

Linear Convergence of Stochastic Nesterov's Accelerated Proximal Gradient method under Interpolation Hypothesis

Author *

July 4, 2025

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

Abstract

This file is for communication purposes between collaborators.

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 $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is an L Lipschitz smooth and $\mu \geq 0$ strongly convex function. The function $g : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ 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 $f : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ be a differentiable function. Then, for all the Bregman divergence $D_f : \mathbb{R}^n \times \operatorname{dom} \nabla f \rightarrow \mathbb{R}$ 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.
{thm:pg-ineq}

Theorem 1.9 (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 1.10 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 ?? with some line search routine. Each of the reference books gives a proof of the theorem. But for the best consistency in notations, see Theorem 2.3 in Li and Wang [2].

{thm:jesen}

Theorem 1.11 (Jensen’s inequality) *Let $F : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ 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 1.12 If x, y is out of $\text{dom } F$, the inequality still work by convexity.

1.2 Stochastic accelerated proximal gradient

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

{ass:sum-of-many}

Assumption 1.13 (sum of many) Define $F := (1/n) \sum_{i=1}^n F_i$ where each $F_i = f_i + g_i$. Assume that for all $i = 1, \dots, n$, each $f_i : \mathbb{R}^n \rightarrow \mathbb{R}$ are $K^{(i)}$ smooth and $\mu^{(i)} \geq 0$ strongly convex function such that $K^{(i)} > \mu^{(i)}$ and, $g_i : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ is a closed convex proper function. Consequently, the function f can be written as $F = g + f$ with $f = (1/n) \sum_{i=1}^n f_i$, $g = (1/n) \sum_{i=1}^n g_i$ therefore, it also satisfies Assumption 1.1 with $L = (1/n) \sum_{i=1}^n K^{(i)}$ and $\mu = (1/n) \sum_{i=1}^n \mu^{(i)}$.

This assumption is stronger than Assumption 1.1. It still appears in practice, for example if F_i are all indicator function of convex set, then it solves feasibility problem $\bigcap_{i=1}^n C_i$ and, in this case, the proximal gradient operator becomes a projection onto the convex set C_i . In practice, each of the strong convexity constant $\mu^{(i)}$ may not be easily accessible. And we further note that if $\mu > 0$ strongly convex, then there exists at least one $\mu^{(i)} \geq 0$.

The interpolation hypothesis from Machine Learning stated that the model has the capacity to perfect fit all the observed data. The following assumption state the interpolation hypothesis in our context.

{ass:interp-hypothesis}

Assumption 1.14 (interpolation hypothesis) Suppose that $F := (1/n) \sum_{i=1}^n F_i$ satisfying Assumption 1.13. 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 $i = 1, \dots, n$ it satisfies $0 = f_i(\bar{x})$.

{def:snapg-v2}

Definition 1.15 (SNAPG-V2) *Let F satisfies Assumption 1.13. Let $(I_k)_{k \geq 0}$ be a list of i.i.d random variables uniformly sampled from set $\{0, 1, 2, \dots, n\}$. 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:

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

{lemma:snapg-v2-seq-range}

Lemma 1.16 (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 satisfies recursively the equality:

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

And, the following items are true:

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

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

- (ii) ...

This part is not finished yet.

{thm:snapg2-one-step}

Theorem 1.17 (SNAPG-V2 one step convergence) *Let F satisfies assumption 1.14. Suppose that an algorithm satisfying Definition 1.15 uses this F . Let \mathbb{E}_k denotes the expectation conditioned on I_0, I_1, \dots, I_{k-1} . Then, for all $k \geq 1$, it has the following inequality*

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

And for $k = 0$, it has

$$\mathbb{E} [F_{I_0}] - F(\bar{x}) + \frac{L_0}{2} \mathbb{E} [\|\bar{x} - x_0\|^2] \leq \frac{L_0 - \mu}{2} \|\bar{x} - v_{-1}\|^2.$$

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 1.9 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 1.11 with $z = z_k$ on F_i .
- (c) The sequence $(\alpha_k)_{k \geq 0}$ has $(L_{k-1}/L_k)(1 - \alpha_k)\alpha_{k-1}^2 = \alpha_k(\alpha_k - \mu/L_k)$. It is a special case of Definition 1.7 with $\rho_{k-1} = L_{k-1}/L_k$.
- (d) From Definition 1.15 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 1.15 it has: $(\forall k \geq 1) \ z_k - x_k = \alpha_k(\bar{x} - v_k)$.
- (f) Using direct algebra, we have for all $k \geq 1$:

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

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

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

(h) Because we assumed interpolation hypothesis in Assumption 1.14, 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:

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

And we have the following chain of equalities:

