Nesterov's estimate sequence: 1. What is it and how to construct one

Andersen Ang

ECS, Uni. Southampton, UK

andersen.ang@soton.ac.uk

Homepage angms.science

Version: July 28, 2023

First draft: Nov 21, 2021

Content

Nesterov's estimate sequence: $\Big\{\phi_k(\boldsymbol{x}), \lambda_k\Big\}_{k=0}^{\infty}$, $\lambda_k \geq 0$ that

$$\lambda_k \xrightarrow{k\to\infty} 0, \qquad \phi_k(\boldsymbol{x}) \ \leq \ (1-\lambda_k) f(\boldsymbol{x}) + \lambda_k \phi_0(\boldsymbol{x})$$

Why estimate sequence: $f(\boldsymbol{x}_k) - f^* \ \leq \ \lambda_k\Big(\phi_0(\boldsymbol{x}^*) - f^*\Big) \ \xrightarrow{k\to\infty} \ 0.$

How to construct an estimate sequence for str-cvx smooth f

Reference

Yurii Nesterov, Introductory lectures on convex optimization: a basic course, Kluwer Academic Publishers, 2003.

Yurii Nesterov, Lectures on convex optimization. Vol. 137. Berlin: Springer, 2018.

Problem setup: unconstrained convex smooth optimization

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

- $f : \mathbb{R}^n \to \mathbb{R}$ 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 ($\mu = 0$)
  - 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.
  - $\mathcal{X}^* := \underset{\boldsymbol{x}}{\operatorname{argmin}} \ f(\boldsymbol{x})$ — solution set, assumed nonempty
  - $\boldsymbol{x}^* \in \mathcal{X}^*$ — minimizer
  - $f^* := f(\boldsymbol{x}^*)$ — optimal function value

Nesterov's estimate sequence: the definition

- Also called Nesterov's estimating sequence[1]
- **Definition 2.2.1** A sequences pair $\Big\{\phi_k(\boldsymbol{x}), \lambda_k\Big\}_{k=0}^{\infty}$ 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)	$\phi_k(\boldsymbol{x})$	$\leq$	$(1-\lambda_k)f(\boldsymbol{x}) + \lambda_k\phi_0(\boldsymbol{x})$	$(\forall k)\big(\forall \boldsymbol{x} \in \mathbb{R}^n\big)$	$\{\phi_k\}_{k\in\mathbb{N}} \leq$ "convex combination" of f, ϕ_0

- At this stage
  - We haven't specify what is λ_0
    - If $\lambda_0 > 1$ then Def2 is not convex combination but linear combination. That's why we put quote "convex combination"
  - We haven't specify how we get λ_k
  - We haven't specify what is ϕ_0
  - We haven't specify what property ϕ_k has
- At this stage, from Definition 2.2.1, we only know $\{\lambda_k\}_{k\in\mathbb{N}}$ converges to 0. But we don't know how it converges to 0, we also don't know is $\{\lambda_k\}_{k\in\mathbb{N}}$ monotonically converges to 0.
  - For example, the following oscillating sequence fulfills Def0 and Def1

$$\frac{\sin x + 1}{x + 0.1}, x \geq 0 \;:\; \Big\{1.6, 0.9, 0.3, 0.05, 0.008, 0.11, 0.23, ... \text{ for } x = \{1, 2, 3, 4, ...\}\Big\}$$

[1] Nesterov used the term "estimate sequence" in his 2003 book and then used "estimating sequence" in his 2018 book.

Nesterov's estimate sequence: the λ_k

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

$$
\begin{array}{llcll}
\text{(Def0)} & \lambda_k & \geq & 0 & (\forall k) \\
\text{(Def1)} & \lambda_k & \xrightarrow{k \to \infty} & 0 & (\forall k) \\
\text{(Def2)} & \phi_k(\boldsymbol{x}) & \leq & (1-\lambda_k) f(\boldsymbol{x}) + \lambda_k \phi_0(\boldsymbol{x}) & (\forall k)\big(\forall \boldsymbol{x} \in \mathbb{R}^n\big)
\end{array}
$$

- **Lemma 2.2.2 (Partly)** Assume that

$$
\begin{array}{llll}
\text{(L2.2.2 A4a)} & \alpha_k \in \;]\,0, 1\,[& (\forall k) & \alpha_k \text{ strictly positive and strictly smaller than 1} \\
\text{(L2.2.2 A4b)} & \displaystyle\sum_{k=0}^{\infty} \alpha_k = +\infty & & \{\alpha_k\} \text{ is not a summable sequence} \\
\text{(L2.2.2 A5)} & \lambda_0 := 1 & & \text{we initialize } \lambda_0 \\
\text{(L2.2.2 A6)} & \lambda_{k+1} = (1-\alpha_k)\lambda_k & (\forall k) & \text{define how we update } \lambda_k
\end{array}
$$

- With **Lemma 2.2.2 (Partly)**, now
  - $\{\lambda_k\}_{k \in \mathbb{N}}$ is *monotonically decreasing*:

$$
\lambda_{k+1} \overset{L2.2.2A6}{=} (1-\alpha_k)\lambda_k \overset{L2.2.2A4a}{<} \lambda_k \overset{L2.2.2A6}{=} (1-\alpha_{k-1})\lambda_{k-1} \overset{L2.2.2A4a}{<} \lambda_{k-1} < \ldots < \lambda_0 := 1 \qquad (\#)
$$

Reading (#) from right to left also means that Def 0 is satisfied, i.e., all $\lambda_k \geq 0$

(L2.2.2 A4) to (L2.2.2 A6) imply (Def2) $\lambda_{k+1} \to 0$ is satisfied

Definition 2.2.1 Lemma 2.2.2 (Partly) Assume that

(L2.2.2 A4a)	$\alpha_k \in \,]\,0,1\,[$	$(\forall k)$	α_k strictly positive and strictly smaller than 1
(L2.2.2 A4b)	$\sum_{k=0}^{\infty} \alpha_k = +\infty$		$\{\alpha_k\}$ is not a summable sequence
(L2.2.2 A5)	$\lambda_0 := 1$		we initialize λ_0
(L2.2.2 A6)	$\lambda_{k+1} = (1-\alpha_k)\lambda_k$	$(\forall k)$	define how we update λ_k

