\gamma_0-\mu)\left[e^{(k+1)\delta}-e^{-(k+1)\delta}\right]^2} \quad \forall k && \text{rearrange}\\
&= \frac{4\mu}{(\gamma_0-\mu)\left[\exp\left((k+1)\sqrt{\frac{\mu}{L}}\right)-\exp\left(-(k+1)\sqrt{\frac{\mu}{L}}\right)\right]^2} \quad \forall k \quad \delta := \frac{1}{2}\sqrt{q}, q := \frac{\mu}{L}\\
&= \frac{\mu}{(\gamma_0-\mu)}\frac{1}{\left[\sinh\left((k+1)\sqrt{\frac{\mu}{L}}\right)\right]^2} \quad \forall k
\end{aligned}$$

Thus

$$\lambda_k \leq \frac{\mu}{(\gamma_0-\mu)}\frac{1}{\left[\sinh\left((k+1)\sqrt{\frac{\mu}{L}}\right)\right]^2} \quad \forall k \quad \gamma_0 \in]\mu, 3L+\mu]$$

The smallest γ_0 is $(1+\epsilon)\mu$ and thus

$$\lambda_k \leq \frac{\mu}{\epsilon}\frac{1}{\left[\sinh\left((k+1)\sqrt{\frac{\mu}{L}}\right)\right]^2} \quad \forall k$$

The largest γ_0 is $3L+\mu$ and thus

$$\lambda_k \leq \frac{\mu}{3L}\frac{1}{\left[\sinh\left((k+1)\sqrt{\frac{\mu}{L}}\right)\right]^2} \quad \forall k$$

$$\lambda_k \;\leq\; \frac{\mu}{(\gamma_0-\mu)}\frac{1}{\left[\sinh\left((k+1)\sqrt{\frac{\mu}{L}}\right)\right]^2} \;=\; \frac{4\mu}{(\gamma_0-\mu)\left[\exp\left((k+1)\sqrt{\frac{\mu}{L}}\right)-\exp\left(-(k+1)\sqrt{\frac{\mu}{L}}\right)\right]^2}, \quad \gamma_0 \in]\mu, 3L+\mu]$$

Consider Taylor series of $\exp(\theta)$

$$\begin{aligned}
\exp(\theta) &= \frac{\theta^0}{0!}+\frac{\theta^1}{1!}+\frac{\theta^2}{2!}+\frac{\theta^3}{3!}+\cdots \\
\exp\left((k+1)\sqrt{\tfrac{\mu}{L}}\right) &= 1+(k+1)\sqrt{\tfrac{\mu}{L}}+\frac{(k+1)^2\frac{\mu}{L}}{2}+\frac{(k+1)^3\left(\frac{\mu}{L}\right)^{\frac{3}{2}}}{6}+\cdots \\
\exp\left(-(k+1)\sqrt{\tfrac{\mu}{L}}\right) &= 1-(k+1)\sqrt{\tfrac{\mu}{L}}+\frac{(k+1)^2\frac{\mu}{L}}{2}-\frac{(k+1)^3\left(\frac{\mu}{L}\right)^{\frac{3}{2}}}{6}+\cdots \\
-\exp\left(-(k+1)\sqrt{\tfrac{\mu}{L}}\right) &= -1+(k+1)\sqrt{\tfrac{\mu}{L}}-\frac{(k+1)^2\frac{\mu}{L}}{2}+\frac{(k+1)^3\left(\frac{\mu}{L}\right)^{\frac{3}{2}}}{6}+\cdots \\
\exp\left(-(k+1)\sqrt{\tfrac{\mu}{L}}\right)-\exp\left(-(k+1)\sqrt{\tfrac{\mu}{L}}\right) &= 2\left((k+1)\sqrt{\tfrac{\mu}{L}}+\frac{(k+1)^3\left(\frac{\mu}{L}\right)^{\frac{3}{2}}}{6}+\frac{(k+1)^5\left(\frac{\mu}{L}\right)^{\frac{5}{2}}}{120}+\cdots\right) \\
\left[\exp\left(-(k+1)\sqrt{\tfrac{\mu}{L}}\right)-\exp\left(-(k+1)\sqrt{\tfrac{\mu}{L}}\right)\right]^2 &= 4\left[(k+1)\sqrt{\frac{\mu}{L}}+\underbrace{\frac{(k+1)^3\left(\frac{\mu}{L}\right)^{\frac{3}{2}}}{6}+\frac{(k+1)^5\left(\frac{\mu}{L}\right)^{\frac{5}{2}}}{120}+\cdots}_{>0}\right]^2 \\
&> 4\left[(k+1)\sqrt{\frac{\mu}{L}}\right]^2 \\
&= 4(k+1)^2\frac{\mu}{L}
\end{aligned}$$

