Lorum ipsum

HONGDA LI *

July 2, 2025

This paper is currently in draft mode. Check source to change options.

Abstract

Lorem ipsum dolor sit amet, dicta iudicabit consequat ex vix, veniam legimus appetere has id, an pri graece epicuri detraxit. Ea aliquam expetendis posidonium eos, nam invenire corrumpit imperdiet ei. Et constituto dissentias usu, mel solum erant et. Mel dolorem menandri in. [3]

This is just psuedo text. This is just psuedo text. This is just psuedo text. This is just psuedo text.

2010 Mathematics Subject Classification: Primary 47H05, 52A41, 90C25; Secondary 15A09, 26A51, 26B25, 26E60, 47H09, 47A63. **Keywords:**

1 Nesterov's Accelerated Gradient

1.1 In preparations

{ass:smooth-plus-nonsmooth}

Assumption 1.1 (smooth add nonsmooth) The function $F = f + g$ where $f : \mathbb{R}^n \to \mathbb{R}$ is a L Lipschitz smooth and $\mu \geq 0$ strongly convex function. The function $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ is a closed convex proper function.

{ass:smooth-plus-nonsmooth-x}

Assumption 1.2 (admitting minimizers) Let $F = f + g$ and in addition assume that the set of minimizers $X^+ := \operatorname{argmin}_x F(x)$ is non-empty.

*University of British Columbia Okanagan, Canada. E-mail: alto@mail.ubc.ca.

Definition 1.3 (Proximal gradient operator) *Suppose $F = f + g$ satisfies Assumption 1.1. Let $\beta > 0$. Then, we define the proximal gradient operator T_β as*

$$T_\beta(x|F) = \operatorname{argmin} z \left\{ g(z) + f(x) + \langle \nabla f(x), z - x \rangle + \frac{\beta}{2}\|z - x\|^2 \right\}.$$

Remark 1.4 If the function $g \equiv 0$, then it yields the gradient descent operator $T_\beta(x) = x - \beta^{-1}\nabla f(x)$. In the context where it's clear what the function $F = f + g$ is, we simply write $T_\beta(x)$ for short.

Definition 1.5 (Bregman Divergence) *Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a differentiable function. Then, for all the Bregman divergence $D_f : \mathbb{R}^n \times \operatorname{dom} \nabla f \to \mathbb{R}$ is defined as:*

$$D_f(x, y) := f(x) - f(y) - \langle \nabla f(y), x - y \rangle.$$

Remark 1.6 If, f is $\mu \geq 0$ strongly convex and L Lipschitz smooth then, its Bregman Divergence has for all $x, y \in \mathbb{R}^n$: $\mu/2\|x - y\|^2 \leq D_f(x, y) \leq L/2\|x - y\|^2$.

{def:rwapg-seq}

Definition 1.7 (R-WAPG sequence) *Let $(L_k)_{k\geq 0}$ be a sequence such that $L_k > \mu$ for all k. Let $\alpha_0 \in (0, 1]$, $(\alpha_k)_{k\geq 1}$ has $\alpha_k \in (\mu/L_k, 1)$. Then define for all $k \geq 0$:*

$$\rho_k(1 - \alpha_{k+1})\alpha_k^2 = \alpha_{k+1}(\alpha_{k+1} - \mu/L_k).$$

Remark 1.8 When $\rho_k = 1$, the recursive relation between α_k, α_{k-1} is the same as the well known Nesterov's sequence used in algorithm such as FISTA and Nesterov's accelerated gradient. See Li and Wang [2] for more information.

{def:st-method}

Definition 1.9 (similar triangle representation of NAPG) *Let $(\alpha_k)_{k\geq 0}$ be an R-WAPG sequence. Suppose that the base case $v_{-1}, x_{k-1} \in \mathbb{R}^n$ is given to initialize the algorithm. Then the algorithm produces the sequence of iterates $(y_k, x_k, v_k)_{k\geq 0}$ and auxiliary parameter sequence L_k, τ_k satisfying these inequalities:*

$$\begin{aligned}
&\tau_k = L_k(1 - \alpha_k)(L_k\alpha_k - \mu)^{-1}, \\
&y_k = (1 + \tau_k)^{-1}v_{k-1} + \tau_k(1 + \tau_k)^{-1}x_{k-1}, \\
&D_f(x_k, y_k) \leq L_k/2\|x_k - y_k\|^2, \\
&x_k = T_{L_k}(y_k), \\
&v_k = x_{k-1} + \alpha_k^{-1}(x_k - x_{k-1}).
\end{aligned}$$

{thm:pg-ineq}

The following theorems are critical in analyzing the behavior of algorithm in Definition 1.9.

Theorem 1.10 (proximal gradient inequality) *Let function F satisfies Assumption 1.1, so it's $\mu \geq 0$ strongly convex. For all $x \in \mathbb{R}^n$, define $x^+ = T_L(x)$, then there exists*

a $B \geq 0$ such that $D_f(x^+, x) \leq B/2\|x^+ - x\|^2$. Then, for all $z \in \mathbb{R}^n$ it satisfies proximal gradient inequality at point x:

$$\begin{aligned} 0 &\leq F(z) - F(x^+) - \frac{B}{2}\|z - x^+\|^2 + \frac{B-\mu}{2}\|z - x\|^2 \\ &= F(z) - F(x^+) - \langle B(x - x^+), z - x\rangle - \frac{\mu}{2}\|z - x\|^2 - \frac{B}{2}\|x - x^+\|^2. \end{aligned}$$

Since f is assumed to be L Lipschitz smooth, the above condition is true for all $x, y \in \mathbb{R}^n$ for all $B \geq L$.