- With **Lemma 2.2.2 (Partly)**,

$$\lambda_{k+1} \overset{L2.2.2A6}{=} (1-\alpha_k)\lambda_k \overset{L2.2.2A6}{=} (1-\alpha_k)(1-\alpha_{k-1})\lambda_{k-1} \overset{L2.2.2A6}{=} \dots \overset{L2.2.2A6}{=} \prod_{i=1}^{k}(1-\alpha_i)\lambda_0 \overset{L2.2.2A5}{=} \prod_{i=1}^{k}(1-\alpha_i) \quad (!)$$

- Now we show that L2.2.2 A4 implies $\prod_{k=1}^{\infty}(1-\alpha_k) = 0$.

  Notice that L2.2.2 A4b is a sum but what we want to prove is produce, this gives the hint that we should take log.
  Let $S = \prod_{k=1}^{\infty}(1-\alpha_k) = 0$, now consider

$$\begin{aligned} \log S &= \sum_{k=1}^{\infty} \log(1-\alpha_k) \leq -\sum_{k=1}^{\infty} \alpha_k && \log(1-x) \text{ is concave so it is under its 1st-order Taylor expansion} \\ &= -\infty && L2.2.2A4b \\ \iff \quad S &= e^{-\infty} = 0 \end{aligned}$$

  Therefore, by (!), we have $\lambda_\infty = S = 0$, i.e., $\lambda_k \xrightarrow{k\to+\infty} 0$.

Why study Nesterov's estimate sequence?

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

$$\begin{array}{lcccl} \text{(Def0)} & \lambda_k & \geq & 0 & (\forall k) \\ \text{(Def1)} & \lambda_k & \xrightarrow{k\to\infty} & 0 & (\forall k) \\ \text{(Def2)} & \phi_k(\boldsymbol{x}) & \leq & (1-\lambda_k)f(\boldsymbol{x}) + \lambda_k \phi_0(\boldsymbol{x}) & (\forall k)\left(\forall \boldsymbol{x} \in \mathbb{R}^n\right) \end{array}$$

- **Lemma 2.2.1** IF for a sequence $\{\boldsymbol{x}_k\}_{k\in\mathbb{N}}$ we have

$$f(\boldsymbol{x}_k) \leq \phi_k^* := \min_{\boldsymbol{x}\in\mathbb{R}^n} \phi_k(\boldsymbol{x}), \tag{2.2.3}$$

THEN

$$\underset{\sim\sim\sim}{f(\boldsymbol{x}_k) - f^*} \leq \lambda_k \underbrace{\left(\phi_0(\boldsymbol{x}^*) - f^*\right)}_{\text{a constant}} \xrightarrow{\text{Def1}} 0. \tag{3}$$

  - It forms a global upper bound the of the cost optimality gap $f(\boldsymbol{x}_k) - f^*$
  - This upper bound converges to 0 by Def1. (Note $\phi_0(\boldsymbol{x}^*) - f^*$ is a constant.)

  $\Longrightarrow$ the convergence rate of $\left\{f(\boldsymbol{x}_k) - f^*\right\}_{k\in\mathbb{N}}$ follows that of $\left\{\lambda_k\right\}_{k\in\mathbb{N}}$ the reason why we study estimate sequence

Proof

$$f(\boldsymbol{x}_k) \overset{(2.2.3)}{\leq} \phi_k^* \overset{(2.2.3)}{:=} \min_{\boldsymbol{x}\in\mathbb{R}^n} \phi_k(\boldsymbol{x}) \overset{\text{(Def2)}}{\leq} \min_{\boldsymbol{x}\in\mathbb{R}^n} (1-\lambda_k)f(\boldsymbol{x}) + \lambda_k\phi_0(\boldsymbol{x}) \leq (1-\lambda_k)f(\boldsymbol{x}^*) + \lambda_k\phi_0(\boldsymbol{x}^*)$$

$$\iff f(\boldsymbol{x}_k) - f^* \leq \lambda_k\left(\phi_0(\boldsymbol{x}^*) - f^*\right) \xrightarrow{\text{(Def1)}} 0.$$

□

Nesterov's estimate sequence

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

$$
\begin{array}{lccll}
\text{(Def0)} & \lambda_k & \geq & 0 & (\forall k) \\
\text{(Def1)} & \lambda_k & \xrightarrow{k\to\infty} & 0 & (\forall k) \\
\text{(Def2)} & \phi_k(\boldsymbol{x}) & \leq & (1-\lambda_k)f(\boldsymbol{x}) + \lambda_k\phi_0(\boldsymbol{x}) & (\forall k)\big(\forall \boldsymbol{x} \in \mathbb{R}^n\big)
\end{array}
$$

- **Lemma 2.2.1** IF for a sequence $\{\boldsymbol{x}_k\}_{k\in\mathbb{N}}$ we have

$$f(\boldsymbol{x}_k) \;\leq\; \phi_k^* \;:=\; \min_{\boldsymbol{x}\in\mathbb{R}^n} \phi_k(\boldsymbol{x}), \tag{2.2.3}$$

  THEN

$$f(\boldsymbol{x}_k) - f^* \;\leq\; \lambda_k\Big(\phi_0(\boldsymbol{x}^*) - f^*\Big) \xrightarrow{\text{Def1}} 0. \tag{3}$$

- Now we know estimate sequence is useful to derive convergence rate
- The next questions is: how to construct an estimate sequence?
  - how should we pick ϕ_0?
  - how should we update λ_k and ϕ_k?