$$\lambda_k \;\leq\; \frac{\mu}{(\gamma_0-\mu)}\frac{1}{\Big[\sinh\Big((k+1)\sqrt{\frac{\mu}{L}}\Big)\Big]^2} \;=\; \frac{4\mu}{(\gamma_0-\mu)\Big[\exp\Big((k+1)\sqrt{\frac{\mu}{L}}\Big)-\exp\Big(-(k+1)\sqrt{\frac{\mu}{L}}\Big)\Big]^2}, \quad \gamma_0 \in \,]\mu, 3L+\mu]$$

$$\begin{aligned}
\left[\exp\Big(-(k+1)\sqrt{\tfrac{\mu}{L}}\Big)-\exp\Big(-(k+1)\sqrt{\tfrac{\mu}{L}}\Big)\right]^2 &> 4(k+1)^2\frac{\mu}{L}\\
\frac{1}{\left[\exp\Big(-(k+1)\sqrt{\tfrac{\mu}{L}}\Big)-\exp\Big(-(k+1)\sqrt{\tfrac{\mu}{L}}\Big)\right]^2} &< \frac{1}{4(k+1)^2\frac{\mu}{L}}
\end{aligned}$$

Theorem (Convergence rate) IF $\lambda_0 = 1$, $\lambda_k = \prod_{i=0}^{k-1}(1-\alpha_i)$, $\gamma_0 \in \,]\,\mu, 3L+\mu\,]$, then for the sequence $\{\boldsymbol{x}_k\}_{k\in\mathbb{N}}$ generated by Nesterov's general scheme, we have that for the sequence $\{f(\boldsymbol{x}_k)\}_{k\in\mathbb{N}}$

$$f(\boldsymbol{x}_k)-f^* \;\leq\; \frac{L}{4(\gamma_0-\mu)(k+1)^2}\Big[f(\boldsymbol{x}_0)-f^*+\frac{\gamma_0}{2}\|\boldsymbol{x}_0-\boldsymbol{x}^*\|_2^2\Big]$$

Proof

$$\begin{aligned}
f(\boldsymbol{x}_k)-f^* &\leq \lambda_k\Big[f(\boldsymbol{x}_0)-f^*+\frac{\gamma_0}{2}\|\boldsymbol{x}_0-\boldsymbol{x}^*\|_2^2\Big]\\
&\leq \frac{\mu}{(\gamma_0-\mu)\Big[\sinh\Big((k+1)\sqrt{\frac{\mu}{L}}\Big)\Big]^2}\Big[f(\boldsymbol{x}_0)-f^*+\frac{\gamma_0}{2}\|\boldsymbol{x}_0-\boldsymbol{x}^*\|_2^2\Big]\\
&< \frac{L}{4(\gamma_0-\mu)(k+1)^2}\Big[f(\boldsymbol{x}_0)-f^*+\frac{\gamma_0}{2}\|\boldsymbol{x}_0-\boldsymbol{x}^*\|_2^2\Big]
\end{aligned}$$