Remark 1.11 The theorem is the same as in Nesterov's book [3, Theorem 2.2.13], but with the use of proximal gradient mapping and proximal gradient instead of project gradient hence making it equivalent to the theorem in Beck's book [1, Theorem 10.16]. The only generalization here is parameter B which made to accommodate algorithm that implements Definition 1.9 with some line search routine.

{thm:jesen}

Theorem 1.12 (Jensen's inequality) *Let $F : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a $\mu \geq 0$ strongly convex function. Then, it is equivalent to the following condition. For all $x, y \in \mathbb{R}^n$, $\lambda \in (0, 1)$ it satisfies the inequality*

$$(\forall \lambda \in [0, 1])\ F(\lambda x + (1-\lambda)y) \leq \lambda F(x) + (1-\lambda)F(y) - \frac{\mu\lambda(1-\lambda)}{2}\|y - x\|^2.$$

Remark 1.13 If x, y is out of $\operatorname{dom} F$, the inequality still work by convexity.

1.2 stochastic accelerated proximal gradient

{ass:sum-of-many}

The following assumption about the objective function is fundamental in incremental gradient method for Machine Learning, data science other similar tasks.

Assumption 1.14 (sum of many) Define $F := (1/n)\sum_{i=1}^n F_i$ where $F_i = f_i + g_i$. Assume that for all $i = 1, \ldots, n$, each $f_i : \mathbb{R}^n \to \mathbb{R}$ are $K^{(i)}$ smooth and $\mu^{(i)} \geq 0$ strongly convex function such that $K^{(i)} > \mu^{(i)}$ and, $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ is a closed convex proper function. Consequently, the function f can be written as $F = g + f$ with $f = (1/n)\sum_{i=1}^n f_i$, $g = (1/n)\sum_{i=1}^n f_i + g_i$ there, it also satisfies Assumption 1.1 with $L = (1/2)\sum_{i=1}^n K^{(i)}$ and $\mu = (1/n)\sum_{i=1}^n \mu^{(i)}$.

This assumption is stronger than Assumption 1.1. It still appears everywhere in practice, for example, in the case of convex feasibility problem $\bigcap_{i=1}^n C_i$. In practice, each of the strong convexity constant $\mu^{(i)}$ may not be easily accessible.

The interpolation hypothesis from Machine Learning stated that the model has the capacity to perfect fit all the observed data.

{ass:interp-hypothesis}

Assumption 1.15 (interpolation hypothesis) Suppose that $F := (1/n)\sum_{i=1}^{n} F_i$ satisfying Assumption 1.14. In addition, assuming that it has $0 = \inf_x F(x)$ and, there exists some $\bar{x} \in \mathbb{R}^n$ such that for all $i = 1, \ldots, n$ it satisfies $0 = f_i(\bar{x})$. Obviously, all such $\bar{x}$ forms the set of minimizers of F.

{def:snapg-v2}

Definition 1.16 (SNAPG-V2) *Let F satisfies Assumption 1.14. Let $(I_k)_{k\geq 0}$ be a list of i.i.d random variables uniformly sampled from set $\{0, 1, 2, \cdots, n\}$. Initialize $v_{-1} = x_{-1}, \alpha_0 = 1$. The SNAPG generates the sequence $(y_k, x_k, v_k)_{k\geq 0}$ such that for all $k \geq 0$ they satisfy:*

$$
\begin{aligned}
&(L_{k-1}/L_k)(1-\alpha_k)\alpha_{k-1}^2 = \alpha_k\left(\alpha_k - \mu/L_k\right),\\
&\tau_k = L_k(1-\alpha_k)\left(L_k\alpha_k - \mu^{(I_k)}\right)^{-1},\\
&y_k = (1+\tau_k)^{-1}v_{k-1} + \tau_k(1+\tau_k)^{-1}x_{k-1},\\
&x_k = T_{L_k}(y_k|F_{I_k}) \ s.t.\ D_f(x_k, y_k) \leq L_k/2\|y_k - x_k\|^2,\\
&v_k = x_{k-1} + \alpha_k^{-1}(x_k - x_{k-1}).
\end{aligned}
$$

{lemma:snapg-v2-seq-range}

Lemma 1.17 (range of the momentum sequence in SNAPG-V2) *Suppose that $(L_k)_{k\geq 0}$ is a sequence such that $L_k > 0$ for all $k \geq 0$. Let $(\alpha_k)_{k\geq 0}$ be a sequence such that $\alpha_0 \in (0, 1]$ and, for all $k \geq 1$, it's recursively satisfies equality:*

$$
(L_{k-1}/L_k)(1-\alpha_k)\alpha_{k-1}^2 = \alpha_k\left(\alpha_k - \mu/L_k\right).
$$

And, the following items are true:

(i) *Solution to the equation with $\alpha_k > 0$ is given by:*

$$
\alpha_k = \frac{L_{k-1}}{2L_k}\left(-\alpha_{k-1}^2 + \frac{\mu}{L_{k-1}} + \sqrt{\left(\alpha_{k-1} - \frac{\mu}{L_{k-1}}\right)^2 + \frac{4\alpha_{k-1}^2 L_k}{L_{k-1}}}\right).
$$

{thm:snapg2-one-step}

Theorem 1.18 (SNAPG-V2 one step convergence) *Let F satisfies assumption 1.15. Suppose that an algorithm satisfying Definition 1.16 uses this F. Then, for all $k \geq 1$, it has the following inequality*

$$
\begin{aligned}
&\mathbb{E}_k\left[F_{I_k}(x_k)\right] - F(\bar{x}) + \mathbb{E}_k\left[\frac{L_k\alpha_k^2}{2}\|\bar{x} - v_k\|^2\right]\\
&\leq (1-\alpha_k)\left(\mathbb{E}_k\left[F_{I_k}(x_{k-1})\right] - F(\bar{x}) + \mathbb{E}_k\left[\frac{\alpha_{k-1}^2 L_{k-1}}{2}\right]\|v_{k-1} - \bar{x}\|^2\right)\\
&\quad + \mathbb{E}_k\left[\frac{(\alpha_k - 1)\mu^{(I_k)}\left(L_k\alpha_k - \mu^{(I_k)}\right)}{2\left(L_k - \mu^{(I_k)}\right)}\right]\|x_{k-1} - v_{k-1}\|^2.
\end{aligned}
$$

Proof. Let's suppose that $I_k = i$ and, for all $k \geq 0$ let $z_k = \alpha_k \bar{x} + (1 - \alpha_k) x_{k-1}$ where $\bar{x}$ is a minimizer of F. With following intermediate results the proof can be built easily. Results (d)-(g) is showed at the end.

(a) We can use proximal gradient inequality from Theorem 1.10 with $z = z_k$ because each F_i is K_i Lipschitz smooth and, $\mu^{(i)}$ strongly convex with $K_i \geq \mu^{(i)}$.

(b) We can use Jensen's inequality of Theorem 1.12 with $z = z_k$ on F_i.

(c) The sequence $(\alpha_k)_{k\geq 0}$ has $(L_{k-1}/L_k)(1-\alpha_k)\alpha_{k-1}^2 = \alpha_k\left(\alpha_k - \mu/L_k\right)$. It is a special case of Definition 1.7 with $\rho_{k-1} = L_{k-1}/L_k$.

(d) From Definition 1.16 it has the following equality

$$(\forall k \geq 1)\ z_k - y_k = \frac{L_k\alpha_k - \mu^{(i)}}{L_k - \mu^{(i)}}(\bar{x} - v_{k-1}) + \frac{\mu^{(i)}(1-\alpha_k)}{L_k - \mu^{(i)}}(\bar{x} - x_{k-1}).$$

(e) From Definition 1.16 it has: $(\forall k \geq 1)\ z_k - x_k = \alpha_k(\bar{x} - v_k)$.

(f) Using direct algebra, we have for all $k \geq 1$:

$$\frac{\left(\mu^{(i)}\right)^2(1-\alpha_k)^2}{2(L_k - \mu^{(i)})} - \frac{\mu^{(i)}\alpha_k(1-\alpha_k)}{2} = \frac{(\alpha_k - 1)\mu^{(i)}\left(L_k\alpha_k - \mu^{(i)}\right)}{2\left(L_k - \mu^{(i)}\right)}.$$

(g) Using (c), we have for all $k \geq 1$:

$$\frac{\left(L_k\alpha_k - \mu^{(i)}\right)^2}{2(L_k - \mu^{(i)})} - \frac{\alpha_{k-1}^2 L_{k-1}(1-\alpha_k)}{2} = \frac{\left(L_k\alpha_k - \mu^{(i)}\right)\mu^{(i)}\left(\alpha_k - 1\right)}{2(L_k - \mu^{(i)})} + \frac{\alpha_k(\mu - \mu^{(i)})}{2}.$$

(h) Because we assumed interpolation hypothesis in Assumption 1.15, it has $\mathbb{E}[F_{I_k}(\bar{x})] = F(\bar{x})$ for all $\bar{x}$ that is a minimizer of F.

For all $k \geq 1$, starting with (a) we have:

{ineq:snapg2-one-step-chain1}

$$\begin{aligned} 0 &\leq F_i(z_k) - F_i(x_k) - \frac{L_k}{2}\|z_k - x_k\|^2 + \frac{L_k - \mu^{(i)}}{2}\|z_k - y_k\|^2 \\ &\underset{\text{(b)}}{\leq} \alpha_k F_i(\bar{x}) + (1-\alpha_k)F_i(x_{k-1}) - F_i(x_k) \\ &\quad - \frac{\mu^{(i)}\alpha_k(1-\alpha_k)}{2}\|\bar{x} - x_{k-1}\|^2 - \frac{L_k}{2}\|z_k - x_k\|^2 + \frac{L_k - \mu^{(i)}}{2}\|z_k - y_k\|^2. \end{aligned} \tag{1.1}$$