A way to construct an estimate sequence for smooth convex f

- **Lemma 2.2.2 IF**

(L2.2.2 A1)	f is L-smooth μ-strongly convex		possibly $\mu = 0$ (convex not strongly convex)
(L2.2.2 A2)	$\phi_0(\cdot)$ is a convex function on $\mathbb{R}^n$		any arbitrary convex function
(L2.2.2 A3)	$\{\boldsymbol{y}_k\}_{k=0}^{\infty}$ is a sequence in $\mathbb{R}^n$		any arbitrary sequence
(L2.2.2 A4a)	$\alpha_k \in]0,1[$	$(\forall k)$	α_k strictly positive and strictly smaller than 1
(L2.2.2 A4b)	$\displaystyle\sum_{k=0}^{\infty} \alpha_k = \infty$		$\{\alpha_k\}$ is not a summable sequence
(L2.2.2 A5)	$\lambda_0 := 1$		we initialize λ_0

THEN the sequence-pair $\{\phi_k(\boldsymbol{x}), \lambda_k\}_{k=0}^{\infty}$ defined as

(L2.2.2 A6)	$\lambda_{k+1} = (1-\alpha_k)\lambda_k$	$(\forall k)$	how we update λ_k
(L2.2.2 A7)	$\phi_{k+1}(\boldsymbol{x}) = (1-\alpha_k)\phi_k(\boldsymbol{x}) + \alpha_k\Big(f(\boldsymbol{y}_k) + \langle \nabla f(\boldsymbol{y}_k), \boldsymbol{x}-\boldsymbol{y}_k\rangle + \frac{\mu}{2}\|\boldsymbol{x}-\boldsymbol{y}_k\|_2^2\Big)$	$(\forall k)$	how we update ϕ_k

is an estimate sequence of $f(\boldsymbol{x})$.

- To prove $\{\phi_k(\boldsymbol{x}), \lambda_k\}_{k\in\mathbb{N}}$ is an estimate sequence of $f(\boldsymbol{x})$, we need to show

  $P0$ $\{\lambda_k\}_{k\in\mathbb{N}}$ defined in this way is nonnegative
  $P1$ $\{\lambda_k\}_{k\in\mathbb{N}}$ defined in this way converges to 0
  $P2$ $\{\phi_k(\boldsymbol{x})\}_{k\in\mathbb{N}}$ defined in this way satisfies $\phi_k \leq \phi_k(\boldsymbol{x}) \leq (1-\lambda_k)f(\boldsymbol{x}) + \lambda_k\phi_0(\boldsymbol{x})$ $\forall k$

- Showing $P0$ is simple: by (!) we have $\lambda_{k+1} = \prod_{i=1}^{k}(1-\alpha_i) > 0$.
  Now we have $P0 : \lambda_k \geq 0$.

Proof $P1$: showing $\lambda_k \to 0$

- **Proposition** By definition $\lambda_{k+1} = (1-\alpha_k)\lambda_k$ with assumption $\alpha_k \in \,]0,1[$, the sequence $\{\lambda_k\}_{k\in\mathbb{N}}$ is monotonically decreasing.
  Proof by ratio test

$$\lambda_{k+1} = (1-\alpha_k)\lambda_k \iff \frac{\lambda_{k+1}}{\lambda_k} = 1-\alpha_k$$

$$\iff \frac{\lambda_{k+1}}{\lambda_k} \overset{\alpha_k \in \,]0,1[}{<} 1$$

- By $P0 : \lambda_k \geq 0$, the sequence $\{\lambda_k\}_{k\in\mathbb{N}}$ is bounded below by 0.
- **Theorem (Real analysis 101)**
  - Monotonic decreasing AND bounded below $\implies$ $\{\lambda_k\}_{k\in\mathbb{N}}$ has a limit c
- What we need to do: show $c = 0$.

There are three ways to show $\lambda_k \to 0$

- Way 1: By **Monotone convergence theorem (Real analysis 101)**, $c = \inf\{\lambda_k\}_{k\in\mathbb{N}} = 0$.

- Way 2: By contradiction.
  - By $\lambda_k \geq 0$, suppose the sequence $\{\lambda_k\}_{k\in\mathbb{N}}$ converges to a positive number $c > 0$.
  - Now consider $\lambda_k - \lambda_{k+1} = \lambda_k - (1-\alpha_k)\lambda_k = \alpha_k\lambda_k$. It forms a telescoping sum, sum it from 0 to k gives

$$\lambda_0 - \lambda_{k+1} = \sum_{i=0}^{k} \alpha_i\lambda_i \geq \sum_{i=0}^{k} \alpha_i c = c\sum_{i=0}^{k} \alpha_i \qquad (*)$$

    where the $\geq$ is based on the fact that we assume $\{\lambda_k\}_{k\in\mathbb{N}}$ converges (from above: all $\lambda_k \geq c$ for all k) to c.
  - Now $\lambda_0 - \lambda_{k+1} \overset{(*)}{\geq} c\sum_{i=0}^{k} \alpha_i$. Take limit $k \to \infty$ gives $\lambda_0 - c \geq c\sum_{i=0}^{\infty} \alpha_i$. By $\sum_{i=0}^{\infty} \alpha_i = \infty$ so $\lambda_0 - c \geq +\infty$, which is impossible (because $\lambda_0 := 1$), a contradiction, therefore $c = 0$.

- Way 3: By $\lambda_\infty = S = 0$ using L2.2.2 A4, A5, A6 and property of $\log(1-x)$ as what we did previously.