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

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

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

$$\begin{aligned}
0 & \leq \alpha_k F_i(\bar{x}) + (1 - \alpha_k) F_i(x_{k-1}) - F_i(x_k) \\
& - \frac{L_k}{2} \|z_k - x_k\|^2 + \frac{(\alpha_k - 1) \mu^{(i)} (L_k \alpha_k - \mu^{(i)})}{2 (L_k - \mu^{(i)})} \|x_{k-1} - v_{k-1}\|^2 \\
& + \frac{\alpha_k (\mu - \mu^{(i)})}{2} \|\bar{x} - v_{k-1}\|^2 + \frac{\alpha_{k-1}^2 L_{k-1} (1 - \alpha_k)}{2} \|\bar{x} - v_{k-1}\|^2 \\
& \stackrel{(e)}{=} \alpha_k F_i(\bar{x}) + (1 - \alpha_k) F_i(x_{k-1}) - F_i(x_k) \\
& - \frac{L_k \alpha_k^2}{2} \|\bar{x} - v_k\|^2 + \frac{(\alpha_k - 1) \mu^{(i)} (L_k \alpha_k - \mu^{(i)})}{2 (L_k - \mu^{(i)})} \|x_{k-1} - v_{k-1}\|^2 \\
& + \frac{\alpha_k (\mu - \mu^{(i)})}{2} \|\bar{x} - v_{k-1}\|^2 + \frac{\alpha_{k-1}^2 L_{k-1} (1 - \alpha_k)}{2} \|\bar{x} - v_{k-1}\|^2 \\
& = (\alpha_k - 1) F_i(\bar{x}) + (1 - \alpha_k) F_i(x_{k-1}) - F_i(x_k) + F_i(\bar{x}) \\
& - \frac{L_k \alpha_k^2}{2} \|\bar{x} - v_k\|^2 + \frac{(\alpha_k - 1) \mu^{(i)} (L_k \alpha_k - \mu^{(i)})}{2 (L_k - \mu^{(i)})} \|x_{k-1} - v_{k-1}\|^2 \\
& + \frac{\alpha_k (\mu - \mu^{(i)})}{2} \|\bar{x} - v_{k-1}\|^2 + \frac{\alpha_{k-1}^2 L_{k-1} (1 - \alpha_k)}{2} \|\bar{x} - v_{k-1}\|^2
\end{aligned}$$

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

Recall that $i = I_k$ is the random variable from Definition 1.15. Rearranging the last expression in the above equality chain can be conveniently written as

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

Recall \mathbb{E}_k denotes the conditional expectation on I_0, I_1, \dots, I_{k-1} . Taking the conditional expectation on the LHS of the (1.2) yields:

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

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

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

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

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, \dots, I_{k-1} hence it falls out of the conditional expectation \mathbb{E}_k . At label (2), we used assumption (Assumption 1.13) 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.

Combining the above results on the expectation of RHS, and LHS of (1.2), we have the one-step inequality in expectation:

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

Finally, we show the base case. When $k = 0$, by assumption it had $\alpha_0 = 1$ hence τ_0 in Definition 1.15 has $\tau_0 = 0$ which makes $y_0 = v_{-1} = x_{-1}$. Therefore, it makes $x_0 = T_{L_0}(y_0|F_{I_0}) = T_{L_0}(v_{-1}|F_{I_0})$. Similarly, it has also $z_0 = \bar{x}$. Applying Theorem 1.9 with $z = z_0$ and, assume a successful line search with L_0 , it yields:

$$\begin{aligned}
0 & \leq F_{I_0}(z_0) - F_{I_0}(x_0) - \frac{L_0}{2} \|z_0 - x_0\|^2 + \frac{L_0 - \mu^{(I_0)}}{2} \|z_0 - y_0\|^2 \\
& = F_{I_0}(\bar{x}) - F_{I_0}(x_0) - \frac{L_0}{2} \|\bar{x} - x_0\|^2 + \frac{L_0 - \mu^{(I_0)}}{2} \|\bar{x} - v_{-1}\|^2.
\end{aligned}$$

Re-arranging and taking the expectation it yields:

$$\begin{aligned}
\mathbb{E} \left[F_{I_0}(x_0) - F_{I_0}(\bar{x}) + \frac{L_0}{2} \|\bar{x} - x_0\|^2 \right] & \stackrel{(h)}{=} \mathbb{E} [F_{I_0}] - F(\bar{x}) + \frac{L_0}{2} \mathbb{E} [\|\bar{x} - x_0\|^2] \\
& \leq \frac{L_0 - \mathbb{E} [\mu^{(I_0)}]}{2} \|\bar{x} - v_{-1}\|^2 \\
& = \frac{L_0 - \mu}{2} \|\bar{x} - v_{-1}\|^2.
\end{aligned}$$

Proof of (d). From Definition 1.15, 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 use 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 1.15 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:

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

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

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

■

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:

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

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

- (ii) Roll up the inequality in Theorem 1.17 recursively and, determine the convergence rate through α_k that makes the previous item true. In addition, I have the hunches that the convergence rate 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 relative strong convexity with respect to a quasi-norm. This would extend interpolation hypothesis in Assumption 1.14 where, even if $\mu > 0$, it doesn't mean that F has a unique solution through strong convexity. This is entirely possible and appeared in the literatures before so, I can give you the words of confidence.
- (ii) Show the convergence of the method for objective function based on quasi-strong convexity. This is a much weaker assumption it works well in practice for the common known problems in convex programming.

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.