And we have the following chain of equalities:

$$-\frac{\mu^{(i)}\alpha_k(1-\alpha_k)}{2}\|\bar{x} - x_{k-1}\|^2 + \frac{L_k - \mu^{(i)}}{2}\|z_k - y_k\|^2$$

$$
\begin{aligned}
&\underset{\text{(d)}}{=} -\frac{\mu^{(i)}\alpha_k(1-\alpha_k)}{2}\|\bar{x}-x_{k-1}\|^2 \\
&\quad + \frac{L_k-\mu^{(i)}}{2}\left\|\frac{L_k\alpha_k-\mu^{(i)}}{L_k-\mu^{(i)}}(\bar{x}-v_{k-1})+\frac{\mu^{(i)}(1-\alpha_k)}{L_k-\mu^{(i)}}(\bar{x}-x_{k-1})\right\|^2 \\
&= -\frac{\mu^{(i)}\alpha_k(1-\alpha_k)}{2}\|\bar{x}-x_{k-1}\|^2 \\
&\quad + \frac{(L_k\alpha_k-\mu^{(i)})^2}{2(L_k-\mu^{(i)})}\|\bar{x}-v_{k-1}\|^2+\frac{(\mu^{(i)})^2(1-\alpha_k)^2}{2(L_k-\mu^{(i)})}\|\bar{x}-x_{k-1}\|^2 \\
&\quad + \frac{(L_k\alpha_k-\mu^{(i)})\mu^{(i)}(1-\alpha_k)}{(L_k-\mu^{(i)})}\langle\bar{x}-v_{k-1},\bar{x}-x_{k-1}\rangle \\
&= \left(\frac{(\mu^{(i)})^2(1-\alpha_k)^2}{2(L_k-\mu^{(i)})}-\frac{\mu^{(i)}\alpha_k(1-\alpha_k)}{2}\right)\|\bar{x}-x_{k-1}\|^2 \\
&\quad + \left(\frac{(L_k\alpha_k-\mu^{(i)})^2}{2(L_k-\mu^{(i)})}-\frac{\alpha_{k-1}^2L_{k-1}(1-\alpha_k)}{2}\right)\|\bar{x}-v_{k-1}\|^2+\frac{\alpha_{k-1}^2L_{k-1}(1-\alpha_k)}{2}\|\bar{x}-v_{k-1}\|^2 \\
&\quad + \frac{(L_k\alpha_k-\mu^{(i)})\mu^{(i)}(1-\alpha_k)}{(L_k-\mu^{(i)})}\langle\bar{x}-v_{k-1},\bar{x}-x_{k-1}\rangle \\
&\underset{\text{(f)}}{=} \frac{(\alpha_k-1)\mu^{(i)}\left(L_k\alpha_k-\mu^{(i)}\right)}{2\left(L_k-\mu^{(i)}\right)}\|\bar{x}-x_{k-1}\|^2 \\
&\quad + \left(\frac{(L_k\alpha_k-\mu^{(i)})^2}{2(L_k-\mu^{(i)})}-\frac{\alpha_{k-1}^2L_{k-1}(1-\alpha_k)}{2}\right)\|\bar{x}-v_{k-1}\|^2+\frac{\alpha_{k-1}^2L_{k-1}(1-\alpha_k)}{2}\|\bar{x}-v_{k-1}\|^2 \\
&\quad + \frac{(L_k\alpha_k-\mu^{(i)})\mu^{(i)}(1-\alpha_k)}{(L_k-\mu^{(i)})}\langle\bar{x}-v_{k-1},\bar{x}-x_{k-1}\rangle \\
&\underset{\text{(g)}}{=} \frac{(\alpha_k-1)\mu^{(i)}\left(L_k\alpha_k-\mu^{(i)}\right)}{2\left(L_k-\mu^{(i)}\right)}\|\bar{x}-x_{k-1}\|^2 \\
&\quad + \left(\frac{\left(L_k\alpha_k-\mu^{(i)}\right)\mu^{(i)}\left(\alpha_k-1\right)}{2(L_k-\mu^{(i)})}+\frac{\alpha_k(\mu-\mu^{(i)})}{2}\right)\|\bar{x}-v_{k-1}\|^2 \\
&\quad + \frac{\alpha_{k-1}^2L_{k-1}(1-\alpha_k)}{2}\|\bar{x}-v_{k-1}\|^2+\frac{(L_k\alpha_k-\mu^{(i)})\mu^{(i)}(1-\alpha_k)}{(L_k-\mu^{(i)})}\langle\bar{x}-v_{k-1},\bar{x}-x_{k-1}\rangle \\
&= \frac{(\alpha_k-1)\mu^{(i)}\left(L_k\alpha_k-\mu^{(i)}\right)}{2\left(L_k-\mu^{(i)}\right)}\left(\|\bar{x}-x_{k-1}\|^2+\|\bar{x}-v_{k-1}\|^2-2\langle\bar{x}-v_{k-1},\bar{x}-x_{k-1}\rangle\right) \\
&\quad + \frac{\alpha_k(\mu-\mu^{(i)})}{2}\|\bar{x}-v_{k-1}\|^2+\frac{\alpha_{k-1}^2L_{k-1}(1-\alpha_k)}{2}\|\bar{x}-v_{k-1}\|^2 \\
&= \frac{(\alpha_k-1)\mu^{(i)}\left(L_k\alpha_k-\mu^{(i)}\right)}{2\left(L_k-\mu^{(i)}\right)}\|x_{k-1}-v_{k-1}\|^2
\end{aligned}
$$

$$+\frac{\alpha_k(\mu-\mu^{(i)})}{2}\|\bar{x}-v_{k-1}\|^2+\frac{\alpha_{k-1}^2L_{k-1}(1-\alpha_k)}{2}\|\bar{x}-v_{k-1}\|^2.$$