Proof part 2: on ϕ_k by induction

- Base case $k=0$: $\phi_0(\boldsymbol{x}) \le (1-\lambda_0)f(\boldsymbol{x}) + \lambda_0\phi_0(\boldsymbol{x}) = \phi_0(\boldsymbol{x})$ by $\lambda_0 := 1$.
- Induction hypothesis $\phi_k(\boldsymbol{x}) \le (1-\lambda_k)f(\boldsymbol{x}) + \lambda_k\phi_0(\boldsymbol{x})$
- Case $k+1$

$$
\begin{aligned}
\phi_{k+1}(\boldsymbol{x}) &= (1-\alpha_k)\phi_k(\boldsymbol{x}) + \alpha_k\Big(f(\boldsymbol{y}_k) + \langle\nabla f(\boldsymbol{y}_k), \boldsymbol{x}-\boldsymbol{y}_k\rangle + \frac{\mu}{2}\|\boldsymbol{x}-\boldsymbol{y}_k\|_2^2\Big) && \text{by A7 def of } \phi_{k+1} \\
&\le (1-\alpha_k)\phi_k(\boldsymbol{x}) + \alpha_k f(\boldsymbol{x}) && \text{A1: } f\ \mu\text{-str cvx, } \mu \ge 0 \\
&\overset{\text{tricky}}{=} (1-\alpha_k)\Big(\phi_k(\boldsymbol{x}) + \underline{(1-\lambda_k)f(\boldsymbol{x})} - (1-\lambda_k)f(\boldsymbol{x})\Big) + \alpha_k f(\boldsymbol{x}) \\
&= (1-\alpha_k)\Big(\underbrace{\phi_k(\boldsymbol{x}) - (1-\lambda_k)f(\boldsymbol{x})}_{\le \lambda_k\phi_0(\boldsymbol{x})}\Big) + \underbrace{(1-\alpha_k)\underline{(1-\lambda_k)f(\boldsymbol{x})} + \alpha_k f(\boldsymbol{x})}_{=\big((1-\alpha)-(1-\alpha)\lambda\big)f} \\
&\le (1-\alpha_k)\lambda_k\phi_0(\boldsymbol{x}) + \Big(1-(1-\alpha_k)\lambda_k\Big)f(\boldsymbol{x}) && \text{case } k\ \&\ \alpha_k \overset{A4a}{<} 1 \\
&= \lambda_{k+1}\phi_0(\boldsymbol{x}) + (1-\lambda_{k+1})f(\boldsymbol{x}). && \lambda_{k+1} \overset{A6}{=} (1-\alpha_k)\lambda_k
\end{aligned}
$$

So case $k+1$ is true. By induction, the proof ϕ_k is completed. □

The framework carries over to convex but not strongly convex f

- In the proof

$$\begin{aligned} \phi_{k+1}(\boldsymbol{x}) &= (1-\alpha_k)\phi_k(\boldsymbol{x}) + \alpha_k\Big(f(\boldsymbol{y}_k) + \langle \nabla f(\boldsymbol{y}_k), \boldsymbol{x}-\boldsymbol{y}_k\rangle + \frac{\mu}{2}\|\boldsymbol{x}-\boldsymbol{y}_k\|_2^2\Big) && \text{by A7 def of } \phi_{k+1} \\ &\le (1-\alpha_k)\phi_k(\boldsymbol{x}) + \alpha_k f(\boldsymbol{x}) && \text{A1: } f\ \mu\text{-str cvx, } \mu \ge 0 \end{aligned}$$

  The argument holds if f convex but not strongly convex

- In fact the whole framework assume $\mu \geq 0$, which includes the case $\mu = 0 \iff$ f is convex but not strongly convex

- When f is convex but not strongly convex, we construct ϕ_{k+1} as L2.2.2 A7 with $\mu = 0$, i.e.,

$$\begin{aligned} \phi_{k+1}(\boldsymbol{x}) &= (1-\alpha_k)\phi_k(\boldsymbol{x}) + \alpha_k\Big(f(\boldsymbol{y}_k) + \langle \nabla f(\boldsymbol{y}_k), \boldsymbol{x}-\boldsymbol{y}_k\rangle + \tfrac{\mu}{2}\|\boldsymbol{x}-\boldsymbol{y}_k\|_2^2\Big) && L2.2.2A7 \\ &= (1-\alpha_k)\phi_k(\boldsymbol{x}) + \alpha_k\Big(f(\boldsymbol{y}_k) + \langle \nabla f(\boldsymbol{y}_k), \boldsymbol{x}-\boldsymbol{y}_k\rangle\Big) && \mu = 0 \end{aligned}$$

The framework carries over to nondifferentiable convex f

- Note that in the whole proof we never explicitly make use of the assumption that f is L-smooth
- The only place we make use of f is differentiable is where we assume $\nabla f(\boldsymbol{x})$ exists at $\boldsymbol{y}_k$
- In the proof

$$
\begin{aligned}
\phi_{k+1}(\boldsymbol{x}) &= (1-\alpha_k)\phi_k(\boldsymbol{x}) + \alpha_k\Big(f(\boldsymbol{y}_k) + \langle \nabla f(\boldsymbol{y}_k), \boldsymbol{x}-\boldsymbol{y}_k\rangle + \frac{\mu}{2}\|\boldsymbol{x}-\boldsymbol{y}_k\|_2^2\Big) && \text{by A7 def of } \phi_{k+1} \\
&\leq (1-\alpha_k)\phi_k(\boldsymbol{x}) + \alpha_k f(\boldsymbol{x}) && \text{A1: } f \ \mu\text{-str cvx, } \mu \geq 0
\end{aligned}
$$

  The argument holds if f is convex but not differentiable

- When f is convex but not differentiable, we replace ∇f by subdifferential / subgradient

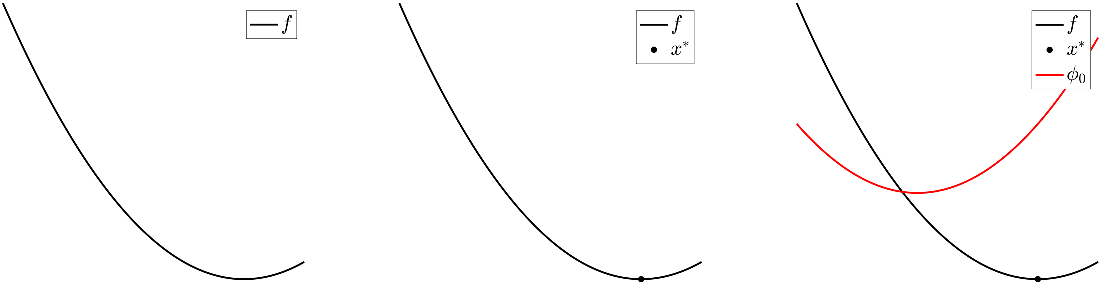
$$
\phi_{k+1}(\boldsymbol{x}) = (1-\alpha_k)\phi_k(\boldsymbol{x}) + \alpha_k\Big(f(\boldsymbol{y}_k) + \langle \partial f(\boldsymbol{y}_k), \boldsymbol{x}-\boldsymbol{y}_k\rangle\Big)
$$

What are these $\phi_k, \alpha_k, \lambda_k$ actually?

$$\phi_{k+1}(\boldsymbol{x}) \;=\; (1-\alpha_k)\phi_k(\boldsymbol{x}) + \alpha_k \underbrace{\left(f(\boldsymbol{y}_k) + \langle \nabla f(\boldsymbol{y}_k), \boldsymbol{x} - \boldsymbol{y}_k\rangle + \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{y}_k\|_2^2\right)}_{\psi(\boldsymbol{x})} \tag{L2.2.2 A7}$$

- λ_k is defined by α_k so you can treat them as the same thing under different expression
- By A4, A5, A6, we can think of λ_k as the coefficient of convex combination and thus think of ϕ_{k+1} as convex combination of ϕ_k and $\psi(\boldsymbol{x})$
  - What is ψ: an global support / global under-estimator of f at a point $\boldsymbol{y}_k$
  - Therefore $\phi_{k+1} \;=\; \mathrm{cvx}\big(\phi_k, \psi\big) \;=\; \mathrm{cvx}\big(\phi_k, \text{lower estimate of } f\big)$

Understand ϕ_k through pictures ... 1/2

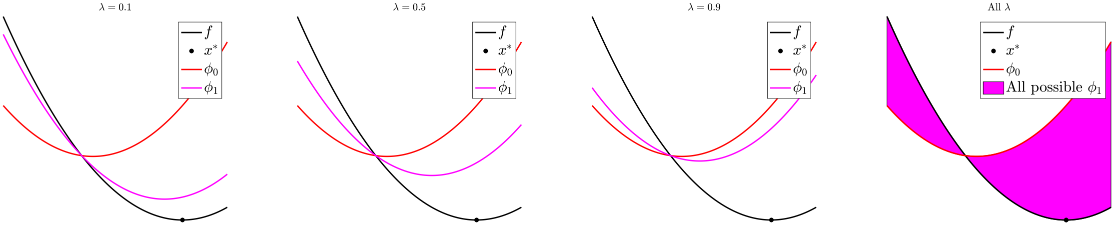


Recall

- We can pick any ϕ_0 as long as it is convex
- If f is strongly-convex, a simple ϕ_0 is a quadratic
- If f is convex, a simple ϕ_0 is a affine function (a line here)
- If f is convex and nondifferentiable, a simple ϕ_0 is a affine function (where we replace gradient by subgradient)

Understand ϕ_k through pictures ... 2/2

You build ϕ_1 using ϕ_0 and f via a convex combination with weights λ_1



- An observation: in all the cases, $\left\{\text{minimizer of } \phi_1\right\}$ is closer to $\boldsymbol{x}^*$ than $\left\{\text{minimizer of } \phi_0\right\}$ to $\boldsymbol{x}^*$
- Therefore, if we can somehow find $\left\{\text{minimizer of } \phi_1\right\}$ and then use it to construct/update to ϕ_2, we move closer to $\boldsymbol{x}^*$
- By Lemma 2.2.1, the convergence speed of such process is bounded above by how fast λ_k approaches to 0

Small summary

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

$$
\begin{array}{lcclc|l}
\text{(Def0)} & \lambda_k & \geq & 0 & (\forall k) & \{\lambda_k\}_{k\in\mathbb{N}} \text{ is nonnegative} \\
\text{(Def1)} & \lambda_k & \xrightarrow{k\to\infty} & 0 & (\forall k) & \{\lambda_k\}_{k\in\mathbb{N}} \text{ converges to 0} \\
\text{(Def2)} & \phi_k(\boldsymbol{x}) & \leq & (1-\lambda_k)f(\boldsymbol{x}) + \lambda_k\phi_0(\boldsymbol{x}) & (\forall k)\big(\forall \boldsymbol{x}\in\mathbb{R}^n\big) & \{\phi_k\}_{k\in\mathbb{N}} \leq \text{"convex combination" of } f,\ \phi_0
\end{array}
$$

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

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

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

IF

- A1 f L-smooth μ-strongly cvx
- A2 $\phi_0(\cdot)$ a cvx function
- A3 $\big\{\boldsymbol{y}_k\big\}_{k=0}^{\infty}$ is a sequence
- A4a $\alpha_k \in \,]\,0,1\,[\ \ \forall k$
- A4b $\displaystyle\sum_{k=0}^{\infty} \alpha_k = \infty$
- A5 $\lambda_0 := 1$

THEN $\big\{\phi_k(\boldsymbol{x}), \lambda_k\big\}_{k=0}^{\infty}$ defined by A6 A7 is an estimate sequence of f

- A6 $\lambda_{k+1} = (1-\alpha_k)\lambda_k$
- A7 $\phi_{k+1}(\boldsymbol{x}) = (1-\alpha_k)\phi_k(\boldsymbol{x}) + \alpha_k\Big(f(\boldsymbol{y}_k) + \langle \nabla f(\boldsymbol{y}_k), \boldsymbol{x}-\boldsymbol{y}_k\rangle + \frac{\mu}{2}\|\boldsymbol{x}-\boldsymbol{y}_k\|_2^2\Big)$

- **What now**
  - What is ϕ_0?
  - Well we can use any convex ϕ_0 by A2
  - We can use a simple function, like a quadratic

A simple quadratic ϕ_0

- We can just define ϕ_0 as

$$\phi_0(\boldsymbol{x}) \;:=\; \phi_0^* + \frac{\gamma_0}{2}\|\boldsymbol{x} - \boldsymbol{v}_0\|_2^2. \tag{Phi-0}$$