Substituting the above back to the tail of Inequality (1.1) it gives:

$$\begin{aligned}
0 &\le \alpha_k F_i(\bar{x})+(1-\alpha_k)F_i(x_{k-1})-F_i(x_k)\\
&\quad-\frac{L_k}{2}\|z_k-x_k\|^2+\frac{(\alpha_k-1)\mu^{(i)}\left(L_k\alpha_k-\mu^{(i)}\right)}{2\left(L_k-\mu^{(i)}\right)}\|x_{k-1}-v_{k-1}\|^2\\
&\quad+\frac{\alpha_k(\mu-\mu^{(i)})}{2}\|\bar{x}-v_{k-1}\|^2+\frac{\alpha_{k-1}^2L_{k-1}(1-\alpha_k)}{2}\|\bar{x}-v_{k-1}\|^2\\
&\underset{\text{(e)}}{=} \alpha_k F_i(\bar{x})+(1-\alpha_k)F_i(x_{k-1})-F_i(x_k)\\
&\quad-\frac{L_k\alpha_k^2}{2}\|\bar{x}-v_k\|^2+\frac{(\alpha_k-1)\mu^{(i)}\left(L_k\alpha_k-\mu^{(i)}\right)}{2\left(L_k-\mu^{(i)}\right)}\|x_{k-1}-v_{k-1}\|^2\\
&\quad+\frac{\alpha_k(\mu-\mu^{(i)})}{2}\|\bar{x}-v_{k-1}\|^2+\frac{\alpha_{k-1}^2L_{k-1}(1-\alpha_k)}{2}\|\bar{x}-v_{k-1}\|^2\\
&= (\alpha_k-1)F_i(\bar{x})+(1-\alpha_k)F_i(x_{k-1})-F_i(x_k)+F_i(\bar{x})\\
&\quad-\frac{L_k\alpha_k^2}{2}\|\bar{x}-v_k\|^2+\frac{(\alpha_k-1)\mu^{(i)}\left(L_k\alpha_k-\mu^{(i)}\right)}{2\left(L_k-\mu^{(i)}\right)}\|x_{k-1}-v_{k-1}\|^2\\
&\quad+\frac{\alpha_k(\mu-\mu^{(i)})}{2}\|\bar{x}-v_{k-1}\|^2+\frac{\alpha_{k-1}^2L_{k-1}(1-\alpha_k)}{2}\|\bar{x}-v_{k-1}\|^2\\
&= (1-\alpha_k)\left(F_i(x_{k-1})-F_i(\bar{x})+\frac{\alpha_{k-1}^2L_{k-1}}{2}\|v_{k-1}-\bar{x}\|^2\right)\\
&\quad-\left(F_i(x_k)-F_i(\bar{x})+\frac{L_k\alpha_k^2}{2}\|\bar{x}-v_k\|^2\right)\\
&\quad+\frac{\alpha_k(\mu-\mu^{(i)})}{2}\|\bar{x}-v_{k-1}\|^2+\frac{(\alpha_k-1)\mu^{(i)}\left(L_k\alpha_k-\mu^{(i)}\right)}{2\left(L_k-\mu^{(i)}\right)}\|x_{k-1}-v_{k-1}\|^2.
\end{aligned}$$

Recall that $i=I_k$ is the random variable from Definition 1.16. Rearranging the last expression in the above equality chain can be conveniently written as

{ineq:snapg2-one-step-presult1}

$$\begin{aligned}
&F_{I_k}(x_k)-F_{I_k}(\bar{x})+\frac{L_k\alpha_k^2}{2}\|\bar{x}-v_k\|^2\\
&\le(1-\alpha_k)\left(F_{I_k}(x_{k-1})-F_{I_k}(\bar{x})+\frac{\alpha_{k-1}^2L_{k-1}}{2}\|v_{k-1}-\bar{x}\|^2\right)\\
&\quad+\frac{\alpha_k(\mu-\mu^{(I_k)})}{2}\|\bar{x}-v_{k-1}\|^2+\frac{(\alpha_k-1)\mu^{(I_k)}\left(L_k\alpha_k-\mu^{(I_k)}\right)}{2\left(L_k-\mu^{(I_k)}\right)}\|x_{k-1}-v_{k-1}\|^2.
\end{aligned}\tag{1.2}$$

Recall $\mathbb{E}_k$ denotes the conditional expectation on $I_0,I_1,\ldots,I_{k-1}$. Taking the conditional

expectation on the LHS of the (1.2) yields:

$$\begin{aligned}
&\mathbb{E}_k\left[F_{I_k}(x_k) - F_{I_k}(\bar{x}) + \frac{L_k\alpha_k^2}{2}\|\bar{x} - v_k\|^2\right] \\
&\underset{\text{(h)}}{=} \mathbb{E}_k\left[F_{I_k}(x_k)\right] - F(\bar{x}) + \mathbb{E}_k\left[\frac{L_k\alpha_k^2}{2}\|\bar{x} - v_k\|^2\right].
\end{aligned}$$