  - We now introduce three new things: ϕ, γ and $\boldsymbol{v}$
  - ϕ_0, γ_0 are scalars and $\boldsymbol{v}_0 \in \mathbb{R}^n$ is a vector
  - ϕ_0^* is a shifting parameter, shifting the parabola up and down
  - γ_0 is a slope parameter
  - $\boldsymbol{v}_0$ is a shifting parameter, shifting the parabola horizontally
  - ϕ_k, γ_k and $\boldsymbol{v}_k$ are all sequence that keep changing

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

  Why ϕ_k, γ_k and $\boldsymbol{v}_k$ are updated this way is not intuitive and can be considered as black magic by Nesterov.
  - Here μ is the strong convexity parameter of f

Lemma on ϕ_k

A7 : $\phi_{k+1}(\boldsymbol{x}) = (1-\alpha_k)\phi_k(\boldsymbol{x}) + \alpha_k\Big(f(\boldsymbol{y}_k) + \langle \nabla f(\boldsymbol{y}_k), \boldsymbol{x}-\boldsymbol{y}_k\rangle + \frac{\mu}{2}\|\boldsymbol{x}-\boldsymbol{y}_k\|_2^2\Big)$

- **Lemma 2.2.3** If

$$\phi_0(\boldsymbol{x}) \ := \ \phi_0^* + \frac{\gamma_0}{2}\|\boldsymbol{x}-\boldsymbol{v}_0\|_2^2. \tag{Phi-0}$$

Then $\big\{\phi_k(\boldsymbol{x})\big\}_{k=0}^{\infty}$ defined by A7 in Lemma 2.2.2 preserves the canonical form of $\{\phi_k(\boldsymbol{x})\}_{k\in\mathbb{N}}$

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

where ϕ_k, γ_k and $\boldsymbol{v}_k$ are defined as

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

- **Proof** First by definition (Phi-0) gives

$$\nabla^2\phi_0(\boldsymbol{x}) \ \overset{\text{(Phi-0)}}{=} \ \gamma_0\boldsymbol{I}_n. \tag{†}$$

What next we show $\nabla^2\phi_k(\boldsymbol{x})$ has the same form as $\nabla^2\phi_0(\boldsymbol{x})$, i.e., we want to show

$$\nabla^2\phi_k(\boldsymbol{x}) = \gamma_k\boldsymbol{I}_n.$$

We do so by induction.

Prove $\nabla^2 \phi_k(\boldsymbol{x}) = \gamma_k \boldsymbol{I}_n$.

$$
\begin{array}{ll}
\nabla^2 \phi_0(\boldsymbol{x}) \overset{\text{(Phi-0)}}{=} \gamma_0 \boldsymbol{I}_n & (\dagger) \\
\phi_{k+1}(\boldsymbol{x}) = (1-\alpha_k)\phi_k(\boldsymbol{x}) + \alpha_k \Big(f(\boldsymbol{y}_k) + \langle \nabla f(\boldsymbol{y}_k), \boldsymbol{x} - \boldsymbol{y}_k \rangle + \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{y}_k\|_2^2 \Big) & (A7) \\
\gamma_{k+1} = (1-\alpha_k)\gamma_k + \alpha_k \mu & (i)
\end{array}
$$

- Base case is proved by ($\dagger$)
- Induction hypothesis: $\nabla^2 \phi_k(\boldsymbol{x}) = \gamma_k \boldsymbol{I}_n$
- Case $k+1$

$$
\begin{array}{rcll}
\phi_{k+1}(\boldsymbol{x}) & = & (1-\alpha_k)\phi_k(\boldsymbol{x}) + \alpha_k \Big(f(\boldsymbol{y}_k) + \langle \nabla f(\boldsymbol{y}_k), \boldsymbol{x} - \boldsymbol{y}_k \rangle + \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{y}_k\|_2^2 \Big) & \text{by def (A7)} \\
\nabla \phi_{k+1}(\boldsymbol{x}) & = & (1-\alpha_k)\nabla\phi_k(\boldsymbol{x}) + \alpha_k \Big(\nabla f(\boldsymbol{y}_k) + \mu(\boldsymbol{x} - \boldsymbol{y}_k) \Big) & \\
\nabla^2 \phi_{k+1}(\boldsymbol{x}) & = & (1-\alpha_k)\nabla^2\phi_k(\boldsymbol{x}) + \alpha_k \mu \boldsymbol{I} & \\
 & = & (1-\alpha_k)\gamma_k \boldsymbol{I}_n + \alpha_k \mu \boldsymbol{I} & \text{induction hypothesis} \\
 & = & \Big((1-\alpha_k)\gamma_k + \alpha_k \mu \Big) \boldsymbol{I} & \\
 & = & \gamma_{k+1} \boldsymbol{I} &
\end{array}
$$

- Hence now we have showed $\nabla^2 \phi_k(\boldsymbol{x}) = \gamma_k \boldsymbol{I}_n$. This equation means that if we perform the antiderivative twice we get

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

  for a scalar ϕ_k^* and a vector $\boldsymbol{v}_k$. Our remaining tasks are to
  - shows $\boldsymbol{v}_k$ satisfies (ii) in Lemma 2.2.3
  - shows ϕ_k^* satisfies (iii) in Lemma 2.2.3

Proving $\boldsymbol{v}_{k+1}$.

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

$$\phi_{k+1}(\boldsymbol{x}) = (1-\alpha_k)\phi_k(\boldsymbol{x}) + \alpha_k\Big(f(\boldsymbol{y}_k) + \langle \nabla f(\boldsymbol{y}_k), \boldsymbol{x} - \boldsymbol{y}_k\rangle + \tfrac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{y}_k\|_2^2\Big) \qquad (A7)$$

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

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

- First combine (A7) and $(\dagger\dagger)$

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

- What are we going to do now is to find the minimizer of ϕ_{k+1} and denote it as $\boldsymbol{v}_{k+1}$. I.e., find $\boldsymbol{v}_{k+1} = \text{argmin } \phi_{k+1}$. This is basically the idea from the pictures of ϕ_k we previously seen.
  - Take gradient

$$\nabla\phi_{k+1}(\boldsymbol{x}) = (1-\alpha_k)\gamma_k(\boldsymbol{x} - \boldsymbol{v}_k) + \alpha_k\Big(\nabla f(\boldsymbol{y}_k) + \mu(\boldsymbol{x} - \boldsymbol{y}_k)\Big)$$

  - Consider at minimizer $\boldsymbol{v}_{k+1}$ that $\nabla\phi_{k+1}(\boldsymbol{v}_{k+1}) = \boldsymbol{0}$

$$\begin{aligned}
& (1-\alpha_k)\gamma_k(\boldsymbol{v}_{k+1} - \boldsymbol{v}_k) + \alpha_k\nabla f(\boldsymbol{y}_k) + \alpha_k\mu(\boldsymbol{v}_{k+1} - \boldsymbol{y}_k) = \boldsymbol{0} \\
\iff\quad & \big((1-\alpha_k)\gamma_k + \alpha_k\mu\big)\boldsymbol{v}_{k+1} + \alpha_k\nabla f(\boldsymbol{y}_k) - (1-\alpha_k)\gamma_k\boldsymbol{v}_k - \alpha_k\mu\boldsymbol{y}_k = \boldsymbol{0} \\
\iff\quad & \gamma_{k+1}\boldsymbol{v}_{k+1} = (1-\alpha_k)\gamma_k\boldsymbol{v}_k + \alpha_k\mu\boldsymbol{y}_k - \alpha_k\nabla f(\boldsymbol{y}_k) \\
\iff\quad & \boldsymbol{v}_{k+1} = \frac{(1-\alpha_k)\gamma_k\boldsymbol{v}_k + \alpha_k\mu\boldsymbol{y}_k - \alpha_k\nabla f(\boldsymbol{y}_k)}{\gamma_{k+1}} \\
\iff\quad & (ii)
\end{aligned}$$

Proving ϕ_{k+1}^*.

$$\begin{aligned}
\phi_k(\boldsymbol{x}) &= \phi_k^* + \frac{\gamma_k}{2}\|\boldsymbol{x}-\boldsymbol{v}_k\|_2^2. && (\dagger\dagger)\\
\phi_{k+1}(\boldsymbol{x}) &= (1-\alpha_k)\phi_k(\boldsymbol{x}) + \alpha_k\Big(f(\boldsymbol{y}_k) + \langle \nabla f(\boldsymbol{y}_k), \boldsymbol{x}-\boldsymbol{y}_k\rangle + \tfrac{\mu}{2}\|\boldsymbol{x}-\boldsymbol{y}_k\|_2^2\Big) && (A7)\\
\boldsymbol{v}_{k+1} &= \frac{(1-\alpha_k)\gamma_k\boldsymbol{v}_k + \alpha_k\mu\boldsymbol{y}_k - \alpha_k\nabla f(\boldsymbol{y}_k)}{\gamma_{k+1}} && (ii)\\
\gamma_{k+1} &= (1-\alpha_k)\gamma_k + \alpha_k\mu && (i)
\end{aligned}$$

- $(\dagger\dagger) = (A7)$ at $k+1$

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

- Put $\boldsymbol{x} = \boldsymbol{y}_k$

$$\begin{aligned}
\phi_{k+1}^* + \frac{\gamma_{k+1}}{2}\|\boldsymbol{y}_k-\boldsymbol{v}_{k+1}\|_2^2 &= (1-\alpha_k)\phi_k(\boldsymbol{y}_k) + \alpha_k f(\boldsymbol{y}_k)\\
&\overset{(\dagger\dagger)}{=} (1-\alpha_k)\Big(\phi_k^* + \frac{\gamma_k}{2}\|\boldsymbol{y}_k-\boldsymbol{v}_k\|_2^2\Big) + \alpha_k f(\boldsymbol{y}_k) \qquad (\dagger\dagger\dagger)
\end{aligned}$$

- By (ii)

$$\begin{aligned}
\boldsymbol{v}_{k+1}-\boldsymbol{y}_k &= \frac{(1-\alpha_k)\gamma_k\boldsymbol{v}_k + \alpha_k\mu\boldsymbol{y}_k - \alpha_k\nabla f(\boldsymbol{y}_k)}{\gamma_{k+1}} - \boldsymbol{y}_k\\
&= \frac{(1-\alpha_k)\gamma_k\boldsymbol{v}_k + \alpha_k\mu\boldsymbol{y}_k - \gamma_{k+1}\boldsymbol{y}_k - \alpha_k\nabla f(\boldsymbol{y}_k)}{\gamma_{k+1}}\\
&\overset{(i)}{=} \frac{(1-\alpha_k)\gamma_k(\boldsymbol{v}_k-\boldsymbol{y}_k) - \alpha_k\nabla f(\boldsymbol{y}_k)}{\gamma_{k+1}}\\
\frac{\gamma_{k+1}}{2}\|\boldsymbol{v}_{k+1}-\boldsymbol{y}_k\|_2^2 &= \frac{(1-\alpha_k)^2\gamma_k^2\|\boldsymbol{v}_k-\boldsymbol{y}_k\|_2^2 - 2\big\langle(1-\alpha_k)\gamma_k(\boldsymbol{v}_k-\boldsymbol{y}_k), \alpha_k\nabla f(\boldsymbol{y}_k)\big\rangle + \alpha_k^2\|\nabla f(\boldsymbol{y}_k)\|_2^2}{2\gamma_{k+1}} \quad (\dagger\dagger\dagger\dagger)
\end{aligned}$$

- Put $(\dagger\dagger\dagger\dagger)$ into $(\dagger\dagger\dagger)$ will give (iii), trust me.