On the RHS of (1.2), using the linearity property while taking the conditional expectation yields:

$$\begin{aligned}
&\mathbb{E}_k\left[(1-\alpha_k)\left(F_{I_k}(x_{k-1}) - F_{I_k}(\bar{x}) + \frac{\alpha_{k-1}^2 L_{k-1}}{2}\|v_{k-1} - \bar{x}\|^2\right)\right] \\
&\quad + \mathbb{E}_k\left[\frac{\alpha_k(\mu - \mu^{(I_k)})}{2}\|\bar{x} - v_{k-1}\|^2\right] + \mathbb{E}_k\left[\frac{(\alpha_k - 1)\mu^{(I_k)}\left(L_k\alpha_k - \mu^{(I_k)}\right)}{2\left(L_k - \mu^{(I_k)}\right)}\|x_{k-1} - v_{k-1}\|^2\right] \\
&\underset{(1)}{=} (1-\alpha_k)\left(\mathbb{E}_k\left[F_{I_k}(x_{k-1})\right] - \mathbb{E}_k[F_{I_k}(\bar{x})] + \mathbb{E}_k\left[\frac{\alpha_{k-1}^2 L_{k-1}}{2}\right]\|v_{k-1} - \bar{x}\|^2\right) \\
&\quad + \mathbb{E}_k\left[\frac{\alpha_k(\mu - \mu^{(I_k)})}{2}\right]\|\bar{x} - v_{k-1}\|^2 + \mathbb{E}_k\left[\frac{(\alpha_k - 1)\mu^{(I_k)}\left(L_k\alpha_k - \mu^{(I_k)}\right)}{2\left(L_k - \mu^{(I_k)}\right)}\right]\|x_{k-1} - v_{k-1}\|^2. \\
&\underset{\text{(h)}}{=} (1-\alpha_k)\left(\mathbb{E}_k\left[F_{I_k}(x_{k-1})\right] - F(\bar{x}) + \mathbb{E}_k\left[\frac{\alpha_{k-1}^2 L_{k-1}}{2}\right]\|v_{k-1} - \bar{x}\|^2\right) \\
&\quad + \mathbb{E}_k\left[\frac{\alpha_k(\mu - \mu^{(I_k)})}{2}\right]\|\bar{x} - v_{k-1}\|^2 + \mathbb{E}_k\left[\frac{(\alpha_k - 1)\mu^{(I_k)}\left(L_k\alpha_k - \mu^{(I_k)}\right)}{2\left(L_k - \mu^{(I_k)}\right)}\right]\|x_{k-1} - v_{k-1}\|^2 \\
&\underset{(2)}{=} (1-\alpha_k)\left(\mathbb{E}_k\left[F_{I_k}(x_{k-1})\right] - F(\bar{x}) + \mathbb{E}_k\left[\frac{\alpha_{k-1}^2 L_{k-1}}{2}\right]\|v_{k-1} - \bar{x}\|^2\right) \\
&\quad + \mathbb{E}_k\left[\frac{(\alpha_k - 1)\mu^{(I_k)}\left(L_k\alpha_k - \mu^{(I_k)}\right)}{2\left(L_k - \mu^{(I_k)}\right)}\right]\|x_{k-1} - v_{k-1}\|^2
\end{aligned}$$

We note that at label (1), we used the fact that α_k is a constant and, x_{k-1}, v_{k-1} only depends on random variable $I_0, I_1, \ldots, I_{k-1}$ hence it falls out of the conditional expectation $\mathbb{E}_k$. At label (2), we used assumption (Assumption 1.14) that the averages of all the $\mu^{(I_k)}$ on each F_{I_k} equals to μ hence, the expectation evaluates to zero by linearity of the expected value operator.

Combing the above results on the expectation of RHS, and LHS of (1.2), we have the one

step inequaity in expectation:

$$\begin{aligned}
&\mathbb{E}_k\left[F_{I_k}(x_k)\right] - F(\bar{x}) + \mathbb{E}_k\left[\frac{L_k\alpha_k^2}{2}\|\bar{x}-v_k\|^2\right] \\
&\leq (1-\alpha_k)\left(\mathbb{E}_k\left[F_{I_k}(x_{k-1})\right] - F(\bar{x}) + \mathbb{E}_k\left[\frac{\alpha_{k-1}^2 L_{k-1}}{2}\right]\|v_{k-1}-\bar{x}\|^2\right) \\
&\quad + \mathbb{E}_k\left[\frac{(\alpha_k-1)\mu^{(I_k)}\left(L_k\alpha_k-\mu^{(I_k)}\right)}{2\left(L_k-\mu^{(I_k)}\right)}\right]\|x_{k-1}-v_{k-1}\|^2.
\end{aligned}$$

Finally, we show the base case. When $k=0$, by assumption it had $\alpha_0=1$ hence τ_0 in Definition 1.16 has $\tau_0=0$ which makes $y_0=v_{-1}=x_{-1}$.