$$\phi_{k+1}^* + \frac{\gamma_{k+1}}{2}\|\boldsymbol{y}_k - \boldsymbol{v}_{k+1}\|_2^2 = (1-\alpha_k)\Big(\phi_k^* + \frac{\gamma_k}{2}\|\boldsymbol{y}_k - \boldsymbol{v}_k\|_2^2\Big) + \alpha_k f(\boldsymbol{y}_k) \qquad (\dagger\dagger\dagger)$$

$$\frac{\gamma_{k+1}}{2}\|\boldsymbol{v}_{k+1} - \boldsymbol{y}_k\|_2^2 = \frac{(1-\alpha_k)^2\gamma_k^2\|\boldsymbol{v}_k - \boldsymbol{y}_k\|_2^2 - 2\langle(1-\alpha_k)\gamma_k(\boldsymbol{v}_k - \boldsymbol{y}_k), \alpha_k\nabla f(\boldsymbol{y}_k)\rangle + \alpha_k^2\|\nabla f(\boldsymbol{y}_k)\|_2^2}{2\gamma_{k+1}} \qquad (\dagger\dagger\dagger\dagger)$$

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

What to do: show $\Big\{(\dagger\dagger\dagger) \text{ and } (\dagger\dagger\dagger\dagger)\Big\} - (iii) = 0$

$$\begin{aligned}
& (1-\alpha_k)\Big(\phi_k^* + \frac{\gamma_k}{2}\|\boldsymbol{y}_k - \boldsymbol{v}_k\|_2^2\Big) + \alpha_k f(\boldsymbol{y}_k) - \frac{\gamma_{k+1}}{2}\|\boldsymbol{y}_k - \boldsymbol{v}_{k+1}\|_2^2 \\
& -\Big\{(1-\alpha_k)\phi_k^* + \alpha_k f(\boldsymbol{y}_k) - \frac{\alpha_k^2}{2\gamma_{k+1}}\|\nabla f(\boldsymbol{y}_k)\|_2^2 + \frac{\alpha_k(1-\alpha_k)\gamma_k}{\gamma_{k+1}}\Big(\frac{\mu}{2}\|\boldsymbol{y}_k - \boldsymbol{v}_k\|_2^2 + \langle\nabla f(\boldsymbol{y}_k), \boldsymbol{v}_k - \boldsymbol{y}_k\rangle\Big)\Big\} \\
= \; & (1-\alpha_k)\frac{\gamma_k}{2}\|\boldsymbol{y}_k - \boldsymbol{v}_k\|_2^2 - \frac{(1-\alpha_k)^2\gamma_k^2\|\boldsymbol{v}_k - \boldsymbol{y}_k\|_2^2 - 2\langle(1-\alpha_k)\gamma_k(\boldsymbol{v}_k - \boldsymbol{y}_k), \alpha_k\nabla f(\boldsymbol{y}_k)\rangle + \alpha_k^2\|\nabla f(\boldsymbol{y}_k)\|_2^2}{2\gamma_{k+1}} \\
& -\Big\{-\frac{\alpha_k^2}{2\gamma_{k+1}}\|\nabla f(\boldsymbol{y}_k)\|_2^2 + \frac{\alpha_k(1-\alpha_k)\gamma_k}{\gamma_{k+1}}\Big(\frac{\mu}{2}\|\boldsymbol{y}_k - \boldsymbol{v}_k\|_2^2 + \langle\nabla f(\boldsymbol{y}_k), \boldsymbol{v}_k - \boldsymbol{y}_k\rangle\Big)\Big\} \\
= \; & \frac{(1-\alpha_k)\gamma_k}{2}\Big[1 - \frac{(1-\alpha_k)\gamma_k}{\gamma_{k+1}}\Big]\|\boldsymbol{y}_k - \boldsymbol{v}_k\|_2^2 + \frac{(1-\alpha_k)\alpha_k\gamma_k}{\gamma_{k+1}}\langle\boldsymbol{v}_k - \boldsymbol{y}_k, \nabla f(\boldsymbol{y}_k)\rangle \\
& -\Big\{\frac{\alpha_k(1-\alpha_k)\gamma_k}{\gamma_{k+1}}\frac{\mu}{2}\|\boldsymbol{y}_k - \boldsymbol{v}_k\|_2^2 + \frac{\alpha_k(1-\alpha_k)\gamma_k}{\gamma_{k+1}}\Big(\langle\nabla f(\boldsymbol{y}_k), \boldsymbol{v}_k - \boldsymbol{y}_k\rangle\Big)\Big\} \\
= \; & 0 \text{ by (i) } \gamma_{k+1} = (1-\alpha_k)\gamma_k + \alpha_k\mu
\end{aligned}$$

Last page

- **[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 $\{\boldsymbol{x}_k\}_{k\in\mathbb{N}}$:

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

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

IF
- A1 f L-smooth μ-strongly cvx
- A2 $\phi_0(\cdot)$ a cvx function
- A3 $\{\boldsymbol{y}_k\}_{k=0}^{\infty}$ is a sequence
- A4a $\alpha_k \in \,]\,0,1\,[\ \ \forall k$
- A4b $\displaystyle\sum_{k=0}^{\infty}\alpha_k = \infty$
- A5 $\lambda_0 := 1$

THEN $\{\phi_k(\boldsymbol{x}), \lambda_k\}_{k=0}^{\infty}$ defined by A6 A7 is an estimate sequence of f
- A6 $\lambda_{k+1} = (1-\alpha_k)\lambda_k$
- A7 $\phi_{k+1}(\boldsymbol{x}) = (1-\alpha_k)\phi_k(\boldsymbol{x}) + \alpha_k\Big(f(\boldsymbol{y}_k) + \langle\nabla f(\boldsymbol{y}_k), \boldsymbol{x}-\boldsymbol{y}_k\rangle + \frac{\mu}{2}\|\boldsymbol{x}-\boldsymbol{y}_k\|_2^2\Big)$

- **[Lemma 2.2.3 (a quadratic ϕ_0)]** **IF** $\phi_0(\boldsymbol{x}) := \phi_0^* + \dfrac{\gamma_0}{2}\|\boldsymbol{x}-\boldsymbol{v}_0\|_2^2$ **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

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

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

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