Proof of (d). From Definition 1.16, it has

$$(1+\tau_k)^{-1} = \left(1+\frac{L_k(1-\alpha_k)}{L_k\alpha_k-\mu^{(i)}}\right)^{-1} = \left(\frac{L_k\alpha_k-\mu^{(i)}+L_k(1-\alpha_k)}{L_k\alpha_k-\mu^{(i)}}\right)^{-1} = \frac{L_k\alpha_k-\mu^{(i)}}{L_k-\mu^{(i)}}.$$

Therefore, for all $k\geq 0$, y_k has

$$\begin{aligned}
0 &= (1+\tau_k)^{-1}v_{k-1} + \tau_k(1+\tau_k)^{-1}x_{k-1} - y_k \\
&= \frac{L_k\alpha_k-\mu^{(i)}}{L_k-\mu^{(i)}}\left(v_{k-1}+\frac{L_k(1-\alpha_k)}{L_k\alpha_k-\mu^{(i)}}x_{k-1}\right) - y_k \\
&= \frac{L_k\alpha_k-\mu^{(i)}}{L_k-\mu^{(i)}}v_{k-1} + \frac{L_k(1-\alpha_k)}{L_k-\mu^{(i)}}x_{k-1} - y_k \\
&= \frac{L_k\alpha_k-\mu^{(i)}}{L_k-\mu^{(i)}}v_{k-1} + (1-\alpha_k)x_{k-1} + \left(\frac{L_k(1-\alpha_k)}{L_k-\mu^{(i)}} - (1-\alpha_k)\right)x_{k-1} - y_k \\
&= \frac{L_k\alpha_k-\mu^{(i)}}{L_k-\mu^{(i)}}v_{k-1} + (1-\alpha_k)x_{k-1} + (1-\alpha_k)\left(\frac{L_k-L_k+\mu^{(i)}}{L_k-\mu^{(i)}}\right)x_{k-1} - y_k \\
&= \frac{L_k\alpha_k-\mu^{(i)}}{L_k-\mu^{(i)}}v_{k-1} + (1-\alpha_k)x_{k-1} + \frac{\mu^{(i)}(1-\alpha_k)}{L_k-\mu^{(i)}}x_{k-1} - y_k.
\end{aligned}$$

Therefore, we establish the equality

$$(1-\alpha_k)x_{k-1} - y_k = -\frac{L_k\alpha_k-\mu^{(i)}}{L_k-\mu^{(i)}}v_{k-1} - \frac{\mu^{(i)}(1-\alpha_k)}{L_k-\mu^{(i)}}x_{k-1}.$$

On the second equality below, we will the above equality, it goes:

$$\begin{aligned}
z_k - y_k &= \alpha_k \bar{x} + (1-\alpha_k)x_{k-1} - y_k \\
&= \alpha_k \bar{x} - \frac{L_k\alpha_k - \mu^{(i)}}{L_k - \mu^{(i)}} v_{k-1} - \frac{\mu^{(i)}(1-\alpha_k)}{L_k - \mu^{(i)}} x_{k-1} \\
&= \frac{L_k\alpha_k - \mu^{(i)}}{L_k - \mu^{(i)}}(\bar{x} - v_{k-1}) + \left(\alpha_k - \frac{L_k\alpha_k - \mu^{(i)}}{L_k - \mu^{(i)}}\right)\bar{x} - \frac{\mu^{(i)}(1-\alpha_k)}{L_k - \mu^{(i)}} x_{k-1} \\
&= \frac{L_k\alpha_k - \mu^{(i)}}{L_k - \mu^{(i)}}(\bar{x} - v_{k-1}) + \left(\frac{\alpha_k L_k - \alpha_k\mu^{(i)} - L_k\alpha_k + \mu^{(i)}}{L_k - \mu^{(i)}}\right)\bar{x} - \frac{\mu^{(i)}(1-\alpha_k)}{L_k - \mu^{(i)}} x_{k-1} \\
&= \frac{L_k\alpha_k - \mu^{(i)}}{L_k - \mu^{(i)}}(\bar{x} - v_{k-1}) + \frac{\mu^{(i)}(1-\alpha_k)}{L_k - \mu^{(i)}}\bar{x} - \frac{\mu^{(i)}(1-\alpha_k)}{L_k - \mu^{(i)}} x_{k-1} \\
&= \frac{L_k\alpha_k - \mu^{(i)}}{L_k - \mu^{(i)}}(\bar{x} - v_{k-1}) + \frac{\mu^{(i)}(1-\alpha_k)}{L_k - \mu^{(i)}}(\bar{x} - x_{k-1}).
\end{aligned}$$

Proof of (e). From Definition 1.16 it has directly:

$$\begin{aligned}
z_k - x_k &= \alpha_k \bar{x} + (1-\alpha_k)x_{k-1} - x_k \\
&= \alpha_k \bar{x} + x_{k-1} - x_k - \alpha_k x_{k-1} \\
&= \alpha_k(\bar{x} - \alpha_k^{-1}(x_k - x_{k-1}) - x_{k-1}) \\
&= \alpha_k(\bar{x} - v_k).
\end{aligned}$$

Proof of (f). The proof is direct algebra and, it has:

$$\begin{aligned}
&\frac{\left(\mu^{(i)}\right)^2(1-\alpha_k)^2}{2(L_k - \mu^{(i)})} - \frac{\mu^{(i)}\alpha_k(1-\alpha_k)}{2} \\
&= \frac{1}{2\left(L_k - \mu^{(i)}\right)}\left(\left(\mu^{(i)}\right)^2(1-\alpha_k)^2 - \left(L_k - \mu^{(i)}\right)\mu^{(i)}\alpha_k(1-\alpha_k)\right) \\
&= \frac{1-\alpha_k}{2\left(L_k - \mu^{(i)}\right)}\left(\left(\mu^{(i)}\right)^2 - \left(\mu^{(i)}\right)^2\alpha_k - \left(L_k\mu^{(i)}\alpha_k - \left(\mu^{(i)}\right)^2\alpha_k\right)\right) \\
&= \frac{1-\alpha_k}{2(L_k - \mu)}\left(\left(\mu^{(i)}\right)^2 - L_k\left(\mu^{(i)}\right)\alpha_k\right) \\
&= \frac{(1-\alpha_k)\mu^{(i)}\left(\mu^{(i)} - L_k\alpha_k\right)}{2\left(L_k - \mu^{(i)}\right)} \\
&= \frac{(\alpha_k - 1)\mu^{(i)}\left(L_k\alpha_k - \mu^{(i)}\right)}{2\left(L_k - \mu^{(i)}\right)}.
\end{aligned}$$

Proof of (g). From the property of the α_k sequence stated in item (c), we have:

$$
\begin{aligned}
&\frac{\left(L_k\alpha_k - \mu^{(i)}\right)^2}{2(L_k - \mu^{(i)})} - \frac{\alpha_{k-1}^2 L_{k-1}(1-\alpha_k)}{2} \\
&= \frac{\left(L_k\alpha_k - \mu^{(i)}\right)^2}{2(L_k - \mu^{(i)})} - \frac{L_k\alpha_k(\alpha_k - \mu/L_k)}{2} \\
&= \frac{\left(L_k\alpha_k - \mu^{(i)}\right)^2}{2(L_k - \mu^{(i)})} - \frac{L_k\alpha_k(\alpha_k - \mu^{(i)}/L_k)}{2} + \frac{L_k\alpha_k(\alpha_k - \mu^{(i)}/L_k)}{2} - \frac{L_k\alpha_k(\alpha_k - \mu/L_k)}{2} \\
&= \frac{\left(L_k\alpha_k - \mu^{(i)}\right)^2}{2(L_k - \mu^{(i)})} - \frac{\alpha_k\left(L_k\alpha_k - \mu^{(i)}\right)}{2} + \frac{L_k\alpha_k}{2}\frac{\left(\mu - \mu^{(i)}\right)}{L_k} \\
&= \frac{L_k\alpha_k - \mu^{(i)}}{2(L_k - \mu^{(i)})}\left(L_k\alpha_k - \mu^{(i)} - \left(L_k - \mu^{(i)}\right)\alpha_k\right) + \frac{\alpha_k(\mu - \mu^{(i)})}{2} \\
&= \frac{L_k\alpha_k - \mu^{(i)}}{2(L_k - \mu^{(i)})}\left(\mu^{(i)}\alpha_k - \mu^{(i)}\right) + \frac{\alpha_k(\mu - \mu^{(i)})}{2} \\
&= \frac{\left(L_k\alpha_k - \mu^{(i)}\right)\mu^{(i)}\left(\alpha_k - 1\right)}{2(L_k - \mu^{(i)})} + \frac{\alpha_k(\mu - \mu^{(i)})}{2}.
\end{aligned}
$$

■

1.3 So, what to do next?

Hi Arron would you like to add me for the co-authorship to continue this line of work and see how Nesterov's Accelerated Technique may work out for the stochastic gradient method? These results are solid results but they are still partial results and, below are the potential I foresee for this these ideas.

(i) Narrow down the sequence α_k and make sure that it can allow the quantity:

$$
\mathbb{E}_k\left[\frac{(\alpha_k - 1)\mu^{(I_k)}\left(L_k\alpha_k - \mu^{(I_k)}\right)}{2\left(L_k - \mu^{(I_k)}\right)}\right]\|x_{k-1} - v_{k-1}\|^2
$$

is negative, or at least bounded. I am not sure how this will work out, but I have some solid ideas aroud it.

(ii) Roll up the inequality in Theorem 1.18 recursively and, determine the convergence rate through α_k that makes the previous item true. In addition, I have the hunches that the convergence reate involves the variance of $\mu^{(I_k)}$ and, it will slower than the non-stochastic case of the algorithm.

For the future we can:

(i) Extend the definition of strong convexity to quasi-norm.

(ii) Show the convergence of the method for objective function that is not necessarily strong convex, but quasi-strongly convex. This is a much weaker assumption it works well in practice for well known problems.

References

[1] A. Beck, *First-order Methods in Optimization*, MOS-SIAM Series in Optimization, SIAM, 2017.

[2] H. Li and X. Wang, *Relaxed Weak Accelerated Proximal Gradient Method: a Unified Framework for Nesterov's Accelerations*, Apr. 2025. arXiv:2504.06568 [math].

[3] Y. Nesterov, *Lectures on Convex Optimization*, vol. 137 of Springer Optimization and Its Applications, Springer International Publishing, 2018.